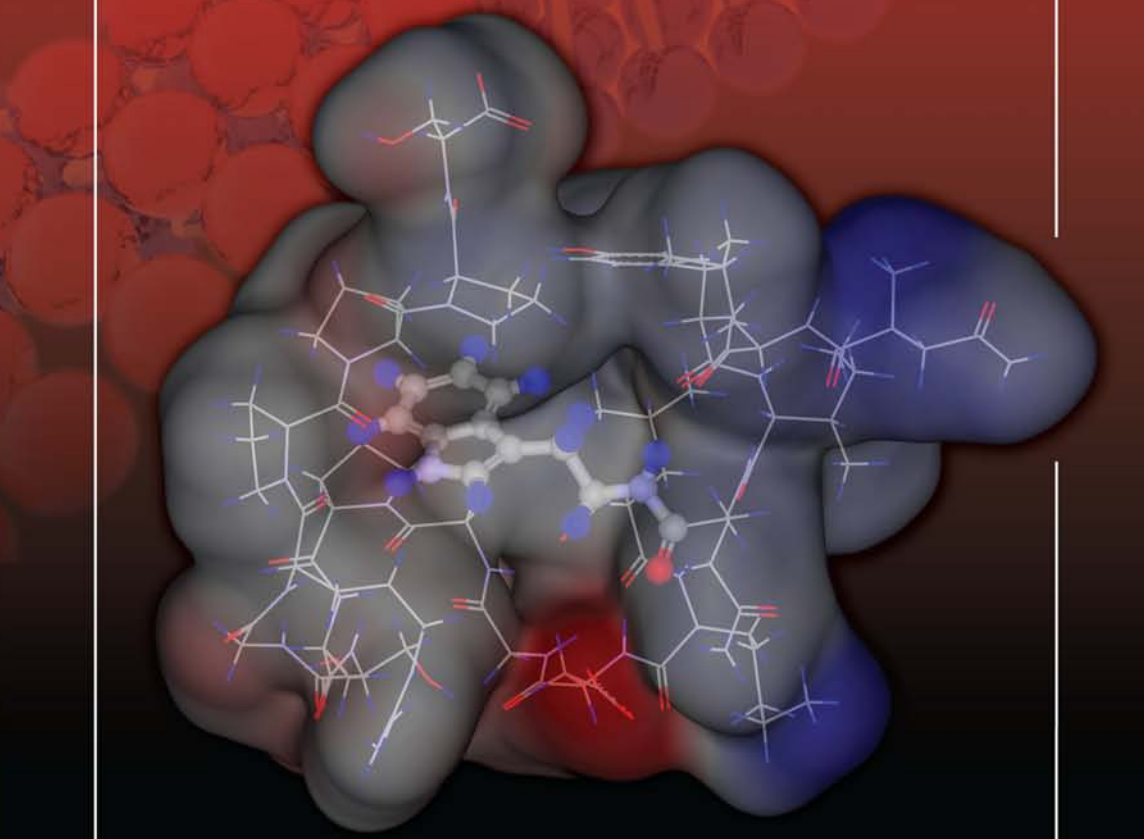
The FRAGMENT MOLECULAR ORBITAL METHOD

PRACTICAL APPLICATIONS TO LARGE MOLECULAR SYSTEMS



CRC Press
Taylor & Francis Group

EDITED BY
Dmitri G. Fedorov
Kazuo Kitaura

The FRAGMENT MOLECULAR ORBITAL METHOD

PRACTICAL APPLICATIONS TO LARGE MOLECULAR SYSTEMS

The FRAGMENT MOLECULAR ORBITAL METHOD

PRACTICAL APPLICATIONS TO LARGE MOLECULAR SYSTEMS

EDITED BY
Dmitri Fedorov
Kazuo Kitaura

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4200-7848-0 (Hardcover)

Library of Congress Cataloging-in-Publication Data

The fragment molecular orbital method : practical applications to large molecular systems / editors, Dmitri Fedorov, Kazuo Kitaura.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4200-7848-0 (hardcover : alk. paper)
1. Molecular orbitals. 2. Macromolecules. 3. Biomolecules. I. Fedorov, Dmitri. II. Kitaura, Kazuo. III. Title.

QD461.F795 2009
541'.22--dc22 2009006361

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Editors

Dmitri G. Fedorov received his M.S. in quantum chemistry from St. Petersburg State University in Russia in 1993. He was awarded a Ph.D. in physical chemistry at Iowa State University in 1999, working under the guidance of Mark Gordon. He spent 2 years at the University of Tokyo as a Japan Society for the Promotion of Science (JSPS) Postdoctoral Research Fellow, where he worked with Kimihiko Hirao. He then moved to the Research Institute for Computational Sciences (RICS) at the National Institute of Advanced Industrial Science and Technology (AIST) in Japan in 2002, where he is currently employed as senior research scientist. Dr. Fedorov's research interests include the relativistic effects in chemistry as well as the quantum-mechanical method development for describing large systems, such as proteins.

