Biological and Medical Physics, Biomedical Engineering

Nikolay Dokholyan *Editor*

Computational Modeling of Biological Systems

From Molecules to Pathways

Springer

BIOLOGICAL AND MEDICAL PHYSICS, BIOMEDICAL ENGINEERING

BIOLOGICAL AND MEDICAL PHYSICS, BIOMEDICAL ENGINEERING

The fields of biological and medical physics and biomedical engineering are broad, multidisciplinary and dynamic. They lie at the crossroads of frontier research in physics, biology, chemistry, and medicine. The Biological and Medical Physics, Biomedical Engineering Series is intended to be comprehensive, covering a broad range of topics important to the study of the physical, chemical and biological sciences. Its goal is to provide scientists and engineers with textbooks, monographs, and reference works to address the growing need for information.

Books in the series emphasize established and emergent areas of science including molecular, membrane, and mathematical biophysics; photosynthetic energy harvesting and conversion; information processing; physical principles of genetics; sensory communications; automata networks, neural networks, and cellular automata. Equally important will be coverage of applied aspects of biological and medical physics and biomedical engineering such as molecular electronic components and devices, biosensors, medicine, imaging, physical principles of renewable energy production, advanced prostheses, and environmental control and engineering.

For further volumes:
http://www.springer.com/series/3740

Nikolay Dokholyan
Editor

Computational Modeling of Biological Systems

From Molecules to Pathways

Springer

Editor
Nikolay Dokholyan
Department of Biochemistry and Biophysics
University of North Carolina at Chapel Hill
Genetics Medicine
120 Mason Farm Road
Chapel Hill, NC 27599-7260, USA
dokh@med.unc.edu

ISSN 1618-7210
ISBN 978-1-4614-2145-0 e-ISBN 978-1-4614-2146-7
DOI 10.1007/978-1-4614-2146-7
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2012930648

Contents

Part I
Molecular Modeling

Introduction to Molecular Dynamics: Theory and Applications in Biomolecular Modeling

Yi Wang and J. Andrew McCammon

1 Introduction

Since the first molecular dynamics (MD) simulation of a protein was performed over 30 years ago [87], MD has been used to study a variety of biomolecular systems, including proteins, nucleotides, lipid bilayers, and carbohydrates [16, 64, 88, 101]. Today, the problems tackled by MD range from large conformational changes in proteins to free energy differences associated with subtle modifications in ligands [46, 62, 65, 127]. Since the high spatial and temporal resolution of MD is rarely achieved in conventional experimental techniques, MD is increasingly used in combination with various experimental methods to provide a multiscale description of the structure, dynamics, and function of a biomolecule.

In a nutshell, MD is a method to integrate the classical (Newtonian) equations of motion for a set of particles [5, 38]. The result is a trajectory of the system over a certain period of time, usually tens to hundreds of nanoseconds. Various structural and dynamic properties of the system can then be calculated from the trajectory, some of which may be directly compared with experimental results. In Fig. 1, we have shown a typical simulation system, consisting of a protein surrounded by solvent water molecules. The system is used to study the enzyme neuraminidase from the avian influenza virus H5N1 and was simulated for 100 ns [78]. At each step of this 100-ns simulation, the force "felt" by every atom is calculated according to a predefined potential energy function. It is then used to solve the equations of

Y. Wang • J.A. McCammon (✉)
Center for Theoretical Biological Physics, Howard Hughes Medical Institute,
Department of Chemistry and Biochemistry, Department of Pharmacology,
University of California, San Diego, La Jolla, CA 92093, USA
e-mail: yiwang@ucsd.edu; jmccammon@mail.ucsd.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_1,

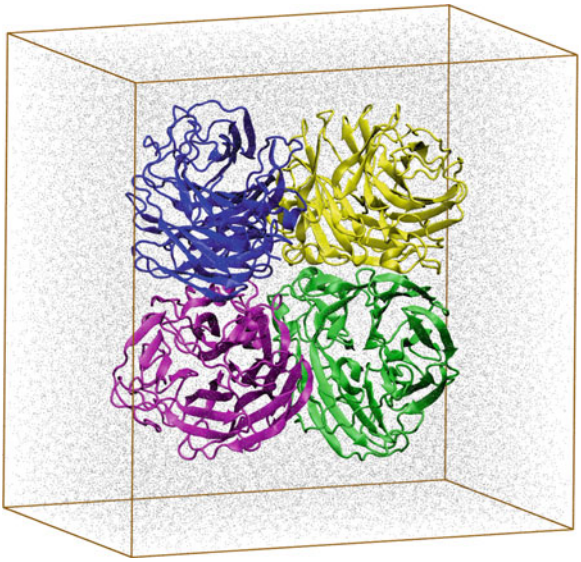

Fig. 1 The simulation system of the neuraminidase tetramer from the avian influenza virus H5N1. The four monomers of the neuraminidase are colored *blue*, *yellow*, *pink* and *green*, respectively. Water molecules are shown as *gray dots*, and the boundaries of the simulation box are highlighted. Figure was created using structures from Lawrenz et al. [78]

motion and generate the new velocity and position of the atom for the next step. The 100-ns trajectory is obtained by repeating the above calculation 5×10^7 times. We will discuss these calculations in more detail in Sect. 3.

Theory and development of MD are deeply rooted in the principles of statistical mechanics. Although users today normally do not need to write their own MD code, it is still very helpful to understand these principles, which can be essential to ensure the proper applications of the method on various complex biomolecular systems. In this chapter, we will introduce the basic statistical mechanics background of MD, the various components of a potential energy function, and the algorithm used to integrate the equations of motion. We will then give some practical examples of MD, followed by a few tips on how to avoid common pitfalls in the preparation of a simulation. In the last section, we will briefly introduce some advanced simulation techniques, such as free energy calculation and enhanced sampling methods. Our goal is to give an overview of MD, rather than discussing any specific aspect of the method in great detail. Therefore, we will provide references to important theories and applications throughout the text for readers to further explore the corresponding topics.

2 Statistical Mechanics Background

MD simulations can generate a very detailed picture of the system under study, i.e., they allow the calculation of microscopic properties, such as positions and velocities of each individual atom in the system. These microscopic properties, however, are

often of less interest or practical value to us than the macroscopic properties of a system, which are the only properties that most experiments can measure. For instance, knowing the exact locations of individual water molecules in a simulation box may be less interesting or important than knowing the average rate at which they are conducted by a channel protein across a lipid membrane [55, 120].

The relationship between microscopic and macroscopic properties of a system is the subject of statistical mechanics. On this topic, many excellent reference books are available [22, 68, 89], including some that offered discussions in the specific context of computer simulations and biomolecular systems [5, 34, 38]. Below, we will give a brief overview of some of the key concepts, and we encourage the readers to the aforementioned books for more information.

2.1 Microstates and the Ensemble Theory

A microscopic state of a system is specified by the positions and momenta of all particles in the system. For a system with N particles, we may write its Hamiltonian H as a sum of the kinetic energy K and the potential energy V, which are functions of the Cartesian momentum $\mathbf{p}_i$ and coordinate $\mathbf{r}_i$ of each particle i, respectively:

$$\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N), \tag{1}$$

$$\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_N), \tag{2}$$

$$H(\mathbf{r}, \mathbf{p}) = K(\mathbf{p}) + V(\mathbf{r}). \tag{3}$$

Usually, the kinetic energy takes the familiar quadratic form:

$$K = \sum_{i=1}^{N} \frac{1}{2m_i}(p_{ix}^2 + p_{iy}^2 + p_{iz}^2), \tag{4}$$

where m_i is the mass of particle i, and p_{ix}, p_{iy}, p_{iz} are the x, y, and z components of its momentum $\mathbf{p}_i$. The potential energy V has a much more complicated form and will be discussed in more detail later. For now, it suffices to say that once the form of V is determined, the time evolution of the system, governed by the Hamiltonian H, can be determined by solving the equations of motion in a MD simulation. If we think of the positions and momenta of all particles in the system as coordinates in a $6N$-dimension space, which we refer to as the phase space, then at any given time, the system corresponds to a point in this multidimensional space. The time evolution of the system, thereby, corresponds to a trajectory in the phase space.

As mentioned earlier, we are interested in calculating certain macroscopic properties of the system. Instead of following the trajectory of a single system in the phase space, the conventional approach used in statistical mechanics is to consider, at any given time, a collection of systems with the same macroscopic properties,

e.g., number of particles N, volume V, energy E, temperature T, or pressure P. These systems have the same Hamiltonian, and, therefore, can be considered as replicas of each other. Each of them corresponds to a point in the phase space, and the collection of these points constitutes a statistical ensemble. We can calculate an observable A by averaging its values in all the members of the ensemble,

$$A_{\text{obs}} = \langle A \rangle_{\text{ens}}, \tag{5}$$

where $\langle A \rangle_{\text{ens}}$ is called an ensemble average of A.

In the rest of this section, we will introduce the equations used to calculate ensemble averages. Although the derivations of these equations are beyond the scope of this chapter, it is probably worthwhile to mention, albeit very briefly, a key assumption behind them. In general, at any given time, there will be many different microscopic states corresponding to a particular set of macroscopic conditions. In other words, the collection of systems with the same macroscopic properties will form a hypersurface in the $6N$-dimension phase space. As an example, if we specify that the total energy of the system is E, there are multiple ways we can distribute E to the N particles in the system. If we do not specify any other conditions, we should be able to perform such distributions in any possible way, i.e., there is no reason to prefer one distribution scheme over another. Likewise, a fundamental assumption in statistical mechanics is that when there are no other constraints, the system is equally likely to be in any one of the microstates corresponding to a macrostate. This "equal a priori probabilities" postulate forms the backbone of statistical mechanics, and many powerful equations can be derived from this simple but highly nontrivial assumption.

2.2 The Ensemble Average

We will use the canonical (*NVT*) ensemble to demonstrate how the ensemble average of an observable can be calculated. Treatment of other statistical ensembles, such as the microcanonical (*NVE*), the isothermal–isobaric (*NPT*), and the grand canonical (μVT) ensemble, can be found in the reference books mentioned earlier. The quantities in parentheses represent the thermodynamic properties kept constant in an ensemble, with N, V, E, T, and P defined earlier, and μ standing for the chemical potential of a given species of particles.

A key concept in statistical mechanics is the partition function Q, which has the following form in the canonical ensemble

$$Q_{\text{NVT}} = \sum_{\mathbf{r},\mathbf{p}} \exp(-\beta\, H(\mathbf{r}, \mathbf{p})), \tag{6}$$

where $\beta = 1/k_{\text{B}}T$ and k_{B} stands for the Boltzmann constant. The term $\exp(-\beta\, H(\mathbf{r}, \mathbf{p}))$ is called the "Boltzmann factor," which is related to the probability $\text{P}(\mathbf{r}, \mathbf{p})$

that a certain microstate can be visited in a canonical ensemble. The partition function is a sum of the Boltzmann factors from all microstates, and is used as a normalization factor to give the formula of $\mathrm{P}(\mathbf{r}, \mathbf{p})$

$$\mathrm{P}(\mathbf{r}, \mathbf{p}) = \frac{\exp(-\beta\, H(\mathbf{r}, \mathbf{p}))}{Q_{\mathrm{NVT}}}. \tag{7}$$

In practice, because the Hamiltonian is a sum of the kinetic energy, which only depends on $\mathbf{p}$, and the potential energy, which only depends on $\mathbf{r}$, the partition function Q can be expressed as a product of the kinetic (ideal gas) contribution $Q^{\mathrm{id}}_{\mathrm{NVT}}$ and the potential (excess) contribution $Q^{\mathrm{ex}}_{\mathrm{NVT}}$ [5]. The former can be integrated analytically, leaving the latter our main target of calculation. In reality, instead of $Q^{\mathrm{ex}}_{\mathrm{NVT}}$, we often use the configurational partition function

$$Z_{\mathrm{NVT}} = \sum_{\mathbf{r}} \exp(-\beta\, V(\mathbf{r})). \tag{8}$$

Correspondingly, the probability of visiting a configurational microstate $\mathbf{r}$ is

$$\mathrm{P}(\mathbf{r}) = \frac{\exp(-\beta\, V(\mathbf{r}))}{Z_{\mathrm{NVT}}}. \tag{9}$$

With (9), we can now calculate the ensemble average of any observable A:

$$\langle A \rangle_{\mathrm{NVT}} = \sum_{\mathbf{r}} A(\mathbf{r})\, \mathrm{P}(\mathbf{r}) = \sum_{\mathbf{r}} \frac{A(\mathbf{r})\, \exp(-\beta\, V(\mathbf{r}))}{Z_{\mathrm{NVT}}}. \tag{10}$$

One goal of an MD simulation is to generate the proper phase space distribution according to (9), from which the ensemble average of various observables can be calculated using (10). A somewhat subtle point is that once we have generated the correct phase space distribution, we will be able to calculate an observable A as a time average from a simulation trajectory,

$$A_{\mathrm{obs}} = \langle A \rangle_{\mathrm{time}}. \tag{11}$$

The equivalence of (11) and (5) relies on the so-called "ergodic assumption." Interested readers can find more on this topic in the book by Allen and Tildesley [5].

Before we move on to the next section, where we will discuss how to generate the desired phase space distribution in a MD simulation, we should say a few more words about (8). As the partition function Z_{NVT} contains all the information about the microstates of a system, it's very tempting to evaluate it directly according to (8). However, this remains a daunting task for most biomolecular systems. The reason is that there are too many microstates for a typical biomolecular system, which makes the direct evaluation of (8) unfeasible. To expedite the sampling of the configurational space, various enhanced sampling methods have been developed, and we will discuss some of these methods in Sect. 6.

3 Molecular Dynamics: The Theory

Let's go back to the Hamiltonian H defined in (3), which governs the time evolution of a system. We have shown that the kinetic part of H usually has a simple quadratic form (4). Now it's time to look at the more complicated potential part of H. Unfortunately, the actual form of the potential energy V is so complicated that even with the power of modern computers, some approximations are needed before we can calculate it efficiently. In this section, we will first go through these approximations, and then discuss how to use MD simulations to generate the desired phase space distribution of a statistical ensemble.

3.1 The Force Field

In MD, the specific form of the potential energy function V is given as a force field, where V is broken down into terms characterizing different types of interactions in a bimolecular system. Examples of some of the most commonly used force fields are AMBER [54, 117], CHARMM [69, 79–81], GROMOS [92, 102], and OPLS [29, 63, 96]. Although these force fields were initially developed by different groups, most of them have similar functional forms, i.e., the potential energy V is divided into bonded and nonbonded terms, the former of which includes the bond, angle, dihedral, and improper interaction terms, while the latter includes the Van der Waals (vdW) and electrostatic interaction terms. As an example, energy functions from the CHARMM force field are shown below:

$$V = V_{\text{bond}} + V_{\text{angle}} + V_{\text{dihedral}} + V_{\text{improper}} + V_{\text{vdw}} + V_{\text{elec}}, \tag{12}$$

$$V_{\text{bond}} = K_b(b - b_0)^2, \tag{13}$$

$$V_{\text{angle}} = K_\theta(\theta - \theta_0)^2, \tag{14}$$

$$V_{\text{dihedral}} = K_\chi\big[1 + \cos(n\chi - \delta)\big], \tag{15}$$

$$V_{\text{improper}} = K_\psi(\psi - \psi_0)^2, \tag{16}$$

$$V_{\text{vdw}} = \varepsilon\Big[\left(\frac{R_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{ij}}{r_{ij}}\right)^{6}\Big], \tag{17}$$

$$V_{\text{elec}} = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}. \tag{18}$$

In Fig. 2, we have given the schematic representation of each energy term listed above, along with a plot showing the corresponding energy value for certain selected atom types. While the bond and angle terms are rather straightforward to understand, the rest of the terms may require some explanation. Mathematically, a dihedral and an improper both involve four atoms, and are defined as the angle between the

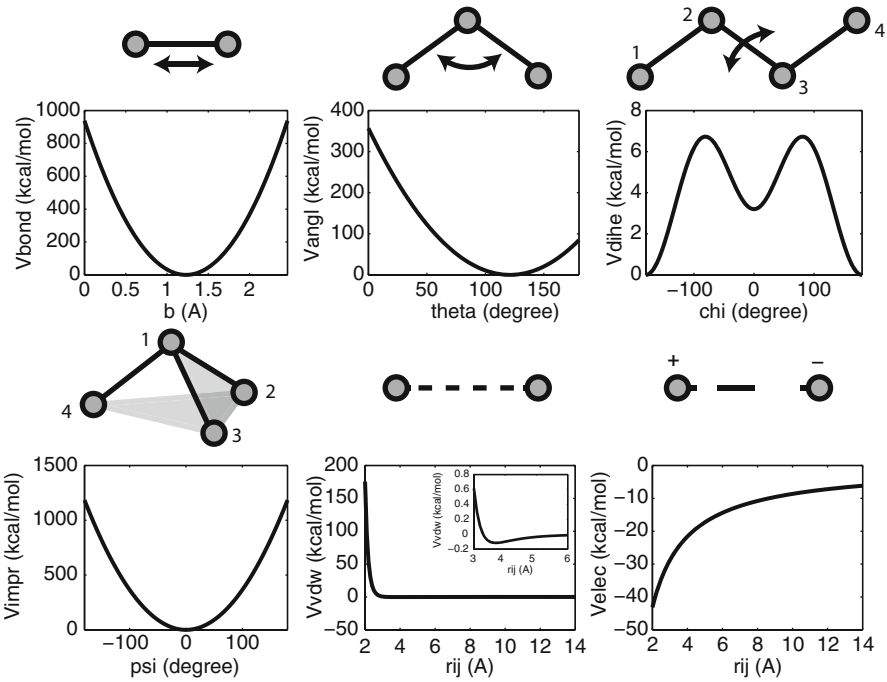

Fig. 2 Potential energy terms in a force field. Schematic representations are shown for the bond, angle, dihedral, improper, vdW and electrostatic interactions. The corresponding energy values for selected atom types in the CHARMM force field are plotted, including the C–O bond, the CA–C–O angle, the CA–C–N–CA (Φ) dihedral and the C–CA–N–O (peptide bond) improper. To demonstrate the nonbonded interactions, we also plotted the vdW and electrostatic energy values for a pair of C–O atoms. The atom names used here are consistent with the naming convention of protein data bank, where CA represents the Cα atom of the protein backbone

plane containing the first three atoms and the plane containing the last three atoms. A dihedral controls the rotation about the bond between the second and the third atom, while an improper controls the "planarity" of the four atoms. For instance, the Φ–Ψ backbone dihedrals of proteins are primarily controlled by dihedral terms, whereas the planarity of a peptide bond (–C(=O)NH–) is controlled by an improper term. As for the nonbonded interactions, the electrostatic term describes the familiar Coulombic interactions between two charged atoms, while the vdW term describes interactions arising from induced dipoles and excluded volumes of pairs of atoms. The vdW potential is attractive at long distance, but quickly becomes repulsive at very short distance between two atoms, the latter of which has the effect of a "hard core" potential and prevents atoms from overlapping with each other.

Among the bonded interactions, the bond, angle and improper terms all have the form of a harmonic potential. Their corresponding spring constants are usually quite large, which means that small changes in the above quantities can result in a huge difference in the corresponding energy. For instance, increasing the length

of the C–O bond in a protein backbone by only 0.5Å from its equilibrium value will increase the bond energy by 155 kcal/mol. Such a drastic change in energy is highly unfavorable in an MD simulation. As a result, the bonds, angles, and most impropers of a system are often found to be very close to their equilibrium values. In comparison, the dihedral potential (and some improper potential) is much "softer". As shown in Fig. 2, the protein backbone dihedral Φ is governed by an energy function in the range of 0–7 kcal/mol. This soft potential allows the backbone dihedral to adopt a broad range of values, and, therefore, gives biomolecules the flexibility to undergo large conformational changes.

3.2 Long-Range Interactions

In most MD simulations, the bonded interaction terms can be calculated rather efficiently, since they only involve atoms connected by one to three covalent bonds. Meanwhile, the nonbonded interactions, which occur between every pair of atoms in a system, are much more expensive to calculate. A closer look at (17) and (18) tells us that the vdW and electrostatic potential functions have different dependence on r. Generally, a potential is considered to be a short-range interaction if it decays faster than $r^{-\mathrm{d}}$, where d is the dimensionality of the system [5]. Under this criterion, the vdW potential is a short-range interaction, since it decays as r^{-6} at long distances, while the electrostatic potential, which decays as r^{-1}, is a long-range interaction.

For the short-range vdW interaction, we can use a cutoff scheme to perform the calculation efficiently: interactions between atoms within the cutoff distance are calculated, while interactions between atoms separated by a distance longer than the cutoff are simply neglected. Typical cutoff distances used in MD simulations are in the range of 8–12Å. The associated approximation is acceptable because a short-range interaction rapidly decays to zero as the distance increases. For instance, the vdW interaction energy between a carbon and an oxygen atom (C–O) is only −0.0002 kcal/mol when they are 12Å apart. Such a small value allows us to truncate the vdW potential at the cutoff distance.

In practice, the simple truncation scheme described above is replaced by a slightly more complicated algorithm, which is needed to avoid a sudden change in the vdW forces caused by the discontinuity in the derivative of the vdW potential at the cutoff distance. In many MD softwares, it is also possible to add a "correction term" to the final result, in order to approximate the neglected vdW potential energy beyond the cutoff distance. We'll leave interested readers to explore these more advanced topics by themselves.

Usually, we cannot calculate the long-range electrostatic potential using the same cutoff scheme described above. This can be seen from the example mentioned earlier—when the C–O atoms are 12Å apart, their Coulombic interaction energy is still −7.2 kcal/mol. This value is four orders of magnitude greater than the vdW energy at the same distance. As a result, we may introduce substantial errors

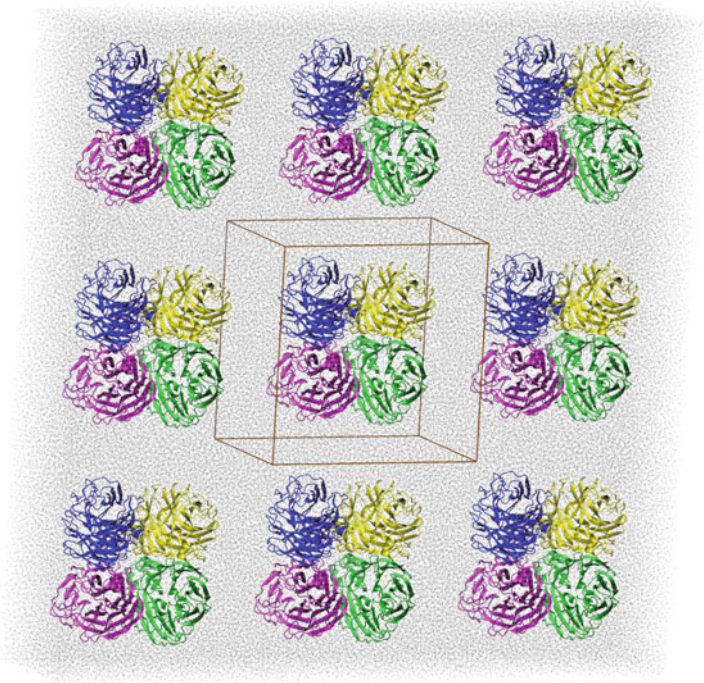

Fig. 3 The periodic boundary conditions. The original simulation box in the center is replicated throughout space to form an infinite lattice. For clarity, only eight replicas are shown in the figure

into the simulation results if we simply truncate the electrostatic potential at the cutoff distance. To solve the problem, several methods have been proposed, and a commonly used method is the Ewald summation [5, 38]. The basic idea of the method is to introduce a neutralizing charge distribution for every point charge in the system. The resulting electrostatic potential, which decays much faster than r^{-1}, can then be calculated using a cutoff scheme. Of course, we have to calculate the electrostatic potential of the neutralizing charge distribution and remove it from the final result. Due to the slowly varying nature of this potential, this part of the calculation can be performed in the reciprocal space via Fourier transform, where we can use the cutoff scheme once again.

The application of the Ewald summation requires the periodic boundary conditions (PBC), i.e., the cubic box containing the original simulation system is replicated throughout space to form an infinite lattice, and atoms leaving the box from one side will enter from the opposite side (Fig. 3). Apart from enabling the Ewald calculation, the PBC have many advantages. For instance, the surface effect of a finite-sized system is eliminated, since no atom is on the surface of an infinite lattice. However, the artificial periodicity introduced by PBC inhibits the occurrence of long-wavelength fluctuations [5], and has been found to reduce the magnitude

of ionic solvation energy [56]. Despite these artifacts, the PBC are generally considered to have little impact on the equilibrium thermodynamic properties of a system [5], and are routinely used in MD simulations of biomolecules.

3.3 Equations of Motion

Now that we have discussed the various components of V and outlined the procedure of their calculation, the next step is to derive the time evolution of our system by integrating the equations of motion. The goal is to calculate, at timestep $n+1$, the coordinates X_{n+1}, velocities V_{n+1} and forces F_{n+1} of all atoms, given the corresponding values of these quantities at the previous timestep n. The Verlet algorithm [114], which belongs to the class of finite difference methods, is commonly used to perform such calculations. In practice, we often use the velocity form of the Verlet algorithm [110], which has improved numerical accuracy over the original method. This algorithm contains the following equations,

$$V_{n+\frac{1}{2}} = V_n + \frac{\Delta t}{2} M^{-1} F_n, \tag{19}$$

$$X_{n+1} = X_n + \Delta t V_{n+\frac{1}{2}}, \tag{20}$$

$$F_{n+1} = F(X_{n+1}), \tag{21}$$

$$V_{n+1} = V_{n+\frac{1}{2}} + \frac{\Delta t}{2} M^{-1} F_{n+1}. \tag{22}$$

As shown above, the velocity of the system at step $n+\frac{1}{2}$ is first calculated, followed by the calculation of the coordinates at step $n+1$. Based on the new coordinates, the potential energy function is evaluated and new forces are obtained. The velocity is then advanced by another half a timestep to produce the new value at $n+1$.

A key parameter in the above equations is the timestep Δt, which determines how frequently we perform the integration. Ideally, we would like to use a timestep as large as possible to minimize the computational cost. In reality, we are often limited to a timestep that is rather small, e.g., 1 fs (10^{-15} s), because the timestep must be small enough to allow for accurate evaluation of the fastest motion in a system, which is the vibration of the bond length between two atoms. Using constraint methods, such as the SHAKE algorithm [99], we can fix the bond lengths and increase the timestep from 1 fs to 2 fs. Even with these algorithms, however, the timescale we can routinely access with MD is currently limited to the submicrosecond range. Compared with most experimental techniques, the limited timescale accessible by MD remains a bottleneck of the method.

In (19)–(22), we have used the Verlet algorithm to integrate the Newtonian equations of motion. Since these equations conserve the total energy of a system, the phase space distribution generated above is that of a microcanonical (NVE) ensemble. In order to simulate other statistical ensembles, such as the canonical

(NVT) and the isothermal-isobaric (NPT) ensembles, the Newtonian equations of motion must be modified. Many techniques are available for this purpose [5, 38], almost all of which can be found in some commonly used MD simulation packages, such as AMBER [18], CHARMM [15], GROMACS [107], NAMD [94], and DESMOND [14]. Regardless of the details of these methods, most of them rely on the Verlet algorithm described above to perform the integration of the modified Newtonian equations of motion.

4 Applications of MD: A Few Examples

So far, we have concentrated on the theoretical aspects of MD: its statistical mechanics background, the functional form of the potential energy, and how to integrate the equations of motion. In this section, we will use a few examples to showcase the applications of MD in biomolecular modeling. Due to space limitations, we cannot hope to be comprehensive on this subject, hence, we refer the readers to recent review articles for more application examples of MD [1,66,71,98].

4.1 Calculation of Water Diffusion

As discussed earlier, we can calculate various macroscopic properties of the system from a MD simulation. One such macroscopic property often calculated is the diffusion coefficient of water (D), which describes the mobility of the solvent molecules in the system. It can be calculated from a simulation trajectory in one of the two following ways:

$$D = \frac{1}{3}\int_0^\infty dt\,\langle \mathbf{v}_i(t)\cdot\mathbf{v}_i(0)\rangle, \tag{23}$$

$$D = \frac{1}{6t}\langle|\mathbf{r}_i(t)-\mathbf{r}_i(0)|^2\rangle, \tag{24}$$

where $\mathbf{v}_i(t)$ and $\mathbf{r}_i(t)$ are the velocity and coordinates of a single water molecule at time t, respectively. Equation (24) is derived from (23) via the "Einstein relation," which also holds for other transport coefficients, such as the shear viscosity. These transport coefficients can be calculated from a simulation with equations similar to (23) and (24) [5].

To reduce the usage of disk space, the velocity trajectory is usually not saved in a simulation, and (24) is used in the calculation of D. Here, we should emphasize that (24) is only valid at long time intervals compared with the correlation time of $\mathbf{r}$. This calculation is performed for all the water molecules in the system and the results are averaged to improve the statistical precision. Another commonly used

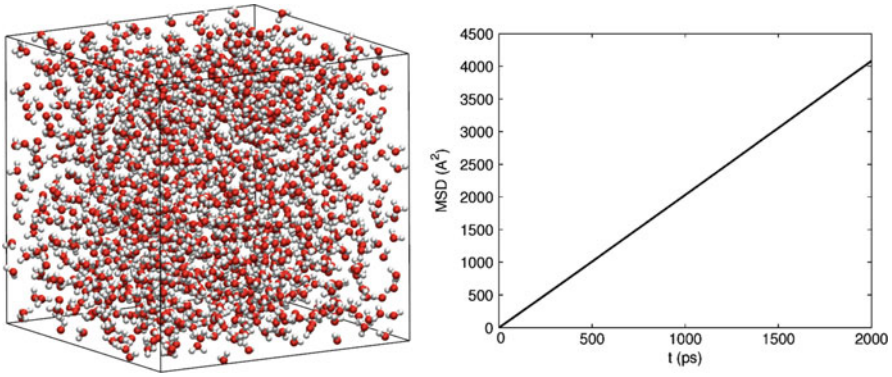

Fig. 4 Calculation of water diffusion coefficient. The simulation system is a 35Å by 35Å by 35Å box, containing 1,437 water molecules. The mean square displacement *MSD* of water is plotted with respect to the time interval t

trick is to shift the time origin $t = 0$ along the simulation trajectory, in order to make use of all data points. For instance, $|\mathbf{r}_i(500) - \mathbf{r}_i(0)|^2$ and $|\mathbf{r}_i(600) - \mathbf{r}_i(100)|^2$ are both included in the mean square displacement (MSD) calculation for the time interval 500 (the unit of the time interval depends on how frequently the simulation trajectory is written). In Fig. 4, we have plotted the MSD with respect to the time interval t in a 10-ns simulation of 1,437 water molecules. The diffusion coefficient, obtained as 1/6 of the slope of the line, is $3.4\times10^{-5}\,\mathrm{cm}^2/\mathrm{s}$.

Although the above example is a very simple application of MD, the methods used here provide the basis to study more complicated biomolecular systems. For instance, a family of channel proteins called aquaporins (AQPs) are responsible for the rapid conduction of water across a lipid membrane [2]. Using MD simulations, the dynamics and selectivity of these channel proteins have been studied extensively [21, 33, 44, 111]. Based on (24) and theories from nonequilibrium statistical mechanics, we can obtain the water permeability coefficient (p_f) through AQPs from relatively short equilibrium MD simulations [128]. This result allows the comparison of water conduction rate in various AQPs to be performed under conditions similar to those used in experiments [51].

4.2 Characterization of Receptor Flexibility in Virtual Screening

In recent years, computer-aided drug design (CADD) has become an indispensable part of the modern drug design process [61]. A widely used technique in CADD

is virtual screening (VS) [67, 70], where a library of compounds are docked into the active site of a protein receptor to identify those molecules with a high binding affinity. The target protein is often an enzyme from a pathogenic organism, and the goal is to identify inhibitors that can block the active site of the enzyme, and, thereby, kill the pathogen. The identified inhibitors can be tested experimentally, and active compounds can be further optimized and developed into new drugs.

Traditionally, VS is performed using only the crystal structure of a protein, which is usually kept rigid during the docking process. The drawback of this scheme is that the protein flexibility is not included in the modeling of receptor-ligand binding. This is now recognized as one of the major issues with the method, since a receptor may adapt to the shape of a specific ligand, e.g., through an "induced fit" mechanism [73], and the resulting complex may not be identified using a rigid protein model [37, 105, 112]. Over the past few years, MD simulations have been successfully applied to provide a description of the receptor flexibility [17, 123]. The relaxed complex scheme (RCS) [6, 7], for instance, is a protocol that combines MD with VS.

In RCS, an MD simulation is used to generate an ensemble of structures for the receptor. Using either RMSD [24] or QR factorization [8] based clustering analysis method, a representative subset of this ensemble is created and then used in subsequent VS. The RCS method offers an efficient approach to incorporate receptor flexibility in VS. Compared with a static crystal structure used in traditional VS, the subset of structures extracted from a NVT or NPT ensemble allow us to have a comprehensive understanding of the receptor active site. For instance, in the VS of HIV-1 integrase, a 2 ns MD simulation revealed a new binding "trench" next to the original active site [100]. This study demonstrates that integrase inhibitors may bind with different orientations in the active site, which proved to be invaluable to the design of new inhibitors with unique resistance profiles [52]. Later, these findings contributed to the discovery of raltegravir [109], the first HIV-integrase inhibitor approved by FDA for treatment of HIV infection.

Another example of RCS is the neuraminidase enzyme from the avian influenza virus H5N1 [24]. In this study, a 40-ns simulation was performed on the neuraminidase tetramer, and RMSD-based clustering was used to extract $\sim$15 structures representing the ensemble generated in the MD simulation. These structures revealed a new "opening" in the neuraminidase active site (Fig. 5), which is completely occluded in the crystal structure. Based on the simulation results, $\sim$1,400 compounds were screened and 25 identified inhibitors were tested experimentally. Out of the 10 confirmed active compounds, 7 were only selected using the MD-generated structures. These potential drug candidates would have been missed if only the crystal structure was used in the VS. For more detailed discussions of the RCS and related methods, we refer the readers to a recent review on this subject [7].

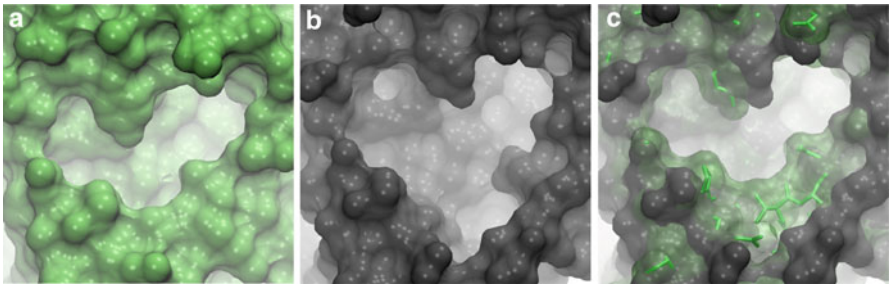

Fig. 5 Receptor flexibility revealed by MD simulations. (**a**) The neuraminidase active site in the crystal structure. (**b**) Representative snapshot from a 40 ns MD simulation. (**c**) The two structures are overlapped to highlight the difference in their active site conformations. Residues from the crystal structure are shown in *green stick* representation. Figure was created using structures from Cheng et al. [24]

5 Running a Simulation: Preparations and Precautions

As shown in the previous section, MD simulations can provide great detail and valuable insight into the biomolecular system under investigation. However, one should use great caution when setting up and analyzing a simulation, especially since MD programs generally have only limited ability to check the "soundness" of simulation results. Hence, an MD simulation may finish without any error after hundreds of CPU hours have been spent, but the result will be of little use if the initial conditions, such as protonation states of residues in an enzymatic site, are not set properly. It is up to the users to take the necessary precautions and avoid such "garbage-in-garbage-out" scenarios. Below, we briefly go over the preparation of an MD simulation and discuss some common pitfalls in this process. The examples we choose are based on simulations of proteins, although most of the principles should apply equally well to simulations of other biomolecules such as nucleotides.

5.1 System Preparation

5.1.1 Choosing Initial Structures

The majority of MD simulations performed today start with an atomic-resolution structure of the biomolecule under investigation. When multiple structures are available for the same biomolecule, the structure best representing the system under investigation, e.g., wild type vs. mutant, should be chosen, and structures with higher resolutions are often preferred over lower-resolution ones. When no structure is available, homology modeling may be used to construct a protein structure based on its similarity to other proteins with known structures. However, a relatively

high sequence identity (>30%) between the reference and target protein is usually required to ensure good quality of the homology model [85], and some expertise may be needed in the refinement of the model [19].

5.1.2 pKa Calculation

Once a protein structure is chosen, the next step is to determine the protonation state of each titratable protein residue. The protonation state of a residue, characterized by its pKa value, is influenced by hydrogen bonding, desolvation effect and Coulombic interactions in its local environment, and, therefore, can be very different from the corresponding standard amino acid [32]. Many good programs are available for predicting and assigning protonation states of protein residues, such as MCCE [3,4], MEAD [10], PROPKA [9], and UHBD [82]. Performing pKa calculation using at least one of these programs should become a routine in the preparation of a simulation system.

5.1.3 Adding Water and Ions

Currently, the majority of MD simulations are performed using explicit water molecules, although implicit solvent simulations have proven very useful in the study of certain biomolecular systems [23, 36, 41]. In an explicit water simulation, the number of water molecules needed is determined by the size of the simulation box. As discussed in Sect. 3, PBC are usually used to avoid the surface effect of a finite-sized system. In these simulations, a rule of thumb is that the biomolecule should never "see" its periodic image. This means that the simulation box has to be large enough so that two neighboring periodic images of the molecule are separated by at least the cutoff distance. In practice, a layer of water at least 10–15Å wide is often added to each side of the protein. However, if the protein is expected to undergo large conformational changes, such as unfolding, the simulation box should be chosen large enough to accommodate the changes.

Apart from water, the buffer solution used in most biological experiments contain various ions. Due to the limited availability of force field parameters, we cannot hope to reproduce the exact experimental conditions in simulations. Nevertheless, it is desirable to include ions, such as Na^+, Cl^-, or K^+ in a system, to provide a similar ionic strength in the simulation box as in the experiments. The added ions should neutralize the net charge of the biomolecule, so that the total charge of the simulation box is zero. However, the Ewald summation method described in Sect. 3 introduces, by design, a homogeneous neutralizing background charge to the system [56]. As a result, systems with nonzero net charges can be simulated without any apparent error. Despite this result, it's generally considered a good practice to keep the simulation system neutral, unless the goal is to simulate a charged system, such as in the calculation of ionic solvation energy.

5.2 Simulation Conditions

5.2.1 Designing Simulation Protocols

Prior to any production MD runs, the system prepared above is put through a short energy minimization run, which removes any large steric clashes or close contacts within the system. Following the minimization, a short simulation with restraints on protein atom positions may be desirable to gradually bring the system to the target conditions, e.g., the desired temperature and pressure. The hope is to introduce as little perturbation to the structure as possible at this stage of the simulation, while allowing the environment to relax around the biomolecule. Although many protocols are used in various simulation studies, for a small protein (<500 aa), a 500-ps run is usually considered sufficient, during which water molecules travel on average ~30Å, giving them enough time to relax in the immediate vicinity of the protein.

The ensembles used in a MD simulation should be chosen according to the nature of the biomolecular system under investigation. The microcanonical (NVE) ensemble, where the total energy of the system is kept constant, is mostly used to examine the energy conservation performance of a MD program, while the canonical (NVT) and the isothermal-isobaric (NPT) ensembles, which better resemble experimental conditions, are often preferred. However, once the system is brought to equilibrium, calculations of thermodynamic properties based on different ensembles can produce very similar results [5]. Therefore, one can start the simulation in the NPT ensemble, bring the system to the equilibrium temperature and pressure, and then continue the production run in either the NVT or the NPT ensemble.

At this stage, a natural question one may ask is how long a simulation should last. This question would become trivial if there is no limitation to our computational resources. Unfortunately, most researchers won't have this luxury, and the answer to the above question will depend closely on the purpose of the study. Diffusion of water and side chain reorientation of protein residues usually take place in tens to hundreds of picoseconds, while large-scale conformational changes, such as folding and unfolding, could take hundreds of nanoseconds, microseconds, to milliseconds. Most simulations performed today are in the nanosecond to microsecond range. Therefore, the user should decide the simulation time based on the target problem. If, for instance, large conformational changes are the goal, which may not occur in a 10-ns simulation, the computational resources may be better spent on enhanced sampling methods, where advanced simulation techniques are used to force the biomolecule to undergo certain structural changes. We will discuss some of these techniques in the next section.

A question related to the simulation length is whether one should run a single long simulation or several short ones, given the same amount of computational resources. This topic remains an interesting and active area of research. Recent simulations and free energy calculation studies have suggested that in certain cases short multiple runs can better capture the dynamics of the protein than a single long run [20, 77].

5.2.2 Dealing with Errors

New users of any MD program will likely encounter some errors in their first few simulation runs. Although no solution can be given without knowing the specific errors, it is possible to offer general guidance: one of the most error-prone steps in an MD simulation is the system preparation. Any small mistake at this stage could cause the simulation to crash within the first few picoseconds or render its result invalid. For instance, a new residue or ligand may have been introduced without the corresponding force field parameters. The transformation matrix used to generate an oligomer of the protein may have been entered in the wrong order, and the resulting protein monomers may be placed too close to each other. In such cases, one must go back to the system preparation step and fix any problems with the starting structure. Missing force field parameters may be obtained using tools provided by the corresponding MD package or force field [113,118,119]; and the transformation matrix can be checked against the initial pdb file.

The cause of some errors may be more difficult to decipher, an example of which is the common error message "atom moving too fast," given by the MD program NAMD [94]. This error means that a certain atom, the index of which is given along with the error message, has a velocity greater than the maximum value defined by the program. Since the maximum value is set to be much higher than the velocities from any realistic MD simulation, this error usually indicates structural defects in the system. For instance, two atoms may be placed right on top of each other and the resulting vdW force, which is repulsive at very short distance, will cause the two atoms to "fly away" at a very high speed. Usually, the minimization step described earlier can eliminate these close contacts effectively. However, some structural defects may be too great to be completely removed by minimization, in which case manual correction of the initial structure is required. It is also likely that only a small part of the initial structure is incorrect, which tends to be the case when new residues or ligands are involved. In these cases, the problematic region in the structure can be difficult to spot, since the majority of the system will behave normally. Often, when one couldn't locate the problematic structure, it is useful to repeat the simulation with the trajectory written every step. Such a "slow motion" picture of the system could help to pinpoint the exact cause of the problem.

6 Advanced Simulation Techniques

In this section, we will briefly introduce some advanced simulation techniques, including enhanced sampling and free energy calculation methods. As the calculation of free energy often requires enhanced sampling to be performed along a specific reaction coordinate, the distinction between these two types of methods is not always clear. Therefore, in the following discussions, we will not attempt a strict classification, but will focus on a few commonly used techniques and their specific

features. Due to space limitations, our accounts of these methods are by no means comprehensive. Therefore, we refer readers to the many review articles and books on these topics for more detailed discussions [13, 25, 27, 40, 72, 86, 95, 124].

6.1 Accelerated Molecular Dynamics

As discussed in Sect. 2, the direct evaluation of Z_{NVT} in a canonical ensemble is not feasible using regular MD simulations. Since the microstates are sampled according to a Boltzmann distribution in the canonical ensemble, the basic idea of enhanced sampling methods is to escape from this distribution and sample the configurational space in a "non-Boltzmann" way. Some examples of enhanced sampling methods are accelerated molecular dynamics (aMD) [49], conformational flooding [45, 76] and hyperdynamics [115, 116]. Here, we will use the aMD method as an example to illustrate the principles behind enhanced sampling methods.

In the original aMD method [49], when the system's potential energy falls below a threshold energy, E, a bias potential is added, such that the modified potential, $V^*(\mathbf{r})$, is related to the original potential, $V(\mathbf{r})$, via

$$V^*(\mathbf{r}) = V(\mathbf{r}) + \Delta V(\mathbf{r}), \tag{25}$$

where $\Delta V(\mathbf{r})$ is the bias potential,

$$\Delta V(\mathbf{r}) = \begin{cases} 0 & V(\mathbf{r}) \geq E \\ \frac{(E - V(\mathbf{r}))^2}{\alpha + E - V(\mathbf{r})} & V(\mathbf{r}) < E. \end{cases} \tag{26}$$

In the above equation, E is the threshold energy specified by the user, which controls the portion of the potential surface affected by the bias. The acceleration factor α determines how "aggressive" the modification to the potential surface is: the smaller α, the more flattened the energy surface becomes.

Under the influence of the bias potential ΔV, the sampling in an aMD simulation will not follow a Boltzmann distribution. Instead, the energy barriers between adjacent low-energy states are lowered, and the system can explore the configurational space more efficiently. Like other enhanced sampling methods, the effect of this bias potential must then be removed from the final result. In aMD, this is achieved by reweighing the simulation trajectory in the calculation of the ensemble average $\langle A \rangle$:

$$\langle A \rangle = \frac{\langle A(\mathbf{r}) \exp(\beta \Delta V(\mathbf{r})) \rangle^*}{\langle \exp(\beta \Delta V(\mathbf{r})) \rangle^*}, \tag{27}$$

in which $\langle ... \rangle$ and $\langle ... \rangle^*$ represent the ensemble average in the original (unbiased) and the aMD (biased) ensembles, respectively.

The aMD method has been successfully applied to study a number of peptide or protein systems, including HIV-1 protease [47], the proteins GB3 [83] and ubiquitin [84], as well as the GTPase protein Ras [42]. Recently, several extensions of this method have been developed [35,50,90,121,122], and interested readers may find more discussions of aMD in a couple of review articles [43,48].

6.2 Free Energy: The Concept

The free energy change associated with a chemical or biological process largely determines the equilibrium properties of the system under investigation. The protein-ligand binding, for instance, is governed by the free energy change associated with the formation of a complex by the two molecules [39]. In statistical mechanics, the excess or configurational Helmholtz free energy, A, which is the thermodynamic potential usually associated with a canonical ensemble, can be expressed as

$$A = -\beta^{-1} \ln Z_{\mathrm{NVT}}, \tag{28}$$

where Z_{NVT} is defined in (8). Just like we cannot use (8) to calculate Z_{NVT}, the direct evaluation of A using the above equation is also unfeasible. Fortunately, in most cases, we are only interested in the difference between the free energies of two states, e.g., state 0 and state 1. Assuming state 0 and state 1 are characterized by the partition functions Z_0 and Z_1, respectively, the difference in their free energies is given by

$$\Delta A = -\beta^{-1} \ln Z_1/Z_0. \tag{29}$$

Equation (29) can be used to describe a large number of free energy calculation problems. Perhaps the most relevant one to biomolecule modeling is the calculation of binding affinity in a protein–ligand complex. These calculations often involve the creation or annihilation of a ligand molecule, i.e., the energy terms involving the ligand are gradually added or removed from the total Hamiltonian. Such calculations are often referred to as "alchemical transformations," and are used to obtain solvation-free energy or binding-free energy [39,127].

Equation (29) suggests that in order to obtain ΔA, we only need to calculate the ratio between the two partition functions, rather than each individual Z_{NVT}. This observation provides the basis for the various free energy calculation methods. Below, we will focus on two of the most widely used methods, namely, free energy perturbation (FEP) and thermodynamic integration (TI).

6.3 Free-Energy Perturbation

In the FEP approach, we start by combining (8) and (29),

$$\Delta A = -\beta^{-1} \ln \frac{\sum \exp(-\beta V_1)}{\sum \exp(-\beta V_0)} \tag{30}$$

$$= -\beta^{-1} \ln \frac{\sum \exp[-\beta (V_1 - V_0)] \exp(-\beta V_0)}{\sum \exp(-\beta V_0)} \tag{31}$$

$$= -\beta^{-1} \ln \langle \exp[-\beta (V_1 - V_0)] \rangle_0. \tag{32}$$

For simplicity, we have omitted the dependence of V on $\mathbf{r}$ in the above equations, and the summation is to be understood as being performed over all the configurations $\mathbf{r}$. The transition from (31) to (32) is made using the definition of ensemble average in (10), and $\langle ... \rangle_0$ represents an ensemble average performed in state 0. If we further define $\Delta V = V_1 - V_0$, then (32) can be simplified to

$$\Delta A = -\beta^{-1} \ln \langle \exp(-\beta \Delta V) \rangle_0. \tag{33}$$

In essence, the above equation can be thought of as calculating the exponential of ΔV when the system is "frozen" at a particular configuration in state 0, and the obtained exponentials are averaged over all configurations in state 0 to give ΔA. Note that we can get an equivalent formula by using the ensemble average in state 1,

$$\Delta A = \beta^{-1} \ln \langle \exp(\beta \Delta V) \rangle_1. \tag{34}$$

The above two equations, which form the basis of the FEP approach, are often referred to as the "forward" and "reverse" calculation, respectively. Combining both forward and reverse calculations using the Benett acceptance ratio (BAR) method [11] has been demonstrated to yield the most accurate result [106]. Since in many MD programs, performing calculations in both directions can be done nearly as efficiently as performing the calculation in a single direction, combining forward and reverse FEP has been recommended as a standard practice [95].

6.4 Thermodynamic Integration

In the TI approach, a parameter λ is used to describe the transition of the system from state 0 to state 1. For instance, in an alchemical transformation where a ligand is annihilated, λ_0 and λ_1 will correspond to the system with and without the ligand, respectively. To calculate ΔA using TI, we start by differentiating (28),

$$\frac{\mathrm{d}A}{\mathrm{d}\lambda} = -\frac{1}{\beta} \frac{1}{Z_{\mathrm{NVT}}} \frac{\partial Z_{\mathrm{NVT}}}{\partial \lambda}. \tag{35}$$

If λ is a parameter in the potential energy function, we have

$$\frac{\mathrm{d}A}{\mathrm{d}\lambda} = \frac{\sum \exp(-\beta V)\frac{\partial V}{\partial \lambda}}{\sum \exp(-\beta V)} = \left\langle \frac{\partial V}{\partial \lambda} \right\rangle_{\lambda}. \tag{36}$$

The free energy difference between state 0 and state 1 is obtained by integrating $\mathrm{d}A/\mathrm{d}\lambda$ in the range of λ_0 to λ_1.

$$\Delta A = \int_{\lambda_0}^{\lambda_1} \left\langle \frac{\partial V}{\partial \lambda} \right\rangle_{\lambda}. \tag{37}$$

The derivative of the potential energy V with respect to λ can be performed analytically in many cases, and the calculation of $\mathrm{d}A/\mathrm{d}\lambda$ amounts to obtain the numerical value of $\mathrm{d}V/\mathrm{d}\lambda$ for each configuration sampled at a particular λ, and then calculate the ensemble average $\langle \mathrm{d}V/\mathrm{d}\lambda \rangle_{\lambda}$ by summing over all the configurations. The integration in (37) is performed numerically, often using either the trapezoidal or the Simpson's rule. The performance of these two integration methods has been compared in a recent study, which shows that the Simpson's rule tends to generate smaller systemic errors in the results [60].

An important development that significantly improves the accuracy of alchemical transformation calculations is the soft-core potential [12, 126], which can be used in combination with both the TI and FEP methods. Such a potential effectively removes the singularities in the potential energy function when the distance between the ligand and surrounding atoms approaches zero during the creation or annihilation process. Since the numerical accuracy of the result is significantly improved, the soft-core potential should always be used in alchemical transformations.

6.5 *Umbrella Sampling and Other Techniques*

Apart from FEP and TI, another commonly used free energy calculation method is the umbrella sampling (US) approach. This method divides the transition from λ_0 to λ_1 into multiple windows, and uses a biasing potential to restrain the system at a particular λ in each window. The probability distribution along λ in each window is collected as a histogram, which is combined to give the complete free energy profile using the weighted histogram analysis method (WHAM) [74, 97].

The US method is often used as a benchmark to evaluate the performance of new free energy calculation techniques. Recently, several such techniques have been developed, including metadynamics [75] and adaptive biasing force (ABF) method [30, 31, 53]. Both of them have been applied to biomolecular systems and demonstrated superior performance than the US method. Another recent development that significantly improves the efficiency of protein-ligand binding affinity calculation is the enveloping distribution sampling (EDS) method [26, 91],

which allows the simultaneous evaluation of the binding free energy for multiple ligands. Interested readers can find more discussions of these methods in the corresponding research articles.

So far, all the methods discussed here are based on equilibrium MD simulations. With recent theory advancement in nonequilibrium statistical mechanics [28,59], we can also calculate free energy changes using nonequilibrium simulations. This type of method is based on the Jarzynski equality [58,59], and can be combined with the steered molecular dynamics (SMD) simulation [57, 93] technique to produce free energy profiles along a reaction coordinate.

7 Outlook

In this chapter, we have discussed the various theoretical aspects of MD and provided examples of its recent applications on biomolecular systems. Compared with the first MD simulation on a protein [87], which lasted about 9 ps, the development of MD has come a long way. The method has made significant contributions to our understanding of the behaviors of complex biomolecules. Today, with the power of supercomputers and the progress of MD softwares, we can readily perform simulations on millions of atoms for tens to hundreds of nanoseconds. The recent launch of specialized machines [103,104] and the usage of graphics processing units (GPUs) [108,125] have initiated a new round of exciting method advancement. With these new technologies, we can expect MD simulations to make even more significant contributions to our understanding of biomolecular systems in future.

Acknowledgments This work has been supported in part by the National Science Foundation, the National Institutes of Health, Howard Hughes Medical Institute, Center for Theoretical Biological Physics, the National Biomedical Computation Resource, and the NSF supercomputer centers.

References

1. Adcock, S.A., McCammon, J.A.: Molecular dynamics: survey of methods for simulating the activity of proteins. Chem. Rev. **106**, 1589–1615 (2006)
2. Agre, P.: The aquaporin water channels. Proc. Am. Thorac. Soc. **3**, 5–13 (2006)
3. Alexov, E., Gunner, M.: Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. Biophys. J. **72**, 2075–2093 (1997)
4. Alexov, E., Gunner, M.: Calculated protein and proton motions coupled to electron transfer: electron transfer from QA- to QB in bacterial photosynthetic reaction centers. Biochemistry **38**, 8253–8270 (1999)
5. Allen, M.P., Tildesley, D.J.: Computer Simulation of Liquids. Oxford University Press, New York (1987)
6. Amaro, R., Baron, R., McCammon, J.: An improved relaxed complex scheme for receptor flexibility in rational drug design. J. Comp.-Aided Mol. Design **22**, 693–705 (2008)

7. Amaro, R., Li, W.: Emerging ensemble-based methods in virtual screening **10**, 3–13 (2010)
8. Amaro, R.E., Schnaufer, A., Interthal, H., Hol, W., Stuart, K.D., McCammon, J.A.: Discovery of drug-like inhibitors of an essential RNA-editing ligase in *trypanosoma brucei*. Proc. Natl. Acad. Sci. USA **105**, 17,278–17,283 (2008)
9. Bas, D.C., Rogers, D.M., Jensen, J.H.: Very fast prediction and rationalization of pK(a) values for protein-ligand complexes. Proteins: Struct. Func. Bioinf. **73**, 765–783 (2008)
10. Bashford, D.: An object-oriented programming suite for electrostatic effects in biological molecules: an experience report on the MEAD project. ISCOPE97. Proccedings **1343**, 233–240 (1997)
11. Bennett, C.H.: Efficient estimation of free energy differences from Monte Carlo data. J. Comp. Phys. **22**, 245–268 (1976)
12. Beutler, T., Mark, A., van Schaik, R., Gerber, P., van Gunsteren, W.: Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. Chem. Phys. Lett. **222**, 529–539 (1994)
13. Beveridge, D.L., DiCapua, F.M.: Free energy via molecular simulation: Applications to chemical and biological systems. Annu. Rev. Biophys. Biophys. Chem. **18**, 431–492 (1989)
14. Bowers, K.J., Chow, E., Xu, H., Dror, R.O., Eastwood, M.P., Gregerson, B.A., Klepeis, J.L., Kolossvary, I., Moraes, M.A., Sacerdoti, F.D., Salmon, J.K., Shan, Y., Shaw, D.E.: Scalable algorithms for molecular dynamics simulations on commodity clusters. In: Proceedings of the ACM/IEEE SC06 Conference. ACM (2006)
15. Brooks, B.R., III, C.L.B., Mackerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Caflisch, S.B.A., Caves, L., Cui, Q., Dinner, A.R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R.W., Post, C.B., Pu, J.Z., Schaefer, M., Tidor, B., Venable, R.M., Woodcock, H.L., Wu, X., Yang, W., York, D.M., Karplus, M.: CHARMM: The biomolecular simulation program. J. Comp. Chem. **30**, 1545–1615 (2009)
16. Brooks, C.L., Karplus, M., Pettit, B.M.: Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics. Wiley, New York (1989)
17. Carlson, H.A.: Protein flexibility and drug design: how to hit a moving target. Curr. Opin. Chem. Biol. **6**, 447 (2002)
18. Case, D.A., Darden, T.A., Cheatham, T.E., Simmerling, C.L., Wang, J., Duke, R.E., Luo, R., Crowley, M., Walker, R.C., Zhang, W., Merz, K.M., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossváry, I., Wong, K.F., Paesani, F., Vanicek, J., Wu, X., Brozell, S., Steinbrecher, T., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Mathews, D.H., Seetin, M.G., Sagui, C., Babin, V., Kollman, P.A.: AMBER 10. University of California, San Francisco (2008)
19. Cavasotto, C.N., Phatak, S.S.: Homology modeling in drug discovery: current trends and applications. Drug Discov. Today **14**, 676–683 (2009)
20. Caves, L., Evanseck, J., Karplus, M.: Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin. Prot. Sci. **7**, 649–666 (1998)
21. Chakrabarti, N., Tajkhorshid, E., Roux, B., Pomès, R.: Molecular basis of proton blockage in aquaporins. Structure **12**, 65–74 (2004)
22. Chandler, D.: Introduction to Modern Statistical Mechanics. Oxford University, New York (1987)
23. Chen, J., III, C.L.B., Khandogin, J.: Recent advances in implicit solvent-based methods for biomolecular simulations. Curr. Opin. Struct. Biol. **18**, 140–148 (2008)
24. Cheng, L.S., Amaro, R.E., Xu, D., Li, W.W., Arzberger, P.W., McCammon, J.A.: Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. J. Med. Chem. **51**, 3878–3894 (2008)
25. Chipot, C., Pohorille, A.: Free Energy Calculations. Theory and applications in chemistry and biology. Springer, Berlin (2007)
26. Christ, C.D., van Gunsteren, W.F.: Simple, efficient, and reliable computation of multiple free energy differences from a single simulation: a reference hamiltonian parameter update scheme for enveloping distribution sampling (EDS). J. Chem. Theor. Comp. **5**, 276–286 (2009)

27. Christen, M., van Gunsteren, W.: On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: a review. J. Comp. Chem. **29**, 157–166 (2008)
28. Crooks, G.E.: Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. Phys. Rev. E **60**, 2721–2726 (1999)
29. Damm, W., Frontera, A., Tirado-Rives, J., Jorgensen, W.L.: The OPLS all-atom force field for carbohydrates. J. Comp. Chem. **18**, 1955–1970 (1997)
30. Darve, E., Pohorille, A.: Calculating free energies using average force. J. Chem. Phys. **115**, 9169–9183 (2001)
31. Darve, E., Rodríguez-Gómez, D., Pohorille, A.: Adaptive biasing force method for scalar and vector free energy calculations. J. Chem. Phys. **128**, 144,120 (2008)
32. Davies, M.N., Toseland, C.P., Moss, D.S., Flower, D.R.: Benchmarking pka prediction. BMC Biochem. **7**, 18 (2006)
33. de Groot, B.L., Grubmüller, H.: Water permeation across biological membranes: mechanism and dynamics of aquaporin-1 and GlpF. Science **294**, 2353–2357 (2001)
34. Dill, K.A., Bromberg, S.: Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology. Garland Science, New York (2002)
35. Fajer, M., Swift, R., McCammon, J.: Using multistate free energy techniques to improve the efficiency of replica exchange accelerated molecular dynamics. J. Comp. Chem. **30**, 1719–1725 (2009)
36. Feig, M., Brooks, C.L.: Recent advances in the development and application of implicit solvent models in biomolecule simulations. Curr. Opin. Struct. Biol. **14**, 217–224 (2004)
37. Fradera, X., de la Cruz, X., Silva, C.H.T.P., Gelpí, J.L., Luque, F., Orozco, M.: Ligand-induced changes in the binding sites of proteins. Bioinformatics **18**, 939–948 (2002)
38. Frenkel, D., Smit, B.: Understanding Molecular Simulation From Algorithms to Applications. Academic Press, California (2002)
39. Gilson, M., Given, J., Bush, B., McCammon, J.: The statistical-thermodynamic basis for computation of binding affinities: a critical review. Biophys. J. **72**, 1047–1069 (1997)
40. Gilson, M., Zhou, H.X.: Calculation of protein-ligand binding affinities. Annu. Rev. Biophys. Biomol. Struct. **36**, 21–42 (2007)
41. Gilson, M.K., McCammon, J.A., Madura, J.D.: Molecular dynamics simulation with a continuum electrostatic model of the solvent. J. Comp. Chem. **16**(9), 1081–1095 (1995)
42. Grant, B., Gorfe, A., McCammon, J.: Ras conformational switching: simulating nucleotide-dependent conformational transitions with accelerated molecular dynamics. PLoS Comput. Biol. **5**, e1000,325 (2009)
43. Grant, B., Gorfe, A., McCammon, J.: Large conformational changes in proteins: signaling and other functions. Curr. Opin. Struct. Biol. **20**, 142–147 (2010)
44. de Groot, B.L., Grubmüller, H.: The dynamics and energetics of water permeation and proton exclusion in aquaporins. Curr. Opin. Struct. Biol. **15**, 1–8 (2005)
45. Grubmüller, H.: Predicting slow structural transitions in macromolecular systems: conformational flooding. Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics **52**, 2893–2906 (1995)
46. van Gunsteren, W., Dolenc, J., Mark, A.: Molecular simulation as an aid to experimentalists. Curr. Opin. Struct. Biol. **18**, 149–153 (2008)
47. Hamelberg, D., McCammon, J.: Fast peptidyl cis-trans isomerization within the flexible gly-rich flaps of HIV-1 protease. J. Am. Chem. Soc. **127**, 13,778–13,779 (2005)
48. Hamelberg, D., McCammon, J.A.: Accelerating conformational transitions in biomolecular systems **2**, 221–232 (2006)
49. Hamelberg, D., Mongan, J., McCammon, J.: Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J. Chem. Phys. **120**(24), 11,919–11,929 (2004)
50. Hamelberg, D., de Oliveira, C., McCammon, J.: Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. J. Chem. Phys. **127**, 155,102 (2007)

51. Hashido, M., Ikeguchi, M., Kidera, A.: Comparative simulations of aquaporin family: AQP1, AQPZ, AQP0 and GlpF. FEBS Lett. **579**, 5549–5552 (2005)
52. Hazuda, D., Anthony, N., Gomez, R., Jolly, S., Wai, J., Zhuang, L., Fisher, T., Embrey, M., Guare JP, J., Egbertson, M., Vacca, J., Huff, J., Felock, P., Witmer, M., Stillmock, K., Danovich, R., Grobler, J., Miller, M., Espeseth, A., Jin, L., Chen, I., Lin, J., Kassahun, K., Ellis, J., Wong, B., Xu, W., Pearson, P., Schleif, W., Cortese, R., Emini, E., Summa, V., Holloway, M., Young, S.: A naphthyridine carboxamide provides evidence for discordant resistance between mechanistically identical inhibitors of HIV-1 integrase. Proc. Natl. Acad. Sci. USA **101**, 11,233–11,238 (2004)
53. Hénin, J., Fiorin, G., Chipot, C., Klein, M.: Exploring multidimensional free energy landscapes using time-dependent biases on collective variables. J. Chem. Theor. Comp. **6**, 35–47 (2010)
54. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., Simmerling, C.: Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins **65**, 712–725 (2006)
55. Hub, J., de Groot, B.: Mechanism of selectivity in aquaporins and aquaglyceroporins. Proc. Natl. Acad. Sci. USA **105**, 1198–203 (2008)
56. Hünenberger, P.H., McCammon, J.A.: Ewald artifacts in computer simulations of ionic solvation and ion-ion interaction: a continuum electrostatics study. J. Chem. Phys. **110**, 1856–1872 (1999)
57. Isralewitz, B., Gao, M., Schulten, K.: Steered molecular dynamics and mechanical functions of proteins. Curr. Opin. Struct. Biol. **11**, 224–230 (2001)
58. Jarzynski, C.: Equilibrium free-energy differences from nonequilibrium measurements: a master equation approach. Phys. Rev. E **56**, 5018–5035 (1997)
59. Jarzynski, C.: Nonequilibrium equality for free energy differences. Phys. Rev. Lett. **78**, 2690–2693 (1997)
60. Jorge, M., Garrido, N.M., Queimada, A.J., Economou, I.G., Maced, E.A.: Effect of the integration method on the accuracy and computational efficiency of free energy calculations using thermodynamic integration. J. Chem. Theor. Comp. **6**, 1018–1027 (2010)
61. Jorgensen, W.L.: The many roles of computation in drug discovery. Science **303**, 1813–1818 (2004)
62. Jorgensen, W.L.: Efficient drug lead discovery and optimization. Acc. Chem. Res. **42**, 724–733 (2009)
63. Jorgensen, W.L., Maxwell, D.S., Tirado-Rives, J.: Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc. **118**, 11,225–11,236 (1996)
64. Karplus, M.: Molecular dynamics: applications to proteins. In: J.L. Rivail (ed.) Modelling of Molecular Structures and Properties, Studies in Physical and Theoretical Chemistry, vol. 71, pp. 427–461. Elsevier Science Publishers, Amsterdam (1990). Proceedings of an International Meeting
65. Karplus, M., McCammon, J.A.: Molecular dynamics simulations of biomolecules. Nat. Struct. Biol. **265**, 654–652 (2002)
66. Khalili-Araghi, F., Gumbart, J., Wen, P.C., Sotomayor, M., Tajkhorshid, E., Schulten, K.: Molecular dynamics simulations of membrane channels and transporters. Curr. Opin. Struct. Biol. **19**, 128–137 (2009)
67. Kitchen, D.B., Decornez, H., Furr, J.R., Bajorath, J.: Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug Disc. **3**, 939–945 (2004)
68. Kittel, C., Kroemer, H.: Thermal Physics. W. H. Freeman, San Francisco (1980)
69. Klauda, J.B., Venable, R.M., Freites, J.A., O'Connor, J.W., Tobias, D.J., Mondragon-Ramirez, C., Vorobyov, I., MacKerell Jr., A.D., Pastor, R.W.: Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. J. Phys. Chem. B **114**(23), 7830–7843 (2010)
70. Klebe, G.: Virtual ligand screening: strategies, perspectives and limitations. Drug Discov. Today **11**, 580–594 (2006)

71. Klepeis, J.L., Lindorff-Larsen, K., Dror, R.O., Shaw, D.E.: Long-timescale molecular dynamics simulations of protein structure and function. Curr. Opin. Struct. Biol. **19**, 120–127 (2009)
72. Kollman, P.: Free energy calculations: applications to chemical and biochemical phenomena. Chem. Rev. **93**, 2395–2417 (1993)
73. Koshland, D.E.: Application of a theory of enzyme specificity to protein synthesis. Proc. Natl. Acad. Sci. USA **44**, 98–104 (1958)
74. Kumar, S., Bouzida, D., Swendsen, R.H., Kollman, P.A., Rosenberg, J.M.: The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J. Comp. Chem. **13**, 1011–1021 (1992)
75. Laio, A., Parrinello, M.: Escaping free energy minima. PNAS **99**(20), 12,562–12,566 (2002)
76. Lange, O., Schäfer, L., Grubmüller, H.: Flooding in GROMACS: accelerated barrier crossings in molecular dynamics. J. Comp. Chem. **27**, 1693–1702 (2006)
77. Lawrenz, M., Baron, R., McCammon, J.A.: Independent-trajectories thermodynamic-integration free-energy changes for biomolecular systems: determinants of H5N1 avian influenza virus neuraminidase inhibition by peramivir. J. Chem. Theor. Comp. **5**, 1106–1116 (2009)
78. Lawrenz, M., Wereszczynski, J., Amaro, R., Walker, R., Roitberg, A., McCammon, J.A.: Impact of calcium on N1 influenza neuraminidase dynamics and binding free energy. Proteins: Struct. Func. Bioinf. **78**(11), 2523–2532 (2010)
79. MacKerell Jr., A.D., Bashford, D., Bellott, M., Dunbrack, J.R.L., Evanseck, J., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph, D., Kuchnir, L., Kuczera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Roux, B., Schlenkrich, M., Smith, J., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., Karplus, M.: Self-consistent parameterization of biomolecules for molecular modeling and condensed phase simulations. FASEB J. **6**(1), A143–A143 (1992)
80. MacKerell Jr., A.D., Bashford, D., Bellott, M., Dunbrack Jr., R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph, D., Kuchnir, L., Kuczera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher, I.W.E., Roux, B., Schlenkrich, M., Smith, J., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., Karplus, M.: All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B **102**, 3586–3616 (1998)
81. MacKerell Jr., A.D., Feig, M., Brooks III, C.L.: Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J. Comp. Chem. **25**, 1400–1415 (2004)
82. Madura, J., Briggs, J., Wade, R., Davis, M., Luty, B., Ilin, A., Antosiewicz, J., Gilson, M., Bagheri, B., Scott, L., McCammon, J.: Electrostatics and diffusion of molecules in solution: simulations with the university of houston brownian dynamics program. Comput. Phys. Commun. **91**, 57–95 (1995)
83. Markwick, P., Bouvignies, G., Blackledge, M.: Exploring multiple timescale motions in protein gb3 using accelerated molecular dynamics and nmr spectroscopy. J. Am. Chem. Soc. **129**, 4724–4730 (2007)
84. Markwick, P., Bouvignies, G., Salmon, L., McCammon, J., Nilges, M., Blackledge, M.: Toward a unified representation of protein structural dynamics in solution. J. Am. Chem. Soc. **131**, 16,968–16,975 (2009)
85. Martí-Renom, M.A., Stuart, A.C., Fiser, A., Sánchez, R., Melo, F., Šali, A.: Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. **29**, 291–325 (2000)
86. McCammon, J.: Free energy from simulations. Curr. Opin. Struct. Biol. **1**, 196–200 (1991)
87. McCammon, J.A., Gelin, B.R., Karplus, M.: Dynamics of folded proteins. Nature **267**, 585–590 (1977)
88. McCammon, J.A., Harvey, S.C.: Dynamics of Proteins and Nucleic Acids. Cambridge University Press, Cambridge (1987)
89. McQuarrie, D.A.: Statistical Mechanics. Harper and Row, New York (1976)

90. de Oliveira, C., Hamelberg, D., McCammon, J.: Coupling accelerated molecular dynamics methods with thermodynamic integration simulations. J. Chem. Theor. Comp. **4**, 1516–1525 (2008)
91. Oostenbrink, C., van Gunsteren, W.F.: Free energies of ligand binding for structurally diverse compounds. Proc. Natl. Acad. Sci. USA **102**, 6750–6754 (2005)
92. Oostenbrink, C., Villa, A., Mark, A.E., Gunsteren, W.F.V.: A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J. Comp. Chem. **25**, 1656–1676 (2004)
93. Park, S., Schulten, K.: Calculating potentials of mean force from steered molecular dynamics simulations. J. Chem. Phys. **120**, 5946–5961 (2004)
94. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L., Schulten, K.: Scalable molecular dynamics with NAMD. J. Comp. Chem. **26**, 1781–1802 (2005)
95. Pohorille, A., Jarzynski, C., Chipot, C.: Good practices in free-energy calculations. J. Phys. Chem. B **114**, 10,235–10,253 (2010)
96. Price, M., Ostrovsky, D., Jorgensen, W.L.: Gas-phase and liquid-state properties of esters, nitriles, and nitro compounds with the OPLS-AA force field. J. Comp. Chem. **22**, 1340–1352 (2001)
97. Roux, B.: The calculation of the potential of mean force using computer simulations. Comput. Phys. Comm. **91**, 275–282 (1995)
98. Roux, B.: Ion conduction and selectivity in K^+ channels. Annu. Rev. Biomol. Struc. Dyn. **34**, 153–171 (2005)
99. Ryckaert, J.P., Ciccotti, G., Berendsen, H.J.C.: Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. J. Comp. Phys. **23**, 327–341 (1977)
100. Schames, J.R., Henchman, R.H., Siegel, J.S., Sotriffer, C.A., Ni, H., McCammon, J.A.: Discovery of a novel binding trench in HIV integrase. J. Med. Chem. **47**, 1879–1881 (2004)
101. Schlick, T.: Molecular Modeling and Simulation: An Interdisciplinary Guide, 2nd edn. Springer, New York (2010)
102. Schuler, L., Daura, X., van Gunsteren, W.: An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. J. Comp. Chem. **22**, 1205–1218 (2001)
103. Shaw, D.E., Dror, R.O., Salmon, J.K., Grossman, J., Mackenzie, K.M., Bank, J.A., Young, C., Deneroff, M.M., Batson, B., Bowers, K.J., Chow, E., Eastwood, M.P., Ierardi, D.J., Klepeis, J.L., Kuskin, J.S., Larson, R.H., Lindorff-Larsen, K., Maragakis, P., Moraes, M.A., Piana, S., Shan, Y., Towles, B.: Millisecond-scale molecular dynamics simulations on Anton. pp. 39:1–39:11. ACM, New York, NY, USA (2009)
104. Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R.O., Eastwood, M.P., Bank, J.A., Jumper, J.M., Salmon, J.K., Shan, Y., Wriggers, W.: Atomic-level characterization of the structural dynamics of proteins. Science **330**, 341–346 (2010)
105. Sherman, W., Day, T., Jacobson, M.P., Friesner, R.A., Farid, R.: Novel procedure for modeling ligand/receptor induced fit effects. J. Med. Chem. **49**, 534–553 (2005)
106. Shirts, M.R., Bair, E., Hooker, G., Pande, V.S.: Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. Phys. Rev. Lett. **91**, 140,601 (2003)
107. van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., Berendsen, H.J.C.: Gromacs: fast, flexible, and free. J. Comp. Chem. **26**, 1701–1718 (2005)
108. Stone, J.E., Phillips, J.C., Freddolino, P.L., Hardy, D.J., Trabuco, L.G., Schulten, K.: Accelerating molecular modeling applications with graphics processors. J. Comp. Chem. **28**, 2618–2640 (2007)
109. Summa, V., Petrocchi, A., Bonelli, F., Crescenzi, B., Donghi, M., Ferrara, M., Fiore, F., Gardelli, C., Gonzalez Paz, O., Hazuda, D., Jones, P., Kinzel, O., Laufer, R., Monteagudo, E., Muraglia, E., Nizi, E., Orvieto, F., Pace, P., Pescatore, G., Scarpelli, R., Stillmock, K., Witmer, M., Rowley, M.: Discovery of raltegravir, a potent, selective orally bioavailable HIV-integrase inhibitor for the treatment of HIV-AIDS infection. J. Med. Chem. **51**, 5843–5855 (2008)

110. Swope, W.C., Andersen, H.C., Berens, P.H., Wilson, K.R.: A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small water clusters. J. Chem. Phys. **76**, 637 (1982)
111. Tajkhorshid, E., Nollert, P., Jensen, M.Ø., Miercke, L.J.W., O'Connell, J., Stroud, R.M., Schulten, K.: Control of the selectivity of the aquaporin water channel family by global orientational tuning. Science **296**, 525–530 (2002)
112. Teodoro, M.L., E., K.L.: Conformational flexibility models for the receptor in structure based drug design **9**, 1419–1431 (2003)
113. Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., Jr., A.D.M.: CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J. Comp. Chem. **31**(4), 671–690 (2010)
114. Verlet, L.: Computer 'experiments' on classical fluids: I. Thermodynamical properties of Lennard-Jones molecules. Phys. Rev. **159**, 98–103 (1967)
115. Voter, A.: Hyperdynamics: accelerated molecular dynamics of infrequent events. Phys. Rev. Lett. **78**, 3908–3911 (1997)
116. Voter, A.: A method for accelerating the molecular dynamics simulation of infrequent events. J. Chem. Phys. **106**, 4665 (1997)
117. Wang, J., Cieplak, P., Kollman, P.: How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J. Comp. Chem. **21**, 1049–1074 (2000)
118. Wang, J., Wang, W., Kollman, P.A., Case, D.A.: Automatic atom type and bond type perception in molecular mechanical calculations. J. Mol. Graph. Model. **25**, 247–260 (2006)
119. Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., Case, D.A.: Development and testing of a general AMBER force field. J. Comp. Chem. **25**, 1157–1174 (2004)
120. Wang, Y., Shaikh, S.A., Tajkhorshid, E.: Exploring transmembrane diffusion pathways with molecular dynamics. Physiology **25**, 142–154 (2010)
121. Wereszczynski, J., McCammon, J.A.: Using selectively applied accelerated molecular dynamics to enhance free energy calculations. J. Chem. Theor. Comp. **6**, 3285–3292 (2010)
122. Williams, S., de Oliveira, C., McCammon, J.: Coupling constant ph molecular dynamics with accelerated molecular dynamics. J. Chem. Theor. Comp. **6**, 560–568 (2010)
123. Wong, C., McCammon, J.: Protein flexibility and computer-aided drug design. Annu. Rev. Pharm. Tox. **43**, 31 (2003)
124. Wong, C.F., McCammon, J.A.: Computer simulation and the design of new biological molecules. Isr. J. Chem. **27**, 211–215 (1986)
125. Xu, D., Williamson, M., Walker, R.: Advancements in molecular dynamics simulations of biomolecules on graphical processing units **6**, 2–19 (2010)
126. Zacharias, M., Straatsma, T., McCammon, J.: Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. J. Chem. Phys. **100**, 9025–9031 (1994)
127. Zhou, H.X., Gilson, M.K.: Theory of free energy and entropy in noncovalent binding. Chem. Rev. **109**, 4092–4107 (2009)
128. Zhu, F., Tajkhorshid, E., Schulten, K.: Collective diffusion model for water permeation through microscopic channels. Phys. Rev. Lett. **93**, 224,501 (2004). (4 pages)

The Many Faces of Structure-Based Potentials: From Protein Folding Landscapes to Structural Characterization of Complex Biomolecules

Jeffrey K. Noel and José N. Onuchic

1 Introduction

Structural biology techniques, such as nuclear magnetic resonance (NMR), x-ray crystallography, and cryogenic electron microscopy (cryo-EM), have provided extraordinary insights into the details of the functional configurations of biomolecular systems. Recent advances in x-ray crystallography and cryo-EM have allowed for structural characterization of large molecular machines such as the ribosome, proteasome, and spliceosome. This deluge of structural data has been complemented by experimental techniques capable of probing dynamic information, such as Förster resonance energy transfer (FRET) and stopped flow spectrometry. While these experimental studies have provided tremendous insights into the dynamics of biomolecular systems, it is often difficult to combine the low resolution dynamical data with the high-resolution structural data into a consistent picture. Computer simulation of these biomolecular systems bridges static structural data with dynamic experiments at atomic resolution (Fig. 1).

Since the first molecular dynamics simulations of bovine pancreatic trypsin inhibitor 35 years ago [38], molecular simulations have become indispensable tools in biophysics. Molecular dynamics simulations of biomolecules treat the molecule as a collection of classical particles interacting through a potential energy function called a force field [1]. The molecule's dynamics are propagated through time by

J.K. Noel
Department of Physics and Center for Theoretical Biological Physics,
University of California, La Jolla, CA 92093, USA
e-mail: jknoel@ucsd.edu

J.N. Onuchic (✉)
Department of Physics and Center for Theoretical Biological Physics,
Rice University, Houston, TX 77005, USA
e-mail: jonuchic@rice.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_2,

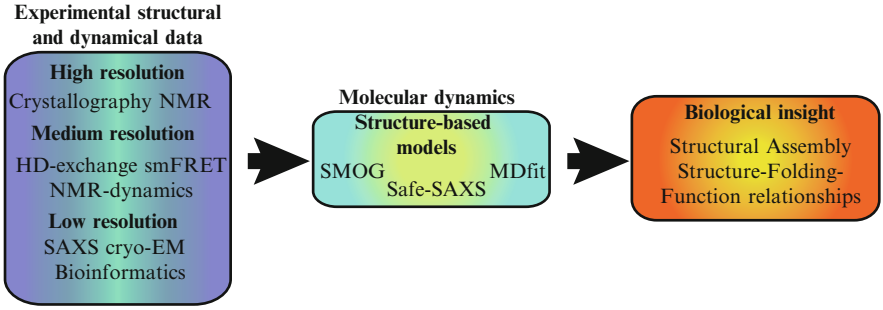

Fig. 1 Structure-based models bridge static high-resolution structural data with lower resolution dynamical and structural data at the all-atom level. Many experimental inputs can be combined to form a coherent picture of a biological process

numerical integration of Hamilton's equations resulting in a molecular trajectory. This trajectory can be used to gain a kinetic and thermodynamic understanding of the system. Simulations can be performed using empirically parameterized force fields that include explicit solvent. In principle, the chemistry-based representation should reproduce the structure and dynamics of a biomolecular system without requiring input from experimental structural data. In practice, making contact with experimental observables poses harsh challenges for these force fields both due to the level of accuracy required and the long time scales needed [54, 66]. In order to integrate experimental data in a consistent manner, biomolecular models with robust potential energy functions able to access long time scales are necessary. The energy landscape theory of protein folding provides the theoretical underpinning for *structure-based models* (SBM) [47]. These models impose a *native bias* by explicitly including structural data in the Hamiltonian. The structural data is derived from experimental techniques that are able to discern a representative structure of a molecule in a deep free energy basin, e.g., a protein native state. The native bias dramatically reduces the complexity of the resulting force field. These simplifications allow for a clear physical understanding of a system and open up biologically relevant timescales while retaining the essential dynamical features. SBM have been validated by their application to protein dynamics, such as folding, stretching, oligomerization, and functional transitions. Multiple experimental inputs can be naturally included, e.g., by extending the single native bias to include information from multiple conformers to explore conformational transitions. Fueled by the introduction of an all-atom (AA) SBM, prospective new applications for SBM are being explored in areas such as RNA folding, molecular machines, and prediction of protein–protein interactions. This chapter will present the basics of SBM and explain how a publicly available SBM, SMOG (Structure-based MOdels in GROMACS http://smog.ucsd.edu), has been used to explore the dynamics of systems as disparate as folding knots in proteins and accommodation in the ribosome.

2 Structure-Based Models

2.1 Foundations in Energy Landscape Theory

The inclusion of a native bias, the hallmark of a SBM, has a rigorous footing in the energy landscape theory of protein folding [8, 33, 47]. Protein folding is a self-organizing process whereby a protein transitions from a highly disordered ensemble (unfolded) to a structured ensemble (folded/native state). The relatively short timescale with which the folded state is reached implies that any competing nonnative states (traps) are shallow compared with the overall energy bias to folding. If these traps are sufficiently shallow, the nonnative interactions can be grouped into an effective diffusion [9, 17]. In addition, the uniqueness of the folded state implies that it corresponds to the global minimum in the free-energy landscape. The *principle of minimal frustration* states that evolution has achieved this folding robustness by selecting for sequences where the interactions present in the native structure are mutually supportive, i.e., attractive. The interactions are minimally frustrated or, in other words, maximally consistent. This organization leads to the protein folding on a *funneled landscape* where the energy on average decreases as it forms structures similar to the native structure.

Minimal frustration and the funneled energy landscape give the theoretical foundation for SBMs. A structure-based potential dramatically reduces the biomolecular Hamiltonian's complexity by stabilizing interactions that are spatially close in the native configuration. While real protein funnels have residual energetic frustration caused by nonnative interactions, the SBMs discussed here are "perfectly funneled" models, since in the force field *all* interactions stabilize the native structure. Nonnative interactions are strictly repulsive. In such a framework, any barriers to folding must be free energy barriers arising from the various ways energy and entropy compensate during folding. The ability of perfectly funneled models to reproduce experimental folding trends and mechanisms shows that geometrical effects like chain connectivity have an enormous influence on protein dynamics [5, 11, 47]. Since the precise energetics are secondary to the geometry of the protein molecule, this idea leads to the commonly held notion that geometry determines the folding mechanism.

Even though SBMs were formulated in the context of protein folding, their applications are widespread. Folding is only a first step in the lives of proteins which go on to perform a myriad of functions in the cell. The funneled energy landscape upon which the protein folds is the same landscape that controls functional protein motions. Multiple functional conformational states captured by experiment can be naturally included by extending the funneled landscape to have multiple basins. Structured RNAs must also have evolutionary pressure to reduce the level of frustration or they would encounter their own "Levinthal's paradox." The robust dynamics of large molecular complexes such as the ribosome and proteasome must

depend even less on the precise atomic energetic details and more on emergent properties controlled by the geometry of their constituents. While all these systems will have residual levels of frustration, the use of SBMs as a baseline is crucial to partition the global properties, those largely dependent on structure, from the details dependent on specific energetics.

2.2 Structure-Based Model as a Baseline

Simplified models have a long history of elucidating the organizing principles governing complex systems. A key question is how sensitive a model is to its underlying parameters. Determining the correct value for a parameter is often equally important as understanding the sensitivity to perturbations in that parameter. Since molecular geometry has a central influence on the motions leading to molecular function, simplified models based on low free energy structures are a natural starting point. The simplest models look at the normal modes of an energy landscape created by replacing all short range interactions in a native structure by Hookean springs [61]. These models can capture relevant rigid body motions. SBMs provide an important generalization by allowing the possibility for "cracking," [24, 25, 40, 68] allowing interactions to break and reform, since the springs are replaced by short range potentials. Thus, SBM can capture motion on all scales from native basin dynamics to unfolding.

The straightforward formulation of a structure-based potential allows for sensitivity analysis of the force field parameters [69] and their simplicity makes them extremely fast to compute. The force field is readily extensible allowing the introduction of complicated effects to be explored parametrically. For example, the effects of electrostatics can be explored by perturbative addition of Coulomb interactions [4,14,35], or the effects of solvent probed by the perturbative addition of desolvation barriers [12]. A crucial question in the protein folding field has been how proteins manage to achieve such smooth energy landscapes, or equivalently, why do AA empirical force fields and structure prediction schemes have difficulty achieving the level of specificity seen in proteins? Using structure-based potentials with AA geometries, we can begin to address this question. These models completely partition energetic effects from geometric effects, and through careful investigation, may discern to what extent energetics contribute to the apparent native specificity in protein structure, folding, and function. While processes like the formation of nonnative intermediates during folding [18, 53, 60] and protein misfolding are clearly cases that perfectly funneled SBM will be unable to fully describe, through adding complexity in a piecemeal fashion to a robust baseline model, a more complete understanding of the interplay between geometry and energy in even these complicated systems will result.

3 Implementation of Structure-Based Models

SBMs have a long history in the protein folding field. The folding dynamics of minimally frustrated sequences were first tested in lattice models. Bryngelson et al. [10] and Socci et al. [56] investigated a minimally frustrated lattice model with three types of beads. They found that the dynamics could be well described by diffusion along a small number of collective coordinates on an effective free energy surface defined by those coordinates. As the structural correspondence between cubic lattices and actual proteins is low, Nymeyer et al. implemented an off-lattice, coarse-grained model of a protein-like structure. They compared the folding dynamics of an energetically frustrated [62] versus a completely unfrustrated β-barrel [45]. They showed that the completely unfrustrated model, effectively a SBM, exhibited the characteristics of a good folder, specifically, having exponential folding kinetics on a funnel-shaped landscape that is robust to reasonable perturbations. Following these successes, Clementi et al. [15] introduced the popular "C_α model," which also had a coarse-grained representation of the protein. This model reproduced the transition-state ensembles (TSE) of several small two- and three-state proteins. The C_α model has since been adopted by several investigators to explore myriad topics in protein folding (see these references for some highlights [2, 11, 12, 22, 26, 28, 29, 52, 59]). The off-lattice geometry allowed clear representation of protein structures, making comparisons to experimentally determined dynamics possible. In order to capture geometric effects like side chain packing, Whitford et al. introduced an AA SBM [69]. This model is being used to represent proteins [69], RNA/DNA [64] and ligands in a consistent fashion for both dynamics [42, 43, 66] and molecular modeling [27, 50, 51]. These two models, AA and C_α, are currently in wide use and are available on the SMOG web server [44].

Before the two available models are described in detail, we review the key components common to any SBM. The defining characteristic is that the parameters are determined from a native structure. The potential V is composed of three contributions,

$$V = \underbrace{V^{\text{Bonded}} + V^{\text{Repulsive}}}_{\text{Maintain geometry}} + \underbrace{V^{\text{Attractive}}}_{\text{Tertiary structure}} . \tag{1}$$

V^{Bonded} includes interactions that maintain the covalently bonded structure and torsional angles. This term also ensures correct chirality. $V^{\text{Repulsive}}$ contains spherically symmetric hard wall repulsions that enforce excluded volume and prevent chain crossings. Collectively, these two terms maintain the protein's structure and allowed conformational diversity. $V^{\text{Attractive}}$ contains short range, attractive interactions between atoms (or residues if coarse graining) close in the native state. These interactions are the core of the SBM and are discussed in the next section.

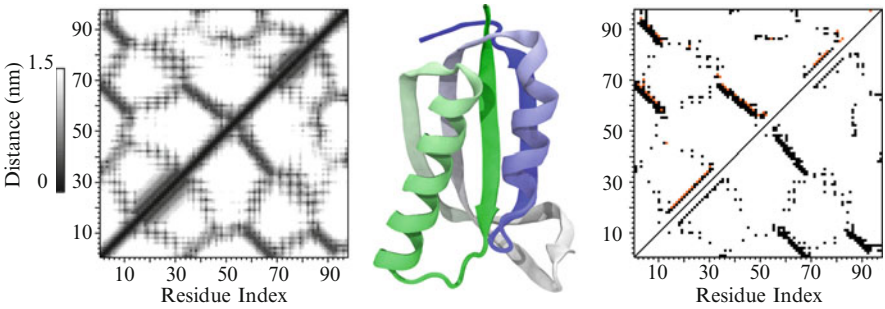

Fig. 2 Native contact map of ribosomal protein S6 (PDB code: 1RIS). Structure of the α/β protein S6 is shown with the N-terminus (residue 1) colored *green*. *Left panel* shows the proximity of the nearest atomic contact for each residue pair up to a maximum of 1.5 nm. *Right panel* compares two coarse-grained native contact maps. A pair of residues are considered a native contact if they share a native atom–atom contact. *Top triangle*: 6Å cutoff. *Bottom triangle*: a 6Å cutoff with geometric occlusion using Shadow [44]. The contacts which are excluded by Shadow are colored *orange*

3.1 Native Contact Map

Atoms that are spatially near in the native state are considered *contacts* and together the set of all contacts composes a *native contact map* (Fig. 2). A contact map is a binary symmetric matrix that encodes which atom pairs ij are given attractive interactions in the SBM potential. In the context of a SBM, the native contact map should approximate the distribution of stabilizing enthalpy in the native state that is provided by short range interactions like van der Waals forces, hydrogen bonding, and salt bridges. Any long range interactions or nonlocal effects are taken into account in a mean field way through the native bias. For example, the hydrophobic effect is encoded by the density of native contacts being larger on the interior of the protein than on the surface.

Methods for constructing contact maps are based on the heavy atom distances in the native structure. There are three widely used techniques: heavy atom cutoffs [16], van der Waals radii overlaps [15, 55, 58], and geometric occlusions [44, 71]. Heavy atom cutoff maps define a cutoff distance R_C, typically 4–6.5Å, and consider all heavy atoms within R_C of each other in contact. van der Waals radii cutoff maps increase the radii of all the heavy atoms by either a multiplicative constant ($\sim$1.25) or an additive constant ($\sim$1.4Å). Any atoms that then overlap are considered to be in contact. The rationale for the multiplicative constant comes from overlapping electron clouds, or "soft spheres." The additive constant represents the size of one water molecule. Half the diameter of water is added to each atomic radius, and if atoms then overlap it means that a water cannot be placed between them. The set of atom pairs excluding water from each other are presumed to interact, and thus considered contacts (the software package CSU [55] uses this approach). Geometric occlusion maps take the output of a heavy atom cutoff contact map, $R_C \gtrsim 6$Å, and then remove any contacts that are geometrically obstructed. $R_C > 4.5$Å introduces

many unphysical or "occluded" contacts where atoms are interacting through an intervening atom. Since these interactions are mostly induced dipole interactions, electron screening effects should dampen the occluded interactions. van der Waals radii overlaps and geometric occlusions both provide the short range, first layer of atomic contacts. Geometric occlusion maps add longer range water- or cofactor-mediated contacts up to the cutoff distance. The advantage of geometric occlusion is that atoms separated by voids, or those coordinated by water and metals not explicitly included in a protein simulation, can be accounted for without introducing spurious occluded contacts.

van der Waals radii overlaps and geometric occlusions provide contact maps that behave similarly in protein-folding simulations. Simulations with these maps consistently predict cooperative, protein-like transitions for globular proteins. They also reproduce thermodynamic folding intermediates for proteins with known intermediates [15]. In the authors' experience, heavy atom cutoff maps are not robust in protein-folding simulations. Short-range cutoffs miss longer range contacts, leaving the contact map sensitive to the precise packing of the native state, and thus overweight regions of the contact map. This reduces the cooperativity of the transition, leading to spurious thermodynamic intermediates. Longer cutoffs reduce the sensitivity to packing by adding larger numbers of contacts, but this introduces many unphysical contacts where atoms are interacting through an intervening atom. This overabundance of contacts, by reducing the relative strength of each individual contact, also tends to decrease cooperativity. SMOG uses a geometric occlusion contact map called Shadow [44] for proteins. On the SMOG server, the default for RNA/DNA systems is a 4Å heavy atom cutoff, but there are indications that Shadow is also sensible for RNA folding.

Single bead per residue coarse-grained contact maps are generally derived from the corresponding atomic structure. Coarse-grained contact maps could conceivably be generated from the coarse-grained structure using C_α–C_α distance cutoffs (generally 7–12Å). Since the coarse-grained structure ignores side chain packing, this metric poorly predicts the enthalpic contributions to the native state [39]. For the C_α model, SMOG considers two residues in contact if they share at least one atomic contact.

3.2 SBM Potential

The SMOG structure-based forcefield is available in two grainings, a coarse-grained (C_α) model [15] and AA model [64, 69].

3.2.1 C_α Model

The C_α model coarse grains the protein as single bead of unit mass per residue located at the position of the α-carbon. $\vec{\mathbf{x}}_\mathbf{0}$ denotes the coordinates (usually obtained

from the Protein Data Bank http://www.rcsb.org) of the native state and any subscript 0 signifies a value taken from the native state. The potential is given by

$$V_{C\alpha}(\vec{\mathbf{x}}, \vec{\mathbf{x_0}}) = \sum_{\text{bonds}} \epsilon_r (r - r_0)^2 + \sum_{\text{angles}} \epsilon_\theta (\theta - \theta_0)^2 + \sum_{\text{backbone}} \epsilon_{\text{D}} F_{\text{D}}(\phi - \phi_0) + \sum_{\text{contacts}} \epsilon_{\text{C}} C(r_{ij}, r_0^{ij}) + \sum_{\text{non-contacts}} \epsilon_{\text{NC}} \left(\frac{\sigma_{\text{NC}}}{r_{ij}} \right)^{12}, \tag{2}$$

where the dihedral potential F_{D} is,

$$F_{\text{D}}(\phi) = [1 - \cos(\phi)] + \frac{1}{2}[1 - \cos(3\phi)]. \tag{3}$$

The coordinates $\vec{\mathbf{x}}$ describe a configuration of the α-carbons, with the bond lengths to nearest neighbors r, three body angles θ, four body dihedrals ϕ, and distance between atoms i and j given by r_{ij}. C denotes the contact potentials given to the native contacts (see Sect. 3.2.3). Protein contacts that are separated by less than 3 residues are neglected. Excluded volume is maintained by a hard wall interaction giving the residues an apparent radius of $\sigma_{\text{NC}} = 4$Å. The native bias is provided by using the parameters from the native state $\vec{\mathbf{x}}_{\mathbf{0}}$. Setting the energy scale $\epsilon \equiv k_{\text{B}} T^* = 1$, the coefficients are given the homogeneous values: $\epsilon_r = 100\epsilon$, $\epsilon_\theta = 40\epsilon$, $\epsilon_{\text{D}} = \epsilon_{\text{C}} = \epsilon_{\text{NC}} = \epsilon$.

3.2.2 All-Atom Model

The AA potential is quite similar to the C_α potential, although representing the AA geometry requires some additional terms. In the AA model, all heavy (nonhydrogen) atoms are explicitly represented as beads of unit mass, so each interaction is now between atoms as opposed to residues. Bonds, angles, and dihedrals therefore have their traditional chemical meanings. In each residue, there is an additional backbone dihedral and, except for glycine, many side chain dihedrals. Improper dihedrals maintain backbone chirality and, when necessary, side chain planarity. The AA potential V_{AA} is

$$V_{\text{AA}}(\vec{\mathbf{x}}, \vec{\mathbf{x_0}}) = \sum_{\text{bonds}} \epsilon_r (r - r_0)^2 + \sum_{\text{angles}} \epsilon_\theta (\theta - \theta_0)^2 + \sum_{\text{impropers/planar}} \epsilon_\chi (\chi - \chi_0)^2 + \sum_{\text{backbone}} \epsilon_{\text{BB}} F_{\text{D}}(\phi - \phi_0) + \sum_{\text{sidechains}} \epsilon_{\text{SC}} F_{\text{D}}(\phi - \phi_0) + \sum_{\text{contacts}} \epsilon_{\text{C}} C(r_{ij}, r_0^{ij}) + \sum_{\text{non-contacts}} \epsilon_{\text{NC}} \left(\frac{\sigma_{\text{NC}}}{r_{ij}} \right)^{12}. \tag{4}$$

As in the C_α model, the coefficients are given homogeneous values: $\epsilon_r = 100\epsilon$, $\epsilon_\theta = 20\epsilon$, $\epsilon_\chi = 40\epsilon$, $\epsilon_{\rm NC} = 0.01\epsilon$, and $\sigma_{\rm NC} = 2.5$Å. The effective repulsive size for the atoms becomes $\sigma_{\rm eff} = (0.01)^{1/12}\sigma_{\rm NC} \approx 1.7$Å. Again, protein contacts that are separated by less than 3 residues are neglected. A technical issue is normalizing the dihedral energy around each bond. When assigning dihedral strengths, we first group dihedral angles that share the middle two atoms. For example, in a protein backbone, one can define up to four dihedral angles that possess the same C−C_α covalent bond as the central bond. Each dihedral in the group is scaled by $1/N_{\rm D}$, where $N_{\rm D}$ is the number of dihedral angles in the group.

Two ratios determine the distribution of dihedral and contact energies, contact to dihedral ratio $R_{C/D}$ and backbone to side chain ratio $R_{BB/SC}$. In proteins $R_{BB/SC} = \epsilon_{BB}/\epsilon_{SC} = 2$ [69] and in RNA/DNA $R_{BB/SC} = \epsilon_{BB}/\epsilon_{SC} = 1$ [64]. The contacts and dihedrals are scaled relative to their total contributions, $R_{C/D} = \frac{\sum \epsilon_C}{\sum \epsilon_{BB} + \sum \epsilon_{SC}} = 2$. Lastly, the total contact and dihedral energy is set equal to the number of atoms $\epsilon N_{\rm atoms} = \sum \epsilon_C + \sum \epsilon_{BB} + \sum \epsilon_{SC}$. This choice gives folding temperatures near 1 in reduced units ensuring a consistent parameterization.

Notice that every term is based on the native structure except the noncontact excluded volume term. In the C_α model, all the residues have a homogeneous shape, but in the AA model each residue has its unique geometry explicitly represented. This gives the AA model structure independent sequence information that adds heterogeneity to the model. This heterogeneity adds geometric frustration to the model, since interactions can only be satisfied if the side chains are correctly oriented [43]. A question of current interest is whether this sequence-dependent information adds constraints to the folding dynamics, allowing the native bias to be relaxed [3, 69].

3.2.3 Contact Potential

All of the pair interactions defined in the native contact map interact through a short range, attractive potential, denoted in the SBM potential by $C(r_{ij}, r_0^{ij})$. The contact potential has a minimum at r_0^{ij}, the distance between the pair in the native state. Traditionally, a contact is defined through a Lennard–Jones (LJ) type potential, since the LJ shape is readily available in molecular dynamics packages. In the C_α model a "10–12" LJ potential is used for contacts with the minimum set at the separation between the C_α pair in the native state r_0^{ij},

$$C_{\rm CA}(r_{ij}, r_0^{ij}) = 5\left(\frac{r_0^{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{r_0^{ij}}{r_{ij}}\right)^{10}, \tag{5}$$

and in the AA model a "6–12" LJ potential with the minimum set at the separation between a contacting atomic pair in the native state,

$$C_{\rm AA}(r_{ij}, r_0^{ij}) = \left(\frac{r_0^{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{r_0^{ij}}{r_{ij}}\right)^{6}. \tag{6}$$

Different LJ potentials are used because the native contact distances r_0^{ij} can be much longer in the C_α model. The contacts are coarse-grained to be between the C_α atoms, which can be as distant as 14Å. The r^{-6} is much broader than the r^{-10} and can lead to unphysical structures in unfolded states as native pairs interact at long distances.

The LJ potentials are well tested and work for many systems, but they have limitations for certain applications because the LJ potential has an excluded volume that moves with the minimum. The effective size of two atoms in contact grows with r_0^{ij}. This additional excluded volume has little effect on the entropy of unfolded conformations since mostly noncontacts are interacting, but has a large effect on the entropy of the folded ensemble where most contacts are formed. In cases where the user wants to control the excluded volume term [32,43], an attractive Gaussian well coupled with a fixed hard wall-excluded volume is used,

$$C_{\mathrm{G}}(r_{ij}, r_0^{ij}) = \left(1 + \left(\frac{\sigma_{\mathrm{NC}}}{r_{ij}}\right)^{12}\right)\left(1 + G(r_{ij}, r_0^{ij})\right) - 1, \tag{7}$$

where

$$G(r_{ij}, r_0^{ij}) = -\exp\left[-(r_{ij} - r_0^{ij})^2/(2\sigma^2)\right]. \tag{8}$$

This unusual construction anchors the depth of the minimum at -1. The width of the Gaussian well σ is determined to model the variable width of the LJ potential. $C_{\mathrm{AA}}(1.2r_0^{ij}, r_0^{ij}) \sim -1/2$ so σ is defined such that $G(1.2r_0^{ij}, r_0^{ij}) = -1/2$ giving $\sigma^2 = (r_0^{ij})^2/(50 \ln 2)$. If σ_{NC} is significantly smaller than r_0^{ij}, (7) reduces to the more pedagogical form,

$$C_{\mathrm{G}}(r_{ij}, r_0^{ij}) \to \left(\frac{\sigma_{\mathrm{NC}}}{r_{ij}}\right)^{12} + G(r_{ij}, r_0^{ij}) \quad \text{for } \sigma_{\mathrm{NC}} \ll r_0^{ij}. \tag{9}$$

The flexibility of the Gaussian approach also allows for multiple basin contact potentials for energy landscapes with multiple minima (see Sect. 4.3). Using multiple LJ potentials with different locations of the minima is not a viable option because the longest LJ potential would occlude the others with its excluded volume term. A multibasin Gaussian potential C_{MB} for minima taken from two structures $\vec{\mathbf{x}}_\alpha$ and $\vec{\mathbf{x}}_\beta$ is given by [32],

$$C_{\mathrm{MB}}(r_{ij}, r_\alpha^{ij}, r_\beta^{ij}) = \left(1 + \left(\frac{\sigma_{\mathrm{NC}}}{r_{ij}}\right)^{12}\right)\left(1 + G(r_{ij}, r_\alpha^{ij})\right)\left(1 + G(r_{ij}, r_\beta^{ij})\right) - 1. \tag{10}$$

Analogous to (7), this construction fixes the depth of both minima at -1.

All of the various potential shapes are presented in Fig. 3. It should be noted that the folding temperature (defined in Sect. 4.1.1) is typically 0.2–0.3 reduced units higher for the Gaussian potential as compared to LJ because the extra excluded volume in the LJ potential destabilizes the native state.

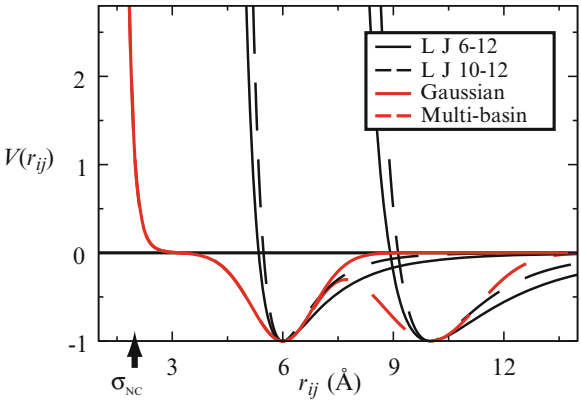

Fig. 3 Comparison of Lennard–Jones and Gaussian contact potentials. *Black curves* show LJ contact potentials with minima at 6Å and 10Å. The Gaussian contact potential shown in *red* has an excluded volume σ_{NC} that can be set independently of the location of the minimum. The *dotted red line* shows how the Gaussian contact would change as another minimum at 10Å is added

3.3 Molecular Dynamics with SBM

Molecular dynamics uses Newtonian mechanics to evolve the motions of atoms in time. The interactions defined in the SBM potential define the various forces on the atoms since force is given by the negative gradient of the potential energy. The system is evolved through time in discrete steps. The NVT canonical ensemble is maintained using a thermostat. Thermostats including a drag term, such as stochastic dynamics or Langevin dynamics are used for implicit solvent systems like SBMs. Velocity-rescaling thermostats can introduce heating artifacts when not coupled to an explicit solvent [41]. Langevin dynamics has been used to model the viscosity of a solvent [25, 57]. The output of a molecular dynamics simulation is a trajectory, a time-ordered series of snapshots of the atomic coordinates. The trajectory can be analyzed as a function of time to uncover kinetic properties or, by application of the ergodic theorem, as an ensemble to compute thermodynamic properties.

A molecular dynamics trajectory contains the coordinates of all the atoms in the system, a massive amount of information. Therefore, the trajectory is reduced down to one or a few reaction coordinates that monitor the progress of the dynamics under investigation. For protein folding, a useful reaction coordinate would differentiate between the unfolded ensemble, folding intermediates, and the folded ensemble. A reaction coordinate for studying a conformational transition would differentiate the various conformers. A natural reaction coordinate for SBMs is Q, the fraction of native contacts formed. A contact between the native pair ij is considered formed if it satisfies $r_{ij} < \gamma r_0^{ij}$, where $\gamma \approx 1.2$–1.4. In protein folding, low Q would correspond to the unfolded ensemble, medium Q would contain the transition state ensemble (TSE) and any intermediates, and high Q the folded ensemble. To investigate a conformational transition between two structures A and B, monitoring switching between high Q_A and high Q_B would indicate transitions. Other possible reaction coordinates are root mean square deviation from a reference structure or radius of gyration. An exciting possibility is to monitor the position of an explicitly represented FRET probe in order to compare with experimental data [66].

After the choice of reaction coordinate is made, the value of the coordinate during the trajectory (or several concatenated trajectories) can be histogrammed to obtain a potential of mean force (PMF) along the reaction coordinate. If the chosen coordinate adequately separates two basins, it can be used to identify the transition state at the peak on the free energy landscape. Q has been shown to be a suitable coordinate for protein transitions and thus the peaks in $F(Q)$ can be identified as TSEs [13] (see Fig. 5). Great care must be exercised when making quantitative predictions of thermodynamic and kinetic quantities from simplified models. The kinetics of the system are not simply determined by the free energy landscape, but are highly dependent on diffusion rates. Diffusion rates vary for different molecular systems and must be calibrated separately. For discussion of diffusion in SBM see [30,46,66]. Secondly, the precise values of free energy barriers and thermal stability are a fine balance and depend on the details of the SBM potential. This said, given a constant parameterization, kinetic and thermodynamic quantities tend to scale in a consistent fashion. Fast-folding proteins will consistently have smaller free energy barriers than slow-folding proteins [11, 69]. Some quantities are robust to perturbations, in particular the TSE and other so-called geometrical features of the energy landscape [32, 69].

3.4 SMOG: Automated Generation of SBM

Molecular dynamics simulations have benefited from years of research on computer algorithms constructed with one goal in mind: speed. Molecular dynamics suites like GROMACS [23], NAMD [49] and Desmond [7], package all the necessary algorithms to run stable molecular dynamics and the ability to scale the calculations to thousands of processors. These packages have made homegrown molecular dynamics codes built to run SBMs obsolete. SMOG, Structure-based models in GROMACS, is a publicly available web server located at http://smog.ucsd.edu [44]. Any PDB structure consisting of standard amino acids, RNA, DNA, and common ligands, can be uploaded to SMOG, which outputs the necessary coordinate, topology, and parameter files to run a SBM in GROMACS. This provides the flexibility necessary to implement efficient and highly scalable SBMs. SMOG in conjunction with GROMACS version 4.5 scales easily to 128 processors when simulating a ribosome, $\sim$150,000 atoms. Protein-folding simulations of much smaller systems scale to $\sim$100 atoms per core on a single motherboard.

3.5 Choosing a Graining: C_α or All-Atom

The C_α and AA model are both able to describe the properties of the molecular scaffold's geometry. When comparing the two models, C_α and AA, the main advantage of C_α is its speed. Because the AA model has roughly eight times more

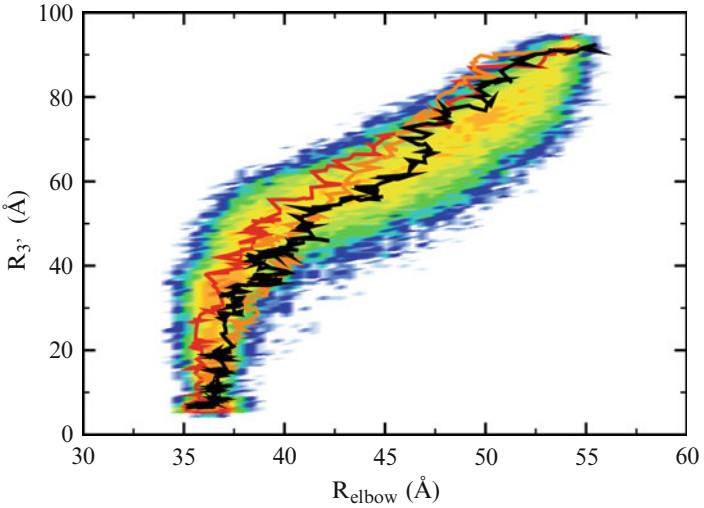

Fig. 4 Comparison of SBM and explicit solvent simulations of tRNA accommodation in the ribosome. Trajectories of three 4 ns explicit solvent-targeted molecular dynamics (TMD) overlay the probability distribution of 704μs structure-based TMD runs. With such a short sampling time, the explicit solvent TMD is dominated by steric interactions between the ribosome and the tRNA. The SBM naturally captures the sterics and is consistent with the detailed model. $R_{3'}$ and R_{elbow} monitor the position of the tRNA along the accommodation pathway. Simulations were started from the A/T state (high $R_{3'}$ and R_{elbow}) and stopped at the accommodated (A/A) state. See [66] for details

atoms and has slower diffusion due to side chain interactions, the C_α model runs significantly faster than AA. This speed is important for studying processes with large barriers, like folding and oligomerization. AA can narrow the speed gap with parallelization, but not close it completely. Nonetheless, AA has been used to fold small single domain proteins [69] and even proteins with complex topologies [43]. Many processes without large activation barriers, e.g., native basin dynamics, have energy landscapes that are easily sampled, and thus the performance hit of AA is of no consequence.

The explicit representation of atomic coordinates is advantageous, even for simplified models like SBM. A clear benefit is acting as a bridge between minimalist models and empirical force fields. Any conformations realized during a simulation of an AA SBM can be compared with, and used as input for, empirical force fields with an explicit solvent. Since the sterics are correct, any process that is dominated by large-scale structural fluctuations should be well represented by an AA SBM [42, 66]. Figure 4 shows targeted molecular dynamics (TMD) simulations of the tRNA accommodation process in the ribosome, a massive ribonucleoprotein molecular machine ($\sim$2.4 MDa). The trajectories from explicit solvent simulations overlay the AA SBM trajectories. On a smaller scale, the AA geometry opens the door to studying side chain degrees of freedom during folding and binding

simulations. Constricted conformations like polypeptide slipknots, found in coarse-grained models, are shown to be sterically possible with the AA geometry [43]. Lastly, the AA geometry allows a clear way to add perturbative nonnative chemical effects like hydrogen bonding [3] and partial charges.

4 Applications

SBMs are being applied to diverse problems, and in the remaining sections we describe a representative sample of how perfectly funneled SBMs are currently in use. In each case, the SBM can be constructed and implemented using SMOG and GROMACS. In Sects. 4.1–4.3, molecular dynamics is used to describe a system at thermodynamic equilibrium. In this case, it is necessary to adequately sample configuration space until the quantities of interest have converged. Finally, in Sect. 4.4 molecular dynamics is used to find deep energetic minima in perturbed structure-based potentials for molecular modeling applications.

4.1 Folding

4.1.1 Protein Folding

The most-established application of SBM is to the study of protein folding. Determining the TSE, the shape and size of free energy barriers, and the existence of folding intermediates are all topics of interest. Figure 5 shows the result of AA SBM folding simulations for two of the most thoroughly studied proteins, chymotrypsin inhibitor-2 (CI2) and the SH3 domain. These two proteins are two-state folders, meaning the protein only populates two basins spanned by a cooperative transition.

Figure 5a,d shows representative traces of Q versus time during constant temperature molecular dynamics near folding temperature T_F. T_F is the temperature such that the folding and unfolding basins are equally populated. Simulations are performed at T_F because it maximizes the sampling rate of the folding transition. T_F is determined by running simulations at high and low temperatures, and iteratively converging on a temperature where both folding and unfolding is observed. Q is defined as the fraction of native residue pairs with at least one atom–atom contact within 1.2 times its native separation. Alternative definition of Q, such as the fraction of atom–atom contacts formed, may shift the locations of basins in the resulting free energy landscape, but will preserve the heights of any barriers.

Q traces from long molecular dynamics trajectories at various temperatures can be combined using weighted histogram analysis (WHAM) [31], to obtain an optimal density of states. The density of states can then be used to extrapolate $F(Q)$ at any temperature (Fig. 5b, e). Always, care must be taken to ensure that the trajectories

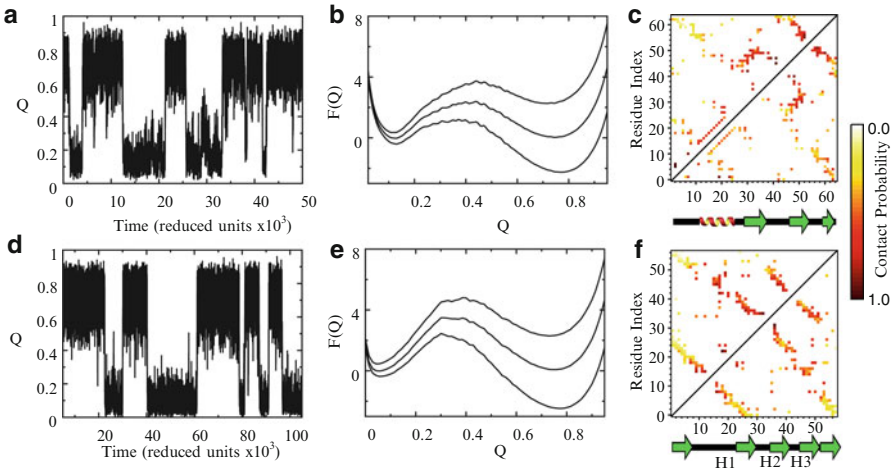

Fig. 5 All-atom structure-based simulations of two state-folding proteins CI2 (*top*) and SH3 domain (*bottom*). PDB codes: 1FMK, 1YPA. (**a,d**) The reaction coordinate Q plotted as a function of time for a typical simulation near T_F. Both proteins exhibit transitions between a folded ensemble at $Q \sim 0.8$ and an unfolded ensemble at $Q \sim 0.1$. (**b,e**) Free energy $F(Q)$ for temperatures $0.98T_F$, T_F, and $1.02T_F$ calculated by weighted histogram analysis of long constant temperature MD trajectories. A set of "long" trajectories typically contain 30 folded to unfolded transitions. (**c,f**) Transition state ensemble (TSE) for the two proteins. Contact formation probabilities are calculated by an unweighed average of all configurations $0.40 < Q < 0.45$. The *upper triangle* shows results from the C_α model and the *lower triangle* shows the AA model. Secondary structure is denoted below the contact maps as are the positions of the three hairpin turns in SH3. CI2 has a diffuse TSE that resembles the native state. The contact probability is largely predicted by sequence separation. SH3 has a more polarized TSE with contacts from the first ten residues largely absent. For both proteins, the introduction of energetic and structural heterogeneity through the AA geometry creates a more specific and less diffuse TSE. The simulations were prepared using SMOG v1.0.6 [44] with default parameters

reflect equilibrium. One easy method is to chop all trajectories in half and verify that $F(Q)$ and the TSE are the same for both halves. The TSE is the ensemble of structures that compose the bottleneck to folding. CI2 and SH3 each have a single TSE that connects the unfolded state to the folded state defined by the structure populating the top of $F(Q)$. Figure 5c,f shows the average contact maps of the structures with $0.4 < Q < 0.45$. The contact formation probabilities can be connected to Φ-value analysis, an experimental technique that estimates the contribution of a particular residue's contacts to the TSE [19]. In simulation, Φ_i is given by

$$\Phi_i = \frac{P_i^{\mathrm{TSE}} - P_i^{\mathrm{U}}}{P_i^{\mathrm{F}} - P_i^{\mathrm{U}}}, \tag{11}$$

where P_i is the probability that residue i forms its contacts and U/F refers to the unfolded/folded ensembles [36]. Φ_i near 1 means that residue i is very native-like in the TSE and a Φ_i near 0 means that residue i is still unfolded in the TSE.

Since the TSE is a simple average over structures, it can hold hidden complexity. For some proteins, the TSE is composed of multiple routes through the TSE [6, 22]. Consider SH3; its TSE could be composed of two routes, a major route where hairpin 2 and hairpin 3 form first and a minor route where hairpin 1 and hairpin 2 form first (Fig. 5f). Multiple routes can be identified by clustering the contact maps of TSE structures using the number of shared contacts as a similarity measure [6]. These routes represent entropically viable routes through the TSE. Thus, two real proteins that fold to the same structure may follow seemingly very different paths due to minor energetic differences.

4.1.2 Multimeric Folding and Binding

Many important biological processes are regulated by the homo- or hetero-oligomers that are formed when proteins bind [70]. A large survey of protein dimers showed that the binding mechanisms found in experiments were reproduced by SBMs [36], which gives strong evidence that protein binding is controlled by protein geometry. The energy landscapes of these proteins exhibited a rich variety of folding routes and binding mechanisms. The interplay of folding and binding can be explored in SBMs by introducing interface contacts into the native contact map. The contact map of crystallographic structures of protein dimers are analyzed in the same way as for monomers, atoms spatially close between the protomers are considered native contacts. Folding trajectories of protomers A and B will have three relevant order parameters, Q_A, Q_B, and Q_{AB}. Note that when analyzing the TSE and folding routes of homo-oligomeric proteins, clustering the TSE is crucial [6]. This is because the structural symmetry is broken by the requirement of labeling the protomers, i.e., protomer A folds then binds protomer B is the same route as B folds then binds A.

Observing binding in simulations is complicated by the entropy loss of binding. In order to observe binding events, the effective concentration of monomers is often much higher than in vivo. The concentration of monomers is imposed either by a linker between the monomers [36], periodic boundary conditions [64], or an umbrella potential [6, 43, 52] (all available in GROMACS). The umbrella potential would be implemented as a harmonic center of mass constraint, making the simulated potential

$$V_{\text{dimer}} = V_{\text{AA}} + k\left(r_{\text{CM}} - r_0^{\text{CM}}\right)^2, \tag{12}$$

where r_{CM} is the distance between the centers of mass and r_0^{CM} is the distance in the native state. k is calibrated to be as weak as possible while still observing binding. Varying k can model varying protomer concentration. The stability of the dimer versus the monomers can be controlled by scaling the strength of the interface contacts.

4.2 Native Basin Dynamics

Entropically driven motions accessible via thermal fluctuations are important components of functional protein dynamics [20]. These motions are difficult or impossible to intuit from rigid crystallographic structure analysis [42]. Analysis of small-angle x-ray scattering (SAXS) on C-terminal Src Kinase (Csk) indicated that Csk occupies extended conformations in solution, whereas the crystal structure showed a compact arrangement of Csk's SH2, SH3, and kinase domains [27]. Typically, a candidate structure for the protein structure is determined by fitting a rigid body model to the SAXS data, but this presumes that Csk assumes a relatively static structure in solution. In order to characterize the Csk solution structure, constant temperature molecular dynamics simulations of the Csk native basin were performed using the AA SBM. Theoretical scattering curves were computed from the resulting native ensemble and compared with the experimental scattering data. Jamros et al. [27] showed that in all cases, theoretical scattering curves generated from mixed populations of Csk structures fit the empirical SAXS data better than any rigid model. This suggests that Csk populates a broad ensemble of structures in solution, adopting conformations not observed in the crystal structure. More pertinently, an SBM is able to suggest a solution ensemble of structures for Csk using only information from the crystal structure. This procedure, termed Safe-SAXS, should be widely applicable to analyzing solution structures of biological macromolecules.

4.3 Multiple Basin Models

When a protein is able to be crystallized in substantially different conformations, it implies the energy landscape has multiple minima. This behavior can be seen in systems with a high degree of structural symmetry. A dual basin-funneled landscape solved the mystery of the Rop dimer, a dimer of two helix bundles that switched from a parallel arrangement to an antiparallel arrangement upon optimization of the hydrophobic core [21, 34, 52]. An SBM was used that combined the two crystal structure contact maps into a single native contact map. Thermodynamic sampling of the landscape showed that the parallel and antiparallel structures were of similar stability, so small experimental perturbations could tip the balance between the structures [52].