Kazuo Kitaura received his Ph.D. in quantum chemistry in 1976 from Osaka City University. After working in Japan as a research associate at the Institute for Molecular Science and Osaka City University, he became associate professor in 1989 at the Institute for Molecular Science. In 1993, Dr. Kitaura was appointed professor at Osaka Prefecture University. In 2001, he moved to the Research Institute for Computational Sciences at AIST, where he served as group leader and director, and is presently the principal research scientist. Since 2006, he has also been a professor at the School of Pharmaceutical Sciences, Kyoto University. Dr. Kitaura's research interests include intermolecular interactions, structure and reactivity of transition metal complexes, and electronic structure theory for large molecules and molecular clusters.

Contributors

Shinji Amari
Computational Science Department
Science and Technology Systems Division
Ryoka Systems, Inc.
Japan

Mahito Chiba
Research Institute for Computational Sciences (RICS)
National Institute of Advanced Industrial Science and Technology (AIST)
Japan

Kaori Fukuzawa
Mizuho Information and Research Institute, Inc.
Japan

Tomohiro Hashimoto
Faculty of Regional Studies
Gifu University
Japan

Hideharu Ishida
Department of Applied Bioorganic Chemistry
Gifu University
Japan

Toyokazu Ishida
Research Institute for Computational Sciences (RICS)
National Institute of Advanced Industrial Science and Technology (AIST)
Japan

Takeshi Ishikawa
Division of Prion Research
Center for Emerging Infectious Disease
Gifu University
Japan

Akifumi Kato
Mizuho Information and Research Institute, Inc.
Japan

Makoto Kiso
Department of Applied Bioorganic Chemistry
Gifu University
Japan

Yuto Komeiji
Research Institute for Computational Sciences (RICS)
National Institute of Advanced Industrial Science and Technology (AIST)
Japan

Ikuo Kurisaki
Graduate School of Science and Technology
Kobe University
Japan

Yuji Mochizuki
Department of Chemistry, and Research Center for Smart Molecules, Faculty of Science
Rikkyo University
Japan

Isao Nakanishi
Department of Pharmaceutical Sciences
Kinki University
Japan

Hirofumi Nakano
Department of Chemistry
Aichi University of Education
Japan

Tatsuya Nakano
Division of Medicinal Safety Science
National Institute of Health Sciences
Institute of Industrial Science
University of Tokyo
Japan

Motohiro Nishio
CHPI Institute
Japan

Kosuke Okazaki
Central Research Laboratory
Kissei Pharmaceutical
Japan

Tomonaga Ozawa
Central Research Laboratory
Kissei Pharmaceutical
Japan

Toshihiko Sawada
College of Life and Health Sciences
Chubu University
Japan

Tohru Suzuki
United Graduate School of Agricultural Science
Gifu University
Japan

Yasuo Suzuki
College of Life and Health Sciences
Chubu University
Japan

Naoki Taguchi
Department of Chemistry, Faculty of Science
Rikkyo University
Japan

Shigenori Tanaka
Graduate School of Human Development and Environment
Kobe University
Japan

Hiroaki Tokiwa
Department of Chemistry, Faculty of Science
Rikkyo University
Japan

1 Introduction

Kazuo Kitaura and Dmitri G. Fedorov

CONTENTS

Due to the continuous increase in computer power as well as the tedious efforts invested in method development, electronic structure calculations are now becoming possible for larger and larger systems, expanding the application territory to systems that were previously exclusively treated by the classical force field methods. The *ab initio* molecular orbital calculations of a hundred atoms are routinely performed at the level of Hartree–Fock (HF) theory and density functional theory (DFT) with moderate basis sets. Although large-scale DFT calculations of several thousands of atoms have been reported,[1] it seems difficult to go beyond that size, because the computational time and resources scale steeply with the system size N; the scaling is $O(N^3)$ for HF and DFT, and it is even higher for correlated wavefunction theories.

The need to treat larger systems has prompted the development of linear scaling or order N methods,[2] whose computational cost increases linearly with system size, and fragment-based approaches, which have a long history in quantum chemistry, and account for the majority of them. Recently, various new methods have been actively introduced, which are reviewed in References [1] and [3] and in Chapter 2. Most of these methods rely on the transferability of the electron density distributions or properties of groups of atoms in a molecule: the densities or properties are transferred from those in small reference molecules to the corresponding groups in a large target molecule. Among these methods, the fragment molecular orbital (FMO)[4] method has a distinctive feature — it is derived from the energy decomposition analysis (EDA)[5] applied to many-body molecular interactions.[6] A number of performance tests have been conducted to establish its accuracy relative to *ab initio* properties and the computational efficiency on parallel computers. We believe that the FMO method has reached the production stage providing means to study electronic structures of very large and complex molecules and molecular clusters.

For computational studies, it is very important to employ a proper model describing real systems (e.g., considering solvent and temperature). The all-electron *ab initio* calculations have their practical limitations, and a combined use of other methods is indispensable, for instance, solvation effects can be taken into account with the polarizable continuum model (PCM).[7] There are other significant terms in the model such as the free energy change of the solute, which can be supplemented from molecular dynamic (MD) simulations to obtain the solvated free energy values.

Thus, most large molecules are subjected to multiphysics treatment, and a proper computational method should be utilized for each aspect of the problem at hand. The reader will find some attempts along these lines presented in several chapters.

This book is intended for those eager to obtain useful information from electronic structure calculations of large systems, and also those who wish to know what can be elucidated with present calculations and those in the near future. The emphasis of this book is on *practical* aspects described for a general scientific reader, including the foundations of the method with as little mathematical detail as possible. Much practical advice and real examples are provided for performing applications, beginning with modeling, followed by calculations, and the resulting visualizations. Some chapters dealing with the development of new capabilities of FMO methods contain sufficient mathematical details that will be of interest to method developers. Moreover, to reduce tedious work during the input data preparation and in the visualization of calculated results, free modeling software, Facio,[8] in which FMO-related functions are implemented, is provided on the included CD-ROM along with many examples and usage hints.

In order to encourage readers to perform FMO calculations, some features of the freely available FMO programs (GAMESS[9] and ABINIT-MP[10]) are described in Chapters 2 and 3, respectively. It can be noted that due to the independent method development in several FMO groups, some differences in the FMO terminology exist. In this book, an attempt has been made to list several variants when applicable, while allowing the authors to use their preferred terms, although we would like to see more uniform FMO terminology in the future.

In Chapters 4 and 5, after a detailed introduction of the FMO approaches for excited states with configuration interaction and time-dependent DFT, respectively, several calculations of photoactive proteins and small molecules in solution are introduced. In Chapter 6, the basic scheme of molecular dynamics simulation using the FMO energy and force (FMO-MD) is described, including its applications to several chemical reactions in solution.

Actual examples of FMO applications will provide readers with an idea of how to use FMO methods and what information is obtained from the calculations. Since the FMO was applied to a real-size protein at the early stage of its development, it continues to be extensively used to study protein–ligand interactions, some examples of which are presented in Chapters 7, 8, 9, and 10, to guide future applications of a similar kind. In these chapters, the importance of structure modeling is emphasized to obtain a meaningful interaction energy between a protein and a ligand and to shed light on the molecular recognition mechanism of the protein, which is essential for understanding the function in living bodies.

Chapter 11 is concerned with enzymatic reactions, where the free energy change along a reaction coordinate is simulated with a hybrid quantum mechanics and molecular mechanics (QM/MM) method, and the FMO method is used to identify the determining factor of the protein environment in stabilizing the transition state. The approach described in this chapter is considered to be a representative example of multiphysics treatment.

Our main purpose has been to continuously develop our method so as to make it available to general chemists, physicists, biologists, and other scientists, who more

and more routinely use quantum-mechanical computations as a standard tool, devoid of its former aura of something esoteric available only to the initiated few. To achieve this goal, we put great effort into making available readily usable computational, modeling, and visualization software, and it will be our pleasure to see future applications of real-life problems made easier through the use of the methods presented in this book.

REFERENCES

1. For a brief review of methods developed for large systems, see Fedorov, D.G., Kitaura, K. 2007. Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. *J. Phys. Chem. A* 111:6904–6914.
2. Sato, F., Yoshihiro, T., Era, M., Kashiwagi, H. 2001. Calculation of all-electron wavefunction of hemoprotein cytochrome *c* by density functional theory. *Chem. Phys. Lett.* 341:645–651.
3. Goedecker, S. 1999. Linear scaling electronic structure methods. *Rev. Mod. Phys.* 71:1085–1123.
4. Kitaura, K., Ikeo, E., Asada, T., Nakano, T., Uebayasi, M. 1999. Fragment molecular orbital method: an approximate computational method for large molecules. *Chem. Phys. Lett.* 313:701–706.
5. Kitaura, K., Morokuma, K. 1976. A new energy decomposition scheme for molecular interactions within the Hartree–Fock approximations. *Int. J. Quant. Chem.* 10:325–340.
6. Fedorov, D.G., Kitaura, K. 2004. The importance of three-body terms in the fragment molecular orbital method. *J. Chem. Phys.* 120:6832–6840.
7. Tomasi, J., Mennucci, B., Cammi, R. 2005. Quantum mechanical continuum solvation models. *Chem. Rev.* (Washington, DC) 105:2999–3093.
8. Suenaga, M. 2008. Development of GUI for GAMESS/FMO calculation. *J. Comput. Chem. Jpn.* 7:33–54 (in Japanese).
9. Schmidt, M.W., Baldridge, K.K., Boatz, J.A. et al. 1993. General atomic and molecular structure system. *J. Comput. Chem.* 14:1347–1363.
10. Nakano, T., Mochizuki, Y., Fukuzawa, K., Amari, S., Tanaka, S. 2006. Developments and applications of ABINIT-MP software based on the fragment molecular orbital method. In *Modern methods for theoretical physical chemistry of biopolymers*, eds. E.B. Starikov, J.P. Lewis, and S. Tanaka, 39–52. Amsterdam: Elsevier.