Combining multiple structures into a single landscape has also been used to study conformational transitions in adenylate kinase (AKE) [26, 67, 68]. AKE has two domains, LID and NMP, that must undergo large conformational changes during its enzymatic function (Fig. 6). The conformational change is captured by two crystal structures, one in the open state and the other in a closed state, with native contact maps M_{O} and M_{C}, respectively. The contacts that are in both maps is given by

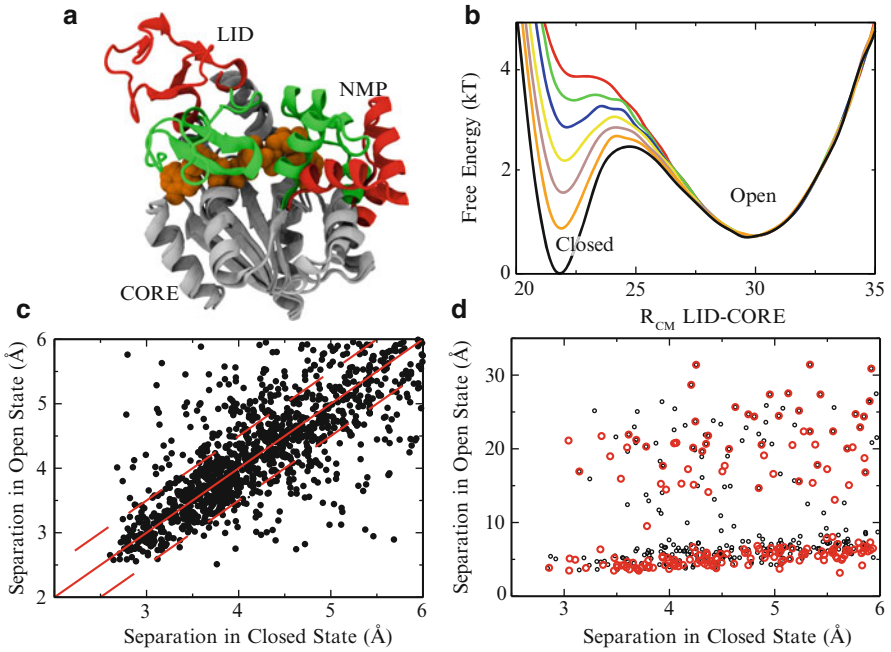

Fig. 6 Modeling conformational transitions in adenylate kinase (AKE). (**a**) AKE contains two domains, NMP and LID, that undergo >25Å motions between open (*red*) and closed (*green*) states. These motions are coupled with ligand (shown as *orange spheres*) binding as it catalyzes ATP+AMP⇆2ADP. The model is built using structures with PDB codes 1AKE and 4AKE. (**b**) The relative occupation of the closed and open states can be tuned to experimental data by varying the strength of the subset of contacts only existing in the closed state M_C. M_{BB} is scaled by 0.6 (*red*) to 1.2 (*black*) relative to the open contacts. (**c**) The subset of atomic contacts existing in both states M_{same}. The *dotted lines* designate a deviation of less than 0.5Å between states. Contacts that have significant shifts between structures may impart strain on the protein and can be included with double minima Gaussian potentials. (**d**) The subset of atomic contacts existing only in the closed state. Black circles show contacts of atoms in the LID domain and red circles show contacts of atoms in the NMP domain. See [68] for details

$M_{same} = M_O \bigcap M_C$ and the complement of M_{same} are the contacts that are in either map but not both M_{diff}. Results from a SBM with native contact map M_{diff} is shown in Fig. 6b. The relative stabilities of the two states can be easily tuned in the SBM. The distance between contacts that exist in both states (Fig. 6c) may change between structures and can be included with double minima Gaussian potentials (Sect. 3.2.3). How to handle multiple dihedral angle values is less obvious. Whitford et al. [68] simply used the dihedrals from the open state, viewing the closed state as an excitation of the open state. Similar methods have been used to look at conformational changes in protein kinase A [24] and kinesin [25].

4.4 Molecular Modeling

SBM are structurally robust, which makes them ideal candidates for molecular modeling applications. During molecular dynamics the native bias maintains a native-like configuration but all interactions are malleable. Under molecular dynamics, a system populates the lowest free energy basins, and coupled with simulated annealing can even search for the lowest potential energy minima [63]. Through the introduction of external biasing potentials, AA SBMs built from high-resolution structures can reveal candidate AA structures from low resolution experimental data.

In a recent study of the ribosomal elongation cycle, Ratje et al. [50] used multiparticle cryoelectron microscopy analysis to capture subpopulations of EF-G-ribosome complexes at subnanometer resolution. While this resolution is not fine enough to achieve atomic details, the known crystallographic structure can be used to obtain atomic models of the microscopy data with a procedure termed MDFIT [65]. MDFIT biases the AA SBM with an energetic term developed in Orzechowski and Tama [48], which uses the correlation between the simulated and experimental electron density. The overall potential function therefore becomes

$$V_{\text{model}} = V_{\text{AA}} + V_{\text{map}} = V_{\text{AA}} + W \sum_{ijk} \rho_{ijk}^{\text{sim}} \rho_{ijk}^{\text{exp}}, \tag{13}$$

where W is an overall weight and ρ_{ijk}^{sim} and ρ_{ijk}^{exp} are the normalized electron densities at voxel (i, j, k) and V_{AA} is the AA SBM potential. A molecular dynamics simulation initialized at the crystallographic structure will distort to maximize the overlap between the simulated structure and the experimental electron density. The structure-based potential naturally maintains tertiary contacts present in the crystal structure without the need for ad hoc restraints.

The electron density map works well as a global bias, but local biases can also be introduced. Candidate structures for protein–protein complexes can be derived by introducing interprotein contacts from bioinformatic analysis and minimizing the resulting structure-based potential with molecular dynamics. Schug et al. [51] were able to predict the structure of the Spo0B/Spo0F two-component signal transduction (TCS) complex within 2.5Å of an existing crystal structure. TCS is ruled by transient interactions, posing harsh challenges to obtain atomic resolution structures. These transient interactions though have bioinformatic signatures, which provide the external biasing potential needed for modeling. Short-range contact potentials were introduced between correlated residues and the resulting potential

$$V_{\text{model}} = V_{\text{AA}} + k(r_{\text{CM}})^2 + \sum_{\{i,j\}} C_{\text{AA}}(r_{ij}, \bar{r}), \tag{14}$$

where r_{CM} is the distance between the proteins' centers of mass, $\{i, j\}$ denotes the correlated residues, C_{AA} is Eqn. 6, r_{ij} the distance between those residues' C_α atoms and $\overline{r} = 7$ Å. A weak center of mass constraint, as with multimeric folding (see Sect. 4.1.2), is a common method of encouraging two molecules to dock. The resulting structure from the AA SBM simulations can be directly used as input to an AA empirical force field for additional minimization.

5 Concluding Remarks

The principle of minimal frustration and the funneled landscape provide the theoretical framework for SBMs. We have presented numerous applications of SBMs, including protein folding and oligomerization, structure–function relationships in protein conformational transitions and structural modeling of protein–protein and ribonucleoprotein complexes. These models are publicly available at SMOG http://smog.ucsd.edu. Recent technical improvements in computer hardware for molecular dynamics simulations should allow for a new level of collaboration between simplified protein models and explicit solvent models. Protein folding simulations on the millisecond timescale will enable quantitative characterization of the roughness of the folding energy landscape [37, 54]. As experimentalists continue pushing boundaries in the characterization of molecular machines at the single molecule level, further theoretical investigation is needed to assess how the interplay of global properties with specific energetic details shapes the dynamics of these large macromolecular complexes [66]. We expect the importance of large-scale structural fluctuations, largely controlled by geometry, to be a central theme in the discussion of molecular machines in the years to come.

Acknowledgments JKN would like to thank Joanna Sułkowska for many helpful discussions and Paul Whitford and Ryan Hayes for a careful reading of the chapter. This work was supported by the Center for Theoretical Biological Physics sponsored by the national science foundation (NSF) (Grant PHY-0822283) and NSF Grant NSF-MCB-1051438.

References

1. Adcock, S.A., McCammon, J.A.: Molecular dynamics: survey of methods for simulating the activity of proteins. Chem. Rev. **106**(5), 1589–1615 (2006)
2. Andrews, B.T., Gosavi, S., Finke, J.M., Onuchic, J.N., Jennings, P.A.: The dual-basin landscape in gfp folding. Proc. Nat. Acad. Sci. USA **105**(34), 12283–12288 (2008)
3. de Araujo, A.F.P., Onuchic, J.N.: A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins. Proc. Nat. Acad. Sci. USA **106**(45), 19001–19004 (2009)
4. Azia, A., Levy, Y.: Nonnative electrostatic interactions can modulate protein folding: molecular dynamics with a grain of salt. J. Mol. Biol. **393**(2), 527–542 (2009)

5. Baker, D.: A surprising simplicity to protein folding. Nature **405**(6782), 39–42 (2000)
6. Baxter, E.L., Jennings, P.A., Onuchic, J.N.: Interdomain communication revealed in the diabetes drug target mitoneet. Proc. Nat. Acad. Sci. USA **108**(13), 5266–5271 (2011)
7. Bowers, K.J., Chow, E., Xu, H., Dror, R.O., Eastwood, M.P., Gregersen, B.A., Klepeis, J.L., Kolossvary, I., Moraes, M.A., Sacerdoti, F.D., Salmon, J.K., Shan, Y., Shaw, D.E.: Scalable algorithms for molecular dynamics simulations on commodity clusters. In: Proceedings of ACM/IEEE, p. 43 (2006)
8. Bryngelson, J., Wolynes, P.: Spin glasses and the statistical mechanics of protein folding. Proc. Nat. Acad. Sci. USA **84**, 7524 (1987)
9. Bryngelson, J., Wolynes, P.: Intermediates and barrier crossing in a random energy model (with applications to protein folding). J. Phys. Chem. **93**, 6902–6915 (1989)
10. Bryngelson, J.D., Onuchic, J.N., Socci, N.D., Wolynes, P.G.: Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins: Struct. Funct. Bioinf. **21**(3), 167–195 (1995)
11. Chavez, L.L., Onuchic, J.N., Clementi, C.: Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. J. Am. Chem. Soc. **126**(27), 8426–8432 (2004)
12. Cheung, M.S., García, A.E., Onuchic, J.N.: Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. Proc. Nat. Acad. Sci. USA **99**(2), 685–690 (2002)
13. Cho, S., Levy, Y., Wolynes, P.G.: P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. Proc. Nat. Acad. Sci. USA **103**(3), 586–591 (2006)
14. Cho, S.S., Weinkam, P., Wolynes, P.G.: Origins of barriers and barrierless folding in bbl. Proc. Nat. Acad. Sci. USA **105**(1), 118–123 (2008)
15. Clementi, C., Nymeyer, H., Onuchic, J.N.: Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J. Mol. Biol. **298**(5), 937–953 (2000)
16. Clementi, C., García, A.E., Onuchic, J.N.: Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein l. J. Mol. Biol. **326**(3), 933–954 (2003)
17. Clementi, C., Plotkin, S.S.: The effects of nonnative interactions on protein folding rates: theory and simulation. Protein Sci. **13**(7), 1750–1766 (2004)
18. Ferguson, N., Schartau, P.J., Sharpe, T.D., Sato, S., Fersht, A.R.: One-state downhill versus conventional protein folding. J. Mol. Biol. **344**(2), 295–301 (2004)
19. Fersht, A.R.: Characterizing transition-states in protein-folding - an essential step in the puzzle. Curr. Opin. Struct. Biol. **5**(1), 79–84 (1995)
20. Frauenfelder, H., Sligar, S.G., Wolynes, P.G.: The energy landscapes and motions of proteins. Science **254**(5038), 1598–1603 (1991)
21. Gambin, Y., Schug, A., Lemke, E.A., Lavinder, J.J., Ferreon, A.C.M., Magliery, T.J., Onuchic, J.N., Deniz, A.A.: Direct single-molecule observation of a protein living in two opposed native structures. Proc. Nat. Acad. Sci. USA **106**(25), 10153–10158 (2009)
22. Gosavi, S., Chavez, L.L., Jennings, P.A., Onuchic, J.N.: Topological frustration and the folding of interleukin-1 beta. J. Mol. Biol. **357**(3), 986–996 (2006)
23. Hess, B., Kutzner, C., van der Spoel, D., Lindahl, E.: Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theo. Comput. **4**(3), 435–447 (2008)
24. Hyeon, C., Jennings, P.A., Adams, J.A., Onuchic, J.N.: Ligand-induced global transitions in the catalytic domain of protein kinase A. Proc. Nat. Acad. Sci. USA **106**(9), 3023–3028 (2009)
25. Hyeon, C., Onuchic, J.N.: Mechanical control of the directional stepping dynamics of the kinesin motor. Proc. Nat. Acad. Sci. USA **104**(44), 17382–17387 (2007)
26. Okazaki, K., Koga, N., Takada, S., Onuchic, J.N., Wolynes, P.G.: Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: struc-based molecular dynamics simulations. Proc. Nat. Acad. Sci. USA **103**(32), 11844–11849 (2006)

27. Jamros, M.A., Oliveira, L.C., Whitford, P.C., Onuchic, J.N., Adams, J.A., Blumenthal, D.K., Jennings, P.A.: Proteins at work: a combined small angle x-ray scattering and theoretical determination of the multiple structures involved on the protein kinase functional landscape. J. Biol. Chem. **285**(46), 36121–36128 (2010)
28. Kaya, H., Chan, H.S.: Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: how adequate is native-centric topological modeling? J. Mol. Biol. **326**(3), 911–931 (2003)
29. Koga, N., Takada, S.: Roles of native topology and chain-length scaling in protein folding: a simulation study with a go-like model. J. Mol. Biol. **313**(1), 171–180 (2001)
30. Kouza, M., Li, M.S., O'brien, E.P., Hu, C.-K., Thirumalai, D.: Effect of finite size on cooperativity and rates of protein folding. J. Phys. Chem. A **110**(2), 671–676 (2006)
31. Kumar, S., Rosenberg, J., Bouzida, D., Swendsen, R.H.: The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J. Comput. Chem. **13**(8), 1011 (1992)
32. Lammert, H., Schug, A., Onuchic, J.N.: Robustness and generalization of structure-based models for protein folding and function. Proteins: Struct. Funct. Bioinf. **77**(4), 881–891 (2009)
33. Leopold, P.E., Montal, M., Onuchic, J.N.: Protein folding funnels - a kinetic approach to the sequence structure relationship. Proc. Nat. Acad. Sci. USA **89**(18), 8721–8725 (1992)
34. Levy, Y., Cho, S.S., Shen, T., Onuchic, J.N., Wolynes, P.G.: Symmetry and frustration in protein energy landscapes: a near degeneracy resolves the rop dimer-folding mystery. Proc. Nat. Acad. Sci. USA **102**(7), 2373–2378 (2005)
35. Levy, Y., Onuchic, J.N., Wolynes, P.G.: Fly-casting in protein-dna binding: frustration between protein folding and electrostatics facilitates target recognition. J. Am. Chem. Soc. **129**(4), 738–739 (2007)
36. Levy, Y., Wolynes, P.G., Onuchic, J.N.: Protein topology determines binding mechanism. Proc. Nat. Acad. Sci. USA **101**(2), 511–516 (2004)
37. Lindorff-Larsen, K., Piana, S., Dror, R.O., Shaw, D.E.: How fast-folding proteins fold. Science 334, 517–520 (2011)
38. McCammon, J.A., Gelin, B.R., Karplus, M.: Dynamics of folded proteins. Nature **267**(5612), 585–590 (1977)
39. Mittal, A., Jayaram, B.: Backbones of folded proteins reveal novel invariant amino acid neighborhoods. J. Biomol. Struct. Dyn. **28**(4), 443–454 (2011)
40. Miyashita, O., Onuchic, J.N., Wolynes, P.G.: Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. Proc. Nat. Acad. Sci. USA **100**(22), 12570–12575 (2003)
41. Mor, A., Ziv, G., Levy, Y.: Simulations of proteins with inhomogeneous degrees of freedom: the effect of thermostats. J. Comput. Chem. **29**(12), 1992–1998 (2008)
42. Nechushtai, R., Lammert, H., Michaeli, D., Eisenberg-Domovich, Y., Zuris, J.A., Luca, M.A., Capraro, D.T., Fish, A., Shimshon, O., Roy, M., Schug, A., Whitford, P.C., Livnah, O., Onuchic, J.N., Jennings, P.A.: Allostery in the ferredoxin protein motif does not involve a conformational switch. Proc. Nat. Acad. Sci. USA **108**(6), 2240–2245 (2011)
43. Noel, J.K., Sulkowska, J.I., Onuchic, J.N.: Slipknotting upon native-like loop formation in a trefoil knot protein. Proc. Nat. Acad. Sci. USA **107**(35), 15403–15408 (2010)
44. Noel, J.K., Whitford, P.C., Sanbonmatsu, K.Y., Onuchic, J.N.: Smog@ctbp: simplified deployment of structure-based models in gromacs. Nucleic Acids Res. 38, W657 (2010)
45. Nymeyer, H., García, A.E., Onuchic, J.N.: Folding funnels and frustration in off-lattice minimalist protein landscapes. Proc. Nat. Acad. Sci. USA **95**(11), 5921–5928 (1998)
46. Oliveira, R.J., Whitford, P.C., Chahine, J., Wang, J., Onuchic, J.N., Leite, V.B.P.: The origin of nonmonotonic complex behavior and the effects of nonnative interactions on the diffusive properties of protein folding. Biophys. J. **99**(2), 600–608 (2010)
47. Onuchic, J.N., Wolynes, P.G.: Theory of protein folding. Curr. Opin. Struct. Biol. **14**(1), 70–75 (2004)

48. Orzechowski, M., Tama, F.: Flexible fitting of high-resolution x-ray structures into cryo-electron microscopy maps using biased molecular dynamics simulations. Biophys. J. **95**(12), 5692–5705 (2008)
49. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L., Schulten, K.: Scalable molecular dynamics with namd. J. Comput. Chem. **26**(16), 1781–1802 (2005)
50. Ratje, A.H., Loerke, J., Mikolajka, A., Brünner, M., Hildebrand, P.W., Starosta, A.L., Dönhöfer, A., Connell, S.R., Fucini, P., Mielke, T., Whitford, P.C., Onuchic, J.N., Yu, Y., Sanbonmatsu, K.Y., Hartmann, R.K., Penczek, P.A., Wilson, D.N., Spahn, C.M.T.: Head swivel on the ribosome facilitates translocation by means of intra-subunit trna hybrid sites. Nature **468**(7324), 713–716 (2010)
51. Schug, A., Weigt, M., Onuchic, J.N., Hwa, T., Szurmant, H.: High-resolution protein complexes from integrating genomic information with molecular simulation. Proc. Nat. Acad. Sci. USA **106**(52), 22124–22129 (2009)
52. Schug, A., Whitford, P.C., Levy, Y., Onuchic, J.N.: Mutations as trapdoors to two competing native conformations of the rop-dimer. Proc. Nat. Acad. Sci. USA **104**(45), 17674–17679 (2007)
53. Scott, K.A., Batey, S., Hooton, K.A., Clarke, J.: The folding of spectrin domains i: wild-type domains have the same stability but very different kinetic properties. J. Mol. Biol. **344**(1), 195–205 (2004)
54. Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R.O., Eastwood, M.P., Bank, J.A., Jumper, J.M., Salmon, J.K., Shan, Y., Wriggers, W.: Atomic-level characterization of the structural dynamics of proteins. Science **330**(6002), 341–346 (2010)
55. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., Edelman, M.: Automated analysis of interatomic contacts in proteins. Bioinformatics **15**(4), 327–332 (1999)
56. Socci, N.D., Onuchic, J.N., Wolynes, P.G.: Diffusive dynamics of the reaction coordinate for protein folding funnels. J. Chem. Phys. **104**(15), 5860–5868 (1996)
57. Sułkowska, J., Sułkowski, P., Szymczak, P., Cieplak, M.: Tightening of knots in proteins. Phys. Rev. Lett. **100**(5), 058106 (2008)
58. Sułkowska, J.I., Cieplak, M.: Selection of optimal variants of gō-like models of proteins through studies of stretching. Biophys. J. **95**(7), 3174–3191 (2008)
59. Sułkowska, J.I., Sułkowski, P., Onuchic, J.: Dodging the crisis of folding proteins with knots. Proc. Nat. Acad. Sci. USA **106**(9), 3119–3124 (2009)
60. Sutto, L., Lätzer, J., Hegler, J.A., Ferreiro, D.U., Wolynes, P.G.: Consequences of localized frustration for the folding mechanism of the im7 protein. Proc. Nat. Acad. Sci. USA **104**(50), 19825–19830 (2007)
61. Tirion, M.: Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett. **77**(9), 1905–1908 (1996)
62. Veitshans, T., Klimov, D., Thirumalai, D.: Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. Folding and Design **2**(1), 1–22 (1997)
63. Wales, D.J.: Energy Landscapes. Cambridge University Press, Cambridge (2003)
64. Whitford, P., Schug, A., Saunders, J., Hennelly, S., Onuchic, J., Sanbonmatsu, K.: Supplementary-nonlocal helix formation is key to understanding s-adenosylmethionine-1 riboswitch function. Biophys. J. **96**(2), L7–L9 (2009)
65. Whitford, P.C., Ahmed, A., Yu, Y., Hennelly, S.P., Tama, F., Spahn, C.M.T., Onuchic, J., Sanbonmatsu, K.Y.: Excited states of ribosome translocation revealed through integrative molecular modeling. Proc. Nat. Acad. Sci. USA **108**(47), 18943–18948 (2011)
66. Whitford, P.C., Geggier, P., Altman, R.B., Blanchard, S.C., Onuchic, J.N., Sanbonmatsu, K.Y.: Accommodation of aminoacyl-trna into the ribosome involves reversible excursions along multiple pathways. RNA **16**(6), 1196–1204 (2010)
67. Whitford, P.C., Gosavi, S., Onuchic, J.N.: Conformational transitions in adenylate kinase. Allosteric communication reduces misligation. J. Biol. Chem. **283**(4), 2042–2048 (2008)

68. Whitford, P.C., Miyashita, O., Levy, Y., Onuchic, J.N.: Conformational transitions of adenylate kinase: switching by cracking. J. Mol. Biol. **366**(5), 1661–1671 (2007)
69. Whitford, P.C., Noel, J.K., Gosavi, S., Schug, A., Sanbonmatsu, K.Y., Onuchic, J.N.: An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. Proteins: Struct. Funct. Bioinf. **75**(2), 430–441 (2009)
70. Wodak, S.J., Janin, J.: Structural basis of macromolecular recognition. Adv. Protein Chem. **61**, 9–73 (2002)
71. Wu, L., Zhang, J., Qin, M., Liu, F., Wang, W.: Folding of proteins with an all-atom go-model. J. Chem. Phys. **128**(23), 235103 (2008)

Discrete Molecular Dynamics Simulation of Biomolecules

Feng Ding and Nikolay V. Dokholyan

1 Introduction

Biological molecules are highly dynamic and coexist in multiple conformations in solution [1]. Molecular motions are observed on a broad range of time and length scales using spectroscopy and hydrogen–deuterium exchange experiments [2–5]. The internal motions and resulting conformational changes of these molecules play an essential role in their function. Sampling the structural and dynamic properties of biomolecules remains a challenge due to the large range of time and length scales associated with molecular life. Molecular modeling, especially molecular dynamics simulations of biomolecules and molecular complexes, has played a crucial role in bridging time and length scale gaps and has been pivotal to our understanding of the dynamic aspect of biomolecules [6].

Molecular dynamics (MD) is a computational simulation algorithm, where atoms move according to the laws of classical mechanics. Energetic interactions between atoms are modeled with empirical functions (a "force field") of varying complexities, usually composed of bonded terms representing chain connectivity (bonds, angles, and dihedrals) and nonbonded terms representing van der Waals (VDW) and electrostatic interactions. The dynamic trajectory of the molecular system can be obtained by integrating the equations of motions over a small time step ($\sim$1–2 fs). Analysis of the trajectories from MD simulations can provide great detail concerning the motions of individual particles as a function of time. Thus, these trajectories can be used to address specific questions about properties of a model system that are often inaccessible to experiments. For many aspects of biomolecular

F. Ding (✉) • N.V. Dokholyan
Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina, Chapel Hill, NC, USA
e-mail: fding@unc.edu; dokh@unc.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_3, © Springer Science+Business Media, LLC 2012

function, it is exactly these details that are of the highest interest and utility. MD simulations allow for the generation of experimentally testable hypotheses, and experiments play an essential role in validating simulation methodology.

The first MD simulation of a fluid system was reported by Alder and Wainwright in 1957 [7]. In a hard sphere fluid system, the authors found evidence of a solid–fluid phase transition that had not been observed in previous Monte Carlo simulations. The subject of hard sphere simulations falls in the general category of discrete potential MD (DMD), which is also called event-driven molecular dynamics, discontinuous molecular dynamics, or discrete molecular dynamics. The DMD methodology is continuously under development for hard-sphere and polymer systems [8–15], and has recently seen an increase in applications for studying biomolecules [16–22]. The development of continuous potentials for MD simulations has facilitated the inclusion of detailed aspects of atomic interactions [23, 24], which is the most common form of MD in current practice. Since the publication of the first MD simulation of bovine pancreatic trypsin inhibitor (BPTI) in 1977 [25], the application of MD simulations to study the structure, dynamics, and function of biomolecules has been increasing steadily. However, the time scales currently accessible in MD simulations are typically 10–100 ns, which restrict their application to many biological processes with large time and length scales (e.g., protein folding occurs in milliseconds to seconds). Even utilizing worldwide computing resources [26] or specialized high-performance computers dedicated to MD simulations (such as Anton [27, 28]), the time scale reached by MD is still in the range of microseconds. Conversely, with the recent development of DMD for biological systems, including the DMD force field [21], all-atom protein models [29–31], and hydrogen bond modeling [18], DMD simulations of realistic biomolecular systems can reach microsecond time scales on personal computers. All-atom DMD simulations have been applied to study protein folding [21, 30], protein design [32, 33], protein structure optimization [34], and post-translational modification of proteins [35]. In this chapter, we focus on DMD simulations of biomolecules. We briefly discuss the DMD algorithm and recent optimization approaches, important developments of DMD methodology for biomolecules, and several applications of all-atom DMD for biomolecules.

2 Discrete Molecular Dynamics

2.1 Algorithm

DMD simulations are based on pairwise interaction potentials that are discontinuous functions of the interatomic distance, r (Fig. 1). We assign for each atom a specific type—A, B, C, . . .—that determines its interaction with other atoms. The interaction potential between two atoms i (type A) and j (type B) is characterized by distances $r^{\mathrm{AB}}{}_{\mathrm{min}} < r^{\mathrm{AB}}{}_{1} < r^{\mathrm{AB}}{}_{2\ldots} < r^{\mathrm{AB}}{}_{k\ldots} < r^{\mathrm{AB}}{}_{\mathrm{max}}$, where $r^{\mathrm{AB}}{}_{\mathrm{min}}$ corresponds to the

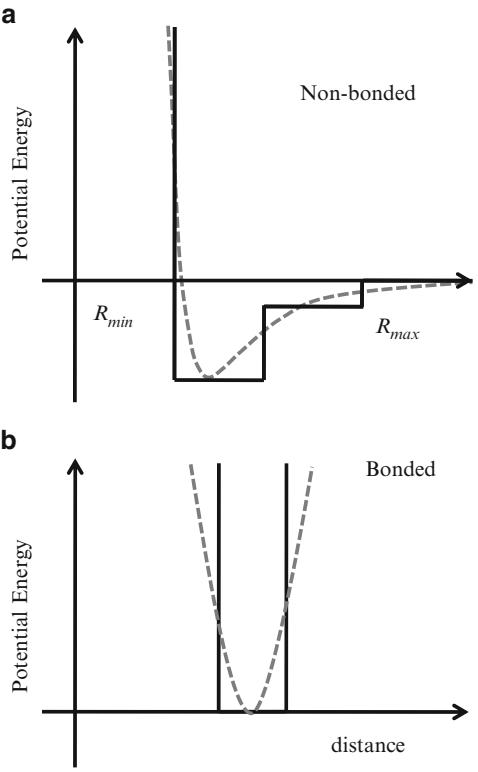

Fig. 1 A schematic diagram of DMD potentials. (**a**) Interaction between nonbonded atom pairs. $R_{\min}$ corresponds to the hard-core distance, $R_{\max}$ corresponds to the interaction range. (**b**) Interaction between bonded atoms. In both cases, *gray dashed lines* correspond to the continuous potential in traditional MD

hardcore collision distance and ${r^{\mathrm{AB}}}_{\max}$ corresponds to the maximal interaction range between the two atoms. If ${r^{\mathrm{AB}}}_k < r_{ij} < {r^{\mathrm{AB}}}_{k+1}$, the pairwise potential energy is assigned as $U_{ij} = {U^{\mathrm{AB}}}_k$. If $r_{ij} < {r^{\mathrm{AB}}}_{\min}$, $U_{ij} = \infty$ so that the two atoms do not come closer than the hard core distance; and if $r_{ij} > {r^{\mathrm{AB}}}_{\max}$, $U_{ij} = 0$ such that two atoms will not interact with each other. If atoms i and j are linked by a bond, the potential energy $U_{ij} = \infty$ when $r_{ij} > {r^{\mathrm{AB}}}_{\max}$. As the result, the two atoms will not escape from each other beyond ${r^{\mathrm{AB}}}_{\max}$, mimicking the bond (Fig. 1b). In DMD simulations, each atom moves with a constant velocity until its distance to another neighboring atom becomes equal to a potential step ${r^{\mathrm{AB}}}_k$, where the potential energy is not continuous. At this moment in time their velocities change instantaneously in accordance with the laws of energy, momentum, and angular momentum conservation. When the kinetic energy of the particles is not sufficient to overcome the potential barrier ${\epsilon^{\mathrm{AB}}}_k = {U^{\mathrm{AB}}}_{k-1} - {U^{\mathrm{AB}}}_k$ (only when the potential change is positive), the atoms undergo a hardcore reflection with no change in potential energy. Each of these events is termed as a collision. At each collision, positions and velocities are updated only for the two colliding atoms, and potential collisions with their neighboring atoms are recomputed. By iterating these calculations, the trajectory of the system is computed as a set of consecutive collision events.

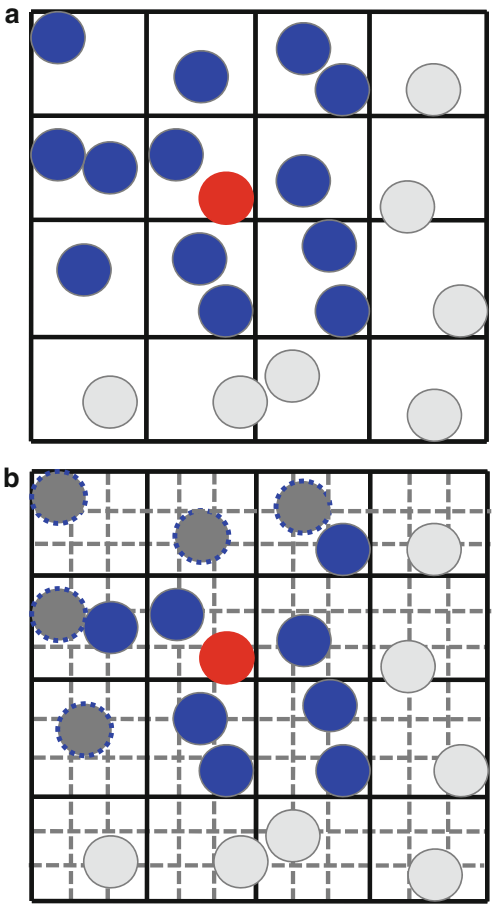

Fig. 2 Grid approach to facilitate the search of neighboring atoms. (**a**) The traditional approach to divide the simulation box into smallest cells, with cell dimension larger than the maximum interaction range. Only the atoms in the neighboring 27 cells (*in blue*) are counted as the neighboring atoms of the atom in *red*. (**b**) The new approach to further divide each cell into a finer grid. By dividing each dimension of the cell by three, the number of neighboring atoms can be greatly reduced (*dark gray spheres*)

In order to efficiently simulate collisions, Rapaport [8] proposed to divide the simulation box into subcells, with the dimension of the cell assigned as the largest interaction range of all the atom pairs and wall-crossing events treated as collisions. As the result, for each atom i, only the collisions between atom i and the atoms in the neighboring $3^3 = 27$ cells are required to be computed for predicting the next collisions of atom i (Fig. 2a). Assuming the average number of atoms in each cell is N_g, the average number of possible collisions to be evaluated for each atom is $27N_g$. To facilitate the evaluation of all possible collisions and prediction of the next collisions, Rapaport [9] proposed a priority tree containing all possible collisions between neighboring atoms ($\sim 27N_gN$), where N is the total number of atoms. The priority tree is sorted according to the collision time with computational complexity $O(\ln(27N_gN))$. As an alternative to this multievent scheduling, Allen and Tildesley [36] proposed a single-event scheduling approach, where only the soonest collision for each atom is stored in a fixed-length binary tree ($\sim N$) with

sorting time $O(\ln(N))$. Smith et al. [15] compared these two scheduling methods and found that in simulations of a polymeric system, the single-event scheduling approach is more efficient than multievent scheduling due to avoiding the insertion and deletion of superfluous potential collisions in the priority tree. Next, we discuss several additional optimization approaches.

2.2 Fine Grid

In DMD, the majority of calculation is the re-evaluation of collision times between a colliding atom and its neighbors. When the dimension of the cell ($l_c \sim \mathrm{IR}_{\max}$, the maximum interaction range) is large compared to the hardcore diameter, as in soft sphere systems (Fig. 1a), the number of atoms in each cell is often more than one. As discussed above, the number of atoms in the neighboring 27 cells is approximately $27\, l_c^3 \rho$, where ρ is the number density. However, assuming l_c approximately equal to the interaction range, the number of atoms inside the interaction range is $\sim(4\pi/3) l_c^3 \rho$ which is much less than $27\, {l_c}^3 \rho$. Therefore, many unnecessary atom pairs are included in the current scheme. We propose to divide each cell into a finer grid with each dimension divided evenly by a number, N_f (e.g., $N_f = 3$ in Fig. 2b). For each cell, we assign an integer address $(C_x,\ C_y,\ C_z)$. If the two cells have the address difference $(\Delta C^x,\ \Delta C^y,\ \Delta C^z)$ and

$$\sum_{d=x,y,z} (\max\{\Delta C^d - 1, 0\} \times l_c^d / N_f)^2 < l_c^2, \tag{1}$$

we consider the two cells as neighbors, and hence the atoms inside the cells are neighbors. Here, ${l_c}^d$ are the cell dimensions. As N_f increases, the number of atoms inside the neighboring cells asymptotically approaches $(4\pi/3){l_c}^3 \rho$, approximately 16% of the original number of neighboring atoms. As the result, the computational efficiency under the new scheme can be increased by as much as 6.4 times. On the other hand, the frequency of cell crossing and the corresponding CPU time spent are correspondingly increasing with this increase in N_f. Therefore, it is possible to find an optimal number of N_f for each type of DMD simulation system. In our all-atom protein model for DMD simulations, we use $N_f = 6$. *We find that in dense-packing cases such as folded proteins, we can improve the simulation efficiency by three to four times by using a finer grid.*

2.3 Reduce the Unnecessary Square Root Calculation

The most expensive calculation in the DMD algorithm is performed after each collision, when the DMD algorithm re-evaluates the collision times between the colliding atoms and their neighboring atoms. Because of the costly square root

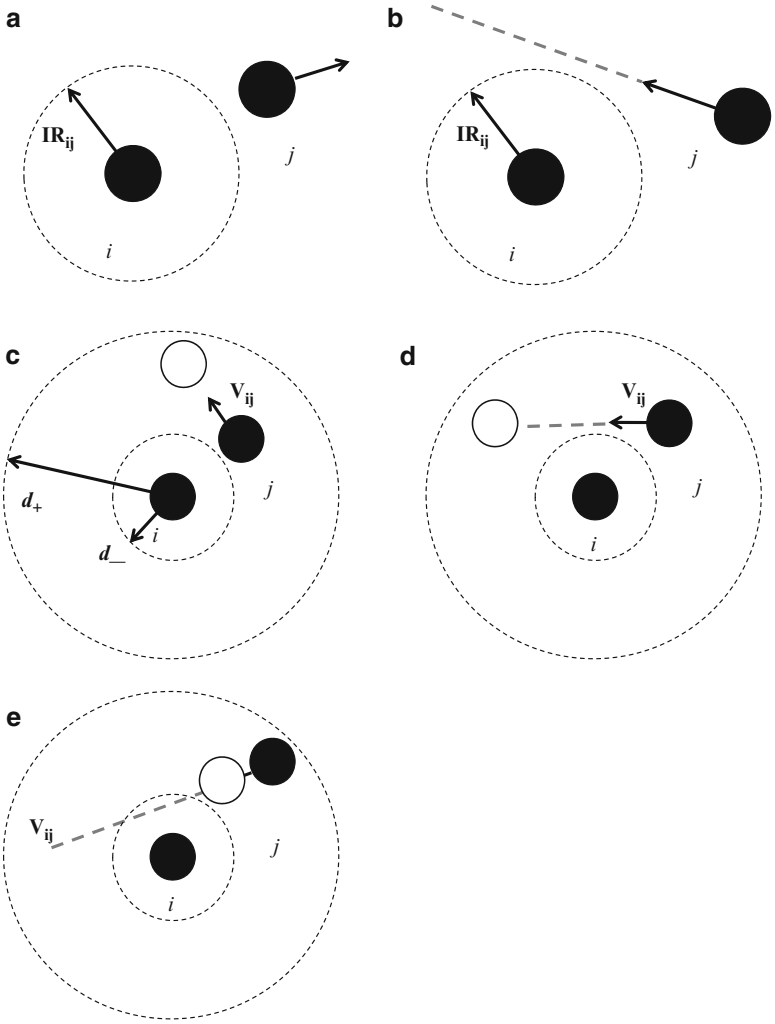

Fig. 3 Cases where the square root calculation to predict the next collisions is not necessary. If (**a**) two noninteracting (beyond the interaction range IR_{ij}) atoms are moving away from each other, and (**b**) two approaching, noninteracting atoms with the minimal distance larger than IR_{ij} (see the *dashed line* in **b**), the square root calculation is *not* necessary, since the operand is negative (collision will not happen). However, even if a collision can happen as in C, D, and E, the collision will not take place if some other event with respect to either i or j happens first. The *open sphere* along the direction of the relative velocity V_{ij} indicates the new position of atom j with respect to atom i

calculation involved in these calculations, it is important to devise a method to reduce the number of unnecessary collision time evaluations. For example, usually under two conditions (Fig. 3a, b), the predicted collision will not happen: (1) when the two atoms are moving away from each other ($\mathbf{R}_{ij} \cdot \mathbf{V}_{ij} > 0$) and the pairwise

distance is larger than the interaction range, IR_{ij} (Fig. 3a) and (2) when the two atoms are approaching the interaction range but the minimum distance is still larger than IR_{ij} $\left(\mathbf{R}_{ij}^2 - (\mathbf{R}_{ij} \cdot \mathbf{V}_{ij})^2/\mathbf{V}_{ij}^2 > IR_{ij}^2\right)$ (Fig. 3b). Here, $\mathbf{V}_{ij}$ is the relative velocity and $\mathbf{R}_{ij}$ is the relative displacement.

We developed a new approach to reduce further unnecessary square root calculations. During the recalculation of potential collisions (see Sect. 2.1), we assume a cutoff time Δt for each atom D. Within such a cutoff time, a collision will always happen to the atom of interest. Therefore, we may simply evaluate the pairwise displacement $\mathbf{R}_{ij}+\mathbf{V}_{ij}\Delta t$, with the pairwise distance R_{ij} within the potential steps $(d_-,\ d_+)$. In the following cases, collision will not occur:

1. Atoms moving away from each other $(\mathbf{R}_{ij} \cdot \mathbf{V}_{ij} > 0)$, but the two atoms do not collide within Δt at d_+, $(\mathbf{R}_{ij} + \mathbf{V}_{ij}\Delta t)^2 < {d_+}^2$ (Fig. 3c)
2. Atoms approaching each other $(\mathbf{R}_{ij} \cdot \mathbf{V}_{ij} < 0)$ with a minimum distance larger than d_- $\left(\mathbf{R}_{ij}^2 - (\mathbf{R}_{ij} \cdot \mathbf{V}_{ij})^2/\mathbf{V}_{ij}^2 > {d_-}^2\right)$, but the two atoms do not collide within Δt at d_+, $\mathbf{R}_{ij} + \mathbf{V}_{ij}\Delta t)^2 < {d_+}^2$ (Fig. 3d)
3. Atoms approaching each other $(\mathbf{R}_{ij} \cdot \mathbf{V}_{ij} < 0)$ with a minimum distance smaller than d_- $\left(\mathbf{R}_{ij}^2 - (\mathbf{R}_{ij} \cdot \mathbf{V}_{ij})^2/\mathbf{V}_{ij}^2 < {d_-}^2\right)$, but the two atoms do not collide within Δt, $[(\mathbf{R}_{ij} + \mathbf{V}_{ij}\Delta t) \cdot \mathbf{V}_{ij}]^2/\mathbf{V}_{ij}^2 > {d_-}^2 - \left[\mathbf{R}_{ij}^2 - (\mathbf{R}_{ij} \cdot \mathbf{V}_{ij})^2/\mathbf{V}_{ij}^2\right]$ (Fig. 3e)

and the collision time can be safely assumed to be infinity. The remaining question is how to define the cutoff time Δt. There are two types of events, the cell crossing and the random collision for the Anderson's thermostat [37], which can be used as the reference events since one of them will always happen if no pairwise collision takes place before these two events. We use the shorter time of these two events to define the cutoff time for each atom. Alternatively, one can dynamically define the cutoff time Δt for a given atom based on the atom's average collision time, $< t_{\text{col}} >$, which can be updated periodically. We set $\Delta t = 4 < t_{\text{col}} >$. *We find that such an optimization can improve the efficiency of simulation by 20–30%.*

2.4 Paul's O(1) Sorting Approach

In DMD, the next collision is obtained by sorting, using either the priority tree in the Rapaport approach (multievent scheduling [9]) or the binary tree in the Allen and Tildesley approach (single-event scheduling [36]). In both cases, the computational complexity is in the order of $O(\ln N)$. Recently, Paul [38] proposed a new sorting approach for DMD with a computational complexity of $O(1)$. In Paul's approach, a fixed length array (N_p) is used to hold the collision times, and the array is head–tail connected for repeated use (Fig. 4). The total time of the array is δt and the time step is $\delta t / N_p$. The pointer (index P_t) corresponds to the "current time" (t_C) in units of $\delta t / N_p$. Each collision at time t is added to the array with respect to the "current time": $[P_t + (t - t_C)/(\delta t / N_p)]\ \%\ N_p$. Each element in the array can hold more than one event since each element corresponds to a time window of $\delta t / N_p$. All the events

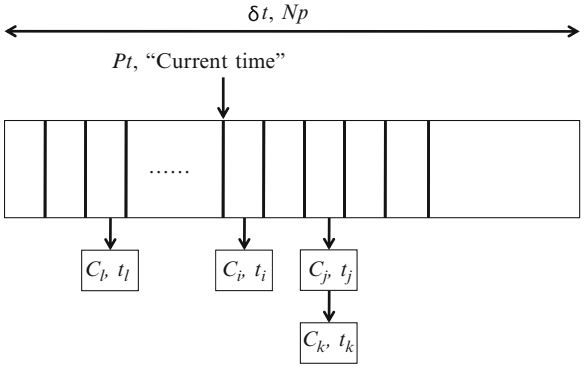

Fig. 4 Schematic for the O(1) sorting approach of collision events by Paul. The linear array of length N_g corresponds to the time interval δt. The array is head–tail connected for repeated usage. A pointer indicates the current time t_c in units of $\delta t/N_g$. Each collision time is inserted into the array with respect the current time: $[P_t + (t - t_c)/(\delta t/N_g)]\%\ N_g$. An element can hold more than one event connected by a simple linked list. The next collision is obtained by advancing the pointer until an occupied array is encountered, and choosing the event with the shortest collision time. By carefully select δt and N_g, the number of events in each element is small and the next collision can be found by a simple bubble sort

within this time window are linked by a simple "linked list." The next collision is obtained by moving the current time pointer forward to the first nonempty element, within which the soonest collision can be found by a simple "bubble sort" approach if the number of events within each element is small. One can define δt and N_p in such a way that the number of events within each element is small. *We find that when the system is large* ($\sim 10^5$ to 10^6 atoms), *sorting takes a significant amount of CPU time (~20% of total computation time). In this case, Paul's sorting approach greatly reduces the percentage of CPU time for sorting from ~20% to only 1–2%.*

Therefore, by carefully selecting the size of the fine grid, reducing the number of unnecessary square-root calculations, and adopting an $O(1)$ sorting algorithm, DMD simulation efficiency can be greatly improved over the traditional approach [9, 36], allowing for the simulation of biomolecular systems with realistic models and force fields. Next, we describe recent developments in the DMD force field and high-resolution molecular models

3 Development of DMD Force Field for Biomolecules

3.1 Hydrogen Bonds

The hydrogen bond interaction is the driving force for secondary structure formation in proteins and nucleotides. In contrast to the model used in continuous MD simulations, hydrogen bond interactions cannot be modeled as dipole–dipole interactions in DMD simulations. Liu and Elliot [39, 40] first proposed a hydrogen

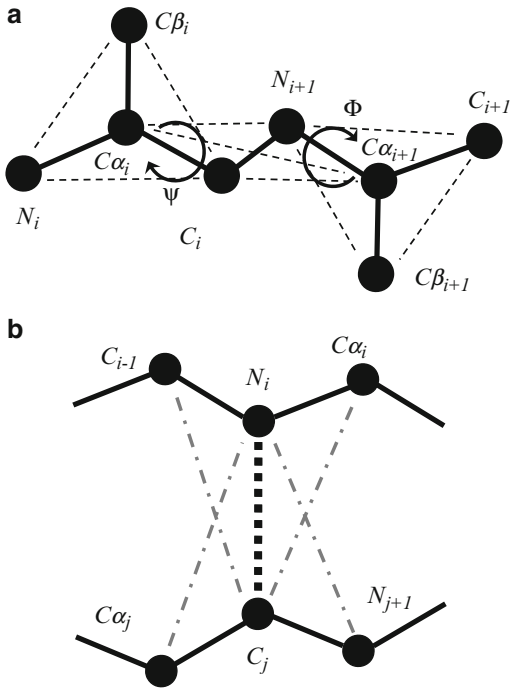

Fig. 5 Model of a hydrogen bond in a simple protein backbone model. (**a**) The four-bead model of a polypeptide. Backbone carbonyl oxygen and amide hydrogen are not explicitly modeled. (**b**) The schematic of a hydrogen bond between carbonyl carbon and amide nitrogen. The *gray dot-dashed lines* correspond to the auxiliary bonds

bond interaction model for DMD, where a hydrogen bond donor (proton) and acceptor (lone electron pair) are explicitly modeled as small attracting atoms positioned inside the hard spheres of the bonding atoms. As the result, the orientation dependence of the hydrogen bond is effectively modeled [39, 40]. However, the explicit modeling of hydrogens and lone electron pairs significantly reduces the computational efficiency of the simulations. Smith and Hall proposed [13] a different approach to model hydrogen bonds in a coarse-grained protein backbone model (alpha carbon C_a, backbone carbonyl carbon C, and nitrogen N; Fig. 5a). Although the backbone carbonyl oxygen O and amide H forming the hydrogen bond are not explicitly modeled, their coordinates can be computed based on the coordinates of existing backbone heavy atoms. A hydrogen bond is formed between N and C when they approach within a certain distance of the hypothetical O and H and are aligned collinearly based on angles of $N-H-O$ and $H-O-C$. When this linear alignment is changed, the hydrogen bond is allowed to dissociate, ignoring the impact of the dissociation energy on the dynamics and thus violating the energy conservation law. To overcome the energy conservation violation problem, we proposed an alternative approach to model the hydrogen bond [18]. The approach is based on a "*reaction*" algorithm in DMD: Two reactant atoms A and B can change their types to A' and B' upon collision at a given reaction interaction range. The total potential energy change ΔE associated with the atom type change is evaluated by summing over all interacting atoms. If the kinetic energy

is sufficient to overcome the potential energy change in the case of $\Delta E > 0$, the reaction takes place. Similarly, the reverse reaction can occur when the two atoms dissociate at the reaction interaction range. The reaction is intrinsically a multibody interaction model.

We explicitly model the hydrogen bond interaction using the reaction algorithm. For example, using the same coarse-grained backbone model as Smith and Hall (Fig. 5a), we assign auxiliary atoms for each hydrogen-bonding atom N and C that correspond to the nearest neighboring atoms along the backbone (Fig. 5b). If two atoms N_i and C_j form a hydrogen bond, we will explicitly assign a hydrogen bond between these two atoms and also assign auxiliary bonds between the auxiliary atoms of the donor and acceptor (gray lines in Fig. 5b). The two atoms then change their type to N_i' and C_j'. The auxiliary bonds will retain the alignment of the hydrogen bond during the simulation. The hydrogen bond and the corresponding auxiliary bonds will dissociate when the two hydrogen-bonded atoms move away from the reaction interaction range with a kinetic energy able to overcome the potential energy change. Upon dissociation, the atoms will revert to their original types. During both hydrogen bond formation and dissociation, the total energy change associated with type change and bond formation and breaking is evaluated. If the two approaching atoms cannot form a hydrogen bond, they will proceed with their regular predicted collision. The DMD potential function for hydrogen and auxiliary bonds can be derived from statistical analysis of the hydrogen bonds in high-resolution protein structures. Using this method, we were able to directly observe in silico a secondary structure transition between alpha helix and beta sheet, in which transition plays a crucial role in disease-associated protein misfolding and aggregation [18].

3.2 All-Atom Protein Model

In previous years, DMD has mainly been associated with coarse-grained modeling. Recently, we have developed an all-atom protein model for use in DMD simulations [21], where all heavy atoms and polar hydrogen atoms are explicitly represented, which is often referred to as the united-atom model. The all-atom model allows for the study of high-resolution conformational dynamics on the atomic level.

In the all-atom protein model (Fig. 6a), bonded interactions are modeled using distance constraints for the covalent bond length, bond angles, and dihedral angles (Fig. 6b). For covalently bonded atom pairs and also the bond angles, the interactions are modeled by a square-well potential (Fig. 1b). Dihedral interactions between atoms i and $i + 3$ are modeled by multistep potential functions [19] of pairwise distance. The set of distance parameters ($d_{\min}$, d_0, d_1, d_2, $d_{\max}$) for these potentials are experimentally determined from distance distributions in a nonredundant database of high-resolution protein structures (Fig. 6b).

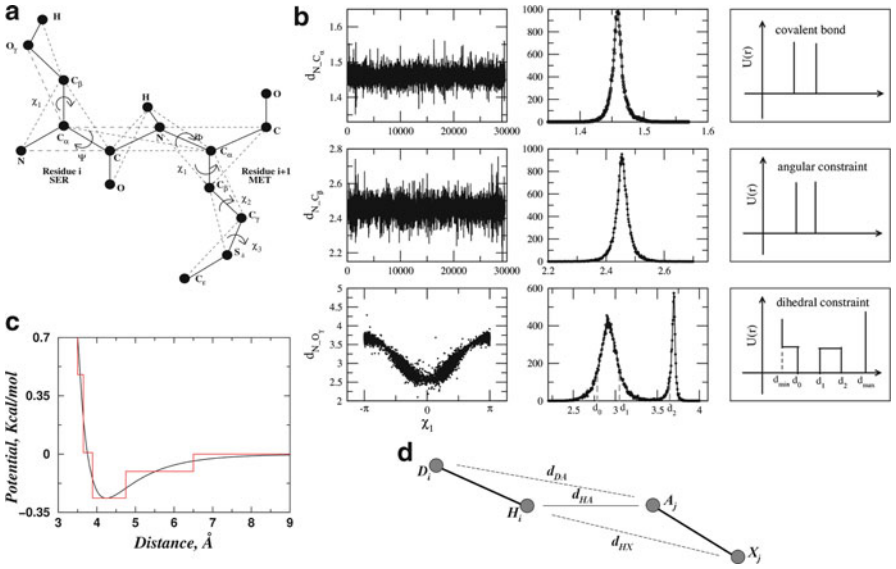

Fig. 6 All-atom protein model. (**a**) Schematic diagram for the all-atom protein model. Only two consecutive residues are shown. The *solid thick lines* represent the covalent and the peptide bonds. The *thin dashed lines* denote the effective bonds that are needed either to fix the bond angles, model the side chain dihedral angles, or to maintain the planarity of the peptide bonds. (**b**) Parameterization of the bonded interactions for representative atom pairs. The first column shows the distribution of the distances in serine between $N-C_\alpha$, $N-C_\beta$, and $N-O_\gamma$, respectively. The second column shows the corresponding histogram for the distribution of each atom pair. The third column shows the resulting constraint potentials schematically. For bonds (e.g., $N-C_\alpha$) and bond angles (e.g., $N-C_\beta$), the left and right boundaries of the constraint potential correspond to $d-\sigma$ and $d+\sigma$, respectively. Here, d is the average length and σ is the standard deviation of the distance distribution. (**c**) Parameterization of nonbonded interactions in all-atom DMD. The continuous *red line* corresponds to the van der Waals and solvation interaction between two carbon atoms. The *black step* function is the discretized potential for DMD. (**d**) A schematic for the hydrogen bonding interaction between hydrogen H_i and acceptor A_j. Atom D_i is the donor and X_j is the heavy atom directly bonded to A_j. Besides the distance between the hydrogen and the acceptor d_{HA}, we also assess the auxiliary distances d_{DA} (distance between atoms D_i and A_j) and d_{HX} (distance between atoms H_i and X_j)

In order to accurately represent nonbonded interactions, we discretized the continuous Medusa force field [34], in which the VDW and solvation interactions are included. VDW interactions use the standard Lennard-Jones potential, and solvation interactions are modeled by the Lazaridis–Karplus (LK) solvation model [41], which is expressed as the sum of pairwise distance-dependent effective solvation energies (EEF1). The discrete potential functions mimic the continuous potential $E_{ij}(d) = E_{ij}^{\mathrm{VDW}}(d) + E_{ij}^{\mathrm{LK}}(d)$ by capturing the attractions and repulsions while using a minimal number of steps (Fig. 6c). By trial and error in test simulations, we adopted the following discretization protocol: (1) we choose an interaction range of

6.5 Å, where the interaction potential attenuates in all atom pairs; (2) we assign a potential step between the distances corresponding to the energy minimum (force is zero) and the interaction action range (force approaching zero), where the force is maximum; (3) we choose the hard sphere distance with VDW–EEF1 energy equal to the minimum energy plus $2k_BT \sim 1.2$ kcal/mol, since thermodynamically the probability to find two atoms within this distance is very low. We choose the next repulsion step with VDW–EEF1 energy equals to the minimum energy plus $k_BT \sim 0.6$ kcal/mol, and the third repulsive step before the energy minimum with the repulsive force ~20 pN, a relative strong force in biology. The energy at each step of the potential is computed as the average of the continuous VDW–EEF1 function, except for the region corresponding to the energetic minimum.

We model the hydrogen bonding interaction using the reaction algorithm, which has been adapted to the all-atom representation (Fig. 6d). All possible interactions between backbone–backbone, backbone–side chain, and side chain–side chain atoms are included. Long-range electrostatic interactions were not included in the previous work [21]. Recently, we have included the electrostatic interaction between formal charges using the Debye–Hückel approximation, which results in better prediction of protein–peptide and protein–ligand interactions (unpublished work).

Other efforts in methods development of all-atom DMD model include those by Borreguero et al. [29], Emperador et al. [31], and Luo et al. [30]. However, these models are either nontransferable with structure-based interaction models [30] and constraints for specific secondary structure [31], or not systematically benchmarked [29].

3.3 Extension of the Force Field for Small Molecules

Recently, we have extended the Medusa force field in order to model small molecule ligands [42] by introducing new atom types and parameterizing the pairwise VDW and EEF1 interactions. We performed a benchmark of the new force field by predicting the binding affinities of a large set of protein–ligand complexes. The correlation coefficient between the computational and experimental affinities is approximately 0.6, which is comparable to other existing computational approaches. Additionally, we developed a flexible ligand docking method using the new force field for both ligand and pose selection [43]. The results of the docking benchmark are comparable to or better than those of other flexible docking programs on the market [43]. Therefore, the extended Medusa force field is useful in modeling small molecules.

We discretized the small molecule Medusa force field extension in order to model small molecules in DMD simulations. Using a similar discretization protocol to that described above for VDW–EEF1, we can readily obtain the nonbonded interactions for small molecules. Since there are an insufficient number of high-resolution small molecule structures to determine the parameters for the bonded terms, we

simply use the accepted average length R_0 and a fixed ratio $\sigma = 0.02$ to model the covalent bond and bond angles, $[R_0(1-\sigma),\ R_0(1+\sigma)]$. For the dihedral angles, we first determine the hybridization of the two central atoms, which determines the symmetry of the dihedral angle: threefold symmetry for sp^3–sp^3, twofold for sp^2–sp^2, and continuous for sp^2–sp^3. For simplicity, we assume a variation of 36° for each ideal angle and compute the multistep potential accordingly, with the energy barrier (ΔE) set as $2k_BT \sim 1.2\,\text{kcal/mol}$ to ensure enough transition between different rotamers ($p \sim \exp(-\Delta E/k_BT)$). Using the extended DMD force field, we are able to perform simulations of the interactions between proteins and small molecules. Since the extended force field also includes nucleotides, we are also able to model both DNA and RNA in DMD.

4 DMD Simulations of Biomolecules

4.1 Folding of Small, Fast-Folding Proteins

Given the vast conformational space available to proteins, the ability to capture protein native states provides an important, milestone benchmark test for all-atom DMD simulations. We performed ab initio folding simulations of six structurally diverse proteins using all-atom DMD with implicit solvation: Trp-cage (20 residues; a mini α/β protein); WW domain (26 residues; the central three strand β-sheet [Gly5-Glu30] of the all-β protein), villin head-piece (35 residues; an all-α protein); GB1 domain (56 residues; an α/β protein); bacterial ribosomal protein L20 (60 residues; an all-α protein); and the engrailed homeodomain (54 residues; an all-α protein). We demonstrate that, using our method, proteins can achieve the native or near-native states in all cases. For three small proteins—Trp-cage, WW domain, and villin headpiece—multiple folding transitions are observed, and the computationally characterized thermodynamics are in qualitative agreement with experiments. For example, our simulation reproduces the apparent two-state folding thermodynamics of WW domain (Fig. 7a), as observed in previous experiments [44, 45]. Additionally, following the folding trajectory in DMD simulations allows us to examine the folding pathway in detail. For the typical folding trajectory of WW (http://dokhlab.unc.edu/research/Abinitio/), we find that the initial folding event features the formation of the first two β-strands. This finding is consistent with experimentally observed kinetics, where the first two strands are more ordered in the folding transition state than the rest of the protein [46]. Such a kinetic folding intermediate was observed only recently in microsecond-long MD simulations with explicit solvent using the state-of-the-art Anton supercomputer, which is optimized specifically for MD simulations [28]. In contrast, our simulations were performed on personal computers, highlighting the computational efficiency of DMD simulations.

Due to the complex nature of protein folding and the fact that the tested proteins are small in size with relatively simple topologies, we do not expect our method to fully resolve the protein folding problem. We do posit that our all-atom DMD

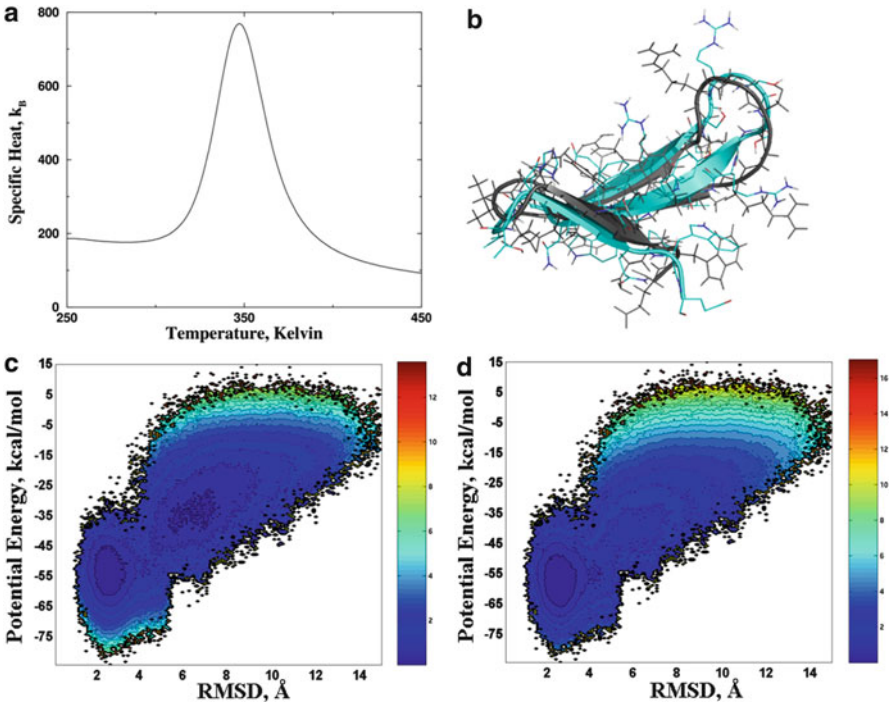

Fig. 7 All-atom DMD simulation of the WW domain. (**a**) Specific heat computed from simulations exhibits a sharp peak at $T \approx 350$ K. (**b**) The alignment between the native state and the representative folded structure in simulations. The contour plot of the 2D-PMF is plotted as the function of potential energy and RMSD at $T = 348$ K (**c**) and $T = 320$ K (**d**)

method can be used for the accurate sampling of conformational space for proteins and protein–protein complexes, which is crucial for protein engineering and the design of protein–protein and protein–ligand interactions.

4.2 Protein–Protein Design

Yin et al. used all-atom DMD simulations in *de novo* protein–protein interface design, where the amino acid sequences of a scaffold protein (human hyperplastic discs protein) were designed to bind a target protein (p21-activated kinase, PAK1). In the design protocol, DMD simulations were utilized for fast conformational sampling, and the RosettaDesign[93] software was used for sequence sampling. The DMD and RosettaDesign steps were performed iteratively in order to attain optimal protein designs that are at global energetic minima in both conformational and sequence spaces. We found that introducing DMD simulations allows for the effective sampling of the protein backbone conformation, which in turn remarkably enriched

the sequence space compatible with the target complex structure. Compared to the initial design obtained without using DMD, the final design had significant backbone (RMSD = 0.82 Å) and rigid-body (RMSD = 3.8 Å) movement. As a result of the backbone movement, 19 out of the 21 interface sites had different amino acids in the final design as compared to the initial design. The final design was experimentally verified to have a binding affinity of ~100 μM to the target protein, and significantly improved solubility as compared to the wild-type human hyperplastic discs protein [32] (Fig. 8).

4.3 Protein Dynamic Coupling and Allosteric Engineering of Kinases

The ability to modulate protein activity in a living cell with temporal control is crucial for our understanding of biological function. We hypothesize that protein dynamics is highly heterogeneous with long range dynamic coupling, and that perturbing distal regions dynamically coupled to the functional site can regulate a protein's function. Such an allosteric regulation is commonly utilized by cell, where the binding of a ligand on one site of the protein can turn the protein's function on or off. We performed DMD simulations of the catalytic domain of focal adhesion kinases (FAK). Based on the simulation trajectory, we found that the catalytically important loop, the G-loop, is strongly coupled to a loop (the insertion loop) that is connected by a β-hairpin (Fig. 9a, b) [33]. We reengineered the insertion loop by inserting a rationally designed unstable FK506-binding protein (iFKBP) domain. This intrinsically metastable domain is stabilized upon the addition of the drug rapamycin (or its analogs) in the presence of FRB. Using the DMD force field extended to include small molecules, we performed DMD simulations in order to study the impact of ligand binding on the conformational dynamics of the catalytic domain of FAK. We showed that the allosteric coupling of FKBP and the catalytic loop allows FAK to be activated via stabilization of FKBP by drug binding (Fig. 9c). *In vivo* experiments using the engineered FAK kinases showed that the protein's kinase function can indeed be regulated by the addition of the ligand. We have demonstrated the transferability of this design approach with other kinases, such as Src and p38 [33]. Therefore, using the allosteric interactions uncovered by DMD, we created a transferrable toolkit for creating regulatable kinases.

5 Conclusion

DMD was originally developed for simple hard sphere systems. In the past, DMD simulations were often associated with coarse-grained molecular systems. With the recent development of a high-resolution DMD force field as well as advances in DMD efficiency, DMD simulations have been applied to study the dynamics of

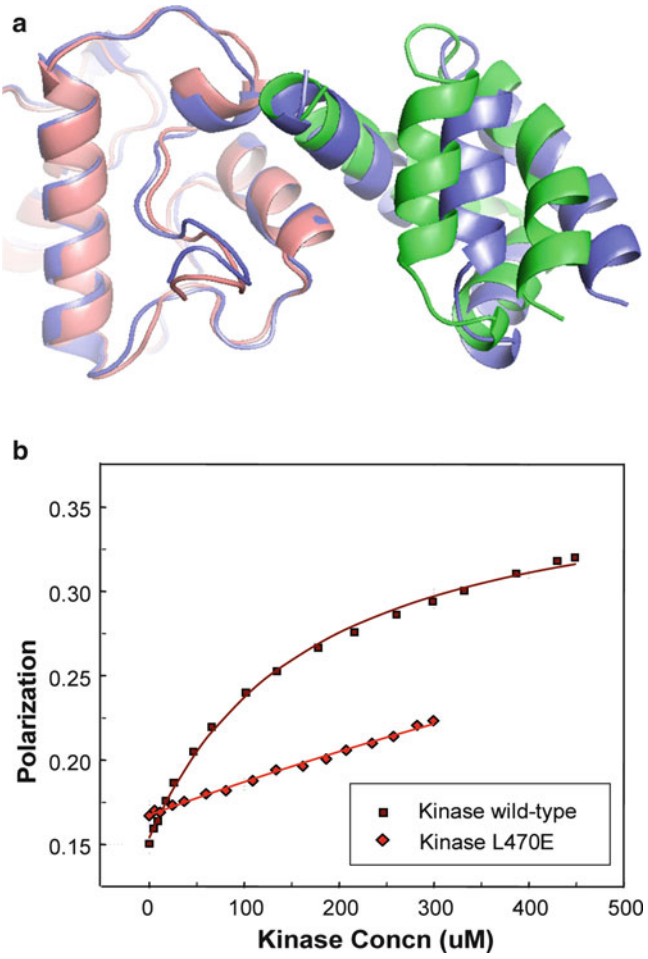

Fig. 8 De novo protein–protein interface design using DMD and Rosetta. (**a**) Starting from the initial structure (*blue*), the DMD assisted design has significant backbone movement in both the scaffold (*green*) and target (*magenta*) proteins. (**b**) The experimental binding assay of the protein–protein complex redesigned using DMD-Rosetta. The redesigned scaffold protein has a binding affinity of $\sim$100 μM with the wild-type target protein. No binding is found in the control experiment with PAK1 mutant L470E, indicating that the actual binding interface is the same as predicted

biological macromolecules. With the continuous development of the methodology, including the parallelization of simulation approaches [47, 48], in the future the DMD engine will be extended to sample the dynamics of ever larger molecules and molecular complexes with even longer time scales. With its ability to efficiently

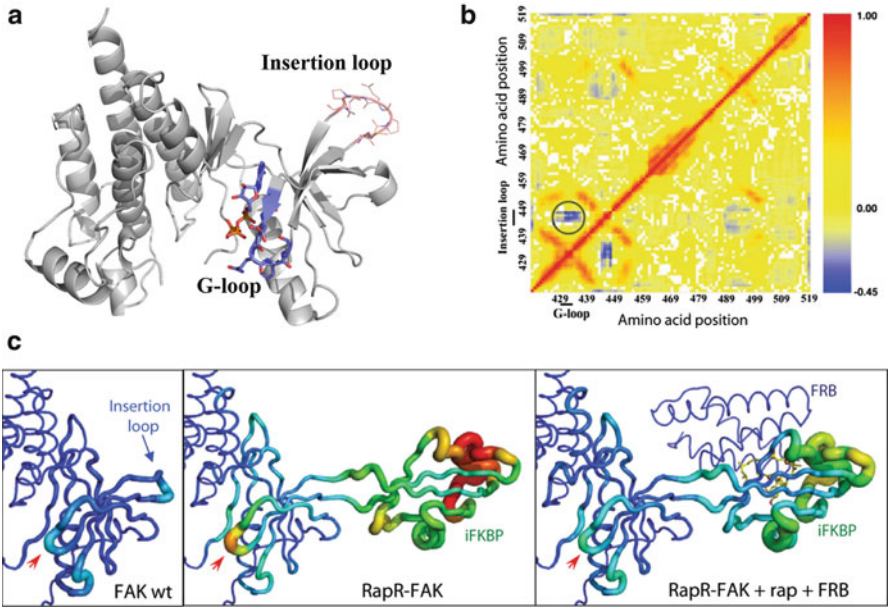

Fig. 9 Mechanism of regulation by iFKBP; Src regulation. (**a**) The portion of the FAK catalytic domain targeted for insertion of iFKBP (*blue*) and the G-loop (*red*). (**b**) Dynamic correlation analysis of the wild-type FAK catalytic domain (*red*, positive correlation; *blue*, negative correlation). The *circled region* indicates strong negative correlation between the movement of the insertion loop and the G-loop. (**c**) Tube representation depicting changes in the dynamics of the N-terminal lobe of the FAK catalytic domain, based on DMD simulations. Warmer colors and thicker backbone correspond to higher root mean squared fluctuation (RMSF) values, reflecting the degree of free movement within the structure. The *red arrows* point to the G-loop

sample the conformational dynamics of complicated systems, DMD simulations will play an important role in our understanding of biology and the effort to combat human diseases.

References

1. Bernado, P., Blackledge, M.: Structural biology: proteins in dynamic equilibrium. Nature **468**, 1046–1048 (2010)
2. Hvidt, A., Nielsen, S.O.: Hydrogen exchange in proteins. Adv. Protein. Chem. **21**, 287–386 (1966)
3. Linderstrom-Lang, K.U.: Deuterium exchange and protein structure. Methuen, London 1958
4. Englander, S.W., Kallenbach, N.R.: Hydrogen exchange and structural dynamics of proteins and nucleic acids. Q Rev. Biophys. **16**, 521–655 (1983)
5. Ishima, R., Torchia, D.A.: Protein dynamics from NMR. Nat. Struct. Biol. **7**, 740–743 (2000)
6. Karplus, M., McCammon, J.A.: Molecular dynamics simulations of biomolecules. Nat. Struct. Biol. **9**, 646–652 (2002)

7. Alder, B.J., Wainwright, T.E.: Phase transition for a hard sphere system. J. Chem. Phys. **27**, 2 (1957)
8. Rapaport, D.C.: Molecular-dynamics simulation of polymer-chains with excluded volume. J. Phys. A Math. Gen. **11**, L213–L217 (1978)
9. Rapaport, D.C.: Event scheduling problem in molecular dynamic simulation. J. Comput. Phys. **34**, 184–201 (1980)
10. Denlinger, M.A., Hall, C.K.: Molecular dynamics simulation results for the pressure of hard-chain fluids. Mol. Phys. **71**, 541–559 (1990)
11. Alejandre, J., Chapela, G.A.: Molecular-dynamics for discontinuous potentials.3. compressibility factors and structure of hard polyatomic fluids. Mol. Phys. **61**, 1119–1130 (1987)
12. Chapela, G.A., Martinezcasas, S.E., Alejandre, J.: Molecular-dynamics for discontinuous potentials.1. General-method and simulation of hard polyatomic-molecules. Mol. Phys. **53**, 139–159 (1984)
13. Smith, A.V., Hall, C.K.: alpha-helix formation: discontinuous molecular dynamics on an intermediate-resolution protein model. Proteins-Struct. Func. Genet. **44**, 344–360 (2001)
14. Smith, S.W., Hall, C.K., Freeman, B.D.: Large-scale molecular-dynamics study of entangled hard-chain fluids. Phys. Rev. Lett. **75**, 1316–1319 (1995)
15. Smith, S.W., Hall, C.K., Freeman, B.D.: Molecular dynamics for polymeric fluids using discontinuous potentials. J. Comput. Phys. **134**, 16–30 (1997)
16. Zhou, Y., Karplus, M.: Folding thermodynamics of a model three-helix-bundle protein. Proc. Natl. Acad. Sci. U. S. A. **94**, 14429–14432 (1997)
17. Zhou, Y., Karplus, M.: Interpreting the folding kinetics of helical proteins. Nature **401**, 400–403 (1999)
18. Ding, F., Borreguero, J.M., Buldyrey, S.V., Stanley, H.E., Dokholyan, N.V.: Mechanism for the alpha-helix to beta-hairpin transition. Proteins **53**, 220–228 (2003)
19. Ding, F., Buldyrev, S.V., Dokholyan, N.V.: Folding Trp-cage to NMR resolution native structure using a coarse-grained protein model. Biophys. J. **88**, 147–155 (2005)
20. Ding, F., Dokholyan, N.V., Buldyrev, S.V., Stanley, H.E., Shakhnovich, E.I.: Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. J. Mol. Biol. **324**, 851–857 (2002)
21. Ding, F., Tsao, D., Nie, H., Dokholyan, N.V.: Ab initio folding of proteins with all-atom discrete molecular dynamics. Structure **16**, 1010–1018 (2008)
22. Peng, S., Ding, F., Urbanc, B., Buldyrev, S.V., Cruz, L., Stanley, H.E., Dokholyan, N.V.: Discrete molecular dynamics simulations of peptide aggregation. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **69**, 041908 (2004)
23. Rahman, A.: Correlations in motion of atoms in liquid argon. Phys. Rev. A Gen. Phys. **136**, A405-A411 (1964)
24. Stilling, F.h., Rahman, A.: Improved simulation of liquid water by molecular-dynamics. J. Chem. Phys. **60**, 1545–1557 (1974)
25. McCammon, J.A., Gelin, B.R., Karplus, M.: Dynamics of folded proteins. Nature **267**, 585–590 (1977)
26. Shirts, M., Pande, V.S.: COMPUTING: screen savers of the world unite! Science **290**, 1903–1904 (2000)
27. Piana, S., Sarkar, K., Lindorff-Larsen, K., Guo, M., Gruebele, M., Shaw, D.E.: Computational design and experimental testing of the fastest-folding beta-sheet protein. J. Mol. Biol. **405**, 43–48 (2010)
28. Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R.O., Eastwood, M.P., Bank, J.A., Jumper, J.M., Salmon, J.K., Shan, Y., Wriggers, W.: Atomic-level characterization of the structural dynamics of proteins. Science **330**, 341–346 (2010)
29. Borreguero, J.M., Urbanc, B., Lazo, N.D., Buldyrev, S.V., Teplow, D.B., Stanley, H.E.: Folding events in the 21–30 region of amyloid beta-protein (Abeta) studied in silico. Proc. Natl. Acad. Sci. U. S. A. **102**, 6015–6020 (2005)

30. Luo, Z., Ding, J., Zhou, Y.: Folding mechanisms of individual beta-hairpins in a Go model of Pin1 WW domain by all-atom molecular dynamics simulations. J. Chem. Phys. **128**, **225103** (2008)
31. Emperador, A., Meyer, T., Orozco, M. Protein flexibility from discrete molecular dynamics simulations using quasi-physical potentials. Proteins **78**, 83–94 (2009)
32. Jha, R.K., Leaver-Fay, A., Yin, S., Wu, Y., Butterfoss, G.L., Szyperski, T., Dokholyan, N.V., Kuhlman, B.: Computational design of a PAK1 binding protein. J. Mol. Biol. **400**, 257–270 (2010)
33. Karginov, A.V., Ding, F., Kota, P., Dokholyan, N.V., Hahn, K.M.: Engineered allosteric activation of kinases in living cells. Nat. Biotechnol. **28**, 743–747 (2010)
34. Ding, F., Dokholyan, N.V.: Emergence of protein fold families through rational design. PLoS Comput. Biol. **2**, e85 (2006)
35. Proctor, E.A., Ding, F., Dokholyan, N.V.: Structural and thermodynamic effects of post-translational modifications in mutant and wild type Cu, Zn Superoxide Dismutase. J. Mol. Biol. **408**, 555–567 (2011)
36. Allen, M.P., Tildersley, D.J.:. Computer simulation of liquids. Clarendon Press, New York, (1989)
37. Andersen, H.C.: Molecular-dynamics simulations at constant pressure and-or temperature. J. Chem. Phys. **72**, 2384–2393 (1980)
38. Paul, G.: A complexity O(1) priority queue for event driven molecular dynamics simulations. J. Comput. Phys. **221**, 615–625 (2007)
39. Liu, J.X., Bowman, T.L., Elliott, J.R.: Discontinuous molecular-dynamics simulation of hydrogen-bonding systems. Ind. Eng. Chem. Res. **33**, 957–964 (1994)
40. Liu, J.X., Elliott, J.R.: Screening effects on hydrogen bonding in chain molecular fluids: thermodynamics and kinetics. Ind. Eng. Chem. Res. **35**, 2369–2377 (1996)
41. Lazaridis, T., Karplus, M.: Effective energy function for proteins in solution. Proteins **35**, 133–152 (1999)
42. Yin, S., Biedermannova, L., Vondrasek, J., Dokholyan, N.V.: MedusaScore: an accurate force field-based scoring function for virtual drug screening. J. Chem. Inf. Model **48**, 1656–1662 (2008)
43. Ding, F., Yin, S., Dokholyan, N.V.: Rapid flexible docking using a stochastic rotamer library of ligands. J. Chem. Inf. Model **50**, 1623–1632 (2010)
44. Ferguson, N., Berriman, J., Petrovich, M., Sharpe, T.D., Finch, J.T., Fersht, A.R.: Rapid amyloid fiber formation from the fast-folding WW domain FBP28. Proc. Natl. Acad. Sci. U. S. A. **100**, 9814–9819 (2003)
45. Ferguson, N., Johnson, C.M., Macias, M., Oschkinat, H., Fersht, A.: Ultrafast folding of WW domains without structured aromatic clusters in the denatured state. Proc. Natl. Acad. Sci. U. S. A. **98**, 13002–13007 (2001)
46. Deechongkit, S., Nguyen, H., Powers, E.T., Dawson, P.E., Gruebele, M., Kelly, J.W.: Context-dependent contributions of backbone hydrogen bonding to beta-sheet folding energetics. Nature **430**, 101–105 (2004)
47. Miller, S., Luding, S.: Event-driven molecular dynamics in parallel. J. Comput. Phys. **193**, 10 (2004)
48. Herbordt, M.C., Khan, M.A., Dean, T.: Parallel discrete event simulation of molecular dynamics through event-based decomposition. In Application-specific Systems, Architectures and Processors, 2009. ASAP 2009. 20th IEEE International Conference, Boston, MA (2009)

Small Molecule Docking from Theoretical Structural Models

Eva Maria Novoa, Lluis Ribas de Pouplana, and Modesto Orozco

1 Docking as a Method for Drug Design

Structural approaches to rational drug design rely on the basic assumption that pharmacological activity requires, as necessary but not sufficient condition, the binding of a drug to one or several cellular targets, proteins in most cases. The traditional paradigm assumes that drugs that interact only with a single cellular target are specific and accordingly have little secondary effects, while promiscuous molecules are more likely to generate undesirable side effects. However, current examples indicate that often efficient drugs are able to interact with several biological targets [1]

E.M. Novoa
Joint IRB-BSC Research Program in Computational Biology, Barcelona Supercomputing Center and Institute for Research in Biomedicine, IRB, Josep Samitier 1–5, Barcelona 08028, Spain

Cell and Developmental Biology, Institute for Research in Biomedicine, Josep Samitier 1–5, Barcelona 08028, Spain
e-mail: eva.novoa@irbbarcelona.org

L.R. de Pouplana
Cell and Developmental Biology, Institute for Research in Biomedicine, Josep Samitier 1–5, Barcelona 08028, Spain

Institució Catalana per la Recerca i Estudis Avançats, Passeig Lluis Companys 23, Barcelona 08010, Spain
e-mail: lluis.ribas@irbbarcelona.org

M. Orozco (✉)
Joint IRB-BSC Research Program in Computational Biology, Barcelona Supercomputing Center and Institute for Research in Biomedicine, IRB, Josep Samitier 1–5, Barcelona 08028, Spain

Institució Catalana per la Recerca i Estudis Avançats, Passeig Lluis Companys 23, Barcelona 08010, Spain

Structural Bioinformatics Node Instituto Nacional de Bioinformática, Institute of Research in Biomedicine, Josep Samitier 1–5, Barcelona 08028, Spain
e-mail: modesto@mmb.pcb.ub.es

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_4, © Springer Science+Business Media, LLC 2012

and in fact some dirty drugs,[1] such as chlorpromazine, dextromethorphan, and ibogaine exhibit desired pharmacological properties [2]. These considerations highlight the tremendous difficulty of designing small molecules that both have satisfactory ADME properties and the ability of interacting with a limited set of target proteins with a high affinity, avoiding at the same time undesirable interactions with other proteins. In this complex and challenging scenario, computer simulations emerge as the basic tool to guide medicinal chemists during the drug discovery process.

Since early works in the 1980s, molecular docking has arised as a leading simulation technique to facilitate the drug design. The traditional paradigm of docking, known as rigid-body docking approach, assumes implicitly the Fisher's lock-and-key model [3], and considers that the ligand-induced structural changes of the protein are negligible [4]. However, drugs generally exhibit a certain degree of flexibility, and the bioactive conformation might not be the most stable conformation in solution [5, 6]. This fact leads to the need of considering drug flexibility for a successful docking simulation. Furthermore, analysis of the Protein Data Bank [7] reveals that ligand binding can introduce non-negligible changes in protein structure which often affect the binding site, raising tremendous difficulties for docking techniques, especially in cases where structural changes are not only binding-specific, but also drug-specific [8]. A second limitation in docking experiments arises from the evaluation of the ligand-binding free energy. Free-energy simulation techniques are expensive calculations that remain impractical for the evaluation of large numbers of ligands [9]. Current docking strategies are based on the combination of very fast functions, which intend to predict binding poses and rank them by means of a more complex equation (the "scoring function"), which has been parameterized to reproduce experimental binding data of protein–drug complexes [10]. However, scoring functions implemented in docking programs make various assumptions and simplifications, and do not fully account for all phenomena that determine molecular recognition.

Despite all the challenges, the major practical limitation for docking procedures does not emerge from technical uncertainties in the evaluation or scoring of docking poses, but comes from the lack of experimentally solved protein structures. Indeed, despite the massive effort focused in the experimental resolution of protein structures, 2010 version of the PDB contains less than 4,000 unique human proteins, while RefSeq [11] suggests the existence of nearly 100,000 human proteins, twice or more if splicing variants are considered. Therefore, the current version of PDB is covering only around 4% of the known human proteome [12]. This sequence-structure gap becomes even larger if we consider proteins from virus, bacteria, or other pathogens for which less amount of structural information exists.

The evaluation of the potential interactions of drugs with multiple targets is severely limited if the analysis relies exclusively on experimentally solved structures. Fortunately, this limitation can be partially solved with the use of predicted models of proteins as templates for docking (Fig. 1). In this chapter, we very briefly

[1] Drugs that bind to several molecular targets or receptors, and therefore tend to have a wide range of effects and possibly negative side effects.

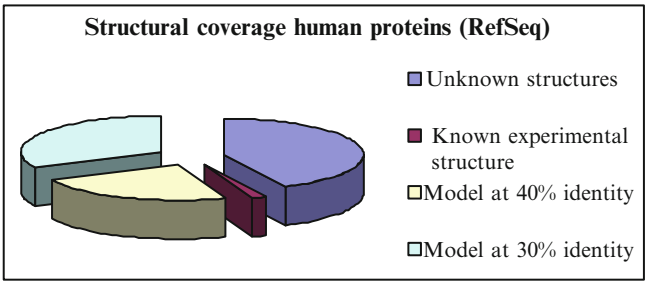

Fig. 1 Structural coverage of human proteins according to RefSeq without including splicing variants

review the state-of-the-art of docking procedures, making special emphasis on the potential use of ensembles of structural protein models derived from homology modeling in high-throughput docking experiments.

2 Docking Algorithms

There is a plethora of docking algorithms and strategies that have been implemented in a large variety of computer programs, some local and used by a restricted community, and others commercially available that have a wide user community. It is out of our scope to review all of them here, and we just outline the basic formalism behind the most popular ones. The reader is addressed to excellent reviews to gain a more complete view on current algorithms [10, 13–16].

In principle, all docking algorithms follow a stepwise procedure: (1) several estimates of the ligand–protein complex (binding poses) are proposed, and (2) these poses are then ranked using a scoring function and offered to the user, who typically focuses his/her attention to the best scored ones. Given that scoring functions are fitted against experimental binding data, scoring values have "free energy of binding" units. Therefore, they can be used to differentiate between good and bad drug candidates and even to have an estimate of the binding free energy of the drug.

The differences between the different docking programs rely on (1) the method used to explore the drug-binding landscape, (2) the method used to introduce flexibility, and (3) the nature and the parameterization of the scoring function. For example, DOCK [17], one of the first widely used docking programs, performs a geometrically based docking of the ligands based on isomorphic subgraph matching algorithms [18], which is later refined by considering the chemical nature of the ligand and the binding site. Different scoring functions—mostly in the AMBER [19] force-field—are used during the different stages of the fitting and ranking process, including complex physical functions calling to atomistic force-field calculations coupled to Generalized Born or Poisson–Boltzmann calculations. The popular AUTODOCK program [20] offers a variety of optimizers including Monte Carlo simulated annealing and different genetic algorithms using smoothed potential

energy terms precomputed in a regular grid.[2] Scoring is performed considering ligand-entropic terms and desolvation contributions in addition to ligand–protein interaction terms. GOLD [21], another very popular program, uses a sampling protocol similar to the genetic algorithm implemented in AUTODOCK and a very wide range of well-validated scoring functions, which include specific corrections such as those for metal ions and covalent interactions [22]. This program includes also specific scoring functions for kinases and offers the possibility to incorporate user-refined scoring functions. The program FLExX [23], which has also an excellent record of success, uses a geometry-fitting algorithm derived from computer vision engineering, where drugs grow in optimum orientations and conformations at the binding site from an original seed fragment. The program permits the introduction of knowledge-based pharmacological restraints and the incorporation of essential water molecules and crucial metal ions in the binding site. Scoring is based on a simple physical scoring function based on OPLS [24] force-field parameters. ICM [25], a powerful program to fit small ligands to proteins, uses a smoothed atomistic energy function coupled with a Monte Carlo algorithm in internal coordinates to sample the drug–protein binding space. Its scoring function contains the usual contributions plus two desolvation correction terms. GLIDE [26], a widely used docking program in the pharmaceutical industry, uses a "funnel strategy" where each pose passes a series of hierarchical filters that evaluate the ligand–receptor interactions, including spatial fit, complementarity of interactions using a grid-based method, and finally an evaluation and minimization using OPLS-AA nonbonded ligand–receptor interaction energy. GLIDE incorporates a variety of scoring functions with increasing computational complexity. MedusaDock [27], a recently developed software, is a docking method which models both ligand and receptor flexibility in a rapid manner by using sets of discrete rotamers, obtaining quite good results with targets which are known to be very flexible.

In addition to those implemented in standard programs, many other scoring functions have been developed (for a review see [28]), using experimentally calibrated master equations similar to that in (1).

$$\Delta G_{\text{binding}} = \alpha E_{\text{ele}} + \beta E_{\text{vW}} + \chi E_{\text{HBond}} + \delta G_{\text{desolv}} + \varepsilon S_{\text{lig}} + \phi E_{\text{dist}}^{\text{lig}} + \varphi G_{\text{others}}, \quad (1)$$

where E_{ele} and E_{vW} stand for usual electrostatic and van der Waals terms—typically smoothed to avoid nuclei discontinuities. Hydrogen bonds contribution is sometimes explicitly included in E_{Hbond}, while in others it is captured by E_{ele} and E_{vW}. The ligand and protein desolvation contribution (typically computed from occluded surface/volumes) are included in G_{desolv}, the loss of ligand entropy upon binding is introduced in S_{lig} (typically roughly approximated by counting the number of rotable bonds in the ligand), and the constrained energy is captured by E_{lig}. Other additional terms can be included, such as corrections for covalent interactions,

[2]Representation of the receptor energetic contributions (mainly electrostatic and van der Waals) to be read during the ligand scoring.

cation–π contacts, special metal–ligand interactions, presence of buried waters in the binding cavity, and many others. All these different terms are weighted using parameters that are fitted against empirical data. As discussed above, different programs offer the user the possibility of using family specific scoring functions and to incorporate his/her own scoring functions. However, the large number of available scoring functions has generated an obvious confusion in the users community and has driven to the popularization of strategies based on consensus or meta-scoring functions. Future work needs to be done by the community to order this explosion of different scoring strategies.

Flexibility is treated at different levels by various programs. Ligands with potential drug-like properties tend to be small and moderately flexible, which facilitates the determination of the optimum docking conformation by different methods such as energy minimization, Monte Carlo, genetic algorithms, molecular dynamics, and many others. The complexity here arises from the need to determine which is the optimum geometry in solution [6]. As noted above, the incorporation of the protein flexibility is much more difficult due to the large number of protein degrees of freedom, and none "final" algorithm has been yet developed. Many programs allow the user to refine a reduced number of residues in the protein—generally limited to side chains—by using rotamer libraries [29], Monte Carlo [30], or restrained molecular dynamics [31]. Nevertheless, one of the most popular strategies consists in the "ensemble" docking approach, which assumes that the effect of target flexibility in docking can be represented by using a Boltzmann ensemble of conformations for the protein instead of just a single rigid structure. Different methods for generating ensembles have been proposed, including molecular dynamics from a known experimental structure of the target [32, 33], crystallographic (X-ray) [34–37], and spectroscopic (NMR) [38, 39]-derived structures.

A common feature in most descriptions of new docking methods is the claim that it is more accurate than the competitors. In our experience, the performance of docking algorithms changes in each version and depends quite significantly on the nature of the problem and the skills of the modeler running the project, factors that hinder the validity of the conclusions derived from blind test experiments [40]. An estimate of the market share taken by the different docking algorithms is also difficult to determine, particularly in a scenario of site-licenses, cost-related decisions in the selection of docking engines and where publication is not often a priority. However, a simple analysis of the literature (ISI CITATION MANAGER) in 2009 reveals that the market is quite equally divided among different codes (see Fig. 2).

3 Scenario for Docking Use

The literature is full of examples of use of docking algorithms in drug design procedures, and the documentation accompanying the different computer programs illustrates many examples where docking has been crucial to derive significant results. Even though most docking studies are done inside pharmaceutical industries

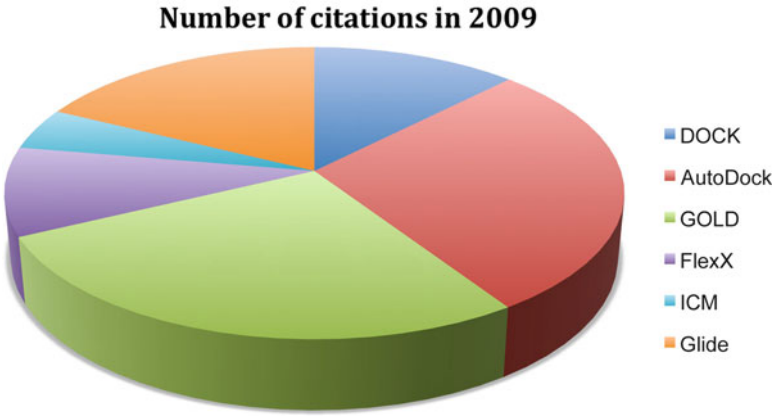

Fig. 2 Number of citations in scientific literature of commonly used docking algorithms in 2009

and are never published, analysis of the literature reveals that the word "docking" has been used in the title or abstract in 1,565 publications during 2009.

Docking can be done in quite different scenarios, where objectives and success criteria can be quite different:

1. Derivation of structural binding mode for a known binder
2. Determination of primary or secondary targets for a drug
3. Virtual high-throughput screening (vHTS)

The derivation of a structural binding mode for a small molecule is probably the most traditional use of docking algorithms. Within this paradigm, the process starts after high-throughput experimental studies (or alternative methods) that detect one or several small molecules which display activity against a given target. However, there are many factors that determine whether these "hits" can become "leads" or can be modified to improve their properties. Such a lead optimization process requires a quite detailed knowledge of the binding mode, something that only in silico docking can provide with the required velocity. In this context, the use of docking methods is defined by the limited number of drugs to consider and by the existence of a single target protein. The accuracy is, however, crucial since errors in the placement of the drug can completely misguide the lead optimization process. A basic metric commonly used for evaluating the accuracy of the predicted binding modes of docking programs is the root mean square deviation (RMSD) between the predicted conformation and the native pose of the ligand:

$$\mathrm{RMSD} = \left(\sum_{N} \frac{(R_i - R_j)^2}{N} \right)^{1/2}, \tag{2}$$

where R stands for the ligand coordinates in the predicted binding mode (i) and in the native pose (j), and N is the total number of atoms. In many practical

cases, the predicted binding mode can be useful even if there is a significant RMSD, provided that some key groups are properly located. Then, it is also convenient to use more case-specific descriptors for the validation of docking methods such as the generalized RMSD:

$$\mathrm{RMSD} = \left(\sum_{N} \frac{\xi_n (R_i - R_j)^2}{N} \right)^{1/2}, \tag{3}$$

where N is the total number of atoms in the drug and the weighting factor ξ_n reflects the importance of the residue n in defining the bioactive drug–protein complex. Many other qualitative measures of structural quality of the docking poses have been suggested [41].

Docking programs do not provide a single pose as an output, but a series of them ranked according to the scoring function. Thus, it is not an uncommon situation that the real binding mode is detected, but not top-ranked by the scoring function. Thus, an additional requirement for the derivation of a structural binding mode is the correct ranking of the good docking solution, which would guarantee that the final user does not disregard it in a further study. A quite common global estimate of the accuracy of the predicted binding mode is the "2 Å RMSD rule," which consists in computing the percentage of predicted binding modes of the ligands that are found at less than 2 Å from the native pose. In a recent study [12], we found that for a selected set of proteins, around 30% of the correctly predicted docked poses are disregarded due to a failure in the scoring of these poses. Thus, instead of correctly predicting the binding mode of 43% of the poses, only 30% of the poses are correctly predicted and scored (see Fig. 3).

The determination of primary or secondary targets for a drug is an increasing field of application for docking algorithms, especially due to the emergence of "drug repositioning" strategies [42], i.e., the identification of new indications for existing drugs. Both new indications and adverse drug reactions are caused by unexpected ligand–protein interactions on secondary targets, and can be explored through docking experiments. The objective here is not necessarily to predict the binding mode with extreme accuracy, but to detect possible targets for a drug.

During the last decades, the dominant philosophy in drug design has been the "one gene, one drug, one disease" paradigm. However, many effective drugs have shown to act via modulation of multiple proteins rather than single targets. Indeed, recent studies suggest that selective compounds compared to multitarget drugs may exhibit lower clinical efficacy [43,44]. In this regard, parallel large-scale multitarget virtual screening is a promising method to derive secondary targets.

The use of docking in vHTS is a common practice in pharmacological research due to its reduced cost compared to experimental HT techniques and to the existence of large virtual chemical libraries—containing over a million of potential ligands—available for screening [10]. The main objective of this type of projects is to mine the original library and derive a small subset of compounds, which has a larger percentage of promising ligand candidates, a process that is known as "enrichment."

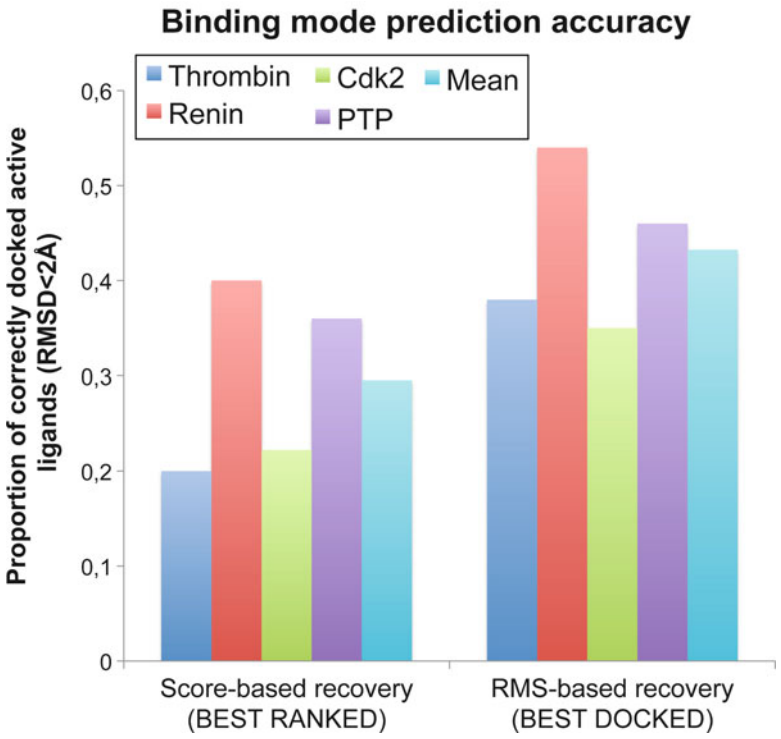

Fig. 3 Binding mode prediction accuracy of for five different human proteins: thrombin, rennin, cyclin-dependent kinase 2 (CDK2), and protein phosphatase 1B (PTP-1B)

Technically, vHTS requires very fast computer strategies, especially in cases where primary and secondary targets are screened simultaneously. Current protocols for vHTS are based on filtering strategies, where basic geometrical or pharmacological criteria are used to obtain a more focused chemical library.

The evaluation of the performance of docking methods is especially important considering the cost of the calculation. Here, the most important objective is to check the ability of the method to discriminate between active compounds and decoys (inactive). A virtual screening run selects a list of molecules (n) from a given database of N entries, which includes both actives (true positive compounds, TP) and decoys (false positive compounds, FP). Actives (A) that have not been found by the screening method are false negatives (FN) and decoys that have not been selected are true negatives (TN). The optimum screening is that able to recover all the true positives, without recovering any false positive. Although it is clear that virtual screening methods can be assessed by their ability to discriminate between active and inactive compounds, assessing the enrichment in a virtual screening procedure is a nontrivial task. Many different enrichment descriptors have been described in the literature [45,46], and they can all provide different information on

the performance of the screening. A combination of several enrichment descriptors is recommended if the aim is to evaluate the performance of a docking algorithm.

The most popular descriptors used to evaluate the quality of docking experiments in this scenario are the *sensitivity* [true positive rate; TPR; see (4)], which indicates the ability of the method to recover the true ligands, and the *specificity* [true negative rate; TNR see (5)], which informs on its ability to avoid decoys.

$$\text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{4}$$

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{\text{FP} + \text{TN}} = 1 - \text{FPR}, \tag{5}$$

where FPR stands for false positive rate. Also, *accuracy* [*Acc*; (6)] describes the percentage of molecules which have been correctly classified by the screening protocol, and the *precision* (positive predictive value; *PPV*) gives accounts for the proportion of true positives among the list of selected compounds given by the docking (7).

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{N}} = \frac{\text{A}}{\text{N}} \times \text{TPR} + \left(1 - \frac{\text{A}}{\text{N}}\right) \times \text{TNR}. \tag{6}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{7}$$

In order to o assess the ability of the models to obtain true actives among the first ranked compounds (an extra requirement in high-throughput docking) [47], the *enrichment factor* [EF, (8)] can be used:

$$\text{EF} = \frac{\text{TP}/n}{A/N}. \tag{8}$$

Recently, receiver operating characteristic (ROC; true positive vs. false positive rates) curves and the associated area under the ROC curves (AUC) have also become very popular to evaluate the discriminatory power of the virtual screening procedure [48–50]. The main advantage of these metrics is that they are independent on the ratio of actives to decoys of the database and accordingly they are good measures of the global performance of a docking algorithm in a vHTS procedure.

4 Protein Structure Prediction

One of the major practical limitations to the use of docking in pharmacological research lies in the need of high accurate structural data for the protein. Fortunately, protein structure can be predicted by a variety of computational methods, homology-modeling (also named comparative modeling) being the most accurate one in cases where there is a clear homolog with known structure [51, 52]. Building a protein structure from homology modeling requires a template—a protein with

similar amino acid sequence—and involves four major steps: fold assignment, sequence alignment, model building, and model refinement. Several computer packages are available to perform all this process automatically, such as the SWISS-MODEL software [53], the 3D-JIGSAW package [54], or the ModWeb tool [55]. Nevertheless, the general consensus [52] is that manually curated models derived from the use of programs, such as MODELLER [56], are more reliable than automatic procedures.

One of the most critical steps in homology modeling is the identification of the proper template. The simplest method that can be used for this purpose is a simple BLAST search [57] against the PDB database. However, methods based on multiple sequence alignments or profiles have demonstrated to be much more sensitive in identifying distantly related homologs [57, 58]. Choosing the best template among the candidates derived from multiple alignments is crucial for the final accuracy of the model and in addition to sequence identity we need to consider that "holo" structures are always better templates than "apo" ones [59]. In the case that several holo candidates are available, we should favor the structure containing a similar ligand to the one that we aim to dock [60, 61].

Another crucial step in the model generation is the alignment of the target with the template(s). This procedure can be done easily with standard alignment algorithms in cases of large identity between template(s) and target protein. However, in difficult cases (below 30% sequence identity), the alignment obtained by standard methods needs to be refined by:

1. Including structural information of the template, i.e., avoiding gaps in secondary structure elements, in buried regions, or between two residues that are far in space [62–65].
2. Building a multiple structure-based alignment of the templates and use them to align the target sequence to it.
3. Calculating the target and template sequence profiles by aligning them with sequences sufficiently similar to the target and template sequences respectively, so that they can be aligned without significant errors. The final target-template alignment is then obtained by aligning the two profiles [66, 67].

In general, the use of multiple structures and multiple sequences benefits from the evolutionary and structural information about the templates and target sequence, and often produces a better alignment for modeling than pairwise alignment methods [68, 69]. In any case, once the template is selected and the target protein is aligned, the structural model can be generated using different approaches. In this context, MODELLER [56], one of the most widely used homology modeling engines, typically builds models by enforcing spatial restraints derived from the template structure(s).

The quality of the structure derived from homology modeling roughly correlates with the sequence identity between the target and the template proteins [70]. Thus, it is accepted that for sequence identities below 30% less than half of the residues have their Cα correctly placed [71, 72]. The percentage of correctly placed residues increases to 85% for identities ranging from 30 to 50% and most of the Cαs are well

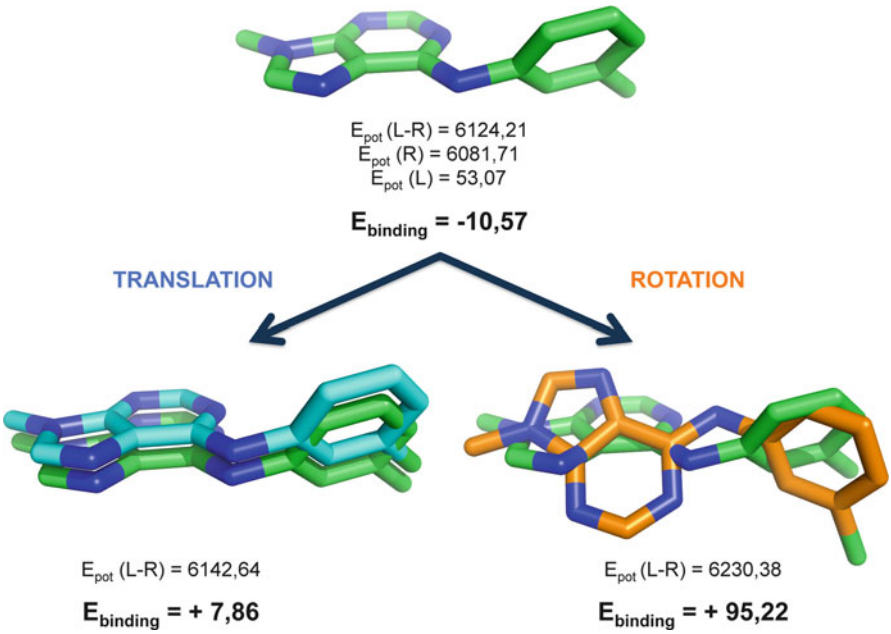

Fig. 4 The ligand–receptor interaction energy is strongly altered by slight translation and/or rotation movements of the ligand. The ligand–receptor binding energy (Ebinding) has been computed as the difference of the potential energy of the complex [Epot(L–R)] with respect to the individual potential energies of the ligand [Epot(L)] and the receptor [Epot(R)]. The ligand shown has been taken from the structure of a human CDK2 (PDB code 1ckp)

positioned for sequence identities above 50%. Inside the high-quality range no direct correlation exists between the accuracy of the model and the sequence identity with the template, and evaluation of the expected quality of a model is still an unsolved problem [73]. In fact, the concept of "accuracy of the model" can be arbitrary, since it depends on its planned use. For example, a model with accuracy around 3.5 Å in backbone positioning may be sufficient for understanding protein function or designing mutations, but is expected to be of small utility for predicting ligand binding [74, 75], since the strong dependence of the ligand–receptor interaction energy on fine geometrical details (Fig. 4) implies that small structural errors might cause a large bias in the binding calculation. A deep discussion on this point is presented in the next section of this chapter.

5 HT Docking from Homology Modeled Structures

The use of homology models in docking calculations has been recently explored by different groups, finding in general quite encouraging results. McGovern and Shoichet [59] performed a high-throughput docking on ten enzymes for which

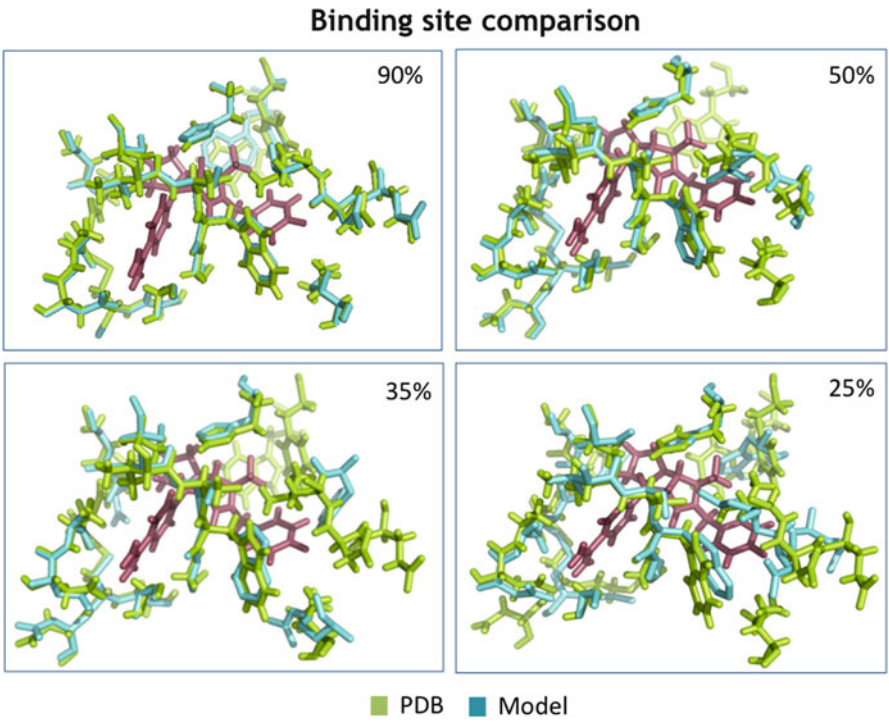

Fig. 5 Binding site comparison of thrombin PDB structure (2cn0, shown in *green*) with homology models of different sequence identity (*blue*). The ligand shown (*magenta*) corresponds to the crystallized ligand in the 2cn0 structure. As can be seen from the figure, even at low sequence identities the binding site structure is still reasonably conserved

apo, holo, and homology model structures were available, suggesting that they were useful for enriching the screening, but in general not as powerful as the holo-crystal structure. Diller and Li [76] reported significant enrichments of the homology models of six kinases with identities in the range of 30–50% when used to screen a large chemical library. Similar results were obtained by Oshiro et al. [77] in the study of two targets (cyclin-dependent kinase 2 (CDK2) and factor VIIa), by Gilson's group [78] with a set of five targets, and by Ferrara and Jacoby [79] in the analysis of insulin growth factor I receptor. All these results suggest that the conservation of the binding sites in modeled structures is sufficient, and does not affect docking accuracy significantly (Fig. 5).

Recently, various groups have suggested [12, 80] using ensembles of homology models as templates, developing automatic strategies valid within the HT-regime (Fig. 6). The use of the ensemble docking approach coupled to homology modeling has two main advantages: (1) there is no need to identify the "best" performing homology model and (2) protein flexibility is implicitly included in the docking run. When using the ensemble docking approach, each homology model is built on

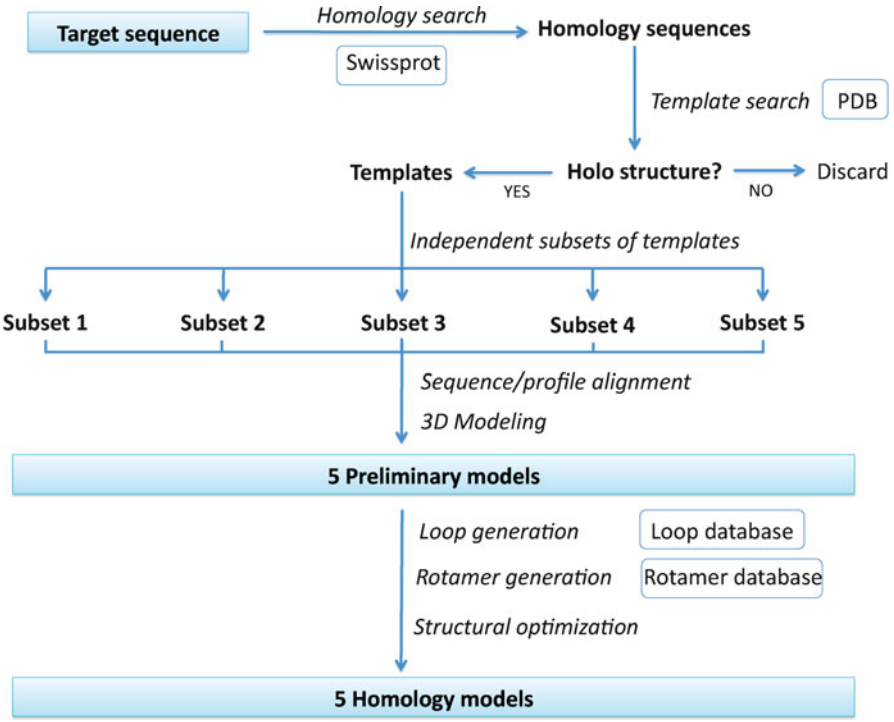

Fig. 6 Example of workflow [12] for building homology models to be used in the ensemble docking approach

a basis of a different template, and thus the binding site is specialized to recognize a different subset of active ligands. As a result, there is an improvement in the probabilities of detecting "true positives" (Fig. 6).

Different studies using ensemble docking with experimental structures have obtained controversial results. Some authors [34] state that ensemble docking clearly improves the performance of the docking process, while others [37, 81, 82] complain about the increase in "false negatives" and suggest that the enrichment of the results using ensembles is not so different when compared to a good-performing crystal structure (although the rules to select "a priori" which is a good-performing crystal structure are not evident). The situation when using homology models is more evident, since in this case the use of ensembles increases very significantly the sensitivity with respect to single models, decreasing only slightly the specificity and leading to an overall clear improvement of the docking results [12]. Figure 7 illustrates the increase in the proportion of correctly predicted binding modes when using ensemble docking compared to single model docking—only homology models are being considered in the figure. In this example, single models produce moderate binding mode predictions, being able to recover 30% of correctly docked

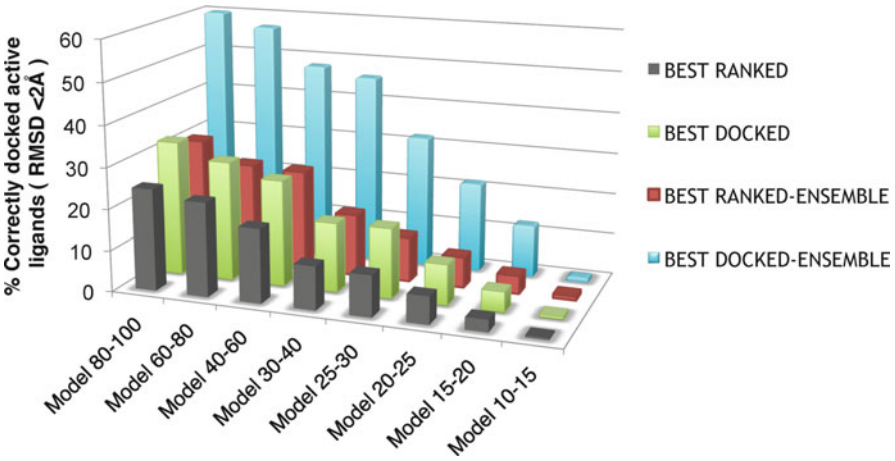

Fig. 7 Recovery of correctly docked active ligands for a selected set of proteins (α-momorcharin, trypsin, p38 kinase, HIV retrotranscriptase, factor Xa, and heat shock protein 90). As can be seen from the figure, the correctly docked ligand recovery is dependent both on the strategy of docking (ensemble docking versus single docking) and on the sequence identity of the template. A ligand is considered as correctly docked when its RMSD with the crystallographic ligand is below 2 Å. Both score-based selection—i.e., best ranked—and RMSD-based selection—i.e., best docked—are shown. Single docking averages are shown in *black* and *green*, whereas ensemble docking averages are shown in *red* and *cyan*. These results were obtained by docking a database containing both known actives and decoys, using Glide docking program in an SP—standard precision—mode (data from [12])

ligands (21% if we only take into account the best ranked solution), whereas the ensemble docking approach increases the correctly docked ligands to 57% (29% when considering the best ranked solution).

Homology modeling-based ensemble docking coupled with good structural models and strict scoring methods can outperform single PDB docking (Fig. 8). Furthermore, the ensemble docking protocol is very robust to the decrease in sequence identity, given that models with sequence identities in the range of 30–40% still provide good results for most proteins.

A better view on the overall quality of the homology-based ensemble docking approach is obtained by analyzing simultaneously its ability for vHTS (i.e., its capability to recover specifically active ligands from the dataset) and in the context of structural determination of binding modes (i.e., its capability to yield good structural solution as the top ranked ones). Results displayed in Figs. 8 and 9 provide evidence on the power of the ensemble docking approach in a wide range of working scenarios.

As a summary, the general accepted "rule" is that only models built with more than 50% sequence identity are accurate enough for docking and the accuracy in docking is higher with holo structures than homology models [75, 77, 83]. However, recent available studies using ensemble docking with homology models

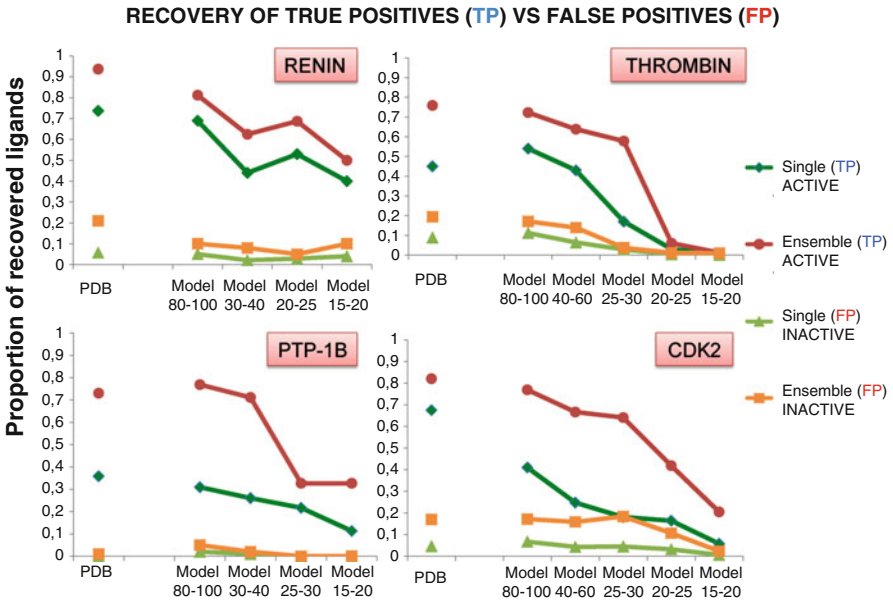

Fig. 8 Ensemble docking versus single docking approach. The performance of both approaches is being compared in terms of recovered active ligands and decoys for four human proteins: renin, thrombin, cyclin- CDK2, and PTP-1B. The single docking approach performance is represented with *green* and lime lines, which correspond to the recovery of active and inactive ligands, respectively. Similarly, the *red* and *orange lines* correspond to the active and inactive ligand recovery, respectively, when using an ensemble docking approach. In all cases, the difference between active (true ligands) and inactive (decoys) recovery is higher when using ensemble docking. Results where obtained using Glide computer program in an extra-precision (XP) mode with a GlideScore (GS) threshold $= -8$ (data from [12])

[12] strongly suggest that models with sequence identity above 30–40% display a considerable ability to specifically recover active ligands, and can even outperform single crystal structures. Although it is difficult to extend results of the small set of proteins used in these studies to the entire proteome, the use of ensemble docking is extremely recommended over single docking, especially when using homology models. Moreover, the use of homology models is not limited to the retrieval of active ligands from a chemical library, but can also provide structural complexes with sufficient accuracy for lead optimization processes.

6 Increasing Coverage

As noted in the beginning of this chapter, despite the tremendous effort focused for many decades in the experimental determination of protein structures, the current version of PDB covers only a small fraction of human proteins. This coverage is

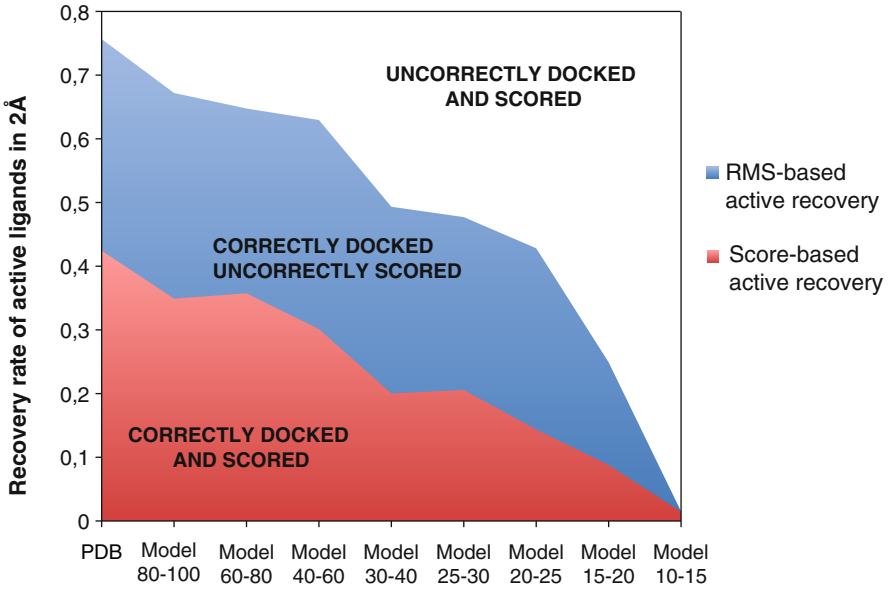

Fig. 9 Expected percentage of success in docking experiments performed using the ensemble docking approach (both for X-ray and homology models, the later obtained from templates with different degrees of sequence identities). Recovery rates have been computed as average recovery rates of four human proteins: renin, thrombin, CDK2, and PTP-1B

even smaller if we focus on protein structures coming from pathogens. Our group and others [12, 80] have suggested that homology models derived from templates with identity ranges of 30–40% can significantly enrich chemical libraries. These results allow us to expand dramatically the universe of use of docking techniques (Fig. 1), especially in the case of human proteins with pharmacological interest (taken from DrugBank database; [84]), which are covered over 75% when using homology models up to 30% identity (Fig. 10).

Thus, with all the required cautions needed in the use of homology models for docking purposes (related mostly to the problems in finding good templates and in determining "a priori" the quality of the model), the use of comparative models can enlarge dramatically the universe of applicability of small-molecule docking approaches. Ensemble docking performed on homology models provides results of similar, or even better quality than those obtained with single crystal structures, leading to a clear enrichment in the chemical libraries, and producing poses of good structural quality, even in cases where ligand binding implies non-negligible changes in protein structure. Altogether ensemble docking from homology modeling appears as a promising alternative to extend the use of docking strategies in drug-design pipelines.