2 Theoretical Background of the Fragment Molecular Orbital (FMO) Method and Its Implementation in GAMESS

Dmitri G. Fedorov and Kazuo Kitaura

CONTENTS

2.1 INTRODUCTION

During recent years, the constant increase in computational power has promoted a very considerable improvement in the quality of computations and in the size of the tractable systems. We consider the advent of multicore central processing units (CPUs) revolutionary for computational sciences, because parallel computing, which

used to be available to a limited number of research groups, has now become ubiquitous, and by purchasing a single computer node with a dual quad-core CPU, any scientist has access to an equivalent of 8 CPUs, one order of magnitude increase from just a few years ago.

Although one can argue that computers that are too fast provide a drive to "crunch numbers" without any time left to analyze them and understand the physics of the system, especially for biological applications there is clearly a need for more powerful computers, because the number of atoms is very large, and the traditional quantum-mechanical (QM) methods scale too steeply with system size. Fortunately, the computational cost to run biochemical calculations with quantum mechanics is still large enough to leave ample time to think about physics and chemistry while waiting for results.

To make large calculations feasible, a number of methods have been suggested, extensively reviewed by us recently.[1] Many of these methods involve fragmentation, and from computations of fragments one can construct the total properties. The fragmentation methods can be divided into three main groups[2]: divide-and-conquer, transferable methods, and *e pluribus unum* approaches. In the latter case, the whole system is included in the individual fragment calculations, usually in the form of the Coulomb field, and total properties are obtained by adding values for fragment and, possibly, many-body corrections. The fragment molecular orbital (FMO) method[3] is an example of the latter group of methods, and it has its roots in energy decomposition analysis (EDA).[4]

The fragment-based methods have become a separate and widely developed area of computational sciences, and now we briefly highlight the important method development not covered in our last review in 2007. The elongation method has been applied to polymers,[5] DNA,[6] and surfaces.[7] A number of other methods have been suggested or improved recently.[8–14]

One can ask: What is the meaning and importance of many-body corrections (computed from fragment pairs and triples)? In a simple fragment approach (there were very many of these suggested in the past), one can ignore almost all interfragment interaction; for instance, to study the ligand–protein binding, one can cap a residue with hydrogens, and do QM calculations of just one such capped residue and a ligand. It is a fast calculation to perform, and it ignores all many-body effects (except for the two-body interaction of just one residue and the ligand). Most importantly, it neglects the polarization of both the residue and the ligand by the remaining system, which can considerably alter the density distribution and properties. The polarization of proteins is one example of the many-body interactions between residues. Another is that there is charge transfer and other QM effects in the protein as well as in the protein–ligand interaction, which cannot be described by considering just one capped residue. Thus, the best approach is to take the whole system and consider all important interactions in it, which constitute the many-body effects. It has been a distinct trend of modern fragment-based methods to include as many many-body interactions as possible.

The progress of fully *ab initio* calculations of large systems has made possible single-point calculations of systems with several hundreds of atoms fairly routinely,

if uncorrelated wavefunctions and moderate basis sets are used. However, the steep scaling of the required memory and calculations as well as parallelization issues pose considerable difficulties to the QM calculations of systems with thousands of atoms. In addition, numeric and convergence problems for the systems with many charge centers such as proteins are severe. On the other hand, fragment-based approaches allude to the chemical idea of parts of the system retaining their identity to a large extent (e.g., functional groups and residues). Thus, fragmentation not only reduces the costs, but it also provides a wealth of information on the properties of fragments and their interactions. In addition, fragment-based methods provide some control over the electronic state (charge localization) of the whole system to which the calculations converge, which for proteins with many charged residues may be an issue to consider.

An important point is accuracy control — that is, whether a fragment-based method can provide means to systematically increase the accuracy. In general, many modern fragment-based methods have some ways to increase the accuracy, mostly by three means:

1. Increasing the fragment size (thus treating larger parts fully *ab initio*).
2. Increasing the threshold to consider more fragment pairs.
3. Adding higher many-body corrections (such as trimers in FMO).

The effect of these changes has to be carefully studied to form some idea as to the numeric effect of their applications. Many such studies were done for FMO.[1]

Several FMO-like methods, different from FMO in minor details, corresponding to a particular way of computing the electrostatic field by the point charge approximation have been developed and used.[15–17] In addition, the kernel energy method employs the FMO energy expansion for hydrogen-capped fragments (kernels) in vacuum (without the electrostatic field of the whole system).[18–25]

Since our review[1] in 2007, the FMO method has been further developed theoretically[26–35] and applied to a number of systems: solid state,[35,36] DNA–protein interaction,[37] structure determination of polypeptides,[38] and protein–ligand binding.[39–44] New, exciting possibilities of extending the usefulness of FMO are developed: visualization of the interactions,[45] receptor-specific scoring function,[46] molecular dynamics,[47–49] and quantitative structure-affinity relationship (QSAR) studies.[50,51] FMO can be used with most common wavefunction types and large systems are tractable: RHF/6-31G* and MP2/6-31G calculations were reported for 20,581[52] and 14,086[33] atoms, respectively.