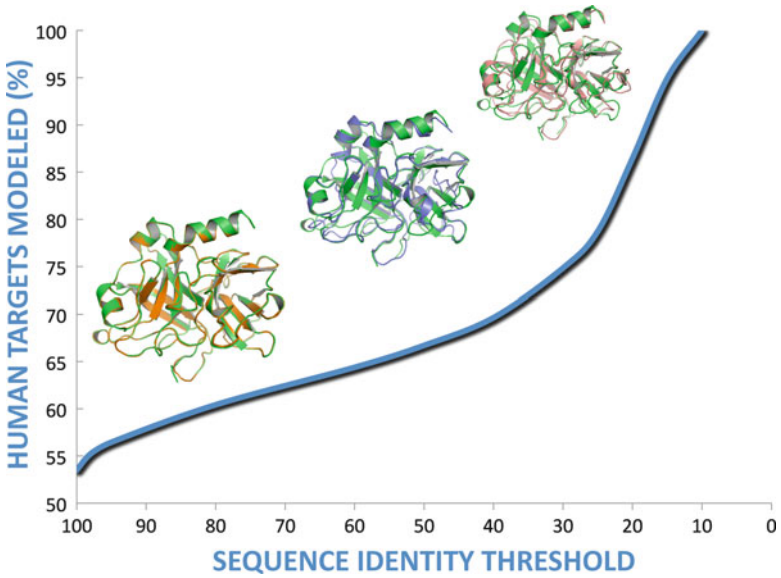

Fig. 10 Structural coverage of human targets of pharmacological interest depending on the sequence identity threshold used in homology modeling. A 30% sequence identity threshold—which still gives very good results when using the ensemble docking approach—allows us to cover 41% more human drug targets, obtaining a final coverage of 75% of the human drug targets. The superimposition of the crystal structure of thrombin (*green*, PDB code 2cn0) and homology models built with different sequence identity—90% (*orange*), 50% (*blue*), and 30% (salmon)—is also shown

References

1. Campbell, S.J., Gold, N.D., Jackson, R.M., Westhead, D.R.: Ligand binding: functional site location, similarity and docking. Curr. Opin. Struct. Biol. **13**(3), 389–395 (2003)
2. Keiser, M.J., et al.: Relating protein pharmacology by ligand chemistry. Nat. Biotechnol. **25**(2), 197–206 (2007)
3. Fisher, E.: Einfluss der Konfiguration auf die Wirkung der Enzyme. Berichte der Deutschen Chemischen Gesellschaft. **27**, 2985–2993 (1894)
4. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., Ferrin, T.E.: A geometric approach to macromolecule–ligand interactions. J. Mol. Biol. **161**(2), 269–288 (1982)
5. Butler, K.T., Luque, F.J., Barril, X.: Toward accurate relative energy predictions of the bioactive conformation of drugs. J. Comput. Chem. **30**(4), 601–610 (2009)
6. Perola, E., Charifson, P.S.: Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. J. Med. Chem. **47**(10), 2499–2510 (2004)
7. Berman, H.M., et al.: The protein data bank. Nucleic. Acids. Res. **28**(1), 235–242 (2000)
8. Cozzini, P., et al.: Target flexibility: an emerging consideration in drug discovery and design. J. Med. Chem. **51**(20), 6237–6255 (2008)
9. Merz, K.M.: Limits of free energy computation for protein–ligand interactions. J Chem. Theory. Comput. **6**(4), 1018–1027 (2010)

10. Kitchen, D.B., Decornez, H., Furr, J.R., Bajorath, J.: Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug. Discov. **3**(11), 935–949 (2004)
11. Pruitt, K.D., Tatusova, T., Maglott, D.R.: NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic. Acids. Res. **33**(Database issue), D501–504 (2005)
12. Novoa, E.M., de Pouplana, L.R., Barril, X., Orozco, M.: Ensemble docking from homology models. J. Chem. Theory. Comput. **6**(8), 2547–2557 (2010)
13. Halperin, I., Ma, B., Wolfson, H., Nussinov, R.: Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins. **47**(4), 409–443 (2002)
14. Leach, A.R., Shoichet, B.K., Peishoff, C.E.: Prediction of protein–ligand interactions. Docking and scoring: successes and gaps. J. Med. Chem. **49**(20), 5851–5855 (2006)
15. Shoichet, B.K., McGovern, S.L., Wei, B., Irwin, J.J.: Lead discovery using molecular docking. Curr. Opin. Chem. Biol. **6**(4), 439–446 (2002)
16. Sousa, S.F., Fernandes, P.A., Ramos, M.J.: Protein-ligand docking: current status and future challenges. Proteins. **65**(1), 15–26 (2006)
17. Shoichet, B.K., Bodian, D.L., Kuntz, I.D.: Molecular docking using shape descriptors. J. Comput. Chem. **13**, 380–397 (1992)
18. Gardiner, E.J., Willett, P., Artymiuk, P.J.: Graph-theoretic techniques for macromolecular docking. J. Chem. Inf. Comput. Sci. **40**(2), 273–279 (2000)
19. Ponder, J.W., Case, D.A.: Force fields for protein simulations. Adv. Protein. Chem. **66**, 27–85 (2003)
20. Morris, G.M. et al.: Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function J. Comput. Chem. **19**, 1639–1662 (1998)
21. Jones, G., Willett, P., Glen, R.C.: Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J. Mol. Biol. **245**(1), 43–53 (1995)
22. Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W., Taylor, R.D.: Improved protein-ligand docking using GOLD. Proteins. **52**(4), 609–623 (2003)
23. Rarey, M., Kramer, B., Lengauer, T., Klebe, G.: A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. **261**(3), 470–489 (1996)
24. Jorgensen, W.L., Tirado-Rives, J.: The OPLS force field for proteins. Energy minimizations for crystals of cyclic peptides and Crambin. J. Am. Chem. Soc. **110**, 1657–1666 (1988)
25. Abagyan, R., Totrov, M., Kuznetsov, D.: ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J. Comput. Chem. **15**, 488–506 (1994)
26. Friesner, R.A., et al.: Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem. **47**(7), 1739–1749 (2004)
27. Ding, F., Yin, S., Dokholyan, N.V.: Rapid flexible docking using a stochastic rotamer library of ligands. J. Chem. Inf. Model. **50**(9), 1623–1632 (2010)
28. Gohlke, H., Klebe, G.: Statistical potentials and scoring functions applied to protein-ligand binding. Curr. Opin. Struct. Biol. **11**(2), 231–235 (2001)
29. Dunbrack, R.L., Jr. Karplus, M.: Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J. Mol. Biol. **230**(2), 543–574 (1993)
30. Holm, L., Sander, C.: Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. Proteins. **14**(2), 213–223 (1992)
31. Brunger, A.T., Kuriyan, J., Karplus, M.: Crystallographic R factor refinement by molecular dynamics. Science. **235**(4787), 458–460 (1987)
32. Armen, R.S., Chen, J., Brooks, C.L.: An evaluation of explicit receptor flexibility in molecular docking using molecular dynamics and torsion angle molecular dynamics. J. Chem. Theory. Comput. **5**(10), 2909–2923 (2009)
33. Paulsen, J.L., Anderson, A.C.: Scoring ensembles of docked protein:ligand interactions for virtual lead optimization. J. Chem. Inf. Model. **49**(12), 2813–2819 (2009)
34. Craig, I.R., Essex, J.W., Spiegel, K.: Ensemble docking into multiple crystallographically derived protein structures: an evaluation based on the statistical analysis of enrichments. J. Chem. Inf. Model. **50**(4), 511–524 (2010)

35. Huang, S.Y., Zou, X.: Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. Proteins. **66**(2), 399–421 (2007a)
36. Rao, S., et al.: Improving database enrichment through ensemble docking. J. Comput. Aided. Mol. Des. **22**(9), 621–627 (2008)
37. Rueda, M., Bottegoni, G., Abagyan, R.: Recipes for the selection of experimental protein conformations for virtual screening. J. Chem. Inf. Model. **50**(1), 186–193 (2010)
38. Damm, K.L., Carlson, H.A.: Exploring experimental sources of multiple protein conformations in structure-based drug design. J. Am. Chem Soc. **129**(26), 8225–8235 (2007)
39. Huang, S.Y., Zou, X.: Efficient molecular docking of NMR structures: application to HIV-1 protease. Protein. Sci. **16**(1), 43–51 (2007b)
40. Hawkins, P.C., Warren, G.L., Skillman, A.G., Nicholls, A.: How to do an evaluation: pitfalls and traps. J. Comput. Aided. Mol. Des. **22**(3–4), 179–190 (2008)
41. Warren, G.L., et al.: A critical assessment of docking programs and scoring functions. J. Med. Chem. **49**(20), 5912–5931 (2006)
42. Yang, L., et al.: Identifying unexpected therapeutic targets via chemical-protein interactome. PLoS ONE. **5**(3), e9568 (2010)
43. Petrelli, A., Giordano, S.: From single- to multi-target drugs in cancer theraphy: when aspecificity becomes an advantage. Curr. Med. Chem. **15**, 422–432 (2008)
44. Wermuth, C.G.: Multitarget drugs: the end of the 'one-target-on-disease' phylosophy? Drug. Discov. Today. **9**, 826–827 (2004)
45. Kirchmair, J., Markt, P., Distinto, S., Wolber, G., Langer, T.: Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—what can we learn from earlier mistakes? J. Comput. Aided. Mol. Des. **22**(3–4), 213–228 (2008)
46. Langer T., Hoffmann RD.: Pharmacophores and Pharmacophore Searches. Wiley-VCH, Weinheim, Germany, pp. 338–343 (2006)
47. Truchon, J.F., Bayly, C.I.: Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J. Chem. Inf. Model. **47**(2), 488–508 (2007)
48. Jain, A.N., Nicholls, A.: Recommendations for evaluation of computational methods. J. Comput. Aided. Mol. Des. **22**(3–4), 133–139 (2008)
49. Nicholls, A.: What do we know and when do we know it? J. Comput. Aided. Mol. Des. **22**(3–4), 239–255 (2008)
50. Witten, I.H., Frank, E.: Credibility: Evaluating what's been learned. In: Data minings: Practical machine learning tools and techniques, 2nd ed; Morgan Kaufmann: San Francisco, CA, pp. 161–176 (2005)
51. Koehl, P., Levitt, M.: A brighter future for protein structure prediction. Nat. Struct. Biol. **6**(2), 108–111 (1999)
52. Marti-Renom, M.A., et al.: Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. **29**, 291–325 (2000)
53. Arnold, K., Bordoli, L., Kopp, J., Schwede, T.: The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics. **22**(2), 195–201 (2006)
54. Bates, P.A., Kelley, L.A., MacCallum, R.M., Sternberg, M.J.: Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins. Suppl. 5, 39–46 (2001)
55. Eswar, N., et al.: Tools for comparative protein structure modeling and analysis. Nucleic. Acids. Res. **31**(13), 3375–3380 (2003)
56. Sali, A., Blundell, T.L.: Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. **234**(3), 779–815 (1993)
57. Altschul, S.F., et al.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic. Acids. Res. **25**(17), 3389–3402 (1997)
58. Wistrand, M., Sonnhammer, E.L.: Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. BMC Bioinformatics. **6**, 99 (2005)
59. McGovern, S.L., Shoichet, B.K.: Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. J. Med. Chem. **46**(14), 2895–2907 (2003)

60. Rockey, W.M., Elcock, A.H.: Structure selection for protein kinase docking and virtual screening: homology models or crystal structures? Curr. Protein. Pept. Sci. **7**(5), 437–457 (2006)
61. Tuccinardi, T., Botta, M., Giordano, A., Martinelli, A.: Protein kinases: docking and homology modeling reliability. J. Chem. Inf. Model. **50**(8), 1432–1441 (2010)
62. Blake, J.D., Cohen, F.E.: Pairwise sequence alignment below the twilight zone. J. Mol. Biol. **307**(2), 721–735 (2001)
63. Jennings, A.J., Edge, C.M., Sternberg, M.J.: An approach to improving multiple alignments of protein sequences using predicted secondary structure. Protein. Eng. **14**(4), 227–231 (2001)
64. Sanchez, R., Sali, A.: Advances in comparative protein-structure modelling. Curr. Opin. Struct. Biol. **7**(2), 206–214 (1997)
65. Shi, J., Blundell, T.L., Mizuguchi, K.: FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol. **310**(1), 243–257 (2001)
66. Marti-Renom, M.A., Madhusudhan, M.S., Sali, A.: Alignment of protein sequences by their profiles. Protein. Sci. **13**(4), 1071–1087 (2003)
67. von Ohsen, N., Sommer, I., Zimmer, R.: Profile-profile alignment: a powerful tool for protein structure prediction. Pac. Symp. Biocomput. 252–263 (2003)
68. Jaroszewski, L., Rychlewski, L., Godzik, A.: Improving the quality of twilight-zone alignments. Protein Sci. **9**(8), 1487–1496 (2000)
69. Sauder, J.M., Arthur, J.W., Dunbrack, R.L., Jr: Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins. **40**(1), 6–22 (2000)
70. Marti-Renom, M.A., Madhusudhan, M.S., Fiser, A., Rost, B., Sali, A.: Reliability of assessment of protein structure prediction methods. Structure. **10**(3), 435–440 (2002)
71. Eswar, N., Sali, A.: (2007) Comparative modeling of drug target proteins. In: Taylor J., Triggle D., Mason J.S., (eds.) Computer-Assisted Drug Design, Comprehensive Medicinal Chemistry II, vol. 4, pp. 215–236. Elsevier, Oxford, UK
72. Sanchez, R., et al.: Protein structure modeling for structural genomics. Nat. Struct. Biol. 7 Suppl. 986–990 (2000)
73. Eramian, D., Eswar, N., Shen, M.Y., Sali, A.: How well can the accuracy of comparative protein structure models be predicted? Protein. Sci. **17**(11), 1881–1893 (2008)
74. Baker, D., Sali, A.: Protein structure prediction and structural genomics. Science. **294**(5540), 93–96 (2001)
75. Cavasotto, C.N., Phatak, S.S.: Homology modeling in drug discovery: current trends and applications. Drug. Discov. Today. **14**(13–14), 676–683 (2009)
76. Diller, D.J., Li, R.: Kinases, homology models, and high throughput docking. J. Med. Chem. **46**(22), 4638–4647 (2003)
77. Oshiro, C., et al.: Performance of 3D-database molecular docking studies into homology models. J. Med. Chem. **47**(3), 764–767 (2004)
78. Kairys, V., Fernandes, M.X., Gilson, M.K.: Screening drug-like compounds by docking to homology models: a systematic study. J. Chem. Inf. Model. **46**(1), 365–379 (2006)
79. Ferrara, P., Jacoby, E.: Evaluation of the utility of homology models in high throughput docking. J. Mol. Model. **13**(8), 897–905 (2007)
80. Fan, H., et al.: Molecular docking screens using comparative models of proteins. J. Chem. Inf. Model. **49**(11), 2512–2527 (2009)
81. Barril, X., Morley, S.D.: Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. J. Med. Chem. **48**(13), 4432–4443 (2005)
82. Birch, L., Murray, C.W., Hartshorn, M.J., Tickle, I.J., Verdonk, M.L.: Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. J. Comput. Aided. Mol. Des. **16**(12), 855–869 (2002)
83. Hillisch, A., Pineda, L.F., Hilgenfeld, R.: Utility of homology models in the drug discovery process. Drug. Discov. Today. **9**(15), 659–669 (2004)
84. Wishart, D.S., et al.: DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic. Acids. Res. **36**(Database issue), D901–906 (2008)

85. Chothia, C., Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. EMBO J. **5**(4), 823–826 (1986)
86. O'Donovan, C., Apweiler, R., Bairoch, A.: The human proteomics initiative (HPI). Trends. Biotechnol. **19**(5), 178–181 (2001)
87. Park, S.J., Kufareva, I., Abagyan, R.: Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles. J. Comput. Aided. Mol. Des. **24**(5), 459–471 (2010)

Homology Modeling: Generating Structural Models to Understand Protein Function and Mechanism

Srinivas Ramachandran and Nikolay V. Dokholyan

1 Homology Models: Need and Applicability

Geneticists and molecular and cell biologists routinely uncover new proteins important in specific biological processes/pathways. However, either the molecular functions or the functional mechanisms of many of these proteins are unclear due to a lack of knowledge of their atomic structures. Yet, determining experimental structures of many proteins presents technical challenges. The current methods for obtaining atomic-resolution structures of biomolecules (X-ray crystallography and NMR spectroscopy) require pure preparations of proteins at concentrations much higher than those at which the proteins exist in a physiological environment. Additionally, NMR has size limitations, with current technology limited to the determination of structures of proteins with masses of up to 15 kDa. Due to these reasons, atomic structures of many medically and biologically important proteins do not exist. However, the structures of these proteins are essential for several purposes, including *in silico* drug design [1], understanding the effects of disease mutations [2], and designing experiments to probe the functional mechanisms of proteins. Comparative modeling has gained importance as a tool for bridging the gap between sequence and structure space, allowing researchers to build structural models of proteins that are difficult to crystallize or for which structure determination by

S. Ramachandran • N.V. Dokholyan (✉)
Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599, USA

Molecular and Cellular Biophysics Program, University of North Carolina, Chapel Hill, NC 27599, USA
e-mail: ramachan@email.unc.edu

N.V. Dokholyan
Center for Computational and Systems Biology, University of North Carolina, Chapel Hill, NC 27599, USA
e-mail: dokh@unc.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_5, © Springer Science+Business Media, LLC 2012

NMR spectroscopy is not tractable. Comparative modeling, or homology modeling, exploits the fact that two proteins whose sequences are evolutionarily connected display similar structural features [3]. Thus, the known structure of a protein (template) can be used to generate a molecular model of the protein (query) whose experimental structure is not known.

The applicability of comparative modeling in structural biology has been validated by the observations of several groups, e.g., that a limited number of protein folds are observed in nature [4, 5] and that nature is able to reuse similar folds for diverse protein functions [6]. Thus, several researchers have used the already available breadth of structural information to build structural models of many proteins whose experimental structures have not been determined. For example, ModBase [7] and SWISS-MODEL [8], repositories of comparative models generated using automated protocols, have structural models for 3.4 million and 2.2 million unique sequences respectively; for comparison, the repository for experimental structures, protein data bank (PDB [9]) has 67,728 experimental structures. The burgeoning number of structural models in repositories such as ModBase and SWISS-MODEL reflects the usefulness of comparative modeling in significantly closing the gap between the number of known sequences and known structures. To further close this gap, the protein structure initiative [10] aims to determine the experimental structures for representative members of protein families that do not yet have any structural templates in the PDB.

Structural models generated by homology modeling can be of direct medical and biological relevance. Structural models can be used to predict the effects of single nucleotide polymorphisms uncovered from genome-wide association studies, helping to delineate the molecular etiology of genetically transmitted diseases [2]. Homology-based structural models have already been used widely in *in silico* drug screening [11–13]. For biological experiments, structural models can be used to design mutations that lead to specific changes in the function or stability of the modeled protein [14, 15]. Importantly, homology models can be used as starting models for molecular replacement in X-ray crystallography [16], leading to better experimental structures. Furthermore, these structural models can be used in conjunction with methods such as FRET that provide interresidue distances and for mapping residue-level experimental data, such as accessibility measured through EPR [17] and H-D exchange mass spectrometry [18, 19].

Thus, to better understand the function and mechanism of a given protein of unknown structure, researchers can generate structural models using comparative modeling. In this chapter, we discuss the process of generating a homology-based structural model of a protein of interest. In particular, we focus on the critical controls and tests to be used at each step of model building to ensure that the final model is physically and biologically reasonable and, most importantly, to determine the extent to which the given model can be used in interpretations of experimental data. Comparative modeling involves several steps, such as identification of the template, sequence threading, processing insertions and deletions, model optimization, quality control, and finally, model interpretation (Fig. 1, Table 1). We discuss each of these steps in the following sections.

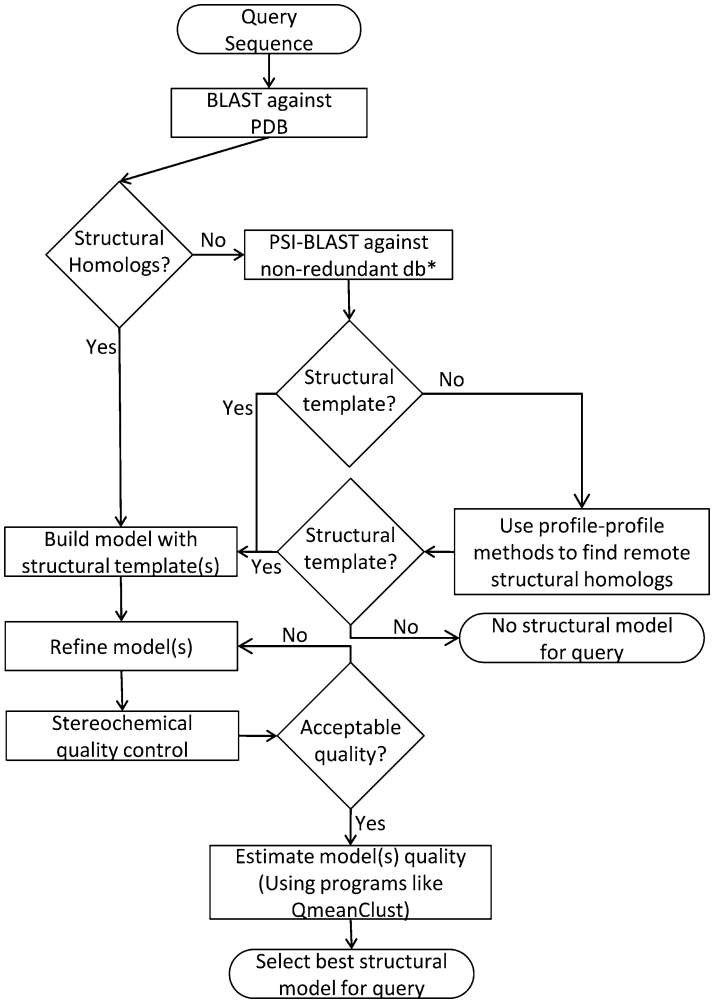

Fig. 1 Flowchart of the steps followed in the construction of a comparative structural model (*database)

2 Template Identification

2.1 Domain Delineation

One of the first steps to be performed with the query sequence is to determine the number of domains in the sequence. In many multidomain proteins, a single structural template covering the whole sequence may not be available. Instead, templates for each of the domains may be available. Many programs that employ

Table 1 Some representative methods for the different steps involved in the construction of a comparative structural model

Procedure	Server
Identify homologous sequences	BLAST [22], PSI-BLAST [23]
Protein family classifications	Pfam [21], InterPro [20]
Profile-based	HMMER [25], HHSearch [27], SAM [26]
Threading + profile-based	FUGUE (based on structure profile created by HOMSTRAD) [32], PROSPECT [33], SPARKS2 [34] and SP3 [35]
Profile-based + secondary structure prediction	PPA [76]
Meta servers	TASSER [77], I-TASSER [40], Bioinfobank [39]
Stereochemical quality control	Gaia [62], WHAT IF [61], PROCHECK [65], MolProbity [64]
Estimating model quality	Qmean [55], QmeanClust [60]

machine learning approaches can be used to delineate domain boundaries in a protein sequence and even identify the potential function of the identified domains. InterProScan [20] and Pfam [21] are two databases available online that one can use to find the domains present in the query sequence. For some multidomain query sequences, one may be able to find structural templates with similar domain architecture, which will be the ideal scenario, but in others one may have to model individual domains separately and look for experimental constraints to model domain–domain orientations.

2.2 Direct Sequence Homology: BLAST and PSI-BLAST

BLAST (basic alignment and search tool) [22] is a powerful and efficient tool to discover the evolutionary connections of a given protein sequence. Given a protein sequence of interest, any current researcher will first and foremost employ BLAST to search for homologs in all available sequence databases to uncover the functional and evolutionary details of the protein sequence. In the context of comparative modeling, BLAST helps in the identification of the structural template on which to base the structural model for a given sequence. While using the protein BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins), one can specify the sequence databases that should be searched; for comparative modeling, one usually chooses the PDB. In the context of BLAST, the PDB sequence database contains all the sequences that have an associated experimental structure. The match between BLAST "hits" and a given sequence are described by three parameters: similarity, coverage, and expect value (E-value). All three parameters are important in selecting the best template for a given sequence. A minimum of 30% similarity between query and template is essential for unambiguous alignment that can be used for generating a homology model. For each domain, at least 70% sequence coverage

is required. The extent of coverage determines the number of residues that need to be modeled without prior knowledge of their backbone coordinates. There are exceptions for the lower bounds of both similarity and query coverage, which will be discussed under remote homology, but if one were to choose a template based on BLAST results alone, the lower bounds for similarity and coverage are to be followed strictly to obtain unambiguous structural models. The E-value provides the statistical significance of a "hit" and describes the number of hits that can be obtained by chance in a given database with a given score. Thus, lower the E-value, greater the significance of a given hit. Generally, E-values less than 0.01 are considered significant for generating homology models.

If no homologs in the PDB are detected using BLAST for a given sequence, the alternative strategy is to use position-specific iterated-BLAST (PSI-BLAST) [23]. PSI-BLAST constructs a position-specific scoring matrix (PSSM) using the multiple sequence alignment of BLAST hits detected above a certain threshold (based on E-value). The PSSM is then used for searching the database. The construction of the PSSM and the subsequent database search are performed iteratively for several rounds until no new sequences are found. By using information from all the BLAST hits of a given iteration, PSI-BLAST helps uncover distant homologs. In determining the optimal template using PSI-BLAST, one uses the same thresholds for sequence coverage, similarity, and E-value that were discussed for BLAST.

Once a suitable template is identified, it is worthwhile to closely analyze the sequence alignment between the query and template. Analysis from several rounds of critical assessment of structure prediction (CASP) [24] has shown that the sequence alignment between query and template is the most important step in comparative modeling. The most prominent inaccuracies in homology models arise from inaccurate sequence alignment rather than errors in subsequent steps of structure building. Significantly, BLAST scoring matrices and PSSMs may not incorporate subtle structural details pertinent to the given protein like the positioning of structurally important cysteine disulphide bridges, proline residues, residues important in protein function, etc. In cases where the positioning of these residues is known to be important based on experimental data, one should manually edit the alignment to ensure that these residue positions are preserved between the query and template. Thus, one should consider all available functional, biochemical, and structural data of all possible residues in the query sequence while scrutinizing and updating the sequence alignment between the query and the template.

2.3 Remote Homology

If a template is not detectable with BLAST or PSI-BLAST, one needs to use programs that are capable of identifying distant evolutionary relationships. It has been shown that two proteins can share a high degree of structural similarity in spite of the lack of detectable sequence similarity [6]. The lack of sequence similarity in these cases highlights high divergence of the sequences and also

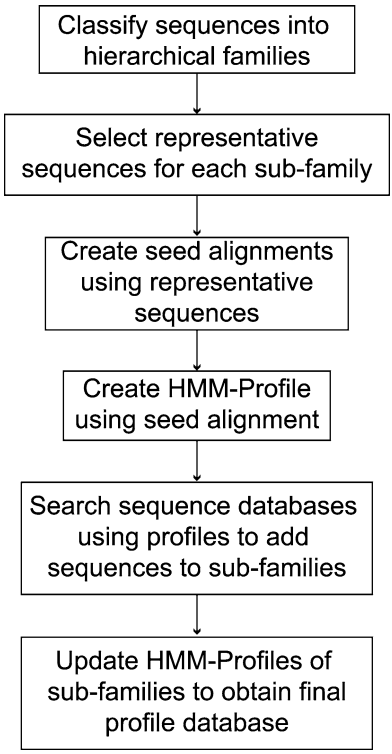

Fig. 2 Construction of profile databases. The scheme illustrates the steps involved in constructing profile databases based on sequence alone

the weakness of our current metric, namely sequence similarity, in identifying distantly related sequences. To account for the observations of distant relationships between protein sequences and to utilize these relationships in protein structure and function prediction, many programs and servers have been developed that detect remote homologs. Even though most of these servers have easy to use interfaces that do not require any knowledge of the underlying computation, in order to discriminate between different identified templates (either by a single program or multiple programs), one needs to have a clear understanding of the algorithms and guiding principles used in these programs. Hence, we give a brief overview of the underlying principles of two important classes of bioinformatics approaches that are used in the detection of remote homologs: sequence-profile-based methods and structural-profile-based methods. Many subsequent approaches have combined the sequence-profile and structural-profile-based methods to increase the robustness of identifying and aligning distantly related proteins.

Using sequence-profiles (Fig. 2) for discovering remote homologs has been achieved using techniques such as hidden markov models (HMM), neural networks, and support vector machines. By far, HMM-based techniques have been used most frequently in direct template detection [25–27]. Other methods have been used in the prediction of subcellular localization [28], secondary structure prediction [29, 30], residue environment prediction [29], and identification of transmembrane segments

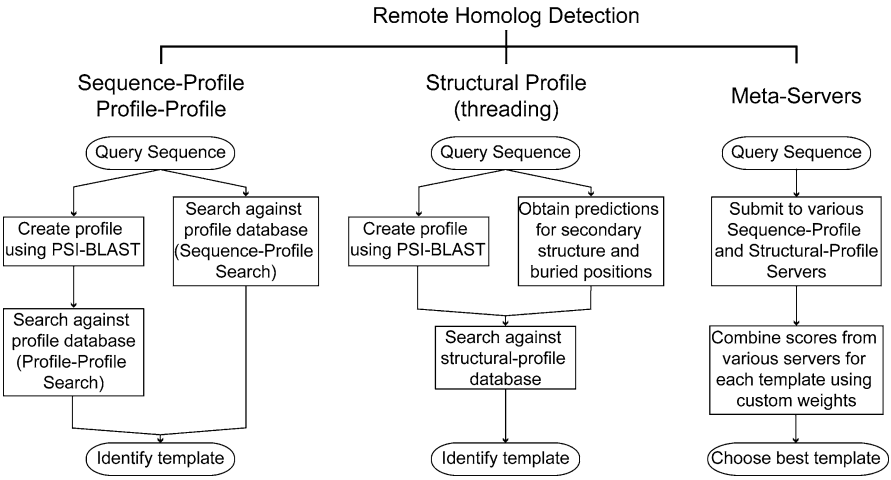

Fig. 3 Remote homolog detection. An outline of different strategies involved in detecting remote homologs: sequence-based and structure-based methods and by using meta-servers. The steps involved in each of these strategies are also outlined as a flowchart

in protein sequence [31]. HMM-based methods rely on constructing an HMM-profile for any given sequence, based on a seed alignment generated either using BLAST or manually. Most of these methods have thousands of such profiles for all known sequences. Using these profiles, for any query sequence, sequence-profile and profile–profile matching can be performed to identify significant structural homologs. All known domains are usually arranged in hierarchical families based on either function or fold to enable quick retrieval of matches. A given sequence is searched against HMM-profiles of families that have at least one representative structure, a process called sequence-profile alignment (Fig. 3). A logical expansion of sequence-profile alignment is profile–profile alignment, where a profile is constructed based on evolutionary conservation of the query sequence. The seed alignment for constructing the profile for the query sequence is usually obtained using PSI-BLAST. Once an HMM-profile is generated for the query sequence using PSI-BLAST-based multiple sequence alignment, this profile is searched against other profiles that have at least one representative structure. Apart from providing a template structure for constructing a homology model, these profile–profile and sequence-profile alignments provide a quick means to predict domain boundaries and possible function of the sequence. For example, scanning the query sequence using Pfam [21] (a database of HMM-profile based domain families) will identify the different domains in the sequence as well as possible functional and structural information of the identified domains.

Structure-based threading [32–35] forms the basis of the second group of protocols (Fig. 4). We can observe high diversity in the specific protocol followed by each structure-based threading program to identify remote homologs. Since each

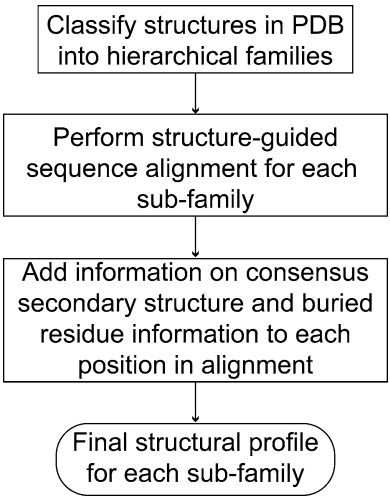

Fig. 4 Construction of structural-profile databases. The scheme illustrates the steps involved in constructing structural-profile databases starting from the structures in PDB

threading program has its own optimized, intricate protocol and scoring system to identify structural templates, we only discuss the general principles underlying these programs rather than the scoring functions of specific programs. There are two groups of data available to the threading programs to generate an optimal alignment: data on the query sequence and data on all possible structural templates. Data on the query side consist of (Fig. 3) (1) the sequence-profile of the query sequence generated either using PSI-BLAST alone or PSI-BLAST and HMM programs and (2) the secondary structure propensity of each position of the query sequence, which can be determined using neural network or HMM-based programs such as PSIPRED [29] or Jpred [30]. Data on the template side are significantly richer. First, all known structures can be grouped into structural families based on structural similarity and a sequence alignment can be performed for sequences in each of these structural families. The sequence alignment, which is primarily based on the structural alignment, gives rise to residue propensities in each position of the fold, which we can denote as the structure-profile (Fig. 4). Second, one can obtain the secondary structure at each position of the fold using the dictionary of protein secondary structure(DSSP) program [36]. Third, one can obtain the environment of each position of the fold—whether it is buried or exposed, whether the backbone or side-chain are involved in any hydrogen bonds (Fig. 4). Fourth, distance or cut-off based residue–residue contact probability can be obtained in each structural family. These four pieces of information are used in a combinatorial fashion by different programs to match the two pieces of information available for the query sequence. Thus, each program uses a combination of terms that are optimally weighted to arrive at a final score that reflects the goodness of fit between a query sequence and a template structure (or a structural family, depending on the program). One way to align structure to sequence can be to match the structure-profile of the template (amino acid propensities in each position of the fold) to the

sequence-profile of the query (amino acid propensities at each position of the sequence based on evolution) using dynamic programming, with the gap-penalties at each template position set by the secondary structure at that position. It has been observed that insertions and deletions (which arise due to gaps in alignments) are minimal in positions that feature helix or strand [37], hence gap penalties can be set higher at positions in the template featuring helix or strand. Similarly, predictions for buried and exposed positions of the query sequence have also been used in sequence-structure alignments. To account for changes in fold topology, many programs also incorporate segmental threading [38] to arrive at a discontinuous sequence-structure alignment.

2.4 Meta Servers

In the previous sections, we have introduced all the commonly used techniques for discovering structural templates. By far, the most successful approach in identifying a structural template for a given query sequence and in determining the best possible alignment between the sequence and structure has been to combine predictions from several diverse approaches. Servers such as 3D-Jury [39] and I-TASSER [40] have developed combined scoring functions that rate each structural template based on its scores in several profile–profile and structure-profile alignment programs to yield a consensus alignment (Fig. 3). For a researcher who is well versed in the biochemical and structural data connected to the query sequence, it is also possible to apply the available biochemical data as additional constraints in refining the consensus alignment between query and template.

3 From Alignment to a Structural Model

3.1 Model Construction

Once a statistically significant alignment is obtained between query and template, construction of the homology model entails converting the template structure into a structural model for the query sequence. Model construction involves two important steps: first, in the regions of the template where alignment with the query exists, the sequence of the template has to be modified to the corresponding sequence of the query. Second, in the regions where alignment do not exist (insertions and deletions), either portions of the structure must be removed (deletion) or new structural fragments need to be built *de novo* (insertion). The first step requires only the modification of the side-chain atoms of the aligned positions as the amino acids in the template structure are morphed into the amino acids corresponding to the query sequence. With knowledge of the coordinates of the protein backbone,

positioning of a new side chain is straightforward [41]. Processing insertions is the most complicated step in model construction, since the positions of backbone atoms are not known for the inserted residues and must be modeled *de novo*. Insertions mostly involve internal loops that are longer in the query compared to the structural template. The methods to build these loops include ModLoop [42] (part of MODELLER [43]) which relies on satisfaction of spatial restraints, Hierarchical Loop Prediction with Surrounding Side chain optimization [44], kinematic closure protocol in Rosetta [45] and constrained all-atom DMD simulations [46]. Processing deletions involves the removal of residues in the template structure whose positions correspond to gaps in the query sequence in the sequence alignment. When a set of residues are removed, the ends of the deletion need to be connected by a peptide bond to ensure continuity of the protein chain. Several programs like all-atom DMD, MD, and Rosetta can be used to create the peptide bond between the ends of the deleted segment, with minimum perturbation to the backbone of the rest of the structure. Once the side chains are modified and the insertions and deletions are processed, one arrives at a complete (albeit initial) structural model for the query sequence. At this stage, if there are several templates that were identified, one can use the steps outlined above to construct several complete structural models for the query based on each of the template structures. In the case of several structural templates, one has to then choose among the many models for the same sequence, the structure that best represents the real structure corresponding to the query sequence, a process that we discuss in the section dealing with model quality. An illustration of the sequence/structure alignment and homology-based structural model is shown in Fig. 5.

3.2 Model Refinement

Once a complete structural model for the query sequence is obtained, there are several possible steps by which the structural model can be refined to approach a physically accurate structure. Just modifying the side-chains and processing insertions and deletions as described above result in a model that has minimal changes from the structural template. However, with limited sequence identity (in many cases, only 30%), the structures of the template and the query will be expected to have significant conformational differences even though they share the same fold. For example, when transitioning from the template to the query sequence, many small to large amino acid changes in the core will require backbone perturbations to accommodate the large amino acid while retaining optimal packing in the core. Furthermore, homology models have been shown to have an excess of steric clashes and structural artifacts caused by unphysical overlap of newly positioned side-chain atoms with other side-chain and backbone atoms (Fig. 6a). Thus, most residues in the core will need to undergo concerted changes in the side-chain rotamers (changes in the χ angles), along with subtle changes in the protein backbone, to form a core that is optimally packed. Even though side-chain

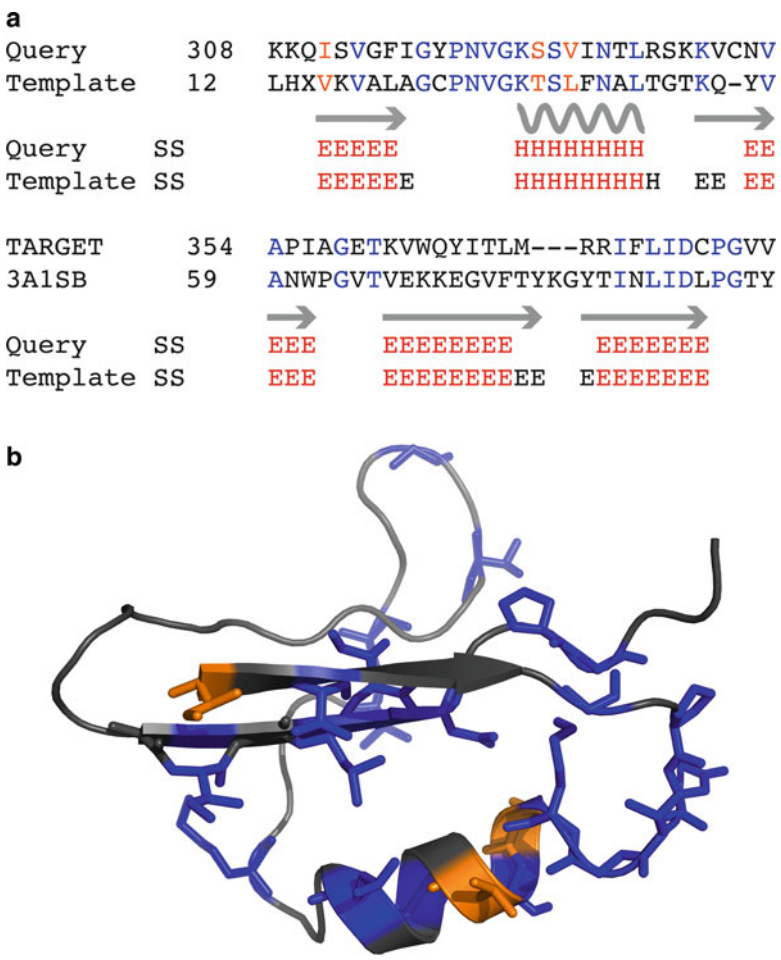

Fig. 5 A homology-based structural model. The alignment (**a**) and the structural model (**b**) of the G-domain of nucleolar GTP-binding protein 2 (Uniprot ID Q13823, query) are shown as a representative homology-based structural model. The template chosen by HHSEARCH [27], as listed in the SWISS-MODEL database [8] has the PDB ID 3A1S. The residues identical between the query and template are colored *blue*. The similar residues are colored *orange*. The predicted secondary structure of the query and the observed secondary structure of the template are also shown; *H* denotes helix and *E* denotes strand. Note the high level of similarity in the predicted and observed secondary structure. The structural model retrieved from SWISS-MODEL database is rendered as a cartoon using PyMol (http://www.pymol.org), with the identical and similar residues rendered as sticks. The positions identical in the alignment are colored *blue* in the structure and the similar residues are colored *orange*

repacking can be performed with great accuracy and efficiency by many programs which fix the backbone position (using knowledge-based rotamer libraries), in this case, the core refinement is useful only when the repacking is coupled with subtle changes in backbone conformations.

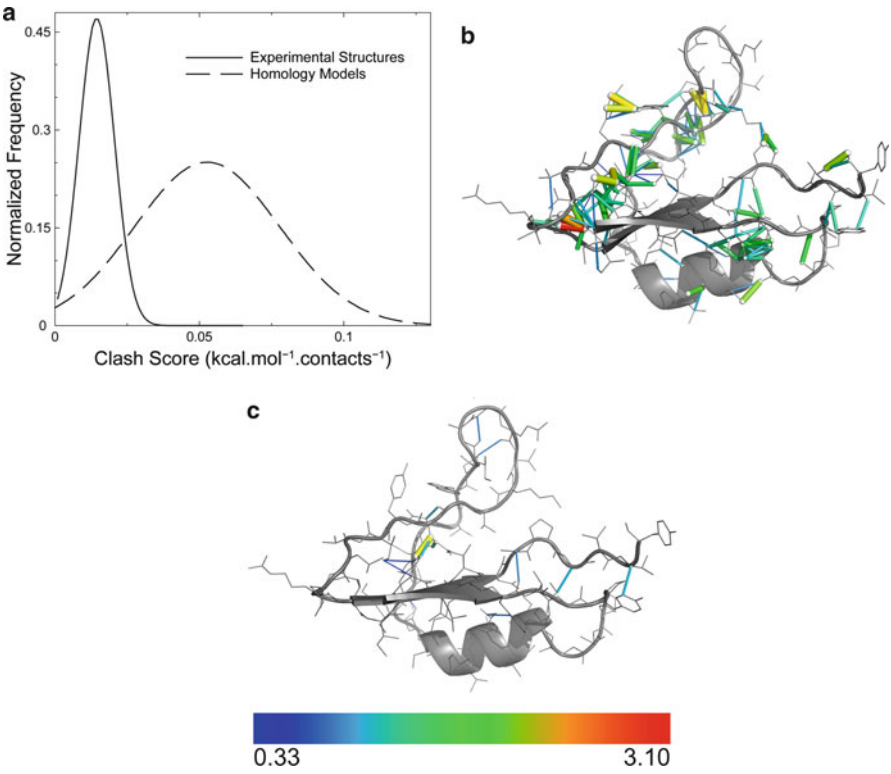

Fig. 6 Steric clashes in homology models. Homology models on an average feature much higher extent of steric clashes when compared to experimental structures (**a**). The distribution of clash-scores (which is a normalized energetic parameter reflecting the extent of steric clashes in a protein structure [50]) of high-resolution crystal structures and representative homology models from SWISS-MODEL database are plotted. The efficacy of Chiron in minimizing clashes in protein structure is demonstrated for the homology model of Q13823, whose initial model (**b**) has a clash-ratio of 0.13, much higher than that seen in experimental structures. The protein structure is shown with the cartoon representation, rendered using PyMol (http://www.pymol.org). Clashes are denoted as *colored cylinders*, where both the colors and the thickness of the cylinders denote the van der Waals repulsion energy. The scale of the repulsion energy is shown as a gradient bar at the *bottom*, with the *numbers* at the ends indicating repulsion energy in kcal/mol. Note the large numbers of cylinders in the initial model, denoting excessive steric clashes. The minimized structure (**c**) has a clash-ratio of 0.018, within one standard deviation of the mean clash-score of high-resolution structures

The refined structural models by definition should feature physically reasonable backbone conformation and a well-packed core that has an acceptable extent of clashes. Minimal backbone perturbation to ensure ideal packing can be achieved by various means including “backrub” and knowledge-based backbone assembly (as used in Rosetta), all-atom DMD simulations and minimization using molecular mechanics forcefields. All these methods refine the structural model to the nearest local minima in the conformational space of the starting structure. Thus, if the

starting model is far away from the actual structure (>5Å root mean square deviation (RMSD)), these methods will be of limited utility in bringing the model closer to the actual structure. Steepest descent/conjugate gradient minimization using all-atom molecular mechanics force fields is the most widely used method to refine a structure while also resolving clashes. However, minimization using molecular mechanics may not resolve severe clashes in some cases, hampering subsequent molecular dynamics simulations. Use of molecular modeling tools such as Rosetta [47] is the alternate avenue for refining structures with severe clashes. These tools use knowledge-based potentials and small backbone moves to resolve clashes. However, these methods work best with smaller proteins (less than 250 residues in size). Tools such as MMTSB [48] and PULCHRA [49] have emerged for structure refinement, which includes removal of clashes during refinement. Chiron [50], an automated server evaluates the extent of clashes in a given structure and if required, minimizes these clashes to the levels seen in high-resolution X-ray structures. Chiron uses all-atom DMD [46, 51] with soft-core potentials. Additionally, Chiron uses a high coefficient of heat exchange of protein atoms with thermostat to ensure minimal perturbation of the protein backbone while resolving clashes in the protein. An example for clash minimization in a homology structural model using Chiron is shown in Fig. 6b, c. While Rosetta couples backbone moves to side-chain repacking, the other methods can also be used iteratively with side-chain repacking programs to achieve rigorous refinement. Side-chain repacking coupled with minimal backbone optimization to improve core packing transitions a complete structural model into a physically realistic model that can be used for further studies.

3.3 Estimating Model Quality

The model quality can be classified into two types: (1) the stereochemical quality of the structural model and (2) the accuracy of the homology-based structural model with respect to its experimental structure. Given the lack of experimental structure for the query sequence, the real accuracy of a homology-based structural model cannot be assessed. To develop methods to predict this accuracy in the absence of known experimental structure, methods have been developed based on benchmark sets of structural models built for proteins with already existing experimental structures. In such cases where the experimental structure is known, there are several measures that estimate the model's quality. RMSD is the widely used measure to estimate the "structural similarity" between any two structures. However, large differences in positions of a small fraction of the proteins being compared can result in high values of RMSD, thus not reflecting the majority of the regions where the structures are highly similar. Thus, CASP competitions use another measure called global distance test (GDT), which is a measure of similarity of two structures with similar amino acid compositions but different tertiary structures. GDT is defined as the largest number of corresponding amino acids' alpha carbon atoms in the

compared structures that fall within a given distance cut-off. Usually, an average of GDT at 1, 2, 4, and 8 Å is used to measure structure quality, and is denoted as GDT_TS. Other variations include corrections to eliminate size-dependence [52] and also to include a negative term for incorrectly positioned amino acids (resulting in non-native contacts) [53]. There are various servers that have been developed to predict GDT_TS of a structural model in the absence of an experimental structure with which to compare.

These quality assessment programs either assess a structural model by itself (single model) [54–56], or in the context of a large reference set of structural models generated for a given sequence using different methods (consensus) [57–59]. The single model methods can compare various structural parameters such as secondary structure and solvent accessibility (whether a given residue is buried or exposed) that are predicted for a given sequence with the corresponding structural parameters of the given model. The scoring functions used in single-model quality assessment programs also include knowledge-based potentials like a sequence-dependent torsion term (usually in the context of three consecutive residues), distance and cut-off based residue–residue interaction potentials and all-atom interaction potentials. Consensus quality assessment programs rely on the idea that if a diverse set of methods were used in generating many structural models for a given query, models that incorporate the best of all methods will be the ones closest to the experimental structure. How do the consensus methods select models based on these criteria? Using a large reference set of models for a given sequence (obtained from the various structure prediction servers), they determine the average distance of a given model to all other models. The distance measure used in most cases is GDT_TS, although specific servers apply various modifications to the distance to obtain better predictive power. Through several rounds of CASP, it is apparent that the models with least average distance to the rest of the reference set feature the best GDT_TS when compared to the experimental structure. We have to emphasize here that there are many mathematical formulations used in modifying the simple average distance to obtain better predictions, but these formulations may not always have a strong physical basis. Interestingly, weighting the average distances with single-model score yields very good prediction of GDT_TS of a model with respect to the experimental structure [60]. Thus, based on the experience gained from CASP competitions, the ideal strategy for constructing a homology-based structural model would entail generating several models using heterogeneous methods, which can also include human intervention during model building and the incorporation of known experimental constraints. Once a handful of models are obtained, one can use the quality assessment programs to obtain a prediction of how close the best model will be to the experimental structure. The quality score will in turn determine the use to which a structural model can be put (discussed below).

Apart from model accuracy, an important criterion for model quality is the stereochemical quality of the given model. The stereochemical quality here broadly defines the acceptable quality of the covalent geometry and the core-packing of a given structural model. The covalent geometry of a structural model is assessed by comparing all its bond lengths, bond-angles, and torsions to standard values.

The standard values of bond lengths and angles are obtained from studies on model small molecules [61] or through surveys of high-resolution crystal structures [62]. The backbone torsions comprise of the φ, ψ, and ω angles. The $\varphi - \psi$ of each residue can be compared to the allowed $\varphi - \psi$ map obtained from survey of high-resolution crystal structures to detect outliers. The side-chain torsions are again compared to the rotamer libraries [63] to detect outliers. All these measures ensure that the covalent geometry of the structural model is physically acceptable. The packing quality can be assessed based on the extent of steric clashes [50], the prevalence of voids, and the scaling of the solvent accessible surface area with protein length [62]. There are several servers that can compare these structural parameters of a model with benchmark distribution to indicate the areas of the protein structure that need further refinement to be physically acceptable [61, 62, 64, 65]. Importantly, the stereochemical quality of the structural model is essential for further studies including molecular simulations.

4 Experimental Constraints to Improve/Verify Homology Models

Any experimental data that can be used as a structural parameter, even indirectly, aid in building a better structural model based on homology [66]. Once a structural model is available, further experiments can be designed using insights from the model. Thus, designing experiments using structural models and building models that satisfy experimental constraints become an iterative process leading to better understanding of a protein's structure–function relationships. The experimental constraints that can be used in model building are diverse and we discuss several examples here. Usually, experimental constraints are sparse and by themselves not enough to lead to an unambiguous structural model. Thus, several models can satisfy a given set of constraints. However, the subset of models that do not satisfy a given experimental constraint can be eliminated from consideration. The experimental constraints can either be at the residue level or provide overall structural information. Some of the residue-level constraints include distance bounds between specific residues obtained by FRET and site-directed cross-linking. Iterative model building is possible using FRET and site-directed cross-linking, since a structural model allows probing a much smaller subset of residue-pairs for distance measurements as opposed to residue-pairs being chosen randomly [2, 67]. Furthermore, these distance measurements provide direct validation of a given structural model. Residue accessibilities obtained by EPR spectroscopy [17] and H-D exchange mass spectrometry [18, 19] also aid in model refinement. Experiments that provide information on the overall protein structure include small angle X-ray scattering (SAXS) [68], cryo-electron microscopy (CryoEM) [69], and circular dichroism (CD) spectroscopy, among others. SAXS provides the molecular envelope or the overall shape of the protein in solution, which can

help in discriminating between structurally diverse templates used in generating the structural model. The utility of CryoEM in atomic-level structural modeling depends greatly on the resolution of the electron density map obtained for a given protein. Recent advances in CryoEM have led to subnanometer resolutions density maps that can be used to directly refine all-atom structural models [70]. CryoEM densities are usually deposited at the electron microscopy data bank (http://www.emdatabank.org/), and programs to perform flexible docking of a structural model to EM densities have been developed [71]. Thus, high-resolution cryoEM currently offers the best alternative to X-ray crystallography and NMR for obtaining accurate atomic structure of a given protein. CD spectroscopy is used to determine the secondary structure content of a given protein and CD measurements can be used to assess the overall accuracy of secondary structure content of the structural model. Indirect structural constraints include mutational studies of the protein that assess changes in function and stability. These constraints can be included in model building only qualitatively, but still provide means to eliminate inaccurate models.

5 Conclusions

Comparative modeling of protein structures offers an efficient alternative to experimental structure determination in cases where there are difficulties in obtaining experimental structures for a given protein. Usually, if one can find a structural template more than 50% identical to the query sequence, a model with an estimated RMSD of 1 Å to the experimental structure can be obtained [72]. Thus, in cases where significant homology to a structural template exists, comparative modeling is a powerful technique to better understand the structure–function relationships and functional mechanisms of a given protein. Importantly, for clinically relevant proteins that are hard to crystallize, like G-protein coupled receptors (GPCRs) and ion channels, landmark structural studies have provided a sufficient number of templates to model many variants. The structural models of these variants have been instrumental in furthering our knowledge of different functional mechanisms (in K^+ channels [73,74]) and in virtual-ligand screening (GPCRs [75]). Advances in structural understanding of GPCRs and ion channels represent the most prominent impact of comparative structural models. These models have been used in numerous other cases to yield biologically useful insights [1]. During the process of model building, we need to undertake several precautions and assess model quality at each step. Most importantly, all structural models need some form of experimental validation to gain relevance. Thus, an iterative cycle of model building and experimental verification provides the best scenario for furthering our understanding of structural and functional aspects of many biological proteins, whose experimental structures remain unsolved.

References

1. Cavasotto, C.N., Phatak, S.S.: Homology modeling in drug discovery: current trends and applications. Drug Discov. Today **14**, 676–683 (2009)
2. Serohijos, A.W., Hegedus, T., Aleksandrov, A.A., He, L., Cui, L., Dokholyan, N.V., Riordan, J.R.: Phenylalanine-508 mediates a cytoplasmic-membrane domain contact in the CFTR 3D structure crucial to assembly and channel function. Proc. Natl. Acad. Sci. U. S. A. **105**, 3256–3261 (2008)
3. Chothia, C., Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. EMBO J. **5**, 823–826 (1986)
4. Finkelstein, A.V., Ptitsyn, O.B.: Why do globular proteins fit the limited set of folding patterns? Prog. Biophys. Mol. Biol. **50**, 171–190 (1987)
5. Zhang, Y., Hubner, I.A., Arakaki, A.K., Shakhnovich, E., Skolnick, J.: On the origin and highly likely completeness of single-domain protein structures. Proc. Natl. Acad. Sci. U. S. A. **103**, 2605–2610 (2006)
6. Todd, A.E., Orengo, C.A., Thornton, J.M.: Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. **307**, 1113–1143 (2001)
7. Pieper, U., Webb, B.M., Barkan, D.T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E.C., Pettersen, E.F., Huang, C.C., et al.: ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res. **39**, D465–474 (2011)
8. Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L., Schwede, T.: The SWISS-MODEL Repository and associated resources. Nucleic Acids Res., **37**, D387–392 (2009)
9. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. Nucleic Acids Res. **28**, 235–242 (2000)
10. Chandonia, J.M., Brenner, S.E.: The impact of structural genomics: expectations and outcomes. Science **311**, 347–351 (2006)
11. Becker, O.M., Dhanoa, D.S., Marantz, Y., Chen, D., Shacham, S., Cheruku, S., Heifetz, A., Mohanty, P., Fichman, M., Sharadendu, A., et al.: An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. J. Med. Chem. **49**, 3116–3135 (2006)
12. Brylinski, M., Skolnick, J.: Q-Dock: low-resolution flexible ligand docking with pocket-specific threading restraints. J. Comput. Chem. **29**, 1574–1588 (2008)
13. Ekins, S., Mestres, J., Testa, B.: In silico pharmacology for drug discovery: applications to targets and beyond. Br. J. Pharmacol. **152**, 21–37 (2007)
14. Labro, A.J., Boulet, I.R., Choveau, F.S., Mayeur, E., Bruyns, T., Loussouarn, G., Raes, A.L., Snyders, D.J.: The S4-S5 linker of KCNQ1 channels forms a structural scaffold with the S6 segment controlling gate closure. J. Biol. Chem. **286**, 717–725 (2011)
15. Szklarz, G.D., Halpert, J.R.: Use of homology modeling in conjunction with site-directed mutagenesis for analysis of structure-function relationships of mammalian cytochromes P450. Life Sci. **61**, 2507–2520 (1997)
16. Claude, J.B., Suhre, K., Notredame, C., Claverie, J.M., Abergel, C.: CaspR: a web server for automated molecular replacement using homology modelling. Nucleic Acids Res. **32**, W606–W609 (2004)
17. Dong, J., Yang, G., McHaourab, H.S.: Structural basis of energy transduction in the transport cycle of MsbA. Science **308**, 1023–1028 (2005)
18. Chung, E.W., Nettleton, E.J., Morgan, C.J., Gross, M., Miranker, A., Radford, S.E., Dobson, C.M., Robinson, C.V.: Hydrogen exchange properties of proteins in native and denatured states monitored by mass spectrometry and NMR. Protein Sci. **6**, 1316–1324 (1997)
19. Engen, J.R., Smith, D.L.: Investigating protein structure and dynamics by hydrogen exchange MS. Anal. Chem. **73**, 256A–265A (2001)
20. Zdobnov, E.M., Apweiler, R.: InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17**, 847–848 (2001)

21. Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., et al.: The Pfam protein families database. Nucleic Acids Res. **38**, D211–D222 (2010)
22. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. **215**, 403–410 (1990)
23. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**, 3389–3402 (1997)
24. Jauch, R., Yeo, H.C., Kolatkar, P.R., Clarke, N.D.: Assessment of CASP7 structure predictions for template free targets. Proteins **69**(Suppl 8), 57–67 (2007)
25. Finn, R.D., Clements, J., Eddy, S.R.: HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. **39**, W29–W37 (2011)
26. Karplus, K., Barrett, C., Hughey, R.: Hidden Markov models for detecting remote protein homologies. Bioinformatics **14**, 846–856 (1998)
27. Soding, J.: Protein homology detection by HMM-HMM comparison. Bioinformatics **21**, 951–960 (2005)
28. Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G.: Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. **10**, 1–6 (1997)
29. Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. **292**, 195–202 (1999)
30. Cole, C., Barber, J.D., Barton, G.J.: The Jpred 3 secondary structure prediction server. Nucleic Acids Res. **36**, W197–W201 (2008)
31. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.: Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. **305**, 567–580 (2001)
32. Shi, J., Blundell, T.L., Mizuguchi, K.: FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol. **310**, 243–257 (2001)
33. Xu, Y., Xu, D.: Protein threading using PROSPECT: design and evaluation. Proteins **40**, 343–354 (2000)
34. Zhou, H., Zhou, Y.: Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins **55**, 1005–1013 (2004)
35. Zhou, H., Zhou, Y.: Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins **58**, 321–328 (2005)
36. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22**, 2577–2637 (1983)
37. Pascarella, S., Argos, P.: Analysis of insertions/deletions in protein structures. J. Mol. Biol. **224**, 461–471 (1992)
38. Wu, S., Zhang, Y.: Recognizing protein substructure similarity using segmental threading. Structure **18**, 858–867 (2010)
39. Ginalski, K., Elofsson, A., Fischer, D., Rychlewski, L.: 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics **19**, 1015–1018 (2003)
40. Roy, A., Kucukural, A., Zhang, Y.: I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. **5**, 725–738 (2010)
41. Dunbrack, R.L., Jr., Karplus, M.: Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J. Mol. Biol. **230**, 543–574 (1993)
42. Fiser, A., Do, R.K., Sali, A.: Modeling of loops in protein structures. Protein Sci. **9**, 1753–1773 (2000)
43. Fiser, A., Sali, A.: Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol. **374**, 461–491 (2003)
44. Sellers, B.D., Zhu, K., Zhao, S., Friesner, R.A., Jacobson, M.P.: Toward better refinement of comparative models: predicting loops in inexact environments. Proteins **72**, 959–971 (2008)

45. Mandell, D.J., Coutsias, E.A., Kortemme, T.: Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat. Methods **6**, 551–552 (2009)
46. Ding, F., Tsao, D., Nie, H., Dokholyan, N.V.: Ab initio folding of proteins with all-atom discrete molecular dynamics. Structure **16**, 1010–1018 (2008)
47. Kaufmann, K.W., Lemmon, G.H., Deluca, S.L., Sheehan, J.H., Meiler, J.: Practically useful: what the Rosetta protein modeling suite can do for you. Biochemistry **49**, 2987–2998 (2010)
48. Feig, M., Karanicolas, J., Brooks, C.L. III.: MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. J. Mol. Graph. Model **22**, 377–395 (2004)
49. Rotkiewicz, P., Skolnick, J.: Fast procedure for reconstruction of full-atom protein models from reduced representations. J. Comput. Chem. **29**, 1460–1465 (2008)
50. Ramachandran, S., Kota, P., Ding, F., Dokholyan, N.V.: Automated minimization of steric clashes in protein structures. Proteins **79**, 261–270 (2011)
51. Dokholyan, N.V., Buldyrev, S.V., Stanley, H.E., Shakhnovich, E.I.: Discrete molecular dynamics studies of the folding of a protein-like model. Fold Des. **3**, 577–587 (1998)
52. Zhang, Y., Skolnick, J.: Scoring function for automated assessment of protein structure template quality. Proteins **57**, 702–710 (2004)
53. Sadreyev, R.I., Shi, S., Baker, D., Grishin, N.V.: Structure similarity measure with penalty for close non-equivalent residues. Bioinformatics **25**, 1259–1263 (2009)
54. Eramian, D., Eswar, N., Shen, M.Y., Sali, A.: How well can the accuracy of comparative protein structure models be predicted? Protein Sci. **17**, 1881–1893 (2008)
55. Benkert, P., Tosatto, S.C., Schomburg, D.: QMEAN: a comprehensive scoring function for model quality assessment. Proteins **71**, 261–277 (2008)
56. Wallner, B., Elofsson, A.: Can correct protein models be identified? Protein Sci. **12**, 1073–1086 (2003)
57. Wallner, B., Elofsson, A.: Pcons5: combining consensus, structural evaluation and fold recognition scores. Bioinformatics **21**, 4248–4254 (2005)
58. McGuffin, L.J., Roche, D.B.: Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics **26**, 182–188 (2010)
59. Cheng, J., Wang, Z., Tegge, A.N., Eickholt, J.: Prediction of global and local quality of CASP8 models by MULTICOM series. Proteins **77**(Suppl 9), 181–184 (2009)
60. Benkert, P., Schwede, T., Tosatto, S.C.: QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. BMC Struct. Biol. **9**, 35 (2009)
61. Hooft, R.W.W., Vriend, G., Sander, C., Abola, E.E.: Errors in protein structures. Nature **381**, 272–272 (1996)
62. Kota, P., Ding, F., Ramachandran, S., Dokholyan, N.V.: Gaia: automated quality assessment of protein structure models. Bioinformatics **27**, 2209–2215 (2011)
63. Dunbrack, R.L., Jr., Cohen, F.E.: Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. **6**, 1661–1681 (1997)
64. Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B. III, Snoeyink, J., Richardson, J.S., et al.: MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucliec. Acids Res. **35**, W375–W383 (2007)
65. Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M.: PROCHECK: a program to check the stereochemical quality of protein structures. J.Appl. Crystallogr., **26**, 283–291 (1993)
66. Alber, F., Forster, F., Korkin, D., Topf, M., Sali, A.: Integrating diverse data for structure determination of macromolecular assemblies. Annu. Rev. Biochem. **77**, 443–477 (2008)
67. Tung, C.S., Wall, M.E., Gallagher, S.C., Trewhella, J.: A model of troponin-I in complex with troponin-C using hybrid experimental data: the inhibitory region is a beta-hairpin. Protein Sci. **9**, 1312–1326 (2000)

68. Schneidman-Duhovny, D., Hammel, M., Sali, A.: FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res. **38**, W540–W544 (2010)
69. Baker, M.L., Zhang, J., Ludtke, S.J., Chiu, W.: Cryo-EM of macromolecular assemblies at near-atomic resolution. Nat. Protoc. **5**, 1697–1708 (2010)
70. Cong, Y., Baker, M.L., Jakana, J., Woolford, D., Miller, E.J., Reissmann, S., Kumar, R.N., Redding-Johanson, A.M., Batth, T.S., Mukhopadhyay, A., et al.: 4.0-A resolution cryo-EM structure of the mammalian chaperonin TRiC/CCT reveals its unique subunit arrangement. Proc. Natl. Acad. Sci. U. S. A. **107**, 4967–4972 (2010)
71. Rusu, M., Birmanns, S., Wriggers, W.: Biomolecular pleiomorphism probed by spatial interpolation of coarse models. Bioinformatics **24**, 2460–2466 (2008)
72. Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., Schwede, T.: Protein structure homology modeling using SWISS-MODEL workspace. Nat. Protoc. **4**, 1–13 (2009)
73. Silva, J.R., Pan, H., Wu, D., Nekouzadeh, A., Decker, K.F., Cui, J., Baker, N.A., Sept, D., Rudy, Y.: A multiscale model linking ion-channel molecular dynamics and electrostatics to the cardiac action potential. Proc. Natl. Acad. Sci. U. S. A. **106**, 11102–11106 (2009)
74. Smith, J.A., Vanoye, C.G., George, A.L. Jr., Meiler, J., Sanders, C.R.: Structural models for the KCNQ1 voltage-gated potassium channel. Biochemistry **46**, 14141–14152 (2007)
75. Katritch, V., Rueda, M., Lam, P.C., Yeager, M., Abagyan, R.: GPCR 3D homology models for ligand screening: lessons learned from blind predictions of adenosine A2a receptor complex. Proteins **78**, 197–211 (2010)
76. Wu, S., Skolnick, J., Zhang, Y.: Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol. **5**, 17 (2007)
77. Zhou, H., Skolnick, J.: Ab initio protein structure prediction using chunk-TASSER. Biophys. J. **93**, 1510–1518 (2007)

Quantum Mechanical Insights into Biological Processes at the Electronic Level

Anastassia N. Alexandrova

Abbreviations

AO	Atomic orbital
AIMS	Ab initio multiple spawning
BOA	Born-oppenheimer approximation
BOMD	Born-oppenheimer molecular dynamics
CASPT2	Complete active space perturbation theory, second order
CASSCF	Complete active space self consistent field
CCSD	Coupled cluster with single and double excitations
CCSD(T)	Coupled cluster with single and double excitations, and triple excitations treated non-variationally
CI	Configuration interaction
CIS	Configuration interaction with single excitations included
CISD	Configuration interaction with single and double excitations included
DFT	Density functional theory
EOM	Equation of motion
FEP	Free energy perturbation
HF	Hartree-Fock
HOMO	Highest occupied molecular orbital
MD	Molecular dynamics
MC	Monte Carlo
MO	Molecular orbital
MO	LCAO, Molecular orbitals written as linear combinations of atomic orbitals

A.N. Alexandrova (✉)
Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, CA 90095-1569, USA
e-mail: ana@chem.ucla.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_6, © Springer Science+Business Media, LLC 2012

MP	Møller-Plesset (perturbation theory)
MRCC	Multireference coupled cluster
MRCI	Multireference configuration interaction
SCF	Self consistent field
PES	Potential energy surface
PT	Perturbation theory
QM/MM	Quantum mechanical molecular mechanical
RHF	Restricted Hartree-Fock
TD-DFT	Time-dependent density functional theory
TISE	Time independent Schrödinger equation
UHF	Unrestricted Hartree-Fock
ZPE	Zero point energy

1 Introduction

The realm of biology is always governed by underlying electronic effects. These effects are often treated implicitly and may go nearly unnoticed in classical biomolecular simulations, such as Monte Carlo or molecular dynamics. It is important to remember, however, that these classical methods always operate on the single, ground electronic potential energy surface (PES). Furthermore, classical methods assume the classical behavior of the atomic nuclei, and thus rely on the so-called Born–Oppenheimer approximation (BAO) heavily used in quantum mechanics, as discussed in detail below. Due to the BAO, the ground PES can be obtained by finding the optimal electronic solution for every position of stationary classical nuclei. The combined electronic and nuclear energy as a function of nuclear coordinates in the PES. The Born–Oppenheimer PES is usually very close to the chemical reality. Parameters of classical force fields are optimized to reproduce this ground PES, either calculated quantum mechanically or derived from the experiment. Thus, electronic structure is always an active player in classical simulations through the parameters of the force field in use. However, when it comes to the assessment of the mechanism of a biochemical reaction that involves breaking and forming of covalent bonds, quantum mechanics is an almost exclusive reliable approach, with a prominent classical exception being the empirical valence bond method. Furthermore, there is a large class of biological processes that simply cannot be assessed without explicit quantum mechanical treatment. An obvious example is electron transfer in enzymes or DNA that plays a pivotal role in every oxidation or reduction event in living cells. Proton or hydrogen transfer is also a process in which quantum effects are not to be ignored, because these light particles tunnel through reaction barriers, which are thereby considerably reduced. As an illustration, it is well-known that the nuclear fusion in the Sun responsible for the heat that the Sun irradiates would be kinetically impossible without H tunneling. The Sun is simply not hot enough to overcome the barrier to the $H + H \rightarrow He$ reaction classically. Electronic structure of transition metals is at the heart of the

catalytic apparatus in metallo-enzymes. Photochemical reactions, such as those in DNA undergoing photodamage, in photosynthetic machinery, or in rhodopsin responsible for our vision, involve electronic excitations. The associated photoinitiated dynamics then takes place on multiple PESs, rather than just one, and near surface crossings nuclei exhibit quantum behavior. Such processes and the quantum mechanical methods that can be used to study them are the subject of this chapter. In the wealth of electronic structure methods, only some are suitable for calculations on biologically related molecules and models, due to the high computational cost of the former and fairly large size of the latter. We address calculations on the ground and excited electronic states, mixed quantum mechanical–molecular mechanical (QM/MM) techniques, methods combining quantum mechanical and statistical mechanical treatment of biological systems, and various types of dynamics.

The sections below address established methods along with examples of biological problems that can be studied using these methods, and they are arranged in order of increasing complexity of the methodology. In each section, the theoretical foundation is discussed first, accompanied by practical suggestions for the user including assessment of accuracy and computational cost, and then a few examples of applications are presented. Some of the presented material is covered in other monographs and textbooks. However, the intention here is to cover the concentrated basics of all the essential relevant methods, so as to efficiently coach the Reader to intelligently use quantum mechanics for biological systems. Finally, the list of described techniques is not exhaustive, and problem-driven variations built upon the key methods continue to emerge.

2 *Ab Initio* Treatment of Biochemical Systems on the Ground State

2.1 Theoretical Foundation

A great insight into molecular structure, and mechanisms and energetics of chemical reactions, such as those catalyzed by biological enzymes, can be gained from ground state ab initio calculations. Using these techniques, various molecular properties can be calculated, such as total energies, geometries, energy gradients, harmonic vibrational frequencies and IR absorption intensities, electric dipole moment, electric polarizability, hyperfine coupling constants, spin-couplings, magnetic susceptibility, nuclear magnetic shielding, etc. Calculated spectroscopic properties can be compared with the experiment, and used for interpretation of the experimental data. Ab initio calculations can yield results of chemical accuracy, if used with knowledge and care. We therefore spend some time covering the basic principles behind major ab initio methods, and we welcome already informed Readers to skip the next ten pages or so. Ab initio quantum chemistry is a foundation of many more complicated techniques discussed later in the chapter.

All quantum mechanical techniques are based on the experimentally observed wave-particle duality intrinsic to all matter. The wave mechanics gave rise to the habit of describing any system in quantum mechanics by its so-called wave function, Ψ, which contains all the information about the system. Ψ itself does not have a physical meaning, but when squared, it has the meaning of the probability density. The quantity $|\Psi|^2 = \Psi \cdot \Psi^*$ multiplied by a volume element, $\mathrm{d}r$, returns the probability of finding the quantum mechanical particle in this volume. Thus, $|\Psi|^2\,\mathrm{d}r$ integrated over the entire space equals 1, as long as the wave function is normalized.

In quantum mechanics, operators are used as an algebraic form of a "measurement." When a quantum mechanical operator acts on Ψ, the result is the product of the value of an observable (physical property) corresponding to this operator and the wave function itself, if Ψ is the eigenfunction of this operator. In particular, the true Ψ of a system (which is by the way often hard to find) is an eigenfunction of the energy operator of the system, called the Hamiltonian, $\hat{H}$. The corresponding eigenproblem is the famous Schrödinger equation, central to quantum mechanics. In the time-dependent form it is written as

$$\hat{H}\Psi = i\hbar\frac{\partial}{\partial t}\Psi. \tag{1}$$

$\hat{H}$ has the kinetic and potential energy terms for all components in the system. For example, for a molecule in vacuum it has nuclear and electronic kinetic energy terms, potential energy terms for the internuclear repulsion, electron–nuclear attraction, and electron–electron repulsion:

$$\hat{H} = -\frac{1}{2}\left[\sum_a \frac{\nabla_a^2}{M_a} + \sum_i \nabla_i^2\right] + \sum_a \sum_{b\neq a} \frac{Z_a Z_b}{|\mathbf{R}_a - \mathbf{R}_b|} - \sum_a \sum_i \frac{Z_a}{|\mathbf{R}_a - \mathbf{r}_i|} + \sum_i \sum_{j\neq i} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}, \tag{2}$$

where indexes a and b enumerate nuclei, i and j enumerate electrons, $\mathbf{R}_\mathrm{a}$ are nuclear coordinates, $\mathbf{r}_i$ are electronic coordinates, and other standard physical constants are omitted because atomic units are used. In the time-independent form of the Schrödinger equation (TISE), $\hat{H}$ acting on Ψ returns the total energy of the system times the wave function:

$$\hat{H}\Psi = E\Psi. \tag{3}$$

Mathematically, the wave function that satisfies the TISE is a standing wave, and the corresponding energy eigenstates are called stationary states. Finding these eigenstates is the primary goal of trying to solve the TISE. This section is concerned with the methods of solving it. Considering a seemingly simple form of the TISE, one may ask why this section is so long. The trick is that for any system more complicated than one-electron atoms or molecules, the TISE cannot be solved exactly. The reason for this is the last term in the Hamiltonian (2), which involves

coordinates of two electrons and represents electron–electron interactions. These terms make TISE inseparable. Hence, approximations are invoked when solving the TISE.

The first key approximation that is made is the BOA. The BOA is based on the fact that electrons are much lighter than nuclei, and so the motion of electrons and nuclei happens on very different time scales. Because of this, electrons are thought to adjust instantaneously to the positions of the nuclei in a molecule. The often used analogy here is the motion of flies on top of a garbage truck: the truck moves so slowly that the flies do not even notice its movement and adjust their positions instantaneously to the position of the truck. The BOA is a good approximation in most cases of chemical relevance. BOA implies that the total wave function of a molecule can be written as a product of the nuclear and electronic parts, and the total Hamiltonian as a sum of the nuclear and electronic Hamiltonians:

$$\Psi_{\text{total}} = \psi_{\text{nuclear}} \cdot \psi_{\text{electronic}}, \tag{4}$$

$$\hat{H}_{\text{total}} = \hat{H}_{\text{nuclear}} + \hat{H}_{\text{electronic}}. \tag{5}$$

Then, the TISE is separable within the BOA, and in particular the electronic TISE can be solved with the nuclear coordinates included as parameters:

$$\hat{H}_{\text{el}}\psi_{\text{el}} = E_{\text{el}}\psi_{\text{el}}, \tag{6}$$

where

$$\hat{H}_{\text{el}} = -\frac{1}{2}\sum_i \nabla_i^2 - \sum_a \sum_i \frac{Z_a}{|\boldsymbol{R}_a - \boldsymbol{r}_i|} + \sum_i \sum_{j \neq i} \frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|}. \tag{7}$$

The first term in the electronic Hamiltonian represents the sum of kinetic energies of electrons, the second term is the potential energy term for the attraction of electrons to the now stationary nuclei, and the third term is the interelectronic repulsion. The nuclear kinetic energy is zero, and the internuclear repulsion is a constant, within the BOA. The solution of the TISE is the total energy and the wave function itself. For the exact wave function, the energy can be found as

$$E_{\text{el}} = \frac{\left\langle \psi_{\text{el}} \middle| \hat{H}_{\text{el}} \middle| \psi_{\text{el}} \right\rangle}{\left\langle \psi_{\text{el}} \middle| \psi_{\text{el}} \right\rangle}. \tag{8}$$

As was mentioned already, solving the TISE exactly is impossible for any system with more than one electron. However, there is a helpful variational principle that tells us that any approximate wave function would return an eigenenergy that is an upper-bound of the true energy:

$$E_{\text{approximate}} \geq E_{\text{el}} = \frac{\left\langle \psi_{\text{el}} \middle| \hat{H}_{\text{el}} \middle| \psi_{\text{el}} \right\rangle}{\left\langle \psi_{\text{el}} \middle| \psi_{\text{el}} \right\rangle} \tag{9}$$

Hence, the "direction" of improving the electronic wave function when solving the TISE using approximate methods is always known: it is that of decreasing the total electronic energy.

The other major approximation is the mean-field approximation dictating that each electron can be viewed as moving in an averaged field created by all other electrons. This approximation gave rise to the Hartree-Fock (HF) method that is the foundation of many approaches in quantum chemistry.

The wave functions for individual electrons in HF are called molecular orbitals (MOs), denoted by ϕ_i. Each ϕ_i is a solution of the one electron equation with the corresponding orbital energy as an eigenvalue. The total electronic wave function in HF is represented as an antisymmmetrized product of MOs, called a Slater determinant. Here is how it looks like for an N-electron system:

$$\psi_{\text{el}} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \varphi_1(\boldsymbol{r}_1) & \varphi_2(\boldsymbol{r}_1) & \cdots & \varphi_N(\boldsymbol{r}_1) \\ \varphi_1(\boldsymbol{r}_2) & \varphi_2(\boldsymbol{r}_2) & \cdots & \varphi_N(\boldsymbol{r}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(\boldsymbol{r}_N) & \varphi_2(\boldsymbol{r}_N) & \cdots & \varphi_N(\boldsymbol{r}_N) \end{vmatrix}, \tag{10}$$

where $\mathbf{r}_i$ are coordinates of individual electrons, and the prefactor in the expression is the normalization constant. Writing the wave function in the form of Slater determinant takes care of the fact that the fermionic wave function should be antisymmetric with respect to permutations of any two fermions (i.e., electrons). In other words, if electrons i and j are swapped, the total electronic wave function should change the sign. Where to get the actual MOs, ϕ_i? Historically, MOs should be written as linear combinations of hydrogen-like atomic orbitals, which is referred to as the MO-LCAO principle: Molecular Orbitals written as Linear Combinations of Atomic Orbitals. MO-LCAO assumes that AOs do not change too much when atoms come together to form a molecule. First of all, this assumption is quite a stretch, and second, actual hydrogenic AOs are hard to deal with computationally. Hence, in practice, MOs are instead written as linear combinations of basis set functions:

$$\varphi_i = \sum_k c_k \phi_k. \tag{11}$$

The basis functions are chosen such that they are mathematically easier to use; for example, they can be Cartesian Gaussians. Obviously, the larger the pool of the basis functions, the more accurately can it represent the MOs. Hence, big basis sets are a good thing. A variety of basis sets of various sizes has been developed for different chemical elements. A large data base of available basis sets is supported by the EMSL of the US Department of Energy (https://bse.pnl.gov/bse/portal).

For each MOs, the one-electron eigen problem is solved:

$$\hat{F}\varphi_i = \varepsilon_i \varphi_i, \tag{12}$$

where ε_i are orbital energies, and $\hat{F}$ is the one-electron Fock operator:

$$\hat{F} = -\frac{1}{2}\nabla_i^2 - \sum_a \frac{Z_a}{|\mathbf{r}_i - \mathbf{R}_a|} + \sum_{j \neq i} (2\hat{J}_j(1) - \hat{K}_j(1)), \tag{13}$$

containing the electronic kinetic energy operator (first term), the electron–nuclear potential energy operator (second term), and the coulomb and exchange pair of operators having to do with electron–electron interactions:

$$\hat{J}(1)\varphi_i(1) = \left\langle \varphi_i(1)\varphi_j(2) \left| \frac{1}{r_{12}} \right| \varphi_i(1)\varphi j(2) \right\rangle = \langle ij \mid ij \rangle . \tag{14}$$

$$\hat{K}(1)\varphi_i(1) = \left\langle \varphi_i(1)\varphi_j(2) \left| \frac{1}{r_{12}} \right| \varphi_j(1)\varphi_i(2) \right\rangle = \langle ij \mid ji \rangle . \tag{15}$$

The latter is a purely quantum mechanical operator that has no analog in classical mechanics, and it appears due to the wave function being written as an antisymmetrized product. Exchange terms are nonzero only for electrons of the same spin. Hence, there are twice as many coulomb terms as there are exchange terms. The full Hamiltonian of the system in HF is then

$$\hat{H}_{\mathrm{HF}} = \sum_i \hat{F}(i). \tag{16}$$

HF is a method of simultaneously solving the one-electron Fock equations for all the MOs and orbital energies through the variational principle. The difficulty is that the one-electron Fock operators themselves depend on the wave functions of all other electrons, which in turn are not known. Hence, the procedure for solving the HF equations must be iterative, and it is called the Self-Consistent Field (SCF) procedure. For a given nuclear geometry, the initial guess for the coefficients c_k in the expansions for the MOs is made at the start. The coefficients are then variationally improved, in the direction of reducing the electronic energy corresponding to this wave function. The Fock equations are solved for the trial MOs, and the MOs are updated, and compared to the ones from the previous step. The convergence criterion is satisfied when the change in the newly produced MOs is within a certain acceptable threshold.

The speed and ultimate success of SCF critically depends on the initial choice for the trial wave function. In quantum chemical calculations, especially for electronically complicated systems, such as those containing multiple transition metal centers, or having low lying excited states, achieving the SCF convergence is a common everyday battle. However, some steps can be taken to help the convergence: having an initial guess for the MOs coming from a simpler calculation and having a fairly accurate starting geometry of the system is a good idea, for example. Also, there is a helpful approach called “level-shifting” in which the unoccupied MOs are artificially brought up in energy so as to discourage their mixing with occupied MOs.

As a result of the HF SCF procedure, a set of orthogonal occupied and unoccupied MOs, whose number equals the number of the basis set functions, and their energies are produced. The total electronic energy of the system is then the sum of the energies of all occupied MOs, with all the double-counted coulomb and exchange terms removed:

$$E_{\text{el}} = \sum_i \varepsilon_i + \sum_i \sum_{j \neq i} \left(J_{ij} - \frac{1}{2} K_{ij} \right). \tag{17}$$

Double-counting happened because J_{ij} and K_{ij} terms contributed to the orbital energies of both ϕ_i, and ϕ_j, and thus half of them need to be removed. SCF is a common approach for all quantum chemical methods, HF and beyond.

Finally, HF exists in two flavors: restricted (RHF) in which each pair of paired electrons is enforced to occupy the MO having the same spatial part, and unrestricted (UHF) in which spatial parts of the MOs occupied by the spin-up and spin-down electrons are treated independently and allowed to be different. Obviously, RHF is much faster than UHF. The use of RHF is permitted only for closed-shell systems.

In quantum chemistry, quality results are those of chemical accuracy, i.e., within 1 kcal/mol from the experimental data. HF is a good starting point and a foundation for many better methods, but in itself it is pretty inaccurate, according to this modern standard. The good HF days were perhaps in the 1970s. What is missing in HF is the so-called electron-correlation energy. In fact, by definition, the correlation energy is the difference between the exact electronic energy and the HF electronic energy. Why does it arise? Because, in reality, the mean-field approximation accounts only for some of the Columbic interaction between electrons (the HF exchange energy is exact). There are two kinds of electron correlation: static and dynamic. Static electron correlation is reflected in the fact that electrons partially occupy virtual MOs in an attempt to avoid each other. Dynamic electron correlation has to do with correlated motion of electrons due to their Coulomb repulsion. In view of this, it is easy to understand why inclusion of electronic correlation always means mixing the reference wave function with some contributions coming from the excited states. Methods of the post-Hartree-Fock type are dedicated to better treatment of electronic correlation, and the degree to which they do it goes together with the computational cost of the method.

The simplest methods of the post-HF category are the Møller-Plesset (MP) perturbation theory (PT) methods, MP2, MP3, MP4, and the rarely used MP5 [1]. These methods add a purturbative correction to the HF solution. MPPT is a variant of the general Rayleigh–Schrödinger PT. The total exact Hamiltonian is viewed as a sum of the HF Hamiltonian, $\hat{H}_{\text{HF}}$, plus the perturbation ($\hat{V}$) responsible for the missing electron correlation:

$$\hat{H} = \hat{H}_{\text{HF}} + \lambda \hat{V}. \tag{18}$$

The part of the total Hamiltonian that is viewed as a perturbation in MPPT is defined as:

$$\hat{V} = \sum_i \sum_{j \neq i} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} - \sum_i \sum_{j \neq i} \left(J_{ij} - \frac{1}{2} K_{ij} \right), \tag{19}$$

i.e., the difference between the true electron–electron interaction and the HF electron–electron interaction energy. The corrected wave function is

$$\psi = \psi_0 + \lim_{n \to \infty} \sum_{i=1}^{n} \lambda^i \psi^{(i)}, \tag{20}$$

where ψ_0 is the HF wave function and $\psi^{(i)}$ are the parts of the electronic wave function due to the perturbation. The total energy is a power-series expansion in λ, too:

$$E = E_0 + \lim_{n \to \infty} \sum_{i=1}^{n} \lambda^i E^{(i)}, \tag{21}$$

where $E^{(0)} = \sum_i \varepsilon_i$. Skipping the details, this particular partitioning of the total Hamiltonian into the unperturbed part and the perturbation yields a first order perturbation theory energy correction equal to zero, with the second order PT correction being:

$$E_0^{(2)} = \sum_{n \neq 0} \frac{\left\langle \psi_0 \left| \hat{V} \right| \psi_n \right\rangle^2}{E_0^{(0)} - E_n^{(0)}}, \tag{22}$$

where ψ_0 is the HF wave function, ψ_n are doubly excited states (terms due to the singly and triply excited states are zero), and $E^{(0)} = \sum_i \varepsilon_i$ for either the ground state, or the n^{th} excited state. If the power series for the total energy is truncated after $E^{(2)}$, the method is called MP2, and it is the most commonly used method of the MP group. Higher order corrections can be derived fairly straightforwardly too, to produce the MP3, MP4, and MP5 formalisms.

There is an important limitation intrinsic to the MP methods. As any PT, MP works well only if the perturbation is small, i.e., the reference HF wave function is already a fairly good solution. If this is not the case, MPPT results will be erroneous. Additionally, one must keep in mind that the perturbation theory correction is nonvariational. In other words, it is not guaranteed that the total energy obtained with the MP methods will be the upper-bound of the true energy. What often comes as a surprise is an oscillatory behavior of the total energies found with MP2, MP3, MP4, and MP5, meaning that with the seeming increase in the amount of electron correlation the energy does not consistently go down but oscillates, often slowly converging to a particular intermediate value. This behavior is also a result of nonvariational treatment. In such cases, the best MP solution can be found by projecting to the limit of MP∞.

The next best, and in fact one of the best quantum mechanical methods is coupled cluster (CC) [2,3]. CC accounts for a good deal of both static and dynamic electron correlation. In CC, the wave function undergoes an expansion over the excited states:

$$|\psi\rangle = e^{\hat{T}} |\psi_o\rangle \,, \tag{23}$$

where $|\psi_o\rangle$ is the reference HF function.

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \cdots \,, \tag{24}$$

is the cluster operator. $\hat{T}_1$ is the operator producing all single excitations, $\hat{T}_2$ is the operator producing all double excitation, etc. The $e^{\hat{T}}$ operator can be expressed as a Taylor series, and produces an expansion of the wave function in terms of many Slater determinants.

$$e^{\hat{T}} = 1 + \hat{T} + \frac{\hat{T}^2}{2!} + \cdots = 1 + \hat{T}_1 + \hat{T}_2 + \frac{\hat{T}_1^2}{2} + \hat{T}_1\hat{T}_2 + \frac{\hat{T}_2^2}{2} + \cdots \tag{25}$$

The expansion is usually truncated at double electronic excitations, which yields the CCSD method. The triple excitations can be added nonvariationally, via one of the few available algorithms. The most popular one has triple excitations treated perturbatively, and the corresponding method is called CCSD(T) [4]. CC is a very expensive method, and requires a lot of memory, disk space, and computer time. In many cases, the method features high accuracy, so the results can be in a quantitative agreement with experiment. Cases where CC, and for that matter all HF-based methods, do not work are discussed next.

So far we have been correcting the HF reference wave function with additional electron correlation terms, i.e., we sincerely hoped that HF is in fact not such a bad starting point. However, sometimes this is not true. Typical systems where HF does not work are those whose ground states are degenerate or nearly degenerate, i.e., have closely lying excited states. Clearly, in these cases, two or more electronic configurations should be able to contribute to the wave function on equal footing (the wave function in such cases is characterized as multiconfigurational). A prototypical case is an open-shell singlet diradical, where exactly two configurations should be equal players: $\phi_1(\alpha)\phi_2(\beta)$ and $\phi_1(\beta)\phi_2(\alpha)$, because choosing only one of these configurations would mean putting "tags" on electrons: this one is spin-up, and that one is spin-down. However, labeling electrons is not permitted, due to their indistinguishablility. HF would exactly "label" electrons by representing the wave function as a single Slater determinant. Hence, another approach, not based on HF is needed to handle such situations. Another classic example is the treatment of homolytic bond cleavage reactions, such as $H_2 \rightarrow H + H$, using HF or other single reference methods. In H_2, there is one occupied bonding σ-MO populated by two electrons. Of course, at the dissociation limit, the two atoms should have one electron each, and the system should acquire a singlet diradical character. However, when the atoms are pulled apart, away from their equilibrium distance in the molecule, HF will keep the electrons paired, because it is unable to deal

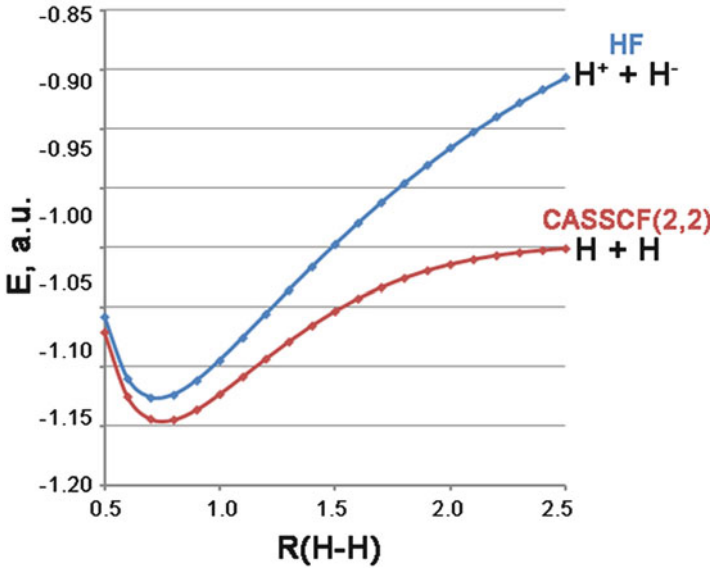

Fig. 1 Scan along the R(H–H) coordinate for the H2 molecule performed at the HF/6–31G* and CASSCF(2,2)/6–31G* levels of theory. It is obvious that HF is unable to properly treat the singlet diradical that forms at large H–H separations, and creates an excited state, $H^+ + H^-$. On the other hand, CASSCF dissociates the molecule properly on two neutral H atoms

with singlet diradicals. As a result, HF will dissociate the H_2 molecule into H^- and H^+, which is an excited state for the system at large distances (Fig. 1). The reason is that at large distances, the system again needs to be represented by two Slater determinants, and not just one.

The methods that fit the bill are called multireference methods. They allow many Slater determinants to contribute to the total wave function. The simplest such method is configuration interaction (CI) [5,6]. The total CI wave function is literally a normalized linear combination of the ground and excited state Slater determinants in which the expansion coefficients c_i are variationally optimized:

$$\Psi = \sum_{i=1}^{n} c_i \psi^{(i)}. \tag{26}$$

CIS includes only the ground and singly excited determinants, CISD also includes doubly excited determinants, and full CI includes all possible excitations, as far as the basis set permits. Notice that the ground state MOs expanded in terms of basis set functions remain fixed in CI, and only the coefficients in front of the Slater determinants are optimized. This is an approximation, because in reality, excited state MOs may be slightly different. CI is a simple method that is acceptable in treatment of open-shell low spin systems.

A slightly more elaborate multireference method is the Multiconfigurational Self-Consistent Field (MCSCF), or its most commonly used variant the Complete Active Space Self-Consistent Field (CASSCF) method [7]. It is similar to CI in that the wave function is a linear combination of Slater determinants. However, unlike in CI, in CASSCF the MOs inside of each determinant are also variationally optimized. In this sense, CASSCF can be called multireference HF. The excited state configurations are generated within the chosen active space. For example, CASSCF(2,4) indicates that there are two active electrons that may be promoted to higher MOs to form excited states, and the total number of MOs over which these two electrons may be distributed is four (that includes the MOs initially occupied by the two active electrons). Bigger active spaces typically mean higher accuracy. If certain included excited states do not contribute to the wave function, the coefficients in front of those determinants will be close to zero, but those included determinants that appear to be important will have a chance to mix with the reference configuration.

CI and CASSCF are methods that include static electron correlation, but are weak in treating dynamic electron correlation. They can provide a hint for whether or not a particular system has a multiconfigurational wave function, but cannot provide very accurate results. Again, the simplest way to improve on a simpler solution is to use the PT. Indeed, much like MP2 and MP3 improve the HF solution, CASPT2 and CASPT3 are used to improve the CASSCF solution [8]. These methods are the second and third order, respectively, complete active space PT. They are known to bring the results closer to the desired chemical accuracy, for species with multiconfigurational wave functions. It is again important to remember that in order for the PT to work, the reference CASSCF solution should be good enough, i.e., capture all the electronic configurations majorly contributing to the wave function. In other words, the active space should be chosen carefully. Another known caveat is that CASPT2 systematically overstabilizes states having more unpaired electrons, and as a result the ground spectroscopic state may be determined incorrectly [9].

Most sophisticated variations that include much of static and dynamic correlation for multireference methods are almost certainly prohibitively expensive for biologically relevant calculations, at least until we learn to take a full advantage of computing on GPU. However, it is worth mentioning that they exist. One such method is multireference CI, MRCI, which forms a CI expansion, but the components in the expansion are CASSCF wave functions, instead of single Slater determinants [10, 11]. Another method is multireference coupled cluster, MRCC [12, 13]. This is a young and promising method of exceptional accuracy, but it is also exceptionally expensive computationally.

One thing to keep in mind about multireference calculations is that all of them besides full-CI are not size consistent. This means that a particular truncated CI active space does not provide equal amounts of electron correlation in calculations of molecules of different sizes. As an implication, the dissociation limits of molecules described by methods that are not size consistent will be slightly off.

What is often viewed as a sanctuary from the computational expense and hardship of ab initio wave function methods, is a principally different approach

based on the density functional theory (DFT) [14]. DFT methods are often a good alternative to electron-correlated methods, because they are relatively inexpensive, and yet fairly accurate in many cases. DFT states that all properties of the system are uniquely determined by its ground state electron probability density, $\rho_0(x, y, z)$, which is a function of only three variables. In particular, the ground state energy is a functional of ρ_0: $E_0 = E_0[\rho_0]$. The theory in principle should completely avoid the use of the wave function. However, the problem is that in practice ρ_0 is usually found from the wave function, which then anyhow needs to be found in the first place.

What makes DFT practical is the use of the so-called Kohn–Sham orbitals. The system of interest can be represented as a system of noninteracting electrons all experiencing the same "external potential," $\upsilon_s(\boldsymbol{r}_i)$, so as to reproduce the exact electron probability density:

$$\rho_0 = \sum_{i=1}^{N} |\varphi_i|^2, \tag{27}$$

where ϕ_i are the spatial Kohn–Sham orbitals. The ground state wave function is then the Slater determinant, in ϕ_i. The exact ground state electronic energy can be written in terms of one-electron Kohn–Sham orbitals and one-electron density as

$$E[\rho] = -\frac{1}{2}\sum_i \int \varphi_i^*(\boldsymbol{r})\nabla_1^2\varphi_i(\boldsymbol{r})\mathrm{d}\boldsymbol{r} - \sum_a \frac{Z_a}{|\boldsymbol{r}-\boldsymbol{R}_a|}\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} + \frac{1}{2}\int \frac{\rho(\boldsymbol{r}_1)\rho(\boldsymbol{r}_2)}{|\boldsymbol{r}_1-\boldsymbol{r}_2|}\mathrm{d}\boldsymbol{r}_1\mathrm{d}\boldsymbol{r}_2 + E_{XC}[\rho]. \tag{28}$$

The last term is the exchange-correlation energy, which is always empirical, and makes DFT not exact. It is also the term that is the most difficult to get right, because it is very problem dependent.

The Kohn–Sham orbitals are found variationally, through iteratively solving the Kohn–Sham equations (SCF procedure). ϕ_i are eigenfunctions of the one-electron operators, $\hat{H}_i^{KS}$, with the eigenenergies being the orbital energies, ε_i:

$$\hat{H}_i^{KS}\psi_i = \varepsilon_i\psi_i, \tag{29}$$

or

$$\left(-\frac{1}{2}\nabla_1^2 - \sum_a \frac{Z_a}{|\mathbf{r}_1-\mathbf{R}_a|} + \int \frac{\rho(\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|}\mathrm{d}\mathbf{r}_2 + V_{XC}(\mathbf{r}_1)\right)\varphi_i(\mathbf{r}_1) = \varepsilon_i\varphi_i(\mathbf{r}_1), \tag{30}$$

where the exchange-correlation potential, $V_{XC}[\rho] = \frac{\delta E_{XC}[\rho]}{\delta\rho}$, was introduced. Since the Kohn–Sham orbitals by definition represent noninteracting electrons, the total Hamiltonian is the sum of the one-electron operators:

$$\hat{H}_s = \sum_{i=1}^{N} \hat{H}_i^{\mathrm{KS}}. \tag{31}$$

The accurate exchange-correlation energy is the holy grail of DFT. It may be calculated via numerous recipes, none of which produce exact results. Hence, the large pool of available DFT functionals, such as BPW91, B3LYP, PBE, TPSSh, MO6, etc. The letters in these abbreviations signify the authors of these functionals: B is for Becke, L is for Lee, P is for Perdew, S is for Scuseria or Staroverov, W is for Wang, and Y is for Yang. Some of these functionals are hybrid, based on linear combinations of several different functionals, and also adding a portion of exact exchange coming from HF (e.g., B3LYP and TPSSh). Hybrid functionals are usually more accurate. TPSSh is a modern functional that includes intermediate range exchange, which appears to be important in the description of transition metal complexes. Choosing an appropriate DFT functional for a given problem is a complicated business frequently guided by experience, and thus may have an alchemic appearance.

Despite the attractiveness of DFT, it is a must to remember the issues that it has: DFT suffers from electron self-interaction error, i.e., it includes the electron–electron interactions for an electron with itself, which is of course unphysical, and has no dispersion built in. As a result, DFT consistently underestimates reaction barriers. It also does poorly for weakly bound complexes, and cannot handle long-range charge transfer. The problem is that the noncoulomb part of exchange usually dies off too quickly at large distances. One solution is range-separated DFT that treats long and short range differently and can capture the dispersion, which is an active area of research. In *Gaussian 09*, a prefix LC- maybe added before the functional name to add a long-range correction (e.g., LC-BLYP).

DFT alone is a subject of several monographs, and will not be explicated any further, beyond its basic principle, advantages, and deficiencies.

A large effort is directed toward the reduction of the computational cost of ab initio calculations. For most methods, various “short-cuts” have been developed, for example, the resolution of identity treatment, the dual basis formalism, and techniques that account for local correlation more than for correlation between distant MOs. These tricks are not discussed in this brief overview. However, it is important to be aware of them, and use them when possible, because they often allow for a considerable reduction of the computational cost and treatment of fairly large molecules with correlated methods.

Finally, ab initio calculations can be coupled with implicit solvation models, such as PCM, to account for solvent effects [15]. While providing some correction to the results due to the presence of the environment, implicit solvation models cannot reproduce the effect of very directional hydrogen bonds that water molecules may form with particular groups in the system. Therefore, caution must be exercised, when exploring the role of water in chemical reactions, for example, using implicit solvation, and if possible, it is better to rely on explicit solvation (to be discussed in Sect. 3).

Any ab initio calculation can be coupled to the optimization of nuclear geometry, where after each SCF cycle, gradients for the nuclear motion are calculated, and the

geometry is adjusted accordingly, to move the species closer to a stationary point on the PES. After that, the electronic TISE is solved again through the SCF procedure to convergence, and so on, until the convergence in geometry is also reached. There is a wealth of geometry optimization algorithms that have been implemented. Optimization is one of the tools most commonly used for biologically relevant molecules. Also, harmonic vibrational frequencies can be calculated to confirm the nature of the found stationary point in the PES (minimum of saddle point), and to compare the calculated IR spectrum to the experiment for the structural characterization. For the latter purpose, frequencies need to be scaled down by a small factor (for example, this factor is 0.9741 for the TPSS DFT functional), in order to empirically account for the anharmonicity of the PES. IR and Raman activities of vibrations are also calculated. Ab initio calculations can provide charge distribution in a molecule, and there are several charge localization schemes, such as CHELPG (CHarges from Electrostatic Potentials using a Grid based method) and NPA (Natural Population Analysis). Those can be used, for example, in the analysis of chemical bonding in the systems, and to find the location of specific electrons participating in a reaction.

In order to use ab initio calculations for the assessment of reaction mechanisms and energetics, one may perform a scan of the PES of the system along a suspected reaction coordinate(s), and this way find the location of the stationary points (minima and transition states) on it. One may also optimize directly to the minima and transition state, if a good guess for the geometries is available to begin with. The nature of a stationary point may be confirmed via vibrational frequency calculations: minima will have all vibrational frequencies nonnegative, whereas transition states will have exactly one imaginary frequency, and the corresponding normal mode of vibration will be along the reaction coordinate. There are also automated methods to search for the transition states of reactions, such as QST2 and QST3 implemented in *Gaussian*.

Some additional quantum mechanical effects may be taken into account ad hoc. For example, geometry optimization will take the system down to the minimum on the PES. However, one must recall that in reality the system never resides exactly there, but rather sits slightly higher in energy due to vibrations, because quantum objects are never at rest. Hence, an additional quantum mechanical correction to the energies should be added due to the so-called zero-point vibrational energy, ZPE. Harmonic ZPE can be obtained from the calculations of vibrational frequencies. Also, when reactions of interest involve transfer of light particles, such as protons, H atoms, or hydrides, empirical tunneling corrections may be added when computing activation barriers.

DFT, HF, and post-HF methods are implemented in several reputable ab initio packages, for example, *Gaussian, GAMESS, TURBOMOLE, ADF, NWChem, Q-Chem*, and *MOLPRO*. Ab initio program packages that have the most developed multireference methods are *MOLCAS, MOLPRO, GAMESS, COLUMBUS*, and also *Gaussian*.

2.2 Navigating Through the Wealth of Ab Initio Methods: Some Quick Recipes

Based on the presented foundation, next, it is important to be able to recognize which approximations should and should not work in each particular case, the accuracy of results one might expect, and how long it would take to complete a particular calculation. Here are some suggestions on how to quickly turn the wealth of ab initio methods into your weapon.

For a molecule or model complex prepared for simulations, it is always advised first to do a quick check for whether or not single determinant methods are applicable at all. In other words, one must be sure that the HF wave function is a good reference function to which electron correlation can be added for a better result. In order to check this, a CASSCF(*n,m*) single point calculation should be performed, and one needs to confirm that the HF Slater determinant is the single main contributor to the CAS expansion (with a coefficient of 0.9 or higher). If the system is single reference, wipe off that cold sweat, as life just became much easier. The next temptation should be to use computationally inexpensive DFT. As a rule of thumb, if the system under consideration does not have weakly bound components and is not an anion, especially carrying more that a single charge, it is usually safe to use DFT methods. For large complexes, long range corrections are desirable though. Sometimes DFT does not work. The famous case is the Cr_2 dimer, whose bond length cannot be predicted correctly by any DFT method. Hence, when using DFT, one needs to exercise caution and confirm that DFT results are not in contradiction with those obtained with correlated methods. So at least some testing of this sort needs to be performed. For more careful treatment of electron correlation, MP2 or CCSD(T) may be used, if the size of the system permits.

Another question is the choice for the appropriate basis set. The basic rule is "the bigger the better." The more basis functions one manages to include in a calculation while keeping it manageable, the more accurate the result will be. The size of the basis set matters particularly for the parts of the system that are involved in hydrogen bonding and other weak interactions, because with small basis sets the fragments of the system will start using the basis functions from the neighboring fragments to lower their energy (this is called the basis set superposition error). As a result, the system will become artificially contracted. For transition metals, especially the ones of the sixth and seventh periods in the Periodic Table, the basis sets with relativistic pseudopotentials should be used.

What if the wave function is discovered to be multiconfigurational? This outcome is suspected for low-spin transition metal complexes, weakly bound systems, regions of the PES near transition states of some reactions, and other situations where two or more Born–Oppenheimer PESs become proximal. Then, it is best to rely on multireference methods. The simplest ones of this category are CI and CASSCF. For additional dynamic correlation, CASPT2 can be used. Other, more expensive methods are probably out of reach for systems of a size meaningful to biology. It could be recommended to optimize the geometry using CI or CASSCF,

and then run a single point energy calculation at CASPT2, to refine the total energy. Multireference methods are often hard to converge. For better convergence and correct results, it is critical to choose the active space carefully. The ambiguity that hides in this practical suggestion occasionally makes multireference methods repugnant to some researchers. One needs to experiment with the size of the active space so as to capture all significant components to the CAS expansion, and yet to not run out of memory. As a qualitative rule, occupied MOs of a certain type that are included in the active space should have vacant counterparts of the same type, so as to enable the promotion of electrons to those vacant MOs. For example, if occupied delocalized π-MOs on a particular fragment are in the active space, it is preferred to have all unoccupied delocalized π-MOs on that fragment also included. If occupied d-AOs on a metal center are in the active space, the unoccupied d-AOs should also be included. Depending on the chemical bonding in the system, this might not be enough though. For example, it was shown that for iron porphyrins the ground electronic state is predicted more accurately by CASPT2, if not only 3d, but also 4s AOs on Fe are included in the active space [9]. This is because inclusion of 3d and 4s AOs allowed for the partial 3d–4s hybridization that enhanced the bonding between Fe and the porphyrin ring. Overall, if the system clearly exhibits a multiconfigurational nature, the only truly justified way to proceed is to use multireference methods. Hopefully, the Reader is now convinced and prepared to argue for this statement, if needed.

If the size of the system prohibits the use of multireference methods with any meaningful size of the active space, and yet the system is not a single reference, unrestricted DFT methods are often used as a compromise. This is so-called broken symmetry approach, and it is often used for the low-spin complexes of transition metals. It is the responsibility of the researcher to check the reliability of UDFT in each particular case!

2.3 Examples of Applications

Pure ab initio methods are often used for the assessment of structure, properties, and reactivity of biologically relevant model complexes.

A large field is the assessment of the mechanism of enzymatic reactions, using small complexes cut out of the protein. When doing this, one must be careful to include the chemically important amino acids in the active site that are in immediate contact with the substrate of the reaction, or hold metal centers. Amino acids are commonly truncated at the Cα atom, which is then capped with H, unless the amino acids is too long, in which case it can be made shorter in the model. There is always some ambiguity as to what residues should be included in the calculations, and which can be ignored. One must find a satisfying compromise between choosing a chemically meaningful model, and having it small enough for the ab initio calculations to be computationally affordable. One may argue that the rest of the protein is playing an important role in catalysis, and there is an ongoing debate

about the dynamics of the entire protein being a player in catalysis. Undoubtedly, nature constructs large proteins for a number of possible reasons, including efficient substrate binding and product release, chemical security and preservation of the architecture of the active site, and the specific overall protein properties. However, the motion of bulky protein parts is a process that occurs on a much slower time scale than does the catalyzed reaction. Therefore, as long as the conformation of the protein does not change dramatically from its crystal structure upon substrate binding, consideration of the catalytic mechanism on a small model is not a bad idea. As a payoff, ab initio calculations yield fairly accurate activation barriers and molecular properties of the reacting system, which can be compared to the experiment.

As an example, consider the mechanistic study of the enzyme urease performed using UDFT [16]. Urease is a di-Ni enzyme that catalyzes the hydrolysis of urea, and many things about the reaction mechanism, such as the role of the second Ni center, the protonation state of the bridging water molecule, the group playing the role of the nucleophile, etc. are still unclear. The model complex prepared by the authors included the two Ni centers, truncated amino acids that bind them, i.e., the immediate coordination shell of the metal ions, and the substrate (Fig. 2). First, the geometry of the active site extracted from the crystal structure was optimized, and the ground state multiplicity was determined. It appeared that the lowest energy state is a quintet. The catalytic mechanism was then explicated using UB3LYP. The initial complex between urea and the active site, and all intermediates on the reaction path were optimized, and vibrational frequency calculations confirmed that they are true minima on the PES. The transition states were also found via geometry optimization to the saddle point between the two minima. All relevant transition state structures were confirmed to be true saddle points, by calculating their vibrational frequencies, and making sure that in each case there is only one imaginary frequency corresponding to the displacements along the reaction coordinate. It was found that there are two competing modes of binding of urea to the active site: bidentate (Fig. 2) and monodentate (Fig. 3). In both complexes, urea can get hydrolyzed by the active site, and the calculated energetics of the two paths renders these mechanisms competitive.

In a similar spirit, Solomon and coworkers [17] performed a systematic and remarkably exhaustive UDFT study of chemically possible peroxo-type intermediates occurring in the nonheme di-iron enzyme class Ia ribonucleotide reductase (RNR). This enzyme is responsible for the oxygen atom removal from RNA building blocks, ribonucleoside diphosphates, to yield corresponding DNA building blocks, deoxyribonucleoside diphosphates, by RNR. Class I RNRs contain two iron atoms in one part of the protein, the R2 subunit, which is separated from the catalytic site in the R1 subunit where the radical chemistry involving the ribonucleotide occurs. The study was conducted on the R2 subunit using spectroscopically calibrated density functional computations of equilibrium structures. Fe–O and O–O stretch frequencies, Mössbauer isomer shifts, absorption spectra, J-coupling constants, electron affinities, and free energies of O_2 and proton or water binding were presented for a series of possible intermediates. The study explored how water or

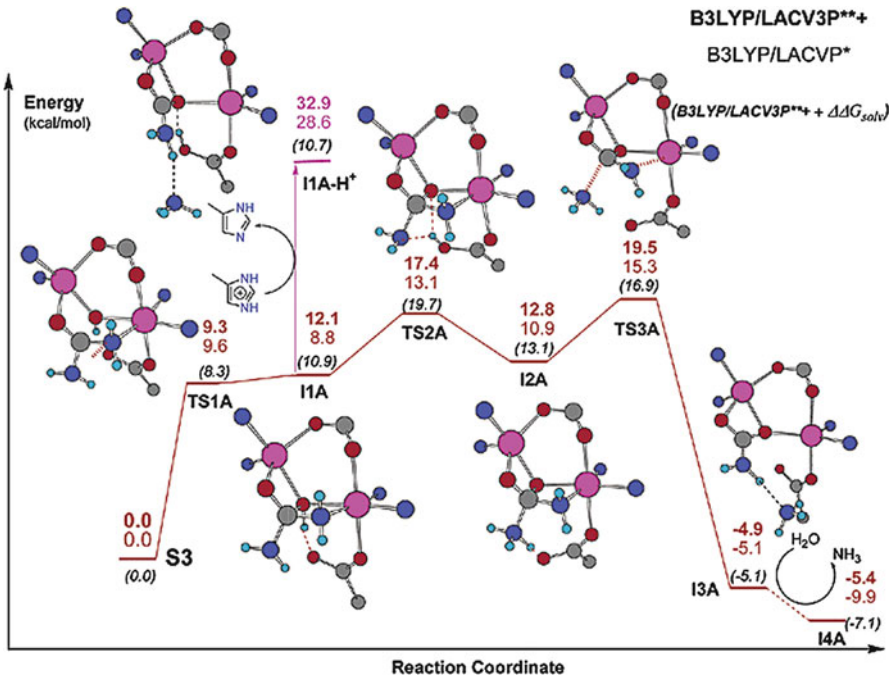

Fig. 2 Hydrolysis of urea catalyzed by urease, starting from the initial complex with bidentate coordination of urea to the active site. Adapted with permission from [16]

a proton can bind to the di-iron site of RNR and facilitate changes that affect the electronic structure of the iron sites and activate the site for further reaction. Two potential reaction pathways were presented: one where water adds to Fe1 of the *cis*-μ-1,2 peroxo intermediate P causing opening of a bridging carboxylate to form intermediate P′ that has an increased electron affinity and is activated for proton-coupled electron transfer to form the Fe(III)Fe(IV) intermediate X; and the other that is more energetically favorable where the P to P′ conversion involves addition of a proton to a terminal carboxylate ligand in the site which increases the electron affinity and triggers electron transfer to form X (Fig. 4). Both pathways provide a mechanism for the activation of peroxy intermediates in binuclear nonheme iron enzymes for reactivity.

One may argue against the use of UDFT for calculations on enzymes such as ureases and ribonucleotide reductases. For example, Ni^{2+} is a metal cation with the $[Ar]4s^2d^6$ electronic configuration. The incomplete population of the d-set of AOs leads to an ambiguity in the number of unpaired electrons on each Ni center. This number depends on the splitting of d-AO in a particular coordination environment of Ni. Several electronic states may lie close in energy, and may change order even in the course of the catalyzed reaction. Hence, the nature and number of singly and doubly occupied d-AOs on Ni may change. In addition, when two Ni centers

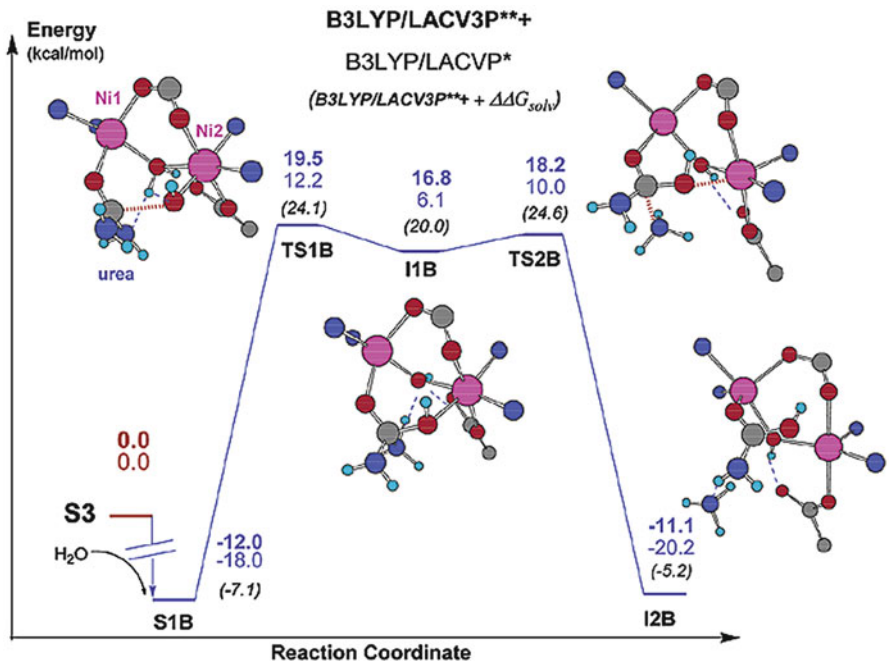

Fig. 3 Hydrolysis of urea catalyzed by urease starting from monodentate coordination of urea to the active site. Notice the similarity of the overall energetics to that characteristic of the reaction path shown in Fig. 2. Adapted with permission from [16]

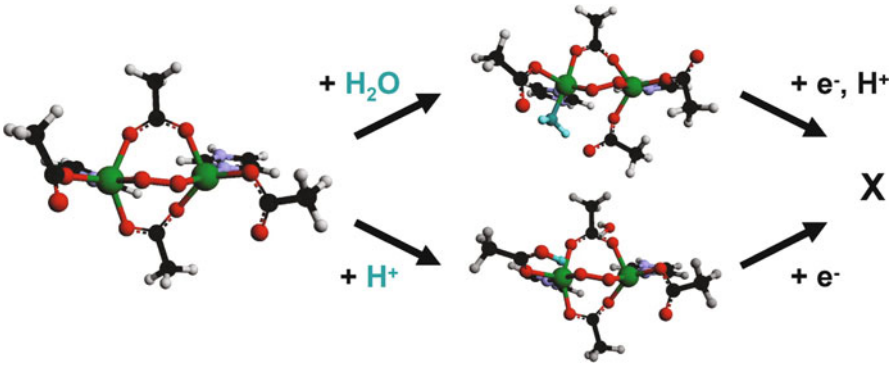

Fig. 4 Two possible reaction pathways for how water or a proton can bind to the di-iron site of ribonucleotide reductase and facilitate changes that affect the electronic structure of the iron sites and activate the site for further reaction. Adapted from [17]

are bridged, they may interact, and unpaired electrons on these centers may couple in a ferromagnetic or antiferromagnetic fashion. Whether this fairly long-distance interaction is adequately described by dispersion-lacking DFT or not remains to be seen. Furthermore, it was found, for example, that the surrounding protein

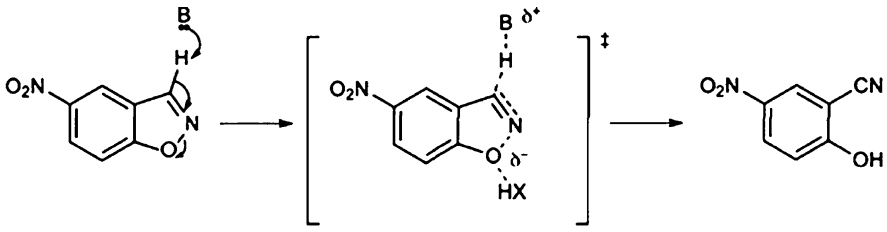

Fig. 5 The reaction of Kemp elimination, subject to catalysis by artificial enzymes designed starting from ab initio calculations

environment can impact the adapted electronic configuration of the metallic site, and the impact is stronger when the prediction is made by multireference methods than when DFT is used [9]. The problem is rooted again in the poor treatment of long-range effects in DFT. As we now know, when multiple electronic states are close in energy, and the ground state is thus nearly degenerate, the best solution is to use high level multireference methods. However, for this large model complex, UDFT is probably a wise compromise, and one may rely on the "broken-symmetry" solution somehow taking care of any multiconfigurationality of the wave function, should it arise. DFT is much easier to use, and it is much faster. As a result, the temptation to use DFT and to never look beyond is high. It is fair to say that DFT is largely overused. A good understanding of the electronic structure and possible problems in the specific system should guide the researchers in their decision for the suitable methodology.

Pure ab initio calculations are also a great asset in the design of artificial enzymes, for example, in the recently developed "inside-out" enzyme design algorithm [18]. In this algorithm, the building of a new enzyme for any reaction of interest starts from the consideration of this reaction in the gas phase. The transition state to this reaction is found, by optimizing to the saddle point on the PES using ab initio methods, usually DFT. The strategic amino acids that will constitute the catalytic machinery in the enzyme are then placed around the transition state, and their conformations and orientations optimal for catalysis are also found using ab initio calculations. The structures thus obtained are called "theo-zymes," for theoretical enzymes. A theo-zyme is then grafted into an existing protein scaffold, and the rest of the enzyme binding pocket is rebuilt and repacked around the theo-zyme, using molecular mechanics and statistical mechanics techniques. The enzymes are then computationally and experimentally tested for catalytic activity. The majority of the work in this effort is done computationally, before any experiment is performed, and at the heart of the process are the ab initio calculations on a small model of the active site. The protocol was successful in the design of numerous active enzymes, for example, for the catalysis of the Kemp elimination reaction (Fig. 5). In this reaction, a base abstracts H from the substrate. The N–O bond opens, putting the negative charge on O, which eventually gets easily protonated. Thus, what was needed for the theo-zyme was a base (Glu, Asp, or Asp/Glu-His diad), and some stabilizers of the negative charge that develops in the reaction. The rest of the binding site should help

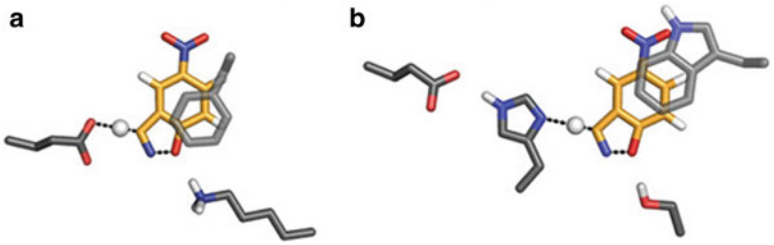

Fig. 6 Theo-zymes predicted by DFT calculations for the catalysis of the Kemp elimination reaction. These structures were then grafted onto existing protein scaffolds, and tested for catalytic activity. The transition state of the reaction is shown in *yellow*. (**a**) Glu is playing the role of the catalytic base, and (**b**) the Asp-His diad is the base in this case. Adapted with permission from [18]

the specificity to the given transition state, and also the hydrophobic environment. In Fig. 6, two types of theo-zymes predicted on the basis of these ideas are shown. In Fig. 6a, Glu plays the role of the catalytic base in the action. Phe is placed for π-stacking purposes, to stabilize the negative charge developing in the transition state and orient the substrate. Lys is placed to form a H-bond to the O atom that acquires the negative charge. The structure in Fig. 6b has Asp as the base, Ser as a H-bond donor, and Trp as a π-stacker. Theo-zymes such as these were installed into existing proteins and catalytic activity was confirmed with a considerable success rate. This demonstrates that small models of the active sites calculated using ab intio techniques are highly relevant to the chemical actuality in the binding site of proteins, even though long-range effects are obviously not included in the model.

Highly correlated electronic structure methods, especially multireference methods are often not needed for the ground state calculations. They are, however, extensively used in more sophisticated calculations, such as those of excited states and conical intersections. We withhold examples of their performance until Sects. 3 and 4.

3 Mixed QM/MM Techniques

3.1 Theoretical Foundation

An obvious deficiency of pure ab initio calculations that one might point out immediately is that calculations are always done on a small model, and the effect of the environment is considered negligible. However, sometimes excluding the effect of the larger biomolecule and solvent surrounding the model is highly undesirable, while it is still needed to describe the small reactive part of the system with quantum mechanical accuracy. The answer to this dilemma is often found in mixed QM/MM approaches [19, 20]. In QM/MM, the system is partitioned into what is thought to be the most chemically significant part, e.g., a reactive center, and the rest of the

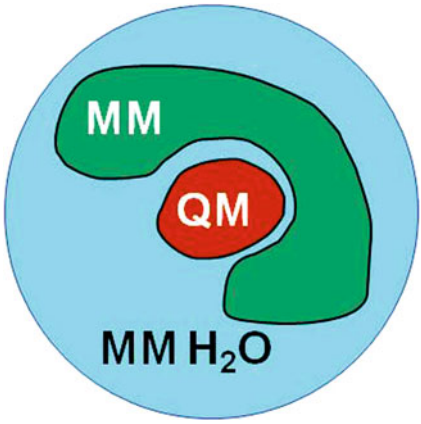

Fig. 7 Partitioning a biological macromolecular system in solution into a small chemically significant QM region and the MM region that includes the rest of the biomolecule and the solvent, and may be viewed as the environment for the QM subsystem

system that can be viewed as a "matrix." The former part of the system is treated at the QM level of choice, and called the QM region. The latter constitutes the MM region and is treated at a MM level, such as a classical force field. The QM region is usually embedded in the MM region (Fig. 7). As a more accurate variation of QM/MM, there exist QM/QM methods, where the two parts of the system are both treated quantum mechanically, but using the methods of different accuracy and hence computational demand, and also three-layer QM/QM/MM. QM/MM is a popular approach, and it is still gaining momentum in biochemical simulations.

The total energy of the QM/MM system is not merely a sum of the QM and MM energies of the corresponding regions, because the two regions are coupled. Instead, there are two schemes for calculating the total energy of the system: subtractive and additive. In the subtractive scheme, the required components are the QM energy of the QM region, $E_{\mathrm{QM}}(\mathrm{QM})$, MM energy of the entire system, $E_{\mathrm{MM}}(\mathrm{system})$, and MM energy of the QM region, $E_{\mathrm{MM}}(\mathrm{QM})$. Then, the total energy is

$$E_{\mathrm{QM/MM}}(\mathrm{system}) = E_{\mathrm{MM}}(\mathrm{system}) + E_{\mathrm{QM}}(QM) - E_{\mathrm{MM}}(\mathrm{QM}). \qquad (32)$$

The method is simple. However, there are two caveats: (1) The interactions between QM and MM regions are treated exclusively at the MM level, which is often inaccurate. In particular, the entire electrostatics of the system will be represented by fixed charges on atoms. (2) The MM parameters are needed for the QM region, and they are often not available, especially for systems containing transition metals or for systems in excited electronic states. A particularly prominent example of the subtractive method is ONIOM, by Morokuma and coworkers [21]. ONIOM has one additional improvement, however, that the MM charges are incorporated into the QM Hamiltonian (electrostatic embedding).

More popular nowadays is the additive approach, which dictates the following expression for the total energy of the system:

$$E_{\mathrm{QM/MM}}(\mathrm{system}) = E_{\mathrm{MM}}(\mathrm{MM}) + E_{\mathrm{QM}}(\mathrm{QM}) - E_{\mathrm{QM-MM}}(\mathrm{QM}, \mathrm{MM}). \qquad (33)$$

The MM calculation is now done only on the MM part. $E_{\mathrm{QM-MM}}$(QM,MM) is the QM–MM coupling term, which defines the communication between the two regions, and also distinguishes between different flavors of QM/MM. This term includes bonded and nonbonded interactions between the regions. Bonded interactions have to do with contributions to the total energy due to bond stretching, valence angle bending, and torsions. Nonbonded interactions include contributions from electrostatics and Van der Waals interactions.

Electrostatic interactions between QM and MM regions constitute an important long-range effect. As an illustrative example, consider a highly charged metal cation inside the QM region interacting with nearby amino acids of the MM region. The strong electrostatic potential provided by the cation should polarize the nearby MM region and may even change the protonation states of amino acids. This, in turn, should affect the protein conformation. At the same time, the impact of the repolarized and repositioned MM region on the QM region should change, too. So treating the system properly becomes quite nontrivial, and this is why it is desired to maximally screen the active center from the potentially inaccurately described QM–MM boundary. Thus, it is recommended to have the QM region as large as possible, which of course forces a compromise on the QM accuracy.

There are several schemes for treating electrostatic interactions, of different levels of sophistication. The mechanical embedding scheme assumes MM-type electrostatic interactions between point charges on atoms, which come from the parameterized force field for the MM part and QM calculations on the QM part. The problem is that QM charge-localization schemes are not necessarily accurate, updating the QM charges as the QM regions changes leads to discontinuities in the PES, and QM charges have no chance to respond to the presence of the MM charges around the QM region. The most popular electrostatic embedding scheme helps the latter problem by adding the electrostatic interactions with the point charges of the MM region into the quantum mechanical Hamiltonian. One potential pitfall of this scheme is that the charges located on the MM region very near the QM region tend to overpolarize the QM region. Yet a better scheme is polarizable embedding, which allows for repolarization of the MM region in response to the presence of the QM region. The repolarized MM region can then act back on the QM region changing its electron density. This QM/MM method requires iterative solving to self-consistency. Generally established fully polarizable force fields are not yet available, however, they are currently under construction.

Nonbonded Van der Waals interactions between QM and MM regions are short range, and usually create fewer problems. They are handled by MM only. For this, Lennard-Jones parameters, ε and σ, for the QM atoms in nonbonded contacts with the MM atoms have to be available. The interactions are approximated with the typical Lennard-Jones potential:

$$V_{\mathrm{LJ}} = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]. \tag{34}$$

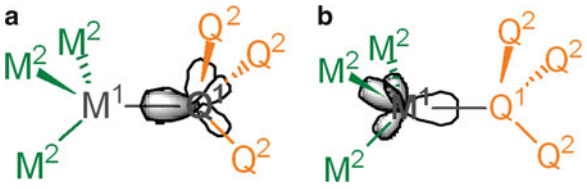

Fig. 8 The frozen-orbital boundary methods, atoms marked with "Q" belong to the QM subsystem, and atoms marked with "M" belong to the MM subsystem. (**a**) A set of localized orbitals is placed on Q1, one of which (*shaded*) is kept frozen and points toward M1. (**b**) GHO method: A set of localized orbitals is placed on M1, one of which (*open*) is active and points toward Q1. Adapted with permission from [19]

If the QM/M boundary does not pass through a covalent bond, that would be it. Sometimes, this is the case, for example, when cofactors are residing in proteins without being covalently attached. If, however, the QM/MM boundary does cut a bond, dangling bonds that appear on either side of the cut need care. Several approaches have been developed for this. The most popular is the link atom approach: an extra atom, usually H, is introduced to satisfy all dangling valencies in QM and MM calculations on the separated subsystems. Sometimes, instead a special boundary atom is placed at the interface, such that on the QM side, it mimics the cut bond and the electronic character of the MM moiety, and in the MM calculation, it behaves as a normal MM atom. The link atoms are exempt from any forces dictated by the calculations. A caveat here is that the link atoms are positioned very close to point charges on the MM atoms, and thus get overpolarized by the MM region. One solution is to use minimal basis sets for the link atoms in the QM calculations, to make them minimally polarizable. Otherwise, the electrostatic terms that are due to the interactions with the link atoms may be artificially diminished in the QM Hamiltonian, or the one-electron integrals on the link atoms can be deleted, or the charges on the MM atoms near the QM region can be redistributed or smeared.

Another approach is to place hybrid localized frozen orbitals at the boundary (Fig. 8). These MOs do not participate in SCF and remain unmixed with the rest of the MO system on the QM fragment. The MOs look the most like lone pairs on atoms at the interface. Some variations on the theme include parameterization for these orbitals. Also, there is a variant called generalized hybrid orbitals (GHO), in which frozen MO is placed on the MM atom rather than the QM atom at the interface, and this MO participates in SCF.

The bonded interactions that involve atoms on both sides of the interface need to be included in the MM calculation, which therefore involves three atoms into the QM region at each interfacial point. It is important to keep the QM region large enough so that no chemical transformations would involve these atoms at the interfaces. It is also important to have the QM–MM cut such that the net charge on the MM region would be zero.

QM/MM calculations can be done in conjunction with geometry optimization. Optimization to a saddle point on the PES (i.e., transition state) is often done through

the calculation of the Hessian on a small core fragment inside the QM region. The rest of the QM region is simply adiabatically optimized to adapt to the changes in the small core. Thus, energy profiles of chemical reactions can be calculated in the context of QM/MM.

QM/MM scheme may be combined with statistical mechanical sampling techniques, such as Monte Carlo (MC), or molecular dynamics (MD). This is very powerful, because it combines the accuracy of the QM treatment with the representation of large biomolecules as a statistical mechanical ensemble at finite temperatures and pressures. In addition, sampling allows for calculations of free energies rather than enthalpies, which is a more relevant quantity. However, it is nearly prohibitively expensive to combine the full sampling on the system with good QM calculations on the QM region at every configuration of MC or MD. The amount of sampling required for good equilibration and statistics accumulation needs to be large for large biomolecules (on the order to tens of millions of MC configurations for a protein consisting of ca. 200 residues in explicit water). To save time, QM/MM MC has been accomplished with the use of semiempirical methods for QM, as implemented in MCPRO [22]. Semiempirical methods rely on heavy parameterization, so that integrals in the QM calculations are either deleted or replaced with parameters. These methods are of a diminished accuracy, and nowadays almost extinct. Also, various tricks may be employed to reduce the amount of sampling on the QM region, or to avoid sampling of the QM region at all.

QM/MM statistical mechanics methods can be used to assess the structures of large biomolecules, and also the mechanisms of chemical reactions. For example, QM/MM MC or MD can be coupled to free energy perturbation (FEP), which among other things allows for driving chemical reactions along chosen reaction coordinates. FEP is based on imposing the chemical change into the system through a slowly introduced perturbation, i.e., in a series of small steps. If the steps are small enough, the assumption is that the statistical mechanical ensemble representing the system does not change much between the start and the end of the step. So the sampling can be done only on one of these points. Based on this approximation, the ΔG per each step can be calculated using the Zwanzig formula:

$$\Delta G\,(1 \rightarrow 2) = G_2 - G_1 = -k_b T \ln \left\langle \exp\left(-\frac{E_2 - E_1}{k_b T}\right)\right\rangle_1 . \tag{35}$$

Then, the changes in free energy per step can be added up to yield the change of free energy to the transition state of the reaction (activation free energy barrier), or all the way to the products of reaction. Thus, the mechanism and energetics of the process become accessible. The QM level of theory used in these calculations should be carefully chosen for the system at hand, and the same rules apply as for stand-alone ab initio calculations.

Reactions occurring on multiple PESs, i.e., in the nonadiabatic regime, can also be assessed in QM/MM, either by itself or coupled to statistical mechanical treatment, usually MD. Nonadiabatic dynamics is considered in Sect. 4.

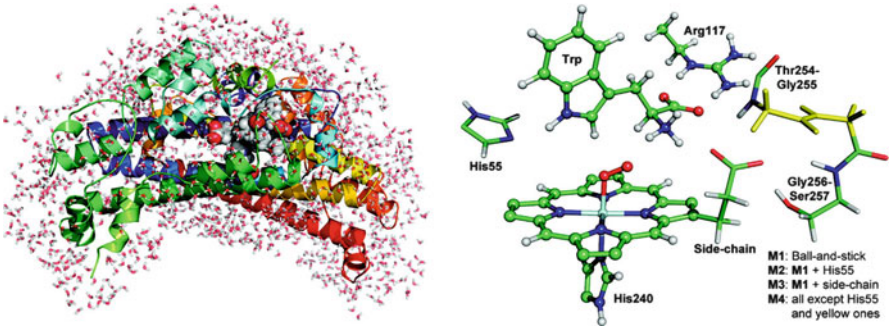

Fig. 9 The solvated tryptophan 2,3-dioxygenase protein and the QM Model (Except Link Atoms) used in the ONIOM calculations. Adapted with permission from [23]

3.2 Examples of Applications

The ONIOM method is of a paramount success and importance for biochemical applications. Among the most recent ONIOM studies is a mechanistic study by Morokuma and coworkers [23] on Heme functionality in bacterial tryptophan 2,3-dioxygenase, which catalyzes the oxidative cleavage of the pyrrole ring of L-Trp. In this mechanistic study, the system consisted of a large portion of the protein and explicit water (the classic rigid body TIP5P water model) (Fig. 9a). The QM region is shown in Fig. 9b. The QM method was B3LYP/6–31G*, and the MM methods was the AMBER force field. The link atoms approach was used: the QM region was capped with H atoms in QM calculations. Only a part of the MM region was allowed to move during the geometry optimization, and the rest of the system provided a static environment.

In this study, the ONIOM calculations do not support the recently proposed mechanisms for xcTDO (via either formation of the dioxetane intermediate or Criegee-type rearrangement) but suggest a rather unique mechanism in the hemes: (1) direct radical addition of a ferric-superoxide intermediate with C2 of the indole of Trp, followed by (2) ring-closure via homolytic O–O cleavage to give epoxide and ferryl-oxo (Cpd II) intermediates, (3) acid-catalyzed regiospecific ring-opening of the epoxide, (4) oxo-coupled electron transfer, and (5) finally, C–C bond cleavage concerted with back proton transfer (Fig. 10). Thus, QM/MM calculations have a potential predictive power, to guide future experimental research.

QM/MM MC simulations are also of a great importance in understanding the mechanisms of enzymatic reactions. For example, QM/MM MC was used to assess the mechanism and performance of artificial enzymes catalyzing Kemp elimination (Fig. 5) that was already mentioned [24]. In these simulations, the QM region was chosen to contain the substrate and the catalytic base up to Cα, and the QM level was semiempirical PDDG-PM3. The rest of the protein and hundreds of explicit solvent water molecules (the TIP4P water model) constituted the MM region treated with OPLS-AA force field (Fig. 11a). The system was extensively sampled with MC (on

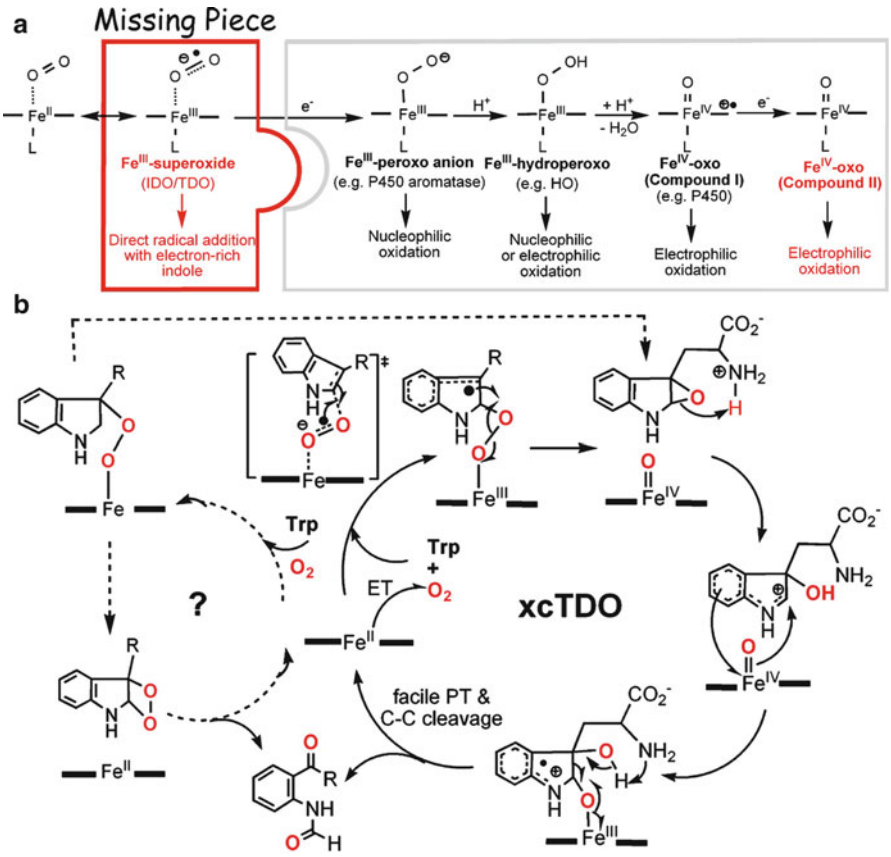

Fig. 10 (**a**) ONIOM calculations were instrumental in identifying the "missing piece" in the sequence of structures produced in the reaction catalyzed by tryptophan 2,3-dioxygenase. (**b**) The new proposed reaction mechanism that involves the "missing piece." Adapted with permission from [23]

the order of 30×10^6 configurations). The simulations were then coupled to the FEP methodology for driving the reaction along the reaction coordinates, which were chosen to be the N–O distance for the opening N–O bond of the substrate, and a combined reaction coordinate for H-transfer, as shown in Fig. 11b. The reaction was slowly driven along these coordinates, and at each point on the reaction path the system was equilibrated with MC. Thus, the free energy map for the reaction was generated (Fig. 11c). The mechanism of the catalyzed reaction is concerted, with the proton transfer occurring slightly earlier and driving the subsequent N–O bond opening. The overall ΔG of the catalyzed reaction is slightly negative. The activation barrier is 13.5 kcal/mol, which, when compared to that for the uncatalyzed reaction in water (ca. 20 kcal/mol), indicates a catalytic effect, in agreement with experimental results.

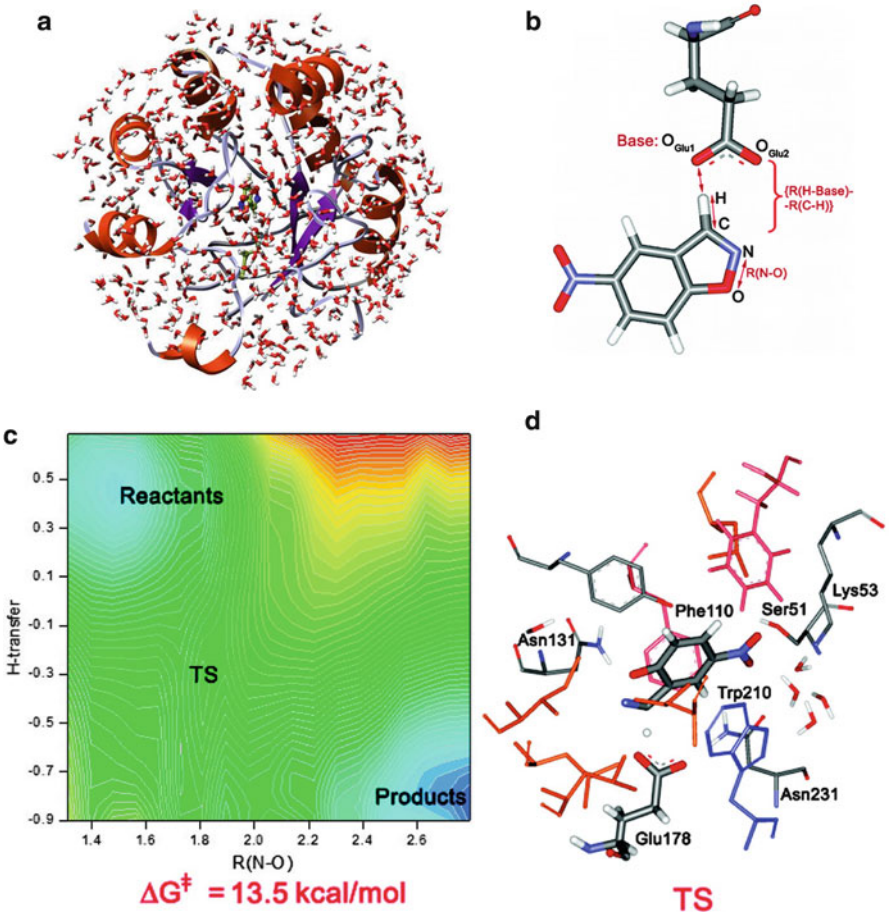

Fig. 11 QM/MM MC modeling of the Kemp elimination reaction catalyzed by an artificial enzyme: (**a**) the entire protein solvated with explicit water, prepared for simulations, (**b**) the choice for the QM region, (**c**) the free energy map, and (**d**) the snap-shot from the MC simulations taken in the region of the free energy surface corresponding to the TS. Adapted with permission from [24]

An additional feature of QM/MM FEP MC method is that one can look at the structure of the protein undergoing the reaction, at any point on the free energy profile. For example, in Fig. 11d, the representative structure of the binding site recorded in the region of the transition state is shown. One may see that the reaction proceeds smoothly, and all the parts of the protein play their designated roles. For example, Trp201 remains in the π-stacking orientation with respect to the substrate throughout the course of the reaction, and Asn131 forms a hydrogen bond to the O atom that acquires negative charge. This structural insight is unique to simulations, and cannot be attained experimentally. This is an important tool for

the analysis of enzymatic mechanisms, and, in the case of artificial enzymes, for the detection of potential structural problems with designs and providing recipes for their improvement and rescue.

The number of applications and variations on the theme of QM/MM is enormous, and the field still gains momentum, suggesting its even greater popularity in the near future. The basic principle of QM/MM is always the same: partitioning the system into the more chemically significant part and the rest, treat the two parts with theory of different accuracy and cost, and take good care of how the two subsystems communicate.

4 Excited States and Electron Detachment

4.1 Theoretical Foundation

Excited electronic states frequently occur in biology. They may form when molecules absorb ultra violet light coming from the Sun, for example. Electronic excitations may happen by promoting an electron from the highest unoccupied MO (HOMO), or from deeper occupied MOs to one of the bound unoccupied MOs. Hence, there is a whole spectrum of excited electronic states accessible to the molecule. Higher energy radiation may induce the photoelectric effect in biomolecules, i.e., electron detachment from one of the valence MOs to the continuum, yielding a photoelectron spectrum. Needless to say that electronic excitation and detachment energies characteristic of molecules can be used as spectroscopic probes for their structure and electronic properties. Furthermore, molecules excited to one of the excited PESs will evolve according to gradients for the nuclear motion characteristic of this PES. This evolution may lead to the formation of various photoproducts, sometimes irreversibly. This is relevant to photodamage of molecules such as, for example, DNA in our cells. Alternatively, the system may fluoresce, if it gets trapped on one of the "dark" exited states. Also, electronic excitations are the key to the catalytic activity of photoactivated enzymes.

A computational description of excited states requires special methods. The least computationally expensive applicable method is a variant of DFT, called time-dependent DFT, or TD-DFT. In TD-DFT, there is a time-dependent potential to which the system is exposed, and it is postulated that this potential uniquely maps onto the time-dependent electron density of the system [25]. TD-DFT is a linear response type of method, which is based on the assumption that the reference ground state is perturbed relatively little upon the presence of the time-dependent field, and so the ground state solutions can be used throughout. The poles in the response function correspond to excitation energies of the system.

The method works well for well-separated ground and excited PESs. However, being intrinsically single-configurational, TD-DFT cannot handle states that are

near-degenerate with the ground state. For example, if TD-DFT is used to scan the ground and excited PESs with the purpose of identifying the mechanism of nonradiative decay of an excited state, the area near the surface-crossing or seam will not be describable. Also, long-range charge-transfer excited states cannot be accurately calculated with TD-DFT, due to the lack of dispersion.

As a cure for the aforementioned problems with TD-DFT, recently, the constrained DFT method (C-DFT) was proposed [26]. The constraint enforces the system to stay in a fixed diabatic state, which is defined by the number of electrons, N_c, populating a particular site or group of atoms in the molecule:

$$\int \Omega(r)\rho(r)\mathrm{d}r = N_c. \tag{36}$$

$\Omega(r)$ is the weighting function that specifies a particular excited state of interest. For example, if a system without charge separation has a long range charge transfer excited state, this excited state would be described as a diabatic state with fixed charges of -1 on the acceptor and $+1$ on the donor of the electron. The total energy of the constrained state is then optimized:

$$E_c(\rho, N_c) = \min\left[E\{\rho(r)\}\right] + \lambda_c\left\{\int \Omega(r)\rho(r)\mathrm{d}r - N_c\right\}, \tag{37}$$

where the constraint is incorporated via an additional Lagrange multiplier, λ_c. The Kohn–Sham equations are solved simultaneously for the orbitals, eigenenergies, and λ_c:

$$\left[-\frac{1}{2}\nabla^2 + V(r) + \int \frac{\rho(r')}{|r-r'|}\mathrm{d}r' + U_{xc}(r) + \lambda_c\Omega(r)\right]\phi_k = \varepsilon_k\phi_k. \tag{38}$$

Thus, C-DFT is a ground state method, but for the state of the modified Hamiltonian that has the desired diabatic constraint incorporated. C-DFT is claimed to allow for more accurate calculations of excited state energies, if long-range charge transfer is involved. In addition, for the mechanisms of photoreactions, the crossing point between the diabatic states can be found, and that can serve for understanding the mechanism of photoreactions. However, in order to perform the dynamics on such surfaces, the coupling between the diabatic states has to be evaluated, which cannot be done rigorously, because again the two obtained states are technically eigenstates of different Hamiltonians. Nevertheless, C-DFT is a promising and inexpensive method; it is currently implemented in *NWChem*.

A more accurate method for the excited state calculations is Equation of Motion Coupled Cluster, EOM-CC [27], although it is quite expensive computationally, $O(N^6)$. EOM is based on a single HF Slater determinant as a reference function. It then finds the excited states by diagonalizing the similarity transformed Hamiltonian:

$$\bar{H} \equiv e^{-T}He^{T}, \tag{39}$$

where

$$\bar{H}R = ER \tag{40}$$

$$L\bar{H} = EL \tag{41}$$

$$L_I R_J = \delta_{IJ}, \tag{42}$$

and T, R, and L^+ are excitation operators with respect to the reference function, i.e., HF. T is usually truncated at single and double excitations (EOM-CCSD). R may or may not conserve the total number of electrons and unpaired spins in the system, thus providing access to radicals with different numbers of unpaired electrons. EOM is also a size-consistent method. The error bar for excitation energies computed with EOM-CCSD is on the order of 0.1–0.3 eV, which allows for the accurate interpretation of electronic spectra. EOM also has access to nearly and completely degenerate excited, Rydberg and valence, and mixed Rydberg-valence states, as long as HF is a good reference function. However, it is important to keep in mind that if the ground state becomes nearly degenerate with an excited state or states, EOM will not be able to handle it. Hence, again, conical intersections with the ground state and other such topographic entities are out of reach.

Finally, what can handle excited states, in both nondegenerate, and degenerate cases, is the multireference methods. The cheapest methods are CI, and state-averaged CASSCF (SA-CASSCF). In a regular CASSCF calculation, the set of active MOs is optimally chosen so as to improve the correlation effects on the ground state. If the same active space is then applied also to the excited states, their treatment will not be as good as that of the ground state. SA-CASSCF differs from regular CASSCF in that the active space is chosen to be an optimal compromise, so as to describe the averaged states with maximally similar accuracy. For more dynamic electron correlation, CASPT2 and CASMRCI may be used. Vertical electronic excitation energies, as well as locations of conical intersections can be calculated using multireference methods.

In order for an electronic transition to be observed in a spectrum, it has to have an appreciable cross section, i.e., transitional probability, which is defined as an integral over electronic degrees of freedom of the transition dipole moment operator sandwiched between the final and initial electronic states of the system. States that have a significant cross section are called “bright” states, and they can be significantly populated through photo-absorption. States that do not have large cross sections are so-called “dark” states. They can get populated primarily through the decay of the prepared bright states. The transitional probability is given through the computed oscillator strengths in calculations, such as TD-DFT and EOM-CCSD.

What is often of additional interest is optimization on excited PESs. At vertical electronic excitation, the system hits the excited PES in a Franck–Condon region (Fig. 12). The excited state PES in the Franck–Condon region has very different curvature than did the ground PES in the minimum, and in fact is rare that the minima of the ground and excited PES would coincide. Hence, after excitation, nuclei experience forces characteristic of the excited PES, and moving under

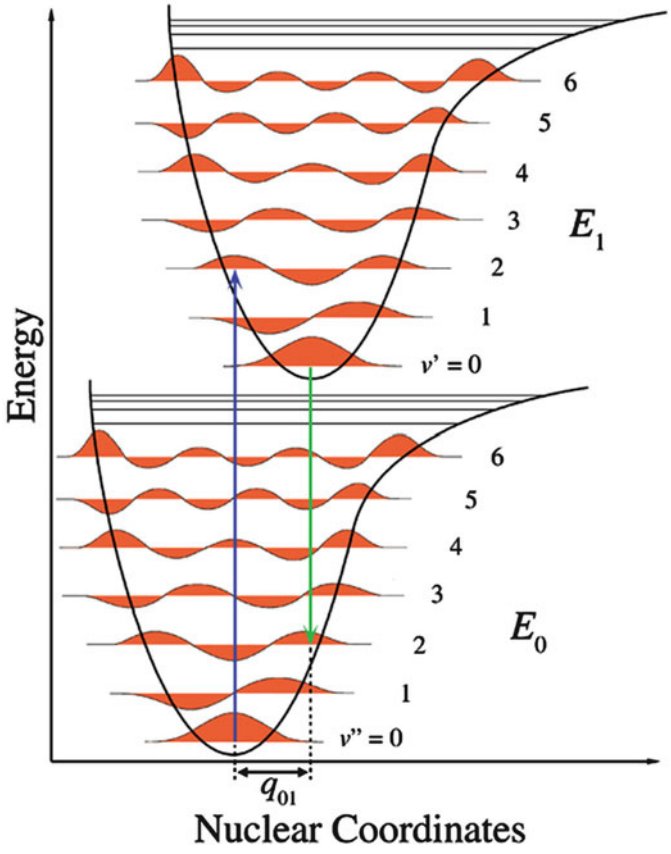

Fig. 12 Vertical electronic excitation from the ground PES (E_0) to the excited PES (E_1). The species hits the excited PES in the Franck–Condon region (vertical arrow pointing upward), located on the slope of the excited PES. Nuclear relaxation would take the system down to the minimum on the excited PES, and q_{01} is the difference between the geometry of the vertically excited species and the new geometry of the minimum on the excited PES

these forces relax to the nearest minimum on the excited PES, or to the region of crossing with another PES. Thus, knowing the position of stationary points on the excited PESs is important in mechanistic studies of photoreactions, and also in understanding and optimizing fluorescent compounds, such as photoprobes. Optimization on the excited PESs can be done with TD-DFT and multireference methods. However, again, if optimization is intended to find the crossing point between two surfaces, TD-DFT would fail, and only multireference methods should be used.

Obviously, the methodology used for excited state calculations can also be used for calculations of vertical electron detachments energies (VDEs). For this, one needs to calculate the first VDE, by subtracting the energies of the ionized

species and the initial species. Then, excited electronic states for the ionized species have to be found, and the excitation energies have to be added to the first VDE to obtain VDEs corresponding to the electron detachment from the MOs deeper than the HOMO. Calculated VDEs can be compared to experimental photoelectron spectra, and good agreement (within 0.1–0.2 eV) would be a structural probe for the molecule.

4.2 Examples of Applications

Ab initio calculations of excited states can be done with chemical accuracy, to explain or challenge existing experiments, and to make predictions. For example, ionization energies of aqueous nucleic acids have been calculated at TD-DFT and CASPT2 with implicit solvation, and compared to experimental values [28]. The lowest vertical ionization energies of aqueous cytidine and deoxythymidine were determined experimentally to be 8.3 eV, corresponding to an electron detaching from the base. Calculations were in quantitative agreement with the experiment. A dramatic effect of the aqueous environment was revealed by the ab initio computations. Namely, bulk water not only modestly lowers the ionization potential of the DNA bases but also makes it insensitive to the presence of sugar or phosphate. This is a very different situation from that observed in the gas phase, where the other DNA components (phosphate in particular) strongly influence the ionization process at the base.

Krylov and coworkers conducted a very detailed study of the electronic structure of the chromophore in the green fluorescent protein (GFP), and its changes upon one and two-electron oxidation [29]. The purpose of this work was to elucidate the mechanism of oxidative redding of the protein. The chromophore in GFP, 4′-hydroxybenzylidene-2,3-dimethylimidazolinone (HBDI),(Fig. 13a) is widely used in bioimaging. Structural changes upon oxidation were minor but sufficient to differentiate the species by their IR absorption spectra. MOs illustrating the electronic changes upon electron detachments are shown in Fig. 13b. Electronic states of relevant species were characterized by the SOS-CIS(D), multireference perturbation theory, and EOM-CCSD calculations, and results obtained with different methods were in a fairly good agreement. The one- and two-electron oxidation processes for deprotonated HBDI were considered. Figure 13c shows an overall energy diagram for the three considered forms of HBDI. Adiabatically, the doubly oxidized form is 9.98 eV above the ground state of the anion. The respective value of VDE corresponding to removing two electrons is 10.37 eV. The adiabatic ionization energy from the ground state of the doublet radical is 7.59 eV computed using the anion's VDE value of 2.54 and 0.15 eV relaxation energy of the neutral radical. Another relevant value is the energy gap between the excited states D_1 and D_2 of the doublet radical and the cation. Using the same values of vertical detachment and relaxation energies, and 1.52 and 3.37 eV for the vertical $D_0 \rightarrow D_{1,2}$ excitation energies, the authors estimated the ionization energy of the electronically excited

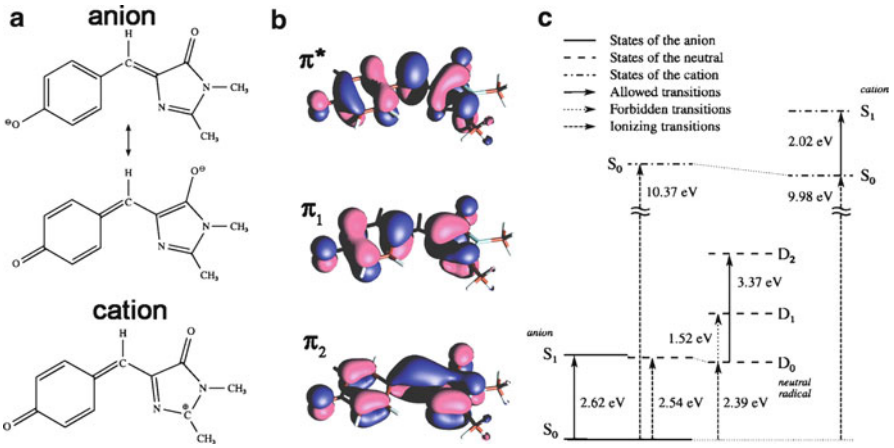

Fig. 13 (**a**) The deprotonated chromophore from the GFP in the anonic and cationic forms. (**b**) Relevant molecular orbitals of deprotonated HBDI (HF/6–311G*). In the ground state of the anion, both π_1 and π_2 are doubly occupied, and the bright state is derived by the $\pi_1 \rightarrow \pi^*$ excitation. Oxidized forms are derived by removing the electrons from π_1. (**c**) Energy diagram for the relevant electronic states of deprotonated HBDI. Adapted with permission from [29]

doublet radical as 6.07 and 4.22 eV, respectively. Finally, the energy gap between the excited singlet state of the anion and the two excited states of the radical are 1.29 and 3.14 eV, respectively. The results suggested that the doubly oxidized species (the deprotonated HBDI cation) may be responsible for the oxidative redding through the following newly proposed mechanism: the first step involves photoexcitation, and the blue light is sufficient to generate this transition. The second and third steps are one-electron oxidation steps. The closed-shell character of the cation is consistent with the relatively chemically stable nature of the red form of GFP. The absorption in the cation (product of two-electron oxidation) is red-shifted by almost 0.6 eV with respect to the anion, and the resulting value of 2.02 eV is in good agreement with the experimental excitation energy of 2.12 eV. The redshift is consistent with the electronic excitation in the cation being the $\pi_1 \rightarrow \pi_2$ process rather than $\pi_1 \rightarrow \pi^*$ in the anion. This mechanism of redding is distinctly different from previously characterized ones in which redding was achieved by extending the π-system of the chromophore.

Excited states in DNA fragments are important for understanding the mechanism of DNA photodamage, and self-rescue due to internal conversion from electronically excited states. Calculations of high accuracy are expensive, as was eluded. Hence, only small fragments, such as one or two nucleic bases can be characterized using high levels of ab intio theory. Roos and coworkers sophisticatedly explicated the electronic spectra of nucleic base monomers [30]. The CASSCF and CASPT2 methods were used. The comparison with the experimental measurements speaks for the stellar qualities of the chosen methodology. For example, for N(9)H-adenine, the computed valence $\pi \rightarrow \pi^*$ excitation energies are 5.1, 5.2 (4.9), 6.2 (5.7–6.1),

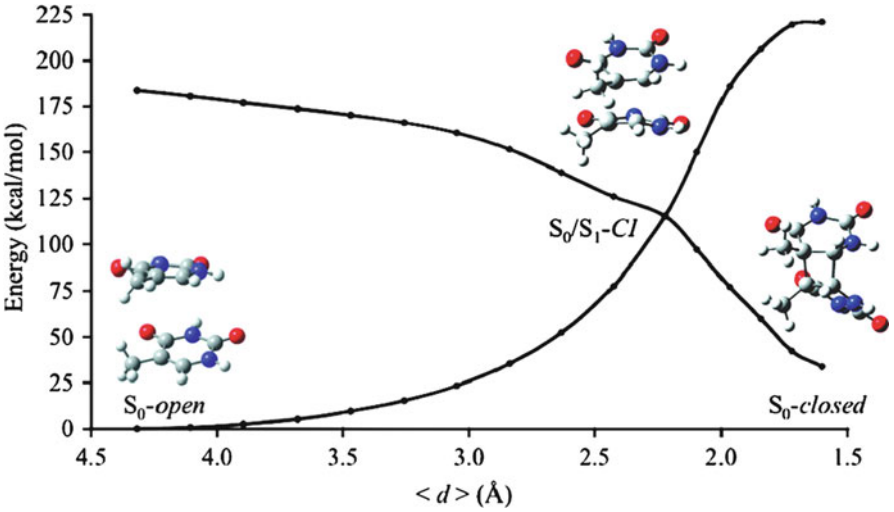

Fig. 14 The ground and excited PESs for the π-stack of two thymine bases, calculated at the CASSCF(12,12)/6–31G* level. The conical intersection is precisely identified, explicating the mechanism of the formation of thymine dimer as a [2 + 2] cycloaddition mechanism. Adapted with permission from [31]

6.7, 7.0 (6.8), 7.6 (7.7) eV, where the numbers in parentheses represent experimental data. For guanine, the numbers are 4.7 (4.5–4.8), 5.1 (4.9–5.0), 6.0 (5.5–5.8), 6.5 (6.0–6.4), 6.6, 6.7 (6.6–6.7), and 6.7 eV. Intensities of the transitions were also predicted.

Robb and coworkers [31] investigated the formation of the thymine dimer, which is a known route of mutagenesis in DNA. The dimer forms between two π-stacking thymine residues that neighbor each other in a strand of DNA. The repair of this photoproduct requires a photoinitiated enzyme, photolyase, so it is a serious type of DNA damage. The process starts from DNA adsorbing a UV photon, and an electronic excitation from the S_0 state to S_1. The scans of the ground and excited PESs were performed at the CASSCF(12,12)/6–31G* and CASPT2/cc-pVDZ levels of theory. In the scans, the chosen reaction coordinate was fixed, and gradually incremented from point to point, and the rest of the internal degrees of freedom were optimized, both on the ground and on the excited PESs. It was found that the molecule on the S_1 PES rapidly evolves, reaching the S_1/S_0 conical intersection (Fig. 14). At the conical intersection it proceeds further along the reaction coordinate toward the new minimum on the ground PES, which is the dimer. The supplementary ground state calculations for this process indicated that it is highly unlikely to proceed adiabatically, without the involvement of the excited state.

When considering photochemistry of small fragments that model larger biological systems, one must remember that the locations of conical intersections can be heavily impacted by the surrounding environment, such as solvent, or the

rest of the biomolecule, whether it is a DNA double helix or a protein. This was demonstrated, for example, in the work by Yamazaki and Kato [32], on nonradiative relaxation and internal conversion of 9H-adenine. The considered photoprocess included an electronic excitation on an isolated 9H-adenine, in the gas phase, water, and acetonitrile. The PESs were explicated with the SA-CASSCF calculations. The equilibrium geometries of the ground and excited ($^1\pi\pi^*$ ^1L_a and ^1L_b, $^1n\pi^*$, and $^1\pi\sigma^*$) states and the conical intersections between them have been found. The dynamic electronic correlation was then added to the found points on the PESs, via multireference perturbation theory. The relative energies of the excited states were found to shift in aqueous solution as compared to the gas phase. As a result, the preferred mechanism of decay changes. Specifically, in water, the ^1L_a and ^1L_b states are very close in energy, and fast to interconvert. The $^1L_a/S_0$ conical intersection is the dominant decay pathway. This conical intersection involves the puckering of the six-membered ring of the 9H-denine molecule. The $^1n\pi^*$ and $^1\pi\sigma^*$ states are pushed higher in energy, and the decay pathway through these conical intersections becomes unfeasible. So in solution, there is no slow component of the decay due to the $^1n\pi^*$ and $^1\pi\sigma^*$ states, unlike in the gas phase.

In order to get the full mechanistic information about photochemistry involving multiple PESs, nonadiabatic dynamics simulations are required. Unlike usual ground state MD, nonadiabatic dynamics must have a possibility for the nuclei to jump from one PES to another near conical intersections and seams. Such simulations are costly and algorithmically more complex. Dynamics is considered in the next section.

5 Ground and Excited States Dynamics

5.1 Theoretical Foundation

The true insight into the mechanisms of reactions, branching ratios, life-times of intermediates, etc. can be gained only from the dynamics simulations. Dynamics may happen on a single PES, in which case it is called adiabatic dynamics. If dynamics happens on multiple PESs, it is called nonadiabatic dynamics, since the adiabatic BAO breaks down in this case.

In reality, both electrons and nuclei are quantum particles, and the most proper approach to dynamics would be to propagate both as quantum objects. The nuclear wave function, $\Omega(R,t)$, is an eigenfunction of the nuclear time-dependent Schrödinger equation:

$$i\hbar\frac{\partial}{\partial t}\Omega(R,t) \cong \left[-\sum_a \frac{\hbar^2}{2M_a}\nabla^2_{R_a} + E(R)\right]\Omega(R,t). \tag{43}$$

In quantum nuclear dynamics, $\Omega(R,t)$ can be represented as a Gaussian wavepacket, which is the object that needs to be propagated in the dynamics:

$$\Omega(R,t+\delta t) \approx \Omega(R,t) + \delta t \frac{\partial}{\partial t}\Omega(R,t) + \frac{(\delta t)^2}{2}\frac{\partial^2}{\partial t^2 \Omega(R,r)} + \ldots$$
$$\approx \Omega(R,t) - i\frac{\delta t}{\hbar}\left[-\sum_a \frac{\hbar^2}{2M_a}\nabla^2_{R_a} + E(R)\right]\Omega(R,t), \qquad (44)$$

However, this kind of dynamics for the entire biological molecule is prohibitively expensive. Usually, nuclei are considered classical objects, as long as the dynamics involves only one PES at a time. As a compromise, there is also mixed quantum-classical nuclear dynamics, where only some of the nuclei in the system are treated as quantum particles (i.e., delocalized wave-packets).

We will first consider the ab initio adiabatic dynamics in which nuclei are treated as classical particles moving on a single adiabatic Born–Oppenheimer PES, one PES at a time. This kind of dynamics also can be approximated by the classical force field formalism, if the force field parameters are available for the system. In ab initio adiabatic dynamics, the Schrödinger equation for the electrons is solved with nuclear coordinates being constant parameters. The nuclei are then exposed to the potential provided by the electrons, and move according to this ab initio potential, obeying the classical equations of motion. The forces acting on nuclei are calculated on-the-fly, at every step of the dynamics by solving the electronic Schrödinger equation, and utilizing the Hellmann–Feynman theorem:

$$F_{X_n} = -\frac{\partial E}{\partial X_n} = -\left\langle \Psi \left| \frac{\partial \hat{H}}{\partial X_n} \right| \Psi \right\rangle. \qquad (45)$$

The ab intio PES may be fitted to an analytical function, in order to solve for the forces. This is the foundation of the Born–Oppenheimer molecular dynamics (BOMD) method [33, 34], implemented in *Gaussian*, and also in the stand-alone classical trajectory program *VENUS*. The typical time-step in BOMD is on the order of a few femtoseconds. Adiabatic molecular dynamics can be coupled with classical treatment of the larger system surrounding the quantum region, in a QM/MM fashion.

Car-Parinello molecular dynamics (CPDM) is another variant [35]. In contrast to BOMD, CPMD explicitly introduces the electronic degrees of freedom (usually provided by DFT) as fictitious dynamical variables. In other words, the Kohn-Sham molecular orbitals are chosen as the dynamical variables to represent the electronic degrees of freedom in the system. Electrons are also assigned a fictitious mass. An extended Lagrangian for the system is then written, and it leads to a system of coupled equations of motion for both nuclei and electrons. The method works in conjunction with DFT and plane-wave basis set. CPMD is computationally expensive, and most likely not usable for sizable systems that might interest biochemists. There is also a cheaper version of CPMD, based on atom-centered density matrix propagation, ADMP, implemented in *Gaussian*. ADMP uses Gaussian basis functions, and works with semiempirical, HF, and pure and hybrid DFT methods.

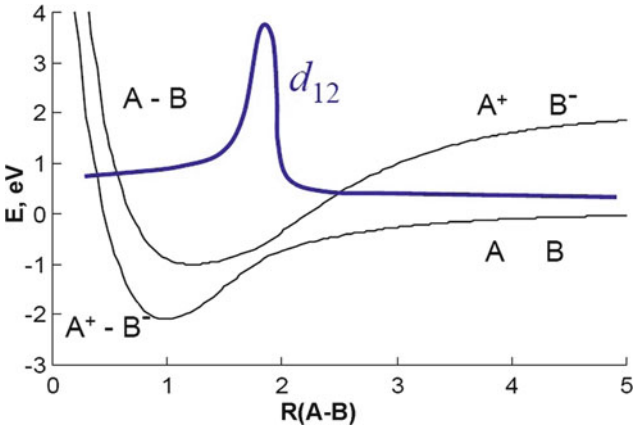

Fig. 15 The onset of nonadiabatic coupling between the two PESs. The present example illustrates the bonding situation in an ionic molecule, like NaCl: at short interatomic distances, the molecules exists as A^+B^-, and the state in which the electron transfer from A to B does not happen is an excited state laying higher in energy; at larger interactomic distances and the dissociation limit, the two atoms are neutral on the ground PES, and the charge transfer state is an excited state. Thus, at some interatomic distance, there is a change of the nature of the ground state. For a diatomic molecule, the region of this change is an avoided crossing. Near this region, the coupling, d_{12}, is the strongest

Adiabatic dynamics works fine, as long as the system is on a single PES, well-isolated in its energy from other PESs. If, however, two or more PESs come close together, they couple to each other, nuclei may hop from one PES to another, the BAO breaks down, and nonadiabatic dynamics is required. A schematic representation of a situation where nonadiabatic dynamics is needed is given in Fig. 15. There are two relevant PESs. At a certain internuclear distance, they get close to each other and exhibit what is called an avoided-crossing. This is a result of the noncrossing rule, valid only for systems with one degree of freedom (diatomic molecules). In systems having more degrees of freedom, the geometry of this area can look like a conical intersection, or a multi-dimensional "seam." For N degrees of freedom in a molecule, the dimensionality of the seam is $N - 2$. The two PESs are coupled. The nonadiabatic coupling, $\mathbf{d}_{12}$, in Fig. 15 is a function of R(A–B). Qualitatively, the evolution of $\mathbf{d}_{12}$ with R(A–B) is indicative of that in most areas on the PESs, the adiabatic approximation works fine, since the nonadiabatic coupling is negligible, but near avoided crossing the coupling rapidly becomes large. Here, we are concerned with the dynamics in the entire space, including the areas of large coupling.

First, let's notice that the total wave function in a nonadiabatic situation is no longer the product of electronic and nuclear parts, but a sum of such products over all accessible electronic states labeled with i:

$$\Psi(r, R) = \sum_i \psi_i(r; R)\Omega_i(R). \tag{46}$$

Substitution of this expression into the time-dependent Schrödinger equation, and integrating over the electronic degrees of freedom yields:

$$-\frac{\hbar^2}{2}\sum_a M_a^{-1}\nabla_{R_a}^2\Omega_j(R)+E_j(R)\Omega_j(R)$$
$$=-\frac{\hbar^2}{2}\sum_i D_{ij}(R)\Omega_i(R)+\hbar^2\sum_{i\neq j}\boldsymbol{d}_{ij}(R)\cdot\nabla_{R_a}\Omega_i(R), \tag{47}$$

where nonadiabatic couplings, D_{ij} and $\mathbf{d}_{ij}$, the components of the full nonadiabatic coupling, are introduced:

$$\boldsymbol{d}_{ij}(R)=-\sum_a M_a^{-1}\int\left\{\psi_i^*(r,R)[\nabla_{R_a}\psi_j(r,R)]\right\}\mathrm{d}r, \tag{48}$$

and

$$D_{ij}(R)=-\sum_a M_a^{-1}\int\left\{\psi_i^*(r,R)\left[\nabla_{R_a}^2\psi_j(r,R)\right]\right\}\mathrm{d}r. \tag{49}$$

Nonadiabaticity means that nuclear motions are capable of causing electronic transitions, and in turn electronic degrees of freedom determine the quantum states. In other words, electronic and nuclear degrees of freedom are no longer separable. There are two representation in which nonadiabatic dynamics simulations can be conducted. First is the adiabatic picture, where the two coupled PESs are adiabatic electronic states, i.e., the Hamiltonian matrix is diagonal, and nuclear coupling terms are zero. The electronic coupling in this case is a vector. The other representation is diabatic. Diabatic states do not diagonalize the Hamiltonian, and the nuclear coupling is a finite number. However, the electronic coupling is strictly zero, by definition. If all quantum effects on electrons and nuclei are properly taken into account, the two representations should give the same result. However, in practice, this is not the case, and usually the adiabatic picture is employed, and algorithms such as surface-hopping are developed for the adiabatic situation.

There are two major quantum-classical approaches to nonadiabatic dynamics. One is Ehrenfest dynamics, where nuclei move on an average PES between the two coupled states. In this way nonadiabatic effects are taken into account. This is a single configuration method, and when the two PESs diverge in energy, the average path has no meaning, and Ehrenfest dynamics is capable of unphysical predictions.

The other algorithm of enormous popularity is already mentioned surface hopping, and its "fewest switches" incarnation in particular [36]. In this approach, nuclei are allowed to instantaneously hop from one PES to another, with a certain probability, when the coupling is strong. After the hop, the component of velocity in the direction of the nonadiabatic coupling vector is adjusted to conserve energy. In the areas far from the seams, nuclei evolve classically, on single PES at a time. Electrons are propagated according to the time-dependent Schrödinger equation,

and the electronic problem should be solved with a multireference method, such as SA-CASSCF, CASPT2, or CASMRCI. The surface-hopping method works really well, though its shortcomings should be mentioned: All surface-hopping trajectories are independent, which is fundamentally different from the propagation of a wave-packet. The hops are a bit too drastic and require a sudden change in velocity, whereas in reality they should move on some effective PESs and the transitions should be smoother. Sometimes hops become forbidden because the system does not have enough energy to hop. Decoherence is not taken into account in surface-hopping, though may be added ad hoc. The surface-hopping algorithm is defined in the context of the adabatic representation of the electronic sates.

Occasionally, additional quantum mechanical treatment for selected nuclei, such as H^+ and hydride, can be added to the dynamics. The selected nucleus or nuclei are represented with vibrational wave functions, which are then propagated in the dynamics as quantum objects. Here, we will not consider these algorithms any further.

Another method of nonadiabatic dynamics was developed by Martínez and coworkers, and is called ab initio multiple spawning (AIMS) [37, 38]. AIMS allows for semiclassical nonadiabatic dynamics simulations, and includes the branching between bifurcating nonadiabatic paths. The method solves the electronic and nuclear Schrödinger equations simultaneously, including all molecular degrees of freedom. The total wave function is expressed as a linear combination of time-dependent, frozen, and localized in phase-space Gaussian basis functions:

$$\psi(R,r,t) = \sum_I \sum_i^{N_I(t)} c_i^I(t)\chi_i^I\left(R;\overline{R}_i^I,\overline{P}_i^I,\gamma_i^I\right)\varphi_I(r;R), \tag{50}$$

where I labels electronic states, $N_I(t)$ is the number of nuclear basis functions associated with Ith electronic state, which may adaptively expand during the dynamics in the areas where potential energies get close and the population may bifurcate, r and R are electronic and nuclear coordinates. The nuclear basis functions χ_i^I are multidimensional products of complex Gaussians and parameterized by their average positions and momenta, $\overline{R}_i^I$ and $\overline{P}_i^I$, as well as a semiclassical phase factor, γ_i^I. The average positions and momenta evolve according to Hamilton's equations, and the semiclassical phase factor evolves as the time integral of the classical Lagrangian. The electronic basis functions, $\phi_I(r;R)$, are defined as solutions of the electronic Schrödinger equation in the adiabatic representation at the nuclear geometry given by R. The complex coefficients, $c_i^I(t)$, evolve according to the time-dependent nuclear Schrödinger equation in the time-evolving basis set, which is solved simultaneously with the equations of motion for $\overline{R}_i^I$, $\overline{P}_i^I$, and γ_i^I:

$$\sum_{kK} S_{jk}^{JK}\dot{c}_k^K = -i\sum_{kK}\left(H_{jk}^{JK} - i\dot{\vec{S}}_{jk}^{JK}\right)c_k^K. \tag{51}$$

The overlap, right-acting time derivative, and Hamiltonian matrix elements in this equation are:

$$S_{jk}^{JK} = \left\langle \chi_j^J \phi_J \middle| \chi_k^K \phi_J \right\rangle \delta_{JK} \tag{52}$$

$$\dot{\tilde{S}}_{jk}^{JK} = \left\langle \chi_j^J \phi_J \middle| \frac{\partial \chi_k^K}{\partial t} \phi_K \right\rangle \delta_{JK} \tag{53}$$

$$H_{jk}^{JK} = \left\langle \chi_j^J \phi_J \middle| \hat{H} \middle| \chi_k^K \phi_K \right\rangle \tag{54}$$

$\widehat{H}$ here is the full molecular Hamiltonian operator including the nuclear kinetic energy, the electronic PES, and nonadiabatic coupling terms. The electronic Schrödinger equation is solved simultaneously with the nuclear dynamics to obtain the PESs and couplings. AIMS on a small fragment can be coupled to purely classical treatment of a large biological molecule surrounding the fragment, i.e., in QM/MM.

In general, nonadiabatic dynamics is computationally expensive, algorithmically complex, but unavoidable for certain kinds of problems in biomolecular simulations. It is important to be able to recognize when nonadiabatic dynamics would play a role in a process. First of all, of course, processes involving excited electronic states are likely candidates. Even fluorescing systems may exhibit excited state population leaks through nonradiative internal conversion. Also, in principle, any time a system crosses an activation barrier, chances are that at the transition state the two surfaces come close enough for nonadiabatic coupling to become significant. A quick check for the adiabaticity of the process is to run a CASSCF calculation and assess the contribution of different states to the CAS expansion.

5.2 Examples of Applications

There are myriads of reported dynamics studies, especially employing ab initio classical adiabatic dynamics. Here, we will highlight some exciting studies that employ more complicated simulation engines, and account for nuclear quantum effects, or go beyond the BAO.

QM/MM adiabatic dynamics simulations were performed on the thymine dimer radical anion splitting in the photoactivated self-repair process in DNA [39]. The simulations were done in explicit water. The QM region was treated with DFT, and the MM region was treated with the AMBER force field. The calculations revealed that the upper-bound of the free energy barrier to this process is 2.5 kcal/mol. The mechanism was found to be asynchronous, with one bond in the dimer breaking earlier in the process (Fig. 16). This work, along with experimental studies contributed to the general appreciation of the general stability and ability to self-repair of the natural DNA structure.

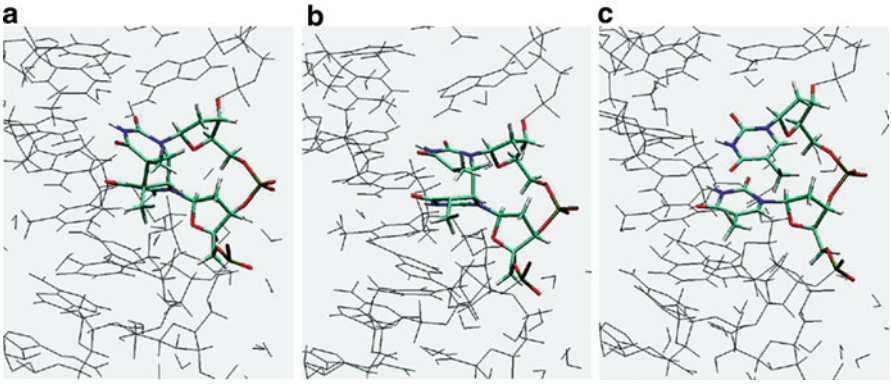

Fig. 16 Snap-shots from QM/MM adiabatic dynamics simulations for thymine dimer radical splitting, and showing the asynchronous mechanism. Adapted with permission from [39]

Nonadiabatic dynamics simulations have been successful, both in the gas phase and in the context of larger biomolecules. Groenoff et al. studied the reaction of the ultrafast deactivation of an excited Cytosine–Guanine (CG) base pair in the DNA double helix consisting of 22 bp [40]. The revealed mechanism is again a demonstration of the apparent capability of DNA to self-rescue, through nonradiative internal conversion from the excited state back to the ground state. The QM region included the two interacting bases. The rest of the DNA molecule and the solvent were treated classically. At the beginning of each nonadiabatic QM/MM MD simulation, the CG pair in its equilibrium ground state geometry was excited from the ground state, S_0, to the first excited state, S_1, and then allowed to evolve according to the gradients provided by S_1. No reaction coordinate was chosen prior to the simulations, and all the forces were determined on-the-fly, during the dynamics. The electronic part of the problem was solved with the CASSCF(2,2) method (implemented in *Gaussian*), and nuclei moved classically on a single PES at a time, as long as they were far away from conical intersections. The MM region evolved via classical MD, as implemented in *GROMACS*. In the areas of strong nonadiabatic coupling, the fewest switches surface hopping algorithm was employed. It appears that the S_1 state is a charge-transfer state, where an electron hops from G to C. As a result of this charge separation, the negatively charged C attracts the proton from the N1 atom of G, and the S_1 PES is repulsive with respect to the H-shuttling motion between the N1 atom in G and the N3 atom in C. Hence, the system on S_1 evolves along this H-transfer coordinate, and the motion is also accompanied by some skeletal deformation of the system. On the S_0 PES, the region corresponding to H being transferred from G to C is high in energy, and so the S_0 and S_1 surfaces cross in this area of configurational space. When nuclei reach this point on the S_1 PES, they transfer to S_0, i.e., the charge hops back from C to G, and the N1 atom on G again becomes attractive to the transferred H. Following the gradients on the S_0 surface, H returns to G, and the system is back in its original

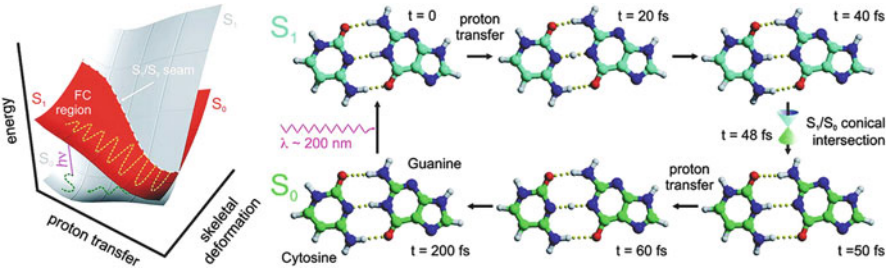

Fig. 17 Nonradiative decay for the CG base pair within the DNA double helix: Potential energy surfaces of the excited and ground states mapped along the proton transfer (N_1–H–N_3) coordinate and the skeletal deformation. The *dashed yellow* and *green lines* represent the path sampled in a typical trajectory. Nonradiative decay occurs along the intersection seam between the surfaces. Relevant structures along the dynamics path are shown at right. Adapted with permission from [40]

equilibrium ground state geometry (Fig. 17). Thus, the H-bonding in the Watson–Crick pair appears to play an important role in the DNA self-protection from the UV damage.

A nonadiabatic dynamics study describing the nonradiative decay in the gas phase cytosine base photoexcited to the bright S_1 state has been reported by Hudock and Martínez [41]. The method in use was AIMS. The electronic structure in the problem was determined on-the-fly, using the SA-CASSCF(2,2) method. The nuclear dynamics was described by the frozen Gaussian trajectory basis functions. In the regions of configurational space where nonadiabatic coupling was large, the nuclear basis set had a chance to adaptively expand: additional basis functions were "spawned" to describe the population transfer between the states involved. The analysis of trajectories showed that all of the known competitive paths of internal conversion in cytosine can be adopted on different time-scales (Fig. 18). This complex picture is different from those in thymine and uracil, which are well characterized by a single mechanism.

6 Summary

Electronic structure plays a critical role in all biological processes. We covered a variety of electronic structure techniques and their applications to biological systems. For molecules on the ground state, various properties can be calculated based on electronic structure methods, for example electronic configuration, IR spectra, polarizability, dipole moment, magnetic shielding, etc. There are many available ab intio techniques, and we emphasized their applicability, benefits, and potential pitfalls. Pure ab initio calculations are typically performed on small molecules or fragments of larger biological molecules. However, they can be coupled to the rest of the biomolecule and solvent in mixed QM/MM formalism, which is a very

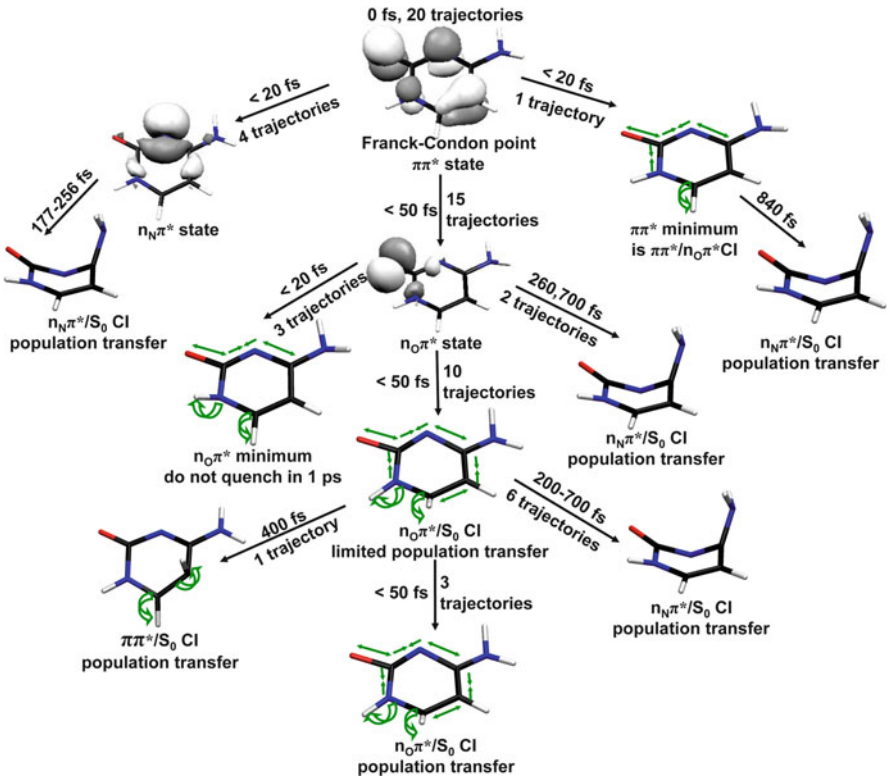

Fig. 18 Summary of pathways of excited cytosine. The *arrows* indicate the passage between different electronic states, with the observed time lapse and the number of trajectories shown. The S_1 electronic state passes through regions with three different characters ($\pi\pi^*$, $n_O\pi^*$, and $n_N\pi^*$), and electronic-state quenching to S_0 occurs through three different conical intersections. Adapted with permission from [41]

popular approach. QM/MM has many different versions. Some of them are static, like ONIOM, and some include statistical mechanical sampling and dynamics. QM/MM calculations are instrumental in the assessment of reaction mechanisms in complex biological molecules in solution. We further considered the treatment of excited electronic states and electronic detachment energies with chemical accuracy. Excited states are involved in photoprocesses characteristic of biomolecules. Also, electronic spectroscopy provides an invaluable means of analysis of biomolecules. Calculations of this kind require methods such as TD-DFT, EOM-CC, and multireference methods. Using these techniques allows for the prediction of electronic spectra within 0.1–0.2 eV from the experimental values. Finally, dynamical simulations were considered as the ultimate approach to gain mechanistic insights into biochemical processes. Dynamical simulations may operate on a single PES (adiabatic dynamics), or on multiple PESs (nonadiabatic dynamics). The techniques

required to properly treat nonadiabatic dynamics were presented in detail. Many of the described methods are implemented in commercially available packages. Some of the methods are less standard, and not out out-the-box. It is also still true and probably will be true for a while that quantum mechanical methods cannot be used as a "black box." One should approach every problem with a good understanding of the electronic nature of the problem, and an idea of which methods should and should not work in each particular case. One also needs to experiment with the system, in order to discover any possible caveats, such as strong multiconfigurational nature of the wave function, or unusual phenomena such as long range charge transfer, or nonadiabatic character of the dynamics. The author hopes that this chapter introduces the main concepts that would enable the Reader to make intelligent choices when using quantum mechanical methods.

References

1. Møller, C., Plesset, M.S.: Note on an approximation treatment for many-electron systems. Phys. Rev. **46**, 0618–22 (1934)
2. Cizek, J.: (1969) In: Hariharan P.C. (ed.) Advances in Chemical Physics, vol. 14, Wiley Interscience, New York. http://www.gaussian.com/g_tech/g_ur/refs.htm
3. Purvis, I.I.I.G.D., Bartlett, R.J.: A full coupled-cluster singles and doubles model—the inclusion of disconnected triples. J. Chem. Phys. **76**, 1910–18 (1982)
4. Pople, J.A., Head-Gordon, M., Raghavachari, K.: Quadratic configuration interaction — a general technique for determining electron correlation energies. J. Chem. Phys. **87**, 5968–75 (1987)
5. Foresman, J.B., Head-Gordon, M., Pople, J.A., Frisch, M.J.: Toward a systematic molecular orbital theory for excited states. J. Phys. Chem. **96**, 135–49 (1992)
6. Pople, J.A., Seeger, R., Krishnan, R.: Variational Configuration Interaction Methods and Comparison with Perturbation Theory. Int. J. Quantum. Chem. Suppl. **Y-11**, 149–63 (1977)
7. Hegarty, D., Robb, M.A.: Application of unitary group-methods to configuration-interaction calculations. Mol. Phys. **38**, 1795–812 (1979)
8. Andersson, K., Malmqvist, P.A., Roos, B.O.: Second-order perturbation theory with a complete active space self-consistent field reference function. J. Chem. Phys. **96**, 1218 (1992)
9. Chen, H., Lai, W., Shaik, S.: Multireference and multiconfiguration ab initio methods in heme-related systems: what have we learned so far? J. Phys. Chem B. **115**, 1727–1742 (2011)
10. Knowles, P.J., Werner, H.J.: An efficient method for the evaluation of coupling coefficients in configuration interaction calculations. Chem. Phys. Lett. **145**, 514–522 (1988)
11. Werner, H.J., Knowles, P.J.: An efficient internally contracted multiconfiguration-reference configuration interaction method. J. Chem. Phys. **89**, 5803 (1988)
12. Evangelista, F.A., Allen, W.D., Schaefer, H.F.: High-order excitations in state-universal and statespecific multireference coupled cluster theories: model systems. J. Chem. Phys. **125**, 154113 (2006)
13. Evangelista, F.A., Allen, W.D., Schaefer, H.F.: Coupling term derivation and general implementation of state-specific multireference coupled cluster theories. J. Chem. Phys. 127, **024102** (2007)
14. Parr, R.G., Yang, W.: Density-Functional Theory of Atoms and Molecules. New York, Oxford University Press (1989)

15. Tomasi, J., Mennucci, B., Cammi, R.: Quantum mechanical continuum solvation models. Chem. Rev. **105**, 2999–3093 (2005)
16. Suárez, D., Díaz, N., Merz, K.M. Jr: Ureases: quantum chemical calculations on cluster models. J. Am. Chem. Soc. **125**, 15324–15337 (2003)
17. Jensen, K.P., Bell, I.I.I.C.B., Clay, M.D., Solomon, E.I.: Peroxo-type intermediates in class i ribonucleotide reductase and related binuclear non-heme iron enzymes. J. Am. Chem. Soc. **131**, 12155–12171 (2009)
18. Rothlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., Dechancie, J., Betker, J., Gallaher, J.L., Althoff, E., Zanghellini, A.A., Dym, O., Albeck, S., Houk, K.N., Tawfik, D.S., Baker, D.: Kemp elimination catalysts by computational enzyme design. Nature. **453**, 109–195, (2008)
19. Senn, H.M., Thiel, W.: QM/MM methods for biomolecular systems. Angew. Chem. Int. Ed. **48**, 1198–1229 (2009)
20. Warshel, A., Levitt, M.: Theoretical studies of enzymatic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J. Mol. Biol. **103**, 227–249 (1976)
21. Froese, R.D.J., Morokuma, K.: (1998) Hybrid methods. In: P.V.R. Schleyer (ed.) Encyclopedia of Computational Chemistry, vol. 2 Wiley, Chichester
22. Jorgensen, W.L., Tirado-Rives, J.: Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. J. Comput. Chem. **26**, 1689–1700 (2005)
23. Chung, L.W., Li, X., Sugimoto, H., Shiro, Y., Morokuma, K.: ONIOM study on a missing piece in our understanding of heme chemistry: bacterial tryptophan 2,3-dioxygenase with dual oxidants. J. Am. Chem. Soc. **132**, 11993–12005 (2010)
24. Alexandrova, A.N., Rothlisberger, D., Baker, D., Jorgensen, W.L.: Catalytic mechanism and performance of computationally designed enzymes for kemp elimination. J. Am. Chem. Soc. **130**, 15907–15915 (2008)
25. Marques, M.A.L., Ullrich, C.A., Nogueira, F., Rubio, A., Burke, K., Gross, E.K.U.(eds.) Time-Dependent Density Functional Theory. Springer, Verlag (2006)
26. Wu, Q., Van Voorhis, T.: Direct optimization method to study constrained systems within dnsityfuncitonal theory. Phys. Rev. A. **72**, 024502-1–024502-4 (2005)
27. Krylov, A.I.: Equation-of-motion coupled-cluster methods for open-shell and electronically excited species: the hitchhiker's guide to fock space. Ann. Rev. Phys. Chem. **59**, 433–462, (2008)
28. Slavíček, P., Winter, B., Faubel, M., Sbadforth, S.E., Jungwirth, P.: Ionization energies of aqueous nucleic acids: photoelectron spectroscopy of pyrimidine nucleosides and ab initio calculations. J. Am. Chem. Soc. **131**, 6460–6467 (2009)
29. Epifanovsky, E., Polyakov, I., Grigorenko, B., Nemukhin, A., Krylov, A.I.: The effect of oxidation on the electronic structure of the green fluorescent protein chromophore. J. Chem. Phys. **132**, 115104 (2010)
30. Fülscher, M.P., Serrano-Andrés, L., Roos, B.O.: A theoretical study of the electronic spectra of adenine and guanine. J. Am. Chem. Soc. **119**, 6168–6176 (1997)
31. Boggio-Pasqua, M., Groenhof, G., Schäfer, L.V., Grubmüller, H., Robb, M.A.: Ultrafast deactivation channel for thymine dimerization. J. Am. Chem. Soc. **129**, 10996–10997 (2007)
32. Yamazaki, S., Kato, S.: Solvent effect on conical intersections in excited-state 9H-adenine: radiationless decay mechanism in polar solvent. J. Am. Chem. Sos. **129**, 2901–2909 (2007)
33. Bunker, D.L.: Classical trajectory methods. Meth. Comp. Phys. **10**, 287 (1971)
34. Raff, L.M., Thompson, D.L.: (1985) Advances in classical trajectory methods. In: Baer M (ed.) Theory of Chemical Reaction Dynamics, CRC, Boca Raton, FL
35. Car, R., Parrinello, M.: Unified approach for molecular-dynamics and density-functional theory. Phys. Rev. Lett. **55**, 2471–74 (1985)
36. Tully, J.C.: Molecular dynamics with electronic transitions. J. Chem. Phys. **93**, 1061 (1990)
37. Ben-Nun, M., Martínez, T.J.: Nonadiabatic molecular dynamics: validation of the multiple spawning method for a multidimensional problem. J. Chem. Phys. **108**, 7244–7257 (1998)

38. Ben-Nun, M.; Martínez, T.J.: Ab initio quantum molecular dynamics. Adv. Chem. Phys. **121**, 439–512 (2002)
39. Masson, M., Laino, T., Tavernelli, I., Rothlisberger, U., Hutter, J.: Computational study of thymine dimer radical anion splitting in the self-repair process of duplex DNA. J. Am. Chem. Soc. **130**, 3443–3450, (2008)
40. Groenhof, G., Schäfer, L.V., Boggio-Pasqua, M., Goette, M., Grubmüller, H., Robb, M.A.: Ultrafast deactivation of an excited cytosine-guanine base pair in DNA. J. Am. Chem. Soc. **129**, 6812–6819 (2007)
41. Hudock, H.R., Martínez, T.J.: Excited-state dynamics of cytosine reveal multiple intrinsic subpicosecond pathways. Chem. Phys. Chem. **9**, 2486–2490 (2008)

Part II
Modeling Macromolecular Assemblies

Multiscale Modeling of Virus Structure, Assembly, and Dynamics

Eric R. May, Karunesh Arora, Ranjan V. Mannige, Hung D. Nguyen, and Charles L. Brooks III

1 Introduction

Viruses are traditionally considered as infectious agents that attack cells and cause illnesses like AIDS, Influenza, Hepatitis, etc. However, recent advances have illustrated the potential for viruses to play positive roles for human health, instead of causing disease [1, 2]. For example, viruses can be employed for a variety of biomedical and biotechnological applications, including gene therapy [3], drug delivery [4], tumor targeting [5], and medical imaging [6]. Therefore, it is important to understand quantitatively how viruses operate such that they can be engineered in a predictive manner for beneficial roles.

Most viruses are nanosized particles that replicate only inside a host cell they infect. A structure of a complete virus particle is made up of a protective coat of protein called a capsid that encloses its nucleic acid, either DNA or RNA. Virus capsids are extremely stable and possess wide-ranging mechanical strengths, which can be characterized in the theoretical framework typically used for characterizing materials [7–9]. Capsids exhibit diversity in not only material properties but also geometric attributes. Capsids across the virosphere display a wide diversity of

E.R. May • K. Arora • C.L. Brooks III (✉)
Department of Chemistry and Biophysics Program, University of Michigan, Ann Arbor, MI 48109, USA
e-mail: ericmay@umich.edu; karunesh@umich.edu; brookscl@umich.edu

R.V. Mannige
Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA
e-mail: ranjanmannige@gmail.com

H.D. Nguyen
Department of Chemical Engineering and Materials Science, University of California, Irving, Irvine, CA 92697, USA
e-mail: hdn@uci.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_7,

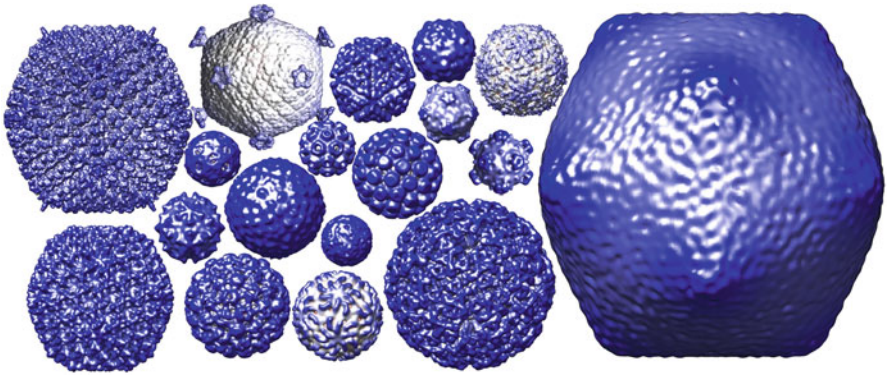

Fig. 1 As evident in the collage above, capsids come in a range of sizes (images represent electron microscopy reconstructions deposited into the virus particle explorer web site: viperdb.scripps.edu)

shapes, sizes, and architectures (Fig. 1), and understanding how these differences affect the material properties will provide design principles for engineering capsids.

Virus capsids are built from spontaneous self-assembly of multiple copies of a single protein or a few distinct proteins arranged in a highly symmetrical manner. Capsid assembly, from individual proteins in a correct, rapid, and spontaneous fashion on a biological timescale, is crucial for spreading an infection *in vivo* [10, 11]. Therefore, elucidating the means by which viral capsid self-assembly occurs may assist in the development of novel approaches to interfere with the assembly process and ultimately prevent viral infections. Following assembly, some virus capsids undergo morphological changes which are critical for the virus maturing into an infectious particle [12,13]. Understanding the dynamic behavior of assembled capsids is equally essential to gain insight into the mechanism associated with the maturation process and may open avenues for rational drug design by providing clues for disrupting the maturation process.

Despite several experimental [14–17] and theoretical efforts [18–20], the underlying principles that govern virus capsid self-assembly and maturation are not well understood. Experimental approaches such as X-ray crystallography and cryo-electron microscopy have provided excellent starting points to begin understanding virus architecture in an intricate manner, but do not provide the dynamical information crucial for understanding the virus life cycle. Other experimental methods probing dynamical and mechanical properties of viruses still lack sufficient resolution in the length and timescales to decipher the movement of individual proteins constituting the virus capsid. Rapid increases in the availability of computer power and algorithmic advances have made possible simulations of complete viruses in atomic detail on the timescale of tens of nanoseconds [21, 22], which are providing some insights into the experiments just noted. However, atomically detailed simulation remains a considerable challenge, at increasingly large time and length scales, for processes of biological importance such as assembly and maturation of virus capsids.

We have applied multiscale computational approaches ranging from topology-based mathematical modeling to physical simulations at different levels of coarse graining to describe the underpinnings of virus function and structural organization. This chapter describes key findings of our group's work in elucidating the underlying principles that govern the assembly and maturation of virus capsids using state-of-the-art multiscale simulation approaches; our focus is on presenting insights gained by various multiscale approaches rather than simulation details, which can be found in individual papers. Following a brief introduction into virus architecture, we describe the biological findings from simple mathematical models concerning the optimal subunit shape for constructing a capsid and the origins of evolutionary discrimination of certain T-numbers. We then describe key results, from self-assembly simulations of virus capsids using coarse-grained modeling, related to the generalized mechanistic description of structural polymorphism often observed *in vitro*. Following, we describe the development of a true multiscale approach linking equilibrium atomic fluctuations with macroscopic elastic properties of virus capsids and apply this approach to investigate the buckling transitions of HK97 bacteriophage. We conclude by outlining future applications and required model developments.

2 Background on Spherical Virus Architecture

A sampling of spherical viruses is shown in Fig. 1 to illustrate the size and shape diversity of virus capsids. Understanding how these structures form, as well as the reasons behind the differences in shape and size of virus structures is fundamentally important. Let us set the stage by providing a brief and historical introduction to the architecture of spherical virus capsids. The foundations of modern structural virology began in the 1950s, in the days before high (subnanometer) resolution imaging was available. During that time, it was becoming clear that the size of any capsid was much larger than the largest protein that the enclosed viral genome could express. Crick and Watson reasoned that one could form such a capsid only if viruses figured out a way to arrange multiple copies of a smaller protein (a "sub"-unit) into the form of a shell. Based on rudimentary crystallographic evidence [23], Crick and Watson had proposed that the capsid would have to assume a high order symmetry group. In doing so, large copies of the same subunit (now known to be a single protein) would possess identical or equivalent positions within the capsid (hence the idea of equivalence between the subunits). The proposed symmetries were the ones displayed by platonic solids [24], of which, icosahedral symmetry, a 60-fold symmetry, is the highest in order. However, new methods (such as negative staining electron microscopy) soon showed that the number of subunits per capsid were in slight disagreement with the Crick–Watson proposal. It was observed that instead of an icosahedral structure with 60 equivalent subunits, spherical capsids, albeit icosahedrally symmetric, were found to be composed of multiples of 60 subunits.

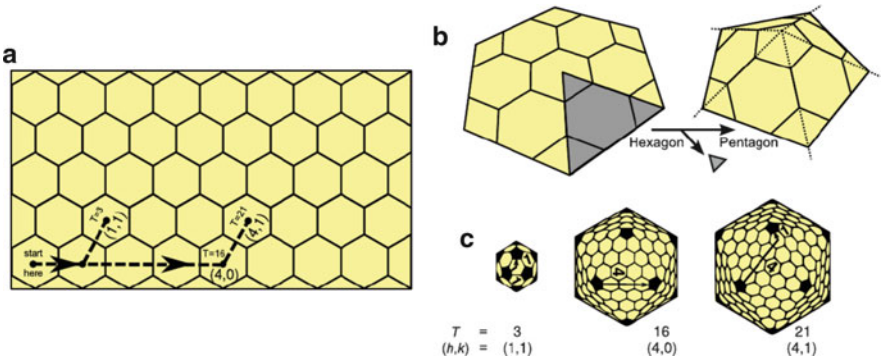

Fig. 2 Making capsid models of various sizes described by (h, k) pairs. The general idea is that all capsids consist of 12 pentamers (*darkened* in (**c**)) and a variable number of hexamers. Starting from only a sheet of hexagons (**a**), where hexagons represent hexamers, and then selectively converting specific hexamers into pentamers (**b**), a complete icosahedron may be constructed

Further advancement in the method of negative staining and electron microscopy led to the observation that capsids are shells formed from repeated pentagon and hexagon-like arrangements. From these early experiments, two groups of structural virologists, Horne and Wildy [25] and Caspar and Klug [26], found an interesting solution to the scalability problem. In particular, both groups recognized that virus capsids of practically any size could be created by combining 12 pentamers (symmetric clusters of five subunits) with a variable number of hexamers (symmetric clusters of six subunits). Caspar and Klug went further to describe a theoretical mechanism to "build an icosahedral capsid shell from a flat lattice of hexagons."

As shown in Fig. 2, a specific capsid can be described by two integers, h and k, representing steps in the h or k direction, respectively. By taking an "h, k walk" on the hexagonal surface (Fig. 2a), one ends up on a hexagon which is to be converted into a pentagon. These hexagons can be converted into pentagons by excising 1/6th of the selected hexagon and gluing the unpaired edges (Fig. 2b). When this procedure is repeated to make 12 such pentagons, one will be left with a three-dimensional model of a complete icosahedral capsid, where pentagons and hexagons represent pentamers and hexamers, respectively.

Although h and k are useful in understanding capsid size and arrangements of pentamers and hexamers, it is not always convenient to deal with two numbers as a descriptor. Conveniently, Caspar and Klug [26] re-introduced a useful descriptor (initially described by Goldberg in the 1940s), the triangulation number,

$$T = h^2 + k^2 + hk. \tag{1}$$

T is useful because it easily describes the number of subunits $(60T)$ and hexamers $(10(T-1))$ in the capsid and the number of distinct symmetry environments present within the capsid (which is T itself). Today, the triangulation number is the ubiquitous descriptor of virus architecture.

3 Mathematical and Geometric Models for Describing Virus Phenomena

In this section, we explore how concepts borrowed from mathematics and geometry may help in understanding structural features of virus capsids. Using these simplistic models, we have addressed problems at various levels of capsid research, ranging from understanding what is the optimal subunit shape for constructing a capsid, to understanding the prevalence of certain T-number structures and the absence of others.

3.1 The Canonical Capsid Model

The utility of simplistic models, which do not account for the specific atomic level interactions (i.e., an all-atom force field) have been useful in explaining various capsid phenomena such as capsid self-assembly [27–33], capsid morphology [34–37], subunit stoichiometry [38–40], mechanical properties [41–43], and symmetry [37, 44]. We will be primarily discussing one such geometric model, the "canonical capsid," that has served as a useful platform for the elucidation of capsid design principles [36, 37, 40].

The concept of the "canonical capsid," which is a surprisingly simple construct, is defined as a polyhedron whose faces, each representing a subunit, must be identical in shape. This model is also known as a "monohedral tiling." This simple model is useful because a large number of capsids found in nature can be represented as monohedral tilings [40]. In addition, these models can shed light on various physical properties of virus capsids that can be described as canonical.

3.2 Prediction of the Optimal Subunit Shape

Given the construct of the canonical capsid, a key question for investigation is which subunit shapes are permitted to exist within the confines of the canonical capsid. Using simple geometry and polyhedral rules [40], we have shown that canonical capsids can only accommodate *one* type of "prototile" (subunit design) consisting of five interacting edges. The bisected trapezoid (Fig. 3a) is one such acceptable prototile design. It is the same subunit shape that appears in all the natural capsids (Fig. 3b) we find to be represented by the canonical capsid model [40]. It has indeed been identified that many viruses share a common subunit protein fold (the double β-barrel), without sharing high sequence identity [45]. It is quite surprising that a simple canonical capsid model predicts such a ubiquitous shape found in viruses infecting almost all domains of life. Apparently, nature may be forcing viral capsid proteins into adopting this very special shape. It is tempting to conjecture that there is an overarching evolutionary pressure that may be acting on virus capsid's design.

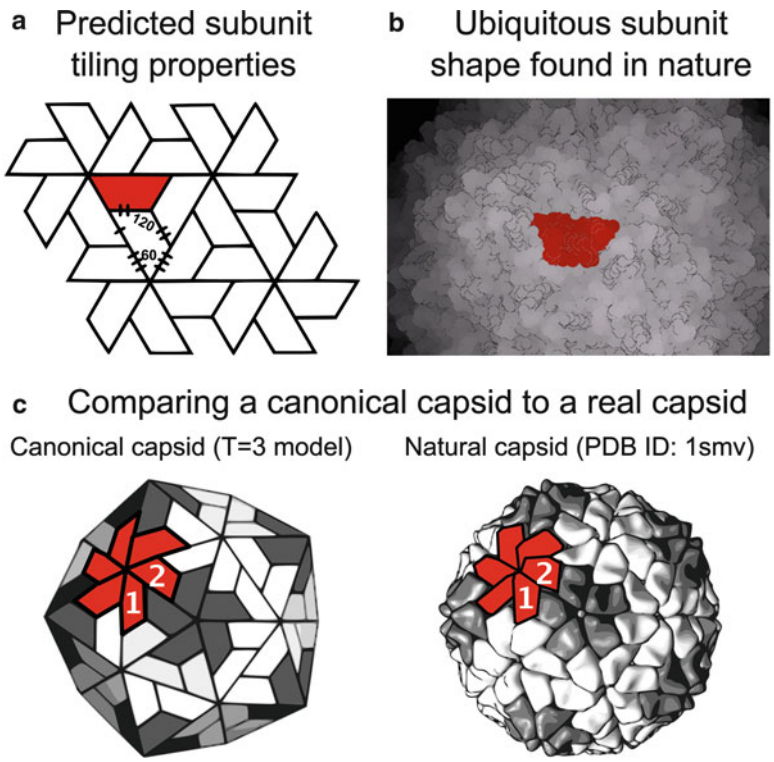

Fig. 3 *Canonical capsids as a model.* The basic subunit prototile–the bisected trapezoid (shaded in (**a**))–predicted from the analysis of the simplistic canonical capsid model [40] bears a strong resemblance to the subunit design ubiquitously found in virus capsids (**b**), indicating a mathematically motivated pressure in maintaining a trapezoidal subunit shape in nature. Apart from explaining the importance of the capsid subunit shape, the strong resemblance between these geometric entities and their real counterparts (exampled in (**c**)), allows for a number of studies in capsid design criteria [36, 37]

3.3 Hexamer Complexity as a Predictor of Capsid Properties

Analysis of the virus structural data collected over the last half century indicates that a very large array of capsid sizes ranging from tens to many thousands of subunits are known to exist in nature (Fig. 1). However, some capsid sizes are rarer than others (such as $T = 12, 19,$ and 27), an observation that has puzzled structural virologists as early as 1961 [25, 26]. The cause for this apparent bias in the distribution of the observed capsid sizes is still not clearly understood. To explore if there is an evolutionary pressure that discriminates against certain capsid shapes, we further investigated intrasubunit interactions within virus capsids using a canonical capsid model. Specifically, we explored how subunits interact and how the angles between subunits can impose constraints on the capsid shape.

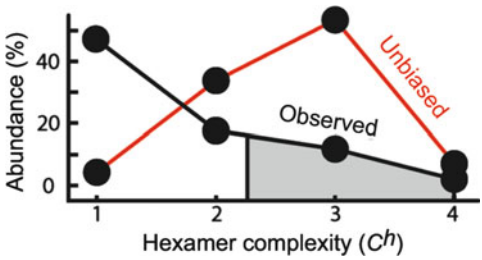

Fig. 4 As predicted by the inverse C^h rule, capsids with high hexamer complexity are underrepresented in nature as evident in the observed versus unbiased capsid abundances (% of families that display capsids of specific C^h)

The subunit–subunit angles present within the pentamers (which we call endo angles) impose constraints on the adjacent hexameric angles, an effect that is termed endo angle propagation [36]. While the shape and number of pentamers is fixed for all T number capsids, the number of hexamers (and therefore the shape) is not. The hexamers experience different environments based upon their adjacency to neighboring pentamers/hexamers. As a result, the angle patterns produced by interacting endo angles within the capsid ensure the emergence of three general morphological classes of capsids that can be differentiated by their h–k relationship [37]: *class 1* (described by the relationship $h > k = 0$), *class 2* ($h > k > 0$), and *class 3* ($h = k$). We have identified the minimum number of distinct hexamer shapes (which we call hexamer complexity C^h) required to form a canonical capsid of specific capsid size (T-number). Each canonical capsid of specific h and k is described by a single C^h value. Thus, C^h is very useful in systematically predicting properties of a group of capsids that were previously thought to be unrelated viruses.

C^h is also an indicator of the ease with which a capsid can be assembled, i.e., a larger number of distinct hexamer shapes would require a more complex assembly mechanism. Indeed, our modeling studies show that the capsids with a high C^h value require more auxiliary control mechanism for their assembly while the capsids with a low C^h value and low $T - number$ ($T = 3, 4,$ or 7) display the ability to assemble with no auxiliary requirements [46, 47]. Thus, the hexamer complexity number (C^h) can be used as tool to predict if a particular capsid assembly requires auxiliary mechanisms or proteins. Accordingly, we predicted that canonical capsids with larger C^h must be present with a lower frequency in nature since they require complex auxiliary assembly mechanisms. This hypothesis is corroborated by surveying all available capsid structures in the literature and virus structure databases. In the scenario that all T number capsids were equally probable, it would be expected that the complex capsids with $C^h > 2$ would represent the majority of the virus families observed in the nature (63%) (Fig. 4 *Unbiased*). However, in actuality, capsids with $C^h > 2$ represent only 5% of the observed capsid structures (Fig. 4 *Observed*). This suggests the existence of an evolutionary pressure which discriminates against viruses with a high hexamer complexity.

3.4 Limitations of the Canonical Capsid Model

A majority of the capsids we have studied display properties of canonical capsids [40]. However, the remaining small percentage of the noncanonical capsids cannot be well described by the canonical capsid model and likely require more sophisticated models for their characterization. For example, many noncanonical capsids possess nontrapezoidal subunit shapes (e.g., the members of the *polyomaviridae* family). It has been shown that these noncanonical capsids can be represented by other simplistic polyhedral models with slight embellishments [38, 39]. Still, there exist a few other noncanonical capsids with holes and large overlaps in their structures for which no simple solutions exist. It is these rule breakers that emphasize the requirement for more sophisticated theoretical models.

These mathematical modeling efforts have served to offer explanations to broad questions in the field of structural virology such as subunit shape and evolutionary discrimination of certain T-numbers. However, questions related to dynamical properties are more suitable to physics-based modeling studies. In the following sections, we will address two fundamental processes in the virus life cycle, capsid assembly (Sect. 4) and maturation (Sect. 5), using physics-based modeling techniques.

4 Self-Assembly of Virus Capsids

Highly specific and spontaneous self-assembly of individual proteins to form symmetric viral capsids inside the infected host cells is crucial for propagating the infection *in vivo* and is one of the most fundamental process of the virus life cycle. In addition, the *in vitro* self-assembly of empty capsids without the viral genome is of significant interest in bionanotechnology for vaccine design, gene therapy, and medical imaging [48]. As a specific example, empty capsids serve as vaccines to prevent cervical cancer, which is caused by the human papilloma virus. The vaccine, which consists of empty capsids of the human papilloma virus, prompts production of appropriate antibodies in the body, thereby priming an effective immune response that could be marshaled during subsequent exposure to the infectious virus [49]. The potency of the cervical vaccine depends strongly upon the degree of capsid self-assembly [50]. However, due to the inability to control assembly in laboratory and manufacturing practices, self-assembly of empty capsids often leads to architectural contaminants (i.e., structural polymorphism) [51]. A clear understanding of the kinetic mechanisms and thermodynamics of icosahedral capsid self-assembly would provide valuable insights into how to control the self-assembly process and is a key prerequisite to their widespread application in medicine.

The quantitative investigation of the virus capsid self-assembly mechanisms presents significant challenges for both experimental and computational approaches. Progress has been made toward understanding the molecular-level mechanisms

driving capsid formation through theoretical studies [19,20,38,44,52–55], structural analysis [56, 57], and *in vitro* self-assembly experiments of empty capsids using only purified capsid proteins [19,58,59]. Still, a detailed mechanistic understanding of the capsid self-assembly process is lacking. Despite rapid increases in the availability of computer power and algorithmic advances, atomically detailed simulations of the self-assembly process have been difficult due to the large system sizes and the long timescales involved in the process. As a consequence, to-date most simulation studies of capsid formation have been performed employing only simple coarse-grained models that significantly reduce the system size [32, 60, 61]. For example, Hagan and Chandler [32] modeled capsid proteins or capsomeres as point particles to simulate the assembly of small shells, Hicks and Henley [61] used an elastic model to represent capsid proteins as deformable triangles and Rapaport simulated the capsid self-assembly of polyhedra structures utilizing trapezoid units as a building block [28].

We investigated the spontaneous self-assembly process of different-sized virus capsids employing a coarse-grained molecular dynamics (MD) simulation approach. To increase the speed and efficiency of the simulations an extremely fast, event-driven method called discontinuous molecular dynamics (DMD) was employed [62–64]. Before performing simulations, we developed a range of geometric models that capture the geometric shape and energetic details of a coat protein without any specific built-in self-assembly rules such as nucleation. Interestingly, our prior two dimensional mathematical modeling studies, as well as, initial exploratory simulation studies by Rapaport [28] had predicted that the trapezoidal shape is a perfect building block to tile a closed icosahedral surface of any capsid size (see Sect. 3) [40]. To test this prediction by way of physical simulations, in our first generation of coat protein models each protein subunit was represented as a set of 24 beads arranged in four layers confined in the trapezoidal geometry (see Fig. 5a). Using a simplified model that exploits the important role of coat protein shape, together with the fast DMD method, allowed us to capture the spontaneous self-assembly of icosahedral capsids of different sizes as well as explore the optimal temperature and protein concentration required for the spontaneous self-assembly of capsids.

By performing over a hundred MD simulations at different temperatures and protein concentrations, we found that the assembly of $T = 1$ and $T = 3$ icosahedral capsids occurs with high fidelity only over a small range of temperatures and protein concentrations [33, 65]. Outside this range, particularly at low temperature or high protein concentration, large enclosed "monster particles" are produced (Fig. 5b). These mis-assemblies are remarkably similar to experimentally observed Turnip crinkle virus monster particles [66] or bacteriophage P22 monster particles [67]. Most importantly, our simulation studies revealed that the capsid assembly dynamics under optimal conditions is a nucleated process [58] involving monomer addition in which building blocks (either monomeric, dimeric, or trimeric species) are glued together in a sequential manner [33]. It is quite remarkable that our simulations employing simple models were able to recapitulate the experimental observations that capsid assembly is a nucleated process [19, 58, 59].

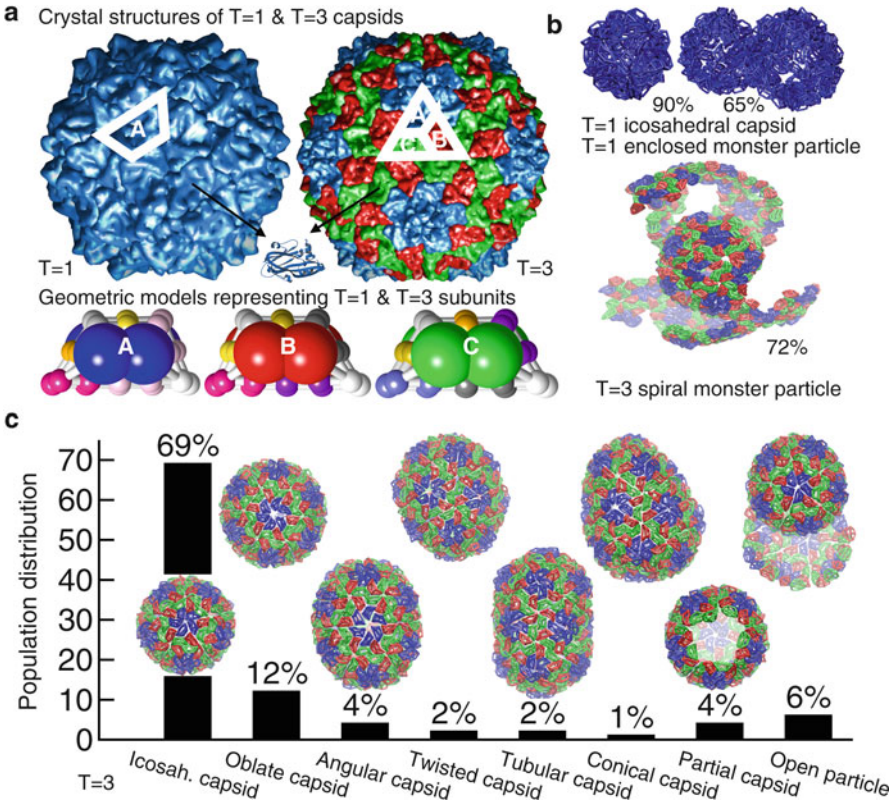

Fig. 5 (**a**) Coarse-grained models capturing the geometric shape of the protein and the protein-protein interactions that occur between proteins in the assembled capsid. (**b**) At low temperatures and high concentrations, assembly is nucleated too rapidly and partial growth leads to the combining of many partial capsids to form "monster particles" that are enclosed in $T = 1$ systems, and spiral-like in $T = 3$ systems, like those seen in cryo-EM experiments. (**c**) The assembly of $T = 3$ capsids under near optimal conditions yields a range of closed capsid forms that are determined by the number of five-to-six fold symmetry dislocations that occur as a result of certain kinetic pathways. The population distributions for supramolecular structures from $T = 3$ systems. The same structural polymorphism is also observed in $T = 1$ systems (Figures 1 and 2 of Nguyen et al. J. Amer. Chem. Soc., 131:2606–14, 2009, copyright 2009, American Chemical Society.)

Interestingly, upon shifting the conditions (protein concentration and temperature) for $T = 1$ and $T = 3$ capsid growth slightly, we observed the self-assembly of not only icosahedral capsids, but also of a well-defined set of nonicosahedral yet completely enclosed and equally stable capsules [65] (Fig. 5c). These nonicosahedral capsules exhibit morphologies similar to particles that have been observed in the mis-assembly of capsids of many viruses [14, 68–74]. These findings demonstrate that structural polymorphism in capsid structure is an inherent property of capsid proteins, is independent of the morphology of constituent subunits, and arises from condition-dependent kinetic mechanisms that are determined by initial assembly conditions.

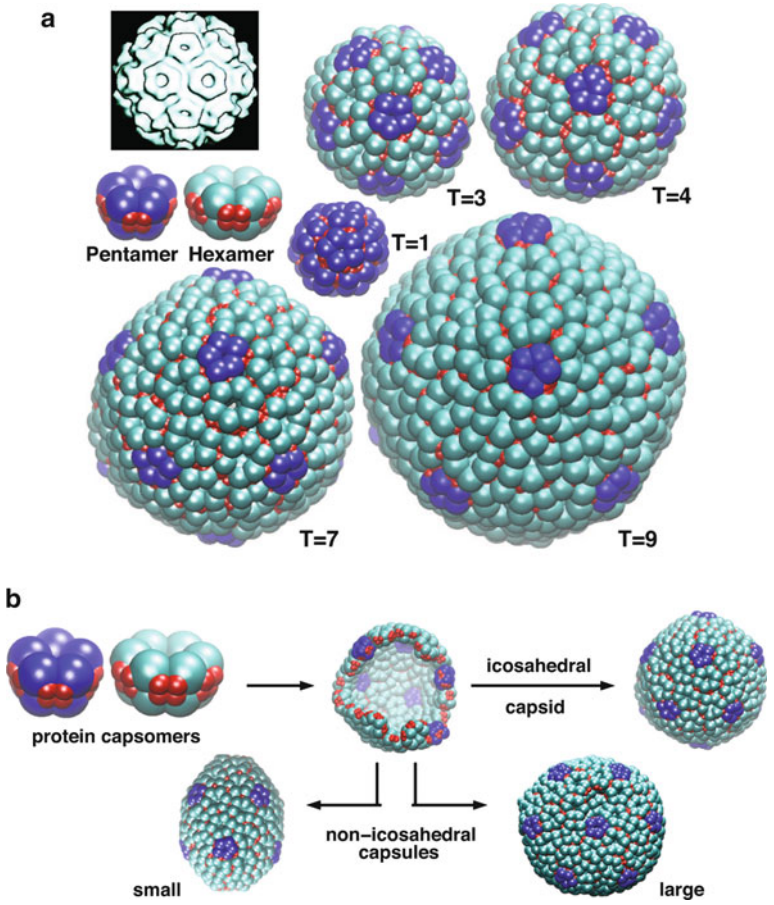

Fig. 6 (**a**) Coarse-grained model representing each pentameric or hexameric capsomer of coat proteins as a pentagonal or hexagonal structure for any T capsid systems, inspired by the presence of pentagonal and hexagonal morphology on many virus particles obtained from cryo-EM experiments. Each $T = 1, 3, 4, 7,$ or 9 capsid obtained from our simulations contains 12 pentamers and $(T - 1)10$ hexamers arranged on icosahedral lattice. (**b**) Different kinetic mechanisms of assembly in $T = 7$ systems were deciphered: sequential addition for icosahedral capsids, condensation of preformed intermediates for large non-icosahedral capsules, and premature collapse of intermediates for small non-icosahedral capsules (Nguyen and Brooks, Nano Lett. 8: 4574–4581, 2008, copyright 2008, American Chemical Society.)

Considering the ubiquitous nature of nonicosahedral capsules observed in our simulations of $T = 1$ and $T = 3$ systems, we predicted that such capsules are also formed in $T > 3$ systems. We confirmed this prediction by developing our second generation of coarse-grained model in which multiples of coat proteins are represented as either pentameric or hexameric capsomers [75] (Fig. 6a). Our model capsomers mimic the building blocks of a few known virus systems such as HK97 capsids [76], which have been shown to assemble from pentamers and hexamers. In

the simulation studies of $T = 1, 3, 4, 7, 9, 12, 13, 16,$ and 19 systems, we observed the formation of a variety of nonicosahedral yet highly ordered and enclosed capsules in addition to the expected icosahedral capsids. These simulations demonstrate that structural polymorphism is independent of the capsid complexity and the elementary kinetic mechanisms of self-assembly. Furthermore, the simulations revealed the existence of two distinctive and comprehensive classes of polymorphic structures. The first class includes aberrant capsules that are larger than their respective icosahedral capsids in $T = 1 - 7$ systems and the second class includes capsules that are smaller than their respective icosahedral capsids in $T = 7 - 19$ systems (Fig. 6b). The kinetic mechanisms responsible for the self-assembly of these two classes of aberrant structures were deciphered, providing insights into how to control the self-assembly of icosahedral capsids. To our knowledge, this is one of the first simulation studies that provided a generalized description of structural polymorphism, which is often observed in *in vitro* experiments [14, 68–70, 72, 73] and vaccine development studies [71].

Simulation studies, as described here, can provide new tools to inform potential strategies in antiviral development, protein design, and the engineering of novel biomaterials. The methodology employed in these studies could also be expanded upon to elucidate the means by which capsid proteins and the viral genome are self-assembled into full viruses. Such studies would enable us to make unique contributions to the field of virology/medicine by suggesting the development of novel ways to interfere with virus assembly and ultimately with viral infections.

5 Maturation and Mechanical Properties of Virus Capsids

An important aspect of designing nanotechnologies is material characterization; understanding how the material responds to stresses and different environmental conditions and ultimately the calculation of the fundamental mechanical moduli. Characterization of the mechanical properties of virus capsids is important for technology design as well as understanding the maturation phenomenon, which is one of the most fundamental process of the virus life cycle.

The $T = 7$ bacteriophage HK97 is a widely studied system [13, 76–81], due to its interesting structural features. This virus assembles into a procapsid structure consisting of 420 copies of a single protein and initially forms a rounded procapsid structure (Prohead II) as shown in Fig. 7 on the left. The seven protein asymmetric unit of Prohead II is also shown in Fig. 8. *In vivo* the structure matures upon packaging of the DNA genome, during which the structure expands, becomes faceted, and iso-peptide bonds form between side chains of different proteins, resulting in the mature (Head II) structure as shown in Fig. 7 on the right. The maturation transition (commonly termed a buckling transition) can also be triggered *in vitro* with empty capsids (genome deficient), by lowering the system pH [77, 82]. This maturation-related structural transition has broad implications for understanding virus behavior [83]. HK97 is believed to share many aspects of its maturation

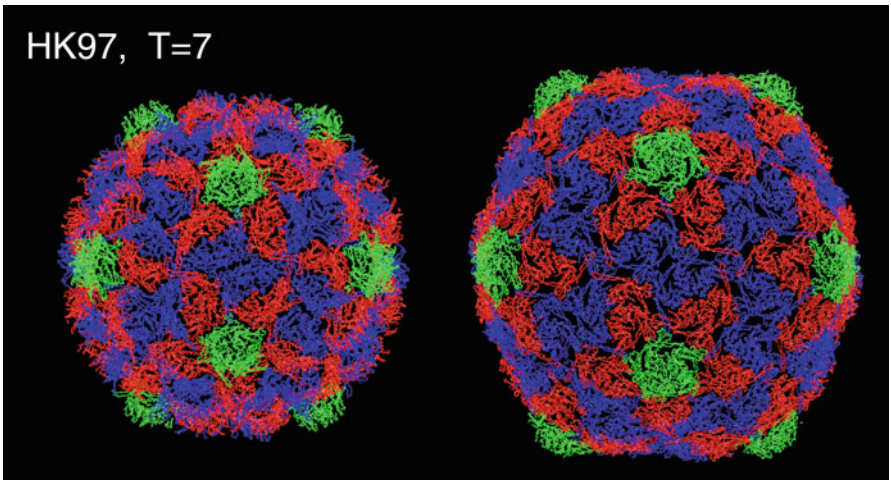

Fig. 7 The structures of HK97 ($T = 7$) in compact and swollen forms (Fig. 1 of Tama et al., J. Mol. Biol., 345: 299–314, 2005, copyright 2005 Elsevier B.V.)

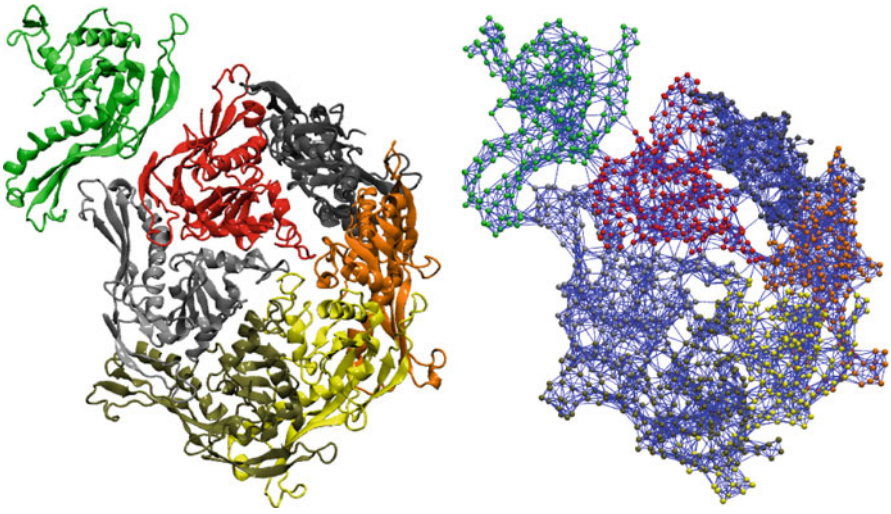

Fig. 8 The seven protein asymmetric unit of the virus HK97, represented with a ribbon drawing (*left*) and as an elastic network (*right*), in which the *lines* represent the harmonic springs of the network connecting C_α atoms within 8Å of each other

process with other double-stranded DNA bacteriophages and with herpes virus [84]. Additionally, HK97 is an ideal system to understand what governs the equilibrium shape of spherical viruses because it exists in both a rounded (immature) and faceted (mature) forms during its life cycle. Understanding structural transitions of HK97 from a rounded to a faceted shape should help explain, in general, why certain

viruses adopt a particular configuration from the range of possible shapes and sizes (see Fig. 1).

Attempts to explain this faceting or buckling phenomena have been made using simplified models. Using the discrete canonical capsid model, we identified that viruses belonging to the *class 2* ($h > k > 0$) morphological group can undergo buckling transitions (See discussion in Sect. 3). The virus structures in this group, which includes HK97 ($T = 7$), have a degree of freedom associated with the hexamer configurations [36]. This is in contrast to the structures in *class 1* and *class 3* which we believe to consist of rigid hexamers, with zero degrees of freedom. This degree of freedom in the *class 2* hexamers was identified through our analysis of endo angle constraints (see Sect. 3), and we find two distinct stable states that the hexamer can sample. The two available hexamer configurations are a *pucker in* and *pucker out* state, corresponding to the faceted and rounded conformation of the capsid, respectively. An alternative explanation to the buckling phenomena has been put forth using purely continuum elastic theory of thin shells, proposed by Lidmar, Mirny, and Nelson (LMN) [41]. According to the LMN theory, the equilibrium configuration of the capsid is governed by a minimization of the elastic energy of the shell. As the elastic energy is dependent on the elastic properties of the shell, shape changes will arise in response to modulation of these properties. While both of these models have offered reasonable explanations for the buckling transition of virus caspids, neither work has incorporated molecular detail into their models. Recently, we have attempted to bridge the discrete and the continuum description of the virus capsid buckling transition by developing a multiscale approach which relates atomic level equilibrium fluctuations to the macroscopic elastic properties of the system [85, 86].

In the LMN theory, a single parameter, the Foppl-von Kármán number (γ), predicts whether a capsid will adopt a rounded or faceted form, and as can be seen in Fig. 1, both states are known to exist in nature. The shape dependence on γ is predicted to have a relatively sharp transition between rounded and faceted states, γ is given by

$$\gamma = \frac{YR^2}{\kappa}, \tag{2}$$

where Y is the two-dimensional Young's modulus, R is the shell radius, and κ is the bending modulus. Determining Y and κ for capsid structures will allow γ to be determined, but it is inherently important to calculate these moduli to better understand the mechanical properties of these systems. Furthermore, the material characterization of capsids should accelerate the development of virus-based nanotechnologies.

It is difficult to measure the elastic properties of nano-sized objects such as virus capsids experimentally because most experimental techniques involve averaging over a large number of particles. However, the single-molecule technique of atomic force microscopy (AFM) is well suited for probing the mechanical strength of capsids through nanoindentation studies. These studies typically are conducted in conjunction with finite-element (FE) simulations in which the three-dimensional

Young's modulus (E) can be estimated by matching the AFM and FE force-displacement curves [87–91]. While these studies address how viruses respond to force loading, they do not directly evaluate the equilibrium mechanical properties of these systems. The loading rates in the AFM studies have been criticized for being too fast [42] and also the assumptions made in the finite element modeling that the capsid is an isotropic and homogenous medium may not be appropriate.

To overcome these limitations, we have developed a multiscale approach for calculating the continuum elastic properties Y and κ. In this approach, we utilize atomically detailed models to compute the equilibrium thermal fluctuations of the capsid, which are then related to the elastic properties of the capsid through an elastic Hamiltonian [85, 86]. The most relevant motions of the system, when trying to connect the atomic model to the continuum level theory, are the low frequency (long wavelength) collective motions. These collective motions arise from the atomic-level interactions and therefore a model is required which incorporates molecular detail. To compute collective motions, we utilize elastic network models (ENM), which incorporate molecular level details and variable density of interatomic interactions; an ENM representation of the asymmetric unit of HK97 is shown in Fig. 8 on the right. In addition, ENMs utilize a simple interaction potential which makes it computationally efficient to capture the collective motions of large macromolecular assemblies [92]. The essence of the ENM is that it is a harmonic approximation to the free energy minimum in which the structure lies. The numerous potential terms (Lennard–Jones, electrostatic, bond, angle, dihedral, etc) in a standard semi-empirical MD force field are replaced by a single harmonic potential term accounting for the vibrations of interacting pairs of atoms [93]. The normal modes of the ENM can be calculated by finding the eigenvectors of the Hessian matrix of second derivatives of the potential. A trajectory of the ENM can then be computed by propagating the network along a set of the lowest frequency normal modes.

From these ENM trajectories, a two-dimensional surface is computed by averaging over the shell thickness, and the fluctuations are projected onto a spherical harmonic basis set. The forces on a 2D elastic shell are known from the early works on continuum elastic theory of shells [94], and from these forces an energy density can be written down in the spherical harmonic basis. The total elastic energy can then be computed by integrating over the shell surface, which, due to the orthogonality properties of the basis set, reduces to a sum over the mode magnitudes

$$E = \frac{1}{2}\sum_{l}\left(8b + \kappa\frac{l(l-1)(l+1)(l+2)}{R^2}\right)|\hat{a}_l|^2, \tag{3}$$

where, $|\hat{a}_l|^2 \equiv \sum_{m=-l}^{+l} a_{lm}a_{lm}^*$, a_{lm} is the magnitude of spherical harmonic l, m, and b is the sum of the Lamé constants (λ, μ), from which Y can be calculated when a value for the Poisson ratio is known (or assumed). Given the quadratic form of the energy, the ensemble averages of $|\hat{a}_l|^2$ can be calculated and a relationship is obtained which contains only measurable surface properties (R, $\langle|\hat{a}_l|^2\rangle$) and elastic

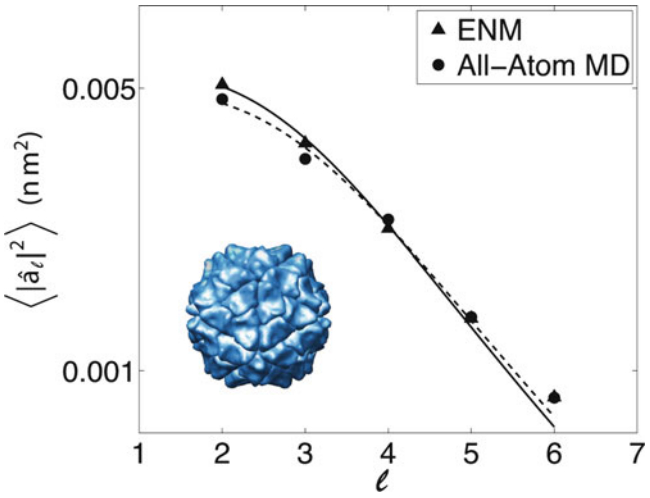

Fig. 9 The spectrum of spherical harmonics describing the equilibrium thermal fluctuations of the capsid surface for the $T = 1$ mutant of the *Sesbania mosaic virus*, shown in *bottom left*. A nearly identical spectrum is produced for an all-atom MD simulation of an entire capsid, as that from and elastic network model (ENM), that is scaled via an MD simulation of only the asymmetric unit. In both cases the fluctuation spectrum is well described by the theoretical model in (4) (Figure 1 of May and Brooks, Phys. Rev. Lett. 106:18801–18804, 2011, copyright, 2011 American Physical Society.)

parameters (λ, μ, κ)

$$\langle |\hat{a}_l|^2 \rangle = \frac{k_B T}{8b + \kappa \frac{l(l-1)(l+1)(l+2)}{R^2}}. \tag{4}$$

Our formulation of the ENM is a nondimensional model, and therefore we use MD to scale the trajectory to make it quantitatively accurate. The MD simulations are performed on the asymmetric unit of the capsid under icosahedral rotational boundary conditions [95]. From the MD simulations a scaling factor is calculated, which is passed to the ENM model to connect the cruder ENM model to the more accurate MD force field. We were able to show that this multiscale approach, combining an ENM with MD on the asymmetric unit, was a good approximation to the fluctuations generated by simulating the entire capsid explicitly with MD. The agreement between the theoretical model and the observed fluctuations, as well as the agreement between the multiscale (ENM) and the brute force MD approach are shown in Fig. 9. From the fits to the data, we are able to determine Y, κ, and γ for the $T = 1$ mutant of *Sesbania mosaic virus* (SeMV). We have applied this multiscale approach to HK97 [85, 86] in the mature and immature forms and predicted a significant change in γ ($\sim$200 immature, $\sim$800 mature) between the states. These values are in agreement with the LMN theory, which predicts structures with $\gamma < 250$ should be spherical and those with higher γ values

should take on faceted forms, as is observed in Fig. 7. Additionally, we calculated a reduction in κ ($\sim$70 k_BT immature, $\sim$30 k_BT mature) between the states. This reduction in κ is functionally important, as it allows the faceted state, with the high curvature corners, to be adopted at a lower energy cost. Furthermore, it can be concluded from this analysis that the interactions which are changing during the transitions, function to reduce κ making the shell more flexible and enable it to reach the infectious state more efficiently. From this analysis we have inferred a mechanical mechanism for the maturation of HK97 by incorporating molecular details and have provided support for the LMN theory of buckling transitions. Further examination of larger ($T > 7$) capsid structures using this multiscale method will allow us to test our predictions from the canonical capsid model that only *class 2* capsids have the propensity to undergo buckling transitions.

6 Conclusions and Future Directions

We have studied several aspects of a virus capsid's behavior ranging from elastic properties to evolutionary pressures using a variety of modeling techniques. These techniques span the range from all-atom molecular simulations, to coarse-grained studies of assembly, to purely mathematical models. Clearly, maturation and capsid assembly, which are fundamental processes of virus life cycles span a wide range of spatial and temporal scales. To make progress, we have explored one virus life cycle process at a time, which allowed us to build models appropriate for the phenomena under investigation. Even within these independent studies, we have used multiscale approaches to bridge molecular level detail to continuum theory (Sect. 5), and incorporate what we learn at one level of description (*subunit shape*, Sect. 3), into our studies of other aspects of the life cycle (*assembly*, Sect. 4). The current work has offered explanations for several features of viruses not currently accessible through experimentation. The goal of all of these works is to gain a better understanding of how viruses operate and it is this knowledge that will further our ability to fight viral infections, develop and manufacture vaccines, and utilize capsids in nanotechnology applications. However, to have an greater impact on health and technology we must continue our exploration to elucidate the intricate and complex processes of virus life cycles.

In future studies, we will explore the transition pathways associated with structural changes of capsids. In an earlier study, normal mode analysis identified the dominant modes characterizing structural transitions of virus capsids, including HK97 [96]. In the case of HK97, two modes were required to describe the configurational change. Using these dominant icosahedral normal modes, pathways were constructed to connect the states of the system. However, these pathways may not be representative of the physical pathway the virus undergoes, because they are not refined against an "accurate" potential function. Computing the energetics of the pathway (free energy barriers, ΔG between stable states) requires using a more detailed potential. These calculations will require advanced sampling techniques to

"find" a minimal energy pathway and then compute the probabilities of the states along the path [97–99]. Understanding how pH effects the energetics of the pathway can also be incorporated into our pathway modeling efforts through constant pH-MD methods [100, 101]. The potential benefit of studying these transition pathways is that molecular interactions that have a drastic effect on the behavior of the system can be identified. For example, specific salt bridges might form at a given pH, but altering the pH could break those salt bridges and change the free energy barrier between the stable states. Identification of these key residues can be tested through mutagenesis studies, and could provide a target for preventing virus maturation. Similarly, the pathway methods can be combined with elasticity calculations such as described above, and residues that are responsible for altering the elastic character of the material can be identified. This knowledge could provide design principles for engineering novel capsids and for modulating the properties of capsids used in nanotechnologies. Understanding transition pathways is just one avenue of further investigation of viruses, other areas of interest include understanding viral protein–host protein interactions [102] and protein–nucleic acid interactions during virus assembly. Viruses have a rich array of features and phenomena that are still poorly understood. However, by building and employing computational and theoretical models that capture the essential physics of the underlying phenomenon, we can shed light on many of these unresolved aspects of the virus life cycle.

Acknowledgments This work has been supported by the NSF through the center for theoretical biological physics (CTBP) at the University of California, San Diego (PHY0216576), by the National Institute of Health for funding through the multiscale modeling tools for structural biology (MMTSB) research resource center RR012255, and research grant GM037555, and by the National Science Foundation through a postdoctoral fellowship to ERM (DBI-0905773).