FMO has been implemented in two freely distributed programs, GAMESS[53,54] and ABINIT-MP,[55] for which graphical user interfaces are available, so that one can readily download the software and apply FMO to a system of interest. In addition, FMO is implemented in a local version of NWChem,[56] a modified version of FMO in GAMESS was created for grid computing (GridFMO[57]), and a new FMO program was developed, known as OpenFMO.[58]

In this chapter, we describe the basics of FMO and its fragmentation scheme and theoretical background, from the practical point of view of answering possible questions and doubts regarding the choice of the level of theory to set up actual

calculations. Much advice in this chapter is general and applies to any FMO implementation, and to exemplify it, in a few places some emphasis on the usage of FMO in GAMESS is made.

2.2 METHODOLOGY

2.2.1 Fragmentation

It is helpful to look at small examples and understand the fragmentation mechanism before proceeding to the mathematical formulation of FMO. Molecular clusters provide the simplest case: one can take each individual molecule as a fragment, and that is usually done when computing water clusters or explicitly solvated molecules (i.e., when solvent molecules are calculated quantum mechanically in FMO similar to the solute). However, water molecules are rather small, and as found by comparison to *ab initio* methods, the FMO error is comparatively large.

One can consider using one such molecule per fragment division for the structure determination (geometry optimization or molecular dynamics), improving the final energetics with two molecules assigned to one fragment. When doing so, it is best to put close molecules together — namely, those that are strongly bound to each other (frequently, by hydrogen bonds). Some examples of grouping solvent molecules can be found among the samples of FMO input files.[59] Alternatively, one can consider using the three-body FMO expansion (*vide infra*) with one molecule per fragment.

Next, we proceed to fragmenting a small polypeptide, capped glycine trimer. In the mechanistic treatment of polypeptides (e.g., in force field methods), it is common to divide them semantically into amino acid residues to define their properties such as interaction. However, such division requires detaching peptide bonds, which have a considerable electron delocalization from the lone pairs on nitrogen and oxygen. Quantum-mechanical methods operate with electron density and introducing a division of the density at this point is rather difficult. In FMO, one typically divides polypeptides at Cα carbons adjacent to peptide bonds, and thus, peptide fragments in FMO (fragment residues) are shifted relative to conventional residues by one CO group.

We frequently distinguish fragment residues by a dash in their symbol (e.g., Ala-10 for fragment residue versus conventional Ala10) or with a mesh (#, as in Ala#10). One should also note that some properties in FMO can be computed for conventional residues, even though fragment residues are used. In FMO, at the end one obtains the properties of the whole system (e.g., energy gradient). Thus, one can define some values for conventional residues from these total properties, for instance, peptide dihedral angles from FMO-optimized geometries[60] or the charges on residues from atomic charges. Other properties, such as dipole moments, polarization, and pair interactions are defined specifically for residue fragments and cannot be easily recomputed for conventional fragments. To compare these properties with force fields, the only way seems to be to define residue fragments in the force field calculations and compute the necessary values for them.

To fragment a molecule in FMO, one should specify pairs of atoms that define the bonds to be detached. In Figure 2.1a, three bonds are detached, for the pairs

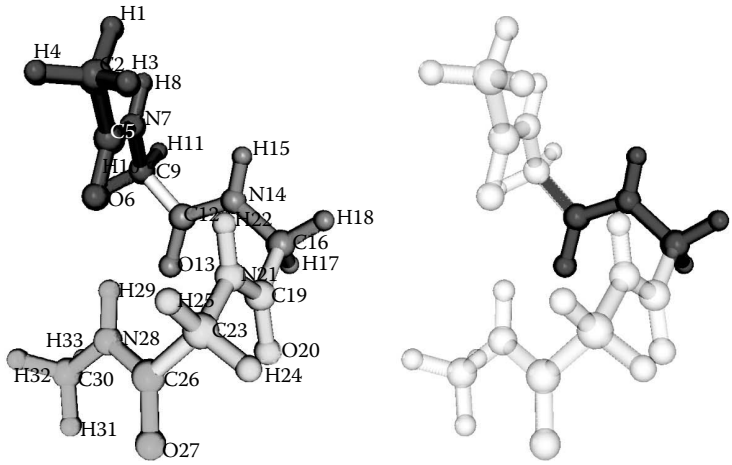


FIGURE 2.1 Fragmentation of capped $(GLY)_3$: (a) the whole system and the fragmentation points and (b) fragment 2. **(See color insert following page 117.)**

of atoms (C9,C12), (C16,C19), and (C23,C26), defining four fragments. It is important to remember that the whole detached bond is assigned to a fragment, and thus, the fragmentation is performed at an atom, not between atoms. For the example in Figure 2.1a, the fragmentation is done at the first atom in the pair (C9, C16, and C23), and these atoms are called bond detached atoms (BDAs). The detached bonds are assigned to the atoms on the other side (C12, C19, and C26), and these atoms are called bond attached atoms (BAAs). The peptide bonds C12–N14, C19–N21, and C26–N28 are left intact within the corresponding fragments. Polypeptides are usually divided into one or two residues per fragment in FMO. (S–S links between cysteins are typically not detached.)