References

1. Uchida, M., Klem, M.T., Allen, M., Suci, P., Flenniken, M., Gillitzer, E., Varpness, Z., Liepold, L.O., Young, M., Douglas, T.: Biological containers: protein cages as multifunctional nanoplatforms. Adv. Mater. **19**(8), 1025–1042 (2007)
2. Maham, A., Tang, Z., Wu, H., Wang, J., Lin, Y.: Protein-based nanomedicine platforms for drug delivery. Small **5**(15), 1706–1721 (2009)
3. Miller, A.D.: Human gene therapy comes of age. Nature **357**(6378), 455–460 (1992)
4. Douglas, T., Young, M.: Host-guest encapsulation of materials by assembled virus protein cages. Nature **393**(6681), 152–155 (1998)
5. Destito, G., Yeh, R., Rae, C.S., Finn, M.G., Manchester, M.: Folic acid-mediated targeting of cowpea mosaic virus particles to tumor cells. Chem. Biol. **14**(10), 1152–1162 (2007)
6. Gupta, S.S., Raja, K.S., Kaltgrad, E., Strable, E., Finn, M.G.: Virus-glycopolymer conjugates by copper(i) catalysis of atom transfer radical polymerization and azide-alkyne cycloaddition. Chem. Commun. (Camb.) (34), 4315–4317 (2005)
7. Smith, D.E., Tans, S.J., Smith, S.B., Grimes, S., Anderson, D.L., Bustamante, C.: The bacteriophage straight phi29 portal motor can package dna against a large internal force. Nature **413**(6857), 748–752 (2001)

8. Ivanovska, I., Wuite, G., Jnsson, B., Evilevitch, A.: Internal dna pressure modifies stability of wt phage. Proc. Natl. Acad. Sci. USA **104**(23), 9603–9608 (2007)
9. Roos, W.H., Bruinsma, R., Wuite, G.J.L.: Physical virology. Nature Phys. **6**(10), 733–743 (2010)
10. Bancroft, J.B., Hills, G.J., Markham, R.: A study of the self-assembly process in a small spherical virus. Formation of organized structures from protein subunits in vitro. Virology **31**(2), 354–379 (1967)
11. Rose, R.C., Bonnez, W., Reichman, R.C., Garcea, R.L.: Expression of human papillomavirus type l1 (HPV-l1) protein in insect cells: in vivo and in vitro assembly of viruslike particles. J. Virol. **67**(4), 1936–1944 (1993)
12. Conway, J.F., Duda, R.L., Cheng, N., Hendrix, R.W., Steven, A.C.: Proteolytic and conformational control of virus capsid maturation: the bacteriophage hk97 system. J. Mol. Biol. **253**(1), 86–99 (1995)
13. Lata, R., Conway, J.F., Cheng, N., Duda, R.L., Hendrix, R.W., Wikoff, W.R., Johnson, J.E., Tsuruta, H., Steven, A.C.: Maturation dynamics of a viral capsid: visualization of transitional intermediate states. Cell **100**(2), 253–263 (2000)
14. Adolph, K.W., Butler, P.J.: Studies on the assembly of a spherical plant virus. I. States of aggregation of the isolated protein. J. Mol. Biol. **88**(2), 327–41 (1974)
15. Rossmann, M.G.: Constraints on the assembly of spherical virus particles. Virology **134**(1), 1–11 (1984)
16. Ceres, P., Zlotnick, A.: Weak protein–protein interactions are sufficient to drive assembly of hepatitis b virus capsids. Biochemistry **41**, 11525–11531 (2002)
17. Johnson, J.M., Willits, D.A., Young, M.J., Zlotnick, A.: Interaction with capsid protein alters rna structure and the pathway for in vitro assembly of cowpea chlorotic mottle virus. J. Mol. Biol. **335**(2), 455–64 (2004)
18. Schwartz, R., Shor, P.W., Prevelige, P.E., Berger, B.: Local rules simulation of the kinetics of virus capsid self-assembly. Biophys. J. **75**, 2626–2636 (1998)
19. Zlotnick, A., Johnson, J.M., Wingfield, P.W., Stahl, S.J., Endres, D.: A theoretical model successfully identifies features of hepatitis b virus capsid assembly. Biochemistry **38**(44), 14644–14652 (1999)
20. Bruinsma, R.F., Gelbart, W.M., Reguera, D., Rudnick, J., Zandi, R.: Viral self-assembly as a thermodynamic process. Phys. Rev. Lett. **90**(24), 248101 (2003)
21. Arkhipov, A., Freddolino, P.L., Schulten, K.: Stability and dynamics of virus capsids described by coarse-grained modeling. Structure **14**(12), 1767–1777 (2006)
22. Zink, M., Grubmller, H.: Mechanical properties of the icosahedral shell of southern bean mosaic virus: a molecular dynamics study. Biophys. J. **96**(4), 1350–1363 (2009)
23. Caspar, D.L.D.: Structure of bushy stunt virus. Nature **177**(4506), 476–477 (1956)
24. Crick, F.H., Watson, J.D.: Structure of small viruses. Nature **177**(4506), 473–475 (1956)
25. Horne, R.W., Wildy, P.: Symmetry in virus architecture. Virology **15**, 348–373 (1961)
26. Caspar, D.L., Klug, A.: Physical principles in the construction of regular viruses. Cold Spring Harb. Symp. Quant. Biol. **27**, 1–24 (1962)
27. Schwartz, R.S., Garcea, R.L., Berger, B.: 'local rules' theory applied to polyomavirus polymorphic capsid assemblies. Virology **268**(2), 461–470 (2000)
28. Rapaport, D.C.: Self-assembly of polyhedral shells: a molecular dynamics study. Phys. Rev. E **70**(5), 1539–1555 (2004)
29. Endres, D., Miyahara, M., Moisant, P., Zlotnick, A.: A reaction landscape identifies the intermediates critical for self-assembly of virus capsids and other polyhedral structures. Prot. Sci. **14**, 1518–1525 (2005)
30. Keef, T., Taormina, A., Twarock, R.: Assembly models for papovaviridae based on tiling theory. Phys. Biol. **2**(3), 175–188 (2005)
31. Keef, T., Micheletti, C., Twarock, R.: Master equation approach to the assembly of viral capsids. J. Theor. Biol. **242**(3), 713–721 (2006)
32. Hagan, M.F., Chandler, D.: Dynamic pathways for viral capsid assembly. Biophys. J. **91**, 42–54 (2006)

33. Nguyen, H.D., Reddy, V.S., Brooks III, C.L. Deciphering the kinetic mechanism of spontaneous self-assembly of icosahedral capsids. Nano Lett. **7**(2), 338–344 (2007)
34. Workum, K.V., Douglas, J.F.: Symmetry, equivalence, and molecular self-assembly. Phys. Rev. E **73**, 031502 (2006)
35. Chen, T., Zhang, Z., Glotzer, S.C.: A precise packing sequence for self-assembled convex structures. Proc. Natl. Acad. Sci. USA **104**(3), 717–722 (2007)
36. Mannige, R.V., Brooks III, C.L.: Geometric considerations in virus capsid size specificity, auxiliary requirements, and buckling. Proc. Natl. Acad. Sci. USA **106**(21), 8531–8536 (2009)
37. Mannige, R.V., Brooks III, C.L.: Periodic table of virus capsids: implications for natural selection and design. PLoS One **5**(3), e9423 (2010)
38. Twarock, R.: A tiling approach to virus capsid assembly explaining a structural puzzle in virology. J. Theor. Biol. **226**(4), 477–482 (2004)
39. Twarock, R.: Mathematical virology: a novel approach to the structure and assembly of viruses. Phil. Trans. R. Soc. A **364**, 3357–3373 (2006)
40. Mannige, R.V., Brooks III, C.L.: Tilable nature of virus capsids and the role of topological constraints in natural capsid design. Phys. Rev. E **77**(5), 051902 (2008)
41. Lidmar, J., Mirny, L., Nelson, D.R.: Virus shapes and buckling transitions in spherical shells. Phys. Rev. E **68**, 051910–051919 (2003)
42. Nguyen, T.T., Bruinsma, R.F., Gelbart, W.M.: Elasticity theory and shape transitions of viral shells. Phys. Rev. E Stat. Nonlin. Soft. Matter Phys. **72**(5 Pt 1), 051923 (2005)
43. Zandi, R., Reguera, D.: Mechanical properties of viral capsids. Phys. Rev. E Stat. Nonlin. Soft. Matter Phys. **72**, 021917 (2005)
44. Zandi, R., Reguera, D., Bruinsma, R.F., Gelbart, W.M., Rudnick, J.: Origin of icosahedral symmetry in viruses. Proc. Natl. Acad. Sci. USA **101**(44), 15556–15560 (2004)
45. Bamford, D.H., Grimes, J.M., Stuart, D.I.: What does structure tell us about virus evolution? Curr. Opin. Struct. Biol. **15**(6), 655–663 (2005)
46. Johnson, J.E., Speir, J.A.: Quasi-equivalent viruses: a paradigm for protein assemblies. J. Mol. Biol. **269**(5), 665–75 (1997)
47. Dokland, T., McKenna, R., Ilag, L.L., Bowman, B.R., Incardona, N.L., Fane, B.A., Rossmann, M.G.: Structure of a viral procapsid with molecular scaffolding. Nature **389**(6648), 308–313 (1997)
48. Douglas, T., Young, M.: Viruses: making friends with old foes. Science **312**(5775), 873–875 (2006)
49. Koutsky, L.A., Ault, K.A., Wheeler, C.M., Brown, D.R., Barr, E., Alvarez, F.B., Chiacchierini, L.M., Jansen, K.U.: A controlled trial of a human papillomavirus type 16 vaccine. N. Engl. J. Med. **347**(21), 1645–1651 (2002)
50. Shank-Retzlaff, M., Wang, F., Morley, T., Anderson, C., Hamm, M., Brown, M., Rowland, K., Pancari, G., Zorman, J., Lowe, R., Schultz, L., Teyral, J., Capen, R., Oswald, C.B., Wang, Y., Washabaugh, M., Jansen, K., Sitrin, R.: Correlation between mouse potency and in vitro relative potency for human papillomavirus type 16 virus-like particles and gardasil vaccine samples. Hum. Vaccin. **1**(5), 191–7 (2005)
51. Shi, L., Sings, H.L., Bryan, J.T., Wang, B., Wang, Y., Mach, H., Kosinski, M., Washabaugh, M.W., Sitrin, R., Barr, E.: Gardasil: prophylactic human papillomavirus vaccine development–from bench top to bed-side. Clin. Pharmacol. Ther. **81**(2), 259–64 (2007)
52. Wales, D.J.: Closed-shell structures and the building game. Chem. Phys. Lett. **141**, 478–484 (1987)
53. Berger, B., Shor, P.W., Tucker-Kellogg, L., King, J.: Local rule-based theory of virus shell assembly. Proc. Natl. Acad. Sci. USA. **91**, 7732–7736 (1994)
54. Zlotnick, A.: To build a virus capsid. An equilibrium model of the self assembly of polyhedral protein complexes. J. Mol. Biol. **241**(1), 59–67 (1994)
55. Endres, D., Zlotnick, A.: Model-based analysis of assembly kinetics for virus capsids or other spherical polymers. Biophys. J. **83**, 1217–1230 (2002)

56. Reddy, V.S., Giesing, H.A., Morton, R.T., Kumar, A., Post, C.B., Brooks III, C.L., Johnson, J.E.: Energetics of quasiequivalence: computational analysis of protein-protein interactions in icosahedral viruses. Biophys. J. **74**(1), 546–558 (1998)
57. Shepherd, C.M., Borelli, I.A., Lander, G., Natarajan, P., Siddavanahalli, V., Bajaj, C., Johnson, J.E., Brooks III, C.L., Reddy, V.S.: Viperdb: a relational database for structural virology. Nucleic Acids Res. **34**, D386–D389. (2006)
58. Zlotnick, A., Aldrich, R., Johnson, J.M., Ceres, P., Young, M.J.: Mechanism of capsid assembly for an icosahedral plant virus. Virology **277**, 450–456 (2000)
59. Casini, G.L., Graham, D., Heine, D., Garcea, R.L., Wu, D.L.: In vitro papillomavirus capsid assembly analyzed by light scattering. Virology **325**, 320–327 (2004)
60. Zhang, T., Schwartz, R.: Simulation study of the contribution of oligomer/oligomer binding to capsid assembly kinetics. Biophys. J. **90**, 57–64 (2006)
61. Hicks, S.D., Henley, C.L.: Irreversible growth model for virus capsid assembly. Phys. Rev. E Stat. Nonlin. Soft. Matter Phys. **74**(3 Pt 1), 031912 (2006)
62. Alder, B.J., Wainwright, T.E.: Studies in molecular dynamics. I. General method. J. Chem. Phys. **31**, 459–466 (1959)
63. Rapaport, D.C.: Molecular dynamics simulation of polymer chains with excluded volume. J. Phys. A **11**, L213–L217 (1978)
64. Bellemans, A., Orban, J., Belle, D.V.: Molecular dynamics of rigid and non-rigid necklaces of hard discs. Mol. Phys. **39**, 781–782 (1980)
65. Nguyen, H.D., Reddy, V.S., Brooks III, C.L. Invariant polymorphism in virus capsid assembly. J. Am. Chem. Soc. **131**(7), 2606–14 (2009)
66. Sorger, P.K., Stockley, P.G., Harrison, S.C.: Structure and assembly of turnip crinkle virus. ii. mechanism of reassembly in vitro. J. Mol. Biol. **191**(4), 639–658 (1986)
67. Earnshaw, W., King, J.: Structure of phage p22 coat protein aggregates formed in the absence of the scaffolding protein. J. Mol. Biol. **126**, 721–747 (1978)
68. Bancroft, J.B., Bracker, C.E., Wagner, G.W.: Structures derived from cowpea chlorotic mottle and brome mosaic virus protein. Virology **38**(2), 324–35 (1969)
69. Salunke, D.M., Caspar, D.L., Garcea, R.L.: Polymorphism in the assembly of polyomavirus capsid protein vp1. Biophys. J. **56**(5), 887–900 (1989)
70. Kanesashi, S.N., Ishizu, K., Kawano, M.A., Han, S.I., Tomita, S., Watanabe, H., Kataoka, K., Handa, H.: Simian virus 40 vp1 capsid protein forms polymorphic assemblies in vitro. J. Gen. Virol. **84**(Pt 7), 1899–905 (2003)
71. Zhao, Q., Guo, H.H., Wang, Y., Washabaugh, M.W., Sitrin, R.D.: Visualization of discrete l1 oligomers in human papillomavirus 16 virus-like particles by gel electrophoresis with coomassie staining. J. Virol. Methods **127**(2), 133–40 (2005)
72. Fu, C.Y., Morais, M.C., Battisti, A.J., Rossmann, M.G., Jr. Prevelige, P.E.: Molecular dissection of o29 scaffolding protein function in an in vitro assembly system. J. Mol. Biol. **366**(4), 1161–1173 (2007)
73. Dong, X.F., Natarajan, P., Tihova, M., Johnson, J.E., Schneemann, A.: Particle polymorphism caused by deletion of a peptide molecular switch in a auasiequivalent icosahedral virus. J. Virol. **72**(7), 6024–6033 (1998)
74. Cusack, S., Oostergetel, G.T., Krijgsman, P.C., Mellema, J.E.: Structure of the top a-t component of alfalfa mosaic virus. A non-icosahedral virion. J. Mol. Biol. **171**(2), 139–55 (1983)
75. Nguyen, H.D., Brooks III, C.L.: Generalized structural polymorphism in self-assembled viral particles. Nano Lett. **8**, 4574–81 (2008)
76. Xie, Z., Hendrix, R.W.: Assembly in vitro of bacteriophage hk97 proheads. J. Mol. Biol. **253**(1), 74–85 (1995)
77. Duda, R.L., Hempel, J., Michel, H., Shabanowitz, J., Hunt, D., Hendrix, R.W.: Structural transitions during bacteriophage hk97 head assembly. J. Mol. Biol. **247**(4), 618–635 (1995)
78. Wikoff, W.R., Liljas, L., Duda, R.L., Tsuruta, H., Hendrix, R.W., Johnson, J.E.: Topologically linked protein rings in the bacteriophage hk97 capsid. Science **289**(5487), 2129–2133 (2000)

79. Conway, J.F., Wikoff, W.R., Cheng, N., Duda, R.L., Hendrix, R.W., Johnson, J.E., Steven, A.C.: Virus maturation involving large subunit rotations and local refolding. Science **292**(5517), 744–748 (2001)
80. Ross, P.D., Conway, J.F., Cheng, N., Dierkes, L., Firek, B.A., Hendrix, R.W., Steven, A.C., Duda, R.L.: A free energy cascade with locks drives assembly and maturation of bacteriophage hk97 capsid. J. Mol. Biol. **364**(3), 512–525 (2006)
81. Gertsman, I., Gan, L., Guttman, M., Lee, K., Speir, J.A., Duda, R.L., Hendrix, R.W., Komives, E.A., Johnson, J.E.: An unexpected twist in viral capsid maturation. Nature **458**(7238), 646–650 (2009)
82. Gan, L., Conway, J.F., Firek, B.A., Cheng, N., Hendrix, R.W., Steven, A.C., Johnson, J.E., Duda, R.L.: Control of crosslinking by quaternary structure changes during bacteriophage hk97 maturation. Mol. Cell **14**(5), 559–569 (2004)
83. Steven, A.C., Heymann, J.B., Cheng, N., Trus, B.L., Conway, J.F.: Virus maturation: dynamics and mechanism of a stabilizing structural transition that leads to infectivity. Curr. Opin. Struct. Biol. **15**(2), 227–236 (2005)
84. Lee, K.K., Gan, L., Tsuruta, H., Hendrix, R.W., Duda, R.L., Johnson, J.E.: Evidence that a local refolding event triggers maturation of hk97 bacteriophage capsid. J. Mol. Biol. **340**(3), 419–433 (2004)
85. May, E.R., Brooks III, C.L.: Determination of viral capsid elastic properties from equilibrium thermal fluctuations. Phys. Rev. Lett. **106**, 188101–188104 (2011)
86. May, E.R., Aggarwal, A., Klug, W.S., Brooks III, C.L.: Viral capsid equilibrium dynamics reveals nonuniform elastic properties. Biophys. J. **100**, L59–L61 (2011)
87. Ivanovska, I.L., de Pablo, P.J., Ibarra, B., Sgalari, G., MacKintosh, F.C., Carrascosa, J.L., Schmidt, C.F., Wuite, G.J.L.: Bacteriophage capsids: tough nanoshells with complex elastic properties. Proc. Natl. Acad. Sci. USA **101**(20), 7600–7605 (2004)
88. Michel, J.P., Ivanovska, I.L., Gibbons, M.M., Klug, W.S., Knobler, C.M., Wuite, G.J.L., Schmidt, C.F.: Nanoindentation studies of full and empty viral capsids and the effects of capsid protein mutations on elasticity and strength. Proc. Natl. Acad. Sci. USA **103**(16), 6184–6189 (2006)
89. Kol, N., Gladnikoff, M., Barlam, D., Shneck, R.Z., Rein, A., Rousso, I.: Mechanical properties of murine leukemia virus particles: effect of maturation. Biophys. J. **91**(2), 767–774 (2006)
90. Carrasco, C., Carreira, A., Schaap, I.A.T., Serena, P.A., Gmez-Herrero, J., Mateu, M.G., de Pablo, P.J.: Dna-mediated anisotropic mechanical reinforcement of a virus. Proc. Natl. Acad. Sci. USA **103**(37), 13706–13711 (2006)
91. Roos, W.H., Gibbons, M.M., Arkhipov, A., Uetrecht, C., Watts, N.R., Wingfield, P.T., Steven, A.C., Heck, A.J.R., Schulten, K., Klug, W.S., Wuite, G.J.L.: Squeezing protein shells: how continuum elastic models, molecular dynamics simulations, and experiments coalesce at the nanoscale. Biophys. J. **99**(4), 1175–1181 (2010)
92. Tama, F., Brooks, C.L.: Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. Annu. Rev. Biophys. Biomol. Struct. **35**, 115–133 (2006)
93. Tirion, M.M.: Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett. **77**(9), 1905–1908 (1996)
94. Landau, L.D., Lifshitz, E.M.: Theory of Elasticity. Pergamon Press, London (1959)
95. Cagin, T., Holder, M., Pettitt, B.M.: A method for modeling icosahedral virions rotational symmetry boundary-conditions. J. Comput. Chem. **12**(5), 627–634 (1991)
96. Tama, F., Brooks III, C.L.: Diversity and identity of mechanical properties of icosahedral viral capsids studied with elastic network normal mode analysis. J. Mol. Biol. **345**(2), 299–314 (2005)
97. Khavrutskii, I.V., Arora, K., Brooks III, C.L.: Harmonic fourier beads method for studying rare events on rugged energy surfaces. J. Chem. Phys. **125**(17), 174108 (2006)
98. Arora, K., Brooks III, C.L. Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. Proc. Natl. Acad. Sci. USA **104**(47), 18496–18501 (2007)

99. Arora, K., Brooks III, C.L. Functionally important conformations of the met20 loop in dihydrofolate reductase are populated by rapid thermal fluctuations. J. Am. Chem. Soc. **131**, 5642–5647 (2009)
100. Lee, M.S., Salsbury, F.R., Brooks III, C.L. Constant-pH molecular dynamics using continuous titration coordinates. Proteins **56**(4), 738–752 (2004)
101. Khandogin, J., Brooks III, C.L. Constant pH molecular dynamics with proton tautomerism. Biophys. J. **89**(1), 141–157 (2005)
102. May, E.R., Armen, R.S., Mannan, A.M., Brooks III, C.L.: The flexible c-terminal arm of the lassa arenavirus z-protein mediates interactions with multiple binding partners. Proteins **78**(10), 2251–2264 (2010)

Mechanisms and Kinetics of Amyloid Aggregation Investigated by a Phenomenological Coarse-Grained Model

Andrea Magno, Riccardo Pellarin, and Amedeo Caflisch

1 Introduction

Amyloid fibrils are ordered polypeptide aggregates that have been implicated in several neurodegenerative pathologies, such as Alzheimer's, Parkinson's, Huntington's, and prion diseases, [1, 2] and, more recently, also in biological functionalities. [3–5] These findings have paved the way for a wide range of experimental and computational studies aimed at understanding the details of the fibril-formation mechanism. Computer simulations using low-resolution models, which employ a simplified representation of protein geometry and energetics, have provided insights into the basic physical principles underlying protein aggregation in general [6–8] and ordered amyloid aggregation. [9–15] For example, Dokholyan and coworkers have used the Discrete Molecular Dynamics method [16, 17] to shed light on the mechanisms of protein oligomerization [18] and the conformational changes that take place in proteins before the aggregation onset. [19, 20] One challenging observation, which is difficult to observe by computer simulations, is the wide range of aggregation scenarios emerging from a variety of biophysical measurements. [21, 22] Atomistic models have been employed to study the conformational space of amyloidogenic polypeptides in the monomeric state, [23–25] the very initial steps of amyloid formation, [26–32] and the structural stability of fibril models. [33–35] However, all-atom simulations of the kinetics of fibril formation are beyond what can be done with modern computers.

A. Magno • R. Pellarin • A. Caflisch (✉)
Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057, Zürich, Switzerland
e-mail: a.magno@bioc.uzh.ch; pellarin@salilab.org; caflisch@bioc.uzh.ch

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_8,

To overcome such computational limitations, simplified models have been developed and used to investigate the kinetics and pathways of oligomerization and fibril formation at different levels of resolution [36]. In this chapter, we first review briefly the simplified models of aggregation. We then present our coarse-grained phenomenological (CGF) model of an amphipathic peptide [37], and its use for studying kinetics and thermodynamics, both in bulk conditions and in presence of other simplified (macro)molecules.

2 Coarse-Grained Models

In coarse-grained models, the complexity of a system (and therefore the computational cost) is reduced by grouping atoms into larger units or "beads," whose mutual interactions are usually approximated by a potential of mean force [38]. Several coarse-grained models of different resolutions have been developed to study aggregation (see Fig. 1). Zhang and Muthukumar [39] have created a cuboid model able of reproducing the features of a nucleation-limited aggregation process. With their so-called "tube" model, Auer and coworkers [40] have shed light upon the conversion of a disordered aggregate into an aggregating nucleus. Higher-resolution models like the one developed by Thirumalai and collaborators [41]

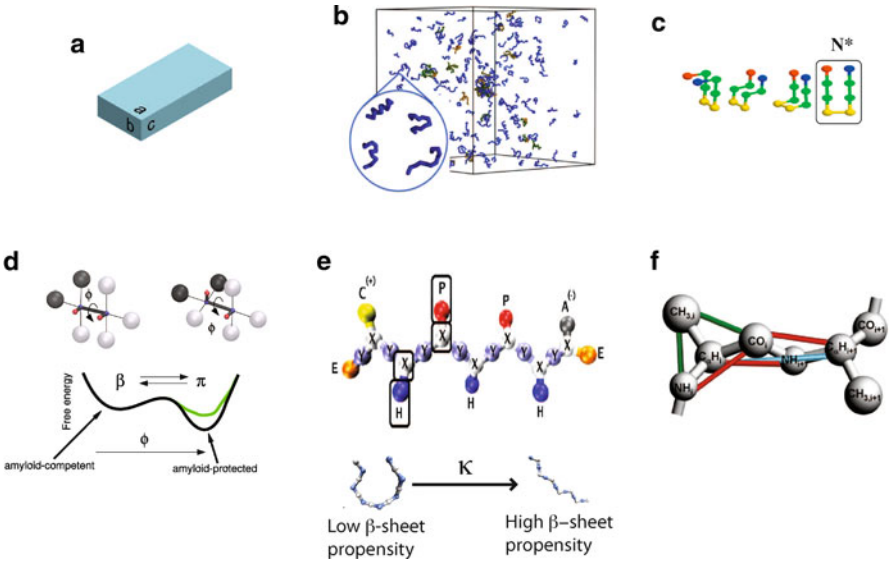

Fig. 1 Main coarse-grained models discussed in Sect. 2. (**a**) Cuboid model [39]; (**b**) tube model [40]; (**c**) lattice model [41]; (**d**) CGF model [37]; (**e**) Shea model [43]; (**f**) Hall model [42]. Reprinted from [36] with permission by Elsevier

or the PRIME model of Nguyen and Hall [42] have also showed disordered aggregates in the early steps of aggregation. Several aggregation scenarios have been described with the three-bead model of Shea and coworkers. [43] In their model, the variation of a parameter related to the dihedral flexibility is able to reproduce different aggregation kinetics and metastable intermediates (amorphous and β-barrel-like), which is in part similar to the CGF model [37]. The main difference between the Shea and CGF models is that the former is based on a coarse-graining from an atomistic description (i.e., "bottom-up" development), whereas the CGF model is purely phenomenological ("top-down") as explained in the next section.

3 The Coarse-Grained Phenomenological Model

The coarse-grained model of an amphipathic peptide developed for studying aggregation kinetics and thermodynamics is a compromise between mesoscopic detail and computational efficiency. It must be stressed that this simplified model does not represent a particular protein sequence, i.e., it has not been generated by grouping into larger beads the atomic structure of a given (poly)peptide. It rather was designed from scratch for emulating the main experimental findings on fibril formation kinetics.

The peptide is approximated by ten spherical beads, four of which represent the "backbone" (small beads) and six the "side chain" (large beads) (Fig. 2, top). The "backbone" beads carry partial charges of $\pm 0.4e$, thereby generating two dipoles; this part of the monomer is designed to interact specifically by intermolecular dipole–dipole interactions. The large beads interact only by van der Waals forces. The nonbonding interaction cutoff is set equal to 20Å. The monomer displays an amphipathic moment, since eight of the ten beads have less favorable van der Waals interactions than the remaining two beads (black spheres in Fig. 2, top). The amphipathicity of the "molecule" allows the formation of amorphous aggregate, such as micellar oligomers, and the assembly of fibrils. In both of these types of aggregates, the hydrophobic spheres are buried and the hydrophilic spheres are exposed. The micellar oligomers are spherical and fluid-like, while the fibrils are ordered and rigid (see below).

The monomer can change its conformation by rotating around the internal dihedral defined by the small beads (Fig. 2, bottom). Using a one-dimensional spline function, [44] a dihedral potential was designed with only two minima separated by a barrier (see Fig. 3). The only parameter that rules the relative populations of the amyloid-prone and amyloid-protected states is the energy difference $dE = E_\pi - E_\beta$ between the conformation with perpendicular dipoles (E_π), which prevents ordered aggregation, and the conformation with parallel dipoles, which is prone to form

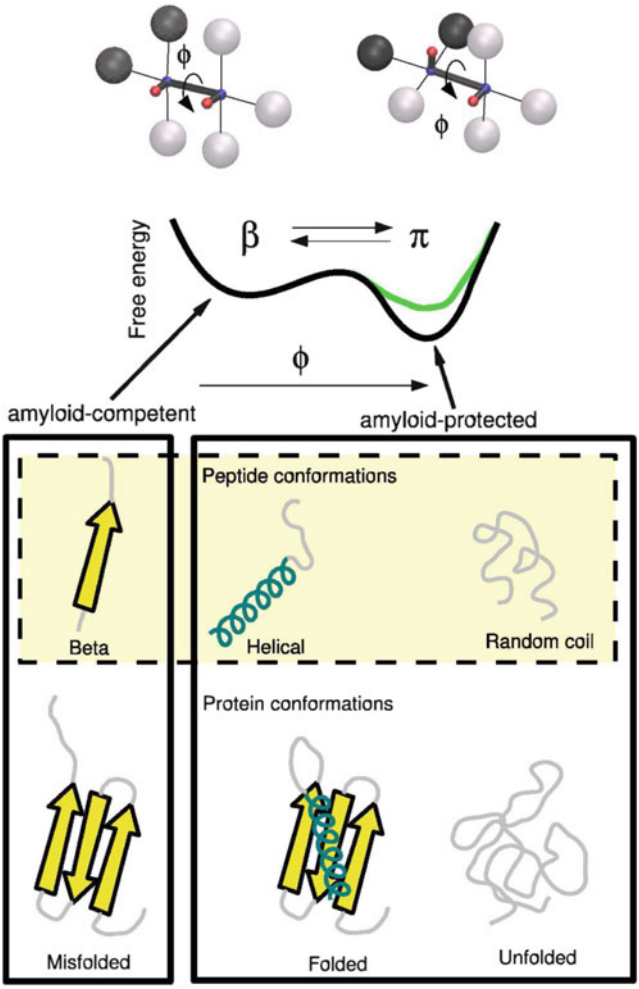

Fig. 2 The CGF model: sticks and beads representations of the monomer in the amyloid-competent state β and the amyloid-protected state π [37]. The large spheres are hydrophobic (*black*) and hydrophilic (*gray*), while the two dipoles are shown with *small red* and *blue spheres*. The size of the spheres does not represent the actual van der Waals radii, which are 2.5Å for the *black* and *gray spheres* and 2.0Å for the *red* and *blue spheres*. The β and π states of the monomer are shown on top of the two corresponding minima of the free energy, plotted as a function of the dihedral angle ϕ of the two dipoles. Reprinted from [50] with permission by Elsevier

fibrils (E_β). The use of a single parameter to model a complex process was inspired by the work of Zhou and Karplus, who have analyzed the folding kinetics of a model protein by varying a single parameter and shown that it is possible to recover several folding scenarios. [45]

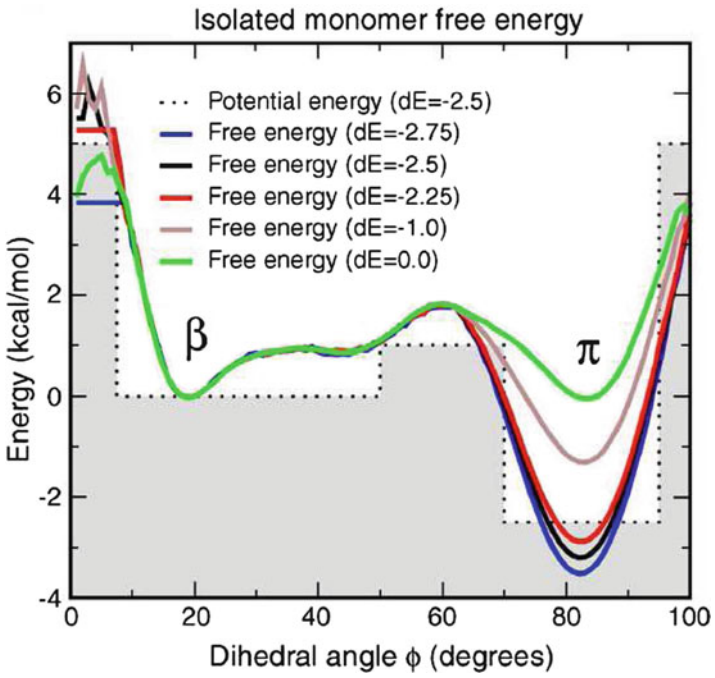

Fig. 3 The *dotted line* is the dihedral potential with a $dE = -2.5$ kcal/mol energy difference between amyloid-protected and amyloid-competent state. The *five continuous lines* represent the free energy profile of the isolated monomer for five different dihedral potentials. Since the peptide has only one degree of freedom, dE is close to the free-energy difference between the two aforementioned states. For instance, when $dE = 0.0$ kcal/mol, the π and β states are equally populated, whereas for $dE = -1.5$, -2.0, and -2.25 kcal/mol, the π state is about 15, 39, and 64 times more populated than the β state, respectively. Reprinted from [37] with permission by Elsevier

4 Aggregation of the CGF Peptide Model in Bulk Solution

Unless specified explicitly, simulations are started from 125 monodispersed monomers of the CGF peptide in a cubic box with a size of 290Å, corresponding to a concentration of 8.5 mM. After minimization and equilibration, simulations are performed with Langevin dynamics at 310 K with a very small friction coefficient of 0.01 ps^{-1} using CHARMM [46].

4.1 Aggregation Kinetics and Pathways

The range of aggregation kinetics of the CGF model is shown in Fig. 4, where the normalized degree of polymerization as a function of time is plotted for different

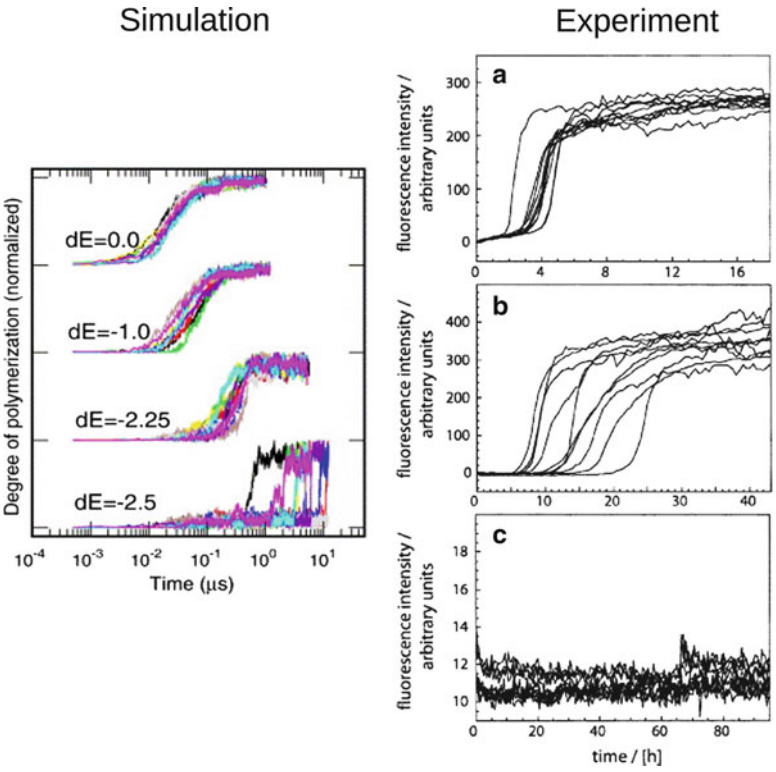

Fig. 4 Influence of amyloidogenic tendency on aggregation kinetics. (*Left*): Time series of the fraction of ordered aggregation evaluated at four values of the amyloidogenic tendency, from very prone to form fibrillar aggregates ($dE = 0.0$ kcal/mol) to marginal propensity ($dE = -2.5$ kcal/mol). Ten independent simulations are shown for each dE value. (*Right*): Fluorescence intensity (degree of aggregation) of V18I (**a**), V18Q (**b**), and V18P (**c**) mutants of Aβ_{40} [48]. Note that the ns-μs timescales in the CGF model simulations are much shorter than in the experiments (hours) because of the much higher concentration in the former (8.5 mM) than in the latter (120 μM). Reprinted from [37] (*left*) and [48] (*right*) with permission by Elsevier (*left*) and Jon Wiley & Sons (*right*)

values of the amyloidogenic propensity dE. The extent of aggregation is controlled by counting the number of polar contacts: a polar contact is formed whenever two dipoles of different monomers are closer than 5Å. Three different kinetic phases are visible: lag, elongation, and final monomer–fibril equilibrium. The variable length of the lag phase and the higher heterogeneity at longer lag times are indicative of a stochastic nucleation [47]. Fibril formation is much slower for the β-unstable models ($dE = -2.5$ and -2.25 kcal/mol) than the β-stable models ($dE = -1.0$ and 0.0 kcal/mol). Both the lag phase and the elongation kinetics are affected by the single free parameter dE of the CGF model. Interestingly, the kinetics of aggregation are qualitatively consistent with the experimental data on single-

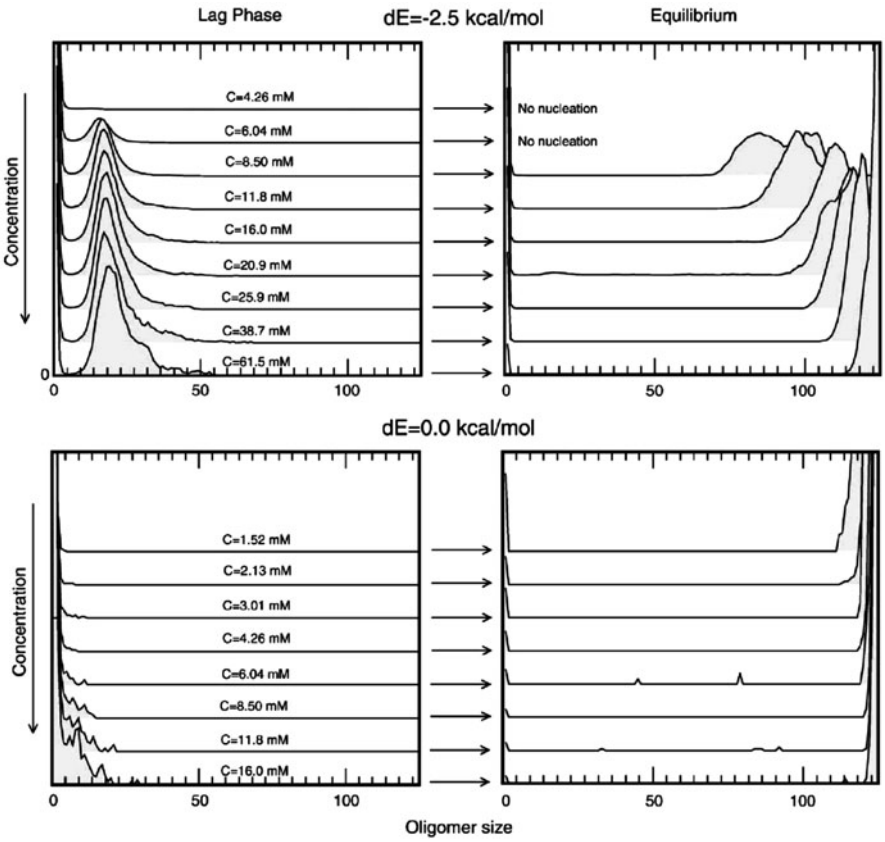

Fig. 5 Oligomer size histograms of the $dE = -2.5$ kcal/mol potential (*top*) and $dE = 0.0$ kcal/mol (*bottom*) calculated at the lag phase (*left*) and at the final equilibrium (*right*). The z-dimension represents the relative probability. Note that most of the results were obtained at a concentration C = 8.5 mM which is the lowest value at which fibril formation takes place within a reasonable simulation time for the $dE = -2.5$ kcal/mol model (10 μs in about 17 days on a single Xeon 5410 processor). Reprinted from [37] with permission by Elsevier

point mutants of Aβ_{40} [48], which have shown that the β-sheet propensity and hydrophobicity affect the features of the aggregation process. This comparison shows that although the CGF model does not represent any particular polypeptide sequence, variations of the single parameter dE emulate the behavior observed for (slightly) different amyloidogenic sequences. Moreover, the anticorrelation between the length of the lag phase and the rapidity of the fibril elongation has also been observed experimentally on several samples prepared from amyloidogenic (poly)peptide sequences. [49]

The distribution function $p(N)$ of the oligomer size N evaluated at the lag phase or at the final equilibrium is depicted in Fig. 5. The monomer peak ranges from $N = 1$ to $N = 7$, the micellar peak from $N = 8$ to $N = 60$, and the fibrillar peak

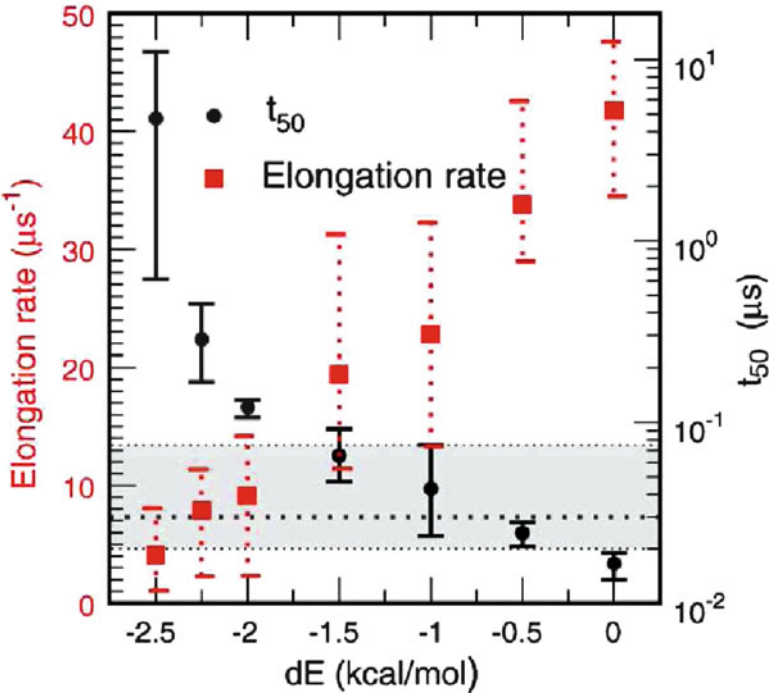

Fig. 6 Influence of the amyloidogenicity parameter dE on the kinetics of the CGF model. The time needed to reach 50% of the maximal amplitude t_{50} (*black circles* and y-axis legend on the *right*) and the elongation rate (*red squares* and y-axis legend on the *left*) are displayed for seven dE values. Symbols represent the average value of ten independent runs, and the error bars are the maximum and minimum values. The *broken line* and the *gray band* indicate the average and the max–min values for the time of micelle formation, respectively. Reprinted from [37] with permission by Elsevier

from $N = 61$ to $N = 125$. The micellar peak is present for the $dE = -2.5$ kcal/mol model at the lag phase, but disappears at the final equilibrium, where the fibril and the monomers are the only co-existing species. For the β-stable potential $dE = 0.0$ kcal/mol, the micellar peak is not observed at any concentration value. Indeed, a comparison of the lag times with the times of micelle formation (Fig. 6) shows that the fibril formation kinetics of the β-unstable and β-stable models are, respectively, slower and faster than micelle formation. In fact, micelles are intermediates consisting mainly of monomers in the π state, whereas the polymerization of β-stable monomers directly yields fibrils.

This observation is confirmed also by the analysis of aggregation pathways. [50] A total of 100 Langevin dynamics simulations for different values of dE were clustered according to three progress variables: the size of the largest aggregate N_{la}, the number of monomers in the β-state within the largest aggregate N_{la}^{β}, and the number of protofilaments in the largest aggregate $N_{\mathrm{la}}^{\mathrm{pf}}$, where a protofilament is defined as a file of monomers with intermolecular dipolar interactions parallel

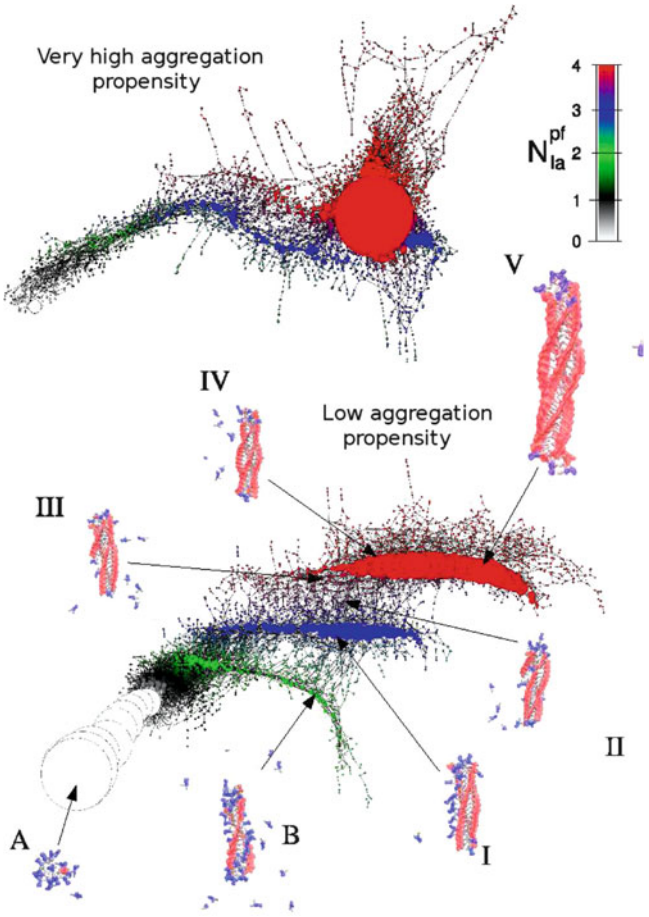

Fig. 7 Aggregation state network. The size of the largest aggregate N_{la} and its number of protofilaments $N_{\mathrm{la}}^{\mathrm{pf}}$ were used to cluster all simulation snapshots into states (i.e., nodes of the network). The size and the color of nodes correspond to the statistical weight and the number of protofilaments $N_{\mathrm{la}}^{\mathrm{pf}}$, respectively. Links are direct transitions within 0.5 ns of Langevin dynamics. Note the much higher heterogeneity of protofibrillar intermediates for the β-unstable ($dE = -2.5$ kcal/mol, *bottom*) as compared to the β-stable ($dE = -1.5$ kcal/mol, *top*) model. The *insets* show (proto)fibrillar structures that are representative of each region of the aggregation state network. In these structures, monomers in the amyloid-competent conformer β and amyloid-protected conformer π are in *red* and *blue*, respectively. Furthermore, hydrophobic spheres are *gray* and hydrophilic spheres are not shown for visual clarity. Reprinted from [50] with permission by Elsevier

with its axis. The aggregation state network (Fig. 7) is a graph in which the states and direct transitions observed during the Langevin dynamics simulations are displayed as nodes and links, respectively. Furthermore, the size of each node reflects the statistical weight of the corresponding state. Micellar oligomers

(white nodes, $N_{\mathrm{la}} \sim 20$, $N_{\mathrm{la}}^{\mathrm{pf}} = 0$) and fibrils (red nodes, $N_{\mathrm{la}} \sim 100$, $N_{\mathrm{la}}^{\mathrm{pf}} = 4$) are the most populated states during the lag phase and the final equilibrium, respectively. Strikingly, a greater variety of aggregation mechanisms emerges for the poorly amyloidogenic CGF peptide model (see Fig. 7, bottom) than the highly amyloidogenic CGF peptide model (see Fig. 7, top). Indeed, the former shows the presence of intermediates, i.e., protofibrils consisting of only two (green nodes) or three (blue nodes) protofilaments. According to this analysis, it is reasonable to expect that a mutation that decreases the β-aggregation tendency could result in a greater variety of prefibrillar aggregates, as in the case of the Arctic mutant (E22G) of the Alzheimer's Aβ peptide and the A30P mutant for α-synuclein, for which a more pronounced in vitro formation of oligomers and protofibrils was observed. [51, 52]

4.2 Mechanism of Nucleation

The nucleation properties of the CGF model are investigated by evaluating the probability of fibril formation for β-subdomains, i.e., the clusters of interacting β-monomers. The nucleus, defined as the oligomer containing a β-subdomain with a 50% probability to form a fibril, shows an increasing size upon destabilization of the β-state. Significantly different nucleation mechanisms are observed upon variation of the amyloidogenicity parameter dE (Fig. 8). For high values of the amyloidogenic propensity ($-2.0 \leq dE \leq 0.0$ kcal/mol), the nucleus size is submicellar, and nucleation is simply the aggregation of monomers in the β-state. On the contrary, for poorly amyloidogenic peptides, nucleus formation requires either spatial proximity of several monomers in the β-state ($dE = -2.25$ kcal/mol) within a micelle or collision of two peptide micelles with merging of their β-subdomains ($dE = -2.5$ kcal/mol). The variety of aggregation scenarios is also observed experimentally. An unstructured peptide with a marginally stable β-prone state like Aβ_{40} [53, 54] visits oligomeric intermediates in the lag phase, and has a very weak dependence of the elongation rate on concentration due to the monomer-micelle equilibrium. This mechanism corresponds to the nucleated conformational conversion proposed by Serio et al. [55] On the other hand, a functional and nonpathological amyloid in mammalians [56] lacks on-pathway intermediates and corresponds to the highly amyloidogenic CGF peptide model. Once more, by varying the only free parameter dE of the CGF model, it is possible to describe the aggregation properties of a wide and diverse range of (poly)peptide sequences.

4.3 Concentration Effects

The dE parameter of the CGF model has a strong influence on the concentration dependence of the fibril-formation kinetics. In agreement with the above-mentioned mechanism of nucleation, CGF peptides poorly prone to aggregation nucleate only

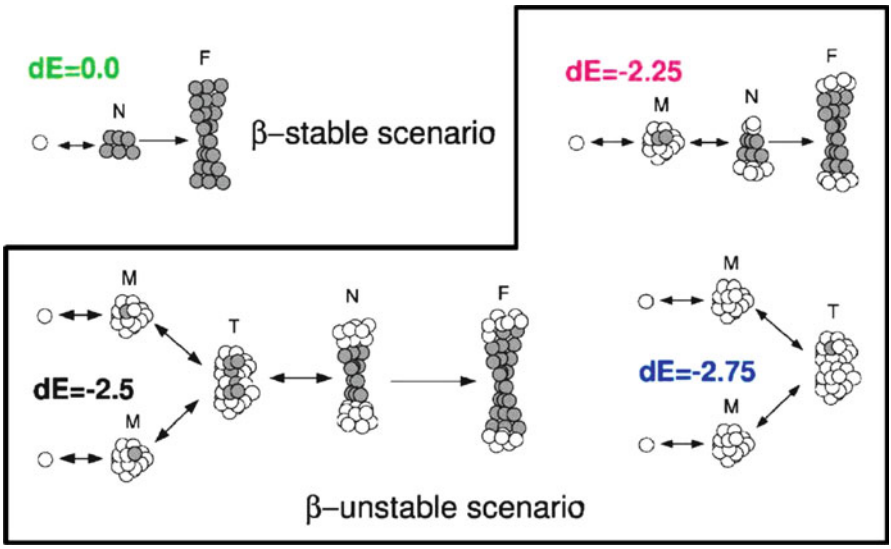

Fig. 8 Observed nucleation scenarios of the CGF peptide model. *Black* and *white* circles represent the amyloid-competent conformer β and amyloid-protected conformer π, respectively. CGF peptides with high values of amyloidogenic tendency nucleate without intermediates, while poorly amyloidogenic CGF peptides can nucleate either through micelle-sized oligomers ($dE = -2.25$ kcal/mol) or transient oligomers larger than a micelle ($dE = -2.5$ kcal/mol). A further stabilization ($dE = -2.75$ kcal/mol) of the protected state prevents fibril formation within the simulation time of about 20 μs. M, micelle; N, nucleus; T, transient oligomer; F, fibril. Reprinted from [37] with permission by Elsevier

at concentration values larger than the critical concentration of peptide micelle formation, whereas CGF peptides with a high value of amyloidogenicity nucleate even at lower concentrations (Fig. 9, left). Furthermore, the dependence of the elongation rate on the concentration is only marginal at low amyloidogenic tendency (Fig. 9b). The reduced concentration dependence originates from competitive polymerizations, i.e., the elongation of the fibril and the presence of micellar oligomers. This observation is a consequence of the monomer–micelle equilibrium of the CGF peptide model, which maintains a nearly constant concentration of isolated monomers [58].

4.4 Amyloid Fibril Polymorphism

Experiments based on electron and atomic force microscopy as well as solid-state NMR spectroscopy revealed that changing the samples conditions, such as the pH [59] or the cosolvent concentration, [60] or introducing a mechanical perturbation [61, 62] results in different amyloid fibril morphologies. Furthermore,

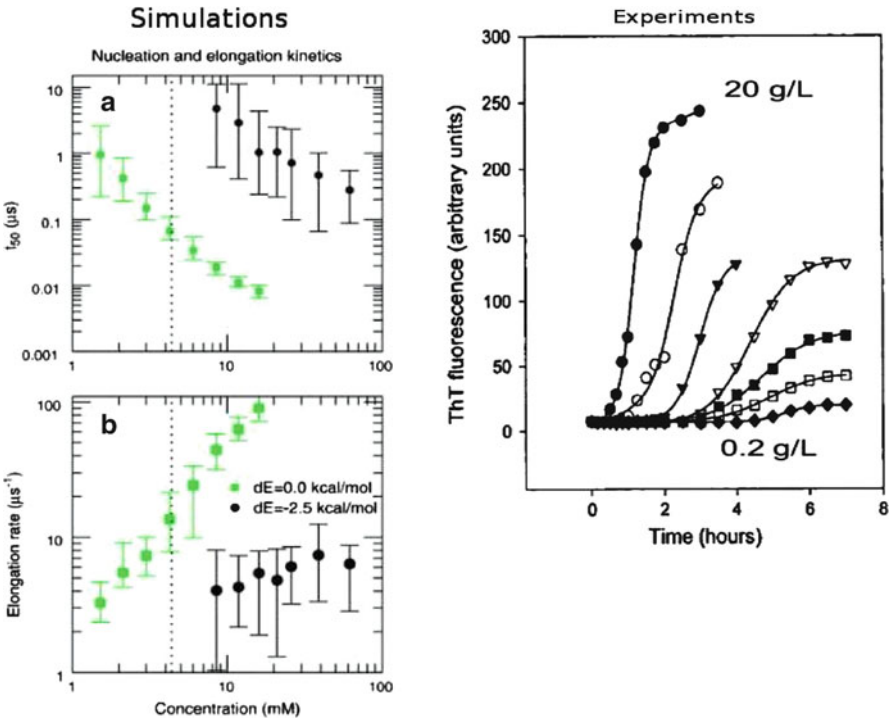

Fig. 9 Influence of peptide concentration on aggregation kinetics. (*Left*) Effect of concentration on the lag phase time t_{50} (**a**) and elongation rate (**b**) for low and high values ($dE = -2.5$ kcal/mol, *black circles*; $dE = 0.0$ kcal/mol, *green squares*, respectively) of the amyloidogenic tendency. The symbols represent the average value calculated from 15 simulations for $dE = -2.5$ kcal/mol and 10 simulations for $dE = 0.0$ kcal/mol. The error bars represent the minimum and the maximum values. The *vertical dotted line* indicates the critical concentration of micelle formation. (*Right*) Influence of the initial monomeric concentration on the kinetics of insulin fibril formation as measured by Thioflavin T fluorescence [57]. Note that the higher the concentration of monomeric insulin is at the beginning of the experiments, the shorter is the lag phase and the faster is the elongation rate. Reprinted from [37] (*left*) and [57] (*right*) with permission by Elsevier (*left*) and American Chemical Society (*right*)

even within the same sample, a number of coexisting morphologies can be detected. [59, 63] Recently, it was observed that the CGF peptide model is able to generate fibrils with distinct morphologies. [64] Interestingly, the populations of the different morphologies are strongly and nontrivially influenced by the amyloidogenic propensity dE, and two main mechanisms for fibril morphogenesis emerge. When the CGF peptide is highly prone to aggregate ($dE = -1.5, -2.0$ kcal/mol), the morphogenesis is under thermodynamic control, meaning that the morphology with the highest stability will emerge with the highest probability. In contrast, when the CGF peptide has a low amyloidogenic tendency ($dE = -2.25, -2.5$ kcal/mol), the fibril morphogenesis is under kinetic control. The morphologies that nucleate

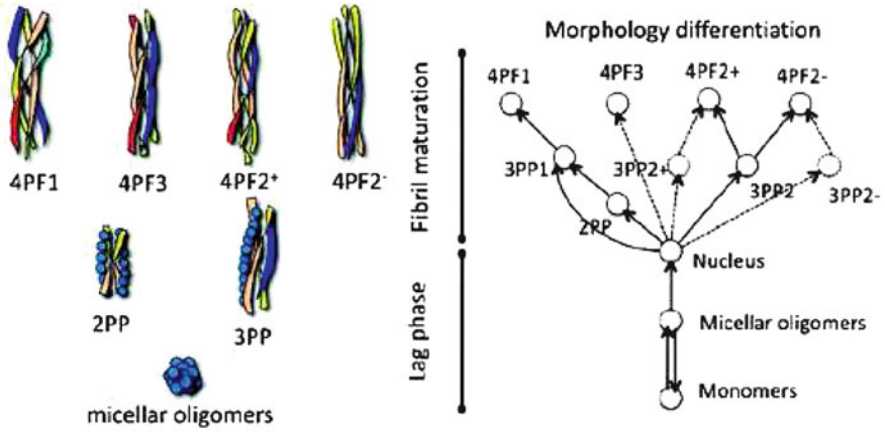

Fig. 10 Morphology differentiation and kinetic control of fibril polymorphism. (*Left*) Morphologies of mature fibrils and prefibrillar species. (*Top*) Mature fibrils display a 4-protofilament structure (4PF). The 4PF morphologies have different orientation of the protofilaments, organization of up and down protofilaments, and thickness of the fibril. (*Bottom*) The prefibrillar species are: the micellar oligomers M, consisting of π-monomers (*blue beads*) aggregated through hydrophobic forces; the 2-protofilament protofibril (2PP), and the 3-protofilament protofibril (3PP), which are early stages of fibril maturation, where the π-monomers are deposited onto the lateral surface of the fibril, and the β-monomers make up the protofilaments (*colored ribbons*). (*Right*) Branched tree illustration of the morphology differentiation process as observed in the simulations. Reprinted from [64] with permission by American Chemical Society

more readily are not necessarily the most stable ones, but those whose precursors are kinetically more accessible, as revealed by the free energy profiles of the fibrillation. [64] For the low amyloidogenic scenario, the process of morphology differentiation can be represented by a branched tree (Fig. 10). During the lag phase, the micellar oligomers are in equilibrium with the dispersed monomers. The early morphology differentiation occurs at the nucleation step, where the formation of the protofibrillar intermediates is regulated by the structural bifurcation of the nucleus. The 2PP and 3PP1 intermediates are competent to 4PF1 fibrils, while the 3PP2 intermediate is competent to 4PF2(+,−) fibrils. Alternatively, the presence of 3PP2+ and 3PP2− intermediates that are directly competent to 4PF2+ and 4PF2− fibrils, respectively, has been observed, although these pathways were not quantitatively analyzed. Finally, the pathway of formation of 4PF3 fibrils was not investigated in detail, due to the small number of nucleation events of this morphology. The multiple-pathways process observed here has a close similarity with the scenario described by Goldsbury et al., [65] where two different morphologies of Aβ have distinct maturation pathways, either with or without the presence of metastable protofibrils.

5 Aggregation in the Presence of Lipid Vesicles and Inert Crowders

Amyloid aggregation in vivo does not occur in bulk solution. Rather, it takes place in the extracellular space, whose composition includes metabolites and proteins, or within the cell, which is usually densely occupied by (macro)molecules like proteins, nucleic acids, and polysaccharides, as well as macromolecular assemblies and organelles. [66] Several research groups have investigated the interactions between lipid vesicles and amyloid aggregates, [67–69] whose accumulation on the surface of lipid bilayers was observed to cause membrane damage. Aggregation has also been studied in crowded media, [70–72] where the thermodynamics and kinetics of aggregation are are expected to sensibly change.

5.1 Effect of Lipid Bilayers on CGF Peptide Aggregation

A three-bead model of a lipid molecule has been developed to study the CGF peptide model aggregation kinetics in the presence of a lipid vesicle. [73,74] Several independent Langevin simulations at 310 K have been performed for four values of dE with 125 peptides initially monodispersed in a cubic box of length 290Å and a preequilibrated unilamellar bilayer vesicle made up of 1,000 lipids. Depending on the lipid/peptide van der Waals coupling parameter λ, between 50% and 80% of the CGF peptides are located on the lipid vesicle surface after the initial equilibration phase, i.e., before fibril formation (Table 1).

The effect of lipid bilayers on aggregation kinetics for different values of amyloidogenicity is reported in Table 2. Highly amyloidogenic peptides fibrillate more rapidly in the presence of lipid vesicles than in their absence, while the opposite is observed for peptides of low amyloidogenicity. The faster aggregation kinetics of highly amyloidogenic peptides is a consequence of their higher effective concentration on the lipid bilayer relative to the bulk. In contrast, despite the same increase of peptide concentration on the vesicle surface, fibrillation of peptides with low amyloidogenic propensity is slower in the presence of lipid

Table 1 Three-bead lipid and surfactants models used with CGF peptide model

Type of molecule	$R_{hydrophilic}$ [nm]	$R_{hydrophobic}$ [nm]	$\epsilon_{hydrophilic}$ [kcal/mol]	$\epsilon_{hydrophobic}$ [kcal/mol]	λ^a	Fraction of CGF peptides bound to lipid vesicles	Reference
Lipid	0.31	0.3	−0.1	−1.265	0.87–0.90	50%	[73]
Lipid	0.31	0.3	−0.1	−1.265	0.95	80%	[74]
Surfactant	0.35	0.3	−0.1	−0.8	1		[80]

Different scaling is used to model different systems, i.e., surfactants ($\lambda = 1$), moderately attractive ($\lambda \leq 0.9$), and strongly attractive ($\lambda \approx 0.95$) lipid bilayers

[a] Scaling factor for the vdW interactions between lipids or surfactants and peptides

Table 2 Characteristic lag-time of aggregation t_{50} for CGF peptide model for different amyloidogenic tendency in the presence or absence of lipid vesicles [73]

Scaling of peptide/ lipid interactions	Amyloidogenicity	Number of runs with fibril formation	Lag time t_{50} [ns]	
			with membrane	without membrane
0.87	High	10/10	11 ± 1	19 ± 3
	Interm.	29/29	89 ± 29	56 ± 15
	Low	17/20	**958** $\pm$ 503	124 ± 28
	Very low	0/10	>**2000**	318 ± 133
0.90	High	10/10	10 ± 1	19 ± 3
	Interm.	30/30	69 ± 23	56 ± 15
	Low	0/20	>**2000**	124 ± 28
	Very low	0/20	>**2000**	318 ± 133

Values in boldface are significantly larger in the presence of the vesicles

vesicles. As mentioned in section III, peptides with low amyloidogenic potential can fibrillate only after aggregating into spherical oligomeric intermediates with hydrophobic interior and hydrophilic surface. In the simulations with lipids, such oligomeric intermediates form in the bulk but not on the vesicle. Fibrillation of low amyloidogenic peptides therefore takes place in the bulk and is slower than in the absence of a vesicle due to the lower effective concentration of peptides in the solvent. These simulation results are consistent with and explain the apparently contradictory experimental observations on faster aggregation of the Aβ [68] or α-synuclein [75] peptides in the presence of lipid surfaces and slower aggregation of insulin (which has lower amyloidogenicity), [76] and have been confirmed by recent studies on the aggregation properties of human islet amyloid polypeptide $hIAPP_{1-19}$ in presence of lipid vesicles [77].

To investigate the influence of the CGF peptides on the lipid bilayer, the simulations were initiated with 20 spherical probes inside the vesicle. It was observed that leakage from the lipid vesicle is enhanced during fibril formation but not by the mature fibril [73]. More precisely, a comparison between the fibrillation and probe release rates (Fig. 11, left) revealed that probe release is fastest during fibril growth, whereas the kinetics of probe release in the presence of mature fibrils is as slow as in the absence of peptides, indicating that mature fibrils do not damage the integrity of the vesicle. Rather, the ongoing process of aggregation on the vesicle results in bilayer surface defects. This observation explains why for some amyloidogenic peptides there exist mutants that form fibrils more rapidly and are more toxic than the wild-type peptides, even though their fibrils are not toxic. [78] Moreover, these computational results are in agreement with the experiments performed by Engel et al. on membrane damage caused by human islet amyloid polypeptide (hIAPP) fibril growth [69] (Fig. 11 right).

It has also been hypothesized that formation of toxic oligomers that induce membrane leakage could be the result of a backward production of oligomers from the mature fibril. [79] Interestingly, by modulating the attraction between the CGF

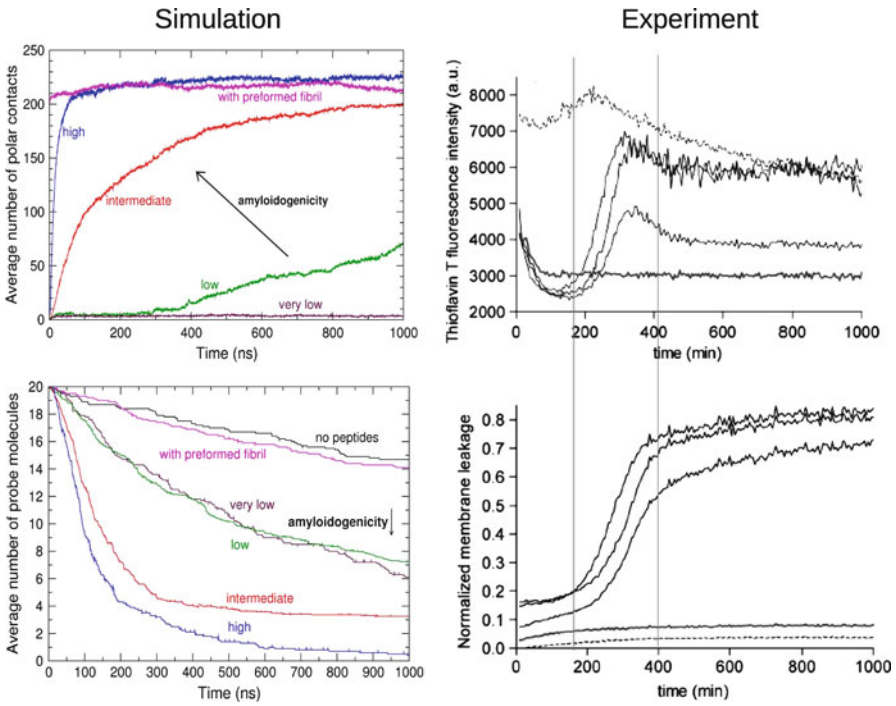

Fig. 11 Comparison of CGF model simulations and experimental data on fibril formation in the presence of lipid vesicles. (*Left*) Simulation results. (*Left*, *top*) Influence of peptide amyloidogenicity on fibril growth kinetics and vesicle leakage. A single parameter, the energy difference between amyloid-competent and amyloid-protected conformations of the peptides, is varied in different simulations to tune amyloidogenicity. Time series of the average number of ordered polar contacts between monomers (corresponding to the degree of fibrillation). (*Left*, *bottom*) Average number of probes inside the vesicle, in the absence (*black*) or presence (colors) of peptides, and for simulations where a preformed fibril was used instead of dispersed peptides. (*Right*) Experimental data: Effect of human islet amyloid polypeptide (hIAPP) fibril growth on membrane leakage. [69] Thioflavin T fluorescence intensity (*Right*, *top*) and induced membrane leakage (*Right*, *bottom*) of three hIAPP samples (*black curves*), together with representative traces for mouse IAPP variant which is known to be non-toxic (*gray lines*) and preformed hIAPP fibrils (*dashed lines*) are shown. The *two vertical lines* are shown to facilitate comparison of the kinetic traces in *top* and *bottom panel*. Reprinted from [73] (*left*) and [69] (*right*) with permission by Elsevier and by Copyright 2008 National Academy of Sciences, U.S.A., respectively

peptides and the membrane, fibril disaggregation into soluble backward oligomers has been observed. [74] The disaggregation process is driven by entropy and results in soluble protofibrillar oligomers. The protofibrillar oligomers are larger, more ordered, and more stable than those observed during the aggregation process and, importantly, are not detected in disaggregation simulations carried out in bulk solution, i.e., in the absence of lipid vesicles.

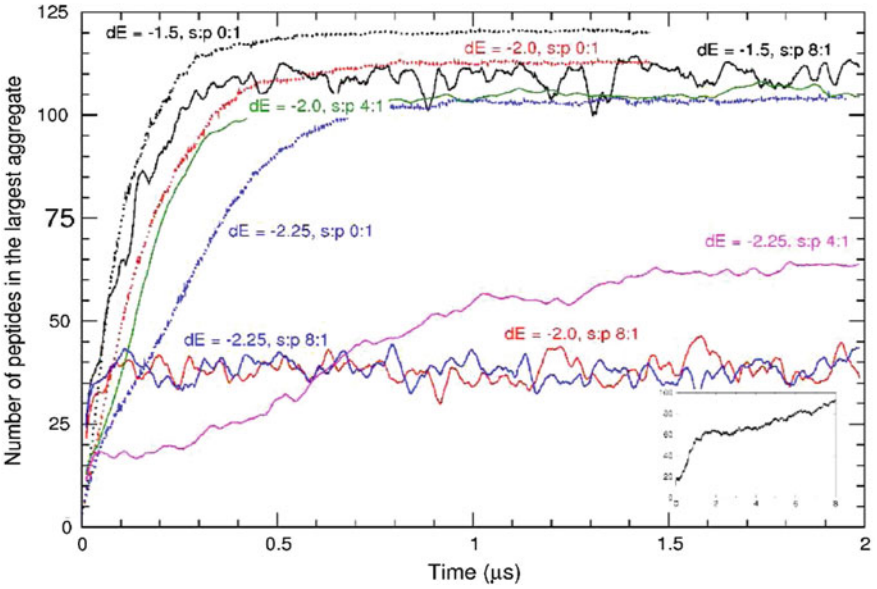

Fig. 12 Amyloid aggregation in the presence and absence of surfactants. The number of peptides in the largest aggregate is averaged over 20 runs at each simulation condition, i.e., for each value of aggregation propensity (dE) and each surfactant/peptide concentration ratio (s:p). Free peptides display fast aggregation without any noticeable lag phase. At surfactant:peptide ratio of 4:1, peptides aggregate into fibrils, but aggregation is much slower for peptides with low amyloidogenicity (see the *inset* for $dE = -2.25$ and surfactant:peptide ratio of 4:1, extended to 8 μs). At a surfactant:peptide ratio of 8:1, aggregation is completely inhibited for $dE = -2.0$ and -2.25 kcal/mol, while highly amyloidogenic peptides ($dE = -1.5$ kcal/mol) are barely affected

5.2 *Effect of Surfactants on CGF-Peptide Aggregation*

Surfactant molecules have been modeled using a similar three-bead model as that used for lipids. The surfactant model differs from lipid models used previously in two parameters (Table 1): the minimum of the van der Waals energy of the two hydrophobic beads is less favorable, and the radius of the hydrophilic bead is larger to enable the formation of amorphous aggregates. Using these parameters, the surfactant solution is not dominated by a micellar phase. Rather, the surfactants are organized either as dispersed monomers or disordered aggregates [80].

In the absence of surfactants, all peptide models form fibrils within 1 μs without any discernible lag phase (Fig. 12, dotted lines). At a surfactant:peptide ratio of 8:1, the fibril formation kinetics of peptides with $dE = -1.5$ kcal/mol are almost unaffected, whereas already at a ratio of 4:1 the ordered self-assembly of peptides with $dE < -2.0$ kcal/mol is significantly slower (Fig. 12, solid lines) mainly because of a longer lag phase. Moreover, for low-amyloidogenic peptides ($dE \leq -2.0$ kcal/mol) no fibrillation is observed within the simulation length of

2 μs at a surfactant:peptide ratio of 8:1, but instead oligomers of $\sim$40 peptides form. These simulation results show that at a fourfold molar excess of surfactant, the inhibition of fibrillation already depends strongly on the amyloidogenicity of the CGF peptide model.

5.3 Macromolecular Crowding Effect on CGF Peptide Aggregation

Simulations with the CGF peptide model together with softly repulsive spheres have been carried out to assess the influence on the aggregation kinetics of excluded volume and hindered peptide diffusion due to macromolecular crowding. [81] As in the case of lipid-bilayer vesicles, the net effect of macromolecular crowding crucially depends on the amyloidogenicity tendency of the CGF peptide. For peptides with low aggregation propensity, the self-association process is transition-state limited, where the kinetic bottleneck is the formation of the fibril nucleus. In this case, since the oligomers, including the nucleus, are thermodynamically favored (with respect to the isolated monomers) by the excluded volume effect, macromolecular crowding accelerates peptide assembly and has an effect analogous to that of an increase in peptide concentration (Fig. 13, left). This trend is analogous to that observed experimentally by Munishkina et al., who have studied the effect of increasing the PEG concentration on the α-synuclein aggregation process. [71]

On the other hand, when the aggregation mechanism is fast and proceeds directly from monomers to fibril, the process is diffusion limited, and the thermodynamic stabilization of oligomers is less important than the reduction in peptide mobility. In this case, the bottleneck is not the formation of the nucleus; the rate-limiting step for peptides that show a direct aggregation mechanism is the elongation of the fibril. Therefore, in this case macromolecular crowding is much less efficient in accelerating the self-association of peptides than an equivalent increase in peptide concentration, since the peptides diffusion is hindered by the crowders (Fig. 13, right).

6 Conclusion

Atomistic simulations of aggregation are limited by short timescale, while experimental approaches to amyloid fibril formation have insufficient spatial resolution. Coarse-grained models of polypeptide aggregation sacrifice atomistic detail to reach timescales that allow the comparison with and interpretation of experimental data. The models presented in this chapter have shed light upon amyloid aggregation kinetics and mechanisms, which is helpful to formulate a unified picture of the available experimental data.

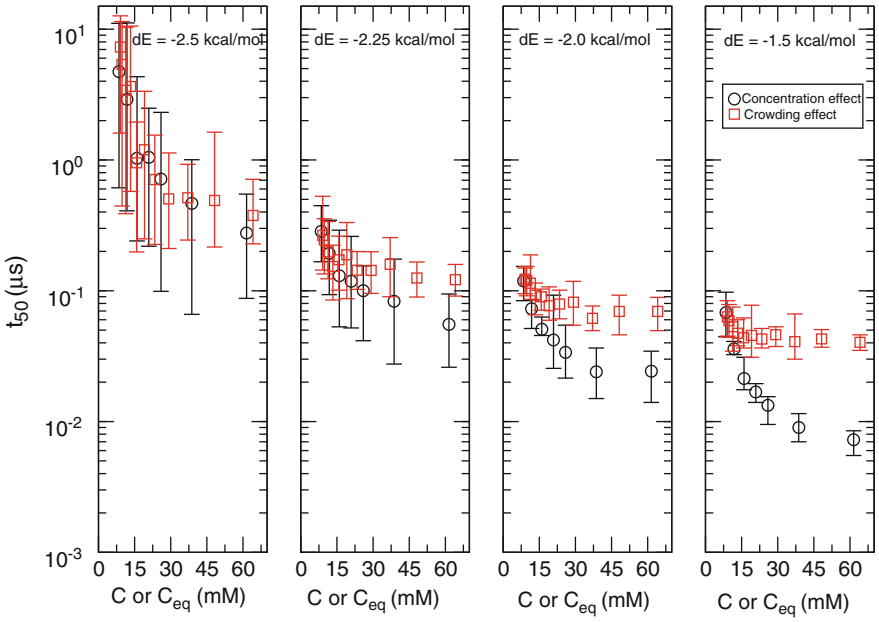

Fig. 13 Differences in aggregation kinetics upon raising peptide concentration or crowders content depend on amyloidogenicity. The time t_{50}, at which the growing fibril has reached 50% of the polar contacts of the mature fibril, is shown as a function of concentration. *Black circles* are t_{50} values calculated at different peptide concentrations in the absence of crowders, while *red squares* are t_{50} values at different equivalent concentrations C_{eq} obtained by varying the number of crowders. Symbols represent the average value of ten independent runs and the error bars are the minimum and maximum values. Reprinted from [81] with permission by American Chemical Society

The CGF model has only one tunable parameter, the difference dE between the energy of the amyloid competent and the amyloid-protected state of the monomer. [37] Variations of this parameter reproduce several aggregation scenarios, both under homogeneous and heterogeneous conditions. It is important to highlight that the CGF model does not mimic any particular amyloid (poly)peptide sequence. However, the different aggregation kinetics obtained with this model can be directly compared with experiments carried out with specific proteins. In Table 3 are reviewed the principal characteristics of the aggregation process for both the high and low amyloidogenic tendency, and in both cases several examples of real amyloid-forming (poly)peptide sequences are listed. It is important to note that (coarse-grained) simulations, e.g., those with the CGF [37] and Shea [43] models, allow for the emulation of conditions and/or phenomena that are not accessible by (standard) experiments. As an example, the possibility to change solely the intrinsic conformational landscape of a monomer without affecting the intermonomer interactions is an advantage of the (coarse-grained) simulation methods with respect to conventional experimental techniques such as mutagenesis

Table 3 Influence of amyloidogenic propensity on the aggregation kinetics and pathways of the CGF model [37]

High amyloidogenicity	Low amyloidogenicity	Reference
Small nucleus	Large nucleus	[37]
Fast fibril formation	Slow fibril formation	[37]
Downhill	Micellar intermediates	[37]
No intermediates	Protofibrillar intermediates	[50]
Single pathway	Multiple pathways	[50]
Strong concentration dependence	Growth rate marginally dependent on concentration	[37]
Polymorphism under thermodynamic control	Polymorphism under kinetic control	[64]
Can promote membrane leakage	Does not promote membrane leakage	[73]
Slightly accelerated by membranes	Decelerated by membranes	[73]
Marginally influenced by surfactants	Decelerated by surfactants	[80]
Not accelerated by macromolecular crowding	Accelerated by macromolecular crowding	[81]
Phe-Phe, GNNQQNY, transthyretin, Aβ_{42}	Aβ_{40}, Sup35, prion protein, myoglobin	-

The last line lists some examples but it must be stressed that amyloidogenic tendency strongly depends on external conditions, so that the same polypeptide sequence can show drastically different amyloidogenic tendency depending on pH, temperature, etc.

and solvent-induced conformational changes, by which it is not possible to decouple changes in intra- from intermolecular interactions.

In conclusion, a slight modification of the free energy profile of an extremely simplified model of an amphipathic peptide is sufficient to observe a wide range of different fibril formation mechanisms, providing a unifying description of the heterogeneity of the experimentally observed kinetics of amyloid fibril formation.

References

1. Dobson, C.M.: Protein folding and misfolding. Nature **426**, 884–890 (2003).
2. Lansbury, P.T., Lashuel, H.A.: A century-old debate on protein aggregation and neurodegeneration enters the clinic. Nature **443**, 774–779 (2006).
3. Fowler, D.M., Koulov, A.V., Balch, W.E., Kelly, J.W.: Functional amyloid–from bacteria to humans. Trends Biochem. Sci. **32**, 217–224 (2007).
4. Maji, S.K., Perrin, M.H., Sawaya, M.R., Jessberger, S., Vadodaria, K., Rissman, R.A., Singru, P.S., Nilsson, K.P.R., Simon, R., Schubert, D., et al.: Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. Science **325**, 328–332 (2009).
5. Greenwald, J., Riek, R.: Biology of amyloid: structure, function, and regulation. Structure **18**, 1244–1260 (2010).
6. Broglia, R.A., Tiana, G., Pasquali, S., Roman, H.E., Vigezzi, E.: Folding and aggregation of designed proteins. Proc. Natl. Acad. Sci. USA **95**, 12930–12933 (1998).

7. Gupta, P., Hall, C.K., Voegler, A.C.: Effect of denaturant and protein concentrations upon protein refolding and aggregation: a simple lattice model. Protein Sci. **7**, 2642–2652 (1998).
8. Harrison, P.M., Chan, H.S., Prusiner, S.B., Cohen, F.E.: Thermodynamics of model prions and its implications for the problem of prion protein folding. J. Mol. Biol. **286**, 593–606 (1999).
9. Urbanc, B., Cruz, L., Yun, S., Buldyrev, S.V., Bitan, G., Teplow, D.B., Stanley, H.E.: In silico study of amyloid beta-protein folding and oligomerization. Proc. Natl. Acad. Sci. USA **101**, 17345–17350 (2004).
10. Sørensen, J., Periole, X., Skeby, K.K., Marrink, S.J., Schiøtt, B.: Protofibrillar assembly toward the formation of amyloid fibrils. J. Chem. Phys. Lett. **2**, 2385–2390 (2011).
11. Jang, H., Hall, C.K., Zhou, Y.: Assembly and kinetic folding pathways of a tetrameric betasheet complex: molecular dynamics simulations on simplified off-lattice protein models. Biophys. J. **86**(1 Pt 1), 31–49 (2004).
12. Dima, R.I., Thirumalai, D.: Exploring protein aggregation and self-propagation using lattice models: phase diagram and kinetics. Protein Sci. **11**(5), 1036–1049 (2002).
13. Malolepsza, E., Boniecki, M., Kolinski, A., Piela, L.: Theoretical model of prion propagation: a misfolded protein induces misfolding. Proc. Natl. Acad. Sci. USA **102**, 7835–7840 (2005).
14. Khare, S.D., Ding, F., Gwanmesia, K.N., Dokholyan, N.V.: Molecular origin of polyglutamine aggregation in neurodegenerative diseases. PLoS Comput. Biol. **1**, 230–235 (2005).
15. Chen, Y., Dokholyan, N.V.: A single disulfide bond differentiates aggregation pathways of beta2-microglobulin. J. Mol. Biol. **354**, 473–482 (2005).
16. Dokholyan, N.V., Buldyrev, S.V., Stanley, H.E., Shakhnovich, E.I.: Discrete molecular dynamics studies of the folding of a protein-like model. Fold. Des. **3**, 577–587 (1998).
17. Ding, F., Buldyrev, S.V., Dokholyan, N.V.: Folding Trp-cage to NMR resolution native structure using a coarse-grained protein model. Biophys. J. **88**, 147–155 (2005).
18. Ding, F., Dokholyan, N.V., Buldyrev, S.V., Stanley, H.E., Shakhnovic, E.I.: Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. J. Mol. Biol. **324**, 851–857 (2002).
19. Ding, F., Borreguero, J.M., Buldyrey, S.V., Stanley, H.E., Dokholyan, N.V.: Mechanism for the alpha-helix to beta-hairpin transition. Proteins **53**, 220–228 (2003).
20. Ding, F., LaRocque, J.J., Dokholyan, N.V.: Direct observation of protein folding, aggregation, and a prion-like conformational conversion. J. Biol. Chem. **48**, 40235–40240 (2005).
21. Gosal, W.S., Morten, I.J., Hewitt, E.W., Smith, D.A., Thomson, N.H., Radford, S.E.: Competing pathways determine fibril morphology in the self-assembly of beta2-microglobulin into amyloid. J. Mol. Biol. **351**, 850–864 (2005).
22. Plakoutsi, G., Bemporad, F., Calamai, M., Taddei, N., Dobson, C.M., Chiti, F.: Evidence for a mechanism of amyloid formation involving molecular reorganisation within native-like precursor aggregates. J. Mol. Biol. **351**, 910–922 (2005).
23. Vitalis, A., Wang, X., Pappu, R.V.: Quantitative characterization of intrinsic disorder in polyglutamine: insights from analysis based on polymer theories. Biophys. J. **93**, 1923–1937 (2007).
24. Vitalis, A., Lyle, N., Pappu, R.V.: Thermodynamics of beta-sheet formation in polyglutamine. Biophys. J. **97**, 303–311 (2009).
25. Vitalis, A., Caflisch, A.: Micelle-like architecture of the monomer ensemble of Alzheimer's amyloid- peptide in aqueous solution and its implications for A_ aggregation. J. Mol. Biol. **403**, 148–165 (2010).
26. Gsponer, J., Haberthr, U., Caflisch, A.: The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. Proc. Natl. Acad. Sci. USA **100**, 5154–5159 (2003).
27. Hwang, W., Zhang, S., Kamm, R.D., Karplus, M.: Kinetic control of dimer structure formation in amyloid fibrillogenesis. Proc. Natl. Acad. Sci. USA **101**, 12916–12921 (2004).
28. de la Paz, M.L., de Mori, G.M.S., Serrano, L., Colombo, G.: Sequence dependence of amyloid fibril formation: insights from molecular dynamics simulations. J. Mol. Biol. **349**, 583–596 (2005).

29. Cecchini, M., Curcio, R., Pappalardo, M., Melki, R., Caflisch, A.: A molecular dynamics approach to the structural characterization of amyloid aggregation. J. Mol. Biol. **357**, 1306–1321 (2006).
30. Strodel, B., Whittleston, C.S., Wales, D.J.: Thermodynamics and kinetics of aggregation for the GNNQQNY peptide. J. Am. Chem. Soc. **129**, 16005–16014 (2007).
31. De Simone, A., Esposito, L., Pedone, C., Vitagliano, L.: Insights into stability and toxicity of amyloid-like oligomers by replica exchange molecular dynamics analyses. Biophys. J. **95**, 1965–1973 (2008).
32. Bellesia, G., Shea, J.E.: What determines the structure and stability of KFFE monomers, dimers, and protofibrils? Biophys. J. **96**, 875–886 (2009).
33. Ma, B., Nussinov, R.: Stabilities and conformations of Alzheimer's beta-amyloid peptide oligomers (Abeta 16–22, Abeta 16–35, and Abeta 10–35): sequence effects. Proc. Natl. Acad. Sci. USA **99**, 14126–14131 (2002).
34. Buchete, N.V., Tycko, R., Hummer, G.: Molecular dynamics simulations of Alzheimer's beta-amyloid protofilaments. J. Mol. Biol. **353**, 804–821 (2005).
35. Wu, C., Bowers, M.T., Shea, J.E.: Molecular structures of quiescently grown and brain-derived polymorphic fibrils of the Alzheimer amyloid abeta9–40 peptide: a comparison to agitated fibrils. PLoS Comput. Biol. **6**, e1000693 (2010).
36. Wu, C., Shea, J.E.: Coarse-grained models for protein aggregation. Curr. Opin. Struct. Biol. **21**, 209–220 (2011).
37. Pellarin, R., Caflisch, A.: Interpreting the aggregation kinetics of amyloid peptides. J. Mol. Biol. **360**, 882–892 (2006).
38. Müller, M., Katsov, K., Schick, M.: Biological and synthetic membranes: what can be learned from a coarse-grained description? Phys. Rep. **434**, 113–176 (2006).
39. Zhang, J., Muthukumar, M.: Simulations of nucleation and elongation of amyloid fibrils. J. Chem. Phys. **130**, 035102 (2009).
40. Auer, S., Dobson, C.M., Vendruscolo, M., Maritan, A.: Self-templated nucleation in peptide and protein aggregation. PLos Comput. Biol. **4**, e1000222 (2008).
41. Li, M.S., Klimov, D.K., Straub, J.E., Thirumalai, D.: Probing the mechanisms of fibril formation using lattice models. J. Chem. Phys. **129**, 175101 (2008).
42. Nguyen, H.D., Hall, C.K.: Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. Proc. Natl. Acad. Sci. USA **101**, 16180–16185 (2004).
43. Bellesia, G., Shea, J.E.: Self-assembly of beta-sheet forming peptides into chiral fibrillar aggregates. J. Chem. Phys. **126**, 245104 (2007).
44. MacKerell, A.D.J., Feig, M., Brooks, C.L.: Improved treatment of the protein backbone in empirical force fields. J. Am. Chem. Soc. **126**, 698–699 (2004).
45. Zhou, Y., Karplus, M.: Interpreting the folding kinetics of helical proteins. Nature **401**, 400–403 (1999).
46. Brooks, B.R., Brooks, C.L., Mackerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S. et al.: CHARMM: the biomolecular simulation program J. Comput. Chem. **30**, 1545–1614 (2009).
47. Fändrich, M.: Absolute correlation between lag time and growth rate in the spontaneous formation of several amyloid-like aggregates and fibrils. J. Mol. Biol. **365**, 1266–1270 (2007).
48. Hortschansky, P., Schroeckh, V., Christopeit, T., Zandomeneghi, G., Fändrich, M.: The aggregation kinetics of Alzheimer's beta-amyloid peptide is controlled by stochastic nucleation. Protein Sci. **14**, 1753–1759 (2005).
49. Christopeit, T., Hortschansky, P., Schroeckh, V., Guhrs, K., Zandomeneghi, G., Fändrich, M.: Mutagenic analysis of the nucleation propensity of oxidized Alzheimer's beta-amyloid peptide. Protein Sci. **14**, 2125–2131 (2005).
50. Pellarin, R., Guarnera, E., Caflisch, A.: Pathways and intermediates of amyloid fibril formation. J. Mol. Biol. **374**, 917–924 (2007).
51. Nilsberth, C., Westlind-Danielsson, A., Eckman, C.B., Condron, M.M., Axelman, K., Forsell, C. et al.: The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced Abeta protofibril formation. Nat. Neurosci. **4**, 887–893 (2001).

52. Conway, K.A., Lee, S.J., Rochet, J.C., Ding, T.T., Williamson, R.E., Lansbury, P.T.: Acceleration of oligomerization, not fibrillization, is a shared property of both alpha-synuclein mutations linked to early-onset Parkinson's disease: implications for pathogenesis and therapy. Proc. Natl. Acad. Sci. USA **97**, 571–576 (2000).
53. Sabate, R., Estelrich, J.: Evidence of the existence of micelles in the fibrillogenesis of betaamyloid peptide. J. Phys. Chem. B **109**, 11027–11032 (2005).
54. Lomakin, A., Chung, D.S., Benedek, G.B., Kirschner, D.A., Teplow, D.B.: On the nucleation and growth of amyloid beta-protein fibrils: detection of nuclei and quantitation of rate constants. Proc. Natl. Acad. Sci. USA **93**, 1125–1129 (1996).
55. Serio, T.R., Cashikar, A.G., Kowal, A.S., Sawicki, G.J., Moslehi, J.J., Serpell, L. et al.: Nucleated conformational conversion and the replication of conformational information by a prion determinant. Science **289**, 1317–1321 (2000).
56. Fowler, D.M., Koulov, A.V., Alory-Jost, C., Marks, M.S., Balch, W.E., Kelly, J.W.: Functional amyloid formation within mammalian tissue. PLoS Biol. **4**, e6 (2006).
57. Lomakin, A., Teplow, D.B., Kirschner, D.A., Benedek, G., Kinetic theory of fibrillogenesis of amyloid beta-protein. Proc. Natl. Acad. Sci. USA **94**, 7942–7947 (1997).
58. Nielsen, L., Khurana, R., Coats, A., Frokjaer, S., Brange, J., Vyas, S., Uversky, V.N., Fink, A.L.: Effect of environmental factors on the kinetics of insulin fibril formation: elucidation of the molecular mechanism. Biochemistry **40**, 6036–6046 (2001).
59. Wasmer, C., Soragni, A., Sabate, R., Lange, A., Riek, R., Meier, B.H.: Infectious and noninfectious amyloids of the HET-s(218–289) prion have different NMR spectra. Angew. Chem. Int. Ed. **47**, 5839–5841 (2008).
60. Dzwolak, W., Grudzielanek, S., Smirnovas, V., Ravindra, R., Nicolini, C., Jansen, R., Loksztejn, A., Porowski, S., Winter, R.: Ethanol-perturbed amyloidogenic self-assembly of insulin: looking for origins of amyloid strains. Biochemistry **44**, 8948–8958 (2005).
61. Petkova, A.T., Leapman, R.D., Guo, Z., Yau, W.M., Mattson, M.P., Tycko, R.: Self-propagating, molecular-level polymorphism in Alzheimer's beta-amyloid fibrils. Science **307**, 262–265 (2005).
62. Paravastu, A.K., Petkova, A.T., Tycko, R.: Polymorphic fibril formation by residues 10–40 of the Alzheimer's beta-amyloid peptide. Biophys. J. **90**, 4618–4629 (2006).
63. Meinhardt, J., Sachse, C., Hortschansky, P., Grigorieff, N., Fändrich, M.: Abeta(1–40) fibril polymorphism implies diverse interaction patterns in amyloid fibrils. J. Mol. Biol. **386**, 869–877 (2009).
64. Pellarin, R., Schuetz, P., Guarnera, E., Caflisch, A.: Amyloid fibril polymorphism is under kinetic control. J. Am. Chem. Soc. **132**, 14960–14970 (2010).
65. Goldsbury, C., Frey, P., Olivieri, V., Aebi, U., Müller, S.A.: Multiple assembly pathways underlie amyloid-beta fibril polymorphisms. J. Mol. Biol. **352**, 282–298 (2005).
66. Ellis, J.R.: Macromolecular crowding: obvious but underappreciated. Trends Biochem. Sci. **26**, 597–604 (2001).
67. Lopes, D., Meister, A., Gohlke, A., Hauser, A., Blume, A., Winter, R.: Mechanism of islet amyloid polypeptide fibrillation at lipid interfaces studied by infrared reflection absorption spectroscopy. Biophys. J. **93**, 3132–3141 (2007).
68. Chi, E., Ege, C., Winans, A., Majewski, J., Wu, G., Kjaer, K., Lee, K.: Lipid membrane templates the ordering and induces the fibrillogenesis of Alzheimer's disease amyloid-beta peptide. Proteins **72**, 1–24 (2008).
69. Engel, M.F.M., Khemtémourian, L., Kleijer, C., Meeldijk, H., Jacobs, J., Verkleij, A. et al.: Membrane damage by human islet amyloid polypeptide through fibril growth at the membrane. Proc. Natl. Acad. Sci. USA **105**, 6033–6038 (2008).
70. Ellis, R.J., Minton, A.P.: Protein aggregation in crowded environments. Biol. Chem. **387**, 485–497 (2006).
71. Munishkina, L.A., Cooper, E.M., Uversky, V.N., Fink, A.L.: The effect of macromolecular crowding on protein aggregation and amyloid fibril formation. J. Mol. Recognit. **17**, 456–464 (2004).

72. Munishkina, L.A., Ahmad, A., Fink, A.L., Uversky, V.N.: Guiding protein aggregation with macromolecular crowding. Biochemistry **47**, 8993–9006 (2008).
73. Friedman, R., Pellarin, R., Caflisch, A.: Amyloid aggregation on lipid bilayers and its impact on membrane permeability. J. Mol. Biol. **387**, 407–415 (2009).
74. Friedman, R., Pellarin, R., Caflisch, A.: Soluble protofibrils as metastable intermediates in simulations of amyloid fibril degradation induced by lipid vesicles. J. Phys. Chem. Lett. **1**, 471–474 (2010).
75. Volles, M.J., Lee, S.J., Rochet, J.C., Shtilerman, M.D., Ding, T.T., Kessler, J.C., Lansbury, P.T.: Vesicle permeabilization by protofibrillar alpha-synuclein: implications for the pathogenesis and treatment of Parkinson's disease. Biochemistry **40**, 7812–7819 (2001).
76. Sharp, J., Forrest, J., Jones, R.: Surface denaturation and amyloid fibril formation of insulin at model lipid-water interfaces. Biochemistry **41**, 15810–15819 (2002).
77. Khemtémourian, L., Engel, M.F.M., Liskamp, R.M.J., Höppener, J.W.M., Killian, J.A.: The N-terminal fragment of human islet amyloid polypeptide is non-fibrillogenic in the presence of membranes and does not cause leakage of bilayers of physiologically relevant lipid composition. Biochim. Biophys. Acta. **1798**, 1805–1811 (2010).
78. Lashuel, H., Lansbury, P.: Are amyloid diseases caused by protein aggregates that mimic bacterial pore-forming toxins? Q. Rev. Biophys. **39**, 167–201 (2006).
79. Martins, I.C., Kuperstein, I., Wilkinson, H., Maes, E., Vanbrabant, M., Jonckheere, W., Gelder, P.V., Hartmann, D., D'Hooge, R., Strooper, B.D. et al.: Lipids revert inert Abeta amyloid fibrils to neurotoxic protofibrils that affect learning in mice. EMBO J **27**, 224–233 (2008).
80. Friedman, R., Caflisch, A.: Surfactant effects on amyloid aggregation kinetics. J. Mol. Biol. **414**, 303–312 (2011).
81. Magno, A., Caflisch, A., Pellarin, R.: Crowding effects on amyloid aggregation kinetics. J. Phys. Chem. Lett. **1**, 3027–3032 (2010).

The Structure of Intrinsically Disordered Peptides Implicated in Amyloid Diseases: Insights from Fully Atomistic Simulations

Chun Wu and Joan-Emma Shea

1 Introduction

Protein aggregation involves the self-assembly of proteins into large β-sheet-rich complexes. This process can be the result of aberrant protein folding and lead to "amyloidosis," a condition characterized by deposits of protein aggregates known as amyloids on various organs of the body [1]. Amyloid-related diseases include, among others, Alzheimer's disease, Parkinson's disease, Creutzfeldt–Jakob disease, and type II diabetes [2–4]. In other instances, however, protein aggregation is not a pathological process, but rather a functional one, with aggregates serving as structural scaffolds in a number of organisms [5].

It is now well-established that the primary end-product of aggregation has a fibril structure, with a cross-β-sheet pattern based on solid state nuclear magnetic resonance (NMR), X-ray diffraction, electron microscope, and dye-binding studies [6–9]. Despite the importance of the aggregation process from a biomedical perspective, several critical questions remain unanswered regarding the nature of the species populated during the fibrillization process. The very starting point of aggregation—the nature of the "misfolded" monomeric species—is unknown, particularly in the case of large class of "intrinsically disordered" or "natively unfolded" proteins that are prone to aggregation [10, 11]. Rather than populating a well-defined three-dimensional stable globular fold, these proteins interconvert among a number of species. As a result, they are very difficult to characterize using traditional ensemble-averaging methods such as NMR and circular dichroism (CD), although recent

C. Wu
Department of Chemistry and Biochemistry, University of California, Santa Barbara, CA 93106, USA
e-mail: cwu@chem.ucsb.edu

J.-E. Shea (✉)
Department of Physics, University of California, Santa Barbara, CA 93106, USA
e-mail: shea@chem.ucsb.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_9,

advances in NMR techniques are making some headway in this direction [12]. Examples of natively unfolded aggregating proteins include the Alzheimer amyloid-β (Aβ) protein implicated in Alzheimer's disease, the α-synuclein protein implicated in Parkinson's disease, and the islet amyloid polypeptide (IAPP) implicated in type II diabetes [13]. Numerous nonfibrillar aggregates (soluble oligomers, micellar species, and amorphous aggregates) that can be on- or off-pathway to fibril formation, and some of these aggregates that may possess toxic properties, have been identified but not thoroughly structurally characterized [14–19]. Small soluble oligomers are difficult to study experimentally, as they correspond to transient, unstable species. Most experimental techniques do not possess the temporal and spatial resolution to yield atomistically detailed information about oligomeric species.

This chapter focuses on the use of fully atomistic simulations to probe the very initial stage of aggregation of intrinsically disordered proteins: the monomeric state. Simulations are uniquely poised to probe the structure of natively unfolded proteins, as they tract individual protein conformations. We focus primarily on two natively disordered peptides (the Aβ peptide [20] and the IAPP peptide [21, 22]) and review recent simulations on these proteins. Although amyloidogenic peptides (e.g., Aβ and IAPP) are defined as "natively disordered" by ensemble-averaging experimental techniques such as CD and NMR, all-atom simulations actually reveal that these on-average "natively unfolded" peptides in fact do have some partial structure. In particular, the simulations that will be presented show that these peptides either populate a small number of β-rich conformations that could serve as direct precursors for the formation of amyloid fibrils or contain some structured elements such as β-hairpin, short helix-coil-helix, salt bridges, and hydrophobic cluster that may serve as nucleus for folding and oligomerization.

2 Simulation Approaches

The primary simulation technique to study the monomeric conformations of such peptides is replica exchange molecular dynamics (REMD) simulation. Conventional Monte Carlo (CMC) and molecular dynamics (CMD) sampling techniques performed under constant temperature condition are prone to getting trapped in local minima and are not suitable methods for a thorough exploration of conformational space. The time required to overcome energy barriers grows exponentially with the barrier height. At physiological temperatures, escape times can easily reach scales that are inaccessible on current computers (seconds or larger). An incomplete sampling of conformational space distorts the statistical picture of conformational ensembles populated under a given set of conditions and can lead to incorrect conclusions regarding both folding mechanisms and conformational preferences of the peptides. A number of enhanced sampling schemes have been recently developed to remedy this sampling problem and facilitate an escape from the local energy minima. One of the most promising methods is the replica exchange

algorithm, initially introduced in the context of spin glasses and later adapted with significant success to the study of peptides and small proteins [23]. Details of the replica exchange formulation for molecular dynamics have been worked out by Sugita and Okamoto [24–27]. In this scheme, a number of identical copies, or "replicas," of the original system are simulated in parallel for a given number of MD steps at different temperatures. Two replicas i and j adjacent in temperatures T_i and T_j, with energies E_i and E_j, are swapped periodically with probability derived from Boltzmann's statistics:

$$p_{ij} = \begin{cases} 1 & \text{for} \quad \Delta \leq 0 \\ \exp(-\Delta) & \text{for} \quad \Delta > 0 \end{cases} \text{ where } \Delta \equiv [(\beta_i - \beta_j)(E_j - E_i)] \text{ and } \beta = \frac{1}{kT}.$$

Since the escape time from local energy minima decreases significantly at elevated temperatures, the replica exchange method enables enhanced sampling by treating the temperature as a dynamical variable. In addition to extensive sampling, the algorithm also ensures that the sampled conformations at a given temperature belong to the canonical statistical ensemble. Data from simulation are clustered [typically by mutual root mean square deviation (RMSD)] for further analysis [28].

3 The Alzheimer Amyloid-β Peptide

The Amyloid plaques found in the brains of Alzheimer's disease (AD) patients consist primarily of fibrillar aggregates of the Alzheimer amyloid-β protein (also known as the Aβ protein or Aβ peptide), a proteolytic by-product of the amyloid precursor protein (APP) [2, 4]. The Aβ protein is produced predominantly in two forms: the 40-residue long Aβ40 protein and the 42-residue long Aβ42 protein (Fig. 1). In addition to appearing as a sporadic disease (associated with the aggregation of the Aβ40 and Aβ42 peptides), AD can also occur as an inherited disease (Familial AD). Familial forms of AD are due to single point mutations in the Aβ protein. The locations of the most common familial mutations are shown in Fig. 1. In addition, Aβ peptides can adopt different conformations in different solvents.

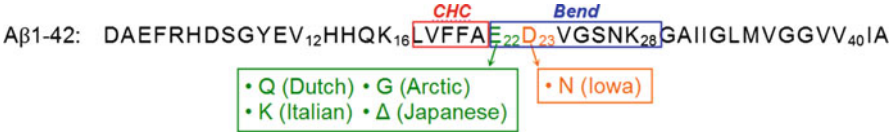

Fig. 1 Sequences for Aβ 1–42. The central hydrophobic core (CHC) spanning residues 17–21 is shown in the *red box* and the bend region (residues 22–28) in the *blue box*. Familial mutations involving residues E22 and D23 are shown in *green* and *orange*, respectively

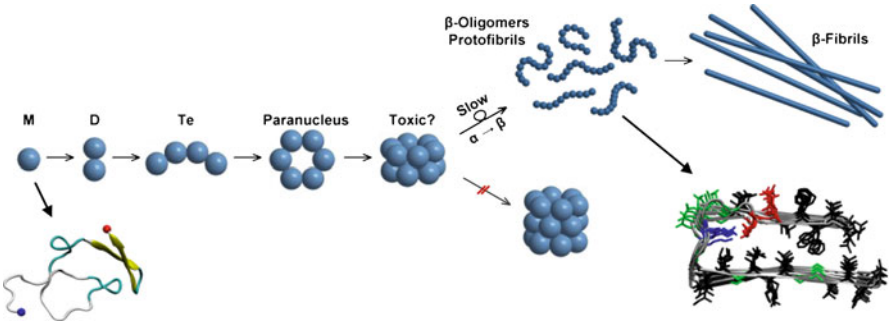

Fig. 2 Aggregation pathways of Aβ42, adapted from [14]. The monomer structure is adapted from [34] and the protofibril structure from [33]

These small chemical–environmental modifications can cause subtle conformation changes in the peptide, which can translate into dramatic differences in the aggregation and toxicity of the Aβ peptides. All-atom simulations [both conventional constant temperature and replica exchange molecular dynamics (CMD/REMD) simulations] are powerful tools to probe these subtle structural effects [29, 30]. For example, simulations were used to examine subtle conformational changes of the Aβ peptides due to the presence of two additional hydrophobic residues (Aβ40 vs. Aβ42), familiar mutations, and solvent effects. We describe the main results of these simulations below.

3.1 Aβ40 and Aβ42

Aβ40 and Aβ42 differ only by the presence of two additional hydrophobic residues (I41 and A42) at the C-terminus of Aβ42. While both species are amyloidogenic and neurotoxic, Aβ42 aggregates faster and is significantly more toxic than Aβ40 [19]. Furthermore, Aβ40 and Aβ42 are known to oligomerize via distinct pathways. While Aβ40 populates stable dimers, trimers, and tetramers prior to fibril formation, Aβ42 oligomerizes further to form hexamers and higher order assemblies [14,31,32] prior to fibril formation [33] (Fig. 2).

Using both CMD and REMD simulations, Garcia and coworkers [34, 35] have found the major conformation difference between the two alloforms is that the C-terminus of Aβ42 is more structured than that of Aβ40, due to the formation of a short β-hairpin in the C-terminal sequence 31IIGLMVGGVVIA42 involving two short strands at residues 31–34 and 38–41, respectively. These structural features are in a quantitative agreement with available NMR data [34] and IR data [36]. Using REMD with explicit solvent, we have studied a number of C-terminal fragments (CTFs) (Aβ($x-42$), $x = 29$–$31, 39$) [37]. Our simulations have revealed that the CTFs adopt a metastable β-structure: a β-hairpin for Aβ($x - 42$),

$x = 29–31$ and an extended β-strand for Aβ(39–42). Furthermore, the β-hairpin of Aβ(30–42) converted into a turn-coil conformation when the last two hydrophobic residues were removed, suggesting that the I41 and A42 residues are critical for stabilizing the β-hairpin in the Aβ42-derived CTFs. Given the critical role of the C-terminus in the self-assembly of full-length Aβ42 peptide, the CTFs were investigated as inhibitors that could potentially disrupt the self-assembly and reduce the Aβ42-induced neural toxicity. Indeed, all of the CTFs did inhibit Aβ-induced neurotoxicity to at least some extent. Interestingly, the smallest CTF, Aβ(39–42), was particularly effective in inhibiting Aβ42-induced cell death and rescuing Aβ42-induced disruption of synaptic activity [38], making it a lead inhibitor for further optimization.

3.2 Familial Forms of AD

Early onset AD refers to cases of the disease diagnosed before the age of 65. This form of AD is present in 5–10% of all AD patients. Approximately, half the cases of early onset AD correspond to familial AD, in which a genetic mutation leads to the early onset of the disease. The majority of familial mutants involve residues E22 and D23 of the Aβ peptide [39–42]. Limited proteolysis showed that the 21–30 region of the Aβ peptide is resistant to proteolysis, and subsequent NMR studies [43, 44] showed that this region adopts a bend structure in isolation. Simulations of the 21–30 fragment using REMD and CMD confirmed that this peptide is structured [45–51]. Importantly, simulations on longer fragments (including the Aβ10–35 peptide) reveal that the 21–30 fragment is structured not only in isolation, but also within the context of the longer sequences [29,30,52–55]. These simulations suggest that the bend may be responsible for nucleating the folding of the Aβ peptide. Experiments by Meredith and coworkers [56] showed that linking residues D23 and K28 by a lactam bridge lead to aggregation rates for Aβ40-lactam(D23/K28) that were 1,000-fold greater than for the wild type Aβ40 peptide. Simulations of Aβ40-lactam(D23/K28) revealed that this peptide populates, to a much greater extent that the wild type Aβ40 peptide, "aggregation-competent" conformations, with a bend spanning D23–G29 and β-strands in the N- and C-terminal regions [51]. A number of computational research groups have investigated the effect of familial mutants located in this bend region (the E22Q Dutch, the E22K Italian, the E22G Arctic, and the D23N Iowa mutants) on the folding of the Aβ21–30 fragment using REMD and CMD [48, 50, 57]. These simulations explored the interplay between hydrophobic and electrostatic interactions in this segment, and revealed that the D23 mutant significantly disrupts the bend region, intimating that the D23 mutant alters the folding nucleation of the Aβ peptide. The E22 mutants, on the other hand, do not affect the bend region, implying that the effect of these mutants is on regions outside of the 21–30 segment. To explore this idea, we extended our simulation to the Aβ15–28 peptide [58]. This peptide encompasses the bend region (residues 22–28) and the central hydrophobic core (residues 17–21), a region critical for aggregation.

Indeed, our simulations showed that although the E22Q mutant did not affect the bend structure in the Aβ E22Q mutant, it weakened the interactions between the CHC and the bend, leading to an increased population of β-structure in the CHC. Our free energy analysis further indicated that the E22Q mutation increases Aβ aggregation rates by lowering the barrier for Aβ monomer deposition onto a fibril.

3.3 Effect of Solution Conditions on Aβ Structure

Whereas on average Aβ42 peptide adopts disordered coil-turn rich structure in aqueous solution, it adopts an ordered helix-rich conformation in apolar solvents [59, 60]. In mixed solvent of hexafluoroisopropanol (HFIP) and water with 80:20 ratio in volume, NMR studies [59] showed it adopts a helix-turn-helix structure, with two helices nearly at a right angle (helix 1: residues 7–27; helix 2: residues 28–39). To understand the solvent effect, simulations with explicit solvent models were used to study the conformational preference of Aβ(1–42) in different solvents [61–63]. The simulations revealed new insights into the conformations populated by the Aβ peptide in nonaqueous solvents, and showed that the apolar solvents HFIP (hexafluoroisopropanol) and TFE (2,2,2-trifluoro-ethanol) promote helix formation, that the polar solvent DMSO (dimethyl sulfoxide) causes the unfolding of the C-terminal part, and that the polar solvent water induces the α-helix to β-sheet transition for the C-terminal part. These simulations may explain why Aβ42 aggregation generally occurs out of apolar membrane and in extracellular water environment, where Aβ42 is converted from nonaggregating helical conformation to aggregation prone coil and sheet-rich conformations.

4 The IAPP Peptide

Type II diabetes is an age and life-style related disease involving insulin resistance and loss of β-cell mass. A hallmark of this disease is the presence of amyloid fibrils of the Islet Amyloid IAPP peptide (also known as amylin) in the β-cells of the pancreas [22]. The IAPP is a 37-residue peptidic by-product of a larger precursor protein, and it is co-secreted along with insulin by the pancreatic islet β-cells. Under pathological conditions, IAPP is over-expressed and aggregates. As in the case of the Aβ peptide implicated in AD, both early oligomers and mature fibrils of IAPP appear to be toxic. The IAPP peptide is present in many mammalian species, including rodents. Although the rat and human IAPP forms (Fig. 3) differ only by six amino acids, only the human form aggregates. Transgenic rodents (rodents with a human form of IAPP) on a high fat diet can develop type II diabetes with the accompanying aggregation of the IAPP peptide [64].

The structures of the human and the rat forms of IAPP monomers were studied independently using REMD by Wu et al. [65] using an implicit solvent (igb5) is the

Fig. 3 Primary sequences of human and rat IAPP. The sequence differences between rat IAPP and human IAPP are *underlined*. The segment in which these mutational differences occur is shown in the *red box*

Amber force field ff96 and by Reddy et al. [66, 67] using an explicit solvent and the Gromos96 53a6 force field. The aim of these simulations was twofold: (1) to obtain structural information about the monomeric states of these peptides that is not available from experiments and (2) to explore whether conformational differences were present in the monomeric form that could explain why the human variant of IAPP aggregated and the rodent form did not. The results of the simulations by the two groups are in remarkable agreement. Representative aggregation prone and nonaggregation prone structures of the human from simulation are shown in Fig. 4. The nonaggregation prone conformations of human IAPP are very similar to those of rat IAPP.

The rat form populates predominantly helix-coil structures, with helicity present in the N-terminal region. The simulation results are in good agreement with NMR studies [68,69] that point to helicity in this region, and with result from the AGADIR program [70] that suggest that the N-terminal region has high helical propensity. The human form, on the other hand, coexists between helix-coil structures, helix-sheet structures, and β-hairpin structures. The human and rat forms share the same N-terminal region, and NMR studies on the human form show that, as in the rat case, there is helical propensity in this region [69, 71]. The presence of β-rich elements is supported by 2D-IR experiments [72] as well as by ion-mobility mass spectrometry (IMMS) studies from the Bowers group [65]. IMMS generates collision-cross sections, a measure of the overall size of the ions under study. The IMMS spectra for the human form differed from the rat form by the presence of an extended feature. Calculation of theoretical cross sections showed that the compact structures seen in experiment were consistent with the helix-coil structures seen in simulation, while the extended feature was consistent with a β-hairpin structure. The simulations suggest that the β-rich hairpin conformation may be a direct precursor to aggregation [65] (Fig. 4). The idea that human IAPP possesses both aggregation prone and nonaggregation prone conformations is supported by earlier CD experiments by the Kayed group [73].

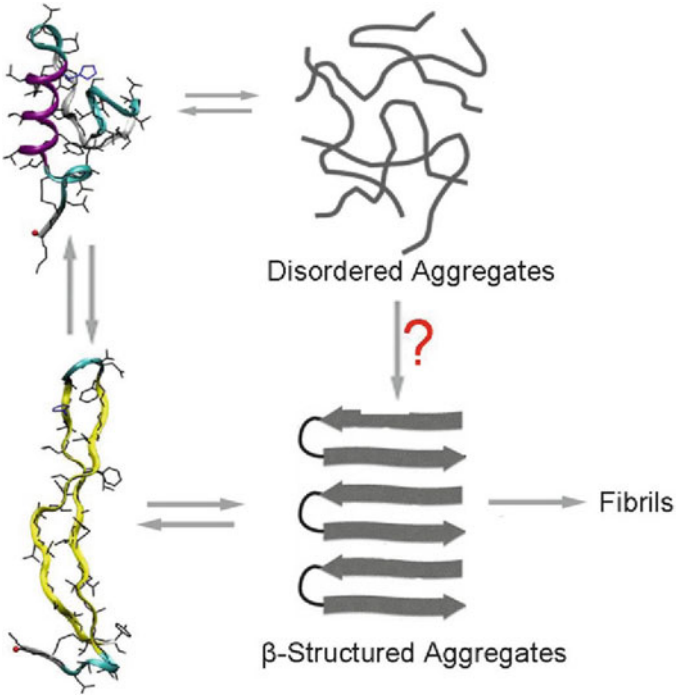

Fig. 4 Schematic representation of a possible oligomerization mechanism, adapted from [65]. Among two interconverting structural families of human IAPP, β-hairpins are proposed to self-assemble into early ordered human IAPP oligomers by side-to-side association. The (?) symbol notes that we have no data pro or con the occurrence of this mechanism and we hence include it for completeness

5 Conclusions

Fully atomic simulations, coupled with enhanced sampling techniques, are emerging as powerful tools for the study of intrinsically disordered peptides. While most experimental techniques only provide ensemble averaged information, simulations are capable of providing detailed atomistic information about the structures populated by these peptides. A highlight of the simulations described in the preceding paragraph is the identification of β-rich conformers, coexisting with non-β-rich structures that may play a role in initiating aggregation. With ever increasing computational power and sophisticated sampling schemes, simulations are now poised to move beyond the investigation of aggregating IDP peptides in a bulk-like environment to the study of the interaction of these peptides with membranes [74–80]. A long-term goal of these simulations is to prove detailed structural information that can guide the rational design of drugs as therapeutics for amyloid diseases.

Acknowledgments This work was funded by the David and Lucile Packard Foundation, the NSF (MCB 0642086), and the NIH (AG027818). The computing time was provided by the Lonestar and Ranger clusters in the Texas Advanced Computing Center (LRAC MCA 05S027).

References

1. Chiti, F., Dobson, C.M.: Protein misfolding, functional amyloid, and human disease. Annu. Rev. Biochem. **75**, 333–366 (2006)
2. Selkoe, D.J.: Cell biology of protein misfolding: the examples of Alzheimer's and Parkinson's diseases. Nat. Cell. Biol. **6**, 1054–1061 (2004)
3. Dobson, C.M.: Protein folding and misfolding. Nature **426**(6968), 884–890 (2003)
4. Harper, J.D., Lansbury, P.T.: Models of amyloid seeding in Alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins. Annu. Rev. Biochem. **66**, 385–407 (1997)
5. Fowler, D.M., Koulov, A.V., Balch, W.E., Kelly, J.W.: Functional amyloid – from bacteria to humans. Trends Biochem. Sci. **32**(5), 217–224 (2007)
6. Tycko, R.: Molecular structure of amyloid fibrils: insights from solid-state NMR. Q. Rev. Biophys. **39**(1), 1–55 (2006)
7. Green, J., Goldsbury, C., Min, T., Sunderji, S., Frey, P., Kistler, J., Cooper, G., Aebi, U.: Full-length rat amylin forms fibrils following substitution of single residues from human amylin. J. Mol. Biol. **326**(4), 1147–1156 (2003)
8. Furumoto, S., Okamura, N., Iwata, R., Yanai, K., Arai, H., Kudo, Y.: Recent advances in the development of amyloid imaging agents. Curr. Top.Med.Chem. **7**(18), 1773–1789 (2007)
9. Shiraham, T., Cohen A.S.: High-resolution electron microscopic analysis of amyloid fibril. J. Cell Biol. **33**(3), 679–708 (1967)
10. Uversky, V.N.: What does it mean to be natively unfolded? Eur. J. Biochem. **269**(1), 2–12 (2002)
11. Radivojac, P., Iakoucheva, L.M., Oldfield, C.J., Obradovic, Z., Uversky, V.N., Dunker, A.K.: Intrinsic disorder and functional proteomics. Biophys. J. **92**(5), 1439–1456 (2007)
12. Salmon, L., Nodet, G., Ozenne, V., Yin, G.W., Jensen, M.R., Zweckstetter, M., Blackledge, M.: NMR characterization of long-range order in intrinsically disordered proteins. J. Am. Chem. Soc. **132**(24), 8407–8418 (2010)
13. Soto, C.: Unfolding the role of protein misfolding in neurodegenerative diseases. Nat. Rev. Neurosci. **4**(1), 49–60 (2003)
14. Bernstein, S.L., Dupuis, N.F., Lazo, N.D., Wyttenbach, T., Condron, M.M., Bitan, G., Teplow, D.B., Shea, J.E., Ruotolo, B.T., Robinson, C.V., Bowers, M.T.: Amyloid-β protein oligomerization and the importance of tetramers and dodecamers in the aetiology of Alzheimer's disease. Nat. Chem. **1**(4), 326–331 (2009)
15. Haass, C., Selkoe, D.J.: Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid Aβ-peptide. Nat.Rev. Mol. Cell. Biol. **8**(2), 101–112 (2007)
16. Cheon, M., Chang, I., Mohanty, S., Luheshi, L.M., Dobson, C.M., Vendruscolo, M., Favrin, G.: Structural reorganisation and potential toxicity of oligomeric species formed during the assembly of amyloid fibrils. PLoS Comp. Biol. **3**(9), 1727–1738 (2007)
17. Smith, A.M., Jahn, T.R., Ashcroft, A.E., Radford, S.E. Direct observation of oligomeric species formed in the early stages of amyloid fibril formation using electrospray ionisation mass spectrometry. J. Mol. Biol. **364**(1), 9–19 (2006)
18. Kayed, R., Head, E., Thompson, J.L., McIntire, T.M., Milton, S.C., Cotman, C.W., Glabel, C.G.: Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. Science **300**(5618), 486–489 (2003)
19. Dahlgren, K.N., Manelli, A.M., Stine, W.B., Baker, L.K., Krafft, G.A., LaDu, M.J.: Oligomeric and fibrillar species of amyloid-β peptides differentially affect neuronal viability. J. Biol. Chem. **277**(35), 32046–32053 (2002)

20. Roychaudhuri, R., Yang, M., Hoshi, M.M., Teplow, D.B.: Amyloid beta-protein assembly and Alzheimer disease. J. Biol. Chem. **284**(8), 4749–4753 (2009)
21. Westermark, P., Wernstedt, C., Obrien, T.D., Hayden, D.W., Johnson, K.H.: Islet amyloid in type-2 human diabetes-mellitus and adult diabetic cats contains a novel putative polypeptide hormone. Am. J. Pathol. **127**(3), 414–417 (1987)
22. Hoppener, J.W.M., Lips, C.J.M.: Role of islet amyloid in type 2 diabetes mellitus. Int. J. Biochem. Cell Biol. **38**(5–6), 726–736 (2006)
23. Swendsen, R.H., Wang, J.S.: Replica Monte Carlo simulation of spin-glasses. Phys. Rev. Lett. **57**, 2607–2609 (1986)
24. Sugita, Y., Okamoto, Y.: Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. **314**(1–2), 141–151 (1999)
25. Bellesia, G., Lampoudi, S., Shea, J.E.: Computational methods in nanostructure design: replica exchange simulations of self-assembling peptides. In: Gazit, E., Nussinov, R. (eds.) Nanostructure Design: Methods and Protocols, vol. 474, Methods in Molecular Biology, pp. 133–152. Humana Press, Totowa, NJ (2008)
26. Nymeyer, H., Gnanakaran, S., Garcia, A.E.: Atomic simulations of protein folding, using the replica exchange algorithm. Methods enzymol. **383**, 119–149 (2004)
27. Penev, E.S., Lampoudi, S., Shea, J.E.: TiREX: replica-exchange molecular dynamics using TINKER. Comput. Phys. Commun. **180**(10), 2013–2019 (2009)
28. Daura, X., van Gunsteren, W.F., Mark, A.E.: Folding-unfolding thermodynamics of a β-heptapeptide from equilibrium simulations. Proteins: Struct. Funct. Genet. **34**(3), 269–280 (1999)
29. Straub, J.E., Thirumalai, D.: Principles governing oligomer formation in amyloidogenic peptides. Curr. Opin. Struct. Biol. **20**(2), 187–195 (2010)
30. Straub, J.E., Thirumalai, D.: Toward a molecular theory of early and late events in monomer to amyloid fibril formation. Annu. Rev. Phys. Chem. **62**, 437–463 (2011)
31. Bitan, G., Kirkitadze, M.D., Lomakin, A., Vollers, S.S., Benedek, G.B., Teplow, D.B.: Amyloid β protein (Aβ) assembly: Aβ40 and Aβ42 oligomerize through distinct pathways. Proc. Natl. Acad. Sci. U.S.A. **100**(1), 330–335 (2003)
32. Bernstein, S.L., Wyttenbach, T., Baumketner, A., Shea, J.E., Bitan, G., Teplow, D.B., Bowers, M.T.: Amyloid β-protein: monomer structure and early aggregation states of Aβ42 and its Pro19 alloform. J. Am. Chem. Soc. **127**(7), 2075–2084 (2005)
33. Luhrs, T., Ritter, C., Adrian, M., Riek-Loher, D., Bohrmann, B., Doeli, H., Schubert, D., Riek, R.: 3D structure of Alzheimer's amyloid-β(1–42) fibrils. Proc. Natl. Acad. Sci. U. S. A. **102**(48), 17342–17347 (2005)
34. Sgourakis, N.G., Yan, Y.L., McCallum, S.A., Wang, C.Y., Garcia, A.E.: The Alzheimer's peptides Aβ40 and 42 adopt distinct conformations in water: a combined MD/NMR study. J. Mol. Biol. **368**(5), 1448–1457 (2007)
35. Sgourakis, N.G., Merced-Serrano, M., Boutsidis, C., Drineas, P., Du, Z.M., Wang, C.Y., Garcia, A.E.: Atomic-level characterization of the ensemble of the Aβ(1–42) monomer in water using unbiased molecular dynamics simulations and spectral algorithms. J. Mol. Biol. **405**(2), 570–583 (2011)
36. Zhuang, W., Sgourakis, N.G., Li, Z.Y., Garcia, A.E., Mukamel, S.: Discriminating early stage Aβ42 monomer structures using chirality-induced 2DIR spectroscopy in a simulation study. Proc. Natl. Acad. Sci. U. S. A. **107**(36), 15687–15692 (2010)
37. Wu, C., Murray, M.M., Bernstein, S.L., Condron, M.M., Bitan, G., Shea J.E., Bowers, M.T.: The structure of Aβ42 C-terminal fragments probed by a combined experimental and theoretical study. J. Mol. Biol. **387**(2), 492–501 (2009)
38. Fradinger, E.A., Monien, B.H., Urbanc, B., Lomakin, A., Tan, M., Li, H., Spring, S.M., Condron, M.M., Cruz, L., Xie, C.W., Benedek, G.B., Bitan, G.: C-terminal peptides coassemble into Aβ42 oligomers and protect neurons against Aβ42-induced neurotoxicity. Proc. Natl. Acad. Sci. U. S. A. **105**(37), 14175–14180 (2008)
39. Van Nostrand, W.E., Melchor, J.P., Cho, H.S., Greenberg, S.M., Rebeck, G.W.: Pathogenic effects of D23N Iowa mutant amyloid β-protein. J. Biol. Chem. **276**(35), 32860–32866 (2001)

40. Nilsberth, C., Westlind-Danielsson, A., Eckman, C.B., Condron, M.M., Axelman, K., Forsell, C., Stenh, C., Luthman, J., Teplow, D.B., Younkin, S.G., Naeslund, J., Lannfelt, L.: The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced Aβ protofibril formation. Nat. Neurosci. **4**(9), 887–893 (2001)
41. Melchor, J.P., McVoy, L., Van Nostrand, W.E.: Charge alterations of E22 enhance the pathogenic properties of the amyloid β-protein. J. Neurochem. **74**(5), 2209–2212 (2000)
42. Lashuel, H.A.: Mixtures of wild-type and a pathogenic (E22G) form of Aβ40 in vitro accumulate protofibrils, including amyloid pores. J. Mol. Biol. **332**, 795–808 (2003)
43. Lazo, N.D., Grant, M.A., Condron, M.C., Rigby, A.C., Teplow, D.B.: On the nucleation of amyloid β-protein monomer folding. Protein Sci. **14**(6), 1581–1596 (2005)
44. Fawzi, N.L., Phillips, A.H., Ruscio, J.Z., Doucleff, M., Wemmer, D.E., Head-Gordon, T.: Structure and dynamics of the Aβ21–30 peptide from the interplay of NMR experiments and molecular simulations. J. Am. Chem. Soc. **130**, 6145–6158 (2008)
45. Murray, M.M., Krone, M.G., Bernstein, S.L., Baumketner, A., Condron, M.M., Lazo, N.D., Teplow, D.B., Wyttenbach, T., Shea, J.E., Bowers, M.T.: Amyloid β-protein: experiment and theory on the 21–30 fragment. J. Phys. Chem. B **113**(17), 6041–6046 (2009)
46. Baumketner, A., Bernstein, S.L., Wyttenbach, T., Lazo, N.D., Teplow, D.B., Bowers, M.T., Shea, J.E.: Structure of the 21–30 fragment of amyloid β-protein. Protein Sci. **15**(6), 1239–1247 (2006)
47. Borreguero, J.M., Urbanc, B., Lazo, N.D., Buldyrev, S.V., Teplow, D.B., Stanley, H.E.: Folding events in the 21–30 region of amyloid β-protein Aβ studied in silico. Proc. Natl. Acad. Sci. U. S. A. **102**(17), 6015–6020 (2005)
48. Cruz, L., Urbanc, B., Borreguero, J.M., Lazo, N.D., Teplow, D.B., Stanley, H.E.: Solvent and mutation effects on the nucleation of amyloid β-protein folding. Proc. Natl. Acad. Sci. U. S. A. **102**(51), 18258–18263 (2005)
49. Chen, J.W., Romero, P., Uversky, V.N., Dunker, A.K.: Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. J. Proteome Res. **5**(4), 879–887 (2006)
50. Tarus, B., Straub, J.E., Thirumalai, D.: Structures and free-energy landscapes of the wild type and mutants of the Aβ(21–30) peptide are determined by an interplay between intrapeptide electrostatic and hydrophobic interactions. J. Mol. Biol. **379**(4), 815–829 (2008)
51. Reddy, G., Straub, J.E., Thirumalai, D.: Influence of preformed Asp23-Lys28 salt bridge on the conformational fluctuations of monomers and dimers of Aβ peptides with implications for rates of fibril formation. J. Phys. Chem. B **113**(4), 1162–1172 (2009)
52. Tarus, B., Straub, J.E., Thirumalai, D.: Dynamics of Asp23-Lys28 salt-bridge formation in Aβ10–35 monomers. J. Am. Chem. Soc. **128**(50), 16159–16168 (2006)
53. Melquiond, A., Xiao, D., Mousseau, N., Derreumaux, P.: Role of the region 23–28 in Aβ fibril formation: insights from simulations of the monomers and dimers of Alzheimer's peptides Aβ40 and Aβ42. Curr. Alzheimer Res. **5**, 244–250 (2008)
54. Kamiya, N., Mitomo, D., Shea, J.E., Higo, J.: Folding of the 25 residue Aβ(12–36) peptide in TFE/water: temperature-dependent transition from a funneled free-energy landscape to a rugged one. J. Phys. Chem. B **111**(19), 5351–5356 (2007)
55. Baumketner, A., Shea, J.E.: The structure of the Alzheimer amyloid β10–35 peptide probed through replica-exchange molecular dynamics simulations in explicit solvent. J. Mol. Biol. **366**(1), 275–285 (2007)
56. Sciarretta, K.L., Gordon, D.J., Petkova, A.T., Tycko, R., Meredith, S.C.: Aβ40-Lactam(D23/K28) models a conformation highly favorable for nucleation of amyloid. Biochemistry (Mosc) **44**(16), 6003–6014 (2005)
57. Krone, M.G., Baumketner, A., Bernstein, S.L., Wyttenbach, T., Lazo, N.D., Teplow, D.B., Bowers, M.T., Shea, J.E.: Effects of familial Alzheimer's disease mutations on the folding nucleation of the amyloid β-protein. J. Mol. Biol. **381**(1), 221–228 (2008)
58. Baumketner, A., Krone, M.G., Shea, J.E.: Role of the familial Dutch mutation E22Q in the folding and aggregation of the 15–28 fragment of the Alzheimer amyloid-β protein. Proc. Natl. Acad. Sci. U. S. A. **105**(16), 6027–6032 (2008)

59. Crescenzi, O., Tomaselli, S., Guerrini, R., Salvadori, S., D'Ursi, A.M., Temussi, P.A., Picone, D.: Solution structure of the Alzheimer amyloid β-peptide (1–42) in an apolar microenvironment – similarity with a virus fusion domain. Eur. J. Biochem. **269**(22), 5642–5648 (2002)
60. Tomaselli, S., Esposito, V., Vangone, P., van Nuland, N.A.J., Bonvin, A., Guerrini, R., Tancredi, T., Temussi, P.A., Picone, D.: The alpha-to-beta conformational transition of Alzheimer's Aβ(1–42) peptide in aqueous media is reversible: a step by step conformational analysis suggests the location of beta conformation seeding. Chembiochem **7**(2), 257–267 (2006)
61. Yang, C., Zhu, X.L., Li, J.Y., Chen, K.: Molecular dynamics simulation study on conformational behavior of Aβ(1–40) and Aβ(1–42) in water and methanol. J. Mol. Struc-Theochem. **907**(1–3), 51–56 (2009)
62. Yang, C., Li, J.Y., Li, Y., Zhu, X.L.: The effect of solvents on the conformations of Amyloid β-peptide (1–42) studied by molecular dynamics simulation. J. Mol. Struc-Theochem. **895**(1–3), 1–8 (2009)
63. Jalili, S., Akhavan, M.: A molecular dynamics simulation study of conformational changes and solvation of Aβ peptide in trifluoroethanol and water. J. Theor. Comput. Chem. **8**(2), 215–231 (2009)
64. Matveyenko, A.V., Butler, P.C.: Islet amyloid polypeptide (IAPP) transgenic rodents as models for type 2 diabetes. ILAR J. **47**(3), 225–233 (2006)
65. Dupuis, N.F., Wu, C., Shea, J.E., Bowers, M.T.: Human islet amyloid polypeptide monomers form ordered β-hairpins: a possible direct amyloidogenic precursor. J. Am. Chem. Soc. **131**(51), 18283–18292 (2009)
66. Reddy, A.S., Wang, L., Lin, Y.S., Ling, Y., Chopra, M., Zanni, M.T., Skinner, J.L., De Pablo, J.J.: Solution structures of rat amylin peptide: simulation, theory, and experiment. Biophys. J. **98**(3), 443–451 (2010)
67. Reddy, A.S., Wang, L., Singh, S., Ling, Y., Buchanan, L., Zanni, M.T., Skinner, J.L., De Pablo, J.J.: Stable and metastable states of human amylin in solution. Biophys. J. **99**(7), 2208–2216 (2010)
68. Williamson, J.A., Miranker, A.D.: Direct detection of transient β-helical states in islet amyloid polypeptide. Protein Sci. **16**(1), 110–117 (2007)
69. Yonemoto, I.T., Kroon, G.J.A., Dyson, H.J., Balch, W.E., Kelly, J.W.: Amylin proprotein processing generates progressively more amyloidogenic peptides that initially sample the helical state. Biochemistry (Mosc) **47**(37), 9900–9910 (2008)
70. Munoz, V., Serrano, L.: Development of the multiple sequence approximation within the AGADIR model of β-helix formation: comparison with Zimm-Bragg and Lifson-Roig formalisms. Biopolymers **41**(5), 495–509 (1997)
71. Mishra, R., Geyer, M., Winter, R.: NMR spectroscopic investigation of early events in IAPP amyloid fibril formation. ChemBioChem **10**(11), 1769–1772 (2009)
72. Shim, S.H., Gupta, R., Ling, Y.L., Strasfeld, D.B., Raleigh, D.P., Zanni, M.T.: Two-dimensional IR spectroscopy and isotope labeling defines the pathway of amyloid formation with residue-specific resolution. Proc. Natl. Acad. Sci. U. S. A. **106**(16), 6614–6619 (2009)
73. Kayed, R., Bernhagen, J., Greenfield, N., Sweimeh, K., Brunner, H., Voelter, W., Kapurniotu, A.: Conformational transitions of islet amyloid polypeptide (IAPP) in amyloid formation in vitro. J. Mol. Biol. **287**(4), 781–796 (1999)
74. Davis, C.H., Berkowitz, M.L.: Structure of the amyloid-β (1–42) monomer absorbed to model phospholipid bilayers: a molecular dynamics study. J. Phys. Chem. B **113**(43), 14480–14486 (2009)
75. Davis, C.H., Berkowitz, M.L.: A molecular dynamics study of the early stages of amyloid-β(1–42) oligomerization: the role of lipid membranes. Proteins: Struct. Funct. Bioinf. **78**(11), 2533–2545 (2010)
76. Strodel, B., Lee, J.W.L., Whittleston, C.S., Wales, D.J.: Transmembrane structures for Alzheimer's Aβ(1–42) oligomers. J. Am. Chem. Soc. **132**(38), 13300–13312 (2010)
77. Lemkul, J.A., Bevan, D.R.: A comparative molecular dynamics analysis of the amyloid β-peptide in a lipid bilayer. Arch. Biochem. Biophys. **470**(1), 54–63 (2008)

78. Lemkul, J.A., Bevan, D.R.: Perturbation of membranes by the amyloid β-peptide – a molecular dynamics study. FEBS J. **276**(11), 3060–3075 (2009)
79. Miyashita, N., Straub, J.E., Thirumalai, D.: Structures of β-amyloid peptide 1–40, 1–42, and 1–55-the 672–726 fragment of APP-in a membrane environment with implications for interactions with gamma-secretase. J. Am. Chem. Soc. **131**(49), 17843–17852 (2009)
80. Miyashita, N., Straub, J.E., Thirumalai, D., Sugita, Y.: Transmembrane structures of amyloid precursor protein dimer predicted by replica-exchange molecular dynamics simulations. J. Am. Chem. Soc. **131**(10), 3438–3439 (2009)

Part III
Modeling Cells and Cellular Pathways

Computer Simulations of Mechano-Chemical Networks Choreographing Actin Dynamics in Cell Motility

Pavel I. Zhuravlev, Longhua Hu, and Garegin A. Papoian

1 Introduction

Cell migration is critical for the development and functioning of many higher organisms, including humans. Among various cellular processes that may be used to drive eukaryotic cell locomotion, by far the most conspicuous mechanism is based on force generation by actin networks. Actin is a medium sized globular protein, which can polymerize into various filaments, depending on its chemical state, such as whether it is bound to ATP or ADP [1]. Actin filament containing subcellular structures are critically important for the survival and motility of eukaryotic cells. In particular, actin-driven motility is implicated in embryonic and organ development, neuronal cone growth, immune response, wound healing, and cancer metastasis among many other biological processes [2–6]. Hence, elucidating the principles of actin-based cell motility is among the most important challenges of modern cell biology.

Intense research efforts starting from the 1950s have led to the realization that the actin protrusion machinery is highly sophisticated, employing complex mechano-chemical networks, with high degree of redundancy and various nonlinear feedbacks. The large volume of experimental work suggested numerous possible mechanisms for various modules of actin polymerization machinery, however, many key aspects of both biology and physics of actin network dynamics in vivo are either not well understood or controversial. In particular, two principal cellular substructures that ubiquitously appear in driving or regulating motility of many different cell lines are *lamellipodia*, thin sheet-like protrusions at leading edge of the cell containing a three-dimensional actin mesh, and *filopodia*, finger-like

P.I. Zhuravlev • L. Hu • G.A. Papoian (✉)
Department of Chemistry and Biochemistry, Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA
e-mail: zhur@umd.edu; lhhu@umd.edu; gpapoian@umd.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_10,

cellular protrusions comprising a bundle of actin filaments [1]. These filamentous networks are highly dynamic, where actin polymerization processes and mechanical interactions continuously remodel the network structure. Actin polymerization in vivo is regulated spatially and temporally by an intricate web of signaling proteins and mechano-chemical feedback. Mechanical interactions include, among others, actin filament buckling, interactions with the cell membrane and adhesion to the outside environment.

Given the enormous complexity of chemical interactions networks, mechanical and transport processes governing in vivo actin dynamics, modeling based on physical principles can be very useful in making sense of sometimes contradictory experimental results, and perhaps more importantly set theoretical foundations for making physically reasonable interpretations. Below, we review recent progress on modeling filopodia and lamellipodia, focusing mainly on simulations and theory at a single molecule resolution. The latter was historically preceded by macroscopic description of actin protrusion dynamics, which has played an important role in formulating the larger framework for understanding cell motility processes. In general, when macroscopic models work, they often provide elegant conceptual understanding of the problem, however, they also fail from time to time, where a recourse to microscopic physics becomes the only solution. A few such examples are discussed below.

In the following, we first describe general aspects of modeling reaction–diffusion processes at the cellular scale. It turns out that inherent microscopic randomness of chemical reactions, which usually averages out on the macroscopic level, may sometime dominate the behavior of cellular signal transduction networks. We also briefly describe mechanical processes necessary for modeling cell motility dynamics. These general mesoscopic modeling sections are followed by discussions focusing on dynamics of filopodia and lamellipodia, respectively. The emerging understanding from modeling various actin-based protrusions suggests that the overall behavior of actin network growth and remodeling dynamics is determined by a subtle interplay among chemical interactions, transport bottlenecks, and mechanical feedbacks. The specific nature of this interplay is discussed in sections on filopodial and lamellipodial dynamics. Finally, various topics in biology and biophysics of actin networks were thoroughly reviewed in a recent book edited by M.-F. Carlier [7]. The present chapter provides discussion largely complementary to the contents of this noteworthy volume, which we recommend as further reading.

2 Mechano-Chemical Networks Regulating Actin Dynamics

Extension and retraction of filopodia and lamellipodia are based on actin polymerization and depolymerization processes, which are, in turn, affected by various regulatory proteins [8]. Actin filaments (*F-actin*) are asymmetric, and, hence, the polymerization–depolymerization rates are different on the two ends called *the*

barbed end and *the pointed end*. In a living cell, typically, the polymerization rate at the barbed end is considerably faster, so a filament can be thought of as mainly growing from the barbed end. Barbed ends in lamellipodia and filopodia are near the cell's membrane effectively pushing it forward during polymerization. In lamellipodia, a key branching agent, *Arp2/3*, can attach to an existing filament and initiate a daughter filament growing at a specific angle with respect to the mother filament [9], thus generating a 3D actin mesh (thickness of a lamellipodium is on the order of 100 nm and transverse dimensions are above the micron length scale). Although this dendritic mesh viewpoint [10, 11], is university accepted, there appeared controversial viewpoints [12]. Another important regulator of actin network dynamics is *capping protein* that attaches to the barbed end and stops polymerization completely until it unbinds [13]. Another obvious factor influencing actin filament polymerization and growth rates is the concentration of actin monomers (G-actin) near barbed ends, which, in turn, is regulated by special sequestering proteins [14]. According to a widely accepted viewpoint, filopodia emerge when lamellipodial actin filaments group together in parallel bundles within the leading edge [15], increasing locally the pressure on the membrane and starting a new membrane-enveloped protrusion. It is thought that the pointed ends of filopodial filaments remain rooted in the lamellipodial actin mesh. Apart from capping and sequestering proteins, there are many others that regulate the growth of filopodia and lamellipodia. For example, anticapping proteins play many roles, including preventing capping proteins from attaching to the barbed ends and also promoting polymerization [16]. Fascins and other cross-linking proteins bundle the parallel filaments together, increasing the mechanical stability of actin meshes in lamellipodia and bundles in filopodia [17]. These and other multiple regulatory proteins form a complex chemical regulatory and signaling network, which allows for intricate, context sensitive control of actin protrusion dynamics, contains numerous redundancies, and, overall, is hard to decipher.

The aforementioned chemical network is tightly coupled to mechanics at the nanometer/micrometer scales. For instance, elastic force from the membrane under tension pushes back the polymerizing barbed ends at the filopodial tip and near the lamellipodial edge. Apart from that, myosin molecular motors in the lamellipodial actin mesh generate active motions, pulling the filaments into the cell body [18]. The total effect of these two (and possibly more) processes is known as *retrograde flow*, which can be clearly observed in imaging studies. Competing with polymerization, it engenders the complicated growth-retraction dynamical behavior. Actin filaments are semiflexible polymers, so they buckle under large enough load. The buckling length of individual filament is on the order of 100 nm for a few pN compressive force [19, 20]. However, parallel actin filaments can be cross-linked by special proteins, which increases the buckling force of the bundle [20]. In addition, in the cells moving over a surface or within a tissue, lamellipodial and filopodial F-actin filaments can attach to the substrate with so-called *focal adhesions*, via complex protein adhesion assembly [5]. Focal adhesions impart both direct mechanical coupling to the actin network, as well as regulate it via chemical signaling.

Summarizing, actin network dynamics in filopodia and lamellipodia arise from a complicated mechano-chemical system representing a great challenge for understanding using both experimental and theoretical approaches. Hence, computer simulations of this mechano-chemical signaling network, based on microscopic physics, may shed light on the mechanisms of network growth regulation by various proteins, complementing and providing guidance to the related experimental efforts.

2.1 Stochastic Simulations of Biological Mechano-Chemical Networks

Chemical part of the signaling network that regulates actin growth dynamics consists of various proteins whose numbers evolve in time due to numerous enzymatic and binary chemical reaction events. Most commonly, chemical reaction dynamics is analyzed by solving the corresponding system of ordinary first-order differential equations, with time as the independent variable and concentrations of interacting species as the time-dependent variables. This approach is known as chemical kinetics. The continuous concentrations in these equations correspond to the average numbers of molecules in a unitary volume. However, in reality, chemical reactions are discrete stochastic processes, where reactants encounter each other randomly, and may react or not in any given collision. Even unary reactions, such as radioactive decay are random events at a single atom or molecule level. However, when the numbers of reacting molecules are large, certainly on the order of Avogadro's number, the relative stochastic fluctuations of these numbers are negligible. In such cases, time evolution of averages gives an accurate description of the system dynamics, and deterministic chemical kinetics can be safely used, as usually done in case of macroscopic and even mesoscopic chemical reaction networks.

2.1.1 Reaction–Diffusion Master Equation

A crucial feature of biological signaling networks is that often the average numbers of molecules of each reacting protein in the relevant spatial region are very low, on the order of several molecules. In many cases, most of the time, there are no molecules of certain protein in that spatial region, with only one or two appearing for short periods of time, which produces average number of molecules lower than 1. In such a case, the fluctuations, which can be roughly estimated as the square root of the number of molecules, are on the same scale as the average and can even exceed it by an order of magnitude (in other words, all appearing molecules are noise). In these cases, chemical kinetics may not provide a physically meaningful description of the time evolution of a biological signaling network, therefore, the dynamics of such network has to be treated stochastically.

In order to illustrate the difference between chemical kinetics and stochastic approaches let us consider a unary reaction where species A can convert to B with rate k_1 and backward with rate k_{-1}:

$$A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} B. \tag{1}$$

Instead of following evolution of concentrations $[A]$ and $[B]$, the stochastic approach treats the dynamics of this chemical system as a random walk on a 1D lattice with nodes corresponding to particular copy numbers for each species involved in the chemical network. If the system starts with n molecules of A and m molecules of B (state $(\mathbf{n}, \mathbf{m})$), at any given moment one of the A molecules can convert to B with rate k_1 per molecule, and one of the B molecules can convert to A with rate k_{-1} per molecule. This translates to a step to the right along the lattice with rate nk_1 or step to the left with rate mk_{-1}:

$$\ldots \underset{(m-1)k_{-1}}{\overset{(n+2)k_1}{\rightleftharpoons}} (n+1, m-1) \underset{mk_{-1}}{\overset{(n+1)k_1}{\rightleftharpoons}} (n, m) \underset{(m+1)k_{-1}}{\overset{nk_1}{\rightleftharpoons}} (n-1, m+1) \underset{(m+2)k_{-1}}{\overset{(n-1)k_1}{\rightleftharpoons}} \ldots \tag{2}$$

The complete description of this random walk is given by the probability distribution $P(n, m, t)$—the probability that the system is in the node (n, m) at time t. This function is a solution of the so-called chemical master equation (CME):

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \widehat{M}\, P, \tag{3}$$

where $\widehat{M}$ is the reaction transition matrix operator [21]. Prior works have shown that noise induced by the discreteness of chemical reactions may result in many interesting biological phenomena, in some cases producing dynamics which is qualitatively different from the one predicted by corresponding deterministic chemical kinetics equations [22–28]. Interestingly, discrete steps in elementary chemical events allows direct mapping of the CME to a quantum field theory (QFT) formalism, with the corresponding creation and annihilation operators [24, 29]. Furthermore, the variational principle frequently used in quantum mechanics carries over, where approximate solutions may be constructed using time-dependent basis functions [30]. This correspondence highlights the computational difficulties one faces when switching from deterministic chemical kinetics (analogous to classical mechanics) to stochastic chemical kinetics (analogous to quantum mechanics).

The example discussed above is zero-dimensional, where it is assumed that reactants are well stirred, and transport is infinitely fast compared with chemical reaction rates. However, even when diffusion is fast, molecules still propagate with finite speed. Hence, for any reaction network, there is a typical distance for a molecule to travel before it reacts. This mean free path is called the *Kuramoto length* [31]: the reactive volumes with linear dimensions below the Kuramoto

length are well stirred, and may be treated with zero-dimensional kinetics, while above the Kuramoto length spatial inhomogeneities may manifest themselves, and diffusion would need to be treated as an explicit additional process. For some of the actin regulating chemical reactions, the Kuramoto length is on the order of 100 nm, hence reaction–diffusion treatment is necessary above that length-scale. In deterministic reaction–diffusion equations, the average concentrations also depend on spatial coordinates, and chemical kinetics equations are complemented by diffusion equations. For reaction volumes with 100 nm side, the copy number of proteins is typically very small, hence randomness of chemical reaction events plays an important role, and should be taken into account. The corresponding, spatially resolved stochastic system may be simulated as a collection of connected Kuramoto volumes, where chemical protein number lattice (2) is replicated in each voxel. The resulting reaction–diffusion master equation (RDME) may be written as [19, 21, 32, 33],

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \left(\widehat{M} + \widehat{D}\right) P, \tag{4}$$

where $\widehat{M}$ and $\widehat{D}$ are the reaction and diffusion operators, respectively. In general, exact analytical solution of these equations is usually beyond reach, and even numerical solution is expected to be computationally formidable. As discussed above, there is direct mapping of these equations into three-dimensional QFT, indicating both potential challenges in simulating 3D stochastic dynamics, and perhaps hinting toward using approximate QFT techniques for accelerating simulations.

2.1.2 Detailed Modeling of Filopodia and Lamellipodia

In the models employed in a series of recent works modeling actin-based organelles [19, 34–36], the space is discretized into compartments, and the basic chemical dynamical variables are the copy numbers of all chemical species in all compartments. Diffusion is realized by reactions of hopping between neighboring compartments, with a compartment size on the order of 100 nm. The rate of reactions corresponding to diffusional hops is equal to D/l_{D}^2, where D is diffusion coefficient, and l_{D} is the compartment length. In a recent work, a careful connection was drawn between the RDME rates and the kinetic rates from a more microscopic description based on Brownian dynamics of reacting particles [32]. To characterize the many-dimensional distribution function, which is the solution of the master equation, multiple realizations of the random process are run and averages, variances, joint distributions, and correlation functions are calculated. This approach is analogous to running multiple Langevin trajectories to obtain characteristics of probability distribution, which is the solution of the related Fokker–Planck equation [21].

Forces modify the corresponding reaction rates. For instance, on general grounds it is expected that the polymerization rate of a filament is exponentially damped by factor of $\exp(-f\delta/kT)$ according to Brownian ratchet model [37], where f is the force acting on the filament from the membrane, and δ is the monomer size. This is explained by the need for the filament or the membrane under load force, f, to fluctuate away from the contact point, so a new monomer can be inserted, where fluctuations are thermally distributed [37]. In general, thermal undulations of both the membrane and the filaments are on a microsecond to millisecond timescale [38–41], and hence, average out during slower reaction–diffusion time steps. The average amplitude of membrane's thermal undulations is on the order of only a few nanometers [19]. This timescale separation allows determination of mechanical objects' instantaneous configuration via corresponding mechanical energy minimization after each reaction–diffusion time step or even less frequently [35].

For filopodia, each filament is represented as a 1D lattice in space, and retrograde flow is realized through shifting this lattice backward by one lattice site (along with any proteins bound to it) in appropriate time interval calculated from the current speed of retrograde flow. The speed can have a constant part (reflecting the activity of myosins in the lamellipodial mesh) and also be influenced by forces from polymerization against the membrane or focal adhesions. The reactions in the filopodial model include protein binding to filaments, protein molecule-binding protein molecule, polymerization of the filaments, and molecular motor steps. In lamellipodia, on the other hand, a dendritic network is established and grows by the branching activity of Arp2/3, as discussed below. Retrograde flow may be considered for the lamellipodial F-actin network as well.

When considering the signaling network-regulating filopodial and lamellipodial growth, the lattice is highly multidimensional, spatially resolved, and in addition, the rates for steps between the nodes are dependent on mechanical degrees of freedom, such as retrograde flow velocity, membrane position, or bending and buckling of the filaments. It is not possible to solve these complicated equations analytically, but stochastic computer simulations may be employed [19, 42]. Hence, in a stochastic Monte Carlo approach (often called the Gillespie algorithm [43]), one draws two random numbers at each step, to decide when and which spatially resolved reaction will occur next, thus realizing the stochastic propagator on the lattice of chemical network. Afterward, the copy numbers of the involved proteins in the corresponding compartments are updated. Finally, the mechanical variables (forces, retrograde flow speed, etc.) are updated and modifications (such as filament lattice shift backward) are made if needed. The new forces change instantaneous chemical reaction rates that will influence the reaction probabilities of the next step. The results of stochastic simulations can be used to interpret various existing experiments, make new predictions, and also guide development of simpler analytical models describing the same processes, for instance, based on reaction–diffusion equations.

3 Filopodia

3.1 Biological Background

Structurally, a filopodium is a bundle of parallel actin filaments (on the order of 10) enveloped by the plasma membrane [20]. Filopodial diameter is about 100–200 nm, and their typical lengths are several microns. However, in some cases, filopodia can grow up to 100 μm in length. The primary role of filopodia is to sense the environment and help to guide the cell's motion. For instance, they are employed by fibroblasts to find their way to the wound in order to cover it [2]. During embryo formation, neurons grow axons, projecting them to macroscopic distances. A front of a growing axon—a neural growth cone—pauses every now and then, extends multiple filopodia to probe the surroundings, helping to decide in which direction it has to turn next [3]. Filopodia are also needed for dorsal closure in drosophila embryos [4] and are implicated in cancer metastasis [6].

Polymerization of F-actin at the barbed ends near the cell's membrane is pushing the membrane forward, forming the protrusion and driving the filopodial growth. The filopodial F-actin growth rate is highly regulated by different mechanisms. First, free G-actin can be sequestered by thymosin-β4 binding, reducing concentration of monomers available to polymerization [14]. Profilin binding to G-actin can promote association of monomers specifically at the barbed end [44]. Opposing that, capping proteins can bind to the F-actin barbed ends and stop their polymerization [13]. To counter that, there are anticapping proteins, like formins, that also bind to the barbed ends and not only prevent capping but also often dramatically increase the polymerization rate [16]. Apart from chemical regulation of the growth rate, a considerable role in filopodial formation and growth is played by mechanics. Elastic membrane counterforce due to membrane tension, on the order of 10 pN, suppresses the polymerization rate. In addition, this membrane force can buckle actin filaments, which are semiflexible polymers with persistence length on the order of 20 μm [45]. However, there are cross-linking proteins, such as fascins [17, 46] or Ena/VASP, which can bind to different filaments simultaneously, so that the buckling force of the filopodial bundle significantly increases. Resistance to bending of the membrane envelope can additionally protect filopodia against buckling [47].

Membrane forces also contribute to the generation of the retrograde flow—a gradual motion of the filaments back to the bulk of the cell. Another important contribution to the retrograde flow comes from special machinery in the cell bulk, where filopodium is rooted. Active motions in the lamellipodial actin mesh, which are driven by myosin molecular motors, pull the filaments into the cell [18]. It turns out that the retrograde flow can dramatically influence the filopodial growth speed, and even turn growth into retraction [19]. However, retrograde flow speed itself can be influenced by chemical interactions among cytoskeletal proteins. For instance, when a cell moves on a surface, actin filaments in the cell's filopodia and lamellipodia can attach to the substrate via focal adhesions [5, 48]—protein

complexes which assemble into a link between F-actin and the substrate. Focal adhesions will counteract the forces generating the retrograde flow, slowing the flow down. The point of attachment to the filament can migrate due to retrograde flow or, if focal adhesion complex includes a myosin motor, due to walking of that motor along the filament. Therefore, the link between the filaments and the focal adhesions will stretch and generate force which pulls both at the substrate and the attachment of the focal adhesion to F-actin [48]. This force will promote focal adhesion release, bringing about yet another round of mechanical regulation. Altogether, these mechanisms are heavily intertwined into a mechano-chemical network of regulatory interaction. Furthermore, apart from these most important players, there are other proteins involved in filopodial regulation, for instance, those that form *the filopodial tip protein complex*, which can be observed on the EM images [15, 46], but remains somewhat mysterious, as neither its exact role nor its detailed protein composition is yet understood.

Why is there a need for such complicated regulation? A filopodium is primarily a probe, so it must be adaptable and sensitive to changes in cell's environment. Fluctuations in the network permit such sensitivity [34]. Experimentally, some filopodia are observed to switch every now and then between growth and retraction phases in response to the environmental cues or to internal random fluctuations.

Last but not the least, the filopodial regulatory mechano-chemical interaction network is spatially extended, so the transport of involved proteins is another crucial aspect of its dynamics. Proteins that have a function at the tip (including G-actin monomers) have to be delivered there, but diffusion becomes less efficient as the filopodium grows. Other possibilities for protein transport are delivery by molecular motors, traveling bound to filaments subject to retrograde flow and in hydrodynamical flows. Transport adds on to the complexity of the regulatory network, as it essentially sets up the baseline concentrations of chemically and mechanically interacting proteins of the network.

Due to complexity of the protein interaction network in filopodia, it is very difficult to devise experiments elucidating the behavior of the network as a whole. However, it is more feasible to find out a microscopic function of any given protein, sometimes, even with measurements of relevant reaction rates. Subsequently, modeling can help to explain the implications of the particular protein on the dynamics of the whole regulatory network and make new predictions. Therefore, such modeling has to incorporate important microscopic details, and include all three interplaying features of the network— chemistry, mechanics, and transport.

3.2 Chemistry, Mechanics, and Transport in Filopodia

Since the filopodium is an elongated organelle, its length and growth speed are its most natural characteristics. They are also natural experimental observables, directly measurable through microscopy. From this perspective, one of the most

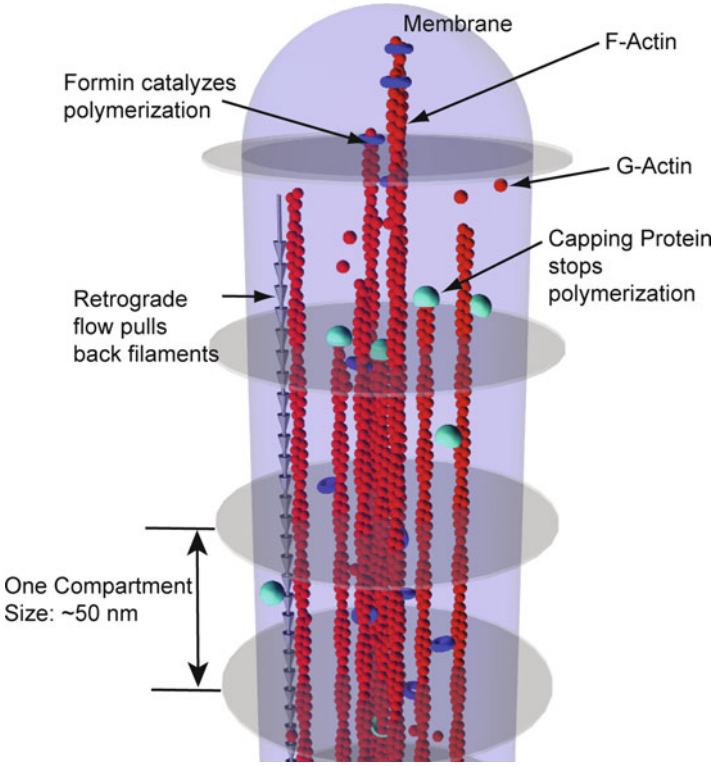

Fig. 1 The following processes were included into a computational model of a matured filopodium [19,34]: (1) monomeric G-actin diffusion along the filopodial tube; (2) (de)polymerization of individual actin filaments; (3) fluctuating membrane under load which slows down the filament polymerization rates; (4) a constant velocity retrograde flow, where actin filaments are pulled into the cell body; (5) capping proteins arrest polymerization; (6) formins accelerate polymerization

fruitful approaches to describing the filopodial dynamics is to look primarily at the protein fluxes. Some fluxes are passive, such as diffusion and certain chemical reactions, for example many binding events, others are driven by energy consumption through ATP hydrolysis, such as retrograde flow, polymerization, or molecular motor walking. Together these fluxes result in the highly nonequilibrium filopodial dynamics. Essentially, in filopodia, the interplay of chemistry, mechanics, and transport translates into the competition of fluxes.

In a formed filopodium (see Fig. 1), there are three main fluxes of actin [19]. First, diffusion of G-actin from the bulk of the cell to the filopodial tip is a transport flux, J_{T}. Based on stationary solution of an 1D diffusion equation, the diffusion flux, J_{D}, may be estimated as,

$$J_{\mathrm{T}} = J_{\mathrm{D}} = D\frac{c_{\mathrm{tip}} - c_{\mathrm{base}}}{L}, \tag{5}$$

where D is the protein diffusion constant, c_{tip} and c_{base} are protein concentrations at the ends of the filopodial tube, and L indicates filopodial length. It is reasonable to assume that the protein's concentration at the filopodial base is maintained at the bulk cell concentrations. Second, polymerization flux at the tip, J_{P}, converting G-actin to F-actin, is a chemical (reaction) flux,

$$J_{\text{P}} = \frac{N}{S_{\text{f}}}\left(k^{+}c_{\text{tip}} - k^{-}\right), \tag{6}$$

where N is the number of filaments, S_{f} is the filopodial cross-section area, k^{+} and k^{-} are instantaneous polymerization and depolymerization rates at the tip. Last, retrograde flow pulling F-actin back to the cell is a mechanical (and also transport) flux,

$$J_{\text{R}} = \frac{N v_{\text{R}}}{S_{\text{f}}\delta}, \tag{7}$$

where v_{R} is the instantaneous retrograde flow speed, and δ corresponds to actin monomer size. When these fluxes are balanced,

$$J_{\text{D}} = J_{\text{P}} = J_{\text{R}}, \tag{8}$$

so that all G-actin diffusing to the tip gets converted to F-actin by polymerization and then pulled back to the cell, the filopodium length is stationary [19]. Solving (5)–(8) leads to the following stationary length of filopodia,

$$L_{\text{stationary}} = \frac{S_{\text{f}}D}{N}\left(\frac{\delta}{v_{\text{R}}}\left(c_{\text{base}} - \frac{k^{-}}{k^{+}}\right) - \frac{1}{k^{+}}\right), \tag{9}$$

where k^{+} is not the bare polymerization rate but is exponentially suppressed by the average membrane force, f,

$$k^{+} = k^{0}\exp\left(-\frac{f\delta}{N k_B T}\right). \tag{10}$$

It turns out that (9) and (10) reproduce detailed stochastic simulations of an actin only filopodial system within 5–10% accuracy [19], where the latter are computationally very intensive. Notice, with retrograde flow rate diminishing, the filopodial length keeps growing, indicating that regulation of retrograde flow rate can powerfully control filopodial dynamics. In addition, increasing the bulk cell concentration of G-actin will linearly increase the filopodial length. Finally, increasing the actin bundle size, N, diminishes the resistance on individual filaments, increasing effective polymerization rate, k^{+}, but also requires more actin transport, hence seriously exacerbating the transport bottleneck problem when the bundle becomes too thick [as evidenced by the N^{-1} term in front of (9)].

Actin-only models of filopodial growth that do not include chemical regulation predict growth of filopodia up to this stationary length with minuscule fluctuations

around the average [19, 20]. In these models, stationary length turns out to be the maximum length as well. If filopodia were to grow longer, diffusion would not be able to provide enough G-actin, as the diffusional flux decreases with filopodial length and polymerization and retrograde flow stay the same [see (5)–(7)]. Additionally, actin-only models predict nearly infinite filopodial lifetimes [19, 20].

Various regulatory proteins can modify these fluxes, hence affecting the filopodial growth speed and steady-state lengths via (8). They can stop, slow down, or accelerate certain reactions, changing the polymerization flux, J_P. For instance, capping protein stops polymerization, formin accelerates it, and G-actin sequestration by thymosin-$\beta 4$ slows it down. Mechanics also modifies chemical fluxes, for example, an increase in membrane force, or an obstacle in front will slow down the polymerization rate. The regulatory proteins can change the mechanics, as in focal adhesion formation or fascins protecting filopodia from buckling instabilities. Finally, additional proteins can change the transport flux, J_T for instance, by additional fluxes supplied by molecular motors carrying cargo.

3.3 Chemistry

Polymerization is the basic and most important reaction in filopodia, so in this section we will focus on chemical modification of the reaction flux, J_P. Formins increase the polymerization rate up to fivefold [16], but capping proteins completely stop polymerization for a particular filament until they fall off. One would therefore expect that, on average, the simultaneous effect of formins and capping proteins on filopodial growth would not be dramatic. However, extensive computer simulations that take into account microscopic details show that the filopodial dynamics is dramatically altered, due to discrete noise in the system arising from fundamental randomness of chemical reaction events [34]. If a filament becomes capped, it starts to retract due to retrograde flow, and might fully retract back to the cell bulk, if the capping protein does not fall off during the time of retraction. As a result, the number of growing filament decreases, which, in turn, decreases polymerization and retrograde flow actin fluxes. Since a lower diffusional flux is needed to maintain these decreased fluxes, filopodia can grow longer [34]. However, as filaments disappear, it becomes harder for them to sustain the membrane force, so at some critical number of filaments (about 3–5) the filopodium starts to retract as a whole [34]. Capping proteins fall off from the barbed ends at a slow rate, so when enough filaments are uncapped to overcome the membrane force, the filopodium switches back to growth. Eventually, at some retraction phase, the filopodium would retract all the way to the cell body and disappear. So, due to an introduction of a simple protein regulation, the computational model predicts finite filopodial lifetime (on the order of several minutes in consistency with experimentally reported values [49–52]) instead of growth to a stationary length.

The model also suggests a possible mechanism for growth–retraction cycles. Capping proteins are present in the filopodium in very low concentrations, so their

fluctuations are important. Capping process is then a discrete (either capped or uncapped) random noise. It is also slow, due to low capping protein concentration. The discreteness and slowness make capping process act as a random switch between two fast processes of growth and retraction. The filopodium thus acts as an amplifier of capping protein binding noise, making it visible on macroscopic temporal ($\sim$100 s) and spatial ($\sim$1 μm) scales. This high susceptibility to tiny fluctuations may be profitably exploited for the sensorial role of filopodia [34].

It should be noted that in order to describe such effects, the model has to be stochastic and microscopic and include molecular details, because it is molecular noise of chemical binding that translates into macroscopic mechanical observables through mechano-chemical amplification.

3.4 Transport

Modeling the interconversion of actin fluxes in a filopodium suggests that the diffusional flux can not sustain the growth above several microns (at biological actin concentrations and retrograde flow speed values), even if the number of polymerizing filaments is decreased by capping proteins [19, 20, 34]. In some experiments filopodia can grow over 80 microns, so there have to be other mechanisms altering actin fluxes to make such growth possible [50]. Also, diffusional flux is not enough to account for the experimentally observed growth speeds of about 10 μm/min [50].

Modeling various mechanisms of flux regulation (even such that are not at present proven to exist from experiments) allows to consider the potential effectiveness and likelihood of various possible scenarios and then help to focus experimental research on the most plausible mechanisms. Since modeling requires less resources than experiments, this is a great way to move forward our understanding of the regulatory processes in filopodia.

Downregulating polymerization flux or retrograde flow will slow down growth and/or retraction speeds. Therefore, to provide enough actin for long, fast growing and retracting filopodia, additional flux of G-actin to the filopodial tip is needed besides diffusion. A "standard" biological solution for underperformance in diffusional transport is the use of molecular motors. Some molecular motors can walk on actin filaments in a directed fashion, that is, the shift in their spatial position is proportional to time, not square root of time, as in diffusion. They can bind cargo, and drag it along while walking, thus realizing active transport of the cargo [53, 54].

In fact, Myosin X molecular motors, which can walk along actin filaments, have been observed inside filopodia and shown to influence filopodial formation [55]. Their exact role is not known, and they have not been observed to carry G-actin experimentally. This is where the modeling can help: it is straightforward to introduce into simulations myosins with ability to walk, diffuse, and bind G-actin, and see if it will considerably increase the filopodial lengths or growth speeds. The picture of active transport suggested by cartoons in biology textbooks shows cargo loaded onto motors which walk forward and unload cargo at the destination,

much like a conveyor belt. It turns out, that motors are not working in this way for two reasons [36]. First, since they bind G-actin, the latter becomes unavailable for the polymerization, so the motors end up *sequestering* their cargo. The more motors in filopodia, the more actin they would bind, including when motors are in the cytosol, which means more sequestration apart from additional transport flux. The second reason is that the excessive amount of motors will create "traffic jam" on the filament thus impeding the additional active transport flux. Together, these two phenomena prevent active transport flux to increase the filopodial length by more than 30%, which is still on the scale of 10 μm. However, with some help from chemistry, the active transport could do more. For instance, Ena/VASP is known to be transported by Myosin X inside filopodia [56] and also cross-links the actin filaments near the tip [17]. As Ena/VASP also has G-actin binding domains, it is plausible to propose that Ena/VASP is needed as an adaptor between G-actin and Myosin X. Again, modeling allows to see the consequences of this assumption. If Ena/VASP is also consumed for cross-linking near the tip, it would promote release of G-actin from the motors near the tip, hence alleviating the sequestration problem. In such case, modeling predicts severalfold increase in the filopodial length [36]. In a narrow range of parameters, the active transport actin flux by itself is enough to overcome the retrograde flow flux, so even thinning of the diffusional flux in a continuously elongating filopodium does not stop the growth [36]. Summarizing, active transport can considerably promote the growth but only with a help of intricate chemical reaction network involving multifunctional proteins (Fig. 2).

3.5 Mechanics

Mechanical timescales in the actin dynamics, such as thermal undulations, are generally much faster than those for reaction–diffusion events. In the case of filopodia, this time-scale separation allows to solve mechanical part of the problem separately and then use the solution to affect chemistry and transport. The most fundamental question is the mechanical behavior of actin filaments themselves. As semiflexible polymers, they can be described by the wormlike chain model [57]. A bundle of cross-linked filaments under load is considerably more complicated system with a wider spectrum of behaviors. The major questions are the following: at what lengths and forces bundle will buckle? How does it depend on number of filaments and character and number of cross-links? What shape will the bundle adapt?

A large body of studies were dedicated to answering these questions with the help of modeling. One of the first estimations based on classic elastic theory enhanced by simulations predicted a mechanical limit on filopodial length [20] in addition to diffusional transport limit. The buckling force was found to be proportional to the number of filaments N if they are strongly cross-linked, and to $N^{1/2}$ if they are weakly cross-linked. The N-dependent mechanical limit on the lengths becomes especially important if capping proteins decrease the amount of filaments. Taking

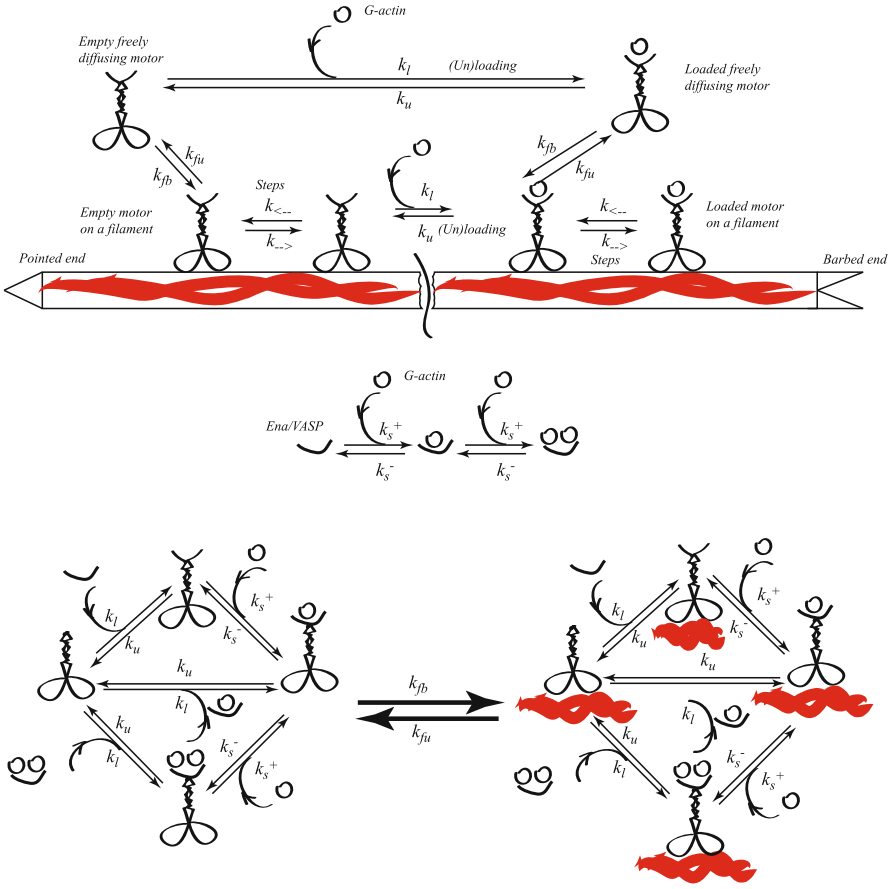

Fig. 2 Models of chemical reaction networks for motor transport in filopodia. (*Upper*) Simple scheme with motors diffusing, walking, binding to the filaments and loading G-actin as a cargo is drawn. (*Middle*, *Lower*) A scheme with Ena/VASP playing the role of adaptor between a motor and G-actin monomer is shown

the enveloping membrane into account can make the model more complicated. A couple of recent works discussed if a particular buckling shape could help obviate Euler instability, making longer filopodia possible [47, 58]. A model for cross-linked bundles based on local fields for filaments deformation and cross-links through numerical simulations and scaling analysis predicts very intricate bending behavior of the bundles, depending on bundle dimensions and sheer stiffnesses of filaments and cross-links [59]. The latter model also considers a case where the bundle consists of fractured filaments that is, overlapping pieces of filaments that do not run the whole length of the bundle.

In the filopodial models, the force against which the polymerization is occurring mainly decreases the polymerization rates of filaments and contribute to generating the retrograde flow. This resistance may come from the force from mechanical

obstacle or be a pure membrane force due to surface tension. Thermal dynamics of the membrane is much faster than average interval between chemical reactions, so the fluctuations of the membrane can be averaged to find, for instance, distribution of the force between the filaments. Bending energy of the membrane also tries to minimize the area of stretched membrane, thus inducing attraction and merging of small F-actin bundles that have slightly protruded the membrane [60, 61]. Actin growth against obstacle is another broad topic for modeling. Intracellular parasite *Lysteria* hi-jacks cell's actin and uses parallel bundle polymerization for propulsion [62]. Filopodial growth against a spherical obstacle in vitro influences orientation of the filaments [63].

Finally, force from focal adhesions can affect retrograde flow. Given the paramount role of retrograde flow in setting up transport fluxes, this aspect of filopodial dynamics is also important. Due to retrograde flow, the focal adhesions stretch (as points of attachment to the filament slide back) and pull at the substrate and at the filaments. The stretching force influences focal adhesion disengagement rate and the field of substrate deformation. Deforming substrate affects stretching of the focal adhesions which, in turn, affects the instantaneous retrograde flow. Even the simplest model of this interplay of forces shows interesting physics, and is consistent with the corresponding experimental measurements [48].

4 Lamellipodia

4.1 Introduction

When placed on a substrate, eukaryotic cells crawl by projecting forward flat sheet-like protrusion structures called lamellipodia which contain a dynamically remodeling three-dimensional actin mesh network [1, 64]. A lamellipodium is composed of dendritically branched actin filaments, which elongate through polymerization at their barbed ends to push cell membrane forward, and new filaments are nucleated from the existing filaments [10, 65]. While the dendritic nucleation/array treadmilling model [10, 11] provides a conceptual model for lamellipodial protrusion, understanding in microscopic detail how cells coordinate the enormous number of molecules involved in motility process to achieve optimal movement remains a challenging task. Mathematical modeling and computer simulations have been increasingly applied to help advance the understanding of growth and force generation in dendritic actin networks [9, 35, 37, 42, 60, 66–82]. In this section, we highlight recent progress on the stochastic simulations of lamellipodial protrusion dynamics, based on detailed chemistry and physics, which have considerably advanced our understanding of the microscopic physics of cellular motility.

Cell motility based on dendritic actin network growth and remodeling is a complex process. It is useful to study it with simple cells such as fish keratocytes, an excellent model system due to the simplicity of their canoe-like shape and

persistent and fast motion [65,77,83,84]. Moreover, despite the complexity of actin-based motility, it has been found that simplified models with reconstituted in vitro system consisting of only the most essential elements can offer great insight on the physics behind actin-based motility [81,85]. Thus, it should be feasible to construct physically based simple computational models with a relatively small number of components to study actin-based motility. Taking into account these considerations, a recent 3D model of the growth dynamics of lamellipodia-like mesh network was developed to investigate how capping and anti-capping proteins regulate growth dynamics of such branched filamentous networks [35,80] (see Fig. 3 for a schematic illustration of the model). It was observed in the stochastic simulations that with introducing capping proteins and anti-capping proteins, the density of membrane leading edge filaments changes correspondingly, which in turn leads to the change of the number of G-actin monomers available for polymerization and the load on polymerizing filaments. Thus, the density of leading edge filaments is a simple but critical quantity controlling protrusion dynamics, since it governs both the pool of G-actin monomers available for polymerization and the instantaneous load polymerizing filaments experience. By observing how the density of filament changes with the varying concentration of regulatory proteins, one can gain valuable insight on how the regulatory proteins such as capping and anti-capping proteins control the actin dynamics in lamellipodial-like branched network.

Lamellipodial protrusion dynamics is controlled by complex mechano-chemical feedbacks. The interplay between molecular processes such as the diffusion and reactions of various molecules and mechanical characteristics of the system such as cell membrane, the cytoskeleton, and adhesion to substrate determines the motile behavior of crawling cells. In the following sections, we'll give a brief introduction to these important aspects of actin-based motility in lamellipodia.

4.2 Chemical Feedbacks Regulate Actin Mesh Growth

4.2.1 Elongation vs. Nucleation of Actin Filaments

Monomeric actin is the building block of filaments and its availability is key to the efficient elongation of filaments; moreover, actin is also a key component in Arp2/3-mediated nucleation process, in which the nucleation of a new filament on an existing filament requires both an activated Arp2/3 and a G-actin molecule [86]. Thus, it is essential to understand the interplay between elongation and nucleation processes. In cells, actin is one of the most abundant proteins, and its availability facilitates both the speed of protrusion and the rate of nucleation. However, if the concentration of actin were kept constant, and the concentration of Arp2/3 were varied, it turns out that the rate of filament nucleation changes monotonically, but there exists an optimal Arp2/3 concentration at which the protrusion speed is maximal [35]. Arp2/3 facilitates the nucleation process, which to a large extent determines the density of leading edge filaments. On the one hand, at low Arp2/3

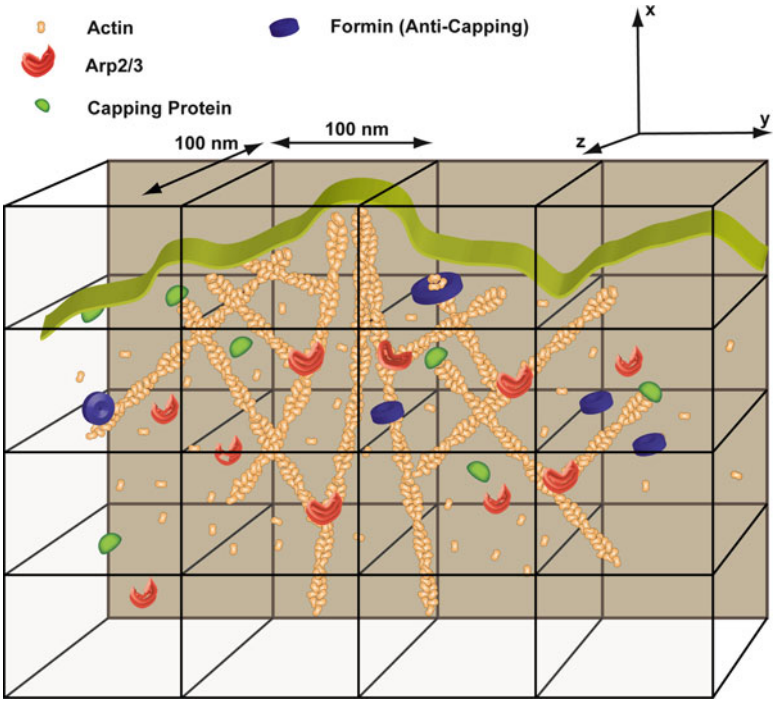

Fig. 3 The following processes were included into our computational model of a lamellipodial protrusion [80]: (1) Stochastic hopping of various monomeric proteins between neighboring voxels; (2) polymerization and depolymerization events for individual F-actin filaments; (3) binding of Arp2/3 to sides of existing filament to nucleate a daughter filament at approximate 70° angle; (4) binding of capping proteins and formins to polymer barbed ends, to correspondingly stop and accelerate the polymerization process. Filaments sterically protrude against the cell membrane, locally deforming it, while the membrane provides some resistance against the bending deformations and an increase of the membrane area. The counteracting membrane push against the filaments slows down polymerization of filament tips that bear the most force. A boundary with bulk reservoir of monomers is placed at the back of the lamellipodium, several microns down in the x direction. This results in establishing of monomeric gradients longitudinally across the lamellipodium for species that are actively consumed in front, and have to be diffusively transported from the rear to replenish the local pool

concentration, filamentous network is sparse, leading to high protrusive resistance on filaments; on the other hand, at high Arp2/3 concentration, actin filaments in the dense network would deplete the local monomeric actin pool. Both cases are unfavorable to the polymerization of actin filaments, although the causes are different. Overall, this observation indicates that having balanced polymerization and nucleation rates is important in order to producing maximal protrusion speeds. The results obtained from microscopic simulations [35] are qualitatively consistent with the theoretical analysis using a set of deterministic reaction–diffusion partial differential equations [66].

4.2.2 The Antagonism Between Capping and Anti-capping Proteins Affects Actin Network Dynamics

Capping proteins and anti-capping proteins such as formin and Ena/VASP are key regulators of actin network dynamics [87, 88]. Capping proteins block the polymerization of actin filaments, thus, one might expect them to inhibit motility. Motility inhibition by capping proteins occurs just as expected at sufficiently high concentrations of capping proteins. However, at modest concentrations, capping proteins can increase the speed of actin-based motility [35, 78, 80, 89–91]. This interesting phenomenon was explained by two fundamentally different ideas: the actin funneling hypothesis [89] proposes that capping proteins increase the rate of individual growing barbed ends by reducing their number, while the monomer gating model [91] suggests that motility enhancement is due to more frequent filament nucleation. Motivated by these studies, stochastic simulations [35, 80] based on microscopic chemical physics have been carried out to investigate the mechanism of capping proteins promoting motility. The advantage of computer simulation is that it offers great details on various microscopic quantities involved in motility process for analysis: it was found that with capping proteins, on average, there are more actin monomers available for polymerization, leading to the faster rate of polymerization at low capping protein concentrations, which is in agreement with the actin funneling hypothesis; although capping proteins indeed promote nucleation of filaments, also the consequence of increased local actin concentration, many filaments become capped and lag behind the leading edge of the membrane, resulting in a diminution of the filament density along the leading edge of the membrane. This, in turn, leads to higher load on polymerizing filaments, unfavorable to polymerization process, especially at high capping protein concentrations.

On the other hand, anti-capping proteins compete with capping proteins for barbed ends binding, thereby affecting actin dynamics by keeping filament density at leading edge from being diminished by capping proteins. This, in turn, affects filament polymerization since the average local actin concentration for polymerization and the average load on polymerizing filaments are highly correlated to the density of leading edge filaments. It should be pointed out that anti-capping proteins such as formins can increase the rate of polymerization dramatically. This polymerization rate enhancement function by anti-capping proteins makes the dynamic behavior of the motility system even more diverse [80]. In particular, it turns out that the coupling of the capping/anti-capping regulation with Arp2/3 nucleation activity allows the cell to robustly achieve maximal protrusion speed under broad set of conditions [80].

In summary, protrusion dynamics in a motility system containing capping proteins may display both the enhanced and inhibited behaviors, which can be significantly affected by the presence of anti-capping proteins. Detailed analysis of microscopic quantities obtained from stochastic simulations offers great insight on the protrusion dynamics of lamellipodial-like branched network, allowing to discriminate among competing qualitative hypotheses. Furthermore, it is known

that the movement of cells is usually robust yet adaptive—cells can respond to the environmental changes efficiently. With the insights obtained from the analysis on how a group of chemical species with opposing action regulate the protrusion dynamics of the model motility system, we can better understand how cells can finely tune the microscopic polymerization process to achieve certain dynamical behaviors.

4.2.3 Transport of Molecules

Molecular processes in cell motility are controlled by the reaction–diffusion of various molecules, thus effective transport of molecules plays an important role in regulating protrusion dynamics.

The rear part of the lamellipodium of the cell is coupled to the lamellum (cell body), which serves as a reservoir for molecules. This can be modeled by coupling the rear part of the lamellipodia to a bulk reservoir of various molecules, whose concentrations are all kept fixed. This coupling is analogous to imposing a boundary condition in deterministic reaction–diffusion equations. Actin monomers are consumed and recycled through the treadmilling process, resulting in a concentration gradient from the bulk region to the leading edge; the nucleation of filaments occurs within the activation zone of the membrane, also leading to a concentration gradient for Arp2/3 molecule from the bulk to the membrane [35]. From the profiles of concentration gradients, one can derive the local concentrations of Arp2/3 and actin in the region close to the membrane, where it is of central importance since it is the location where nucleations occur and also is the main location for polymerizations. With this, we can analyze what limits the growth of lamellipodial network [35]. For example, when actin monomers are abundant, this favors faster nucleation and tends to deplete Arp2/3, whose bulk concentration is assumed to be constant and low in absolute value ($\sim$100 nM). In such case, nucleation is limited by the availability of Arp2/3, and the filamentous network would be sparse. In a sparse filamentous network, there are not enough filaments to push the membrane, and hence, the protrusion speed would be adversely affected. As in the filopodial transport discussed above, the abundance of proteins in the cell body does not necessarily indicate that there would be no problem with protein's availability at cell's leading edge: if the protein is actively consumed, it needs to be continuously transported, hence, significant local depletion may still result due to potential transport bottlenecks [35].

4.3 Mechanical Aspects of Lamellipodial Protrusion

4.3.1 Cell Membrane

The cytoskeleton of eukaryotic cells is enclosed by the cell membrane. Cell migration relies on the force generated from polymerizing actin filaments to push

cell membrane, which in turn exerts the force on the barbed ends of actin filaments and inhibits the growth of filaments. Besides the physical confinement on filaments and the mechanical impedance on the polymerization of actin filaments, the cell membrane also plays an important role in the nucleation of actin filaments since the activation of Arp2/3 relies on signals from receptors on cell membrane. The cell membrane is also involved in cell adhesion connecting the cytoskeleton to the extracellular matrix.

To incorporate the cell membrane in mechanical models of actin-based motility, it is necessary to have a proper representation of the membrane. While membrane may be modeled as a rigid obstacle against which the branched actin network grows [67, 68], models with simplified flexible membrane offer more realistic representation of the membrane behavior [9]. Finite element method has also been applied to model cell membrane in lamellipodial studies [20, 60]. With the representation of the membrane set, the energy of the membrane including bending and tension terms could then be written, and the interaction between the membrane and actin filaments can also be introduced to mimic the force generation process [9, 35]. Since the protruding plasma membrane grows against the external obstacle, this additional resistive force in turn impact the polymerizing filaments below the membrane. The effect of external load can be modeled by introducing an effective external field which acts on the cell membrane [35]. Higher load from membrane would slow down the growth of actin filaments.

4.3.2 Re-organization of the Actin Network: From Lamellipodia to Filopodia

Actin filaments are commonly present in cells, and they may form different structures including the branched network—lamellipodia and the bundled network—filopodia, as reviewed in preceding section of this chapter. These different types of actin networks are mediated by different regulatory proteins: in branched network, filaments are cross-linked by Arp2/3; while in bundled network, filaments are connected by actin-binding proteins such as fascin and Ena/VASP. Elucidating how the different types of actin networks are controlled by physico-chemical factors is of fundamental significance in understanding the shape and motile behaviors of cells. Understanding the mechanism of bundling process may help understand the formation of filopodia from lamellipodia. The convergent elongation model provides an insightful explanation on the mechanism of filopodia initiation by re-organization of the dendritic network [15].

The phase behavior of charged rods in the presence of inter-rod linkers has been studied theoretically as a model for the equilibrium behavior underlying the organization of actin filaments by linker proteins in the cytoskeleton [92]. The phases include bundle-dominant structure, network-dominant structure and phase containing both types of structures. The reconstitution of the transition from lamellipodium (2D aster) to filopodium (star) in membrane-free system has also been carried out: in the motility system containing no fascin, there is spontaneous formation of aster-like structure; and in the presence of fascin, these asters transition

into stars with actin bundles, and capping protein inhibits star formation [73]. Experimental and 3D kinetic Monte Carlo studies of Arp2/3 branched actin network mediating filopodia-like bundles formation in vitro have also been performed [76]. The latter study showed that the energy gain due to fascin bundling outweighs the unfavorable energy to bend the filaments to form a bundle.

5 Summary

Until very recently, actin-based cell motility was investigated mainly experimentally, with a wealth of generated data that was hard to synthesize together into a coherent picture. Hence, many competing and sometimes contradictory interpretations are commonly found in the current literature. Computer simulations based on microscopic chemistry and physics provide a unique opportunity to examine some of the common scenarios of actin polymerization dynamics, with the aim of pointing out more physically plausible hypothesis, and also making new predictions that can be tested experimentally. Deterministic reaction–diffusion calculations work at a coarser level, allowing to see the bigger picture, but occasionally fail due to lack of some critical physical ingredient. Hence, a combination of both stochastic simulations where individual filaments are geometrically resolved and molecules randomly hop around and react, and continuous reaction diffusion models without explicit actin network geometry will be needed to make future progress in deciphering mechano-chemical networks controlling actin dynamics.

In terms of what we have learnt from recent computational studies, it is clear that actin network dynamics is controlled through the fine balance of chemical, mechanical, and transport processes. In particular, because single filaments under tension are thought to grow exponentially slower, the filament number density is a critical physical observable that controls growth speed. When the network is sparse, too few filaments hold the membrane resistance, diminishing polymerization rates. When the network is too dense, the need to feed numerous growing ends creates a severe transport bottleneck, also slowing down protrusion speed. Hence, many regulatory proteins, such as capping proteins, anti-capping proteins, and Arp2/3, modulate the network dynamics by directly or indirectly influencing the density of filaments. Furthermore, stochastic effects can sometimes be dramatic, where for example the molecular noise of capping protein binding and unbinding can be amplified to macroscopic length- and time-scales. Finally, the importance of delivering material to the polymerization front, the diffusional transport bottleneck, is becoming appreciated and new suggestions have been put forward about how active transport by molecular motors could alleviate the local depletion of monomers.

As we gain understanding how the simplest components of actin's mechano-chemical machinery work, future computer simulations will undoubtedly shed more light on the need and mechanisms of action of several dozen additional regulatory proteins that are important in actin-based protrusion dynamics. Furthermore, more realistic coupling of stress distributions in the actin network with chemical processes, and better modeling of interactions with the external substrate will lead to more comprehensive understanding of physics and chemistry of eukaryotic cell motility.

Acknowledgment We are grateful for the support from the National Science Foundation under CAREER award CHE-0846701.

References

1. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P.: Molecular Biology of the Cell, 4th edn. Garland Science, New York (2002).
2. Noselli, S.: Drosophila, actin and videotape—new insights in wound healing. Nat. Cell Biol. **4**(11), E251 (2002).
3. Dent, E.W., Gertler, F.B.: Cytoskeletal dynamics and transport in growth cone motility and axon guidance. Neuron **40**(2), 209 (2003).
4. Jacinto, A., Wood, W., Balayo, T., Turmaine, M., Martinez-Arias, A., Martin, P.: Dynamic actin-based epithelial adhesion and cell matching during drosophila dorsal closure. Curr. Biol. **10**, 1420 (2000).
5. Yamazaki, D., Kurisu, S., Takenawa, T.: Regulation of cancer cell motility through actin reorganization. Cancer Sci. **96**(7), 379 (2005).
6. Lorenz, M., Yamaguchi, H., Wang, Y., Singer, R.H., Condeelis, J. Imaging sites of n-wasp activity in lamellipodia and invadopodia of carcinoma cells. Curr. Biol. **14**(8), 697 (2004).
7. Carlier, M.F. (ed.): Actin-Based Motility: Cellular, Molecular and Physical Aspects, 1st edn. Springer, New York (2010).
8. Mejillano, M., Kojima, S.I., Applewhite, D.A., Gertler, F.B., Svitkina, T.M., Borisy, G.G.: Lamellipodial versus filopodial mode of the actin nanomachinery: pivotal role of the filament barbed end. Cell **118**, 363 (2004).
9. Schaus, T.E., Taylor, E.W., Borisy, G.G.: Self-organization of actin filament orientation in the dendritic-nucleation/array-treadmilling model. Proc. Natl. Acad. Sci. USA **104**(17), 7086 (2007).
10. Mullins, R.D., Heuser, J.A., Pollard, T.D.: The interaction of arp2/3 complex with actin: nucleation, high affinity pointed end capping, and formation of branching networks of filaments. Proc. Natl. Acad. Sci. USA **95**, 6181 (1998).
11. Pollard, T.D., Borisy, G.G.: Cellular motility driven by assembly and disassembly of actin filaments., Cell **112**(4), 453 (2003).
12. Yang, C., Svitkina, T.: Visualizing branched actin filaments in lamellipodia by electron tomography. Nat. Cell Biol. **13**(9), 1012 (2011).
13. Schafer, D.A., Jennings, P.B., Cooper, J.A.: Dynamics of capping protein and actin assembly in vitro: uncapping barbed ends by polyphosphoinositides. J. Cell. Biol. **135**(1), 169 (1996).
14. Cassimeris, L., Safer, D., Nachmias, V.T., Zigmond, S.H.: Thymosin beta 4 sequesters the majority of g-actin in resting human polymorphonuclear leukocytes. J. Cell. Biol. **119**(5), 1261 (1992).
15. Svitkina, T.M., Bulanova, E.A., Chaga, O.Y., Vignjevic, D.M., Kojima, S.I., Vasiliev, J.M., Borisy, G.G.: Mechanism of filopodia initiation by reorganization of a dendritic network. J. Cell Biol. **160**(3), 409 (2003).
16. Vavylonis, D., Kovar, D.R., O'Shaughnessy, B., Pollard, T.D.: Model of formin-associated actin filament elongation. Mol. Cell **21**(4), 455 (2006).
17. Schirenbeck, A., Arasada, R., Bretschneider, T., Stradal, T.E.B., Schleicher, M., Faix, J.: The bundling activity of vasodilator-stimulated phosphoprotein is required for filopodium formation. Proc. Natl. Acad. Sci. USA **103**(20), 7694 (2006).
18. Heid, P.J., Geiger, J., Wessels, D., Voss, E., Soll, D.R.: Computer-assisted analysis of filopod formation and the role of myosin ii heavy chain phosphorylation in dictyostelium. J. Cell Sci. **118**(Pt 10), 2225 (2005).
19. Lan, Y., Papoian, G.A.: The stochastic dynamics of filopodial growth. Biophys. J. **94**(10), 3839 (2008).

20. Mogilner, A., Rubinstein, B.: The physics of filopodial protrusion. Biophys. J. **89**(2), 782 (2005).
21. van Kampen, N.G.: Stochastic Processes in Physics and Chemistry. North-Holland, Amsterdam, (1992).
22. Lan, Y., Papoian, G.A.: Stochastic resonant signaling in enzyme cascades. Phys. Rev. Lett. **98**(22), 228301 (2007).
23. Kepler, T.B., Elston, T.C.: Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. Biophys. J. **81**, 3116 (2001).
24. Sasai, M., Wolynes, P.G.: Stochastic gene expression as a many-body problem. Proc. Natl. Acad. Sci. USA **100**(5), 2374 (2003).
25. Korobkova, E., Emonet, T., Vilar, J.M.G., Shimizu, T.S., Cluzel, P.: From molecular noise to behavioural variability in a single bacterium. Nature **428**, 574 (2004).
26. Walczak, A.M., Onuchic, J.N., Wolynes, P.G.: Absolute rate theories of epigenetic stability. Proc. Natl. Acad. Sci. USA **102**, 18926 (2005).
27. Weinberger, L.S., Burnett, J.C., Toettcher, J.E., Arkin, A.P., Schaffer, D.V.: Stochastic gene expression in a lentiviral positive-feedback loop: Hiv-1 tat fluctuation drive phenotypic diversity. Cell **122**, 169 (2005).
28. Thattai, M., van Oudenaarden, A.: Stochastic gene expressions in fluctuating environments. Genetics **167**, 523 (2004).
29. Doi, M.: Second quantization representation for classical many-particle system. J. Phys. A **9**(9), 1465 (1976).
30. Lan, Y., Wolynes, P.G., Papoian, G.A.: A variational approach to the stochastic aspects of cellular signal transduction. J. Chem. Phys. **125**, 124101 (2006).
31. Kuramoto, Y.: Chemical Oscillations, Waves and Turbulence. Springer, New York (1984).
32. Fange, D., Berg, O.G., Sjöberg, P., Elf, J.: Stochastic reaction-diffusion kinetics in the microscopic limit. Proc. Natl. Acad. Sci. USA **107**(46), 19820 (2010).
33. Tanaka, N., Papoian, G.A.: Reverse-engineering of biochemical reaction networks from spatiotemporal correlations of fluorescence fluctuations. J. Theor. Biol. **264**(2), 490 (2010).
34. Zhuravlev, P.I., Papoian, G.A.: Molecular noise of capping protein binding induces macroscopic instability in filopodial dynamics. Proc. Natl. Acad. Sci. USA **106**(28), 11570 (2009).
35. Hu, L., Papoian, G.A.: Mechano-chemical feedbacks regulate actin mesh growth in lamellipodial protrusions. Biophys. J. **98**(8), 1375 (2010).
36. Zhuravlev, P.I., Der, B.S., Papoian, G.A.: Design of active transport must be highly intricate: a possible role of myosin and ena/vasp for g-actin transport in filopodia. Biophys. J. **98**(8), 1439 (2010).
37. Peskin, C., Odell, G., Oster, G.: Cellular motions and thermal fluctuations. Biophys. J. **65**, 316 (1993).
38. Lin, L.C.L., Brown, F.L.H.: Brownian dynamics in fourier space: membrane simulations over long length and time scales. Phys. Rev. Lett. **93**(25), 256001 (2004).
39. Pécréaux, J., Döbereiner, H.G., Prost, J., Joanny, J.F., Bassereau, P.: Refined contour analysis of giant unilamellar vesicles. Eur. Phys. J. E. **13**(3), 277 (2004).DOI 10.1140/epje/i2004–10001-9.
40. Gov, N.S., Safran, S.A.: Red blood cell membrane fluctuations and shape controlled by atp-induced cytoskeletal defects. Biophys. J. **88**(3), 1859 (2005).DOI 10.1529/biophysj.104.045328.
41. Safran, S., Gov, N., Nicolas, A., Schwarz, U., Tlusty, T.: Physics of cell elasticity, shape and adhesion. Physica A **352**(1), 171 (2005).
42. Schaus, T.E., Borisy, G.G.: Performance of a population of independent filaments in lamellipodial protrusion. Biophys. J. **95**, 1393 (2008).
43. Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. **81**, 2340 (1977).
44. Bugyi, B., Carlier, M.F.: Control of actin filament treadmilling in cell motility. Annu. Rev. Biophys. **39**, 449 (2010).

45. DiDonna, B.A., Levine, A.J.: Unfolding cross-linkers as rheology regulators in f-actin networks. Phys. Rev. E Stat. Nonlin. Soft. Matter Phys. **75**(4 Pt 1), 041909 (2007).
46. Mellor, H.: The role of formins in filopodia formation. Biochim. Biophys. Acta. **1803**(2), 191 (2010).
47. Pronk, S., Geissler, P.L., Fletcher, D.A.: Limits of filopodium stability. Phys. Rev. Lett. **100**(25), 258102 (2008).
48. Chan, C.E., Odde, D.J.: Traction dynamics of filopodia on compliant substrates. Science **322**(5908), 1687 (2008).
49. Gomez, T.M., Robles, E., Poo, M., Spitzer, N.C.: Filopodial calcium transients promote substrate-dependent growth cone turning. Science **291**(5510), 1983 (2001).
50. Miller, J., Fraser, S.E., McClay, D.: Dynamics of thin filopodia during sea urchin gastrulation. Development **121**(8), 2501 (1995).
51. Portera-Cailliau, C., Pan, D.T., Yuste, R.: Activity-regulated dynamic behavior of early dendritic protrusions: evidence for different types of dendritic filopodia. J. Neurosci. **23**(18), 7129 (2003).
52. Varnum-Finney, B., Reichardt, L.F.: Vinculin-deficient pc12 cell lines extend unstable lamellipodia and filopodia and have a reduced rate of neurite outgrowth. J. Cell Biol. **127**(4), 1071 (1994).
53. Kolomeisky, A.B., Fisher, M.E.: Molecular motors: a theorist's perspective. Ann. Rev. Phys. Chem. **58**, 675 (2007).
54. Howard, J.: Mechanics of Motor Proteins and the Cytoskeleton. Sinauer Associates, Sunderland (2001).
55. Berg, J.S., Cheney, R.E.: Myosin-x is an unconventional myosin that undergoes intrafilopodial motility. Nat. Cell Biol. **4**(3), 246 (2002).
56. Tokuo, H., Ikebe, M.: Myosin x transports mena/vasp to the tip of filopodia. Biochem. Biophys. Res. Commun. **319**(1), 214 (2004).
57. Kratky, O., Porod, G.: Röntgenuntersuchung gelöster fadenmoleküle. Rec. Trav. Chim. Pays-Bas. **68**, 1106 (1949).
58. Daniels, D.R.: Effect of capping protein on a growing filopodium. Biophys. J. **98**(7), 1139 (2010).
59. Bathe, M., Heussinger, C., Claessens, M.M.A.E., Bausch, A.R., Frey, E.: Cytoskeletal bundle mechanics. Biophys. J. **94**(8), 2955 (2008).
60. Atilgan, E., Wirtz, D., Sun, S.X.: Mechanics and dynamics of actin-driven thin membrane protrusions. Biophys. J. **90**(1), 65 (2006).
61. Liu, A.P., Richmond, D.L., Maibaum, L., Pronk, S., Geissler, P.L., Fletcher, D.A., Membrane-induced bundling of actin filaments. Nat. Phys. **4**(10), 789 (2008).
62. Prost, J., Joanny, J.F., Lenz, P., Sykes, C.: Cell Motility, Biological and Medical Physics, Biomedical Engineering, pp. 1–30. Springer, New York (2008).
63. Lee, K.C., Gopinathan, A., Schwarz, J.M.: Modeling the formation of in vitro filopodia. J. Math. Biol. 63, 229–261 (2011).
64. Lauffenburger, D.A., Horwitz, A.F.: Cell migration: a physically integrated molecular process. Cell **84**(3), 359 (1996).
65. Svitkina, T., Borisy, G.G.: Arp2/3 complex and actin depolymerizing factor/cofilin in dendritic organization and treadmilling of actin filament array in lamellipodia. J. Cell. Biol. **145**, 1009 (1999).
66. Mogilner, A., Edelstein-Keshet, L.: Regulation of actin dynamics in rapidly moving cells: a quantitative analysis. Biophys. J. **83**(3), 1237 (2002).
67. Carlsson, A.: Growth of branched actin networks against obstacles. Biophys. J. **81**, 1907 (2001).
68. Carlsson, A.E.: Growth velocities of branched actin networks. Biophys. J. **84**, 2907 (2003).
69. Rubinstein, B., Jacobson, K., Mogilner, A.: Multiscale two-dimensional modeling of a motile simple-shaped cell. Multiscale Model. Simul. **3**, 413 (2005).
70. Atilgan, E., Wirtz, D., Sun, S.X.: Morphology of lamellipodium and organization of actin filaments at the leading edge of crawling cells. Biophys. J. **89**, 3589 (2005).

71. Gov, N.S., Gopinathan, A.: Dynamics of membranes driven by actin polymerization. Biophys. J. **90**(2), 454 (2006).
72. Veksler, A., Gov, N.S.: Phase transitions of the coupled membrane-cytoskeleton modify cellular shape. Biophys. J. **93**, 3798 (2007).
73. Haviv, L., Brill-Karniely, Y., Mahaffy, R., Backouche, F., Ben-Shaul, A., Pollard, T.D., Bernheim-Groswasser, A.: Reconstitution of the transition from lamellipodium to filopodium in a membrane-free system. Proc. Natl. Acad. Sci. USA **103**, 4906 (2006).
74. Maree, A., Jilkine, A., Dawes, A., Grieneisen, V.A., Edelstein-Keshet, L.: Polarization and movement of keratocytes: a multiscale modelling approach. Bull. Math. Biol. **68**, 1169 (2006).
75. Huber, F., Kas, J., Stuhrmann, B.: Growing actin networks form lamellipodium and lamellum by self-assembly. Biophys. J. **95**, 5508 (2008).
76. Ideses, Y., Brill-Karniely, Y., Haviv, L., Ben-shaul, A., Bernheim-Groswasser, A.: Arp2/3 branched actin network mediates filopodia-like bundles formation in vitro. PLoS One. **3**, e3297 (2008).
77. Lacayo, C.I., Pincus, Z., Vanduijn, M.M., Wilson, C.A., Fletcher, D.A., Gertler, F.B., Mogilner, A., Theriot, J.A.: Emergence of large-scale cell morphology and movement from local actin filament growth dynamics. PLoS Biol. **5**, 2035 (2007).
78. Lee, K.C., Liu, A.J.: New proposed mechanism of actin-polymerization-driven motility. Biophys. J. **95**, 4529 (2008).
79. Ditlev, J.A., Vacanti, N.M., Novak, I.L., Loew, L.M.: An open model of actin dendritic nucleation. Biophys. J. **96**, 3529 (2009).
80. Hu, L., Papoian, G.A.: How does the antagonism between capping and anti-capping proteins control actin network dynamics? J. Phys. Condens. Matter **23**, 374101 (2011).
81. Pollard, T.D., Berro, J.: Mathematical models and simulations of cellular processes based on actin filaments. J. Biol. Chem. **284**(9), 5433 (2009).
82. Mogilner, A.: Mathematics of cell motility: have we got its number? J. Math. Biol. **58**, 105 (2009).
83. Theriot, J.A., Mitchison, T.J.: Actin microfilament dynamics in locomoting cells. Nature **352**, 126 (1991).
84. Karen, K., Pincus, Z., Allen, G.M., Barnhart, E.L., Marriott, G., Mogilner, A., Theriot, J.A.: Mechanism of shape determination in motile cells. Nature **453**, 475 (2008).
85. Pantaloni, D., Clainche, C.L., Carlier, M.F.: Mechanism of actin-based motility. Science **292**(5521), 1502 (2001).
86. Beltzner, C.C., Pollard, T.D.: Pathway of actin filament branch formation by arp2/3 complex. J. Biol. Chem. **283**, 7135 (2008).
87. Bear, J.E., Svitkina, T.M., Krause, M., Schafer, D.A., Loureiro, J.L., Strasser, G.A., Maly, I.V., Chaga, O.Y., Cooper, J.A., Borisy, G.G., Gertler, F.B.: Antagonism between ena/vasp proteins and actin filament capping regulates fibroblast motility. Cell **109**, 509 (2002).
88. Bear, J.E., Gertler, F.B.: Ena/vasp: towards resolving a pointed controversy at the barbed end. J. Cell. Sci. **122**, 1947 (2009).
89. Carlier, M.F., Pantaloni, D.: Control of actin dynamics in cell motility. J. Mol. Biol. **269**, 459 (1997).
90. Loisel, T.P., Boujemaa, R., Pantaloni, D., Carlier, M.F.: Reconstitution of actin-based motility of listeria and shigella using pure proteins. Nature **401**, 613 (1999).
91. Akin, O., Mullins, R.D.: Capping protein increases the rate of actin-based motility by promoting filament nucleation by the arp2/3 complex. Cell **133**, 841 (2008).
92. Borukhov, I., Bruinsma, R.F., Gelbart, W.M., Liu, A.J.: Structural polymorphism of the cytoskeleton: a model of linker-assisted filament aggregation. Proc. Natl. Acad. Sci. USA **102**, 3673 (2005).

Computational and Modeling Strategies for Cell Motility

Qi Wang, Xiaofeng Yang, David Adalsteinsson, Timothy C. Elston, Ken Jacobson, Maryna Kapustina, and M. Gregory Forest

1 Introduction

A predictive simulation of the dynamics of a living cell remains a fundamental modeling and computational challenge. The challenge does not even make sense unless one specifies the level of detail and the phenomena of interest, whether the focus is on near-equilibrium or strongly nonequilibrium behavior, and on localized,

Q. Wang (✉) • X. Yang
Department of Mathematics and NanoCenter, University of South Carolina, Columbia, SC 29208, USA
e-mail: qwang@math.sc.edu; xfyang@math.sc.edu

D. Adalsteinsson
Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: david@amath.unc.edu

T.C. Elston
Department of Pharmacology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: telston@med.unc.edu

K. Jacobson
Department of Cell and Developmental Biology and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: frap@med.unc.edu

M. Kapustina
Department of Cell and Developmental Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: mkapust@med.unc.edu

M.G. Forest
Department of Mathematics and Institute for Advanced Materials, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: forest@unc.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_11,

subcellular, or global cell behavior. Therefore, choices have to be made clear at the outset, ranging from distinguishing between prokaryotic and eukaryotic cells, specificity within each of these types, whether the cell is "normal," whether one wants to model mitosis, blebs, migration, division, deformation due to confined flow as with red blood cells, and the level of microscopic detail for any of these processes. The review article by Hoffman and Crocker [48] is both an excellent overview of cell mechanics and an inspiration for our approach. One might be interested, for example, in duplicating the intricate experimental details reported in [43]: "actin polymerization periodically builds a mechanical link, the lamellipodium, connecting myosin motors with the initiation of adhesion sites, suggesting that the major functions driving motility are coordinated by a biomechanical process," or to duplicate experimental evidence of traveling waves in cells recovering from actin depolymerization [35, 42]. Modeling studies of lamellipodial structure, protrusion, and retraction behavior range from early mechanistic models [84] to more recent deterministic [97, 112] and stochastic [51] approaches with significant biochemical and structural detail. Recent microscopic–macroscopic models and algorithms for cell blebbing have been developed by Young and Mitran [116], which update cytoskeletal microstructure via statistical sampling techniques together with fluid variables. Alternatively, whole cell compartment models (without spatial details) of oscillations in spreading cells have been proposed [35, 92, 109] which show positive and negative feedback mechanisms between kinetics and mechanics, and which are sufficient to describe a modality of sustained cell oscillations. The generalization of such a nonlinear limit cycle mechanism to include 3D spatial substructures consistent with cell mechanics, and biochemical kinetics and diffusion, charts a path that our group has elected. Detailed microscopic features are resolved through effective or collective properties of each substructure, which are dynamically updated by chemical species and processes. This choice is guided by a series of developments in the biophysics community on cell structure and rheology (cf. New Journal of Physics, Vol. 9, 2007), together with recent progress on the biochemical feedback mechanisms associated with cell morphological oscillations [20, 35, 58] as well as other dynamic cell modes.

Our approach is likewise guided by multiphase (implying differentiated substructures) modeling and computational tools developed for analogous applications such as biofilms [68, 108, 119, 121] and complex fluid mixtures (polymer dispersed nematic rods [37, 66], liquid crystal drops in viscous fluids [36, 115]). We integrate these approaches to propose a multiphase cell model with an energy-based phase field formulation, which we then simulate to illustrate qualitative phenomena that are possible with such a model. We conclude the chapter with a summary of experimental information and model advances that will be necessary to make the model biologically relevant and applicable to experiments. Our goal is a modeling and numerical framework which captures sufficient biological structure acceptable to cell biologists, which relies upon experimental data to parametrize the model equations for the structure, and which can reproduce single cell dynamic morphology behavior including blebbing, migration, contractile waves, oscillations, membrane-cortex rupture, and division. An early two-phase model of cell motion is developed by Alt and Dembo [2].

We model the cell as a composite of multiple phases or substructures, where each phase has its own material properties and constitutive relations that must be experimentally determined (cf. [79]). In the phase field formalism, the boundary between adjacent phases is diffuse rather than sharp; a phase field variable is introduced to model the thin transition layer, and an energy functional prescribes the momentum and energy exchange in the diffuse interface domain rather than traditional sharp interface elements such as surface tension and normal stress jumps. The cell phases include a bilayer membrane, a nucleus, and the cytoplasm which contains various protein filaments, other organelles, and aqueous cytosol [13]. Permeating the cytosol is a network of protein filaments of varying size and rigidity called the cytoskeleton [41,78,95]. The cytoskeleton not only provides the cell with mechanical integrity, but also provides a pathway for chemical and mechanical transport. Eukaryotic cells contain three main types of cytoskeletal filaments: actin filaments (microfilaments), intermediate filaments, and microtubules [41, 78]. Cytoskeletal elements interact extensively with cellular membranes and extracellular materials through functional and regulatory molecules or molecule complexes to affect cell motion [6, 61, 85, 87–89]. A distinguished phase, the cortical layer, lies between the bilayer membrane and the interior cytosol, and plays a prominent role in our model. Activation and deactivation in the cortical layer, triggered by specific protein families, are fundamental to our model. The phase field formulation allows for dramatic changes in each substructure, such as rupture of the bilayer membrane or cortical layer, separation of the membrane from the cortical layer by influx of cytosol, or even cell division. A long-term goal is to have sufficient biophysical and biochemical resolution to describe any cell morphological dynamic process.

A motile cell can crawl or migrate, especially on a supportable substrate, by protruding its front and retracting its rear [24, 49, 50, 60, 86, 90, 101, 102]. Cell motility is a result of orchestrated dynamical reconstruction and destruction of cytoskeletal structure coupled with cell membrane deformation. This reconstruction process is triggered by cell–substrate interactions through extracellular signalling and intracellular responses. The process of cell protrusion, the prelude of cell motion, is based on the polymerization of G-actin into F-actin filaments and force redistribution along other filament bundles like microtubules [88]. Actin polymerization is a directional or more precisely a polar phenomenon. During this process, the ATP (Adenosine-5'-triphosphate) bound G-actin is added to the barbed end of the existing F-actin filament, then ATP hydrolyzes into ADP; subsequently, the ADP bound actin drops off at the pointed end to depolymerize [13]. The local actin polymerization/depolymerization dynamics are regulated by the local concentration of functioning proteins, in particular, ATP-bound G-actin, ADP-bound G-actin, various accessory proteins, and binding subunits such as WASP proteins, Arp2/3 complexes, ADF/cofilin, profilin, thymosin $\beta 4$, $\alpha-$actinin, etc. [89]. The accessory proteins and binding subunits can inhibit or promote the polymerization/depolymerization process and thereby regulate the cell motility. In our model, we cannot retain full biochemical resolution and dynamics initially comparable to biochemical network models (cf. [1] and references therein), so simplifying choices will be made focusing on the key activation and deactivation species that are implicated in experiments.

In the case of cell migration on a substrate, the dynamic assembly and disassembly of focal adhesions plays a central role [12, 28, 90]. Focal adhesions are specific types of large macromolecular assemblies through which both mechanical and regulatory signals are transmitted. They serve as the mechanical linkages to the extracellular matrix (ECM) and as a biochemical signaling hub to concentrate and direct various signaling proteins at sites of integrin binding and clustering. On the other hand, surface or substrate topography has long been recognized to strongly influence cell adhesion, shape, and motion. Patterning and aligning scaffolds at the micro- and nano-scale with topographical features (indentations or grooves) as well as ligand organization have been reported to influence cell responses, such as adhesion, shape deformation (oriented cell elongation), migration, and growth [22]. The phenomenon of surface topography influencing cell migration is known as "contact cue guidance." [54, 73].

The underpinning issue in the contact cue guidance of motile cells is cell motility via cell–substrate interaction. Theoretical and computational modeling of cell motility continues to evolve in a variety of directions and for diverse purposes. However, given the complexity in cell motility, a whole cell model is still in an immature stage. Significant advances are more focused, such as on local cytoskeletal and actin dynamics [10,45,80,83], chemotaxis [82], membrane shape conformation [29], and simple cell models with idealized microstructural details of the cytoplasm [2,25,26,62,98,99,113]. In studying how actin filaments interact with the membrane locally, there have been a host of interesting local cytoskeletal dynamical models developed [3, 6, 14, 45, 80].

In addition to the local dynamical models for cytoskeletal and membrane dynamics, models have been developed to study cell migration on substrates. One model was devised to study effects of adhesion and mechanics on cell migration incorporating cytoskeletal force generation, cell polarization, and dynamic adhesion for persistent cell movement [26]. In this model, a coarse-grained viscoelastic model was used to describe mechanics of the cell body. Stephanou et al. [98] proposed a whole cell model for the dynamics of large membrane deformations of isolated fibroblasts, in which the cell protrusion was treated as the consequence of the coupling between F-actin polymerization and contractibility of the cortical actomyosin network. A model for the contractility of the cytoskeleton including the effect of stress fiber formation and disassociation in cell motion was developed by Deshpande et al. to investigate the role of stress fibers in the reorganization of the cytoskeleton [25]. Models treating the cytoplasm as active gels were proposed to study cell movement and drug delivery by Wolgemuth et al. [113]. Two-phase fluid models have also been used to study cell motion, in which the motion of the membrane and the local forces due to actin polymerization and membrane proteins are coupled through conservation laws and boundary conditions [2]. The coupling of biochemistry and mechanics in cell adhesion was recently studied by a new model for inhomogeneous stress fiber contraction [10]. A computational cell model for migration coupling the growth of focal adhesions with oscillatory cell protrusion is developed to show more numerical detail in the migration process [99]. A new continuum modeling approach to study viscoelastic cytoskeletal networks

is proposed to model the cytoplasm as a bulk viscoelastic material [62]. Each of these models, and others below, represents a step toward a multipurpose whole cell dynamics model.

Active polar gel models have emerged as a new and exciting topic in soft matter and complex fluids [4,8,65,71]. In an active material system, energy is continuously supplied by internal as well as external sources to drive the movement of the material system. In a living cell, cross-linking proteins bind two or more self-assembled filaments (e.g., F-actin or F-actin and microtubules) to form a dynamical gel, in which motor proteins bind to filaments and hydrolyze nucleotide ATP. This process coupled to a corresponding conformational change of the binding protein turns stored energy into mechanical work, thereby leading to relative motion between bound filaments [72]. Self-propelled gliding motion of certain bacterial species is another example of such an active material system, where molecular motors drive the cellular motion in a matrix of another material [8]. Both continuum mechanical models and kinetic theories have been proposed for active complex fluid systems [4, 8, 21, 63, 65, 71, 91]. The mathematical framework incorporates the source of "active forcing" into an otherwise passive material system. The models are based on free energy considerations, both equilibrium and nonequilibrium, where one can keep track of dissipative and conservative principles, and the challenge for biological fidelity is to construct relevant energy potentials and chemical–mechanical activation functions. These potentials require detailed viscous and elastic properties of the fundamental cell components or phases, for which experimental techniques are now advanced enough to make progress. The energy formulation is likewise compatible with mathematical modeling, numerical algorithms, and simulation tools that have been developed for the hydrodynamics of multiphase complex fluids in evolving spatial domains. The simultaneous modeling of reaction and diffusion of biochemical species is self-consistent with the energetic formulation. These advances lay the groundwork for our approach.

Given the collective advances in membrane and cytoskeletal modeling, cell–substrate coupling, and biochemical kinetics, it is now feasible to develop a whole cell model for migration on substrates. This global cell–substrate model will enable us to investigate cell motility, dynamics of signaling proteins, cytoskeleton–substrate coupling, and contact cue guidance of motile cells. The model predictions will provide qualitative comparisons with cell experiments in the first proof-of-principle stage, and potentially guide future experiments on detailed mechanisms associated with motility. As properties of each substructure become more quantified, the model will be able to make predictions to guide cell motility experiments. Given the complex nature of cell migration on topographically designed substrates, we must adopt a theoretical and computational platform that is applicable to a variety of dynamical modalities. Among the competing mathematical models for multiphase soft matter phenomena, the field phase approach is sufficiently versatile to handle the complexity of this challenge, and to sequentially incorporate additional biological complexity. We take up this topic next.

Phase field models have been used successfully to study a variety of interfacial phenomena like equilibrium shapes of vesicle membranes [29–34,105,120], dynamics of two-phase vesicles [39, 40], blends of polymeric liquids [38, 103, 106, 107],

multiphase flows [16, 36, 52, 53, 69, 70, 74, 110, 111, 114, 115, 117–119], dentritic growth in solidification, microstructure evolution [47, 59, 77], grain growth [17], crack propagation [18], morphological pattern formation in thin films and on surfaces [67, 93], self-assembly dynamics of two-phase monolayer on an elastic substrate [75], a wide variety of diffusive and diffusionless solid-state phase transitions [18, 19, 104], dislocation modeling in microstructure, electromigration, and multiscale modeling [100]. Phase field methods can also describe multiphase materials [39, 40, 110]. Recently, phase field models are applied to study liquid crystal drop deformation in another fluid and liquid films by our group and other groups [36, 52, 69, 70, 74, 110, 111, 114, 115, 117–119]. We will now apply the phase field modeling formalism, treating the substructures of the cell as well as its surrounding environment as distinct complex fluids, including an ambient fluid or solid substrate or another cell(s). Distinct phases are differentiated by phase variables. As a result, the entire material system can be modeled effectively as a multiphase complex fluid in contact with a substrate [29]; the cell membrane is modeled naturally as a phase boundary between the cortical layer and the ambient fluid or substrate. Additional phase variables can be introduced to account for the various complex fluid components (cortical layer, cytosol, nucleus) confined inside the cell membrane; these phase variables can serve as volume fractions for each of the cytoplasm components. The phase field formulation allows the dynamical model developed for each phase of the mixture to be integrated to form the global cell model.

We review an incremental set of models for active fluids of self-propelled microconstituents and active gels, respectively. We will then propose a whole cell model as a framework for proof-of-principle simulations and future development.

2 Models for Active Filaments

In a seminal paper by Simha and Ramaswamy [96], an active stress mechanism for diverse model systems including bacteria, molecular motors, F-actin treadmilling polymerization, and depolymerization mechanisms is formulated. Two fundamental mechanisms are distinguished that lead to macroscopic motion, both of which are tied to the existence of a pair of permanent force dipoles of the moving object. One corresponds to contractile motion, called a puller mechanism by analogy with a breast stroke of a swimmer, and the other is due to a tensile motion on the object, called a pusher by analogy with the kick of a swimmer [44, 91, 96]. The fluid flow field around the moving object in these two different situations exhibits distinct flow patterns, both of which propel at the particle scale. The stress associated with this motion is called the active stress. Since this is the essential part of the theories for active filament material systems, we will give a brief overview of the derivation.

An ensemble of moving objects, including rod macromolecules, bacteria, F-actin filaments, etc., are considered. An object has its center of mass located at $\mathbf{r}_i$ and two permanent force dipoles localized at $\mathbf{r}_i + b\mathbf{n}_i$ and $\mathbf{r}_i - b'\mathbf{n}_i$, respectively, where

$\mathbf{n}_i$ is a unit vector associated with the displacement direction of the ith object. If $b = b'$, the object is called apolar; otherwise, polar. The collective force exerted by the ensemble at location $\mathbf{r}$ is given by

$$\mathbf{f}^{(a)} = f \sum_i \mathbf{n}_i \left[\delta \left(\mathbf{r} - \mathbf{r}_i(t) - b\mathbf{n}_i(t) \right) - \delta \left(\mathbf{r} - \mathbf{r}_i(t) + b'\mathbf{n}_i(t) \right) \right], \tag{1}$$

where f is the magnitude of the force dipole. We expand the $\delta-$ function formally, and the force can be rewritten as:

$$\begin{aligned} \mathbf{f}^{(a)} &= (b + b') f \nabla \cdot \left(\sum_i \mathbf{n}_i \mathbf{n}_i \delta(\mathbf{r} - \mathbf{r}_i) \right) \\ &\quad - \frac{(b + b')(b - b')}{2} f \nabla\nabla : \left(\sum_i \mathbf{n}_i \mathbf{n}_i \mathbf{n}_i \delta(\mathbf{r} - \mathbf{r}_i) \right) + \cdots . \end{aligned} \tag{2}$$

From this force formula, the active stress tensor is deduced,

$$\begin{aligned} \tau^{(a)} &= (b + b') f \sum_i \mathbf{n}_i \mathbf{n}_i \delta(\mathbf{r} - \mathbf{r}_i) \\ &\quad - \frac{(b + b')(b - b')}{2} f \nabla \cdot \left(\sum_i \mathbf{n}_i \mathbf{n}_i \mathbf{n}_i \delta(\mathbf{r} - \mathbf{r}_i) \right) + \cdots . \end{aligned} \tag{3}$$

At leading order, the active stress tensor is given by

$$\tau^{(a)} = \alpha \sum_i \mathbf{n}_i \mathbf{n}_i \delta(\mathbf{r} - \mathbf{r}_i), \tag{4}$$

where $\alpha = (b + b') f$. Positive values correspond to pullers and negative values correspond to pushers.

In the case of ATP-driven polymerization and depolymerization, the active stress is given in the same form, where α is proportional to the energy difference of the chemical potentials of ATP and the product molecules ADT and P_i. This latter expression defines the active stress at leading order in all active filament models discussed below.

2.1 Active Polar Filament Model

We consider a suspension of active polar filaments in a viscous solvent. The active polar filament model of Muhuri et al. [81] uses the concentration of the active polar suspensions c and the polarity vector of the filament particle $\mathbf{p}$, in which a background fluid velocity $\mathbf{v}$ is introduced. The governing system of equations in

this model is summarized below. In this model, the polarity vector is assumed to represent the velocity of the active particle; the background velocity is assumed solenoidal, and inertia is neglected. Without external forces, the governing system of equations consists of:

$$
\begin{aligned}
&\nabla \cdot \mathbf{v} = 0,\\
&\nabla \cdot \sigma = 0, \sigma = \sigma^{\mathrm{a}} + \sigma^{\mathrm{r}} + \sigma^{\mathrm{d}},\\
&\sigma^{\mathrm{d}} = \eta \left(\nabla \mathbf{v} + \nabla \mathbf{v}^{T}\right),\\
&\sigma^{\mathrm{r}} = -\frac{\lambda}{2}(\mathbf{ph} + \mathbf{hp}) + \Pi \mathbf{I},\\
&\sigma^{\mathrm{a}} = Wc(\mathbf{x}, t)\left(\mathbf{pp} - \|\mathbf{p}\|^{2}\frac{I}{3}\right),\\
&\frac{\partial \mathbf{p}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{p} - \frac{1}{2}(\nabla \times \mathbf{v}) \times \mathbf{p} + \left[\lambda_1(\mathbf{p} \cdot \nabla)\mathbf{p} + \lambda_2(\nabla \cdot \mathbf{p})\mathbf{p} + \lambda_3 \nabla \|\mathbf{p}\|^{2}\right]\\
&\quad = \frac{\lambda}{2}\left(\nabla \mathbf{v} + \nabla \mathbf{v}^{T}\right) \cdot \mathbf{p} - \zeta \nabla c + \Gamma \mathbf{h},\\
&\frac{\partial c}{\partial t} + \nabla \cdot (c(\mathbf{v} + \mathbf{p})) = 0,
\end{aligned}
\tag{5}
$$

where $\lambda_{1,2,3}, \lambda, \Pi, W, \zeta, \Gamma$ are model parameters. The sign of W determines the nature of the elementary force dipoles. Here, $\mathbf{h}$ is the molecular field for the polar vector $\mathbf{p}$ and is given by

$$h = c\left[\alpha \mathbf{p} - \beta \|\mathbf{p}\|^{2}\mathbf{p} + K\nabla^{2}\mathbf{p}\right], \tag{6}$$

where α and β are model parameters, and K is the analog of the Frank elastic constant of the Ericksen–Leslie theory for liquid crystals in the one-constant approximation [23]. The stress tensors $\sigma^{\mathrm{d}}, \sigma^{\mathrm{r}}, \sigma^{\mathrm{a}}$ are the dissipative, reversible (or reactive) and active stress, respectively. The reversible stress is due to the response to the polar order gradient. The terms containing $\lambda_{1,2,3}$ and ζ are the symmetry-allowed polar contribution to the nematodynamics of $\mathbf{p}$. The corresponding free energy for the system is identified as

$$F = \int \frac{c}{2}\left[-\alpha \|\mathbf{p}\|^{2} + \beta \|\mathbf{p}\|^{4} + K\| \nabla \mathbf{p}\|^{2}\right] \mathrm{d}\mathbf{x}. \tag{7}$$

The molecular field is defined by $\mathbf{h} = -\frac{\delta F}{\delta \mathbf{p}}$. The moving polar particle velocity and the background fluid flow velocity are fully coupled. With this model, Muhuri et al. studied shear-induced isotropic to nematic phase transition of active filament suspensions as a model of reorientation of endothelial cells. This model neglects the impact of energy changes to migration of polar filaments.

An analogous theory using the same set of hydrodynamical variables was developed by Giomi et al. [44] which involves more sophisticated coupling between the concentration, background fluid flow, and the polarity vector of the polar particles. It extends the previous theory to account for the energetic influence to filament migration. The governing system of equations is summarized below. In this model, inertial effects are retained and attention was paid to the variational structure of the governing system of equations. For instance, the missing asymmetric contribution to reactive stress in the previous model is supplemented.

$$
\begin{aligned}
&\rho\left(\frac{\partial}{\partial t}+\mathbf{v}\cdot\nabla\right)\mathbf{v}=\nabla\cdot\sigma,\\
&\nabla\cdot\mathbf{v}=0,\\
&\sigma=\sigma^{\mathrm{d}}+\sigma^{\mathrm{r}}+\sigma^{\mathrm{a}}+\sigma^{\mathrm{b}},\\
&\sigma^{\mathrm{d}}=2\eta\mathbf{D},\\
&\sigma^{\mathrm{r}}=-\pi\mathbf{I}-\frac{\lambda}{2}(\mathbf{ph}+\mathbf{hp})+\frac{1}{2}(\mathbf{ph}-\mathbf{hp}),\\
&\sigma^{\mathrm{a}}=\frac{\alpha c^2}{T}(\mathbf{pp}+\mathbf{I}),\\
&\sigma^{\mathrm{b}}=\frac{\beta c^2}{T}\left(\nabla\mathbf{p}+\nabla\mathbf{p}^T+\nabla\cdot\mathbf{pI}\right),\\
&\frac{\partial c}{\partial t}+\nabla\cdot\left[c(\mathbf{v}+c\beta_1\mathbf{p})+\Gamma'\mathbf{h}+\Gamma''\mathbf{f}\right]=0,\\
&\left(\frac{\partial}{\partial t}+(\mathbf{v}+c\beta_2\mathbf{p})\cdot\nabla\right)\mathbf{p}+\Omega\cdot\mathbf{p}=\lambda Tr(\nabla\mathbf{v})\mathbf{p}+\Gamma\mathbf{h}+\Gamma'\mathbf{f},
\end{aligned}
\tag{8}
$$

where $\mathbf{h}$ is the molecular field given by $\mathbf{h}=-\frac{\delta F}{\delta p}$, $\mathbf{f}=-\nabla\frac{\delta F}{\delta c}$ is the molecular flux of the active rods, $\mathbf{D}=\frac{\nabla\mathbf{v}+\nabla\mathbf{v}^T}{2}$ is the rate of strain tensor, $\Omega=\frac{1}{2}(\nabla\mathbf{v}-\nabla\mathbf{v}^T)$ is the vorticity tensor, σ^{b} is a dissipative stress (an analogue of σ^{d}), and all the parameters unspecified are model parameters. The free energy of the system is given by

$$
\begin{aligned}
F=\int\Bigg[&\frac{C}{2}\left(\frac{\delta c}{c_0}\right)^2+\frac{a_2}{2}\|\mathbf{p}\|^2+\frac{a_4}{4}\|\mathbf{p}\|^4+\frac{K_1}{2}(\nabla\cdot\mathbf{p})^2+\frac{K_3}{2}(\|\nabla\times\mathbf{p}\|)^2\\
&+B_1\frac{\delta c}{c_0}\nabla\cdot\mathbf{p}+B_2\|\mathbf{p}\|^2\nabla\cdot\mathbf{p}+\frac{B_3}{c_0}\|\mathbf{p}\|^2\mathbf{p}\cdot\nabla c\Bigg]\mathrm{d}\mathbf{x},
\end{aligned}
\tag{9}
$$

where $\delta c=c-c_0$, c_0 is a baseline concentration, C is the compression modulus, and $K_{1,3}$ are the splay and bend elastic constants; the other coefficients depend on both passive and active contributions [44]. This model adds additional fluxes to the transport of the concentration c due to the energetic activity of both polar velocity field $\mathbf{p}$ and the concentration fluctuations of c. The convective effect of the polar velocity $\mathbf{p}$ is added to the transport of both c and $\mathbf{p}$ as well. An additional

"viscous" stress σ^{b} is added analogous to the viscous stress σ^{d}. The free energy contains additional coupling terms between the polar velocity and the concentration gradient.

This model is used to study sheared active polar fluids. An extremely rich variety of phenomena are identified including an effective reduction or increase in the apparent viscosity, depending on the nature of the active stresses and flow alignment property of the particles, nonmonotone stress-strain-rate relationship, and yield stress for large active forcing [44]. In the limit of strongly polarized states where the magnitude of $\mathbf{p}$ is locked, this formulation can be recast in terms of a unit vector $\mathbf{u} = \frac{\mathbf{p}}{\|\mathbf{p}\|}$. The details can be found in [44].

2.2 Active Apolar Filament Models

When the polarity on the moving objects is weak, instead of the polarity vector, a second order nematic tensor can be employed to describe both the nematic order as well as the active stress. For apolar filament fluids, a coarse-grained model can be derived with only the nematic order tensor [15,46]. We summarize the version used by Cates et al. [15] in this section. Let $\mathbf{Q}$ be a traceless second order tensor denoting the nematic order in the active filament fluid. The governing system of equations consist of the following equations.

$$
\begin{aligned}
&\nabla \cdot \mathbf{v} = 0,\\
&\rho\left(\frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla\right)\mathbf{v} = \nabla\cdot(\sigma),\\
&\mathbf{H} = -\frac{\delta F}{\delta \mathbf{Q}} + \frac{1}{3}Tr\left(\frac{\delta F}{\delta \mathbf{Q}}\right)\mathbf{I},\\
&\sigma = -P_0\mathbf{I} + 2\eta\mathbf{D} + 2\xi\left(\mathbf{Q}+\frac{\mathbf{I}}{3}\right)\mathbf{Q}:\mathbf{H} - \xi\mathbf{H}\left(\mathbf{Q}+\frac{\mathbf{I}}{3}\right) - \xi\left(\mathbf{Q}+\frac{1}{3}\mathbf{I}\right)\mathbf{H}\\
&\qquad -\nabla\mathbf{Q}:\frac{\delta F}{\delta\nabla\mathbf{Q}} + \mathbf{Q}\cdot\mathbf{H} - \mathbf{H}\cdot\mathbf{Q} - \zeta\mathbf{Q},\\
&\left(\frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla\right)\mathbf{Q} - \Omega\cdot\mathbf{Q} + \mathbf{Q}\cdot\Omega - [\mathbf{D}\cdot\mathbf{Q} + \mathbf{Q}\cdot\mathbf{D}] = \Gamma\mathbf{H},
\end{aligned}
\tag{10}
$$

where c is the concentration of the apolar active rod assumed constant in this model, ξ is the friction coefficient, P_0 is the hydrostatic pressure, ζ is the activity parameter with $\zeta > 0$ corresponding to extensile and $\zeta < 0$ contractile motion. The free energy density of the material system is given by a simplified Landau-deGennes functional

$$
F = k_{\mathrm{B}}Tc\left[\left(1-\frac{N}{3}\right)\frac{\mathbf{Q}:\mathbf{Q}}{2} - \frac{N}{3}\mathbf{Q}^3 + \frac{N}{4}(\mathbf{Q}:\mathbf{Q})^2 + \frac{K}{2}\left(\nabla\mathbf{Q}\vdots\nabla\mathbf{Q}\right)^2\right], \tag{11}
$$

where k_B is the Boltzmann constant, T is the absolute temperature, N measures the strength of the bulk part of the potential, and K is the one-constant Frank elastic coefficient. This model was used to study sheared active gels close to the isotropic–nematic transition. This model was later extended to add an active term in the nematodynamic equation for $\mathbf{Q}$ and simulated with lattice Boltzmann numerical methods [76].

2.3 Kinetic Models for Active Fluids

In an effort to unify the polar and apolar models for active fluids, Liverpool, Marchetti and collaborators developed a framework for active filament fluids using a polymer kinetic theory formulation [72]. Here, we briefly describe the 2-D formulation of the theory and its coarse-graining procedures to yield the moment equations. Let c be the number density of rigid active rods. The Smoluchowski equation is given by

$$\begin{aligned}
&\frac{\partial c}{\partial t} + \nabla \cdot \mathbf{J} + \mathcal{R} \cdot \mathbf{J}^{\mathrm{R}} = 0, \\
&\mathbf{J} = \mathbf{v}c - D \cdot \nabla c - \frac{1}{k_B T} D \cdot c \nabla V_{\mathrm{ex}} + \mathbf{J}^{\mathrm{a}}, \\
&\mathbf{J}^{\mathrm{R}} = c\boldsymbol{\omega} - D_{\mathrm{r}} \left[\mathcal{R}c - \frac{c}{k_B T} \mathcal{R} V_{\mathrm{ex}} \right] + \mathbf{J}_{\mathrm{r}}^{\mathrm{a}},
\end{aligned} \tag{12}$$

where $\mathbf{J}$ is the translational flux, $\mathbf{J}^{\mathrm{R}}$ is the rotational flux, $D = D_{||}\mathbf{uu} + D_{\perp}(\mathbf{I} - \mathbf{uu})$ is the translational diffusivity, $\mathbf{u}$ is the unit vector in the direction of the molecular velocity, D_{r} is the rotational diffusivity, ω is the angular velocity, $\mathcal{R} = \mathbf{u} \times \frac{\partial}{\partial \mathbf{u}}$ is the rotational gradient operator, the active translational and rotational fluxes are defined by

$$\begin{aligned}
\mathbf{J}^{\mathrm{a}} &= cb^2 m \int \int \mathbf{v}_{\mathrm{a}}(\mathbf{u}, s_1; \mathbf{u}_2, s_2) c(r + \xi, \mathbf{u}_2, t) \mathrm{d}\mathbf{u}_2 \mathrm{d}\xi, \\
\mathbf{J}_{\mathrm{r}}^{\mathrm{a}} &= cb^2 m \int \int \boldsymbol{\omega}_a(\mathbf{u}, s_1; \mathbf{u}_2, s_2) c(r + \xi, \mathbf{u}_2, t) \mathrm{d}\mathbf{u}_2 \mathrm{d}\xi,
\end{aligned} \tag{13}$$

$\mathbf{v}_{\mathrm{a}}$ and ω_{a} are the translational and rotational velocities, respectively. The excluded volume potential is given by the Onsager potential

$$V_{\mathrm{ex}} = k_B T \int \int \|\mathbf{u} \times \mathbf{u}_2\| c(\mathbf{x} + \mathbf{s}, \mathbf{u}_2, t) \mathrm{d}\mathbf{s}\mathrm{d}\mathbf{u}_2, \tag{14}$$

b is a parameter and m is the mass of the rod. The active velocities are given by

$$\begin{aligned}
\mathbf{v}_{\mathrm{a}} &= \frac{1}{2}\mathbf{v}_r + \mathbf{V}_m, \\
\mathbf{v}_r &= \frac{\tilde{\beta}}{2}(\mathbf{u}_2 - \mathbf{u}) + \frac{\tilde{\alpha}}{2l}\xi, \\
\mathbf{V}_m &= A(\mathbf{u} + \mathbf{u}_2) + B(\mathbf{u}_2 - \mathbf{u}), \\
\omega_{\mathrm{a}} &= 2\left[\gamma_{\mathrm{P}} + \gamma_{\mathrm{NP}}(\mathbf{u}\cdot\mathbf{u}_2)\right](\mathbf{u}\times\mathbf{u}_2),
\end{aligned} \tag{15}$$

where γ_{P} and γ_{NP} are the rotational rates, $\tilde{\alpha} = \alpha(1 + \mathbf{u}\cdot\mathbf{u}_2), \tilde{\beta} = \beta(1 + \mathbf{u}\cdot\mathbf{u}_2)$, $A = -(\beta - \alpha(s_1 - s_2)/2)/12$, and $B = \alpha(s_1 - s_2)/24$ for long thin rods [4]. α and β are model parameters. All four parameters can be related to the stiffness of the crosslinkers and to the rate $u(s)$ at which a motor cluster attached at position s steps along a filament toward the polar end [72]. The concentration has a generalized Fourier expansion

$$c\,(\mathbf{x}, \mathbf{u}, t) = \frac{c(\mathbf{x}, t)}{2\pi}\left[1 + 2\mathbf{p}\cdot\mathbf{u} + 4\mathbf{Q} : \mathbf{uu} + \cdots .\right], \tag{16}$$

where the zeroth, first, and second moments are defined by:

$$\begin{aligned}
c(\mathbf{x}, t) &= \int c(\mathbf{x}, \mathbf{u}, t)\mathrm{d}\mathbf{u}, \\
c\mathbf{p}(\mathbf{x}, t) &= \int \mathbf{u}c(\mathbf{x}, \mathbf{u}, t)\mathrm{d}\mathbf{u}, \\
c\mathbf{Q}(\mathbf{x}, t) &= \int \left(\mathbf{uu} - \frac{\|\mathbf{uu}\|^2}{2}\mathbf{I}\right) c(\mathbf{x}, \mathbf{u}, t)\mathrm{d}\mathbf{u}.
\end{aligned} \tag{17}$$

The force due to the stress for the system is given by

$$\nabla\cdot\sigma = \int\int c(\mathbf{x} - \xi, \mathbf{u}, t)\left\langle\delta(\xi - s\mathbf{u})\mathcal{F}^{\mathrm{h}}(s)\right\rangle_s \mathrm{d}\xi\mathrm{d}\mathbf{u}, \tag{18}$$

where $\mathcal{F}^{\mathrm{h}}(s)$ is the hydrodynamic force per unit length exerted by the suspension at position s along the rod. The force can be approximated by the first two terms in the Taylor expansion

$$\nabla\cdot\sigma = \int c(\mathbf{x}, \mathbf{u}, t)\mathbf{F}^h\mathrm{d}\mathbf{u} - \int\left\langle\left(\frac{s}{l}\right)^2\left(\frac{\mathbf{u}\nabla}{l}\right)c\tau^h\right\rangle_s \mathrm{d}\mathbf{u}. \tag{19}$$

We denote $\tilde{\sigma} = \sigma - Tr(\sigma)\frac{\mathbf{I}}{3}$.

$$\begin{aligned}
\tilde{\sigma} &= \tilde{\sigma}^{\mathrm{a}} + \sigma^{\mathrm{d}},\\
\tilde{\sigma}^{\mathrm{a}} &= 2k_{\mathrm{B}}T_a c\left[\left(1-\frac{c}{c_I N}\right)\mathbf{Q} - \frac{c}{c_I N}\left(\mathbf{pp} - \frac{\|\mathbf{p}\|^2}{2}\mathbf{I}\right) + C_1 c\left(\frac{4}{3}\mathbf{Q} + \mathbf{pp} - \frac{\|\mathbf{p}\|^2}{2}\mathbf{I}\right)\right]\\
&\quad + C_2 c\left[\nabla\mathbf{p} - \frac{\nabla\cdot\mathbf{p}}{2}\mathbf{I} - \frac{1}{4}\left(\nabla\mathbf{p} - \nabla\mathbf{p}^T\right)\right],\\
\tilde{\sigma}^{\mathrm{d}} &= C_3\left[\frac{1}{2}\left(\mathbf{D} - \frac{\mathbf{I}}{2}Tr(\mathbf{D})\right) + \frac{1}{3}\left(\mathbf{Q}Tr(\mathbf{D}) - (\mathbf{D}:\mathbf{Q})\mathbf{I}\right) + \frac{1}{3}\left(\mathbf{D}\cdot\mathbf{Q} + \mathbf{Q}\cdot\mathbf{D}\right)\right], \quad (20)
\end{aligned}$$

where C_1, C_2, C_3 are model parameters [72].

This together with the continuity equation for the average velocity $\mathbf{v}$ and the momentum balance equation in the form of Stokes equation constitute the governing system of equations for the kinetic theory.

$$\begin{aligned}
&\nabla\cdot\mathbf{v} = 0,\\
&\nabla\cdot(\sigma + 2\eta\mathbf{D} - P_0\mathbf{I}) = 0, \quad (21)
\end{aligned}$$

where η is the viscosity of the solvent and P_0 is the hydrostatic pressure.

Shelley and Santillian studied dilute active rod particle fluids using a kinetic theory in which only convective transport is accounted for [91];

$$\mathbf{J} = \mathbf{v}c + v_0\mathbf{u}c, \mathbf{J}^{\mathrm{R}} = c\boldsymbol{\omega}. \quad (22)$$

In their model, the active stress tensor is given by

$$\tilde{\sigma}^{\mathrm{a}} = \zeta\mathbf{Q}. \quad (23)$$

Next, we present one of the latest versions of the kinetic theory in which the active flux due to rod–rod binary collisions is carefully considered. Baskaran and Marchetti derived a Smoluchowski equation for self-propelled hard rods in 2-D [9].

$$\begin{aligned}
&\frac{\partial c}{\partial t} + \nabla\cdot\mathbf{J} + \mathcal{R}\cdot\mathbf{J}^{\mathrm{R}} = 0,\\
&\mathbf{J} = c\mathbf{v} + v_0\mathbf{u}c - D^{\mathrm{SP}}\cdot\nabla c - \frac{1}{k_{\mathrm{B}}T}D\cdot c\nabla V_{\mathrm{ex}} - \frac{D_{||}mv_0^2}{2k_{\mathrm{B}}T}\mathbf{I}^{\mathrm{SP}},\\
&\mathbf{J}^{\mathrm{R}} = c\omega - D_{\mathrm{r}}\left[\mathcal{R}c + \frac{c}{k_{\mathrm{B}}T}\mathcal{R}V_{\mathrm{ex}}\right] - \frac{D_{\mathrm{r}}mv_0^2}{2k_{\mathrm{B}}T}\mathbf{I}_r^{\mathrm{SP}}, \quad (24)
\end{aligned}$$

where $\mathbf{J}$ is the translational flux and $\mathbf{J}^{\mathrm{R}}$ is the rotational flux,

$$D^{\mathrm{SP}} = D_{\perp}\mathbf{I} + (D_{||} + D_{\mathrm{S}} - D_{\perp})\mathbf{uu} \quad (25)$$

is the translational diffusivity, $D_S = \frac{v_0^2}{\zeta}$, v_0 is the speed of the moving rod, $D_{||}$ and $D_\perp$ are the diffusivity in the parallel and perpendicular direction of the rod, m is the mass of the rod, and the additional fluxes due to collisions are given below.

$$\begin{aligned}
\mathbf{I}^{\mathrm{SP}} &= \int\int \sin^2(\theta_1-\theta_2)\,[\Theta(\sin(\theta_1-\theta_2))-\Theta(-\sin(\theta_1-\theta_2))] \\
&\quad\times\left[\mathbf{u}_1^{\perp}c\left(\mathbf{x}_1+s\mathbf{u}_1-\frac{l}{2}\mathbf{u}_2,\mathbf{u}_2,t\right)+\mathbf{u}_2^{\perp}c\left(\mathbf{x}_1+s\mathbf{u}_2-\frac{l}{2}\mathbf{u}_1,\mathbf{u}_2,t\right)\right]\mathrm{d}s\mathrm{d}\mathbf{u}_2, \\
\mathbf{I}_r^{\mathrm{SP}} &= \mathbf{z}\int\int \sin^2(\theta_1-\theta_2)\,[\Theta(\sin(\theta_1-\theta_2))-\Theta(-\sin(\theta_1-\theta_2))] \\
&\quad\times\left[sc\left(\mathbf{x}_1+s\mathbf{u}_1-\frac{l}{2}\mathbf{u}_2,\mathbf{u}_2,t\right)+\frac{l}{2}\cos(\theta_1-\theta_2)\right. \\
&\qquad\left.\times\, c\left(\mathbf{x}_1+s\mathbf{u}_2-\frac{l}{2}\mathbf{u}_1,\mathbf{u}_2,t\right)\right]\mathrm{d}s\mathrm{d}\mathbf{u}_2,
\end{aligned} \tag{26}$$

where $\mathbf{z} = \mathbf{u}\times\mathbf{u}_2$, θ_1 and θ are the initial angles of $\mathbf{u}$ and $\mathbf{u}_2$, respectively, before collision. $\Theta(x)$ is the Heaviside function.

Taking the zeroth moment, the first moment, and the second moment of the Smoluchowski equation, the transport equation for the rod density, polarity vector, and the nematic order tensor can be derived [9].

These models are developed for dilute to semidilute suspensions of active filaments and rods in viscous solvents. Inside a cell, the cytoplasm is comprised of various cytoskeletal filaments, microtubules, and intermediate filaments immersed in the cytosol. The resulting network structures and buffer solution behave like a gel. We briefly review new models for active biogels next.

3 Models for Active Gels

In active gels, networks of active filaments can form either temporarily or on a longer timescale. The solvent permeation into the network must be accounted for in the gel. We next describe several relevant models for active gels briefly.

3.1 Isotropic Active Gel Model

Banerjee and Marchetti proposed a phenomenological model for isotropic active gels based on a continuum model for physical gels [7]. The governing system of equations are summarized. We denote by $\mathbf{u}$ the position vector of the network, $\mathbf{v}$ the

velocity of the solvent fluid, c_{b} the concentration of bound motor proteins, c_{u} the concentration of the unbound motors, ρ the mass density of the gel network, and ρ_{f} the density of the fluid. The model is based on the two-component formulation of multiphase fluids, in which network is treated as a viscoelastic material while the solvent is modeled as a viscous fluid. The momentum conservation for each phase is enforced and the mixture is assumed incompressible i.e., the combined velocity is assumed solenoidal. The interaction between the solvent and the network is through a friction term in the momentum balance equations for both materials.

$$
\begin{aligned}
&\rho \frac{\partial^2}{\partial t^2}\mathbf{u} = -\Gamma(\dot{\mathbf{u}} - \mathbf{v}) + \nabla \cdot \sigma,\\
&\sigma = \sigma^e + \sigma^{\mathrm{d}} + \sigma^{\mathrm{a}},\\
&\sigma^e = \mu\left(\nabla\mathbf{u} + \nabla\mathbf{u}^T\right) + \lambda Tr(\nabla\mathbf{u})\mathbf{I},\\
&\sigma^{\mathrm{d}} = \eta_{\mathrm{s}}\left(\nabla\dot{\mathbf{u}} + \nabla\dot{\mathbf{u}}^T\right) + \left(\eta_{\mathrm{b}} - \frac{2\eta_{\mathrm{s}}}{3}\right) Tr(\nabla\dot{\mathbf{u}})\mathbf{I},\\
&\sigma^{\mathrm{a}} = \zeta(\rho, c_{\mathrm{b}})\Delta\mu\mathbf{I},\\
&\rho_{\mathrm{f}}\dot{\mathbf{v}} = \nabla \cdot (-P\mathbf{I} + 2\eta\mathbf{D}) + \Gamma(\dot{\mathbf{u}} - \mathbf{v}),\\
&\frac{\partial c_{\mathrm{b}}}{\partial t} + \nabla(c_{\mathrm{b}}\mathbf{u}) = -k_{\mathrm{u}}c_{\mathrm{b}} + k_{\mathrm{b}}u_u,\\
&\frac{\partial c_{\mathrm{u}}}{\partial t} = D\nabla^2 c_{\mathrm{u}} + k_{\mathrm{u}}c_{\mathrm{b}} - k_{\mathrm{b}}c_{\mathrm{u}},\\
&\nabla \cdot \left((1-\phi_{\mathrm{p}})\mathbf{v} + \phi_{\mathrm{p}}\dot{\mathbf{u}}\right) = 0,
\end{aligned}
\tag{27}
$$

where P is the hydrodynamic pressure for the solvent, η the fluid viscosity, λ and μ the Lame coefficients of the gel network, η_{b} and η_{s} are the bulk and shear viscosity arising from internal friction in the gel, $\Delta\mu$ is the change in chemical potential due to hydrolysis of ATP, ζ is a parameter with units of the number density describing the stress per unit change in chemical potential due to the action of crosslinkers, k_{b} is the bounding rate of the motor molecules, k_{u} is the unbinding rate, D is the diffusion coefficient for the unbound motor, and ϕ_{p} is the volume fraction of the active gel network. A transport equation for ϕ_{p} is needed to complete the system. In their model, the volume fraction is assumed small, $\phi_{\mathrm{p}} \ll 1$, so the incompressibility condition reduces to $\nabla \cdot \mathbf{v} = 0$.