In Figure 2.1b, fragment 2 is shown. It includes the bond between atoms C9 and C12, and excludes the bond between atoms C16 and C19. To describe the C9–C12 bond in fragment 2, one needs to have basis functions on atom C9. Thus, a quasi-atom C9 is added to fragment 2 (see also Figure 2.2), which has carbon atom basis functions. Its atomic charge is set to 1, and C9 in fragment 1 has the charge of 5 (and carbon basis functions). This is a purely formal[2] division of protons, which has no effect upon the total properties, because the fragment calculations are performed in the electrostatic field of all atoms (those inside the fragment, as well as outside it). The main purpose of the proton division is not to affect the total properties (which are invariant) but to define physically meaningful interfragment pair interaction energies, which involve the internal energies, see Section 2.3.

The adopted fragmentation scheme leads to the preservation of charges and multiplicities (because one electron and one proton are reassigned for each detached bond, corresponding to a heterolytic division). The neutral and charged closed-shell (i.e., with an even number of electrons) residues retain their charge and the closed-shell

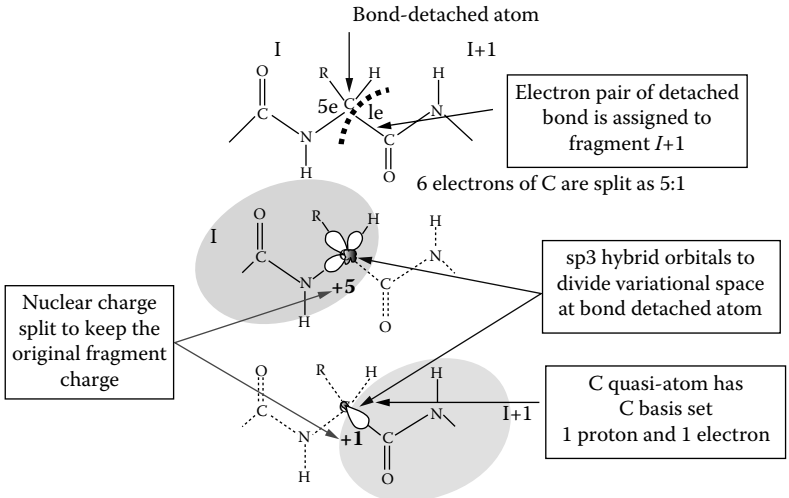


FIGURE 2.2 Details of bond detachment in the FMO method, showing the use of the hybrid orbitals and the formal proton reassignment at the bond detached atom, located on the border of fragments I and $I + 1$. **(See color insert.)**

property when they are made into fragments in FMO. In Figure 2.1, all fragments are neutral and have singlet multiplicity. In the present formulation of FMO, one can have any number of charged residues, but at most one fragment can be nonsinglet. This is accomplished with some excited state methods, multiconfiguration self-consistent field (MCSCF), time-dependent density functional theory (TDDFT), or configuration interaction (CI), as described Section 2.2.6.

Practical advice on fragmentation was given,[2] all of which was based on two simple rules: (1) detach single bonds without electron delocalization and (2) avoid detaching close bonds. So, in general, one should not attempt to divide aromatic rings or peptide bonds. It is especially easy to do fragmentation of standard systems, which are polymers of peptides, nucleotides, and saccharides (and any combination, such as DNA–protein complexes). In this case, there is no need to do manual fragmentation, and some software such as Facio[61] can do it completely automatically. For more sophisticated systems, for example, protein–ligand complexes (if the ligand is not a standard system listed above), first one detaches the peptide bonds automatically, and then, if necessary, adds some manual fragmentation points (specifies BDA and BAA pairs) for ligands using a graphic user interface (GUI), such as Facio.

When deciding upon the number of fragments, one should weigh the desired accuracy and the available computational resources. An additional factor is given by the pair interaction analysis and other fragment properties, which are frequently more convenient with a larger number of fragments, leading to a reduction in the computational time but a decrease in the accuracy. Having larger fragments results in the full quantum-mechanical description of larger areas, which is more accurate but also more expensive. Also, some ways to do an interaction analysis exist for subunits of

fragments without a need to detach bonds in FMO (i.e., by configuration analysis for fragment interactions [CAFI],[62] which can determine properties for molecular orbitals localized on a fragment, and the interaction between them).