The active contribution to the stress is an active pressure on the gel network proportional to $\Delta\mu$. The gel network is modeled as a viscoelastic material subject to an active stress due to ATP activities on the motors bound to the filament. This model leads to spontaneous oscillations at intermediate activity and contractile instability of the network at large activity [7].

3.2 Active Polar Gel Model

In a series of papers, Joanny, Prost, Kruse, Julicher et al. [56,57,63,64,92] studied active gels pertinent to cytoskeletal dynamics. We discuss one of their generic models below. We denote the domain occupied by the gel by Ω, the number density of monomers in the gel by ρ, an average velocity transporting the gel by $\mathbf{v}$. The transport equation for ρ is given by

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\mathbf{v}\rho) = -k_d \delta(S)\rho + k_\mathrm{p}\delta(S), \tag{28}$$

where k_p is the rate of polymerization and h_d is the rate of depolymerization at the gel surface defined by the level surface $\{\mathbf{x}|S = 0\}$. The polymerization and depolymerization in this model are assumed to only take place at the gel surface. Let ρ^a be the number density of diffusing free monomers and the diffusive flux $\mathbf{j}^\mathrm{a}$ of free monomers. The transport equation for ρ^a is

$$\frac{\partial \rho^\mathrm{a}}{\partial t} + \nabla \cdot \mathbf{j}^\mathrm{a} = k_d \delta(S)\rho - k_\mathrm{p}\delta(S). \tag{29}$$

Note that the total number of monomers is conserved

$$\frac{\partial}{\partial t}(\rho + \rho^\mathrm{a}) + \nabla \cdot (\mathbf{v}\rho + \mathbf{j}^\mathrm{a}) = 0. \tag{30}$$

Active processes are mediated by molecular motors. Let $c^{(\mathrm{b})}$ be the concentration of bound motors and $c^{(\mathrm{m})}$ the concentration of the free diffusing motors. The conservation equations for the motors are given by

$$\begin{aligned}
&\frac{\partial c^{(\mathrm{m})}}{\partial t} + \nabla \cdot \mathbf{j}^{(\mathrm{m})} = k_{pff} c^{(\mathrm{b})} - k_\mathrm{on}\left(c^{(\mathrm{m})}\right)^n, \\
&\frac{\partial c^{(\mathrm{b})}}{\partial t} + \nabla \cdot \left(\mathbf{v}c^{(\mathrm{b})}\right) + \nabla \cdot \mathbf{j}^{(\mathrm{b})} = -k_{pff} c^{(\mathrm{b})} + k_\mathrm{on}\left(c^{(\mathrm{m})}\right)^n,
\end{aligned} \tag{31}$$

where k_on and k_off denote the attachment and detachment rate, respectively, and $\mathbf{j}^{(\mathrm{b})}$ and $\mathbf{j}^{(\mathrm{m})}$ are the flux of free motors and the bounded ones relative to the gel motion.

In the timescales considered in their model, the momentum balance is replaced by a force balance equation

$$\nabla \cdot \left(\sigma^\mathrm{total} - \Pi\mathbf{I}\right) + \mathbf{f}^\mathrm{ext} = 0, \tag{32}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{f}^\mathrm{ext}$ is the external force, σ^total denotes the total stress tensor, and Π is the pressure.

Let $\mathbf{p}$ be the polarity vector describing the polar direction of the monomer. The time rate change of the system free energy is given by

$$\dot{F} = -\int \mathrm{d}\mathbf{x}\left[\sigma^{\text{total}} : \nabla\mathbf{v} + \mathbf{h}\cdot\mathbf{P} + \Delta\mu r - c^{\dot{(\mathrm{b})}}\mu^{(\mathrm{b})} - c^{\dot{(\mathrm{m})}}\mu^{(\mathrm{m})} - \dot{\rho}\mu - \dot{\rho}^{\mathrm{a}}\mu^{\mathrm{a}}\right], \tag{33}$$

where $\mathbf{h}$ is the molecular field, $\mathbf{P} = \frac{\partial}{\partial t}\mathbf{p} + \mathbf{v}\cdot\nabla\mathbf{p} + \Omega\cdot\mathbf{p}$ is the corotational derivative of $\mathbf{p}$, $\Delta\mu$ is the chemical force conjugate to the ATP production rate r which determines the number of ATP molecules hydrolyzed per unit time and unit volume. The dot denotes time derivative, $\mu, \mu^{\mathrm{a}}, \mu^{(\mathrm{b})}, \mu^{(\mathrm{m})}$ are the chemical potentials corresponding to $\rho, \rho^{\mathrm{a}}, c^{(\mathrm{b})}, c^{(\mathrm{m})}$, respectively. The total stress is given by

$$\sigma^{\text{total}} = \sigma + \frac{1}{2}(\mathbf{ph} - \mathbf{hp}), \tag{34}$$

where σ is the symmetric part of the stress. The symmetric stress tensor σ, $\mathbf{P}$, and the ATP consumption rate r can be decomposed into reactive part and dissipative part, respectively,

$$\begin{aligned}
\sigma &= \sigma^{\mathrm{r}} + \sigma^{\mathrm{d}},\\
\mathbf{P} &= \mathbf{P}^{\mathrm{r}} + \mathbf{P}^{\mathrm{d}},\\
r &= r^{\mathrm{r}} + r^{\mathrm{d}},
\end{aligned} \tag{35}$$

where the superscripts r denote the reactive response and d the dissipative response. The constitutive equations for the dissipative response are given by

$$\begin{aligned}
\left(1 - \tau^2\frac{D^2}{Dt^2}\right)\sigma^{\mathrm{d}} &= 2\eta\mathbf{D} - \tau\frac{D}{Dt}\left(\frac{\nu_1}{2}(\mathbf{ph} + \mathbf{hp}) + \tilde{\nu}_1(\mathbf{p}\cdot\mathbf{h})\mathbf{I}\right),\\
\left(1 - \tau^2\frac{D^2}{Dt^2}\right)\mathbf{P}^{\mathrm{d}} &= \left(1 - \tau^2\frac{D^2}{Dt^2}\right)\left(\frac{\mathbf{h}}{\gamma_1} + \lambda_1\mathbf{p}\Delta\mu\right)\\
&\quad + \tau\frac{D}{Dt}(\nu_1\nabla\mathbf{v}\cdot\mathbf{p} + \tilde{\nu}_1 Tr(\nabla\mathbf{v})\mathbf{p}),\\
r^{\mathrm{d}} &= \Lambda\Delta\mu + \lambda_1\mathbf{p}\cdot\mathbf{h} + \lambda\mathbf{p}\cdot\nabla\mu^{(\mathrm{b})},
\end{aligned} \tag{36}$$

where $\nu_1, \tilde{\nu}_1, \Lambda, \lambda_1\lambda$ are model parameters and τ is the relaxation time. The fluxes for the monomers and motor molecules, which do not have reactive parts, are given by

$$\begin{aligned}
\mathbf{j}^{(a)} &= -D^{(a)}\nabla\rho^{(a)} + \lambda^{(a)}\Delta\mu\mathbf{p},\\
\mathbf{j}^{(\mathrm{m})} &= -D^{(m)}\nabla c^{(\mathrm{m})} + \lambda^{(m)}\Delta\mu\mathbf{p},\\
\mathbf{j}^{(\mathrm{b})} &= -D^{(b)}\nabla c^{(\mathrm{b})} + \lambda^{(b)}\Delta\mu\mathbf{p},
\end{aligned} \tag{37}$$

where $D^{(i)}$ are the diffusion coefficients, $\lambda^{(i)}$ are coupling parameters.

The reactive fluxes, the polarity vector, and the ATP consumption rate, are given next.

$$\begin{aligned}
\sigma^{\mathrm{r}} &= -\tau\left(\frac{D\sigma^{\mathrm{d}}}{Dt}+\mathbf{A}\right)-\zeta\Delta\mu\mathbf{pp}-\bar{\zeta}\Delta\mu\mathbf{I}-\zeta'\Delta\mu\|\mathbf{p}\|^2\mathbf{I}+\frac{\nu_1}{2}(\mathbf{ph}+\mathbf{hp})+\bar{\nu_1}\mathbf{p}\cdot\mathbf{hI},\\
&\left(1-\tau^2\frac{D^2}{Dt^2}\right)\mathbf{P}^{\mathrm{r}} = -\nu_1\nabla\mathbf{v}\cdot\mathbf{p}-\bar{\nu_1}Tr(\nabla\mathbf{v})\mathbf{p},\\
r^{\mathrm{r}} &= \zeta\mathbf{pp}:\nabla\mathbf{v}+\bar{\zeta}Tr(\nabla\mathbf{v})+\zeta'\|\mathbf{p}\|^2Tr(\nabla\mathbf{v}),
\end{aligned} \tag{38}$$

where $\zeta, \bar{\zeta}, \zeta', \nu_1, \bar{\nu_1}$ are model parameters and

$$\begin{aligned}
\mathbf{A} = {} & \nu_2\left(\nabla\cdot\sigma^{\mathrm{d}}+\sigma^{\mathrm{d}}\cdot\nabla\mathbf{v}\right)+\nu_3Tr(\nabla\mathbf{v})\sigma^{\mathrm{d}}+\nu_4Tr(\nabla\mathbf{v})Tr\left(\sigma^{\mathrm{d}}\right)\mathbf{I}\\
& +\nu_5Tr\left(\sigma^{\mathrm{d}}\right)\nabla\mathbf{v}+\nu_6\nabla\mathbf{v}:\sigma^{\mathrm{d}}\mathbf{I},
\end{aligned} \tag{39}$$

where ν_i are the model parameters analogous to the eight-constant Oldroyd model.

Combining the reactive and dissipative parts, the total stress, polarity vector, and the ATP consumption rate are finally given by

$$\begin{aligned}
2\eta\mathbf{D} = {} & \left(1+\tau\frac{D}{Dt}\right)\left(\sigma+\zeta\Delta\mu\mathbf{pp}+\zeta'\Delta\mu\|\mathbf{p}\|^2\mathbf{I}+\bar{\zeta}\Delta\mu\mathbf{I}+\tau\mathbf{A}\right)\\
& -\frac{\nu_1}{2}(\mathbf{ph}+\mathbf{hp}-\tilde{\nu}_1(\mathbf{p}\cdot\mathbf{h})\mathbf{I}),\\
\left(1-\tau^2\frac{D^2}{Dt^2}\right)\mathbf{P} = {} & \left(1-\tau^2\frac{D^2}{Dt^2}\right)\left(\frac{\mathbf{h}}{\gamma_1}+\lambda_1\mathbf{p}\Delta\mu\right)\\
& -\left(1-\tau\frac{D}{Dt}\right)\left(\nu_1\nabla\mathbf{v}\cdot\mathbf{p}+\tilde{\nu}_1Tr(\nabla\mathbf{v})\mathbf{p}\right),\\
r = {} & \Lambda\Delta\mu+\lambda_1\mathbf{p}\cdot\mathbf{h}+\lambda\mathbf{p}\cdot\nabla\mu^{(\mathrm{b})}+\zeta\mathbf{pp}:\nabla\mathbf{v}+\bar{\zeta}Tr(\nabla\mathbf{v})\\
& +\zeta'\|\mathbf{p}\|^2Tr(\nabla\mathbf{v}).
\end{aligned} \tag{40}$$

By restricting $\mathbf{p}$ to be a unit vector and using a free energy for polar liquid crystals

$$F = \int\left[\frac{K_1}{2}(\nabla\cdot\mathbf{p})^2+\frac{K_3}{2}\|\mathbf{p}\cdot\nabla\mathbf{p}\|^2+k\nabla\mathbf{p}-\frac{h_{||}}{2}\|\mathbf{p}\|^2\right]\mathrm{d}\mathbf{x}, \tag{41}$$

Kruse et al. studied point defects in two dimensions [64]. This model was later extended to a multicomponent active fluid model by Joanny et al. [55].

3.3 Three-Component Active Fluid Model

In this multicomponent model, the active fluid is assumed to consist of three effective components [55]. Let n_0 denote the number density of the monomeric subunits in a polar network, n_1 the number density of the free monomeric subunits, and n_2 the number density of the solvent molecules. The effect of ATP hydrolysis is considered in the model. The conservation equations for the three densities are given by

$$\begin{aligned}
\frac{\partial n_0}{\partial t} + \nabla \cdot \mathbf{J}_0 &= S, \\
\frac{\partial n_1}{\partial t} + \nabla \cdot \mathbf{J}_1 &= -S, \\
\frac{\partial n_2}{\partial t} + \nabla \cdot \mathbf{J}_2 &= 0,
\end{aligned} \tag{42}$$

where the source term S represents the polymerization and depolymerization which leads to the exchange of monomers between the gel and the solvent, the flux constitutive equations are

$$\begin{aligned}
\mathbf{J}_0 &= n_0 \mathbf{v} + \frac{\mathbf{j}_0}{m_0}, \\
\mathbf{J}_1 &= n_1 \mathbf{v} + \frac{\mathbf{j}_1}{m_1}, \\
\mathbf{J}_2 &= n_2 \mathbf{v} - \frac{\mathbf{j}_0}{m_2} - \frac{\mathbf{j}_1}{m_2},
\end{aligned} \tag{43}$$

$m_{0,1,2}$ are the molecular masses of monomers in the gel, free monomers in the solution, and the solvent molecules, respectively. The mass density of the material system is given by $\rho = m_0 n_0 + m_1 n_1 + m_2 n_2$. These equations warrant the conservation of mass

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{44}$$

because $m_0 = m_1$. If the monomeric subunits on the polymer network (m_0) differs from those free ones (m_1), the mass conservation may not be upheld in the model. In this case, the transport equation for n_i must be modified.

The conservation of linear momentum is given by

$$\frac{\partial}{\partial t}(\rho \mathbf{v}) + \nabla \cdot (\rho \mathbf{v}\mathbf{v}) = \nabla \cdot \sigma, \tag{45}$$

where σ is the total stress of the system and external forces are absent. We denote the free energy density by f and the free energy by F, i.e., $F = \int f \mathrm{d}\mathbf{x}$. The time rate of change of the free energy is given by

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \int \left[\partial_t \left(\frac{\rho}{2} \|\mathbf{v}\|^2 \right) + \sum_{i=0}^{2} \mu_i \partial_t n_i - \mathbf{h} \cdot \partial_t \mathbf{p} - r\Delta\mu \right] \mathrm{d}\mathbf{x}, \tag{46}$$

where $\mu_i = \frac{\delta F}{\delta n_i}, i = 0, 1, 2$ are the chemical potentials for the three effective components, respectively, $\mathbf{h} = -\frac{\delta F}{\delta \mathbf{p}}$ is the molecular field, r is the rate at which ATP molecules are hydrolyzed, and $\Delta\mu = \mu_{\mathrm{ATP}} - \mu_{\mathrm{ADP}} - \mu_{\mathrm{P}}$ is the difference in chemical potentials of ATP and the product molecules ADP and P_i, respectively. Using the generalized Gibbs–Duhem relation for a multicomponent polar fluid

$$\nabla \cdot \sigma^{\mathrm{e}} = -\sum_{i=0}^{2} n_i \nabla \mu_i - \nabla \mathbf{p} \cdot \mathbf{h}, \tag{47}$$

where σ^{e} represents the Ericksen stress, the free energy rate of change is rewritten into

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \int \left[-\sigma^{\mathrm{s}} : \nabla \mathbf{v} + \sum_{i=0}^{1} \mathbf{j}_i \cdot \nabla \bar{\mu}_i + (\mu_0 - \mu_1) S - \mathbf{P} \cdot \mathbf{h} - r\Delta\mu \right] \mathrm{d}\mathbf{x}, \tag{48}$$

where σ^{s} is the symmetric part of the stress less the Ericksen stress as well as the anisotropic stress, $\bar{\mu}_i = \frac{\mu_i}{m_i} - \frac{\mu_2}{m_2}$, $\mathbf{P} = \frac{\partial}{\partial t}\mathbf{p} + \mathbf{v} \cdot \nabla \mathbf{p} + \Omega \cdot \mathbf{p}$ is the convected corotational derivative of $\mathbf{p}$, Ω is the vorticity tensor,

$$\begin{aligned} \sigma^{\mathrm{s}} &= \sigma - \sigma^{\mathrm{a}} - \sigma^{\mathrm{e,s}}, \\ \sigma^{\mathrm{a}} &= \frac{1}{2}(\mathbf{ph} - \mathbf{hp}), \\ \sigma^{\mathrm{e,s}} &= sym \left[\left(f - \sum_{i=0}^{2} \mu_i n_i \right) \mathbf{I} - \frac{\partial f}{\partial \nabla \mathbf{p}} \cdot \nabla \mathbf{p} \right], \end{aligned} \tag{49}$$

where sym denotes the symmetric part of the stress. We can identify the generalized force fields $(\nabla \mathbf{v}, -\nabla \bar{\mu}_i, \mathbf{h}, \Delta\mu)$. The corresponding conjugate fluxes are $(\sigma^{\mathrm{s}}, \mathbf{j}_i, \mathbf{P}, r)$, assuming the fluxes are functions of the forces, and expanded to linear order. The force fields are distinguished in that some forces change signs when time is reversed like $\nabla \mathbf{v}$ while others do not. The stress component obeying time reversal is dissipative and the others are reactive. With this, we propose the following phenomenological dissipative fluxes

$$
\begin{aligned}
\sigma^{\mathrm{s,d}} &= 2\eta\left(\nabla\mathbf{v} - \frac{Tr(\nabla\mathbf{v})}{3}\mathbf{I}\right) + \bar{\eta}Tr(\nabla\mathbf{v})\mathbf{I},\\
\mathbf{j}_i^{\mathrm{d}} &= -\sum_{j=0}^{1}\gamma_{ij}\nabla\bar{\mu}_j + \bar{\lambda}_i\mathbf{h} + \kappa_i\mathbf{p}\Delta\mu,\\
\mathbf{P}^{\mathrm{d}} &= +\bar{\lambda}_i\nabla\bar{\mu}_i + \frac{\mathbf{h}}{\gamma_1} - \lambda_1\Delta\mu\mathbf{p},\\
r^{\mathrm{d}} &= -\sum_{i=0}^{1}\kappa_i\mathbf{p}\cdot\nabla\bar{\mu}_i + \lambda_1\mathbf{p}\cdot\mathbf{h} + \Lambda\Delta\mu,
\end{aligned}
\tag{50}
$$

where (γ_{ij}) is nonnegative definite, Λ, γ_1 are nonnegative. The reactive terms are proposed as follows:

$$
\begin{aligned}
\sigma^{\mathrm{s,r}} &= -\sum_{i=0}^{1}\frac{\epsilon_j}{2}\left(\mathbf{p}\nabla\bar{\mu_j} + \nabla\bar{\mu_j}\mathbf{p}\right) - \sum_{j=0}^{1}\bar{\epsilon_j}\mathbf{p}\cdot\nabla\bar{\mu_j}\mathbf{I} + \frac{\nu_1}{2}(\mathbf{ph}+\mathbf{hp})\\
&\quad + \bar{\nu_1}\mathbf{p}\cdot\mathbf{h}\mathbf{I} - \zeta_1\mathbf{pp}\Delta\mu - \zeta_2\Delta\mu\mathbf{I} - \zeta_3\|\mathbf{p}\|^2\Delta\mu\mathbf{I},\\
\mathbf{j}_i^{\mathrm{r}} &= -\epsilon_i\nabla\mathbf{v}\cdot\mathbf{p} - \bar{\epsilon_i}Tr(\nabla\mathbf{v})\mathbf{p},\\
\mathbf{P}^{\mathrm{r}} &= -\nu_1\nabla\mathbf{v}\cdot\mathbf{p} - \bar{\nu_1}Tr(\nabla\mathbf{v})\mathbf{p},\\
r^{\mathrm{r}} &= \zeta_1\mathbf{pp}:\nabla\mathbf{v} + \zeta_2 Tr(\nabla\mathbf{v}) + \zeta_3\|\mathbf{p}\|^2 Tr(\nabla\mathbf{v}),
\end{aligned}
\tag{51}
$$

where $\zeta_{1,2,3}$ are the coefficients for the active terms. By applying the Onsager reciprocal principle and verifying the long and short time asymptotic behavior, the constitutive equation can be extended to account for viscoelastic behavior and chirality.

$$
\begin{aligned}
\sigma^{\mathrm{s,r}} &= -\tau\left[\frac{D}{Dt}\sigma^{\mathrm{s,d}} + \mathbf{A}\right] - \sum_{i=0}^{1}\frac{\epsilon_j}{2}\left(\mathbf{p}\nabla\bar{\mu_j} + \nabla\bar{\mu_j}\mathbf{p}\right) - \sum_{j=0}^{1}\bar{\epsilon_j}\mathbf{p}\cdot\nabla\bar{\mu_j}\mathbf{I}\\
&\quad + \frac{\nu_1}{2}(\mathbf{ph}+\mathbf{hp}) + \bar{\nu_1}\mathbf{p}\cdot\mathbf{h}\mathbf{I} - \zeta_1\mathbf{pp}\Delta\mu - \zeta_2\Delta\mu\mathbf{I} - \zeta_3\|\mathbf{p}\|^2\Delta\mu\mathbf{I}\\
&\quad + \frac{\Pi_1}{2}(\mathbf{p}\times\mathbf{hp} + \mathbf{pp}\times\mathbf{h}) + \sum_{j=0}^{1}\frac{\Pi_2}{2}\left(\mathbf{p}\times\nabla\bar{\mu_j}\mathbf{p} + \mathbf{pp}\times\nabla\bar{\mu_j}\right),\\
\mathbf{j}_i^{\mathrm{r}} &= -\epsilon_i\nabla\mathbf{v}\cdot\mathbf{p} - \bar{\epsilon_i}Tr(\nabla\mathbf{v})\mathbf{p} - \Pi_2\nabla\mathbf{v}\cdot\mathbf{p}\times\mathbf{p},\\
\frac{D}{Dt}\mathbf{p}^r &= \tau\frac{D}{Dt}\frac{\mathbf{h}}{\gamma_1} - \nu_1\nabla\mathbf{v}\cdot\mathbf{p} - \bar{\nu_1}Tr(\nabla\mathbf{v})\mathbf{p} - \Pi_1\nabla\mathbf{v}\cdot\mathbf{p}\times\mathbf{p},\\
r^{\mathrm{r}} &= \zeta_1\mathbf{pp}:\nabla\mathbf{v} + \zeta_2 Tr(\nabla\mathbf{v}) + \zeta_3\|\mathbf{p}\|^2 Tr(\nabla\mathbf{v}),
\end{aligned}
\tag{52}
$$

where Π_1, Π_2 denote the coefficients for the chiral terms. The dissipative parts are given by

$$\left(1-\tau^2\frac{D^2}{Dt^2}\right)\sigma^{\mathrm{s,d}} = 2\eta\left(\nabla\mathbf{v}-\frac{Tr(\nabla\mathbf{v})}{3}\mathbf{I}\right)+\bar{\eta}Tr(\nabla\mathbf{v})\mathbf{I},$$

$$\mathbf{j}_i^{\mathrm{d}} = -\sum_{j=0}^{1}\gamma_{ij}\nabla\bar{\mu}_j+\bar{\lambda}_i\mathbf{h}+\kappa_i\mathbf{p}\Delta\mu+\Pi_3^i\mathbf{p}\times\mathbf{h},$$

$$\mathbf{P}^{\mathrm{d}} = +\bar{\lambda}_i\nabla\bar{\mu}_i+\frac{\mathbf{h}}{\gamma_1}-\lambda_1\Delta\mu\mathbf{p}-\sum_{j=0}^{1}\Pi_3^i\mathbf{p}\times\nabla\bar{\mu}_j,$$

$$r^{\mathrm{d}} = -\sum_{i=0}^{1}\kappa_i\mathbf{p}\cdot\nabla\bar{\mu}_i+\lambda_1\mathbf{p}\cdot\mathbf{h}+\Lambda\Delta\mu, \tag{53}$$

where Π_3^i denotes coefficients for the chiral terms. The reactive and dissipative parts can be combined to yield the constitutive equations for the active gel system. The details are available in [55].

4 A Phase Field Model for a Cell Surrounded by Solvent

We take a simplistic view of the cell structure recognizing the cell membrane as an elastic closed surface, the nucleus/core as a relatively hard, closed 3-D object inside the membrane, the remaining cytoplasm/cytoskeleton as a mixture of ATP bound and ADP bound G-actin, actin filament networks (or polymer-networks), and a third phase material called solvent which includes all other accessory proteins, organelles, and other unaccounted for material in the cytoplasm. In the simplified formulation, we assume the G-actin is available for polymerization at the barbed end and depolymerization at the pointed end [13]. This assumption will be refined later in the following.

We use a single-phase field variable or labeling function $\phi(\mathbf{x},t)$ to denote the material inside or outside the cell. Since the core is always disjoint from the outside of the cell membrane, we simply use $\phi = -1$ to denote or label the material outside the cell membrane and the one inside the core at the same time; whereas the material in the cytoplasmic region is denoted by $\phi = 1$. We treat all materials in the cytoplasm as multiphase complex fluids or complex fluid mixtures. The interfacial free energy at the interfaces associated with the phase field variable ϕ is given by

$$f_{\mathrm{mb}} = \frac{k_{\mathrm{B}}T\kappa_{\mathrm{b}}}{2}\int_{\mathbf{S}}\left[\left[\tau_0+(C_1+C_2-C_0)^2+\kappa_{\mathrm{G}}C_1C_2\right]\mathrm{d}S+\kappa_{\mathrm{d}}\left(\Delta S-\Delta S_0\right)^2\right], \tag{54}$$

where f_{mb} is the Helfrich elastic membrane energy, ϵ is the transitional parameter that scales with the width of the interfacial region, k_{B} is the Boltzmann constant and T the absolute temperature, τ_0 is a constant that is the analog of surface tension of the membrane, κ_{b} is the bending rigidity and κ_{G} is the Gaussian bending rigidity, respectively, C_1 and C_2 are the principle curvatures, respectively, κ_{d} is a constant for the nonlocal bending resistance also related to the area compression modulus of the membrane surface, and $\mathbf{S}$ denotes the membrane surface. For single-layered membranes, $\kappa_{\text{d}} = 0$, whereas it maybe nonzero for bilayers. We notice that the Gaussian bending elastic energy integrates to a constant when the cell membrane does not undergo any topological changes. For simplicity, we will treat it as a constant in this book chapter.

Note that $\int_\Omega \frac{1+\phi}{2} \mathrm{d}\mathbf{x} = V_{\text{c}}$ is the volume of the cytoplasm region. To conserve the volume of this region, we can simply enforce $V(\phi) = \int_\Omega \phi \mathrm{d}\mathbf{x} = V(\phi(t = t_0))$ at some specified time t_0. In addition, the surface area of the membrane can be approximated by the formula

$$A(\phi) = \kappa_{\text{a}} \int_\Omega \left(\|\nabla \phi\|^2 + \frac{(\phi^2 - 1)^2}{2\epsilon^2} \right) \mathrm{d}\mathbf{x}, \tag{55}$$

where κ_{a} is a scaling parameter. In the case of $\kappa_{\text{d}} = 0$, the free energy can be represented by the phase field variable ϕ [29–32, 34]

$$\begin{aligned} f_{\text{mb}} = \frac{k_{\text{B}} T \kappa_{\text{b}}}{k_a \epsilon} \int_\Omega \Bigg[& \tau_0 \left(\frac{\epsilon}{2} \|\nabla \phi\|^2 + \frac{1}{4\epsilon} \left(1 - \phi^2\right)^2 \right) \\ & + \epsilon \left(\Delta \phi - \frac{1}{\epsilon^2} \left(\phi^2 - 1\right) \left(\phi + \sqrt{2} C_0 \epsilon \right) \right)^2 \Bigg] \mathrm{d}x. \end{aligned} \tag{56}$$

If $\kappa_{\text{d}} \neq 0$, we can similarly formulate the last term of (54).

For a weakly compressible and extensible membrane, we modify the elastic energy as following:

$$\begin{aligned} f_{\text{mb}} = \frac{k_{\text{B}} T \kappa_{\text{b}}}{k_a \epsilon} \int_\Omega \Bigg[& \tau_0 \left(\frac{\epsilon}{2} \|\nabla \phi\|^2 + \frac{1}{4\epsilon} \left(1 - \phi^2\right)^2 \right) \\ & + \epsilon \left(\Delta \phi - \frac{1}{\epsilon^2} \left(\phi^2 - 1\right) \left(\phi + \sqrt{2} C_0 \epsilon \right) \right)^2 \Bigg] \mathrm{d}\mathbf{x} \\ & + M_1 \left(A(\phi) - A\left(\phi(t_0)\right) \right)^2 + M_2 \left(V(\phi) - V(\phi(t_0)) \right)^2, \end{aligned} \tag{57}$$

where M_1 and M_2 are penalizing constants. In this formulation, we penalize the volume and surface area difference to limit the variation of the two conserved quantities as in Du et al. [29–32, 34]. We can drop the surface tension term since we are penalizing it in the energy potential already,

$$f_{\mathrm{mb}} = \frac{k_{\mathrm{B}} T \kappa_{\mathrm{b}}}{k_a \epsilon} \int_{\Omega} \left[\epsilon \left(\Delta\phi - \frac{1}{\epsilon^2} \left(\phi^2 - 1\right) \left(\phi + \sqrt{2} C_0 \epsilon \right) \right)^2 \right] \mathrm{d}\mathbf{x}$$
$$+ M_1 \left(A(\phi) - A(\phi(t_0)) \right)^2 + M_2 (V(\phi) - V(\phi(t_0)))^2. \tag{58}$$

This will be the free energy density used in our cell model.

We denote $\mathbf{v}$ as the velocity of the mixture, and p the hydrostatic pressure. We denote by ρ_1 the mass density of the fluid outside the membrane and inside the core and by ρ_2 the mass density of the mixture in the cytoplasm. We assume the material is incompressible in both domains, i.e., ρ_1 and ρ_2 are constants. The density of the mixture is defined as

$$\rho = \frac{\rho_1}{2}(1 - \phi) + \frac{\rho_2}{2}(1 + \phi). \tag{59}$$

From mass conservation, we have

$$\nabla \cdot \mathbf{v} = -\frac{\rho_2 - \rho_1}{\rho_1 + \rho_2} \frac{\mathrm{d}\phi}{\mathrm{d}t}. \tag{60}$$

This is true when

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} + \phi \nabla \cdot \mathbf{v} = -\frac{1}{\lambda} \mu. \tag{61}$$

Here, $\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{\partial \phi}{\partial t} + \mathbf{v} \cdot \nabla \phi$ is the material derivative and μ is the chemical potential of the material system.

If we use

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = -\frac{1}{\lambda} \mu \tag{62}$$

to transport ϕ, the continuity equation should be

$$\nabla \cdot \mathbf{v} = -\frac{\rho_2 - \rho_1}{\rho} \frac{\mathrm{d}\phi}{\mathrm{d}t}. \tag{63}$$

The balance of linear momentum is governed by

$$\rho \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = \nabla \cdot (-p\mathbf{I} + \tau) + \mathbf{F}_{\mathrm{e}},$$
$$\tau = \tau_1 + \tau_2, \tag{64}$$

where τ_1 is the stress tensor for the fluid outside the membrane and inside the core, τ_2 is the stress tensor inside the cytoplasmic region, and $\mathbf{F}_{\mathrm{e}}$ is the external force exerted on the complex fluid.

Constitutive equations:

We assume the ambient fluid material surrounding the cell is viscous, whose extra stress is given by the viscous stress law:

$$\tau_1 = (1-\phi)\eta_1 \mathbf{D}, \tag{65}$$

where $\mathbf{D} = \frac{1}{2}[\nabla \mathbf{v} + \nabla \mathbf{v}^T]$ is the rate-of-strain tensor for the mixture. The viscosities outside the cell and inside the core are distinct.

The extra stress for the cytoplasm is a volume-fraction weighted stress:

$$\tau_2 = (1+\phi)\eta_s \mathbf{D} + \tau_p, \tag{66}$$

where η_s is the zero shear rate viscosity and τ_p is the viscoelastic stress [38].

The total free energy for the complex fluid mixture system is given by

$$f = f_{mb} + f_n, \tag{67}$$

where f_n is the free energy associated to the active cytoplasmic material:

$$f_n = f_n(\phi, \mathbf{Q}, \nabla \mathbf{Q}), \tag{68}$$

where $\mathbf{Q}$ is the orientation tensor in cytoplasm with $tr(\mathbf{Q}) = 0$. It is zero outside the cell. We denote the chemical potential with respect to ϕ by

$$\mu = \frac{\delta f}{\delta \phi}. \tag{69}$$

The time evolution of the membrane interface is governed by the Allen–Cahn equation

$$\frac{d\phi}{dt} = -\frac{1}{\lambda_1}\mu, \tag{70}$$

where λ_1 is a relaxation parameter. The Cahn–Hilliard dynamics can also be used if we assume the volume conservation without additional constraint:

$$\frac{d\phi}{dt} = \nabla \cdot (\lambda_1 \nabla \mu), \tag{71}$$

where λ_1 is the mobility parameter which has different physical units than the analogous parameter in the Allen–Cahn dynamics. In this latter case, the term $V(\phi) - V(\phi_0)$ is identically zero in the energy potential and can be dropped from the surface energy expression.

The transport equation for the orientation tensor $\mathbf{Q}$ is proposed as following

$$\begin{aligned}\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}t} &+\mathbf{W}\cdot\mathbf{Q}-\mathbf{Q}\cdot\mathbf{W}-a\left[\mathbf{D}\cdot\mathbf{Q}+\mathbf{Q}\cdot\mathbf{D}\right]=+\frac{a(1+\phi)\mathbf{D}}{3}\\ &-2a\mathbf{D}:(\mathbf{Q}+(1+\phi)\mathbf{I}/6)(\mathbf{Q}+(1+\phi)\mathbf{I}/6)+\Gamma\mathbf{H}+\lambda_2\mathbf{Q},\end{aligned}\tag{72}$$

where λ_2 is an active parameter, $\mathbf{W}$ is the vorticity tensor, $\mathbf{D}$ is the rate of strain tensor, $\mathbf{H}=-[\frac{\delta f}{\delta\mathbf{Q}}-tr(\frac{\delta f}{\delta\mathbf{Q}})(1+\phi)\mathbf{I}/6]$ is the so-called molecular field and f_n is the free energy density associated with the orientational dynamics given by

$$\begin{aligned}f_\mathrm{n}=A_0\left(\frac{1+\phi}{2}\right)^r&\left[\frac{1}{2}(1-N/3)\,\mathbf{Q}:\mathbf{Q}-\frac{N}{3}tr\left(\mathbf{Q}^3\right)+\frac{N}{4}(\mathbf{Q}:\mathbf{Q})^2\right]\\ &+\frac{(1+\phi)^rK}{2^{r+1}}\left(\nabla\mathbf{Q}\vdots\nabla\mathbf{Q}\right)+f_\mathrm{anch},\end{aligned}\tag{73}$$

where $r=1$ is a positive integer, N is the dimensionless concentration, K is a elastic constant, and f_anch is the anchoring potential [11].

Elastic stress

The elastic stress is calculated by the virtual work principle [27]. Consider a virtual deformation given by $\mathbf{E}=\nabla\mathbf{v}\delta t$. The corresponding change in the free energy is given by

$$\delta f=\mu\frac{\partial\phi}{\partial t}\delta t-\mathbf{H}:\frac{\partial\mathbf{Q}}{\partial t}\delta t.\tag{74}$$

The variation of ϕ, $\mathbf{Q}$ are given, respectively, by

$$\begin{aligned}\delta\phi&=\frac{\partial\phi}{\partial t}=-\nabla(\phi\mathbf{v})\delta t,\\ \delta\mathbf{Q}&=\Big[-\nabla\cdot(\mathbf{vQ})+\mathbf{W}\cdot\mathbf{Q}-\mathbf{Q}\cdot\mathbf{W}+a\left[\mathbf{D}\cdot\mathbf{Q}+\mathbf{Q}\cdot\mathbf{D}\right]\\ &\quad+\frac{a(1+\phi)}{3}\mathbf{D}-2a\mathbf{D}:(\mathbf{Q}+(1+\phi)\mathbf{I}/6)\,(\mathbf{Q}+(1+\phi)\mathbf{I}/6)\Big]\delta t.\end{aligned}\tag{75}$$

So, the elastic stress is calculated as

$$\begin{aligned}\mathbf{F}_\mathrm{e}&=-\phi\nabla(\mu)+\nabla(\mathbf{H}_{ij})\mathbf{Q}_{ij},\\ \tau_\mathrm{p}&=-(\mathbf{H}\cdot\mathbf{Q}-\mathbf{Q}\cdot\mathbf{H})-a(\mathbf{H}\cdot(\mathbf{Q}+(1+\phi)\mathbf{I}/6)+(\mathbf{Q}+(1+\phi)\mathbf{I}/6)\cdot\mathbf{H})\\ &\quad+2a(\mathbf{Q}+(1+\phi)\mathbf{I}/6):\mathbf{H}(\mathbf{Q}+(1+\phi)\mathbf{I}/6)-\zeta\mathbf{Q},\end{aligned}\tag{76}$$

where $\zeta < 0$ (respectively, $\zeta > 0$) represents a contractile (respectively, extensile) filament.

$$\begin{aligned}
\mu &= \frac{\delta f_{\mathrm{n}}}{\delta \phi} + \frac{\delta f_{\mathrm{mb}}}{\delta \phi} \\
&= \frac{\delta f_{\mathrm{n}}}{\delta \phi} - 4M_1 \left(A(\phi) - A(\phi_0)\right) \kappa_{\mathrm{a}} \left(\nabla^2 \phi + \frac{2}{\epsilon^2}\phi(1-\phi^2)\right) + M_2(V(\phi) - V(\phi_0)) \\
&\quad + \frac{2\phi\left(\phi^2-1\right)}{\epsilon^2} + k_{\mathrm{b}}\epsilon \left[\nabla^2 \left(\nabla^2 \phi - \frac{1}{\epsilon^2}\left(\phi^2-1\right)\left(\phi + \sqrt{2}C_0\epsilon\right)\right)\right. \\
&\quad \left. - \frac{2}{\epsilon^2}\left(\nabla^2 \phi - \frac{1}{\epsilon^2}\left(\phi^2-1\right)\left(\phi + \sqrt{2}C_0\epsilon\right)\right)\phi\left(3\phi^2 + 2\sqrt{2}C_0\epsilon\phi - 1\right)\right] \\
\mathbf{H} &= -\frac{A_0(1+\phi)^r}{2^r}\left[(1-N/3)\,\mathbf{Q} - N\mathbf{Q}^2 + N\mathbf{Q} : \mathbf{Q}\left(\mathbf{Q} + \frac{(1+\phi)\mathbf{I}}{6}\right)\right] \\
&\quad + \frac{K}{2^r}\nabla \cdot \left((1+\phi)^r \nabla \mathbf{Q}\right) - \frac{\delta f_{\mathrm{anch}}}{\delta \mathbf{Q}}.
\end{aligned} \tag{77}$$

4.1 Approximate Model

We impose a solenoidal velocity field

$$\nabla \cdot \mathbf{v} = 0. \tag{78}$$

Then, the model can be simplified further.

$$\begin{aligned}
\frac{\mathrm{d}\phi}{\mathrm{d}t} &= -\frac{1}{\lambda_1}\mu, \\
\rho\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} &= \nabla \cdot (-p\mathbf{I} + \tau) + \mathbf{F}_{\mathrm{e}}, \\
\rho &= \frac{(1-\phi)}{2}\rho_1 + \frac{(1+\phi)}{2}\rho_2, \\
\tau &= \tau_1 + \tau_2, \tau_1 = (1-\phi)\eta_1\mathbf{D}, \tau_2 = (1+\phi)\eta_{\mathrm{s}}\mathbf{D} + \tau_{\mathrm{p}}, \\
\mathbf{F}_{\mathrm{e}} &= -\phi\nabla(\mu) + \nabla(\mathbf{H}_{ij})\mathbf{Q}_{ij},
\end{aligned}$$

$$\tau_{\rm p} = -(\mathbf{H}\cdot\mathbf{Q}-\mathbf{Q}\cdot\mathbf{H}) - a(\mathbf{H}\cdot(\mathbf{Q}+(1+\phi)\mathbf{I}/6)+(\mathbf{Q}+(1+\phi)\mathbf{I}/6)\cdot\mathbf{H})$$
$$+2a(\mathbf{Q}+(1+\phi)\mathbf{I}/6):\mathbf{H}(\mathbf{Q}+(1+\phi)\mathbf{I}/6)-\zeta\mathbf{Q},$$
$$\frac{{\rm d}\mathbf{Q}}{{\rm d}t}+\mathbf{W}\cdot\mathbf{Q}-\mathbf{Q}\cdot\mathbf{W}-a[\mathbf{D}\cdot\mathbf{Q}+\mathbf{Q}\cdot\mathbf{D}]=\frac{a(1+\phi)}{3}\mathbf{D}$$
$$-2a\mathbf{D}:(\mathbf{Q}+(1+\phi)\mathbf{I}/6)(\mathbf{Q}+(1+\phi)\mathbf{I}/6)+\Gamma\mathbf{H}+\lambda_2\mathbf{Q},$$
$$\zeta=\zeta_0(1+\phi),\lambda_2=\lambda_2^0(1+\phi),\Gamma=\Gamma_0(1+\phi), \tag{79}$$

where ζ_0 and λ_2^0 are parameters which depend on regulatory proteins such as the Rho family of GTPases for the active gel. *As a proof-of-principle illustration for this article, we assume these activation parameters are prescribed functions of space and time.*

Remark. (i) What should the anchoring condition be at the membrane interface? The anchoring potential density is given by

$$f_{\rm anch}=W_0\left(1-\phi^2\right)\left[\alpha_1\left(\mathbf{Q}+\frac{(1+\phi)\mathbf{I}}{6}\right):(\nabla\phi\nabla\phi)\right.$$
$$\left.+\alpha_2\left(\|\nabla\phi\|^2-\left(\mathbf{Q}+\frac{(1+\phi)\mathbf{I}}{6}\right):(\nabla\phi\nabla\phi)\right)\right]$$
$$=W_0\left(1-\phi^2\right)\left[(\alpha_1-\alpha_2)\left(\mathbf{Q}+\frac{(1+\phi)\mathbf{I}}{6}\right):(\nabla\phi\nabla\phi)+\alpha_2\|\nabla\phi\|^2\right], \tag{80}$$

where $\alpha_2=0$ gives the tangential anchoring and $\alpha_1=0$ gives the normal anchoring. Then, the variations of the potential are given by

$$\frac{\delta f_{\rm anch}}{\delta\mathbf{Q}}=(\alpha_1-\alpha_2)\,W_0\left(1-\phi^2\right)\left(\nabla\phi\nabla\phi-\frac{\mathbf{I}}{3}\|\nabla\phi\|^2\right),$$
$$\frac{\delta f_{\rm anch}}{\delta\phi}=-2\phi W_0\left[(\alpha_1-\alpha_2)\left(\mathbf{Q}+\frac{(1+\phi)\mathbf{I}}{6}\right):(\nabla\phi\nabla\phi)+\alpha_2\left(\|\nabla\phi\|^2\right)\right]$$
$$-2W_0\left(1-\phi^2\right)\left[\alpha_2\nabla^2\phi+(\alpha_1-\alpha_2)\,\nabla\cdot\left(\left(\mathbf{Q}+\frac{(1+\phi)\mathbf{I}}{6}\right)\cdot\nabla\phi\right)\right]$$
$$+\frac{W_0(\alpha_1-\alpha_2)\left(1-\phi^2\right)}{6}\|\nabla\phi\|^2. \tag{81}$$

We need to update the $\mathbf{H}$ and μ using the above equations.

(ii) How do we deal with the core of the cell? If we don't want to introduce additional variables and equations, we could use the same membrane equation at the cytoplasm-core interface. The viscosity at the core will have to be much higher than the viscosity in the fluid outside the cell. An alternative is to introduce a second phase variable ψ to deal with the interface between the cytoplasm and the core.

In the following, we apply the multiphase complex fluid cell model to an active cortical layer near the membrane. Everything outside the layer is treated as a viscous fluid for simplicity. Our goal is to investigate how this cytoskeletal-membrane coupled model responds to an imposed ATP-activated stress in the cortical layer.

5 Numerical Results and Discussion

The coupled flow and structure equations are solved using a spectral method in 2D built from analogous multiphase phase field codes [94, 115, 117]. The computed domain size is $[0, 1] \times [0, 1]$, which we emphasize encompasses the cell and ambient viscous fluid. The number of grid points in each direction for the reported simulations is 256. The parameters used are $k_a = 0.01, K = 0.01, M_1 = 0.1$, $M_2 = 1, k_B T k_b = 1e-9, \lambda_1 = 1, \lambda_2^0 = 0.001, a = 0.8, N = 6, \Gamma = 0.2, \xi_0 = 4$, $\eta_1 = 1, \eta_s = 1, W_0 = 0.01, \epsilon = 0.02$.

5.1 Activation of a Local Domain in the Cortical Layer

In this simulation, we impose the active region on the left side of the cell within the cortical layer. The initial shape of the cell and active region are shown in Fig. 1. As time evolves, the activated region induces a protrusion in the membrane and cortical

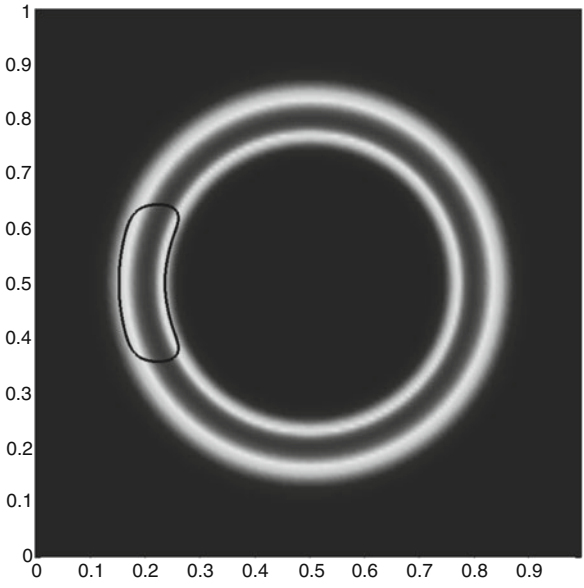

Fig. 1 Initial shape and activation domain are indicated by the contours

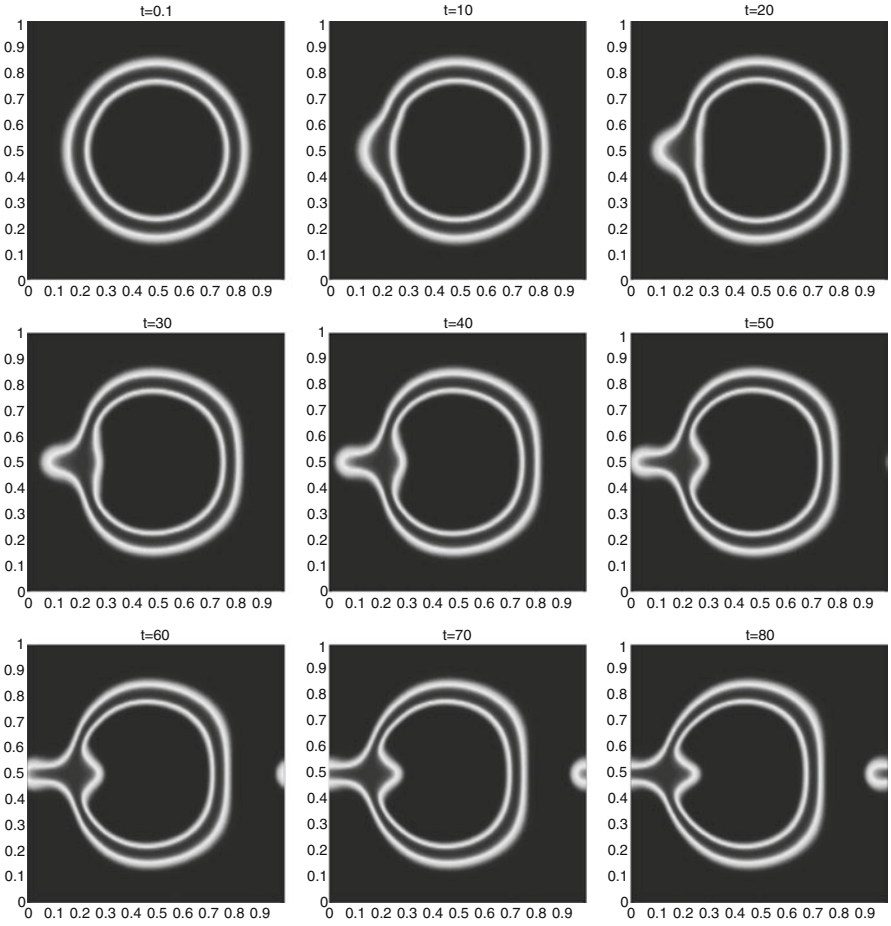

Fig. 2 Snapshots of cell activation and subsequent movement at $t = 0.1, \cdots, 80$. The anchoring condition is not enforced at the membrane ($\alpha_1 = \alpha_2 = 0$). The cell migrates to the direction where the cortex lay is activated

layer due to the activation of the nematic cortical layer. The cell deformation and translational motion are simulated with respect to variations in the energy associated with tangential anchoring conditions in the diffuse interface layer between the membrane and nematic cortical layer: without enforcing an anchoring condition, a weak anchoring condition, and then strong anchoring.

In order to conserve the cell volume, the entire cell undergoes a deformation represented by a passive retraction on the opposite side of the cell, leading to clear cell migration to the left. The activation domain pulls the cell in its direction. Several snapshots are shown in Figs. 2, 3, 4. Figures 5, 6, 7 contrast the cell membrane profile at select times in the interval $t = [0.1, 80]$ to show cell movement. Recall that we track the membrane by the zero level set of the phase variable.

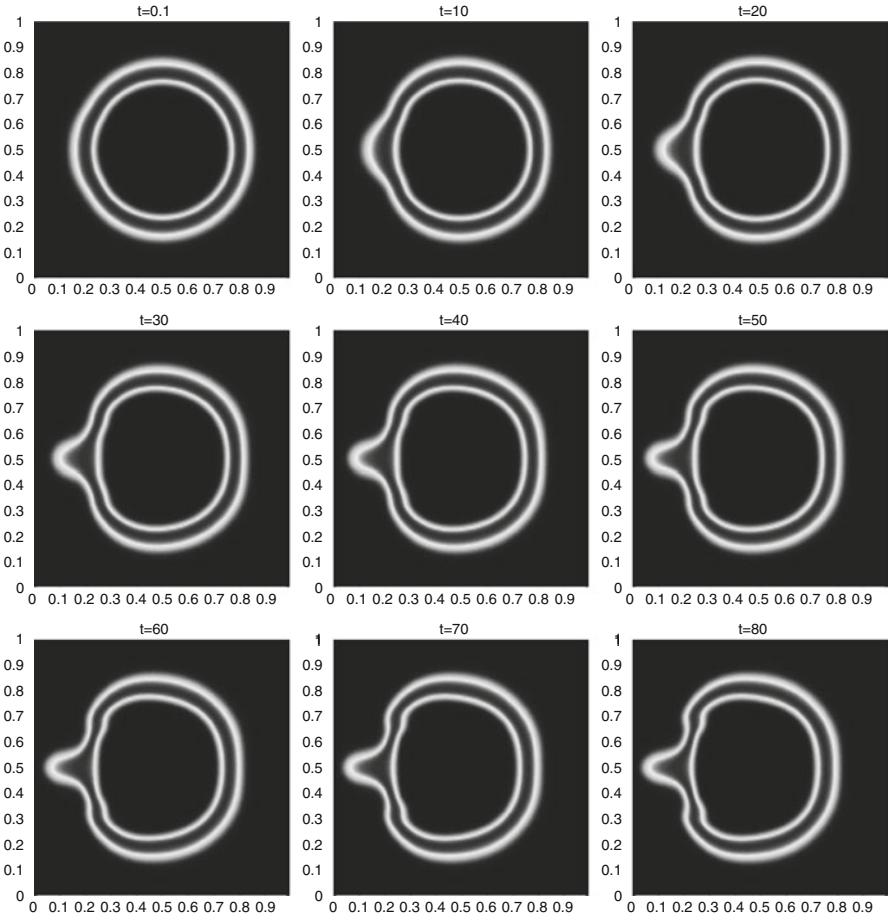

Fig. 3 Snapshots of cell activation and subsequent movement at $t = 0.1, \cdots, 80$. Tangential anchoring energy is enforced in the diffuse interface between the membrane and cortical layer with ($\alpha_1 = 0.1, \alpha_2 = 0$). The cell migrates to the direction where the cortex lay is activated

We first simulate the cell movement under the influence of local activation of the nematic phase in the cortical layer without explicitly enforcing an anchoring boundary condition at the membrane (the diffuse interface). The activation affects both the membrane and the interface between the cortical layer and the interior cytoplasma/cytosol region. Both outward and inward protrusion of the cortical layer are shown in Fig. 2. We then repeat the simulation with the same set of model parameters while allowing for tangential anchoring energy at the membrane. The protrusion is reduced in magnitude. However, the inward invasion nearly disappears while the cell membrane bulges slightly on both sides of the prominent protrusion. This is depicted in Fig. 3 with a few selected snapshots. In the third numerical experiment, we impose the tangential anchoring condition at the membrane with

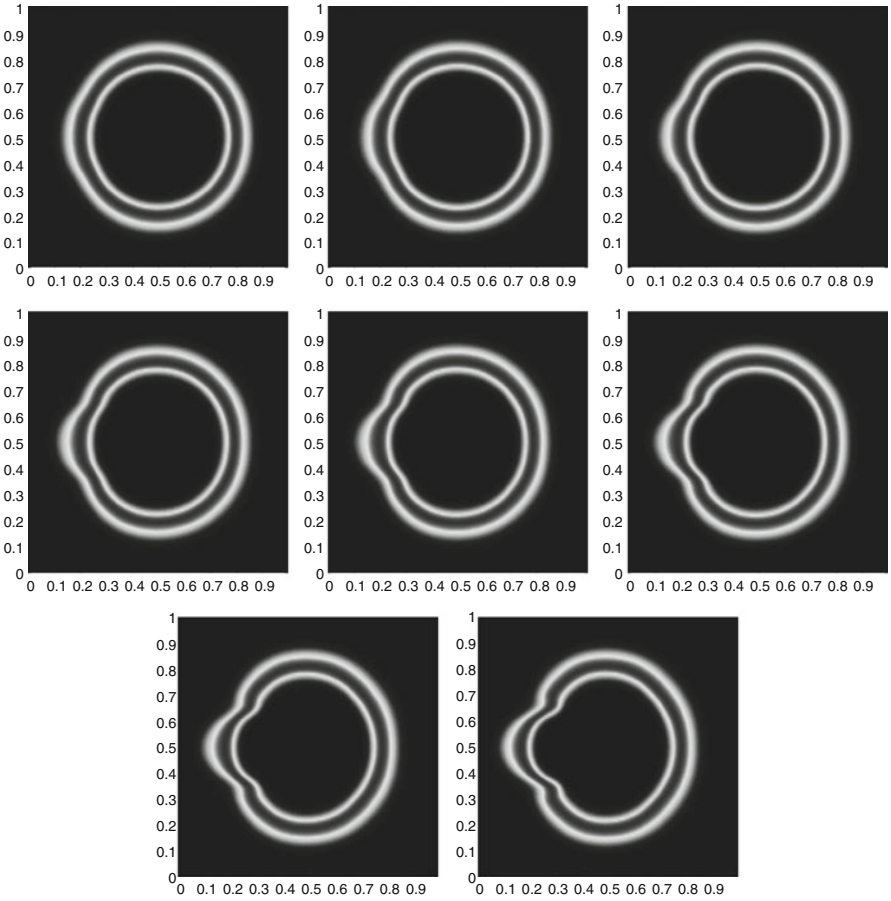

Fig. 4 Snapshots of cell activation and subsequent movement at $t = 0.1, \cdots, 80$. Tangential anchoring energy is enforced in the membrane-cortical layer diffuse interface ($\alpha_1 = 0.5, \alpha_2 = 0$). The cell migrates to the direction where the cortex lay is activated

an enhanced anchoring energy. The resulting deformations of the membrane and cortical layer demonstrate an outward protrusion and a propagation of the cortical layer deformation reminiscent of a slice of a cortical ring contraction wave.

5.2 Active Regions Alternating on Opposing Sides of the Cell

We impose time-dependent activation to two regions located on opposite sides of the cortical layer within the cell membrane. This imposed activation scheme is motivated by the compartment model of [5,58] where there are positive and negative

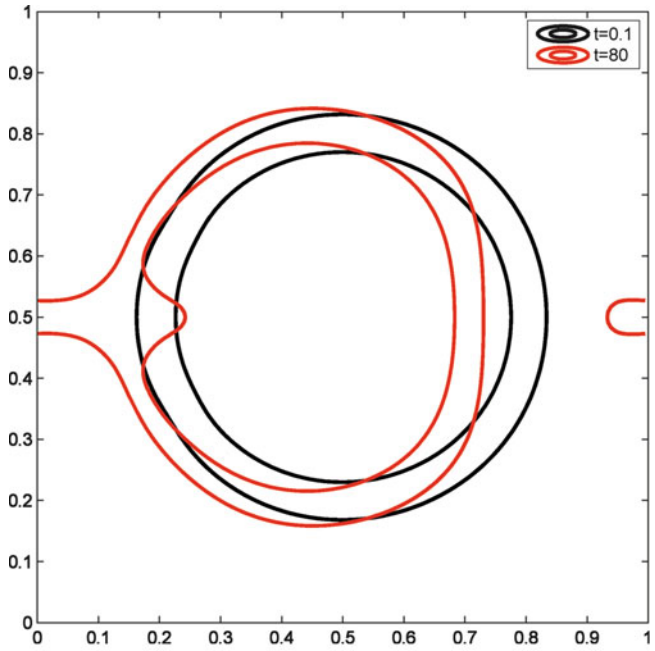

Fig. 5 The profile of the cell conformation at $t = 80$ contrasted with the initial shape at $t = 0$

feedback loops of protein species on either side of the cell. The region on the left is first activated for $t \in (0, 6)$. At $t = 6$, the active region on the left is turned off while an active region on the right is started until the end of the simulation at $t = 40$. The dynamical process is shown in Fig. 8. Due to longer activation at the right, the cell exhibits a protrusion on the right.

This formulation is now amenable to reaction–diffusion of protein species or other components whose concentrations provide the activation potential in the cortical layer. These features are necessary to explore the possible simulation within this framework of the cell oscillation modes identified in the Jacobson lab [20,58,87] and modeled by Allen and Elston [5]. To be biologically useful, many features in these illustrative simulations will need to be based on experimental data. For example, we have not attempted to use consistent cell membrane properties, cortical layer properties, cytosol viscoelastic properties, nor have we introduced a cell nucleus phase. The detailed biochemical species, and their reaction and diffusion rates as well as activation potentials, have to be integrated into the model, as well as constraints for proteins that are bound to the membrane and cortical layer. The addition of substrate boundary conditions instead of an ambient viscous fluid is relatively straightforward to put into the model, yet experimental data on the appropriate surface energies is needed.

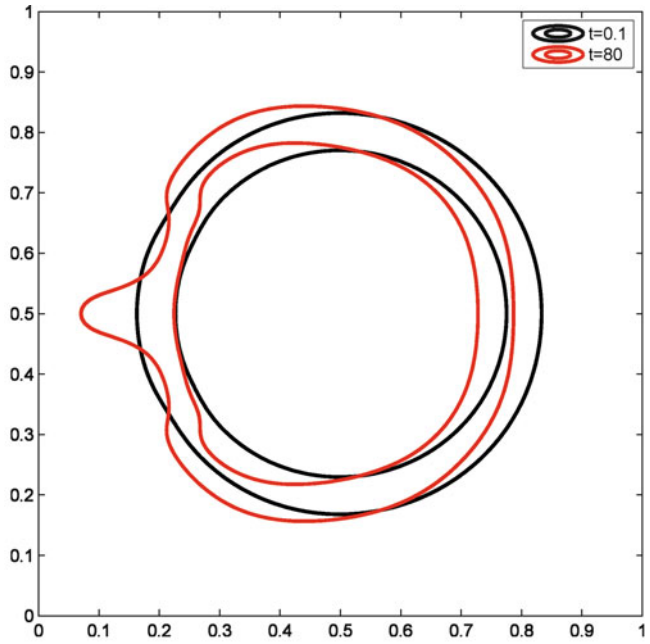

Fig. 6 The profile of the cell conformation at $t = 80$ contrasted with the initial shape at $t = 0.1$ for tangential membrane-cortical layer anchoring energy with $\alpha_1 = 0.1, \alpha_2 = 0$

6 Conclusion

We have surveyed recent theoretical and numerical developments that are relevant to modeling of cell motility. We have integrated many of these advances into a phase field model of the cell with multiple substructures (the ambient fluid, bilayer membrane, nematic cortical layer, and internal cytosol) with an activation potential in the cortical layer that resolves chemical–mechanical transduction. For this chapter, we have imposed the activation domains, amplitudes, and timescales, which in the future will be triggered by biochemical processes. The simulated phase field model exhibits plausible cell morphology dynamics, which are only a cartoon at this point. To make the model and simulations more biologically relevant, we plan to use experimental characterizations of the physical properties of the membrane, cortical layer, cytoplasm, and nucleus, and biochemical kinetics of reacting and diffusing G protein species which trigger activation and deactivation.

Acknowledgments Wang's research is partly supported by National Science Foundation grants CMMI-0819051 and DMS-0908330. Yang's research is supported in part by the army research office (ARO) W911NF-09-1-0389. Forest's research is supported in part by grants NSF DMS-0908423 and DMS-0943851.

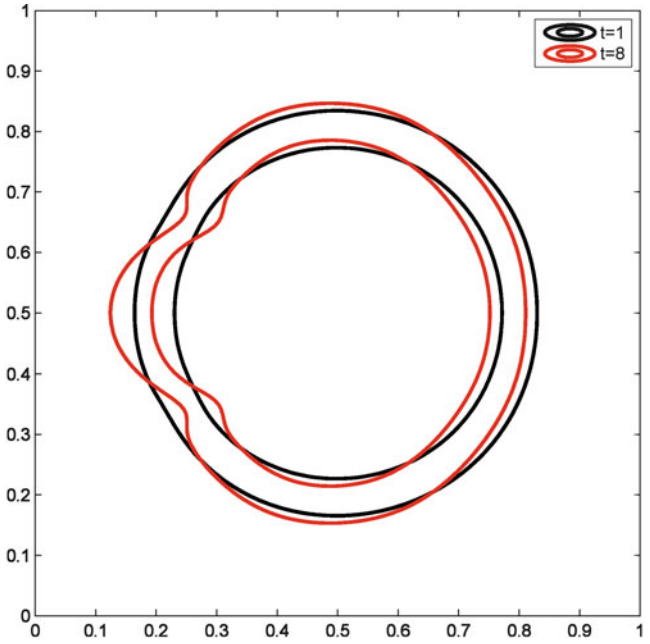

Fig. 7 The profile of the cell conformation at $t = 80$ compared with the shape at $t = 1$ for tangential anchoring in the membrane-cortical layer diffuse interface, where $\alpha_1 = 0.5, \alpha_2 = 0$

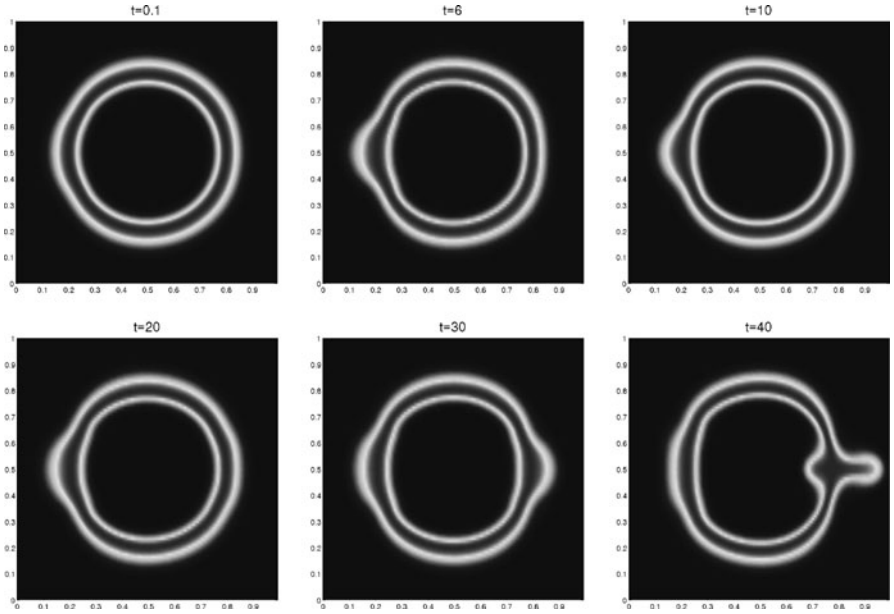

Fig. 8 Activation in the cortical layer on opposing sides of the cell, from $t = 2 - 40$ in equal increments. The active part on the LHS is shut down at $t = 6$ and the RHS is activated for the next 34 time units. Tangential anchoring energy is enforced with $\alpha_1 = 0.1, \alpha_2 = 0$

References

1. Adalsteinsson, D., Elston, T.: www.amath.unc.edu/faculty/Adalsteinsson
2. Alt, W., Dembo, M.: Cytoplasm dynamics and cell motion: two-phase fluid models. Math. Biosci. **156**, 207–228 (1999)
3. Atilgan, E., Wirtz, D., Sun, S.X.: Mechanics and dynamics of actin-driven thin membrane protrusions. Biophys. J. **80**, 65–76 (2006)
4. Ahmadi, A., Marchetti, M.C., Liverpool, T.B.: Hydrodynamics of isotropic and liquid crystalline active polymer solutions, Phys. Rev. E **74**, 061913 (2006)
5. Allen, R., Elston, T.: A compartment model for chemically activated, sustained cellular oscillations, UNC Preprint (2011)
6. Auth, T., Safran, S., Gov, N.: Filament networks attached to membranes: cytoskeletal pressure and local bilayer deformation. New J. Phys. **9**, 430–444 (2007)
7. Banerjee, S., Marchetti, M.C.: Instability and oscillations in isotropic gels. Soft Matter **7**, 463–473 (2011)
8. Baskaran, A., Marchetti, M.C.: Hydrodynamics of self-propelled hard rods, Phys. Rev. E **77**, 031311 (2008)
9. Baskaran, A., Marchetti, M.C.: Nonequilibrium statistical mechanics of self-propelled hard rods. J. Stat. Mech. Theor. Exp. **4**, 04019 (2010)
10. Besser, A., Schwarz, U.S.: Coupling biochemistry and mechanics in cell adhesion: a model for inhomogeneous stress fiber contraction. New J. Phys. **9**, 425 (2007)
11. Bird, B., Armstrong, R., Hassager, O.: Dynamics of Polymeric Liquids, 2nd edn., Vol. 2. Wiley, New York (1987)
12. Bershadsky, A., Kozlov, M., Geiger, B.: Adhesion-mediated mechanosensitivity: a time to experiment, and a time to theorize. Curr. Opin. Cell Biol. **18**, 472–481 (2006)
13. Boal, D.: Mechanics of the Cell. Cambridge University Press, New York (2002)
14. Carlsson, A.: Growth velocities of branched actin networks. Biophys. J. **84**, 2907–2918 (2003)
15. Cates, M.E., Fielding, S.M., Marenduzzo, D., Orlandini, E., Yeomans, J.M.: Shearing active gels close to the isotropic-nematic transition. Phys. Rev. Lett. **101**, 068102 (2008)
16. Chen, C., Ren, M., Srinivasan, A., Wang, Q.: 3-D simulations of biofilm-solvent interaction. Asian J. Appl. Math. **1**, 197–214 (2011)
17. Chen, L.Q., Yang, W.: Computer simulation of the dynamics of a quenched system with large number of non-conserved order parameters. Phys. Rev. B **50**, 15752–15756 (1994)
18. Chen, L.Q.: Phase-field modeling for microstructure evolution. Annu. Rev. Mater. Res. **32**, 113–140 (2002)
19. Chen, L.Q., Wang, Y.: The continuum field approach to modeling microstructural evolution. J. Miner Met. Mater. Soc. **48**, 13–18 (1996)
20. Costigliola, N., Kapustina, M., Weinreb, G., Monteith, A., Rajfur, Z., Elston, T., Jacobson, K.: Rho regulates calcium independent periodic contractions of the cell cortex. Biophys. J. **99**(4), 1053–1063 (2010)
21. Cui, Z., Wang, Q.: Dynamics of chiral active liquid crystal polymers. DCDS-B **15**(1), 45–60 (2011)
22. Curtis, A., Wilkinson, C.: Nanotechniques and approaches in biotechnology. Trends Biotechnol. **19**, 97–101 (2001)
23. De-Gennes, P.G., Prost, J.: The Physics of Liquid Crystals. Oxford Science Publications, Oxford (1993)
24. DeMali, K.A., Barlow, C.A., Burridge, K.: Recruitment of the Arp2/3 complex to vinculin: coupling membrane protrusion to matrix adhesion. J. Cell Biol. **159**, 881–891 (2002)
25. Deshpande, V.S., McMeeking, R.M., Evans, A.G.: A model for the contractibility of the cytoskeleton incldluing the effects of stress-fiber formation and dissociation. Proc. Roy. Soc. A **463**, 787–815 (2007)
26. DiMilla, P.A., Barbee, K., Lauffenburger, D.: Mathematical model for the effects of adhesion and mechanics on cell migration speed. Biophys. J. **60**, 15–37 (1991)

27. Doi, M., Edwards, S.F.: The Theory of Polymer Dynamics. Oxford University Press, Oxford (1986)
28. Doherty, G.J., McMahon, H.T.: Mediation, modulation, and consequences of membrane-cytoskeleton interactions. Annu. Rev. Biophys. **37**, 65–95 (2008)
29. Du, Q., Liu, C., Ryham, R., Wang, X.: Phase field modeling of the spontaneous curvature effect in cell membranes. Comm. Pur. Applied. Anal. **4**, 537–548 (2005)
30. Du, Q., Liu, C., Ryham, R., Wang, X.: A phase field formulation of the Willmore problem. Nonlinearity **18**, 1249–1267 (2005)
31. Du, Q., Liu, C., Ryham, R., Wang, X.: Energetic variational approaches in modeling vesicle and fluid interactions. Physica D **238**, 923–930 (2009)
32. Du, Q., Liu, C., Wang, X.: A phase field approach in the numerical study of the elastic bending energy for vesicle membranes. J. Comp. Phy. **198**, 450–468 (2004)
33. Du, Q., Liu, C., Wang, X.: Retrieving topological information for phase field models. SIAM J. Appl. Math. **65**, 1913–1932 (2005)
34. Du, Q., Liu, C., Wang, X.: Simulating the deformation of vesicle membranes under elastic bending energy in three dimensions. J. Comp. Phys. **212**, 757–777 (2006)
35. Elston, T., Allen, R., Kapustina, M., Jacobson, K.: A compartment chemical-mechanical model for sustained cellular oscillations, University of North Carolina at Chapel Hill preprint (2011)
36. Feng, J.J., Liu, C., Shen, J., Yue, P.: Transient drop deformation upon startup of shear in viscoelastic fluids, fluids. Phys. Fluids **17**, 123101 (2005)
37. Forest, M.G., Liao, Q., Wang, Q.: 2-D kinetic theory for polymer particulate nanocomposites. Comm. Comput. Phys. **7**(2), 250–282 (2010)
38. Forest, M.G., Wang, Q.: Hydrodynamic theories for blends of flexible polymer and nematic polymers. Phys. Rev. E **72**, 041805 (2005)
39. Funkhouser, C.M., Solis, F.J., Thornton, K.: Coupled composition-deformation phase-field method for multicomponent lipid membranes. Phys. Rev. E **76**, 011912 (2007)
40. Funkhouser, C.M., Solis, F.J., Thornton, K.: Dynamics of two-phase lipid vesicles: effects of mechanical properties on morphology evolution. Soft Matter **6**, 3462–3466 (2010)
41. Frixione, E.: Recurring views on the structure and function of the cytoskeleton: a 300-year epic. Cell Motil. cytoskeleton **46**(2), 73–94 (2000)
42. Gerisch, G., Bretschneider, T., Muller-Taubenberger, A., Simmeth, E., Ecke, M., Diez, S., Anderson, K.: Mobile actin clusters and traveling waves in cells recovering from actin depolymerization. Biophys. J. **87**(5), 3493–3503 (2004)
43. Giannone, G., Dubin-Thaler, B.J., Rossier, O., Cai, Y., Chaga, O., Jiang, G., Beaver, W., Dobereiner, H.-G., Freund, Y., Borisy, G., Sheetz, M.P.: Lamellipodial actin mechanically links myosin activity with adhesion-site formation. Cell **128**(3), 561–575 (2007)
44. Giomi, L., Liverpool, T.B., Marchetti, M.C.: Sheared active fluids: thickening, thinning, and vanishing viscosity. Phys. Rev. E **81**, 051908 (2010)
45. Gopinathan, A., Lee, K.-C., Schwarz, J.M., Liu, A.J., Branching, capping, and severing in dynamic actin structures. Phys. Rev. Lett. **99**, 058103 (2007)
46. Hatwalne, Y., Ramaswamy, S., Rao, M., Simha, R.A.: Rheology of active-particle suspensions. Phys. Rev. Lett. **93**, 198105 (2004)
47. Hobayashi, R.: Modeling and numerical simulations of dendritic crystal growth. Physica D **63**, 410–423 (1993)
48. Hoffman, B., Crocker, J.: Cell mechanics: dissecting the physical responses of cells to force. Annu. Rev. Biomed. Eng. **11**, 259–288 (2009)
49. Horwitz, R., Parsons, J.: Cell migration-moving on. Science **286**, 1102–1103 (1999)
50. Horwitz, R., Webb, D.: Cell migration. Curr. Biol. **13**, R756-9 (2003)
51. Hu, L., Papoian, G.A.: Mechano-chemical feedbacks regulate actin mesh growth in lamellipodial protrusions. Biophys. J. **98**, 1375–384 (2010)
52. Hua, J., Lin, P., Liu, C., Wang, Q.: Energy law preserving C^0 finite element schemes for phase field models in two-phase flow computations. J. Comp. Phys. **230**(19), 7115–7131 (2011)

53. Jeong, J., Goldenfeld, N., Dantzig, J.: Phase field model for three-dimensional dendritic growth with fluid flow. Phys. Rev. E **64**, 041602 (2001)
54. Jiang, X., Takayama, S., Qian, X., Ostuni, E., Wu, H., Bowden, N., LeDuc, P., Ingber, D.E., Whitesides, G.M.: Controlling mammalian cell spreading and cytoskeletal arrangement with conveniently fabricated continuous wavy features on poly(dimethylsiloxane). Langmuir **18**, 3273–3280 (2002)
55. Joanny, J.F., Julicher, F., Kruse, K., Prost, J.: Hydrodynamic theory for multi-component active polar gels. New J. Phys. **9**, 1–17 (2007)
56. Joanny, J.F., Julicher, F., Prost, J.: Motion of an adhesive gel in a swelling gradient: a mechanicsm for cell locomotion. Phys. Rev. Lett. **25**(6), 168102 (2003)
57. Julicher, F., Kruse, K., Prost, J., Joanny, J.-F.: Active behavior of the cytoskeleton. Phys. Rep. **449**, 3–28 (2007)
58. Kapustina, M., Weinreb, G., Costigliola, N., Rajfur, Z., Jacobson, K., Elston, T.: Mechanical and biochemical modeling of cortical oscillations in spreading cells. Biophys. J. **94**(12), 4605–4620 (2008)
59. Karma, A., Rappel, W.: Phase-field model of dendritic sidebranching with thermal noise. Phys. Rev. E **60**, 3614–3625 (1999)
60. Kataoka, A., Tanner, B.C.W., Macpherson, J.M., Xu, X., Wang, Q., Reginier, M., Daniel, T., Chase, P.B.: Spatially explicit, nanomechanical models of the muscle half sarcomere: implications for mechanical tuning in atrophy and fatigue. Acta Astronautica **60**(2), 111–118 (2007)
61. Kiehart, D.P., Bloom, K.: Cell structure and dynamics. Curr. Opin. Cell Biol. **19**, 1–4 (2004)
62. Kim, J., Sun, S.: Continuum modeling of forces in growing viscoelastic cytosketal networks. J. Theor. Biol. **256**, 596–606 (2009)
63. Kruse, K., Joanny, J.F., Julicher, F., Prost, J., Seimoto, K.: Asters, vortices, and rotating spirals in active gels of polar filaments: Phys. Rev. Lett. **92**(7), 078101 (2004)
64. Kruse, K., Joanny, J.F., Julicher, F., Prost, J., Sekimota, K.: Generic theory of active polar gels: a paradigm for cytoskeletal dynamics. Eur. Phys. J. E **16**, 5–16 (2005)
65. Kruse, K., Julicher, F.: Actively contracting bundles of polar filaments. Phys. Rev. Lett. **85**(8), 1778–1781 (2000)
66. Li, J., Forest, M.G., Wang, Q., Zhou, R.: A kinetic theory and benchmark predictions for polymer dispersed, semi-flexible nanorods and nanoplatelets. Physica D **240**, 114–130 (2011)
67. Li, Y., Hu, S., Liu, Z., Chen, L.Q.: Phase-field model of domain structures in ferroelectric thin films. Appl. Phys. Lett. **78**, 3878–3880 (2001)
68. Lindley, B., Wang, Q., Zhang, T.: Multicomponent models for biofilm flows. Discrete Continuous Dyn. Syst. Ser. B **15**(2), 417–456 (2011)
69. Liu, C., Shen, J.: A phase field model for the mixture of two incompressible fluids and its approximation by a fourier-spectral method. Physica D **179**, 211–228 (2003)
70. Liu, C., Walkington, N.J.: An Eulerian description of fluids containing visco-hyperelastic particles. Arch. Rat. Mech. Ana. **159**, 229–252 (2001)
71. Liverpool, T.B., Marchetti, M.C.: Bridging the microscopic and the hydrodynamic in active filament solutions. Europhys. Lett. **69**, 846 (2005)
72. Liverpool, T.B., Marchetti, M.C.: Hydrodynamics and rheology of active polar filaments. In: Lenz, P. (ed.) Cell Motility. Springer, NY (2007)
73. Loesberg, W.A., te Riet, J., van Delft, F.C.M.J.M., Schoen, P., Figdor, C.G., Speller, S., van Loon, J.J.W.A., Walboomers, X.F., Jansen, J.A.: The threshold at which substrate nanogroove dimensions may influence fibroblast alignment and adhesion. Biomaterials **28**(27), 3944–3951 (2007)
74. Lowengrub, J., Truskinovsky, L.: Quasi-incompressible Cahn-Hilliard fluids and topological transitions. R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. **454**, 2617–2654 (1998)
75. Lu, W., Suo, Z.: Dynamics of nanoscale pattern formation of an epitaxial monolayer. J. Mech. Phys. Solids **49**, 1937–1950 (2001)

76. Marenduzzo, D., Orlandini, E., Cates, M.E., Yeomans, J.M.: Steady-state hydrodynamic instabilities of active liquid crystals: hybrid lattice Boltzmann simulations. Phys. Rev. E **76**, 031921 (2007)
77. McFadden, G., Wheeler, A., Braun, R., Coriell, S., Sekerka, R.: Phase-field models for anisotropic interfaces. Phys. Rev. E **4**, 2016–2024 (1993)
78. Michie, K., Lowengrub, J.: Dynamic filaments of the bacterial cytoskeleton. Annu. Rev. Biochem. **75**, 467–492 (2006)
79. Mofrad, M., Kamm, R.: Cytoskeletal Mechanics: Models and Measurements. Cambridge University Press, Cambridge (2006)
80. Mogilner, A.: On the edge: modeling protrusion. Curr. Opin. Cell Biol. **18**, 32–39 (2006)
81. Muhuri, S., Rao, M., Ramaswamy, S.: Shear flow induced isotropic to nematic transition in a suspension of active filaments. Europhysics Lett. **78**, 48002 (2007)
82. Murray, J.: Mathematical Biology. Springer, Heidelberg (1989)
83. Nguyen, L., Yang, W., Wang, Q., Hirst, L.: Molecular dynamics simulation of F-actin reveals the role of cross-linkers in semi-flexible filament. Soft Matter **5**, 2033–2036 (2009)
84. Oster, G., Perelson, A.: Cell spreading and motility: a model lamellipod. J. Math. Biol. **21**, 383–388 (1985)
85. Paluch, E., Piel, M., Sykes, C.: Cortical actomyosin breakage triggers shape oscillations in cells and cell fragments. Biophys. J. **89**, 724–733 (2005)
86. Parent, C., Devreotes, P.: A cell's sense of direction. Science **284**, 765–70 (1999)
87. Pletjushkina, O., Rajfur, Z., Pamorski, P., Oliver, T., Vasiliev, J., Jacobson, K.: Induction of cortical oscillations in spreading cells by depolymerization of microtubules. Cell Mot. Cytoskeleton **48**(4), 235–244 (2001)
88. Pollard, T.D., Borisy, G.G.: Cellular motility driven by assembly and disassembly of actin filaments. Cell **112**, 453–465 (2003)
89. Rafelski, S.M., Theriot, J.A.: Crawling toward a unified model of cell motility: spatial and temporal Regulation of actin dynamics. Annu. Rev. Biochem. **73**, 209–239 (2004)
90. Ridley, A.J., Schwartz, M.A., Burridge, K., Firtel, R.A., Ginsberg, M.H., Borisy, G., Parsons, J.T., Horwitz, A.R.: Cell migration: integrating signals from front to back. Science **302**, 1704–1709 (2003)
91. Saintillan, D., Shelley, M.: Instabilities and pattern formation in active particle suspensions: kinetic theory and continuum simulations. Phys. Rev. Lett. **100**, 178103 (2008)
92. Salbreux, G., Joanny, J.F., Prost, J., Pullarkat, P.: Shape oscillation of non-adhering fibroblast cells. Phys. Biol. **4**, 268–284 (2007)
93. Seol, D.J., Hu, S.Y., Li, Y.L., Shen, J., Oh, K.H., Chen, L.Q.: Three-dimensional phase-field modeling of spinodal decomposition in constrained films. Acta Materialia **51**, 5173–5185 (2003)
94. Shen, J., Yang, X.: An efficient moving mesh spectral method for the phase-field model of two phase flows. J. Comput. Phys. **228**, 2978–2992 (2009)
95. Shih, Y.L., Rothfield, L.: The bacterial cytoskeleton, Microbiol. Mol. Biol. Rev. **70**(3), 729–754 (2006)
96. Simha, R.A., Ramaswamy, S.: Hydrodynamic fluctuation and instabilities in ordered suspension of self-propelled particles. Phys. Rev. Lett. **89**(5), 058101 (2002)
97. Stachowiak, M.R., O'Shaughnessy, B.: Kinetics of stress fibers. New J. Phys. **9**, 025002 (2007)
98. Stephanou, A., Chaplain, M.A.J., Tracqui, P.: A mathematical model for the dynamics of large membrane deformation of isolated fibroblasts. Bull. Math. Biol. **66**, 1119–1154 (2004)
99. Stephanou, A., Mylona, E., Chaplain, M., Tracqui, P.: A computational model of cell migration coupling the growth of focal adhesions with oscillatory cell pretrusion. J. Theor. Biol. **253**, 701–716 (2008)
100. Tadmor, E., Phillips, R., Ortiz, M.: Mixed atomistic and continuum models of deformation in solids. Langmuir **12**, 4529–4534 (1996)
101. Van Haastert, P.J., Devreotes, P.N.: Chemotaxis: signalling the way forward. Nat. Rev. Mol. Cell. Biol. **5**(8), 626–634 (2004)

102. Vicente-Manzanares, M., Webb, D.J., Horwitz, A.R.: Cell migration at a glance. J. Cell Sci. **118**, 4917–4919 (2005)
103. Wang, Q.: A hydrodynamic theory of nematic liquid crystalline polymers of different configurations. J. Chem. Phys. **116**, 9120–9136 (2002)
104. Wang, Y., Chen, C.L.: Simulation of microstructure evolution. In: Ksufmann, E.N., Abbaschian, R., Bocarsly, A., Chien, C.L., Dollimore, D., et al. (eds.) Methods in Materials Research, 2a3.1–2a3.23, Wiley, New York (1999)
105. Wang, X., Du, Q.: Modelling and simulations of multi-component lipid membranes and open membranes via diffuse interface approaches. J. Math. Biol. **56**, 347–371 (2008)
106. Wang, Q., E, W., Liu, C., Zhang, P.: Kinetic theories for flows of nonhomogeneous rodlike liquid crystalline polymers with a nonlocal intermolecular potential. Phys. Rev. E **65**(5), 0515041–0515047 (2002)
107. Wang, Q., Forest, M.G., Zhou, R.: A hydrodynamic theory for solutions of nonhomogeneous nematic liquid crystalline polymers with density variations. J. Fluid Eng. **126**, 180–188 (2004)
108. Wang, Q., Zhang, T.Y.: Kinetic theories for biofilms. Discrete Continuous Dyn. Syst. Ser. B in press (2011)
109. Weinreb, G., Kapustina, M., Jacobson, K., Elston, T.: In silico hypothesis generation using casual mapping (CMAP). PLoS One **4**, e5378 (2009)
110. Wheeler, A., McFadden, G., Boettinger, W.: Phase-field model for solidification of a eutectic alloy. Proc. R. Soc. London Ser. A **452**, 495–525 (1996)
111. Wise, S.M., Lowengrub, J.S., Kim, J.S., Johnson, W.C.: Efficient phase-field simulation of quantum dot formation in a strained heteroepitaxial film. Superlattice Microst. **36**, 293–304 (2004)
112. Wolgemuth, C.W.: Lamellipodial contractions during crawling and spreading. Biophys. J. **89**(3), 1643–1649 (2005)
113. Wolgemuth, C.W., Mogilner, A., Oster, G.: The hydration dynamics of polyelectrolyte gels with applications to cell motility and drug delivery. Eur. Biophys. J. **33**, 146–158 (2004)
114. Yang, X., Feng, J., Liu, C., Shen, J.: Numerical simulations of jet pinching-off and drop formation using an energetic variational phase-field method. J. Comput. Phys. **218**, 417–428 (2006)
115. Yang, X., Forest, M.G., Shen, J., Liu, C.: Shear cell rupture of liquid crystal droplets in a viscous fluid. J. Non-Newtonian Fluid Mech. **166**, 487–499 (2011)
116. Young, J.J.: Cytoskeleton micromechanics: a continuum-microscopic approach, Dissertation in Mathematics, UNC Chapel Hill, advised by S. Mitran (2010)
117. Yue, P., Feng, J.J., Liu, C., Shen, J.: A diffuse-interface method for simulating two-phase flows of complex fluids. J. Fluid Mech. **515**, 293–317 (2004)
118. Yue, P., Feng, J.J., Liu, C., Shen, J.: Diffuse-interface simulations of drop coalescence and retraction in viscoelastic fluids. J. Non-Newtonian Fluid Mech. **129**, 163–176 (2005)
119. Zhang, T.Y., Cogan, N., Wang, Q.: Phase field models for biofilms. II. 2-D Numerical simulations of biofilm-flow interaction. Comm. Comput. Phys. **4**, 72–101 (2008)
120. Zhang, J., Das, S., Du, Q.: A phase field model for vesicle-substrate adhesion. J. Comput. Phys. **228**, 7837–7849 (2009)
121. Zhang, T.Y., Wang, Q.: Cahn-Hilliard vs singular Cahn-Hilliard equations in phase field modeling. Comm. Comput. Phys. **7**(2), 362–382 (2010)

Theoretical Analysis of Molecular Transport Across Membrane Channels and Nanopores

Anatoly B. Kolomeisky

1 Introduction

A successful functioning of cellular systems requires that some molecules and ions be transferred out of the cell while other particles should be taken in. The bidirectional flux is accomplished with the help of a complex system of membrane protein channels and pores [1, 2]. It is known that molecular transport across cellular membranes is fast, efficient, selective, and that the functioning of channels is robust with respect to strong nonequilibrium fluctuations in the cellular environment [2]. These observations are especially surprising because in many cases molecular translocation does not involve the use of metabolic energy or significant conformational changes [4]. Although in recent years significant advances in studying molecular transport in biological systems have been achieved, the mechanisms of translocation phenomena are still not well understood.

To develop a comprehensive picture of molecular transport across the membrane, one has to recall that when a molecule enters into the pore its motion is slowed mostly due to entropic barriers and possibly due to other biochemical interactions. Additional forces are needed to overcome these barriers. In biological systems, electric fields, concentration gradients, and chemical interactions are used to speed up the transport. There is increasing experimental evidence that the high efficiency, robustness, speed, and selectivity of many biological and artificial channels are a result of complex processes that involve molecule/pore and intermolecular interactions [3–17]. Recent high-resolution experiments on polypeptide translocations through protein nanopores [14, 15] have opened new possibilities in probing the effect of molecule/pore interactions at the single-molecule level. In these experiments, it was found that changing the location of the binding site in the pore significantly modified

A.B. Kolomeisky (✉)
Department of Chemistry, Rice University, Houston, TX 77005, USA
e-mail: tolya@rice.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_12, © Springer Science+Business Media, LLC 2012

the polypeptide flux across the channel. However, a comprehensive description of the role of interactions in transport through the pores is still not available due to the biochemical and biophysical complexity of the translocation machinery and the lack of structural information [4, 14, 15].

Experimental advances in the investigation of biological transport phenomena have stimulated significant theoretical efforts to describe the fundamental properties of underlying processes [18–31]. Several theoretical methods have been utilized to uncover mechanisms of the motion through cellular and artificial nanopores. Significant efforts have been devoted to developing computational Molecular Dynamics (MD) approaches to calculate translocation dynamics by taking into account all atomic and molecular interactions [18, 19]. However, biological transport systems are so large and complex that current full-atomic MD simulations techniques allow analysis of dynamics for very short times of $\sim$1 ns, while the relevant processes are taking place in the timescales of seconds and minutes. These observations suggest that coarse-grained computer simulations and phenomenological analytical methods probably are more valuable for understanding biological transport mechanisms.

Existing phenomenological approaches mainly follow two directions. Continuum models of channel transport view the translocation as one-dimensional motion in an effective potential created by interactions with nanopores and with other molecules [22–27]. Interactions are typically modeled as square well potentials that occupy the volume of the pore. A different theoretical picture utilizes discrete-state stochastic models in which the translocation dynamics is analyzed as a set of chemical transitions between specific binding sites in the channel [28–31]. By mapping the discrete-state model of molecular transport across the channel to a single-particle hopping along a periodic lattice, a full dynamic description of permeation through the pore can be obtained for arbitrary sets of parameters [28,29]. Theoretical calculations also show that both continuum and discrete approaches are closely related, and the results obtained by these approaches can be mapped into each other [24, 26].

In this chapter, we present a theoretical analysis of membrane channels translocation phenomena utilizing discrete-state stochastic models. Specific attention is devoted to explaining physical/chemical mechanisms that control transport through channels and nanopores.

2 Discrete-State Stochastic Models

The main idea of the phenomenological approach based on discrete-state stochastic models is that the molecular translocation can be described by a one-dimensional free-energy profile with minimal positions corresponding to specific binding sites. Transport of molecules through the nanopore is considered as an effective one-dimensional motion along the discrete lattice of these binding sites, as illustrated in Fig. 1. There are N binding sites in the channel, and the concentrations of molecules to the left or right of the channel are equal to c_1 and c_2, respectively. The molecule

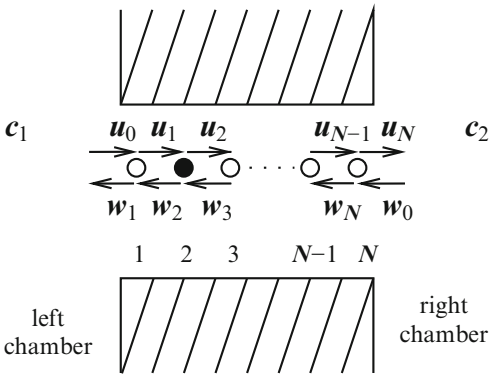

Fig. 1 A general schematic view of the discrete stochastic model of channel-facilitated transport with N binding sites. A cylindrical membrane channel divides the system into three parts: the *left chamber* with particle concentration c_1, the *right chamber* with particle concentration c_2, and the pore that can be occupied by a single particle. *Open circles* correspond to available binding sites in the channel. The *filled circle* describes the position that is currently occupied by the particle

can move into the channel from the left (right) with rate $u_0 = k_{on}c_1$ ($w_0 = k_{on}c_2$), and the particle can move out of the channel with rates w_1 and u_N—see Fig. 1. In the nanopore, the molecule at the site j ($j = 1, 2, \cdots, N$) can jump forward (backward) with the rate u_j (w_j). It is assumed that the molecular size is comparable to the size of the binding site. In the case of polymer translocations, when the size of the molecule is larger than the binding site, this theoretical method can also be extended by properly taking into account free-energy contributions from the corresponding polymer configurations [32].

2.1 Molecule/Channel Interactions

First, we consider the situation when only one particle can be found in the channel. The probability to find a molecule at site j at time t is given by a function $P_j(t)$, and the translocation dynamics is fully described by a set of master equations,

$$\frac{\mathrm{d}P_j(t)}{\mathrm{d}t} = u_{j-1}P_{j-1}(t) + w_{j+1}P_{j+1}(t) - (u_j + w_j)P_j(t), \tag{1}$$

for $j = 1, \cdots, N$; while $P_0(t) \equiv P_{N+1}(t) = 1 - \sum_{j=1}^{N} P_j(t)$ describes the completely empty channel at the time t [28, 29]. It has been shown [28, 29] that this model with N binding sites can be solved exactly at $t \to \infty$ by mapping it into a single-particle random walk model on an infinite periodic lattice (with a period equal to $N + 1$). The size of the period is equal to $N + 1$ because there are N states inside the channel and one state outside of the channel [28, 29]. This mapping can be understood by considering multiple identical channels arranged sequentially and

keeping the constant concentration gradient across each period as $\Delta c = c_1 - c_2$ [28]. Thus, all dynamic properties of molecular transport across nanopores can be calculated explicitly. Specifically, for the uniform channel with a simplifying assumption of zero particle concentration to the right of the pore ($w_0 = 0$) the expression for the particle current is given by

$$J_0(N) = \frac{uu_0}{(N+1)\left(u + \frac{N}{2}u_0\right)}, \tag{2}$$

where $u_j = w_j = u$ ($j = 1, \cdots, N$). One can also write down the corresponding expressions for the particle dispersion [29].

To quantify the effect of interactions, we assume that in one of the binding sites, say the k-th, the particle interacts with the pore with potential ε that differs from other sites. The case of $\varepsilon > 0$ corresponds to attractive interactions at this special site, while negative ε describes the repulsive special binding site. The transition rates near the special binding site must satisfy the detailed balance conditions [28, 29], which lead to

$$\frac{u'_{k-1}}{w'_k} = \frac{u_{k-1}}{w_k}x, \quad \frac{u'_k}{w'_{k+1}} = \frac{u_k}{w_{k+1}}\frac{1}{x}, \tag{3}$$

where u_{k-1}, u_k, w_k, and w_{k+1} correspond to the uniform channel without special interactions, and we define $x = \exp(\varepsilon/k_B T)$. The corresponding explicit expressions for transition rates can now be written as [33],

$$u'_{k-1} = u_{k-1}x^{\theta}, \quad u'_k = u_k x^{\theta-1}, \quad w'_k = w_k x^{\theta-1}, \quad w'_{k+1} = w_{k+1}x^{\theta}, \tag{4}$$

where the coefficient θ ($0 \le \theta \le 1$) describes how the interaction potential modifies the corresponding transition rates [28, 29, 33]. Now, flux through the channel with the special binding site at position k is equal to

$$J_k(N) = \frac{uu_0}{u\left[2x^{-\theta}+N-1\right]+u_0\left[2(k-1)x^{-\theta}+x^{1-\theta}+(N-k)x+\frac{N(N-1)}{2}-k+1\right]}. \tag{5}$$

The effect of interactions can be better understood by analyzing the dimensionless ratio of particle currents,

$$\frac{J_k(N)}{J_0(N)} = \frac{(N+1)\left[(u/u_0)+\frac{N}{2}\right]}{(u/u_0)\left[2x^{-\theta}+N-1\right]+\left[2(k-1)x^{-\theta}+x^{1-\theta}+(N-k)x+\frac{N(N-1)}{2}-k+1\right]}, \tag{6}$$

where $J_0(N)$ is the current for the uniform channel without interactions. This function provides a convenient measure of how the spatial positioning of the special interaction binding site changes the particle current.

The curves presented in Fig. 2 show the effect of the special binding site location on particle fluxes through the channels. For attractive interactions, the most optimal

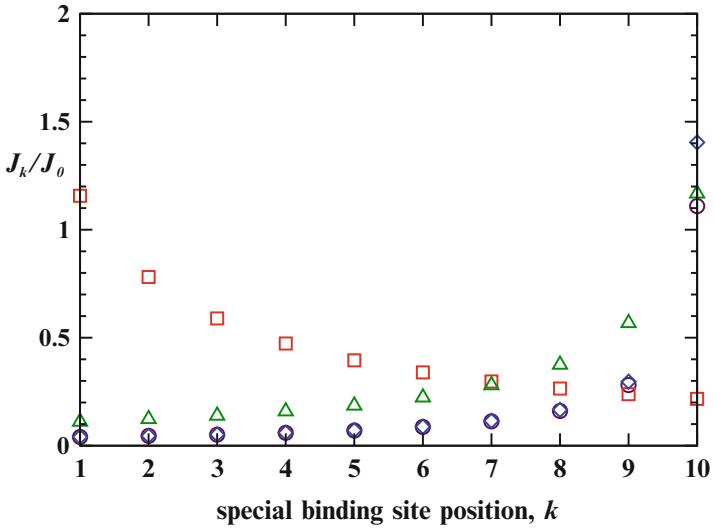

Fig. 2 The ratio of molecular fluxes for different positions of the special binding site k for the channel with $N = 10$ binding sites. *Circles* are for $\varepsilon/k_BT = 5, u/u_0 = 0.1$, and $\theta = 0.5$. *Squares* are for $\varepsilon/k_BT = -5, u/u_0 = 0.1$, and $\theta = 0.5$. *Triangles* are for $\varepsilon/k_BT = 5, u/u_0 = 10$, and $\theta = 0.5$. *Diamonds* are for $\varepsilon/k_BT = 5, u/u_0 = 0.1$, and $\theta = 0.0$

current is reached when the binding site is at the exit from the pore ($k = N$), while it is better to have the repulsive site closer to the entrance ($k = 1$) to accelerate the transport. It can be shown rigorously from (6) that $\frac{\partial J_k(N)}{\partial k} > 0$ for positive ε, and $J_k(N)$ is always a decreasing function for negative ε.

These observations can be understood in the following way. Putting the attractive binding site near the exit increases the probability of finding the particle there, which leads to higher chances to complete the translocation by exiting the nanopore. The repulsive site at the entrance serves as a barrier for the particles that have already passed it, lowering the probability of unsuccessful excursions without the translocation. These results are in full agreement with recent single-molecule experiments on translocation of polypeptides [14, 15]. In these experiments, the mutation in the biological nanopore that increased the molecule/pore interaction have led to faster transport when the mutation site was near the exit. These theoretical results might also shed the light on experimental observations, showing that many biological channels have their binding sites at the entrance and/or at the exit positions, that have not been fully understood so far. To have special binding at these locations will optimize the overall flux [34]. These results can be easily extended to more complex potentials with several attractive and repulsive sites, and it can easily be shown that the most optimal flux is reached when several repulsive sites cluster together near the entrance, while attractive sites must stay closer to the exit to optimize the overall particle flux through the channel.

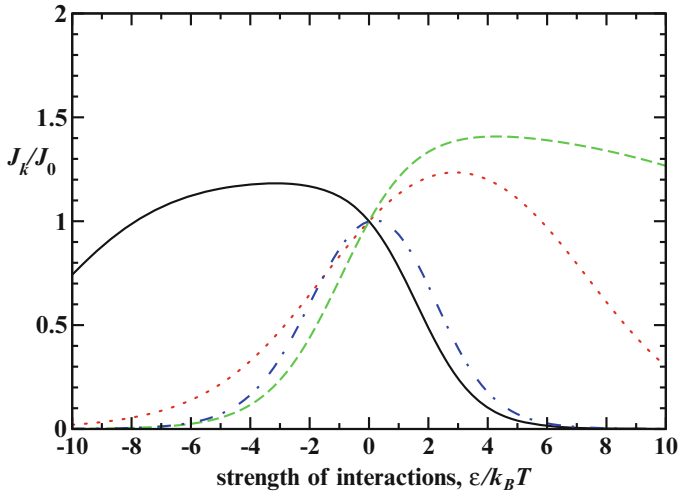

Fig. 3 The ratio of particle currents as a function of the interaction strength for channel with $N = 10$ binding sites. For the *solid curve*, the parameters are $u/u_0 = 0.1$, $k = 1$, and $\theta = 0.5$. For the *dotted curve*, the parameters are $u/u_0 = 0.1$, $k = 10$, and $\theta = 0.5$. For the *dashed curve*, the parameters are $u/u_0 = 0.1$, $k = 10$, and $\theta = 0.8$. For the *dash-dotted curve*, the parameters are $u/u_0 = 0.1$, $k = 5$, and $\theta = 0.9$

The amplitude of interactions at the special site can also affect the flux through the nanopore as shown in Fig. 3, in agreement with previous theoretical predictions for channel-facilitated molecular transport [22–26, 28, 29]. For any set of parameters, there is an optimal interaction strength ε^* that can be obtained from the condition $\frac{\partial J_k(N)}{\partial x}(\varepsilon^*) = 0$, yielding the following equation:

$$2\theta\left[\frac{u}{u_0} + k - 1\right] = (1-\theta)x + (N-k)x^{1+\theta}. \tag{7}$$

Specifically, for the most optimal site $k = N$ (for attractive interactions) we have the following expression for the most optimal interaction strength,

$$\varepsilon^* = k_B T \ln\left[\frac{2\theta}{1-\theta}\left(\frac{u}{u_0} + N - 1\right)\right]. \tag{8}$$

It is interesting to note that the optimal interaction in this case is an increasing function of the size of the channel (or the number of binding sites N). For arbitrary position of the special binding site, from (7) it can be shown that for $\theta = 0$ we have $\varepsilon^* = -\infty$, while for $\theta = 1$ one can obtain

$$\varepsilon^* = \frac{1}{2}k_B T \ln\left[\frac{2\left(\frac{u}{u_0} + k - 1\right)}{N-k}\right]. \tag{9}$$

The optimal interaction strength also depends on the concentration gradient across the channel [28]. In addition, it is important to emphasize that very strong attractive interactions do not benefit the particle transportation, because the molecules become trapped inside the channel. At the same time, very strong repulsive interactions serve as impenetrable barriers for molecular motion through the channel. These theoretical results suggest that the shape and symmetry of free-energy profiles are important factors that modify the molecular transport across membrane channels.

Theoretical arguments for optimization of particle currents can also be extended to analyze optimization with respect to other dynamic properties, e.g., particle fluctuations [29]. It was shown that the largest particle current and the minimal dispersion can be realized for the same interaction strengths only for locally symmetric potentials, while breaking this symmetry leads to different optimal conditions. This result clearly shows the importance of symmetry for understanding molecular transport across membrane channels.

2.2 Intermolecular Interactions

In real biological transport systems, more than one molecule can be found inside the nanopores and channels, and intermolecular interactions might become an important factor of translocation [30, 31]. Earlier theoretical treatments have considered the effect of particle crowding [30, 31], but only hard core exclusion interactions have been assumed, and correlations in channel occupation have also been neglected. However, since molecular permeation through the pores can be viewed effectively as a one-dimensional system the effect of intermolecular interactions and particle correlations could be significant. To investigate explicitly the effect of intermolecular interactions, the method of discrete-state stochastic models is very convenient, but the requirement of having only a single molecule in the pore must be relaxed. In addition, intermolecular interactions will change the free-energy landscape of the system, and this should be taken into account.

Specifically, we consider a simple $N = 2$ model without molecule/pore interactions, as described above, but allowing more than one particle to occupy the channel. This model can serve as a good testing ground for underlying complex biological transport phenomena. There is an energy cost associated with finding two particles next to each other. The configuration with two particles has an energy ε, with $\varepsilon > 0$ ($\varepsilon < 0$) describing attractive (repulsive) interactions. There are four possible configurations in the channel, as plotted in Fig. 4. We label them as (i, j) with $i, j = 0$ ($i, j = 1$) for the empty (occupied) site. It should be noted that the rate to enter the half-filled configuration u_1 and the exit rate from the fully occupied state u_2 are related via the detailed balance,

$$\frac{u_1}{u_2} = \frac{u_0}{u} x, \tag{10}$$

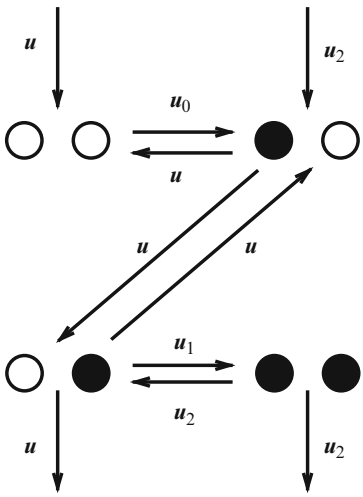

Fig. 4 A general schematic picture for channel with $N = 2$ binding sites, multiple occupancy, and with intermolecular interactions. *Open circles* describe empty sites, while *filled circles* denote occupied sites. Two molecules sitting next to each other interact with energy ε

with $x = \exp(\varepsilon/k_{\mathrm{B}}T)$. The case $\varepsilon = 0$ corresponds to the situation analyzed in Zilman [30, 31]. This observation allows us to write explicit expressions,

$$u_1 = u_0 x^{\theta}, \quad u_2 = u x^{\theta-1}, \tag{11}$$

where the coefficient $0 \leq \theta \leq 1$ again specifies how the interparticle interaction modifies these entrance and exit rates. We can define $P(i, j; t)$ as the probability of finding channel in the state (i, j) at time t, and temporal evolution of the system dynamics can be found by analyzing corresponding master equations:

$$\begin{aligned}
\frac{\mathrm{d}P(0,0;t)}{\mathrm{d}t} &= uP(0,1;t) + uP(1,0;t) - u_0 P(0,0;t),\\
\frac{\mathrm{d}P(1,0;t)}{\mathrm{d}t} &= u_0 P(0,0;t) + u_2 P(1,1;t) + uP(0,1;t) - 2uP(1,0;t),\\
\frac{\mathrm{d}P(0,1;t)}{\mathrm{d}t} &= uP(1,0;t) + u_2 P(1,1;t) - (2u + u_1)P(0,1;t),\\
\frac{\mathrm{d}P(1,1;t)}{\mathrm{d}t} &= u_1 P(0,1;t) - 2u_2 P(1,1;t).
\end{aligned} \tag{12}$$

Solving these equations at large times ($\frac{\mathrm{d}P(i,j;t)}{\mathrm{d}t} \to 0$) yields the expression for the molecular flux,

$$J_2 = \frac{uu_0\left(u + \frac{u_0}{2}x^{\theta}\right)}{3u^2 + \frac{u_0^2}{2}(x + x^{\theta}) + uu_0\left(3 + \frac{x^{\theta}}{2}\right)}. \tag{13}$$

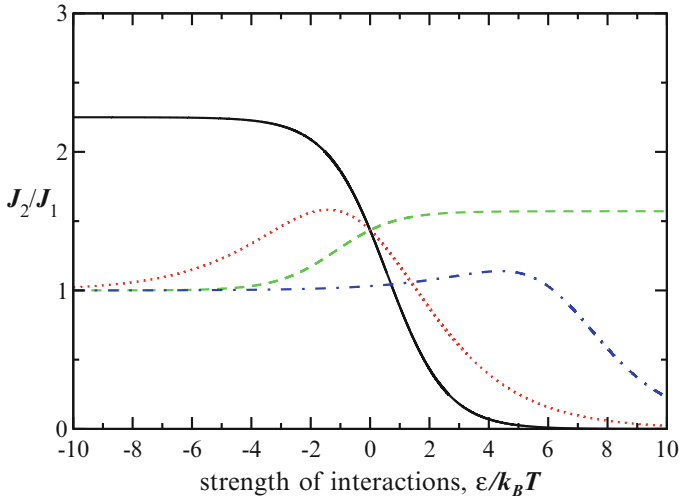

Fig. 5 Ratio of the particle currents as a function of the intermolecular interaction for channel with $N = 2$ binding sites. For the solid curve, the parameters are $u/u_0 = 0.1$ and $\theta = 0$. For the *dotted curve*, the parameters are $u/u_0 = 0.1$ and $\theta = 0.5$. For the *dashed curve*, the parameters are $u/u_0 = 0.1$ and $\theta = 1$. For the *dash-dotted curve*, the parameters are $u/u_0 = 10$ and $\theta = 0.5$

In the limit of $\varepsilon \to -\infty$, only a single molecule can be found in the channel, and (13), as expected, reduces to (2) for $N = 2$ and without molecule/pore interactions ($x = 1$),

$$J_1 = \frac{uu_0}{3(u + u_0)}. \tag{14}$$

The effect of intermolecular interactions on the channel fluxes, as shown in Fig. 5, is rather complex. For $\theta = 0$, the flux is always a decreasing function of the interaction, and single-particle transport is the most optimal. For $\theta = 1$, the trend is reversed: the stronger the interaction, the larger the molecular flux. However, for intermediate values of $0 < \theta < 1$ (which is probably a more realistic situation to describe molecular transport in cellular systems), a nonmonotonous dependence is observed, with the flux reaching a maximum at some optimal interaction strength. The optimal interaction could be positive or negative depending on the parameters of the system. These observations can be explained using the following arguments. For attractive interactions, the presence of the particle in the channel stimulates the entrance of another particle into the pore, but it also slows down the exit of both particles from the channel. For repulsive interactions, partially filled channels serve as barriers for particle entrance, thus lowering particle flux through the system. But the entering particle simultaneously accelerates the exit of the particle already inside the channel, increasing the particle current. The combination of these processes determines the complex behavior in the channel.

This theoretical analysis can be extended to channels with the larger number of binding sites. Although the algebraic expressions and derivations become very complicated, one still expects to have the same mechanisms for molecular translocation of interacting particles.

3 Summary and Conclusions

To summarize, the effect of interactions on molecular transport across channels has been investigated theoretically. Using exactly solvable discrete stochastic models, we have shown that the strength of the interaction and the spatial distribution of binding sites are important parameters that can effectively control molecular translocations through nanopores. It was found that the largest particle current can be achieved when attractive sites are near the exit of the channel, while the most optimal position of repulsive sites are near the entrance. Optimization of other dynamic properties can be done in a similar way. It has been argued that the mechanism of how the interaction affects the transport across the channel is based on controlling local concentration of particles in the channel. Special binding sites serve as local traps or barriers, modifying the overall dynamics. Attractive sites increase the probability of finding the particles at these binding sites, while repulsive sites work as barriers preventing particles already in the channel from moving back. Our theoretical picture agrees well with available single-molecule experiments on translocation of polypeptides, and it also explains qualitatively these observations [14, 15]. In addition, the presented theoretical conclusions also agree with experimental observations on maltoporin channels [8]. One could argue that our theoretical approach explains observed distributions of binding sites in real biological channels [34]. However, it should be noted that biological transport systems most probably are not optimized with respect to the particle current. Our method still allows the analysis of microscopic details of molecular transport across channels.

The presented theoretical method also allowed us to study the role of intermolecular interactions in transport through nanopores. It was found that at some interaction strength, the particle flux can be increased to reach the maximum level. This complex behavior could be explained by the fact that particles already in the channel catalyze or inhibit the entrance of other molecules into the channel. Thus, discrete-state stochastic models present a convenient theoretical framework for investigating complex transport phenomena in biological and artificial channels, and they also serve as a first step for further studies that must include more realistic structural and biochemical information.

References

1. Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., Darnell, J.: Molecular Cell Biology, 4th edn. W.H. Freeman and Company, New York (2002)
2. Hille, B.: Ionic Channels of Excitable Membranes, 3rd edn. Sinauer Associates, Sunderland Massachusetts (2001)
3. Nekolla, S., Andersen, C., Benz, R.: Noise analysis of ion current through the open and the sugar-induced closed state of the LamB channel of *Escherichia coli* outer membrane: evaluation of the sugar binding kinetics to the channel interior. Biophys. J. **66**, 1388–1397 (1994)
4. Wickner, W., Schekman, R.: Protein translocation across biological membranes. Science **310**, 1452–1456 (2005)
5. Hilty, C., Winterhalter, M.: Facilitated substrate transport through membrane proteins. Phys. Rev. Lett. **86**, 5624–5627 (2001)
6. Kullman, L., Winterhalter, M., Bezrukov, S.M.: Transport of maltodextrins through maltoporin: a single-channel study. Biophys. J. **82**, 803–812 (2002)
7. Nestorovich, E.M., Danelon, C., Winterhalter, M., Bezrukov, S.M.: Designed to penetrate: time-resolved interactions of single antibiotic molecules with bacterial pores. Proc. Natl. Acad. Sci. USA **99**, 9789–9794 (2002)
8. Danelon, C., Brando, T. Winterhalter, M.: Probing the orientation of reconstituted maltoporin channels at the single-protein level. J. Biol. Chem. **278**, 35542–35551 (2003)
9. Schwarz, G., Danelon, C., Winterhalter, M.: On translocation through a membrane channel via an internal binding site: kinetics and voltage dependence. Biophys. J. **84**, 2990–2998 (2004)
10. Krasilnikov, O.V., Rodrigues, C.G., Bezrukov, S.M.: Single polymer molecules in a protein nanopore in the limit of a strong polymer-pore attraction. Phys. Rev. Lett. **97**, 018301 (2006)
11. Caspi, Y., Zbaida, D., Elbaum, M.: Synthetic mimic of selective transport through the nuclear pore complex. Nano Lett. **8**, 3728–3734 (2008)
12. Karginov, V.A., Nestorovich, E.M., Moayeri, M., Leppla, S.H., Bezrukov, S.M.: Blocking anthrax lethal toxin at the protective antigen channel by using structure-inspired drug design. Proc. Natl. Acad. Sci. USA **102**, 15075–15080 (2005)
13. Kohli, P., Harrell, C.C., Cao, Z., Gasparac, R., Tan, W., Martin, C.R.: DNA-functionalized nanotube membranes with single-base mismatch selectivity. Science **305**, 984–986 (2004)
14. Wolfe, A.J., Mohammad, M.M., Cheley, S., Bayley, H., Movileanu, L.: Catalyzing the translocation of polypeptides through attractive interactions. J. Am. Chem. Soc. **129**, 14034–14041 (2007)
15. Mohammad, M.M., Prakash, S., Matouschek, A., Movileanu, L.: Controlling a single protein in a nanopore through electrostatic traps. J. Am. Chem. Soc. **130**, 4081–4088 (2008)
16. Niedzwiecki, D.J., Grazul, J., Movileanu, L.: Single-molecule observation of protein adsorption onto an inorganic surface. J. Am. Chem. Soc. **132**, 10816–10822 (2010)
17. Gillespie, D., Boda, D., He, Y., Apel, P., Siwy, Z.S.: Synthetic nanopores as a test case for ion channel theories: the anomalous mole fraction effect without single filing. Biophys. J. **95**, 609–619 (2008)
18. Jensen, M.O., Park, S., Tajkhorshid, E., Schulten, K.: Energetics of glycerol conduction through aquaglyceroporin GlpF. Proc. Natl. Acad. Sci. USA **99**, 6731–6736 (2002)
19. de Groot, B.L., Grubmüller, H.: Water permeation across biological membranes: mechanism and dynamics of Aquaporin-1 and GlpF. Science **294**, 2353–2357 (2001)
20. Chou, T.: How fast do fluids squeeze through microscopic single-file pores? Phys. Rev. Lett. **80**, 85–88 (1998)
21. Chou, T.: Kinetics and thermodynamics across single-file pores: solute permeability and rectified osmosis. J. Chem. Phys. **110**, 606–615 (1999)
22. Berezhkovskii, A.M., Pustovoit, M.A., Bezrukov, S.M.: Channel-facilitated membrane transport: transit probability and interaction. J. Chem. Phys. **116**, 9952–9956 (2002)

23. Berezhkovskii, A.M., Pustovoit, M.A., Bezrukov, S.M.: Channel-facilitated membrane transport: average lifetimes in the channel. J. Chem. Phys. **119**, 3943–3951 (2003)
24. Berezhkovskii, A.M., Bezrukov, S.M.: Channel-facilitated membrane transport: constructive role of particle attraction to the channel pore. Chem. Phys. **319**, 342–349 (2005)
25. Berezhkovskii, A.M., Bezrukov, S.M.: Optimizing transport of metabolites through large channels: molecular sieves with and without binding. Biophys. J. **88**, L17–L19 (2005)
26. Bezrukov, S.M., Berezhkovskii, A.M., Szabo, A.: Diffusion model of solute dynamics in a membrane channel: mapping onto the two-site model and optimizing the flux. J. Chem. Phys. **127**, 115101 (2007)
27. Berezhkovskii, A.M., Pustovoit, M.A., Bezrukov, S.M.: Fluxes of non-interacting and strongly repelling particles through a single conical channel: analytical results and their numerical tests. Chem. Phys. **375**, 523–528 (2010)
28. Kolomeisky, A.B.: Channel-facilitated molecular transport across membranes: attraction, repulsion and asymmetry. Phys.Rev. Lett. **98**, 048105 (2007)
29. Kolomeisky, A.B., Kotsev, S.: Effect of interactions on molecular fluxes and fluctuations in the transport across membrane channels. J. Chem. Phys. **128**, 085101 (2008)
30. Zilman, A.: Effects of multiple occupancy and interparticle interactions on selective transport through narrow channels: theory versus experiment. Biophys. J. **96**, 1235–1248 (2009)
31. Zilman, A., Pearson, J., Bel, G.: Effects of jamming on nonequilibrium transport times in nanochannels. Phys. Rev. Lett. **103**, 128103 (2009)
32. Kolomeisky, A.B., Slonkina, E.: Polymer translocation through a long nanopore. J. Chem. Phys. **118**, 7112–7118 (2003)
33. Kolomeisky, A.B., Fisher, M.E.: Molecular motors: a theorists's perspective. Ann. Rev. Phys. Chem. **58**, 675–695 (2007)
34. Chacinska, A., Pfanner, N., Meisinger, C.: How mitochondria import hydrophilic and hydrophobic proteins. Trends Cell Biol. **12**, 299–303 (2002)

Part IV
Modeling Evolution

Modeling Protein Evolution

Richard Goldstein and David Pollock

1 Why Model Protein Evolution?

The study of biology is fundamentally different from many other scientific pursuits, such as geology or astrophysics. This difference stems from the ubiquitous questions that arise about function and purpose. These are questions concerning why biological objects operate the way they do: what is the function of a polymerase? What is the role of the immune system? No one, aside from the most dedicated anthropist or interventionist theist, would attempt to determine the purpose of the earth's mantle or the function of a binary star. Among the sciences, it is only biology in which the details of what an object does can be said to be part of the *reason* for its existence. This is because the process of evolution is capable of improving an object to better carry out a function; that is, it adapts an object within the constraints of mechanics and history (i.e., what has come before). Thus, the ultimate basis of these biological questions is the process of evolution; generally, the function of an enzyme, cell type, organ, system, or trait is the thing that it does that contributes to the fitness (i.e., reproductive success) of the organism of which it is a part or characteristic. Our investigations cannot escape the simple fact that all things in biology (including ourselves) are, ultimately, the result of an evolutionary process.

The understanding of our evolutionary heritage has a wide range of conceptual, theoretical, and practical applications. First, we are often interested in the evolutionary process because it has specific consequences. To control pandemics, we want

R. Goldstein (✉)
Division of Mathematical Biology, National Institute for Medical Research,
London, NW7 1AA, UK
e-mail: r.goldstein@ucl.ac.uk

D. Pollock
Department of Biochemistry and Molecular Genetics, University of Colorado
School of Medicine, Aurora, CO 80045, USA
e-mail: David.Pollock@UCDenver.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_13, © Springer Science+Business Media, LLC 2012

to know how a pathogen spreads, and we do this by studying the phylogenetic relationships among various pathogen isolates. To understand how to preserve ecosystems, we study how population structure and population size affects the evolution and long-term stability of various species. And to understand and control the evolution of drug resistance or virulence in bacteria, we are interested in understanding how these processes are mediated by horizontal gene transfer.

Second, by observing not just a single instance of something, but also how it varies within and between populations and species, we can learn more about how it works and what is important for maintaining or altering function. It is extremely difficult to analyze patterns of conservation and variation without considering the source of this conservation and variation, i.e., the evolutionary process.

Third, we are interested in evidence of new things that are not contained in our current philosophy. Something that is inconsistent with our current understanding of what would arise through evolution requires us to postulate a new process, phenomenon, or extended theory. For example, an overly slow or overly rapid rate of sequence change, or an overabundance of certain types of network motifs or protein structures, may clash with what we would otherwise expect.

Fourth, evolutionary biology is the story of our creation, the basis of who we are and why we are here on this planet. Because it provides ultimate answers to the "why" questions in biology, evolution serves as an illuminating mechanism to correct erroneous conceptions of ourselves that we have fabricated based on how we would like to justify our existence, rather than on biological evidence. Thus, evolution has justifiably been described as the "universal acid" [1], sculpting and eating away at our most fundamental and dangerous misconceptions of ourselves to reveal the underlying sculptured beauty of our true role in the universe. This is where art and science meet, both "incandescently" and "incestuously" [2].

Conceptual models of evolution often rest on a mapping of genotype (what is evolving) to phenotype (the traits that result). The process of evolution through genotype–phenotype space critically depends on the nature of this mapping; many fundamental disagreements in evolutionary biology result from different conceptualizations of this relationship. Much work has focused on the evolution of biological macromolecules such as DNA, RNA, and proteins. These molecules provide a tractable but sufficiently realistic genotype–phenotype map, where the phenotypical properties such as structure, stability, and functionality can be modeled based on first principles. In addition, while higher-level models representing cells, organisms, and ecosystems are not uninteresting, the properties of biological macromolecules lie at the base of evolutionary processes, and it seems to us that if we are to eventually understand the higher levels, we must first understand the base. In this review, we direct our attention to the modeling of proteins and their evolution. The evolution of proteins can serve as a model for how the activity of the cell is directed, modulated, regulated, accelerated, and controlled.

2 The Challenges

Evolution involves a massively parallel operation occurring in large populations over long periods of time. We seek ways to simulate this process in a reasonable time with limited computational resources. This means we need to greatly speed up and therefore grossly simplify the system. In order to model the evolution of proteins, we need a model of how the proteins behave, a way of characterizing the viability or fitness of a given protein, and a description of the population dynamics of the evolutionary process. Developing simplified representations of these three aspects has specific challenges.

1. Proteins are complex. Proteins can range from hundreds to thousands of amino acids, representing thousands to tens of thousands of atoms. The number of possible conformations exceeds the number of atoms in the universe by many orders of magnitude. Proteins also interact intimately with their heterogeneous environment, which can include water, ions, and complex bilayer membranes, as well as other biomolecules such as nucleic acids and other proteins.
2. The interactions within proteins, and among proteins and their environments, are not well understood. This is especially problematic given that the thermodynamic properties of proteins often represent small differences between large numbers; the large terms must be known to excruciatingly high accuracy in order for the small differences to be meaningful.
3. Proteins have to fulfill multiple requirements, including the abilities to fold into a stable, well-defined structure; maintain solubility; be trafficked to appropriate parts of the cell (or excreted externally); and recognize, interact with, and process other biological and environmental components. The specifics of these constraints, and how they interact, are difficult to determine.
4. The structure and function of proteins can evolve, as well as the sequence. It is difficult to predict how changing structural, functional, and sequence contexts might alter subsequent evolutionary patterns.
5. In general, evolution involves competition between individuals with genomes that encode multiple proteins. The evolutionary dynamics of a given gene variant involves the fitness of those individuals that contain this variant, including how it interacts with other genes and gene products as well as the environment. It is not easy to quantify the complex and heterogeneous relationship between the fitness of an individual and the properties of a single or few proteins encoded by their genomes.
6. Evolution takes place in large, heterogeneous populations. The variation within this population can have a strong effect on the evolutionary dynamics. The effect of a mutation on an individual's fitness depends in part upon the characteristics of other individuals in that population.

A range of different research approaches has been developed to address these difficulties and complications.

2.1 Modeling Protein Energetics

It is generally impractical to explicitly consider the motions and interaction energies of all of the atoms found in a protein and its environment, and so models must be constructed that do not consider all of the atoms and all of the motions. A common strategy is to consider *reduced representations* of proteins, in which each amino acid in the chain is represented by one or a few particles.

While the use of a reduced representation removes many conformational degrees of freedom from the system, the number of possible conformations is still too large to be easily modeled. To reduce the conformation space, a limited number of possible conformations may be considered, one of which represents the native state. These conformations can include known protein structures as well as one of the sets of decoy structures that have been created (e.g., [3]). A more extreme approach is to consider each amino acid as a single particle located at adjacent vertices in a (generally cubic) lattice [4]. The covalent bonds between residues are represented as edges in the lattice, and interactions occur only between adjacent nonbonded residues. Possible conformations correspond to different self-avoiding walks through these lattices; excluded volume is implemented by the requirement that two residues cannot share the same lattice point. The lattice can either be small (so that only the compact states are represented) or it can be big enough that entirely unfolded proteins are possible. The number of possible compact conformations for small proteins is reasonable (a 27-mer protein on a compact $3 \times 3 \times 3$ lattice has 103,346 conformations, excluding reflections and rotations), but the number of conformations increases rapidly with the size of the protein and when noncompact forms are considered. Small three-dimensional proteins on a regular lattice have few residues that are internal or "buried" compared to real proteins, and so some researchers have instead used two-dimensional lattices, which generate a more reasonable fraction of buried residues. Such lattices have advantages for modeling protein thermodynamics but are generally inappropriate for simulations of, for instance, folding dynamics. The ground state can either be specified in advance or allowed to change during the simulation.

The free energy of a protein is the sum of a number of different types of interactions, including (a) the van der Waals contact energy between atoms, (b) Coulomb interactions between charges, (c) hydrogen bonds, (d) the hydrophobic effect, encompassing the entropy loss resulting from the structuring of water near nonpolar groups, (e) bond stretches, bends, and rotations, and (f) changes in the conformational and vibrational entropy. It is difficult to calculate an accurate value for the free energy of a particular conformation of a protein in its full atom representation, especially the entropic contributions. The situation becomes even worse for coarse-grained models, as many of the atoms involved in the interactions are not represented in the structure. Generally, modelers use highly simplified free energy functions such as those based on contact energies, where the free energy $G(S, F)$ of a sequence $S = \{A_1,\ A_2,\ \ldots\ A_n\}$ in fold F can then be expressed as

$$G(S, F) = \sum_{<i,j>} \gamma(A_i, A_j)\ H(r_0 - r_{ij}), \tag{1}$$

where $\gamma(A_i, A_j)$ is the interaction between the type of amino acid found in positions i and j, and $H(r_0 - r_{ij})$ is a Heavyside step function equal to one if and only if r_{ij}, the distance between specified atoms in residues i and j, is less than some cutoff r_0. Early models considered only two types of amino acids, hydrophobic (H) and polar (P), where only hydrophobic residues interact energetically [5]. Energetic parameters for more realistic schemes have been constructed based on knowledge of physical chemistry or, alternatively, extracted through statistical analysis of the available protein structures [6]. The latter approach is based on two incorrect but useful assumptions. The first is that we can integrate over all of the degrees of freedom of the system that we consider unimportant for the simulation (such as conformations of the side chains and solvent molecules) in order to calculate *potentials of mean force*, and that the potentials of mean force can be decomposed into a sum of uncorrelated terms that can be computed independently. The second assumption is that the distribution of states of a single protein in its multiplicity of possible conformations can be represented by the distribution of native states of a multiplicity of known proteins. Under this assumption, the Boltzmann expression relating the probability of observing a state to the free energy of that state can be inverted to determine the free energy of an interaction based on the frequency of that interaction in the database of known protein structures. (Finkelstein and collaborators have promoted an interesting evolutionary justification for this assumption [7, 8]). Because the potential of mean force explicitly includes the sum over all degrees of freedom of the system, including the solvent degrees of freedom, entropic interactions involving the solvent (and thus, the hydrophobic effect) are included in the simulation.

2.2 Modeling Selective Constraints

Once a model of the protein and a representation of its interactions are in place, the selective constraints acting on the protein can be determined. These can be a mixture of structural, thermodynamic, or functional properties. Common selective constraints include:

1. *The need to fold into a well-defined structure*. In general, explicit folding simulations are too slow for all but the simplest representations of proteins or of the evolutionary process [9]. Rather than simulating the folding process, an alternative approach is to evaluate properties that are characteristic of folded proteins. For instance, the presence of a nondegenerate ground state (either any ground state or a prespecified ground state) has been considered adequate [10], although this is difficult to justify at nonzero temperatures. Wolynes and coworkers used concepts from spin-glass physics to suggest that an appropriate measure of "foldability" [11] is the energy gap between the native conformation and the distribution of the energies for random conformations [12]; this can also be equated with a statistical Z-score. These predictions were later verified

using folding simulations [9]. Evolutionary simulations can be performed so that proteins with a foldability larger than some critical value are considered viable, with fitness equal to one, while proteins with lesser foldability are considered nonviable, with fitness equal to zero [13, 14]. (Evolutionary models where fitnesses are either a constant high value or a constant low value depending upon some criterion are called *truncation selection*.)

2. *The need to be sufficiently stable in the native state conformation*. This is quantified as either a free energy of folding ($\Delta G_{\text{Folding}}$) or as the fraction of time spent in the folded state at equilibrium. The added difficulty here is the need to consider the enormous ensemble of unfolded conformations. One approach is to consider a large number of alternative conformations, recognizing that the number of conformations considered represents a vast underestimation of the total possible number of such states. Alternatively, one can extrapolate from a small ensemble of proteins to the properties of a larger ensemble. For instance, imagine that the distribution of free energies of alternative structures $\rho_{\text{U}}(G)$ is approximated by a Gaussian distribution with average $\overline{G}$ and variance σ^2. We can consider sampling a large number of structures from this distribution, representing the entire ensemble of unfolded states, to calculate G_{U}, the free energy of the ensemble of alternative unfolded states

$$G_{\text{U}} = \frac{\sigma^2 - 2kT\overline{G}}{2kT} + kT \ln N_{\text{U}}, \tag{2}$$

where N_{U} is the total number of alternative states, k is the Boltzmann constant, and T is the temperature [15]. Calculation of the free energies of a limited number of alternative structures may be sufficient to characterize $\rho_{\text{U}}(G)$ and allow us to estimate $\overline{G}$ and σ^2.

Another alternative is to consider that we often do not have to calculate stabilities, but only the change in stability ($\Delta\Delta G_{\text{Folding}}$) due to a mutation. Such changes in stability can be calculated using, for instance, FoldX [16]. While this can be more accurate than the simple representation of protein energetics, the complexity of these calculations limits evolutionary simulations.

3. *The need to be functional*. In addition to being foldable and stable, proteins generally fulfill one or more functions. In contrast to foldability and stability, the functional requirements for proteins are highly specific, making modeling of functional constraints difficult. One class of models that have been investigated involves the consideration of binding properties. For instance, Hirst and coworkers considered the ability of a protein to construct an appropriate binding pocket [17, 18]. Alternatively, a protein's fitness can be modeled as a function of how well it binds a specified peptide or other protein [19, 20]. The intermolecular binding interaction can be calculated using the same parameters used for the intramolecular interactions, and we can consider fundamental properties of protein function such as specificity by modeling binding to competing peptides as well. We can also consider what happens in the evolutionary simulation when

the binding partner changes, allowing us to model evolutionary dynamics under changing selective constraints. When fitness involves the binding of multiple proteins, we can consider these interacting proteins as encoded in a single genome. The proteins would then co-evolve as the genome evolves.

The above selective constraints refer to the properties of the amino acid sequence that constitute the protein. Mutations, however, occur at the DNA level, and it is only through the process of translation that these changes are represented at the amino acid level [21]. The accuracy of translation is much lower than that of DNA replication, meaning that there will be a distribution of protein sequences corresponding to a single DNA sequence. This more complicated relationship between gene and gene product, and the consequential complication of the fitness function, can have interesting and important evolutionary consequences [22–24].

2.3 Modeling Evolutionary Dynamics

Evolution is a population phenomenon. In a population of individuals, mutations that change the phenotype of the organism will succeed based on how that phenotype fares in competition with other phenotypes that exist in the population at the same time. A point mutation is the exchange of one nucleotide for another somewhere in the genome. If the exchange occurs in a coding region, it might alter the amino acid that is produced upon translation (a *missense* or *nonsynonymous* mutation), or it might result in a stop codon, thus causing premature truncation of the protein (a *nonsense* mutation). Alternatively, the new codon might encode the same amino acid as the old codon, and thus produce no change in the translated protein sequence (a *silent* or *synonymous* mutation). Mutations that are synonymous are more likely to have a relatively neutral effect on fitness compared to missense or nonsense mutations, because they do not alter the amino acid sequence and thus do not alter the functional properties of the individual protein. More complicated mutations are possible involving the deletion, insertion, duplication, or re-arrangement of single nucleotides, stretches of nucleotides, or larger units of the genome.

By definition, mutations occur in individuals, but in subsequent generations the offspring of the mutated individual will carry the mutation and compete with offspring in the population that do not carry the mutation. Ultimately, this will usually lead either to elimination of the new mutation from the population or to *fixation*, the process whereby the other variants in the population are eliminated. After fixation, the previous "mutant" defines a new "wild type."

In unusual cases, mutants may be neither eliminated nor fixed for long periods of time, during which the population remains polymorphic. For example, in diploid organisms an individual carries two copies of each gene, and an individual with one copy of each variant (a *heterozygote*) may have an advantage over other individuals. This is the case with the mutation causing sickle-cell anemia, in which one copy

of the mutant gene provides some resistance against malaria, while two copies leads to debilitating illness and early death. Such situations are called *heterozygote superiority* or *overdominance*. Alternatively, it may be that there is an advantage to being different from others in the population, so that any new mutant with a phenotypic difference has an advantage as long as it is not overly common in the population (such as being able to utilize a new but limited source of food); this situation, where the fitness of a trait depends upon the frequency of the trait in the population, is called *frequency-dependent selection*. In either of these cases, multiple variants may stably co-exist in the population.

Individuals replicate, reproduce, and have heredity, but proteins by themselves do not. Evolutionary simulations with proteins therefore involve making connections between the characteristics of the proteins and the fitnesses of the organisms in which they reside. Furthermore, evolution occurs in populations of individuals containing proteins, and therefore what matters is the relative fitness of the individual carrying the protein. We can make this connection most simply by assuming that the probability that an individual reproduces is proportional to the probability that a given protein is folded (or that it is folded and able to bind a peptide). Because folding probability is maximal at a probability of 1.0, fitness also reaches a maximum at 1.0. At the extreme, there may be a subset of proteins with relative fitnesses indistinguishably near 1.0, with the remainder of the proteins having relatively poor folding and likely to be quickly eliminated whenever they do arise via mutation. This leads to a fitness landscape that is essentially a *neutral network*, a set of genotypes connected by single mutations that have equal fitness.

It is possible to ignore population dynamics completely and instead follow the movement of point mutations via a *stochastic hill-climbing* procedure, where random mutations are made and accepted if the mutation increases the fitness relative to the previous point. This type of simulation results in a trajectory in the space of possible sequences that moves unceasingly toward higher and higher fitness values. Such studies can provide insight into the nature and structure of the fitness landscape but are not representative of the evolutionary process, and generally will have quite different dynamic and equilibrium properties.

Instead, then, evolution is almost always modeled in a population; but there are many reasonable choices for the characteristics of this population. Common practice is to model the simplest population possible unless there are more complex aspects that one particularly wishes to test or that are critical for a given study. For example, if heterozygote advantage or other properties of a diploid system are not of interest, one would usually simulate a haploid organism, in which case the allele is equivalent to the genotype. Different kinds of nucleotide mutations (for example, thymine to cytosine, versus guanine to cytosine) are often assumed to arise at equal frequency, unless it is deemed critical to reflect more realistic relative mutation patterns. An extreme example of this is the modeling of amino-acid-altering exchanges at equal frequencies, as though mutations occurred at the amino acid level. Although this is an obviously unrealistic fiction, it may still be justifiable if it could be argued that this particular departure from reality does not make any difference in the properties of concern in a particular study.

Another consideration in modeling evolutionary dynamics is the timing of reproduction and selection among individuals. It is common to follow the classic synchronous population model that coincides with the bulk of classical mathematical theory on population genetics dynamics. In such populations, reproduction and changes in allele frequency due to fitness differences occur at the same time for the entire population. Each such set of steps is called a generation. Thus, for a population of N individuals, the allele frequencies at the end of one generation would be multiplied by their relative fitness, and then the entire population would be randomly re-sampled with replacement from the resulting allele frequencies to create the next generation. Mutations can occur during this re-sampling. These can include, in addition to simple mutations, other processes such as mating and recombination. It is traditional to hold the population size constant over the course of a simulation.

There are a number of common variants to this scheme, including allowing the population size to change over time, and allowing individual (or mating pair) variation in reproductive success. Sometimes alternative schemes are utilized. For example, in *tournament selection*, a subset of n_{Tourn} individuals from the population are chosen at random, with replacement, and only the most fit individual in this subset is replicated in the next population. This process is then repeated N times to fill the entire population for the next generation. Selection pressure can be controlled by the value of n_{Tourn}, with $n_{\text{Tourn}} = 1$ corresponding to no selective pressure (the next generation is chosen at random from the genotypes in the previous generation) and $n_{\text{Tourn}} \gg N$ corresponding to stochastic hill climbing where only the most fit individual reproduces. A scheme that better reflects reproduction in many species (including humans) is the *Moran process*, in which reproductive events occur sequentially in the population [25]. Population size is maintained by selecting at each time step one individual to duplicate (with probability proportional to its fitness) and one individual to be eliminated. Again, mutations can be implemented as part of the replication process.

In order to speed up simulations, we can also calculate, rather than simulate, the probability of fixation. This approach relies on the assumption that mutations that are destined to become fixed do not overlap in time, such that each mutation is either eliminated or fixed in the population prior to the arrival of the next mutation. In this case, the probability of fixation of the mutant P_{Fix} has been computed by Kimura [26–28]:

$$P_{\text{Fix}} = \frac{1 - e^{-2s}}{1 - e^{-2\gamma N s}}, \tag{3}$$

where γ is equal to 1 for haploid organisms and 2 for diploid organisms, and the selection coefficient s is given by

$$s = \frac{\omega_{\text{mut}} - \omega_{\text{wt}}}{\omega_{\text{mut}}}, \tag{4}$$

where ω_{mut} and ω_{wt} are the relative fitnesses of the mutant and wild-type alleles, respectively. This equation holds when $s << 1$, in which case we can choose a

mutant to test, and then efficiently accept or reject it based on this probability. While the use of the Kimura formulation results in faster simulations, it eliminates the possibility of observing interesting phenomena that arise from interactions among mutations that are polymorphic in a population. For example, a favorable mutant might not be fixed if, before it does, another mutant happens to arise that is more favorable. Also, in a *selective sweep*, an extremely fit mutation that becomes fixed rapidly can cause linked variants to become fixed as well, even if these other variants are neutral or even deleterious. These additional variants are said to have *hitchhiked* on the fixation of the favorable mutation. Finally, if the mutation rate is especially rapid, then the fitness of a variant may depend partly on how often it spawns deleterious mutations [29, 30]. This can lead to emergent properties such as robustness that would be overlooked when using the Kimura formulation.

3 The Distribution of Observed Protein Structures

As we accumulate an increasing number of protein structures, it is clear that the distribution of proteins among the diverse folds is extremely uneven, with some folds greatly over-represented and other possible folds that have not yet been observed [31–35]. Three classes of explanations for this observation have emerged [36]: (1) Some folds may be more "designable," that is, they can be formed by more sequences, and are therefore more likely to arise in evolution. (2) Some folds are better suited to important or common functionalities, or a greater range of functionalities. For instance, the cleft found in the common TIM Barrel fold might be extremely well suited for catalyzing reactions. (3) Evolutionary dynamics, as modeled as birth–death processes, may naturally lead to uneven distributions of proteins as proteins with common folds are more likely to increase their number through gene duplication events than proteins with rare folds. The first explanation, involving the "sequence entropy" of various structures, has focused the most on the nature of the genotype (sequence)–phenotype (structure) map, and the consequences of this mapping for evolutionary processes.

Parallel to this effort has been the attempt to delineate more specifically the processes that are involved in the creation of new protein folds [37]. It is clear, both experimentally [38, 39] and theoretically [14, 40], that changes in protein structure are extremely slow compared with the rate of change of protein sequences, with many highly divergent pairs of proteins having extremely similar structures. For proteins, investigations with simple models have shown that the neutral network is clustered around a prototype sequence [41]. These neutral networks are isolated from each other [42], manifested by the slow rate of change of structure. This is in striking contrast to the case of RNA, where it is relatively easy to make changes in structure with single changes in sequence [43–45]. (One possible explanation for this difference may lie in how these different systems are modeled. Often an RNA molecule is considered viable as long as it has a nondegenerate ground state, resulting in the vast majority of sequences being viable. In contrast, proteins

are often modeled as requiring a certain degree of stability or foldability, with a corresponding low fraction of viable sequences. This has obvious consequences for the topology of the neutral networks and the resultant ability to move between structures.)

There has been increased interest in structural plasticity, that is, the ability of a single protein to populate multiple structures under physiological conditions. The extreme examples are proteins that fulfill their physiological roles while remaining unstructured, or becoming structured only when binding some other molecule. But other proteins have been experimentally observed to fold into different structures, possibly with different functionalities. These proteins can then form *bridge states*, allowing a protein to evolve from one structure to another through this plastic intermediate [46, 47]. This can greatly enhance the rate of structural evolution.

4 Evolution of Thermodynamic Properties

Natural (that is, evolved) proteins are generally only marginally stable, so that relatively small changes in temperature are sufficient to cause unfolding. This observation, combined with the fact that the stability of the same protein from multiple organisms generally tracks the physiological temperature of each organism, has been interpreted to indicate that marginal stability is an adaptation for enhanced functionality [48–54]. An alternative explanation has been provided by evolutionary simulations, suggesting that marginal stability is a natural consequence of mutation-selection balance: proteins become stabilized up to the point that selective pressure for increased stability is counterbalanced by the tendency of random mutations to decrease stability [20, 55]. Such a model predicts that selection is more efficient in larger populations, as is clear from (3). There is some evidence to support this effect: Bastolla and colleagues estimated that proteins in intracellular bacteria (which have smaller population sizes than most bacteria) have smaller energy gaps between native and alternative compact states than those in free-living bacteria [56]. Similarly, proteins with both short [57] and long [58] disordered regions are more common in eukaryotes than in prokaryotes.

In addition to the biophysical consequences, marginal stability can cause *phenotypic plasticity* because the same sequence will be more likely to fold into multiple structures at thermal equilibrium. This would then favor the presence of bridge states, enhancing the rate of structural evolution, as described above. Another significant effect may be on the topology of biochemical networks; there is a relationship between protein stability, the resulting conformational flexibility, and the number of partners with which a protein can interact [59]. A tendency toward marginal stability might result in highly promiscuous hub proteins, resulting in some proteins with many more interacting partners than would be expected in random networks, reducing the number of intervening interactions necessary to link one protein to another. This is a possible explanation for the observation of these "small world networks" [60].

5 Other Evolutionary Processes

It is clear that many other evolutionary processes involve proteins. Examples include the evolution of biochemical and regulatory networks, the origin of pleiotropy and epistasis, co-evolution between, for instance, hosts and pathogens, hybrid incompatibility and speciation, phenotypic buffering, the existence and impact of neutral networks on evolution, and the evolution of evolvability, modularity, and robustness. There has been a history of exploring these processes with the aid of simple models that tend to ignore the particular (evolved) characteristics of the proteins that mediate these effects. With our growing ability to model protein evolution, these topics will be increasingly amenable to the types of simulations described in this chapter.

6 Conclusion

Our explorations of biological systems rest on our understanding of evolution. In addition to its own importance as a biological process, evolution can provide us with important insights into critical biological phenomena. We can develop our comprehension of evolution by modeling interesting but tractable systems, such as simplified models of proteins. Proteins provide us with a plausible mapping of genotype to phenotype, allowing us to explore topics beyond the "toy model" stage. Conversely, modeling the evolution of proteins also can provide insights regarding the nature of these biomolecules, what properties they would be expected to possess, how to interpret signatures encoded in their sequences, and how they function in their biological context. The properties of proteins arise as a result of the evolutionary process, while the evolutionary process is itself constrained by the properties of the evolving proteins. This circle of causality, influence, and constraint has proven to be a fruitful area for both those interested in proteins and those concerned with evolution. As these evolutionary models become more and more complex, they will have an increasing impact on a wide range of questions not just in evolution and protein biophysics, but also in biochemistry, cell and organismal biology, and ecology.

References

1. Dennett, D.C.: Darwin's Dangerous Idea. Simon and Schuster, New York (1996)
2. Nabakov, V.: Ada, or Ardor: A Family Chronicle. McGraw-Hill, New York (1969)
3. Samudrala, R., Levitt, M.: Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. Protein. Sci. **9**, (7):1399–1401 (2000).
4. Crippen, G.M.: Topology of globular proteins. J. Theor. Biol. **45**, (2):327–338 (1974).

5. Dill, K.A.: Theory for the folding and stability of globular proteins. Biochemistry. **24**, (6):1501–1509 (1985)
6. Miyazawa, S., Jernigan, R.: Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules. **18**, (3):534–552 (1985)
7. Finkelstein, A.V., Gutin, A.M., Badretdinov, A.: Boltzmann-like statistics of protein architectures. Origins and consequences. Subcell. Biochem. **24**, 1–26 (1995)
8. Finkelstein, A.V., Gutin, A.M., Badretdinov, A.Y.: Boltzmann-like Statistics of protein Architectures. Subcell. Biochem. **24**, 1–26 (1995)
9. Gutin, A.M., Abkevich, V.I., Shakhnovich, E.I.: Evolution-like selection of fast-folding model proteins. Proc. Natl. Acad. Sci. USA. **92**, 1282–1286 (1995)
10. Li, H., Helling, R., Tang, C., Wingreen, N.: Emergence of preferred structures in a simple model of protein folding. Science. **273**, 666–669 (1996)
11. Govindarajan, S., Goldstein, R.A.: Searching for foldable protein structures using optimized energy functions. Biopolymers. **36**, 43–51 (1995)
12. Goldstein, R.A., Luthey-Schulten, Z.A., Wolynes, P.G.: Optimal protein folding codes from spin glass theory. Proc. Natl. Acad. Sci. USA. **89**, 4918–4922 (1992)
13. Govindarajan, S., Goldstein, R.A.: Evolution of model proteins on a foldability landscape. Proteins. **29**(4), 461–466 (1997)
14. Govindarajan, S., Goldstein, R.A.: The foldability landscape of model proteins. Biopolymers. **42**(4), 427–438 (1997)
15. Williams, P.D., Pollock, D.D., Blackburne, B.P., Goldstein, R.A.: Assessing the accuracy of ancestral protein reconstruction methods. PLoS. Comput. Biol. **2**(6), e69 (2006)
16. Guerois, R., Nielsen, J.E., Serrano, L.: Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J. Mol. Biol. **320**(2), 369–387 (2002)
17. Blackburne, B.P., Hirst, J.D.: Population dynamics simulations of functional model proteins. J. Chem. Phys. **123**(15), 154907 (2005)
18. Hirst, J.D.: The evolutionary landscape of functional model proteins. Protein. Eng. **12**(9), 721–726 (1999)
19. Williams, P.D., Pollock, D.D., Goldstein, R.A.: Evolution of functionality in lattice proteins. J. Mol. Graph. Model. **19**(1), 150–156 (2001)
20. Williams, P.D., Pollock, D.D., Goldstein, R.G.: Functionality and the evolution of marginal stability in proteins: inferences from lattice simulations. Evol. Bioinformatics Online. **2**, 1–11 (2006)
21. Ellis, N., Gallant, J.: An estimate of the global error frequency in translation. Mol. Gen. Genet. **188**(2), 169–172 (1982).
22. Drummond, D.A., Wilke, C.O.: Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. **134**(2), 341–352 (2008)
23. Drummond, D.A., Wilke, C.O.: The evolutionary consequences of erroneous protein synthesis. Nat. Rev. Genet. **10**(10), 715–724 (2009)
24. Wilke, C.O., Bloom, J.D., Drummond, D.A., Raval, A.: Predicting the tolerance of proteins to random amino acid substitution. Biophys. J. **89**(6), 3714–3720 (2005)
25. Moran, P.A.P.: The Statistical Processes of Evolutionary Theory. Clarendon Press, Oxford (1962)
26. Crow, J.F., Kimura, M.: An Introduction to Population Genetics Theory. Harper & Row, New York (1970)
27. Kimura, M.: Some problems of stochastic processes in genetics. Ann. Math. Stat. **28**, 882–901 (1957)
28. Kimura, M.: On the probability of fixation of mutant genes in a population. Genetics. **47**, 713–719 (1962)
29. Nimwegen, EV., Crutchfield, J.P., Huynes, M.: Neutral evolution of mutational robustness. Proc. Natl. Acad. Sci. USA. **96**, 9716–9720 (1999)
30. Taverna, D.M., Goldstein, R.A.: Why are proteins so robust to site mutations. J. Mol. Biol. **315**, 479–484 (2002)

31. Huynen, M.A., van Nimwegen, E.: The frequency distribution of gene family sizes in complete genomes. Mol. Biol. Evol. **15**(5), 583–589 (1998)
32. Kuznetsov, V.: Statistics of the number of transcripts and protein sequences encoded in the genome. In: Zhang W, Shmulevich I (eds.) Computational and Statistical Approaches to Genomics, pp. 125–171. Kluwer, Boston (2002)
33. Qian, J., Luscombe, N.M., Gerstein, M.: Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J. Mol. Biol. **313**(4), 673–681 (2001)
34. Taylor, W.R., Chelliah, V., Hollup, S.M., MacDonald, J.T., Jonassen, I.: Probing the "dark matter" of protein fold space. Structure. **17**(9), 1244–1252 (2009)
35. Zhang, C., DeLisi, C.: Estimating the number of protein folds. J. Mol. Biol. **284**, 1301–1305 (1998)
36. Goldstein, R.A.: The structure of protein evolution and the evolution of protein structure. Curr. Opin. Struct. Biol. **18**(2), 170–177 (2008)
37. Bornberg-Bauer, E., Huylmans, A.K., Sikosek T How do new proteins arise? Curr. Opin. Struct. Biol. **20**(3), 390–396 (2010)
38. Chothia, C., Lesk, A.M.: The relationship between the divergence of sequence and structure in proteins. EMBO J. **5**, 823–826 (1986)
39. Lesk, A.M., Chothia, C.H.: The response of protein structure to amino-acid sequence changes. Philos. Trans. R. Soc. London Ser. B. **317**, 345–356 (1986)
40. Govindarajan, S., Goldstein, R.A.: Evolution of model proteins on a foldability landscape. Proteins. **29**, 461–466 (1997).
41. Bornberg-Bauer, E., Chan, H.S.: Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. Proc. Natl. Acad. Sci. USA. **96**, 10689–10694 (1999)
42. Bonberg-Bauer, E.: Randomness, structural uniqueness, modularity, and neutral evolution in sequence space of model proteins. Z. Phys. Chem. **216**, 139–154 (2002)
43. Fontana, W., Stadler, P.F., Bornberg-Bauer, E.G., Griesmacher, T., Hofacker, I.L., Tacker, M., Tarazona, P., Weinberger, E.D., Schuster, P.: RNA folding and combinatory landscapes. Phys. Rev. E. Stat. Phys. Plasmas. Fluids. Relat. Interdiscip. Topics. **47**(3), 2083–2099 (1993)
44. Schuster, P.: Genotypes with phenotypes: adventures in an RNA toy world. Biophys. Chem. **66**(2–3), 75–110 (1997)
45. Schuster, P.: Evolution in silico and in vitro: the RNA model. Biol. Chem. **382**(9), 1301–1314 (2001)
46. Babajide, A., Farber, R., Hofacker, I.L., Inman, J., Lapedes, A.S., Stadler, P.F.: Exploring protein sequence space using knowledge-based potentials. J. Theor. Biol. **212**(1), 35–46 (2001)
47. Bornberg-Bauer, E.: How are model protein structures distributed in sequence space? Biophys. J. **73**(5), 2393–2403 (1997)
48. Beadle, B.M., Shoichet, B.K.: Structural bases of stability-function tradeoffs in enzymes. J. Mol. Biol. **321**(2), 285–296 (2002)
49. Daniel, R.M., Dunn, R.V., Finney, J.L., Smith, J.C.: The role of dynamics in enzyme activity. Annu. Rev. Biophys. Biomol. Struct. **32**, 69–92 (2003)
50. DePristo, M.A., Weinreich, D.M., Hartl, D.L.: Missense meanderings in sequence space: a biophysical view of protein evolution. Nat. Rev. Gen. **6**, 678–687 (2005)
51. Fields, P.A.: Review: Protein function at thermal extremes: balancing stability and flexibility. Comp. Biochem. Physiol. Mol. Integr. Physiol. **129**(2–3), 417–431 (2001)
52. Schreiber, C., Buckle, A.M., Fersht, A.R.: Stability and function—2 constraints in the evolution of barstar and other proteins. Structure. **2**(10), 945–951 (1994)
53. Somero, G.N.: Proteins and temperature. Annu. Rev. Physiol. **57**, 43–68 (1995)
54. Zavodszky, P., Kardos, J., Svingor, A., Petsko, G.A.: Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. Proc. Natl. Acad. Sci. U S A. **95**(13), 7406–7411 (1998)
55. Taverna, D.M., Goldstein, R.A.: Why are proteins marginally stable? Proteins. **46**(1), 105–109 (2002)

56. Bastolla, U., Moya, A., Viguera, E., van Ham, R.C.: Genomic determinants of protein folding thermodynamics in prokaryotic organisms. J. Mol. Biol. **343**(5), 1451–1466 (2004)
57. Tokuriki, N., Oldfield, C.J., Uversky, V.N., Berezovsky, I.N., Tawfik, D.S.: Do viral proteins possess unique biophysical features? Trends. Biochem. Sci. **34**(2), 53–59 (2009)
58. Dunker, A.K., Obradovic, Z., Romero, P., Garner, E.C., Brown, C.J.: Intrinsic protein disorder in complete genomes. Genome Informatics Workshop on Genome Informatics. **11**, 161–171 (2000)
59. Fromer, M., Shifman, J.M.: Tradeoff between stability and multispecificity in the design of promiscuous proteins. PLoS. Comput. Biol. **5**(12), e1000627 (2009)
60. Wagner, A., Fell, D.A.: The small world inside large metabolic networks. Proc. Biol. Sci. **268**(1478), 1803–1810 (2001)

Modeling Structural and Genomic Constraints in the Evolution of Proteins

Ugo Bastolla and Markus Porto

1 Molecular Phenotypes

Macromolecules influence the phenotype of the organism where they are expressed through their function, and in particular through their interactions. Nevertheless, it is very difficult to computationally predict protein function and interactions. Moreover, only a few residues take part in them. For these reasons, models of molecular evolution usually represent folded macromolecules such as RNA or proteins and identify the function of the molecule with the folded structure, whose stability determines the modeled fitness.

The first works that modeled a quantitative relationship between genotype and phenotype were pioneered by the Vienna group, and took RNA molecules rather than proteins as a case study [1, 2]. RNA secondary structure can be represented as a list of base pairs interacting through hydrogen bonding. It is possible to assess the stability of RNA secondary structures using reliable empirical free energy functions based on Watson and Crick pairing rules. If pseudoknots are forbidden, fast algorithms allow to determine the most stable secondary structure (low energy ground state) for a given RNA sequence [3]. These studies showed that there is a large set of sequences that have a particular target structure as the ground state. This set is called the neutral network of an RNA secondary structure, where the term neutral refers to the fact that all sequences in the set share the same structure as their ground state. Notice, however, that they are not necessarily neutral under the point of view of misfolding stability because the free energy of the target structure

U. Bastolla (✉)
Centro de Biología Molecular "Severo Ochoa", Consejo Superior de Investigaciones Científicas and Universidad Autónoma de Madrid, Madrid, Spain
e-mail: ubastolla@cbm.uam.es

M. Porto
Institut für Theoretische Physik, Universität zu Köln, Köln, Germany
e-mail: porto@thp.uni-koeln.de

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_14,

and of alternative, competing structures may be quite different. Sequences with the same ground state that can be evolutionarily interconnected through a point mutation are neighbors on the neutral network. We define the neutral fraction $z(\mathbf{S})$ as the fraction of point mutations of sequence $\mathbf{S}$ that do not change the ground state. If the average neutral fraction in the neutral network, $\overline{z(\mathbf{S})}$, is above a critical threshold, random graph theory predicts that the neutral network is connected, whereas below the threshold the neutral network is formed by a giant connected component and several small components. For the four-letter AUGC alphabet, the neutral fraction of some tRNA-like structures were computed and shown to be $\overline{z(\mathbf{S})} \approx 0.29$, slightly smaller than the critical threshold $z_{\rm cr} = 0.37$ [2], so that the giant component dominates the neutral network. Furthermore, it was shown that the neutral networks of any two common structures are close in sequence space in the sense that there are sequences for which both structures are the ground state (i.e., have very similar low energy) that allow to connect the two neutral networks [2]. These studies highlighted the importance of neutrality for molecular adaptation [4]. We note, however, that such multiconformational sequences would not be observed in evolution if selection required that the target structure must be sufficiently stable against misfolded conformations.

We now turn from RNA to proteins, which is the main focus of this review chapter. Motivated by the earlier work on RNA, proteins have been the subject of intense investigation from several groups since the late 1990s [5–14]. In this case, there exists no approach to reliably predict the lowest energy structure of a protein sequence, and we must resort to approximations. It is common to represent a protein structure as a contact matrix whose element C_{ij} equal 1 if the residues at sites i and j are close in the three-dimensional folded structure and zero otherwise. This representation is formally similar to the secondary structure representation of RNA. However, in the latter case, each site can interact at most with another site, whereas in proteins each site has multiple contacts. It has been shown that the contact matrix is sufficient to reconstruct the whole three-dimensional structure of the protein with very high accuracy [15]. In this context, one usually assumes that the free energy of a protein with sequence $\mathbf{A}$ folded into the contact matrix $\mathbf{C}$ is given by the sum of pairwise contact interactions,

$$E(\mathbf{A}, \mathbf{C}) = \sum_{ij} C_{ij}\, U(A_i, A_j), \tag{1}$$

where $U(a, b)$ is the contact interaction matrix that expresses the free energy gained when amino acids a and b are brought in contact. In most of the results reported here, the matrix determined in Bastolla et al. [16] has been used. For proteins that fold with two-state thermodynamics, i.e., for which only the native structure and the unfolded structure are thermodynamically important, stability against unfolding is defined as the free energy difference between the folded and the unfolded state, and it can be estimated as $\Delta G_2 \approx E(\mathbf{A}, \mathbf{C}^{\rm nat}) + sL$, where $\mathbf{C}^{\rm nat}$ is the native structure, L is protein length, and $s = 0.074$ is an entropic parameter that was determined by fitting the above equation to a set of 20 experimentally measured unfolding

free energies, yielding a correlation coefficient $r = 0.92$ (UB, unpublished data). The accuracy of this method for predicting the stability effect of mutations is comparable to state-of-the-art atomistic methods such as Fold-X [17], and its computational simplicity allows for its use in simulating protein evolution for long evolutionary trajectories.

Stability against unfolding is, however, not sufficient to characterize protein stability, since one has also to assess stability against compact, incorrectly folded conformations of low energy that can act as kinetic traps in the folding process and, in many cases, give rise to pathological aggregation. The term positive design indicates sequence features that favor stability against unfolding by making the native structure more stable, obtained either through evolution or through sequence design algorithms. On the other hand, stability against misfolding is thought to be realized in natural proteins by increasing the energy of key contacts that are frequently found in alternative structures, which is termed negative design [18, 19]. Multiple sequence alignments show correlated mutations not only between positions that are in contact in the native structure (as a consequence of positive design) but also between positions that are distant in the native structure, which has been interpreted as evidence for negative design [18], although functional interpretations have also been proposed [20].

Stability against misfolded structures is difficult to estimate, and several models of protein evolution do not consider it, despite its importance being more and more recognized. Two simplified sets of alternative structures are most often used: either the set of approx. 10^5 maximally compact structures on the $3 \times 3 \times 3$ cubic lattice, which can be enumerated with affordable computational effort [21], or the set of all compact matrices of L residues that can be obtained from non-redundant structures in the protein database (PDB). This latter procedure, called threading in bioinformatics jargon, guarantees that the contact matrices fulfill physical constraints on chain connectivity, atomic repulsion, and hydrogen bonding (secondary structure), which are not enforced in the contact energy function. We will use the threading set in most of this chapter. Whatever the set of alternative structures, there are several measures to assess the stability against misfolding. A parameter that is often used is the normalized energy gap $\alpha(\mathbf{A})$, defined as

$$\alpha(\mathbf{A}) = \min_{\mathbf{C}} \frac{E(\mathbf{A}, \mathbf{C}) - E(\mathbf{A}, \mathbf{C}^{\text{nat}})}{|E(\mathbf{A}, \mathbf{C}^{\text{nat}})|\ [1 - q(\mathbf{C}, \mathbf{C}^{\text{nat}})]}\,, \tag{2}$$

where the contact overlap $q(\mathbf{C}, \mathbf{C}^{\text{nat}})$ measures the structural similarity between the native (lowest energy) structure $\mathbf{C}^{\text{nat}}$ and the alternative structure $\mathbf{C}$. A large value of $\alpha(\mathbf{A})$ implies that all the low energy structures are structurally similar to the native one and belong to its attraction basin. This is necessary for thermodynamic stability, since the contact interaction parameters are effective free energy parameters that depend on temperature and, if α is small, a small change in these parameters could completely change the ground state. Similarly, it is necessary for stability against mutation and for fast folding kinetics. $\alpha(\mathbf{A})$ is determined by the lowest energy structure that is structurally unrelated to the ground state and it can be

approximated analytically, without need to use the set of alternative structures, using the random energy model (REM) [22] and neglecting correlations between contacts [23]. However, this latter approximation is not always accurate enough for characterizing negative design (unpublished results).

An alternative way to estimate protein stability, both against unfolding and against misfolding, has been recently proposed by Goldstein [24]. This consists of using the REM to estimate the free energy of unfolded and misfolded structures, computing the difference in free energy of the native state as

$$\Delta G \approx E(\mathbf{A}, \mathbf{C}^{\rm nat}) + k_{\rm B} T s_0 L - \langle E(\mathbf{A}, \mathbf{C}) \rangle + \frac{\sigma_{\rm E}^2}{2 k_{\rm B} T} , \tag{3}$$

where $\langle E(\mathbf{A}, \mathbf{C}) \rangle$ is the mean and $\sigma_{\rm E}^2$ is the variance of the energy of alternative structures.

In modeling protein evolution, we still need to define how protein stability influences fitness, i.e., the reproduction rate of the organism. The simplest possibility is a neutral fitness landscape where the fitness is a binary variable and all proteins with stability above a given threshold are considered equally fit, whereas all proteins below threshold are inviable. We and our collaborators modeled neutral evolution, imposing two conditions on unfolding and misfolding stability: $\Delta G_2(\mathbf{A}) > \Delta G_{\rm thr}$ and $\alpha(\mathbf{A}) > \alpha_{\rm thr}$ [25]. The target native structure is fixed throughout evolution, and the thresholds are chosen proportional to the values of the stability parameters for the starting protein sequence in the PDB. In a neutral fitness landscape, population size has almost no influence on evolution. A more interesting possibility is a model where fitness is a smooth function of stability, in which case there are mutations that produce a small decrease in fitness and are more likely to be fixed in small populations (see next section). We generalized the neutral model through the fitness function $f(x_\alpha, x_G, S) = 1/\left[1 + x_\alpha^{-S} + x_G^{-S}\right]$ if both $x_\alpha > 0$ and $x_G > 0$, and $f(x_\alpha, x_G, S) = 0$ otherwise, where $x_\alpha(\mathbf{A}) = \alpha(\mathbf{A})/\alpha_{\rm thr}$ and $x_G(\mathbf{A}) = \Delta G_2(\mathbf{A})/\Delta G_{\rm thr}$. The parameter S is called neutrality exponent, and we recover the neutral landscape in the limit $S \to \infty$.

An alternative fitness function has been used by Shakhnovich [26] and Goldstein [24] among others, and it defines fitness as the Boltzmann probability to find the protein in the native state,

$$f(\mathbf{A}) = \frac{\exp(-\Delta G(\mathbf{A})/k_{\rm B} T)}{1 + \exp(-\Delta G(\mathbf{A})/k_{\rm B} T)} . \tag{4}$$

In the low temperature limit, the model becomes neutral in that sequences with $\Delta G < 0$ receive fitness one, and other sequences receive fitness zero. Notice, however, that ΔG depends on temperature, so that the low temperature limit might not be well defined (see (3)). Also note that ΔG enforces both unfolding stability $E(\mathbf{A}, \mathbf{C}^{\rm nat}) < k_{\rm B} T s_0 L$ and misfolding stability $E(\mathbf{A}, \mathbf{C}^{\rm nat}) \ll \langle E(\mathbf{A}, \mathbf{C} \rangle$.

2 Population Dynamics and Statistical Physics

If mutations with small fitness effects are considered, generalizing the neutral model, effective population size N becomes a key variable of the evolutionary process, since slightly deleterious mutations are more likely to be fixed in small populations [27–29]. This modeling approach has been pioneered by Ohta, who showed that population size can provide a possible explanation for empirical observations such as the generation time effect [30, 31]. Obligate intracellular lifestyle, such as that of endosymbiotic or parasitic bacteria, implies a strong reduction in effective population size due to bottlenecks upon transmission from one host to another. Inspired by Ohta's theory, computational studies have compared bacterial species displaying an obligate intracellular lifestyle with their free-living relatives, suggesting that the genes of intracellular bacteria evolve faster as a result of relaxed selection [32] (but Itoh et al. [33] give a different interpretation), and that their structural RNAs [34] and their proteins [35] are less stable than the orthologous macromolecules of free-living bacteria. Evolution experiments with virus and bacteria confirm the influence of small population size, demonstrating fitness loss in populations evolving under repeated bottlenecks [36, 37], and show that such a loss can be partly compensated by overexpressing chaperones that assist in protein folding [38]. These findings support the idea that fitness is reduced in small populations as a consequence of the reduction of protein folding stability.

When modeling population dynamics, two variables are of key importance, mutation rate per genome per generation μ and effective population size N. The product $N\mu$ determines the genetic variability of the population. In the rare mutation limit $N\mu \ll 1$, on average at most one mutation arises in the population at each generation, and the timescale for fixation of a neutral mutation, proportional to N, is smaller than the timescale for the appearance of a new mutation in the same genome, proportional to $1/\mu$, so that mutations do not interact. In this limit the population is genetically homogeneous, apart from mutants that arise from time to time. On the contrary, if $N\mu$ is not small, different alleles may exist simultaneously at macroscopic frequencies, and more than one mutation may exist in the same genome, which can produce interactions such as the hitch-hiking effect.

Recent theoretical work has shown that, in the rare mutation limit, the statistical properties of population genetics are formally equivalent to a statistical mechanical system, so that there is an exact analogy between the reduction of fitness for small populations and the increase of entropy for high temperature [39, 40]. Population genetics shows that the probability that the mutation is fixed in the population can be exactly computed as [41]

$$P_{\text{fix}}(\mathbf{A} \to \mathbf{A}') = \frac{1 - \dfrac{f(\mathbf{A})}{f(\mathbf{A}')}}{1 - \left[\dfrac{f(\mathbf{A})}{f(\mathbf{A}')}\right]^{N}}, \tag{5}$$

where N is the effective population size and $f(\mathbf{A})$ is the exponential growth rate of the phenotype associated to sequence $\mathbf{A}$, which will be called fitness in the following. This formula enormously simplifies both the analytic and the numeric study of evolution. It has been noted that the above formula, multiplied by the mutation probability from $\mathbf{A}$ to $\mathbf{A}'$, can be interpreted as the transition probability of a Markov process in sequence space. Such a Markov process admits a stationary distribution in which fitness fluctuates around an equilibrium value, with events of fitness increase and decrease being on the average equally likely. The stationary distribution can be computed analytically [39, 40], and it is given by

$$P_{\text{evol}}(\mathbf{A}) \approx P_{\text{mut}}(\mathbf{A}) \exp[N \log f(\mathbf{A})] , \tag{6}$$

where $P_{\text{mut}}(\mathbf{A})$ is the probability to obtain sequence $\mathbf{A}$ under mutation alone. The factor $\exp[N \log f(\mathbf{A})]$ is equivalent to a Boltzmann distribution, where the effective population size N plays the role of inverse temperature and the logarithm of fitness plays the role of minus energy. Thus the model predicts that smaller populations reach lower fitness, which means that their macromolecules are less stable and the mutational entropy in sequence space is larger.

Wright [28] generalized the stationary distribution (6) to the case where the product $N\mu$ is not small. This stationary distribution has also a deep analogy with statistical physics [42]. However, it has a simple expression only in the case of two alleles, in which case the probability to find the first allele with frequency x_1 and the second one with frequency $x_2 = 1 - x_1$ is

$$P(x_1, x_2) \propto x_1^{V_1 - 1} \mathrm{e}^{x_1 \log f_1} x_2^{V_2 - 1} \mathrm{e}^{x_2 \log f_2} , \tag{7}$$

where f_i is the fitness of allele i and $V_1 = N\mu u_{21}/(u_{12} + u_{21})$. The last factor represents the mutation bias from allele 2 to allele 1.

It would be interesting to further develop the analogy of statistical mechanics also in the case of potentially infinite alleles. This has only been done in the infinite population limit, where population dynamics can also be studied analytically, as a mutation can be fixed only if it is advantageous or neutral. In this limit, a large number of mutants arise at each generation, and the population can be represented as a distribution in the space of all possible genotypes, which is called a quasispecies [43]. Also, in this limit there is a formal analogy between population dynamics and statistical physics, where mutation rate plays the role of temperature [44, 45]. In the single peak landscape, where a unique master sequence has higher fitness than the sea of mutant sequences, this model undergoes a phase transition in which at low mutation rate a sizeable fraction of the population adopts the master sequence, whereas at high mutation rate the population is dispersed in a sea of less stable mutant sequences.

3 Substitution Rate and Mutational Robustness

Already in the early days of molecular evolution studies, it was recognized that the number of amino acid substitutions in the evolution of a protein sequence grows approximately linearly with the time of divergence [46]. This important observation, named the molecular clock, lies at the ground of several methods for reconstructing phylogenetic trees from the comparison of extant protein sequences. In this context, it is essential to distinguish between mutations and substitutions. Mutation is a process at the level of the individual, giving rise to a different genotype and producing a phenotypic effect on fitness (or no effect, if the mutation is neutral). Substitution is a process at the population level—the macroscopic level, in the language of statistical physics—and it consists in a mutation that gets fixed in the population, raising at a frequency of almost 100%.

The first and simplest explanation of the molecular clock was provided by Kimura's neutral model [47, 48], which is still one of the most influential models of molecular evolution. Within this model, a protein can contribute to the fitness of the organism that bears it in only two ways: either it is functional, and the organism is viable, or it is not functional, and the organism is not viable. All viable organisms are considered as equally fit, i.e., there are only two classes of phenotypes and the fitness is a binary variable. Within the neutral model, and provided that the mutation rate is small, the substitution rate does not depend on population size since the number of neutral mutations arising at each generation is $N\mu z$, where z is the fraction of mutations that are neutral, and the probability that a neutral mutation is eventually fixed in the population is $1/N$, in agreement with (5) in the limit $f \to f'$. Thus, the number of substitutions grows linearly with the divergence time with rate μz. The original neutral model also assumes that z is constant on the neutral network. This hypothesis implies that the number of neutral substitutions observed in a time t is a Poisson variable with mean value $\mu z t$. However, the variance of the number of substitutions is significantly larger than expected for a Poisson variable (overdispersion) [49]. This is not in contrast with the neutral model, since simulations of neutral evolution with structure conservation show that the fraction of neutral mutations $z(\mathbf{A})$ fluctuates strongly from one sequence to another (being larger for more stable sequences), and this in turn implies overdispersion [50].

It is instructive to study the neutral model in the infinite population limit [51]. In this limit, analytical and numerical evidence shows that the population is not uniformly distributed in the neutral model, but concentrates at sequences that have a large fraction of neutral neighbors, since this is the stationary limit of the quasispecies distribution. This has the same effect as if the population minimizes the mutation load, i.e., the fraction at which lethal mutants are eliminated. Therefore, in populations where $N\mu$ is large, mutational robustness is expected to spontaneously arise as a result of population dynamics. This in turn establishes a negative relationship between substitution rate and population size even in the case of neutral evolution.

As already stated above, if mutations with small fitness effects are present in the model, effective population size N becomes a key variable of the evolutionary

process, since slightly deleterious mutations are more likely to be fixed in small populations. This is the basis of the nearly neutral model proposed by Ohta [30]. Since small populations are more tolerant to mutations, the frequency z of effectively neutral mutations is larger in small populations, and the substitution rate increases.

Finally, if $N\mu$ is not small, multiple mutations, both beneficial and detrimental, may happen in the same genome. This effect complicates the substitution process considerably, since neutral and slightly deleterious mutations can hitchhike genomes with advantageous mutations that are positively selected, and therefore increase their substitution rate. When the advantageous mutation has been fixed, these passenger mutations are eliminated by purifying selection.

4 Translation Load

As discussed above, if $N\mu$ is not small, mutational robustness arises even in the neutral model as a consequence of population dynamics. Mutational robustness is correlated with stability, since a more stable protein is more tolerant to mutations. Therefore, a large mutation rate is expected to favor tolerance to mutations and a larger stability than would be expected in the monomorphic limit. However, the mutation rate per gene is very small in natural populations of bacteria and eukaryotes, so that it is doubtful that this mechanism is relevant to explain the high tolerance to mutations observed in natural proteins. A related explanation for this mutational robustness has been recently proposed by Wilke and collaborators [52]. When proteins are synthesized on the ribosome, translation errors may happen relatively frequently, since the accuracy of the ribosome is not very high (for a 200 amino-acids protein, a wrong amino acid is incorporated on the average every few replication cycles in *E. coli*). These translation errors may produce wrongly folded proteins that are not functional and, moreover, tend to aggregate; therefore, there is a strong selective pressure enforcing robustness against translation errors. This selective pressure is expected to be stronger for highly expressed proteins. This is in agreement with the observations that highly expressed proteins tend to be codified by optimal codons (which improve the accuracy of translation) and tend to evolve more slowly (since they are subject to stronger selective pressure) [52]. Another element that enforces robustness against translation errors is the genetic code. It has been shown that the standard genetic code is almost optimal for reducing the consequences of translation errors on the physiochemical properties of protein sequences [53]. Using the protein folding model that we adopted for evolutionary studies, we recently verified that the standard genetic code reduces the effects on protein folding stability of frequent translation errors with respect to existing alternative codes [54], although the advantage of the standard code is sometimes reduced for extreme mutation bias. Despite the importance of the translation load for protein evolution, however, this ingredient is seldom taken into account when modeling the fitness.

5 Protein Stability and Mutation Bias

Another key evolutionary variable, which has received little attention, is the nucleotide composition of the genome. In prokaryotes, it varies from extreme adenine plus thymine (AT) content in obligatory intracellular bacteria to extreme guanine plus cytosine (GC) content, for instance in actinobacteria. These differences in GC content are thought to be prevalently due to mutation bias [55, 56]. They are strongest at the third codon position, where GC content barely affects the amino acid composition of the protein, but also influences the coding positions [57, 58]. Due to the structure of the genetic code, a mutation bias favoring thymine at the nucleotide level favors the incorporation of hydrophobic amino acids in the translated protein [35, 59]. Hydrophobicity is a key property for protein folding [60]. Proteins that are too hydrophobic tend to misfold and aggregate, whereas proteins that are too hydrophilic tend to be naturally unfolded [61]. This qualitative trade-off between unfolding and misfolding was confirmed by a computational study of the properties of homologous proteins in the proteomes of several bacterial species, using a model of protein folding stability that correlates well with experimentally measured unfolding stabilities [35]. The trade-off between unfolding and misfolding stability is also clear if we consider the unfolding free energy (3). In this case, we can define $\Delta G = \Delta G_{\mathrm{u}} + \Delta G_{\mathrm{m}}$, with $\Delta G_{\mathrm{u}} = E(\mathbf{A}, \mathbf{C}^{\mathrm{nat}})/2 + k_{\mathrm{B}} T s_0 L$ and $\Delta G_{\mathrm{m}} = E(\mathbf{A}, \mathbf{C}^{\mathrm{nat}})/2 - \langle E(\mathbf{A}, \mathbf{C}) \rangle + \sigma_{\mathrm{E}}^2 / 2 k_{\mathrm{B}} T$. Using the hydrophobic approximation $U(a, b) \approx \varepsilon_{\mathrm{H}}\, h(a)\, h(b)$ (see (9) and (10) below) and writing $g_i = h(A_i)/\langle h \rangle$, where $\langle h \rangle$ is the mean hydrophobicity of the protein sequence, we see that $\Delta G_{\mathrm{u}} \approx -a \langle h \rangle^2 + sL$, with $a > 0$, so that stability against unfolding increases with hydrophobicity, whereas $\Delta G_{\mathrm{m}} \approx -b \langle h \rangle^2 + c \langle h \rangle^4$ with $b > 0$ and $c > 0$, so that stability against misfolding decreases with hydrophobicity when it is large. We and coworkers investigated the relationship between unfolding stability, misfolding stability, and mutation bias using a protein evolution model with a neutral fitness landscape. We indeed found that the mutation bias modulates the trade-off between the two kinds of stability, making proteins evolving under AT mutation bias more stable against unfolding but less stable against misfolding [62].

Interestingly, the two aspects discussed above, effective population size and mutation bias, are correlated in nature. In fact, most bacterial and eukaryotic species that adopted an intracellular lifestyle, with consequent reduction of their effective population size, also shifted their mutation spectrum toward AT [63], as indicated by the strong correlation between reduced genome size, which is a signature of intracellularity, and the AT bias [32, 35]. In order to investigate this relationship, we have modeled protein evolution in a nonneutral fitness landscape where fitness smoothly depends on stability against unfolding and against misfolding [25]. Mutations are randomly drawn at each step of the simulations according to a given mutation bias, and they are fixed in the population according to (5), since we assume that the population is monomorphic. Protein stability increases with effective population size, in agreement with theoretical expectations. Interestingly, for a given effective population size the fitness that can be achieved depends on the

mutation bias, and it is maximal at an optimal mutation bias. Simulations show that the optimal mutation bias favors AT ($AT_{opt} \approx 0.6$) for small effective population size, it favors GC ($AT_{opt} \approx 0.3$) for intermediate effective population size, and we find absence of bias ($AT_{opt} \approx 0.5$) for very large population size. These results may contribute to an explanation for why almost all intracellular bacteria evolving with small effective population size have developed a mutation bias toward AT.

6 Protein Size and Marginal Stability

The model reported above and similar ones also show that protein stability is higher for nonneutral evolution than for neutral evolution. In fact, within the neutral model, folding stability does not affect fitness as long as it is above threshold, so that the stability that is more likely to be observed in evolution is the viable stability more likely to arise by mutation, which is expected to coincide with the minimal stability compatible with viable molecules. This phenomenon may explain why natural proteins only possess a modest stability, which is relatively easy to increase through engineered mutations [24]. The same tendency toward marginal stability is also observed in a model of nonneutral evolution, where the fitness is described by (4) and the fixation probability (5) is used. Stability evolves to a stationary value that is determined by the mutation-selection balance and by temperature, where stability is marginal.

The mutation-selection balance also influences the dependence of protein stability on chain length. Unfolding stability balances an energetic term $E(\mathbf{A}, \mathbf{C}^{\rm nat}) = \sum_{ij} C_{ij}^{\rm nat} U(A_i, A_j)$ that grows with the number of native contacts $N_{\rm C}$, with a conformational entropy term that grows with the number of residues L. For globular proteins, the number of contacts per residue $N_{\rm C}/L$ increases with L as $N_{\rm C}/L \approx C_0 \left(1 - DL^{-1/3}\right)$, with $D \approx 3/2$, due to the reduction of the surface to volume ratio for larger proteins. Therefore, native contacts need to be less strong in longer proteins in order to compensate for conformational entropy loss upon folding, and the tendency of protein toward marginal stability predicts that native interactions are weaker in longer proteins. This hypothesis was confirmed by a statistical analysis of the PDB conducted by one of us and a coworker [23]. Similarly, examination of the energy gap α that measures stability against unfolding leads to the expectation that the Z-score of native interactions with respect to all possible interactions, both native and nonnative, must be more negative for shorter proteins, as it is indeed observed [23],

$$Z_{\rm nat} = \frac{\langle U \rangle_{\rm nat} - \langle U \rangle}{\sqrt{\langle U^2 \rangle - \langle U \rangle^2}} < -\sqrt{\frac{2(A + B/L)}{N_{\rm C}/L}}\,, \tag{8}$$

where $\langle U \rangle = \sum_{ij} \langle C_{ij} \rangle U(A_i, A_j) / \sum_{ij} \langle C_{ij} \rangle$ is the average interaction energy of alternative contacts and $\langle U \rangle_{\rm nat} = \sum_{ij} C_{ij}^{\rm nat} U(A_i, A_j) / \sum_{ij} C_{ij}^{\rm nat}$ is the average

interaction energy of native contacts. Therefore, native interactions tend to be weaker and less optimized in longer proteins. This may be interpreted as yet another manifestation of the tendency of proteins to have marginal stability.

7 Inverse Folding

Closely related to the question of protein evolution is the inverse folding problem. The inverse folding problem can be formulated as the study of the statistical properties of protein sequences that fold into a given target structure, and is of importance for bioinformatics applications such as structure prediction, as well as for theoretical modeling. Interestingly, it is possible to analytically solve the inverse folding problem within the hydrophobic approximation of the energy. This approximation exploits the fact that the contact energy matrix $U(a,b)$ is well approximated by its main eigenvector $h(a)$,

$$U(a,b) \approx \varepsilon_{\mathrm{H}}\, h(a)\, h(b)\,, \tag{9}$$

where $\varepsilon_{\mathrm{H}} < 0$ and the eigenvector $h(a)$ is related to the hydrophobicity of residue a [64,65]. Using this approximation, the native energy can be expressed as

$$E(\mathbf{A}, \mathbf{C}^{\mathrm{nat}}) \approx \varepsilon_{\mathrm{H}} \sum_{ij} C_{ij}^{\mathrm{nat}}\, h_i\, h_j\,, \tag{10}$$

where $h_i = h(A_i)$ is the hydrophobicity profile of sequence $\mathbf{A}$. It is immediately seen that the optimal hydrophobicity profile that minimizes the native energy for a given value of the mean squared hydrophobicity $\langle h^2 \rangle$ coincides with the principal eigenvector of the contact matrix. If we further impose a condition on the mean hydrophobicity $\langle h \rangle$ in order to constrain stability against unfolding, we find that the optimal hydrophobicity profile is proportional to the so-called effective connectivity (EC) profile c_i [66], a structural profile that almost coincides with the principal eigenvector for single domain proteins, and generalizes it for multi-domain proteins. The EC has large components in the core of the protein, where residues have many contacts, and small components on the surface. Thus, the optimal hydrophobicity profile $h_i^{\mathrm{opt}} \propto c_i$ expresses the well-known fact that buried positions tend to be hydrophobic and surface positions tend to be hydrophilic, but in a quantitative fashion. The optimal hydrophobicity profile is in very good agreement both with the hydrophobicity profile averaged over sequences obtained by simulating protein evolution with structural conservation, and with the hydrophobicity profile averaged over positions in the PDB that have similar EC components. We are currently investigating modifications of this framework that allow for enforcing stability against misfolding in a more explicit way.

The sequence that best matches the optimal hydrophobicity profile is analogous to the prototype sequence found by Bornberg-Bauer and Chan in simulations of

protein evolution [7]. Such a sequence is the most stable and most robust against mutations and, despite the fact that it is never encountered in evolution, it deeply influences the statistical properties of the family of sequences that fold into the same target structure.

The above scheme allows for computing site-specific amino acid distributions $P_i(a)$ at each site of a protein by maximizing the entropy in sequence space with a constraint on the average hydrophobicity, which enforces stability against unfolding. One finds that

$$P_i(a) \propto w(a) \exp(-\beta_i\, h(a)), \tag{11}$$

where $w(a)$ is the global frequency of amino acid a due to the mutation process and the genetic code, β_i is a site-specific Lagrange multiplier that enforces the appropriate value of the average hydrophobicity, and $h(a)$ is the hydrophobicity parameter of amino acid a. This framework assumes the independent site approximation, $P(A_1 \cdots A_L) \approx \prod_i P_i(A_i)$. This approximation is inevitable to get analytic insight, however, sites in a protein are not independent, in particular when they are in contact in the native structure. In these cases, compensatory or correlated mutations are revealed by statistical analysis of multiple sequence alignments, and they are often used to predict contacts for proteins of unknown structure [67]. We are currently generalizing the approach presented here using a pairwise approximation of the sequence entropy that allows for analytical prediction of pairwise amino acid distributions for a given protein structure.

8 Protein Structure Evolution

It is often stated that protein structure is much more conserved than sequence in evolution. This principle is the basis of bioinformatics methods that predict protein structure based on homology. It has been shown that proteins can diverge in sequence, reaching similarities typical of randomly related sequences, and yet conserve a similar fold [5, 68, 69]. This is possible because these sequences are not random, but share the same hydrophobic fingerprint [65].

A milestone in the study of protein structure divergence was a paper by Chothia and Lesk, who showed that the root mean square deviation (RMSD) between different globins diverges regularly with the number of amino acid substitutions, up to a limit of low sequence identity, where the RMSD suddenly explodes [70]. This result suggests a generalization of the molecular clock hypothesis to the evolution of protein structure, but has technical limitations since the RMSD can be used as a measure of structural divergence only for aligned residues that have good spatial superimposition. One of us and coworkers recently proposed a measure of structure divergence based on the contact overlap, which is more suitable for such a quantification since it minimizes the dependence on protein length both for evolutionarily related and for unrelated protein pairs [71]. Using this measure, we confirmed that protein structures diverge in a clock-like manner up to a very small

sequence identity ≈ 0.15, where structure divergence explodes. Interestingly, the explosion of structural divergence seems to take place only for proteins performing different functions, whereas proteins that share exactly the same function can diverge in structure only up to a limiting value of divergence, so that functional conservation imposes strong constraints on sequence and structure [71].

The simplest explanation for the explosion of structural divergence is that, below the crossover, sequence identity does not allow for estimation of the evolutionary divergence time, so that protein pairs with identity below the crossover may have diverged for a time much longer than what is inferred from their sequence identity. This simple explanation is supported by the fact that the sequence identity at the crossover decreases with protein length. Nevertheless, it is interesting that a qualitatively similar explosion of structural diversity has been found in a recent study of protein sequence design [72]. In this study, protein sequences were designed by optimizing the folding stability of a target structure. It was found that, when the target structure and the reference structure in the PDB are very similar, the designed sequence has a rather large identity with the reference sequence. However, when the target and the reference structure are more different, as it would be in the case of selection for new function, the designed and reference sequences only share very low identity, on the order of 20%, i.e., slightly more than the average identity of unrelated protein pairs. Therefore, the plot of sequence divergence versus structure divergence of designed protiens shows a crossover very much reminiscent of the one that we observed for evolved proteins and it may help to rationalize it: When two proteins perform the same function, natural selection targets very similar structures, determining sequence and structure conservation, whereas for proteins with significantly different function, natural selection targets different structures, whose typical sequence identities are below the crossover region. This interpretation is consistent with the findings, reported above, that protein function influences evolution by limiting the extent of sequence and structure divergence in the case of function conservation, and by accelerating structure divergence with respect to sequence divergence in the case of function change.

We also found that structure evolution is accelerated upon function change, since protein pairs with different functions diverge in structure at a rate significantly larger than those with the same function even before the explosion of structural divergence. Although not unexpected, this is an interesting result, since it demonstrates a quantitative influence of protein function on the sequence to structure relationship. Moreover, it suggests possible improvements to protein function prediction. In fact, it is known that very small changes in sequence and structure are sufficient to modify protein function, so that sequence and structure conservation are not a sufficient indication of function conservation. Our observation that function change modifies quantitatively the sequence to structure relationship suggests that this information could be used in order to predict function conservation more reliably.

Besides the quantification of the rate of structure divergence, another interesting result concerning protein structure evolution was the observation that protein structures tend to diverge along directions that overlap with the normal modes of low frequency [73]. This observation has been subsequently rationalized using one

of the few quantitative models of protein structure evolution, the linearly forced elastic network model (LFENM) proposed by Echave [74]. In the framework of the LFENM, a protein is modeled as an elastic network of contacts [75] and a mutation in the sequence is represented as a perturbation to the native structure that is directed along the contacts formed by the mutated residue. According to linear response theory, the response of the protein structure to a perturbation that produces a force f is $\Delta r \propto H^{-1} f$, where H is the Hessian matrix of the elastic network model. Therefore, the model predicts that structural changes in evolution have large components along low frequency normal modes, i.e., eigenvectors of the H matrix with small eigenvalue.

It was believed until recently that divergence in protein evolution in most cases conserves the fold, defined in colloquial terms as "the main arrangement of secondary structure elements of a protein." This assumption is at the basis of hierarchical structural classifications of proteins like SCOP [76] and CATH [77], where proteins with demonstrable homology, i.e., common origin, are automatically classified as belonging to the same structural class or fold. The distribution of the number of proteins in each fold is a power law, which is consistent with the view that folds have been populated through divergent evolution from a common ancestor followed by structural and functional differentiation within the given fold [78]. Nevertheless, it is increasingly recognized that the fold of a protein can change through evolution [79, 80], and that the fold as an equivalence class of protein structure is not consistently defined. In fact, proteins in the same fold of SCOP or CATH fail to fulfill the transitivity relation that, if two proteins a and b are both similar to the same protein c, they should be similar to each other. This transitivity relation is violated because protein domains do not only evolve in a monoparental way through gene duplication followed by clock-like structure and function divergence, but they also evolve from multiple parents, i.e., multiple fragments of supersecondary structure that are combined to give rise to a new fold and a new function. If protein c shares a fragment with protein a and a different fragment with protein b, transitivity is violated and proteins cannot be classified in a tree-like structure, but rather they must be described as a network. We have shown that, at high structural similarity, transitivity holds, and the traditional, tree-like view of protein evolution is justified, but for low similarity, structurally related proteins form a network rather than a tree [81].

As already stated, function is essential in the evolution of proteins. Function change is difficult to explain in an evolutionary framework, and several models have been proposed. It is beyond the scope of this chapter to review them here, but only to point out an interesting link to protein stability: The experimental work of the group of Tawfik [82] has recently shown that it is possible to evolve proteins that perform new functions, and that tolerance to mutations is crucial for improving evolvability. Chaperones that assist protein folding buffer the phenotypic effects of reduced folding stability, and in this way they enhance the ability of proteins to evolve new functions.

9 Conformation Changes

Of course, proteins are not static but perform their function dynamically. The study of how evolution modulates protein functional dynamics is still in its infancy, but some interesting observations have already been made. As it happens for structure divergence, conformation changes also tend to happen along directions related to the lowest frequency normal modes of the protein. The physical explanation of this phenomenon does not call evolution into play, but only linear response theory, since ligand binding can be seen as a perturbation to which the protein responds according to the matrix H^{-1}, where H is the Hessian matrix of the structure. According to a maximum Rényi entropy null model, the contribution of mode α to a conformation change produced by a random perturbation, c_α^2, will be proportional to the contribution of the same mode to thermal fluctuations, ω_α^{-2}. It is possible to define a parameter ρ that measures the deviation of the observed perturbation change from this null model [83],

$$\rho = \text{Correlation.coefficient}\left(c_\alpha^2\omega_\alpha^2, \omega_\alpha^2\right). \tag{12}$$

If the perturbation that produces the conformation change is random, the null model predicts that $c_\alpha^2\omega_\alpha^2$ is uncorrelated with the mode frequency. A positive and significant ρ hints that the perturbation is not random, but the normal modes have coevolved with it. In particular, $\rho > 0$ means that low frequency normal modes contribute to the conformation change more than expected on the basis of the null model, with the effect of reducing its free energy barrier and increasing its rate. Interestingly, $\rho \approx 0$ is found for conformation changes that are not functional, whereas positive and significant ρ is found for conformation changes that take place upon binding of transport proteins with their physiological substrate, among others.

10 Disordered Proteins

In this review chapter, we have discussed proteins that fold to a well-defined native state. Nevertheless, it is increasingly recognized that a large fraction of eukaryotic proteins have large unstructured loops that are important for function, and some are totally unstructured in their native state, becoming partially structured upon interaction with their interaction partner [84]. Disordered proteins are preferentially involved in molecular recognition and regulatory function [85]. Folding upon binding is thought to favor high specificity but low affinity interactions that can be finely modulated. Moreover, disordered proteins can form multiple conformations, allowing them to bind multiple partners, which increases the complexity of intermolecular interactions in eukaryotes.

From the point of view of stability, disordered proteins do not possess hydrophobic cores and contain many charged residues. Modulation of the charge

of disordered proteins modulates their size [86] and their conformation. It will be interesting in the future to model protein evolution in such a way that disordered proteins naturally arise at one side of the flexibility versus stability spectrum.

11 Conclusions

Although the models presented in this section still have a very limited ability to represent some crucial aspects of the protein world, such as function, dynamics, disorder, and the complexity of molecular dynamics, in our opinion they clearly show that folding stability, both against unfolding and against misfolding, represents a useful proxy of the genotype to phenotype relationship in proteins, and it allows for rationalization of some important aspects of their molecular evolution while, on the other hand, evolution is an essential ingredient for understanding the thermodynamic properties of natural proteins.

Acknowledgments We gratefully acknowledge our past and present collaborators in this field: David Abia, LLoyd Demetrius, Miriam Fritsche, Raul Méndez, Gonzalo S. Nido, Jonas Minning, Alberto Pascual-García, H. Eduardo Roman, Christoph Schmitt, Stefanie Sammet, Florian Teichert, and Michele Vendruscolo. Our research has been funded by several agencies over the years, and we wish to specifically mention financial support by the Spanish Science and Innovation Ministry ("Ramón y Cajal" and "Acciones Integradas España-Alemania" programs), the Deutscher Akademischer Austauschdienst ("Acciones Integradas España-Alemania" program) and the Deutsche Forschungsgemeinschaft (Normalverfahren and Heisenberg program).

References

1. Schuster, P., Fontana, W., Stadler, P.F., Hofacker, I.L.: From sequences to shapes and back – A case-study in RNA secondary structures. Proc. R. Soc. London B **255**, 279–284 (1994)
2. Schuster, P., Stadler, P.F.: Modeling conformational flexibility and evolution of structure: RNA as an example. In: Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M. (eds.) Structural Approaches to Sequence Evolution, pp. 3–36. Springer, Heidelberg (2007)
3. Hofacker, I.L.: Vienna RNA secondary structure server. Nucl. Ac. Res. **31**, 3429–3431 (2003)
4. Huynen, M.A., Stadler, P.F., Fontana, W.: Smoothness within ruggedness: the role of neutrality in adaptation. Proc. Natl. Acad. Sci. USA **93**, 397–401 (1996)
5. Babajide, A., Hofacker, I.L., Sippl, M.J., Stadler, P.F.: Neutral networks in protein space. Fol. Des. **2**, 261–269 (1997)
6. Govindarajan, S., Goldstein, R.A.: On the thermodynamic hypothesis of protein folding. Proc. Natl. Acad. Sci. USA **95**, 5545–5549 (1998)
7. Bornberg-Bauer, E., Chan, H.S.: Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. Proc. Natl. Acad. Sci. USA **96**, 10689–10694 (1999)
8. Bussemaker, H.J., Thirumalai, D., Bhattacharjee, J.K.: Thermodynamic stability of folded proteins against mutations. Phys. Rev. Lett. **79**, 3530–3533 (1997)
9. Tiana, G., Broglia, R.A., Roman, H.E., Vigezzi, E., Shakhnovich, E.I.: Folding and misfolding of designed proteinlike chains with mutations. J. Chem. Phys. **108**, 757–761 (1998)
10. Mirny, L.A., Abkevich, V.I., Shakhnovich, E.I.: How evolution makes proteins fold quickly. Proc. Natl. Acad. Sci. USA **95**, 4976–4981 (1998)

11. Dokholyan, N.V., Shakhnovich, E.I.: Understanding hierarchical protein evolution from first principles. J. Mol. Biol. **312**, 289–307 (2001)
12. Parisi, G., Echave, J.: Structural constraints and emergence of sequence patterns in protein evolution. Mol. Biol. Evol. **18**, 750–756 (2001)
13. DePristo, M.A., Weinreich, D.M., Hartl, D.L.: Missense meanderings in sequence space: a biophysical view of protein evolution. Nature Rev. Genet. **6**, 678–687 (2005)
14. Bloom, J.D., Silberg, J.J., Wilke, C.O., Drummond, D.A., Adami, C., Arnold, F.H.: Thermodynamic prediction of protein neutrality. Proc. Natl. Acad. Sci. USA **102**, 606–611 (2005)
15. Vendruscolo, M., Kussell, E., Domany, E.: Recovery of protein structure from contact maps. Fol. Des. **2**, 295–306 (1997)
16. Bastolla, U., Farwer, J., Knapp, E.W., Vendruscolo, M.: How to guarantee optimal stability for most representative structures in the protein data bank. Proteins **44**, 79–96 (2001)
17. Guerois, R., Nielsen, J.E., Serrano, L.: Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J. Mol. Biol. **320**, 369–87 (2002)
18. Berezovsky, I.N., Zeldovich, K.B., Shakhnovich, E.I.: Positive and negative design in stability and thermal adaptation of natural proteins. PLoS Comput. Biol. **3**, e52 (2007)
19. Noivirt-Brik, O., Horovitz, A., Unger, R.: Trade-off between positive and negative design of protein stability: from lattice models to real proteins. PLoS Comput. Biol. **5**, e1000592 (2009)
20. Lockless, S.W., Ranganathan, R.: Evolutionarily conserved pathways of energetic connectivity in protein families. Science **286**, 295–299 (1999)
21. Shakhnovich, E., Gutin, A.: Enumeration of all compact conformations of copolymers with random sequence of links. J. Chem. Phys. **93**, 5967–5972 (1990)
22. Derrida, B.: Random energy model: an exactly solvable model of disordered systems. Phys. Rev. B **24**, 2613–2626 (1981)
23. Bastolla, U., Demetrius, L.: Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds. Protein Eng. Des. Sel. **18**, 405–415 (2005)
24. Goldstein, R.A.: The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins **79**(5), 1396–407 (2011)
25. Mendez, R., Fritsche, M., Porto, M., Bastolla, U.: Mutation bias favors protein folding stability in the evolution of small populations. PLoS Comp. Biol. **6**, e1000767 (2010)
26. Zeldovich, K.B., Chen, P., Shakhnovich, E.I.: Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc. Natl. Acad. Sci. USA **104**, 16152–16157 (2007)
27. Muller, H.J.: Some genetic aspects of sex. Am. Nat. **66**, 118–138 (1932)
28. Wright, S.G.: The distribution of gene frequencies in populations of polyploids. Proc. Natl. Acad. Sci. USA **24**, 372–377 (1938)
29. Fisher, R.A.: The Genetical Theory of Natural Selection. Dover, New York (1958)
30. Ohta, T.: Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor. Pop. Biol. **10**, 254–275 (1976)
31. Graur, D., Li, W.H.: Fundamentals of Molecular Evolution. Sinauer, Sunderland (2000)
32. Moran, N.A.: Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. USA **95**, 4458–4462 (1996)
33. Itoh, T., Martin, W., Nei, M.: Acceleration of genomic evolution caused by enhanced mutation rate in endocellular bacteria. Proc. Natl. Acad. Sci. USA **99**, 12944–12948 (2002)
34. Lambert, D.J., Moran, N.A.: Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria. Proc. Natl. Acad. Sci. USA **95**, 4458–4462 (1998)
35. Bastolla, U., Moya, A., Viguera, E., van Ham, R.C.H.J.: Genomic determinants of protein folding thermodynamics. J. Mol. Biol. **343**, 1451–1466 (2004)
36. Duarte, E., Clarke, D., Moya, A., Domingo, E., Holland, J.: Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet. Proc. Natl. Acad. Sci. USA **89**, 6015–6019 (1992)
37. Novella, I.S., Dutta, R.N., Wilke, C.O.: A linear relationship between fitness and the logarithm of the critical bottleneck size in vesicular stomatitis virus populations. J. Virol. **82**, 12589–12590 (2008)
38. Fares, M.A., Ruiz-Gonzalez, M.X., Moya, A., Elena, S.F., Barrio, E.: Endosymbiotic bacteria: GroEL buffers against deleterious mutations. Nature **417**, 398 (2002)

39. Berg, J., Willmann, S., Lässig, M.: Adaptive evolution of transcription factor binding sites. BMC Evol. Biol. **4**, 42 (2004)
40. Sella, G., Hirsh, A.E.: The application of statistical physics to evolutionary biology. Proc. Natl. Acad. Sci. USA **102**, 9541–9546 (2005)
41. Durrett, R.: Probability Models for DNA Sequence Evolution. Springer, Heidelberg (2002)
42. Barton, N.H., Coe, J.B.: On the application of statistical physics to evolutionary biology J. Theor. Biol. **259**, 317–324 (2009)
43. Eigen, M.: Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften **58**, 465–523 (1971)
44. Leuthäusser, I.: Statistical mechanics of Eigen's evolution model. J. Stat. Phys. **48**, 343–336 (1987)
45. Tarazona, P.: Error thresholds for molecular quasispecies as phase transitions: from simple landscapes to spin-glass models. Phys. Rev. A **45**, 6038–6050 (1992)
46. Bromham, L., Penny, D.: The modern molecular clock. Nature Rev. Genet. **4**, 216–224 (2003)
47. Kimura, M.: Evolutionary rate at the molecular level. Nature **217**, 624–626 (1968)
48. Kimura, M.: The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge (1983)
49. Gillespie, J.H.: The Causes of Molecular Evolution. Oxford University Press, New York (1991)
50. Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M.: Connectivity of neutral networks, overdispersion and structural conservation in protein evolution. J. Mol. Evol. **56**, 243–254 (2003)
51. van Nimwegen, E., Crutchfield, J.P., Huynen, M.: Neutral evolution of mutational robustness. Proc. Natl. Acad. Sci. USA **96**, 9716–9720 (1999)
52. Drummond, D.A., Wilke, C.O.: Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell **134**, 341–352 (2008)
53. Freeland, S.J., Knight, R.D., Landweber, L.F., Hurst, L.D.: Early fixation of an optimal genetic code. Mol. Biol. Evol. **17**, 511–518 (2000)
54. Sammet, S.G., Bastolla, U., Porto, M.: Comparison of translation loads for standard and alternative genetic codes. BMC Evol. Biol. **10**, 178 (2010)
55. Muto, A., Osawa, S.: The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA **84**, 166–169 (1987)
56. Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., McAdams, H.: Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. USA **101**, 3480–3485 (2004)
57. Sueoka, N.: Correlation between base composition of the deoxyribonucleic acid and amino acid composition of proteins. Proc. Natl. Acad. Sci. USA **47**, 469–478 (1961)
58. Bernardi, G., Bernardi, G.: Codon usage and genome composition. J. Mol. Evol. **24**, 1–11 (1985)
59. D'Onofrio, G., Jabbari, K., Musto, H., Bernardi, G.: The correlation of protein hydropathy with the base composition of coding sequences. Gene **238**, 3–14 (1999)
60. Kauzmann, W.: Some factors in the interpretation of protein denaturation. Adv. Protein Chem. **14**, 1–63 (1959)
61. Uversky, V.N.: Protein folding revisited. A polypeptide chain at the folding – misfolding – nonfolding cross-roads: which way to go? Cell. Mol. Life Sci. **60**, 1852–1871 (2003)
62. Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M.: A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank. BMC Evol. Biol. **6**, 43 (2006)
63. Silva, F., Latorre, A., Gomez-Valero, L., Moya, A.: Genomic changes in bacteria: from free-living to endosymbiotic life. In: Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M. (eds.) (2007) Structural Approaches to Sequence Evolution, pp. 149–168. Springer, Heidelberg (2008)
64. Li, H., Tang, W.: Nature of driving force for protein folding: a result from analyzing the statistical potentials. Phys. Rev. Lett. **79**, 765–768 (1997)

65. Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M.: Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. Proteins **58**, 22–30 (2005)
66. Bastolla, U., Ortiz, A.R., Porto, M., Teichert, F.: Effective connectivity profile: a structural representation that evidences the relationship between protein structures and sequences. Proteins **73**, 872–888 (2008)
67. Göbel, U., Sander, C., Schneider, R., Valencia, A.: Correlated mutations and residue contacts in proteins. Proteins **18**, 309–317 (1994)
68. Rost, B.: Protein structures sustain evolutionary drift. Fol. Des. **2**, S19–S24 (1997)
69. Bastolla, U., Roman, H.E., Vendruscolo, M.: Neutral evolution of model proteins: diffusion in sequence space and overdispersion. J. Theor. Biol. **200**, 49–64 (1999)
70. Chothia, C., Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. EMBO J. **5**, 823–826 (1986)
71. Pascual-Garcia, A., Abia, D., Méndez, R., Nido, G.S., Bastolla, U.: Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation. Proteins **78**, 181–196 (2010)
72. Ding, F., Dokholyan, N.V.: Emergence of protein fold families through rational design. PLoS Comp. Biol. **2**, e85 (2006)
73. Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., Ortiz, A.R.: An analysis of core deformations in protein superfamilies. Biophys J. **88**, 1291–1299 (2005)
74. Echave, J.: Evolutionary divergence of protein structure: the linearly forced elastic network model. Chem. Phys. Lett. **457**, 413–416 (2008)
75. Tirion, M.M.: Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett. **77**, 1905–1908 (1996)
76. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: SCOP: A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**, 536–540 (1995)
77. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M.: CATH – A hierarchic classification of protein domain structures. Structure **5**, 1093–1108 (1997)
78. Dokholyan, N.V., Shakhnovich, B., Shakhnovich, E.I.: Expanding protein universe and its origin from the biological big bang. Proc. Natl. Acad. Sci. USA **99**, 14132–14136 (2002)
79. Grishin, N.V.: Fold change in evolution of protein structures. J. Struct. Biol. **134**, 167–185 (2001)
80. Viksna, J., Gilbert, D.: Assessment of the probabilities for evolutionary structural changes in protein folds. Bioinformatics **23**, 832–841 (2007)
81. Pascual-García, A., Abia, D., Ortiz, A.R., Bastolla, U.: Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. PLoS Comput. Biol. **5**, e1000331 (2009)
82. Soskine, M., Tawfik, D.S.: Mutational effects and the evolution of new protein functions. Nature Rev. Genet. **11**, 572–582 (2010)
83. Mendez, R., Bastolla, U.: Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins. Phys. Rev. Lett. **104**, 228103 (2010)
84. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., Jones, D.T.: Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. **337**, 635–645 (2004)
85. Uversky, V.N., Dunker, A.K.: Understanding protein non-folding. Biochim. Biophys. Acta **1804**, 1231–1264 (2010)
86. Mao, A.H., Crick, S.L., Vitalis, A., Chicoine, C.L., Pappu, R.V.: Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proc. Natl. Acad. Sci. USA **107**, 8183–8188 (2010)

Modeling Proteins at the Interface of Structure, Evolution, and Population Genetics

Ashley I. Teufel, Johan A. Grahnen, and David A. Liberles

1 Introduction

Biological systems span multiple layers of organization and modeling across layers of organization enables inference that is not possible by analyzing just one layer. An example of this is seen in an organism's fitness, which can be directly impacted by selection for output from a metabolic or signal transduction pathway. Even this complex process is already several layers removed from the environment and ecosystem. Within the pathway are individual enzymatic reactions and protein–protein, protein–small molecule, and protein–DNA interactions. Enzymatic and physical constants characterize these reactions and interactions, where selection dictates ranges and thresholds of values that are dependent upon values for other links in the pathway. The physical constants (for protein–protein binding, for example) are dictated by the amino acid sequences at the interface. These constants are also constrained by the amino acid sequences that are necessary to maintain a properly folded structure as a scaffold to maintain the interaction interface. As sequences evolve, population genetic and molecular evolutionary models describe the availability of combinations of amino acid changes for selection, depending in turn on parameters like the mutation rate and effective population size. As the systems biology level of constraints has not been thoroughly characterized, it is this multiscale modeling problem that describes the interplay between protein biophysical chemistry and population genetics/molecular evolution that we will describe.

There are three main trajectories in multiscale modeling at the interface of protein structure and evolutionary biology. The first trajectory involves simulation and forward evolution to describe both biophysical and evolutionary processes.

A.I. Teufel • J.A. Grahnen • D.A. Liberles (✉)
Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA
e-mail: ateufel@uwyo.edu; jgrahnen@uwyo.edu; liberles@uwyo.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_15, © Springer Science+Business Media, LLC 2012

In this trajectory, the assumptions and limitations are explicit and well-controlled hypotheses can be formulated. The next two trajectories are retrospective. First, there is a growing field adding thermodynamic complexity to phylogenetic models and this will be described. Second, another growing field involves the use of standard population genetic models for interspecific evolution, also incorporating increasing biophysical reality in parameterizing the models. These will be described in turn.

2 Simulation and Forward Evolution

Simulation and comparative genomic analysis represent complementary trajectories in evolutionary analysis. As events become more ancient, reconstruction of evolutionary histories and substitutions underlying the sequence–structure–function link become harder to recover. Further, interpretation of underlying mechanisms and controlling data for numerous potential variables become potentially problematic. Therefore, simulations of evolutionary processes with different mechanisms and explicit assumptions can provide insights into evolutionary pathways that are difficult to obtain from comparative genomic data.

Simulation approaches have a long history in the field of evolutionary biology, for example, in the context of population genetics [1]. Calculations on a population-wide level of variation in allele frequencies due to different fitness effects are not computationally demanding and provide insights into population-level processes. However, because of the lack of biochemical detail in such models, they tell us very little about the mechanisms that underlie fitness changes. To model the effects of mutations on molecular evolution, it is necessary to have at least a minimal representation of protein structure.

The most basic such model is the hydrophobic–polar lattice model [2]. The protein is represented as a series of interconnected beads on a rectangular grid, in either two or three dimensions, each of which represents either a hydrophobic or a polar residue (Fig. 1a). Hydrophobic residues interact favorably, mimicking the solvation pressure to form a hydrophobic core, and other interactions are typically neutral or repulsive. This simple model enables sampling a very large number of configurations rapidly, and in the two-dimensional case it even allows a complete enumeration and examination of the entire sequence–structure–fitness landscape. Bornberg-Bauer and Chan [2] used this technique to examine the distribution of thermodynamic stability for sequences undergoing neutral evolution, and found that the funnel-like behavior of the energy landscape of protein folding is recapitulated by the sequence landscape. Melin and co-workers [3] found two criteria that apply to all protein-like sequences under such a model. The native state is highly designable (robust to mutation) and is well separated on the energy landscape from random configurations. These properties are similar to those observed in real proteins.

The next level of complexity involved the usage of the 20 amino acids in lattice simulations. Sali and co-workers [4] examined the folding process using a 3D

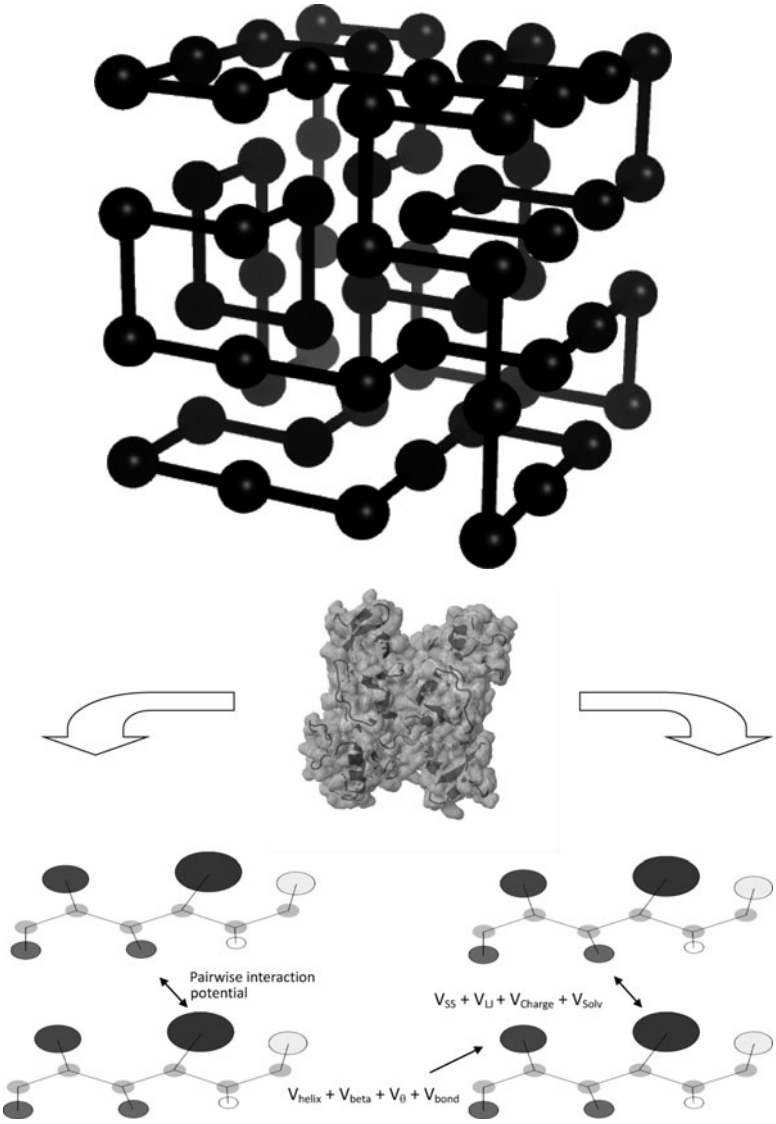

Fig. 1 (**a**) A hydrophobic–polar lattice model of protein structure, where interactions are defined contacts based upon the structure of the lattice and the type of each residue, is shown. (**b**) On *top*, the backbone depicted as a cartoon of the secondary structure threaded through the van der Waals surface of a GRB2 protein (PDB ID: 1GRI) is shown. Below on the *left*, contacts similar to those used in lattice models are used to calculate pairwise interaction potentials. On the *right*, a scoring function with some terms rooted in physics is used to evaluate mutations, based upon both terms that affect both interactions and those that are intrinsic properties of the amino acid, including the propensities to be involved in different secondary structural elements

cubic lattice model with residue–residue contact energies derived from those found in known protein structures [5]. They were able to observe that folding rapidly progresses from random coil to semicompact globule, and then more slowly finds the exact native contacts, suggesting that Levinthal's paradox [6] is solved by the vast reduction in available conformations after the hydrophobic collapse. On the same trajectory, Li and co-workers [7] further refined the lattice model to include side chains and allowed a noncubic shape for the protein.

Although thermodynamic stability and kinetics are both important factors in protein evolution, ultimately the fitness effect due to mutation on any protein is contingent on its function. Protein–protein interaction is a common type of functionality and is amenable to study with the lattice model system. A common strategy is to represent the interaction as occurring between the protein and a small peptide that binds to some portion of the lattice surface [8]. To simplify interpretation it is often assumed that the peptide sequence does not change throughout the simulations, although the methodology can be adapted to study the co-evolution of interacting proteins. Bloom and co-workers [9] examined the effects of selection for several ligands under a similar model in more detail. The evolution of both folding and binding stability were found to be intertwined in this model, providing somewhat opposing forces on the protein, and sequence evolution most often causes small steps in structure space rather than large backbone modifications.

More recently, Rastogi and Liberles [10] employed this class of model to examine the fates of duplicated genes, where binding to extant and novel ligands was the function. This study used lattice models to address a question in evolutionary biology that only indirectly involved protein folding thermodynamics, establishing subfunctionalization as a transition state to neofunctionalization as an important evolutionary mechanism that is also now supported by comparative genomic data [11]. Massey et al. [12] used the lattice modeling framework to address the role of positive selection in driving molecular convergence and its ultimate phylogenetic effect. Here, the sequences that enabled a new binding interaction were independently evolved from different starting points in a sequence simulation on a phylogenetic tree, demonstrating that it was possible for reconstructed phylogenetic trees from such sequences to show a signal for common function rather than common ancestry. A similar effect was subsequently observed in comparative genomic data from snake and lizard mitochondrial genomes [13].

Population dynamics represent an additional consideration when modeling evolution. Most of the work that does not explicitly address evolutionary questions models evolution as a random or adaptive walk on the sequence/fitness landscape by a single sequence reflecting a population size of one. Except for very strong fitness effects, such simulations are likely to bury true evolutionary signal in the stochastic noise of immediate fixation of neutral and slightly deleterious changes. By adding population-level complexity to the simulation, one can also ask additional questions about the influence of population size, mutation rate, and population dynamics on the process of molecular evolution. Taverna and Goldstein [14] used a lattice model to show that protein-like mutational robustness and marginal stability only emerged in simulations when population dynamics were properly considered.

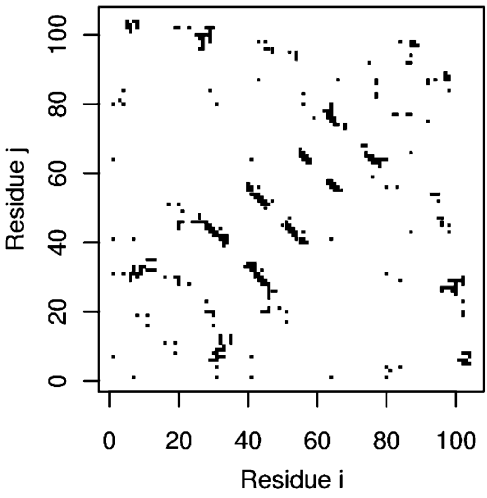

Fig. 2 The symmetrical contact map excluding contacts within four amino acid positions derived from the structure of the SAP protein (PDB ID 1D4T) at a distance of 4.5 Å is shown. This enables the definition of residues in contact with each other for the use of informational potentials

Although lattice models have provided a wealth of information on the rules that govern proteins in general, they are not well suited to describing differences between distinct protein folds with respect to either function or evolution. Simulating the evolution of sequences in real protein conformations or folds requires a higher level of detail. Currently, the most popular approach involves creating a native contact map from an experimentally determined structure (see Fig. 2). This allows one to specify the effect of backbone shape on which residues are near each other, and hence how their proximity and its effect on the overall folding energy constrains their mutual mutational opportunity. A pairwise residue interaction energy is typically assumed [15] (Fig. 1b), although this may also be augmented with a term representing the effects of solvation [16]. Since the contact map never changes, one need only to evaluate the effects of a substitution on a relatively small number of local interactions. Parisi and Echave [17] adopted this level of structural description to simulate the evolution of a left-handed beta-helix domain with a known and specific sequence pattern thought to be due to folding constraints. Using a random walk simulation with an energy difference-based fitness criterion, they demonstrated that it is possible to reproduce the specific pattern of sequences of this fold. This suggests that particular folds exert quite specific evolutionary pressures that constrain the variation of sequences within a protein family. Bastolla and co-workers [18] showed that overdispersed substitution (deviation from the expected Poisson distribution) can result from neutral evolution under structural constraint rather than selection and that the rate of substitution can vary considerably between populations for the same reason.

An interesting contrast to the contact map is presented by the coarse-grained (CG) physics-based approach (Fig. 1b). Instead of relying on pairwise interactions between residues, which inherently ignores multiresidue effects, a variety of descriptors of separate forces is applied to a protein model with a reduced level of

detail to score the probability for a sequence to adopt a particular conformation. This increased level of detail enables study of the effect of both the specific interactions and the specific underlying forces through evolution.

One particular concern that has not been addressed in detail in the literature is the relationship between $\Delta\Delta G$ (change in scoring function value) and evolutionary fitness. Many studies either employ folding thresholds or use a goodness-of-fit criterion, but more attention to this will need to be paid to this relationship in the future.

Rastogi et al. [19] introduced a scoring function for folding and binding based upon that of Mukherjee and Bagchi [20], readapted for evolutionary simulation. This approach included terms for the bonding potential, the bending potential, the torsional potential, the Leonard-Jones potential, and the propensities for amino acids to lie in helical or sheet regions.

Grahnen and co-workers [21] expanded upon this, incorporating additional terms for solvation and for disulphide bridges, using a structural representation where side chain positions are modified during simulation and the folding potential is parameterized for each individual protein. This was found to outperform a pair-wise interaction approach with respect to biophysical realism and fold specificity (although further improvement in the model with regard to packing of the coarse-grained core of the protein will be needed before biological application of these models), and is fast enough to be applied to more complex situations involving multiple proteins and large population sizes, enabling future studies in molecular evolution.

In recapitulating selective pressures on proteins, it is important to consider not only the energy gap between folded and unfolded states and bound and unbound states, but also any selective pressures on what not to fold into or bind. Inherent in considerations of folding is the consideration that the native structure is favored by an energy gap to alternative structures, and this is considered in simulations with a folding decoy. Selective pressures on what not to bind may also be an important aspect of functional specificity in cells and such explicit selective constraints can be incorporated into simulations as well (see [22)].

As more computationally intensive methods become tractable with increases in computational resources, a number of future steps can be envisioned. The coarse-grained approaches can be replaced by all-atom representation of protein structure, possibly using some variation on modern molecular dynamics force fields [23–25]. In the future, one would expect electron-level resolution via density functional theory or other quantum-mechanical approaches to become possible. Also, incorporating simulations of chemical reactions at the electron level would open up the study of enzymatic function, which would be expected to exert different selective pressures than protein–protein interactions. The trajectory involving contact maps could in the near future be augmented by allowing the contact maps to vary throughout the simulation, which would involve some estimation of side chain or backbone angle modifications due to mutation. An intriguing prospect for simulation of very long evolutionary processes, such as evolution between entirely different fold types, is the application of methods such as elastic network models

[26] in a simulation context rather than for comparative purposes. Adopting existing protein structure prediction protocols to model the large effect of insertions and deletions on backbone structure would also be of interest.

In another trajectory, evolutionary systems biology is an emerging field. Simulations with more than one protein can enable evaluation of the interactome, with pathway-level selective pressures rather than selective pressures on individual binding interactions. Ultimately, this would sum up to simulation of a cellular network on an evolutionary timescale. The logical endpoint in that direction is a marriage of sequence–structure simulation to the metabolic and transcriptional rate simulations common in systems biology today. A second goal in this trajectory is the simulation of cross-species interactions, such as those involved in a viral or bacterial infection or in the molecular interactions between organisms that co-exist in an ecosystem. The more interconnections between layers of simulation that are taken into consideration, the richer our description of evolution becomes and the questions that can be asked multiply in both significance and number.

3 Phylogeny and Thermodynamics

Phylogenetic inference is another problem where structural models can enable a greater level of understanding of the evolutionary process. Phylogenetic analysis is a retrospective problem that enables complementary inference of evolutionary processes. Models of phylogeny are generally formulated as continuous time Markov chains, in which branch lengths are a function of the probability of substitutions. As the probability of a mutation being accepted is impacted by the disturbances which it may cause the protein energetically, there has been an effort to incorporate biophysical reality into phylogenetic models in order to improve their accuracy. These efforts have drawn upon the same classes of models as used in the evolutionary simulation field. Using models which consider the constraints of structure [17], one can evaluate the probability of fixing a mutation, resulting in a particular evolutionary history, in the context of thermodynamic stability. This approach lends itself naturally to incorporation in maximum likelihood (ML) or Bayesian methods for phylogenetic tree construction. For instance, Parisi and Echave [27] apply these considerations to phylogenetic inference. They make use of a model, in which sequences are allowed to mutate and then are selected upon based on structural constraints. The selection criterion is based on the difference in contact potential between the position mutated and the nonmutated sequence. This selection criterion can then be scaled by a parameter which is related to the selection pressure. To apply this model in a phylogenetic context, site-specific replacement matrices are calculated based on the contact potential score, amino acid equilibrium frequencies, and a count matrix. The replacement matrices produced by this method are then used to calculate the maximum likelihood of a data set based on a given topology.

A further improved method which takes advantage of biochemical properties with less structural specificity is known as the CAT model [28]. The CAT model uses an infinite mixture model with a Dirichlet prior distribution to describe the substitution process. This mixture model is constructed on the premise that a substitution will generally have similar biochemical properties to the amino acid which it replaces. Positions on the protein which share similar biochemical properties can be grouped into categories. These categories are then defined by stationary probabilities [29].

Rodrigue et al. [30] take another step in the direction of incorporating thermodynamic reality by making use of statistical potentials, similar to the contact potentials employed by Parisi and Echave [27]. In this model, the change in contact potential is scaled such that when the scalar $\beta = 0$, interdependent changes are not considered. Since the Bayes factor can be computed for any value of β, this allows examination of which values of β result in the best fit of the model and the value of β relays the relative importance of the difference from the contact potentials. It was found that models which considered site interdependence always outperformed models which assumed independence.

Nasrallah et al. [31] examined the impact of site-dependent evolution on phylogenetic inference. An approach similar to that of Rodrigue et al. [30] was used to simulate dependent sequence evolution and phylogenies created from this simulated data, finding that increased levels of dependence in the simulated data resulted in decreased levels of accuracy in the constructed phylogenies. In Rodrigue et al. [30], the complexity of structure has essentially been reduced to two parameters, and limitations exist as to what only two measures of a vastly complicated structure can reconstitute. It is suggested that a more accurate description of fitness is needed for such methods to reach their potential.

Recently, based on measurements inferred from coarse grain models similar to those of Rastogi et al. [19], Kleinman et al. [32] proposed an energy score that is based on combinations of a set of factors including secondary structure, contacts, interresidue distance, solvent accessibility, torsion angles, and flexibility. These factors were considered linearly and parameters related to them can be optimized. Models which assume different combinations of these can then be compared based on the β-value which results in the best fit.

Despite the increase in model fitness, Kleinman et al. [32] find that their model is not a significant improvement over site-independent models. Specific assumptions made about the relationship between fitness and change in scoring function, the lack of consideration of negative design, and potentials that are too general may underlie the performance of this method.

Current amino acid-level models for phylogenetic analysis do not offer a sufficient level of biophysical realism and structural models for this purpose are in their infancy. Further development of models like those of Kleinmann et al. [32] and Grahnen et al. [21] will enable this field to advance. Better amino acid-level models will enable more accurate reconstruction of ancient evolutionary histories that are less likely to be confounded by problems like functional convergence.

4 Population Genetics and Biophysical Constraints in Models for Interspecific Evolution

To understand the process of interspecific molecular evolution, the factors which influence substitution and selection must be recognized and the interplay between them understood. There exists a broad range of features which can affect substitution and selection, which may be more or less influential dependent on a given scenario. To measure selective strength, some models examine the ratio of nonsynonymous to synonymous changes, where an acceleration of the nonsynonymous substitution rate is indicative of positive selection. Similar comparisons of the ratio of these rates can be made to determine if a site is undergoing neutral evolution or purifying selection. Due to the interest in these ratios, it follows that codon-level models where synonymous and nonsynonymous changes can be observed have begun to proliferate. These models allow for formulations of the substitution and the mutation-selection process to be combined in a single framework which can be related back to known data from genomes. Measures of selective pressures or selective strength can then be estimated using maximum likelihood or Bayesian methods.

A basic implementation of this type of model is given by Halpern and Bruno [33]. This model assumes the mutation process (rates) is constant across all codons but natural selection acts differently depending on the position and the nature of the amino acid side chain. The model does not explicitly consider structure. Halpern and Bruno represent the substitution rate between an amino acid i and amino acid j as

$$R_{ij} = \mu\mu_{ij} NP(Z_{ij}),$$

where μ is a proportionality constant, μ_{ij} is the rate at which i mutates to j, N is the haploid population size, and Z_{ij} is the probability of fixation of the new mutation. For the mutation-selection portion of their model, they use the probability of fixation of a new mutation in a haploid population as that given by the classic Kimura [34] formulation.

$$P(Z_{ij}) = \frac{1 - \mathrm{e}^{-2s_j}}{1 - \mathrm{e}^{-2Ns_j}}.$$

While conceptually clear, this model does not consider any other forces beside population size in computing the substitution rate.

As suggested above, the substitution rate may differ depending on the types of changes made to a codon. A reasonable assumption could be that nonsynonymous substitutions are more likely to be selected for. Should one be interested in examining the ratio of nonsynonymous to synonymous changes, the differences in nonsynonymous transition, nonsynonymous transversion, synonymous transition, and synonymous transversion can be considered. The first models formulated to capture these dynamics were outlined by Goldman and Yang [35] and many models have built upon this foundation,

$$R_{ij} = \begin{cases} 0, & \text{if the two codons differ at more that one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ k\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega k\pi_j, & \text{for nonsynonymous transition,} \end{cases}$$

where k is the transition transversion ratio, ω is the ratio of nonsynonymous to synonymous changes, and π_j is the equilibrium frequency of codon j.

This initial framework given by Goldman and Yang has been expanded to incorporate a more complex view of the mutation-selection process taken from the realm of population genetics, which allows for the consideration of other parameters influential on selection. Important features of the physical environment such as the structure of the protein and the solvent accessible surface area of the position can affect the replacement process. These elements are incorporated into the Robinson et al. [16] model, where differences in a solvent accessibility score and pairwise interaction score are considered in the calculation of nonsynonymous changes. In order to compute these scores, the 3D structure of a protein must be known and is assumed to be identical across the set of sequences being analyzed; in this case, it is inferred using a threading-based approach. Methods like these are novel because they incorporate the phenotypic property of sequence–structure compatibility in accounting for the substitution process.

Another extension of codon models involving structure for the detection of positive selection was the invention of tertiary windowing [36–38]. In this approach, standard codon models were applied independently in a contact sphere delineated by protein structure to detect regions of a protein that were co-evolving under positive selection.

However, neither the Robinson model nor the tertiary windowing approach captures the whole picture of the substitution process, as both leave out population-level factors. In fact by relating the Robinson model back to the Halpern–Bruno model, a $2Ns_j$ term can be computed. The $2Ns_j$ term is often referred to as the scaled selection coefficient. The scaled selection coefficient relates the evolutionary importance of a trait in a way such that it can be compared to those of other traits which arose in populations of different sizes.

Amino acid substitution matrices are typically empirical, drawn from existing data rather than parameterized for use on any one data set. Koshi and Goldstein [39] built such matrices explicitly incorporating structural considerations. Assuming that sequences with the same structural properties have the same substitutional properties, structural context-specific substitution matrices were created. Properties considered included secondary structure and solvent accessibility. Interestingly, it was found that different residue position classes produce different optimal substitution matrices, suggesting the importance of the incorporation of structural data in models of the substitution process.

Koshi et al. [40] extended this approach, examining the influence of some of the physicochemical properties of the amino acids on the substitution process by introducing hydrophobicity and bulk as parameters in a suitability function into their model. Wong et al. [41] expanded further on this idea by examining the impact of volume, hydrophobicity, charge, and polarity independently. It is recognized that these forces are not independent and that detection of selection for one factor does not imply that the other factors are not influential. To implement this concept, they expand the Goldman and Yang [35] model and include a category for the change of physicochemical properties. While this model represents another step forward in the inclusion of biophysically relevant parameters and yields interesting results concerning physicochemical selective pressures, it ignores population-level effects.

Nielsen and Yang [42] proposed a codon model that would estimate the effective population size along with selection coefficients. They cast ω in a population genetic framework

$$w_{ij} = \frac{2N_j s_i}{1 - e^{-N_j s_i}},$$

and substitute this new value of w_{ij} into their familiar 1994 formulation. From their analysis of mitochondrial protein-coding genes, they found that allowing for the variation of N among lineages increases model fit.

Huzurbazar et al. [43] constructed a model which considers both population size and some basic elements of protein physiochemistry. Building on Kimura's framework, they explore the probability of fixation of classes of substitutions on effective population size. These classes of substitution are defined by partitions based on physicochemical data from the Grantham matrix [44]. The probability of fixation is given in a manner easily relatable to Kimura's model as

$$F_{ij} = \frac{(\mu_j S_{ij})/(1 - e^{-2N_i S_{ij}})}{\sum_j (\mu_j S_{ij})/(1 - e^{-2N_i S_{ij}})},$$

where μ_j is the relative mutational opportunity, j is the index of partitions, and i indexes the populations. Applying this model to a set of seven species with vastly different population sizes, it was found that selective coefficients decline as population size increases and decline with more radical amino acid substitutions. This unexpected result could be caused by a number of factors, such as the complexity of the mechanisms by which positive selection acts, linkage of substitutions, failure to control for the underlying distribution of protein folds and corresponding substitution patterns, compensatory processes at a systems level, failure to account for segregating variation differentially averaged with fixed changes, or other population-level forces. More than anything, this model suggests that numerous factors on a broad range of levels influence the substitution and selection process and should ultimately be considered explicitly.

Though this model includes aspects of physicochemical and population-level properties, they are both incorporated on a relativity basic level. Future substitution models will have to consider these factors in a more detailed manner. At a structural

level, the Grantham matrix can be replaced by a scoring function like that in Grahnen et al. [21]. This can also enable explicit testing of the relationship between changes in values of the scoring function and probability of fixation (fitness).

Another physically based process that has yet to be considered is that of linkage. It has been shown that proteins that are encoded for by genes which are located nearly one another do not evolve independently [45]. Aspects other than protein structure must be taken into consideration in order to account for this dependence. Nielsen and Yang [42] and Huzurbazar et al. [43] acknowledge the implications of recombination but avoid the complication of incorporating this complexity by making the assumption that their models are only valid in a particular population genetics context where the mutation rate is low enough to make concurrent polymorphisms improbable and where the recombination rate is high enough to insure that polymorphisms are in linkage equilibrium.

In order to make inferences about the selective process in populations in which these assumptions are not the case, linkage should be considered. Hill–Robertson effects are not unlikely as it has been found that the probability that a beneficial mutation will drag another allele into fixation with it is normally less than twice the selective advantage inferred by the beneficial mutation [46]. Nevo et al. [47] suggested that fixation effects account for increased standing variation in small population size organisms compared with expectations from site-independent models. Should a pair of mutations arise which both confer a similar fitness change in the absence of recombination, these new mutations could experience competition. This phenomenon has been observed in *Drosophila melanogaster* and *Drosophila simulans*, where it was found that regions of low recombination had reduced rates of evolution [48]. The classic formulation for linkage is dependent on both molecular and population-level processes, with the recombination rate based on the crossover process and the population size. While it has been suggested that the variation of recombination rates across organisms may have a limited effect on the substitution process [49], this conclusion may change as the population size varies. In fact, as suggested by Huzurbazar et al. [43], there appears to be a complex interplay between effective population size and other parameters in the substitution process, possibly even including the underlying protein fold distribution found in different species.

5 Concluding Thoughts

Advances in the development of protein models have generated the ability to incorporate multiple factors based on structure, evolution, and population genetics. The implementations of these models in fields as divergent as phylogenetics, comparative genomics, and evolutionary simulation have enabled addressing questions involving how both population-level dynamics and physicochemical/structural properties are influential in the evolutionary process. The development of more complex models which further consider these multidimensional intricacies in both forward and retrospective trajectories is vital in furthering our knowledge of

the process of molecular evolution. The three current research trajectories have developed more independently than would be desirable, as models from each subfield can be useful in other subfields. Further, incorporation of population genetic models into molecular evolution is more advanced than the incorporation of structural and functional considerations. However, it is with this growing multiscale modeling trajectory that a key understanding of the evolution of cell and molecular systems can be generated, both in the details and in the processes.

Acknowledgements D.A.L. is funded by NSF DBI-0743374 and NIH-INBRE award P20 RR016474. A.I.T. is supported by the aforementioned NSF award, while J.A.G. is supported by the aforementioned NIH-INBRE award.

References

1. Gillespie, J.H.: Some properties of finite populations experiencing strong selection and weak mutation. Am. Nat. **121**, 691–708 (1983)
2. Bornberg-Bauer, E., Chan, H.S.: Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. PNAS. **96**, 10689–10694 (1999)
3. Melin, R., Li, H., Wingreen, N.S., Tang, C.: Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study. J. Chem. Phys. **110**, 1252–1262 (1999)
4. Sali, A., Shakhnovich, E., Karplus, M.: How does a protein fold? Nature. **369**, 248–251 (1994)
5. Miyazawa, S., Jernigan, R.: Estimation of effective interresidue contact energies from protein crystal structures- Quasi-chemical approximation. Macromol. **18**, 534–552 (1985)
6. Levinthal, C.: Are there pathways for protein folding? J. Chim. Phys. **65**, 44–45 (1968)
7. Li, L., Mirny, L.A., Shakhnovich, E.I.: Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. Nat. Struct. Mol. Biol. **7**, 336–342 (2000)
8. Williams, P., Pollock, D., Goldstein, R.: Evolution of functionality in lattice proteins. J. Mol. Graph. Model. **19**, 150–156 (2001)
9. Bloom, J., Wilke, C., Arnold, F.A.C.: Stability and the evolvability of function in a model protein. Biophys. J. **86**, 2758–2764 (2004)
10. Rastogi, S., Liberles, D.: Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol. Biol. **5**, 28 (2005)
11. He, X., Zhang, J.: Rapid Subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics. **169**, 1157–1164 (2005)
12. Massey, S.E., Churbanov, A., Rastogi, S., Liberles, D.A.: Characterizing positive and negative selection and their phylogenetic effects. Gene. **418**, 22–26 (2008)
13. Castoe, T.A., de Koning, A.P.J., Kim, H.M., Gu, W., Noonan, B.P., Naylor, G., Jiang, Z.J., Parkinson, C.L., Pollock, D.D.: Evidence for an ancient adaptive episode of convergent molecular evolution. PNAS. **106**, 8986–8991 (2009)
14. Taverna, D., Goldstein, R.: Why are proteins marginally stable? Proteins. **46**, 105–109 (2002)
15. Bastolla, U., Farwer, J., Knapp, E.W., Vendruscolo, M.: How to guarantee optimal stability for most representative structures in the protein data bank. Protein. Struct. Funct. Bioinf. **44**, 79–96 (2001)
16. Robinson, D.M., Jones, D.T., Kishino, H., Goldman, N., Thorne, J.L.: Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. **20**, 1692–1704 (2003)
17. Parisi, G., Echave, J.: Structural constraints and emergence of sequence patterns in protein evolution. Mol. Biol. Evol. **18**, 750–756 (2001)
18. Bastolla, U., Porto, M., Eduardo Roman, M.H., Vendruscolo, M.H.: Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. J. Mol. Evol. **56**, 243–254 (2003)

19. Rastogi, S., Reuter, N., Liberles, D.A.: Evaluation of models for the evolution of protein sequences and functions under structural constraint. Biophys. Chem. **124**, 134–144 (2006)
20. Mukherjee, A., Bagchi, B.: Correlation between rate of folding, energy landscape, and topology in the folding of a model protein HP-36. J. Chem. Phys. **118**, 4733–4747 (2003)
21. Grahnen, J.A., Nandakumar, P., Kubelka, J., Liberles, D.A.: Evaluating protein threading and protein–protein interaction on proteomic and evolutionary scales. BMC Evol. Biol. **11**, 361 (2011)
22. Liberles, D.A., Tisdell, M.D.M., Grahnen, J.A.: Binding constraints on the evolution of enzymes and signaling proteins: The important role of negative pleiotropy. Philos. Trans. R. Soc. London B. **278**, 1930–1935 (2011)
23. Baker, C., MacKerell, A.: Polarizability rescaling and atom-based Thole scaling in the CHARMM Drude polarizable force field for ethers. J. Mol. Model. **16**, 567–576 (2010)
24. Christen, M., Hünenberger, P.H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D.P., Heinz, T.N., Kastenholz, M.A., Kräutler, V., Oostenbrink, C., Peter, C., Trzesniak, D., van Gunsteren, W.F.: The GROMOS software for biomolecular simulation: GROMOS05. J. Comput. Chem. **26**(16), 1719–1751 (2005)
25. Ponder, J.W., Wu, C., Ren, P., Pande, V.S., Chodera, J.D., Schnieders, M.J., Haque, I., Mobley, D.L., Lambrecht, D.S., DiStasio, R.A., Head-Gordon, M., Clark, G.N.I., Johnson, M.E., Head-Gordon, T.: Current status of the AMOEBA polarizable force field. J. Phys. Chem. B. **114**, 2549–2564 (2010)
26. Maguid, S., Fernandez-Alberti, S., Echave, J.: Evolutionary conservation of protein vibrational dynamics. Gene. **422**, 7–13 (2008)
27. Parisi, G., Echave, J.: The structurally constrained protein evolution model accounts for sequence patterns of the LbetaH superfamily. BMC Evol. Biol. **4**, 41 (2004)
28. Lartillot, N., Philippe, H.: A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. **21**, 1095–1109 (2004)
29. Lartillot, N., Philippe, H.: Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos. Trans. R. Soc. London B. **363**, 1463–1472 (2008)
30. Rodrigue, N., Philippe, H., Lartillot, N.: Assessing site-interdependent phylogenetic models of sequence evolution. Mol. Biol. Evol. **23**, 1762–1775 (2006)
31. Nasrallah, C.A., Mathews, D.H., Huelsenbeck, J.P.: Quantifying the impact of dependent evolution among sites in phylogenetic inference. Syst. Biol. **60**, 60–73 (2011)
32. Kleinman, C.L., Rodrigue, N., Lartillot, N., Philippe, H.: Statistical potentials for improved structurally constrained evolutionary models. Mol. Biol. Evol. **27**, 1546–1560 (2010)
33. Halpern, A.L., Bruno, W.J.: Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. **15**, 910–917 (1998)
34. Kimura, M.: On the probability of fixation of mutant genes in a population. Genetics. **47**, 713–719 (1962)
35. Goldman, N., Yang, Z.: A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**(5), 725–736 (1994)
36. Berglund, A.C., Wallner, B., Elofsson, A., Liberles, D.A.: Tertiary windowing to detect positive diversifying selection. J. Mol. Evol. **60**, 499–504 (2005)
37. Liang, H., Zhou, W., Landweber, L.F.: SWAKK: a web server for detecting positive selection in proteins using a sliding window substitution rate analysis. Nucleic. Acids. Res. **34**, W382–W384 (2006)
38. Suzuki, Y.: Three-dimensional window analysis for detecting positive selection at structural regions of proteins. Mol. Biol. Evol. **21**, 2352–2359 (2004)
39. Koshi, J.M., Goldstein, R.A.: Context-dependent optimal substitution matrices. Protein. Eng. **8**, 641–645 (1995)
40. Koshi, J.M., Mindell, D.P., Goldstein, R.A.: Beyond mutation matrices: physical-chemistry based evolutionary models. Genome Informatics. Refereed conference proceeding (1998)
41. Wong, W., Sainudiin, R., Nielsen, R.: Identification of physicochemical selective pressure on protein encoding nucleotide sequences. BMC Bioinformatics. **7**, 148 (2006)

42. Nielsen, R., Yang, Z.: Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol. Biol. Evol. **20**, 1231–1239 (2003)
43. Huzurbazar, S., Kolesov, G., Massey, S.E., Harris, K.C., Churbanov, A., Liberles, D.A.: Lineage-specific differences in the amino acid substitution process. J. Mol. Biol. **396**, 1410–1421 (2010)
44. Grantham, R.: Amino acid difference formula to help explain protein evolution. Science. **185**, 862–864 (1974)
45. Birky, C.W., Walsh, J.B.: Effects of linkage on rates of molecular evolution. PNAS. **85**, 6414–6418 (1988)
46. Lynch, M.: (2007) The Origins of Genome Architecture. Sinauer Associates, Sunderland, MA
47. Nevo, E., Kirzhner, V., Beiles, A., Korol, A.: Selection versus random drift: long–term polymorphism persistence in small populations (evidence and modelling). Philos. Trans. R. Soc. London Biol. **352**, 381–389 (1997)
48. Betancourt, A.J., Presgraves, D.C.: Linkage limits the power of natural selection in Drosophila. PNAS. **99**, 13616–13620 (2002)
49. Pal, C., Papp, B., Lercher, M.J.: An integrated view of protein evolution. Nat. Rev. Genet. **7**, 337–348 (2006)

Index

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7, © Springer Science+Business Media, LLC 2012