The properties in FMO thus depend upon the particular fragmentation scheme, although usually very weakly because of the pair corrections (see Equation 2.3). Nevertheless, chemical knowledge should assist in performing fragmentation, which is not to be done mechanistically with a disregard for the physical properties of the system. To give an example, strong electron donors or acceptors should not be put into stand-alone fragments, especially if the total properties such as the energy are of importance. Metal cations are a simple example, and they are usually bound to some electron donor group. If one defines a metal cation as a separate fragment, then the electron density distribution at the monomer level will not be physically reasonable, because the charge is fixed as this level, and only pair corrections introduce charge transfer, which is very large in this case, and pair corrections are not sufficiently accurate to account for it.

Another somewhat problematic case is given by salt bridges (e.g., between close pairs of residues of opposite charge). Typically, when polypeptides are solvated, charged residues in salt bridges are separated by some solvent molecules, thus eliminating charge transfer between the two residues. However, if the structure is not properly refined, for instance, if it is optimized in gas phase, then a very short distance between charged residues can promote a large charge transfer between them. If the total properties are of interest (i.e., not just the pair interaction analysis), one can consider merging these two residues into one fragment. In FMO, there is no formal restriction in terms of the number of detached bonds per fragment, and one can have several detached bonds between a fragment and the rest of the system. FMO has well-established means to improve the accuracy of calculations (by increasing the fragment size and the degree of the many-body expansion, see Section 2.4).

One way to check the appropriateness of the fragmentation is to look at the interfragment charge transfer values:

$$\Delta\bar{Q} = \sum_{I}^{N} |\Delta Q_I| = \sum_{I}^{N} \left| \sum_{I \neq J}^{N} \Delta Q_{IJ} \right| \tag{2.1}$$

$$\Delta Q = \sum_{I>J}^{N} |\Delta Q_{IJ}| \tag{2.2}$$

where ΔQ_{IJ} is the charge transfer between fragments I and J. (There is some ambiguity in the definition of ΔQ_{IJ} for the connected dimers.)[63] ΔQ_I is the charged transfer from fragment I to all other fragments. N is the total number of fragments.

The two definitions ΔQ and $\Delta\bar{Q}$ are similar. The difference between them is easy to understand on the example of the cyclic water trimer.[63] $\Delta\bar{Q}$ is nearly zero, because charged transfer to each water molecule ΔQ_I is almost zero: the charge coming in and out cancels out. ΔQ, on the other hand, is considerable: three times the value for a hydrogen bond (about 0.03 a.u. each, 6-31G*). Although $\Delta\bar{Q}$ for

water predicts the observed good FMO accuracy in this case (which correlates with the interfragment charge transfer), in general ΔQ may be more reliable, as charge transfer occurs pairwise (ΔQ_{IJ}), and averaging it in $\Delta \bar{Q}$ may hide the actual charge redistribution and the associated accuracy loss of the FMO description.

Typically, hydrogen bonds contribute about 0.03 a.u. to ΔQ. Salt bridges have larger values, especially if the distance is too close (often from an inappropriate structure). The values of ΔQ_{IJ} for connected dimers are also usually about 0.03 to 0.06 a.u. Charged residues have a tendency for a larger charge transfer. Because the polarizable continuum model (PCM) effectively screens the strong Coulomb interaction, one can expect that the addition of solvent would decrease the charge transfer and thus increase the accuracy of FMO.

To give a practical example, for the Trp-cage protein (PDB ID: 1L2Y, 304 atoms) the 6-31G* values of ΔQ and $\Delta \bar{Q}$ are 1.20 and 0.80 a.u., respectively, and for 6-311G* they become 1.78 and 0.97 a.u., respectively. Comparing to *ab initio*, the two-body FMO2 method (*vide infra*) has the errors of 4.6 and 35.9 kcal/mol for 6-31G* and 6-311G*, respectively. Thus, larger ΔQ values hint at a decrease in FMO accuracy; it is necessary to proceed to the three-body FMO3 method (then the error is decreased to 0.3 and 5.2 kcal/mol, for 6-31G* and 6-311G*, respectively). To reduce the error even further, two residues per fragment can be assigned, with the error less than 0.2 kcal/mol (for both basis sets).

One should design fragmentation with the whole physical process in mind. For instance, if it is a chemical reaction involving a small molecule and a large one (e.g., an enzymatic reaction), one can consider adding the small molecule to some fragment in the enzyme forming the active reaction center. This would result in a better description of the reaction as a whole and more accurate energetics. It is possible, however, to employ two fragmentation schemes in the same project, which in fact is quite common in FMO applications. Namely, accurate energetics are obtained with a smaller number of fragments (e.g., two residues per fragment), whereas the structure optimization and the pair interaction analysis are often done at a coarser level (one residue per fragment).

In some cases, the initial fragmentation may prove to be not very appropriate during a geometry optimization, if the optimized structure becomes drastically different from the original one. In this case, it may be advisable to restart the optimization with an adjusted fragmentation, which usually does not happen in practice. (One example is gas phase geometry optimization of proteins, when a hydrogen atom seldom shifts between two fragments forming a salt bridge.) Another complication occurs for molecular dynamics, which can result in considerably distorted or else very different geometries, and some way to dynamically adjust the fragmentation was developed.[48] In addition, when individual atomic properties are of high concern, to better describe the vicinity of the detached bonds, two fragmentation schemes shifted relative to one another by one residue were employed to compute the nuclear magnetic resonance (NMR) chemical shifts with FMO.[28]

Finally, in some systems it is difficult to apply FMO, at least in its current methodology. All of them share an underlying global delocalized character, and one can name metallic clusters as a typical example. Another is presented by porphyrine with a two-dimensional network of aromatic rings, especially for the properties involving

delocalized orbitals (e.g., excited states). Similarly, trying to mimic the infinite limit band gap by extending the linear cluster size (e.g., in polymers) would be difficult with FMO, although the molecular orbital definition[64] for FMO may help treat this property (also well described by the elongation method[6]). For properties involving the delocalized orbitals (porphyrine, etc.), the three-body expansion in FMO can be expected to considerably reduce the error, especially with a larger fragment size.

2.2.2 Basic Theory

When the fragmentation details are specified, one can perform FMO calculations. The basis scheme[2] of FMO is simple: perform self-consistent fragment calculations followed by fragment pairs. For higher accuracy, one can add fragment triple calculations. Fragments, their pairs, and their triples are called *monomers*, *dimers,* and *trimers*, also abbreviated as *n*-mers ($n = 1$, 2, and 3, respectively).

Each of these *n*-mer calculations is performed in the Coulomb field exerted by the remaining fragments. This field is also called the *electrostatic potential* (ESP). For instance, fragment 2 in Figure 2.1b is immersed in the Coulomb field due to the electron density of fragments 1, 3, and 4, as well as their nuclei. Thus, ESP adds the electron–electron repulsion and the electron–nucleus attraction between fragment 2 and the rest of the system.

Comparing to the full *ab initio* calculation, only the exchange interaction between the given *n*-mer and the rest of the system is ignored. It is necessary to do so, because this exchange interaction is connected to charge transfer, and the latter is prohibited (i.e., *n*-mers have a constant electron count). Some attempts to introduce exchange to ESP in FMO were suggested,[65] which we find questionable. Other than the addition of ESP, *n*-mer calculations are conducted in the usual *ab initio* fashion, with the full exchange and charge transfer accounted for within the given *n*-mer, and these dimer and trimer calculations introduce many-body effects to the interfragment interactions beyond electrostatics.

The electrostatic field in ESP is always computed from the monomer densities. Thus, monomer densities converged in SCF affect the field they exert upon other fragments, and hence the monomer calculations have to be repeated self-consistently. This procedure was referred to as self-consistent charge (SCC) or monomer self-consistent field (SCF) in FMO publications. Starting from some initial guess (usually, extended Hückel) for the fragment electron densities, each fragment calculation is performed in the field of the electron densities of other fragments from the previous iteration, until the energy of each fragment converges. Because ESP is determined from the monomer densities, dimer and trimer calculations are performed once.

Dimers and trimers are constructed automatically as a union of all atoms in the corresponding fragments. That is, if there is a detached covalent bond between them, in the corresponding dimer the bond is intact (undivided). Such dimers are called *connected*, and they introduce explicit corrections for the detached bonds. Numerically, these corrections are large (for C, about –15 a.u.), because formally they contain the intra-atomic interaction energy, because one electron and one proton from the BDA are assigned to the other fragment. The interfragment interaction energy for all other dimers (called *unconnected*) is on the order of the usual intermolecular interactions

(typically 0 to 20 kcal/mol, but for charged fragments in vacuum it can be as large as 100 to 150 kcal/mol).

The energy expression in the two-body FMO expansion (FMO2) is:

$$E^{\mathrm{FMO2}} = \sum_{I}^{N} E_I + \sum_{I>J}^{N} (E_{IJ} - E_I - E_J) \tag{2.3}$$

That is, the total energy E of the full system is written as the sum of the monomer energies E_I, and the pair corrections $E_{IJ} - E_I - E_J$, where E_{IJ} is the energy of the dimer made of two fragments I and J. To obtain better accuracy, one can define the three-body FMO expansion (FMO3):

$$\begin{aligned} E^{\mathrm{FMO3}} = E^{\mathrm{FMO2}} + \sum_{I>J>K}^{N} \{&(E_{IJK} - E_I - E_J - E_K) - (E_{IJ} - E_I - E_J) \\ &- (E_{JK} - E_J - E_K) - (E_{KI} - E_K - E_I)\} \end{aligned} \tag{2.4}$$

which includes the triple corrections computed from trimer energies E_{IJK}.

Individual n-mers (X) energies are obtained from solving the necessary *ab initio* equations, where the *ab initio* Hamiltonian **H** is modified by the addition of ESP **V** and the projection operator **P** (X is I, IJ, or IJK, for n = 1, 2, and 3, respectively):

$$\tilde{H}^X_{\mu\nu} = H^X_{\mu\nu} + V^X_{\mu\nu} + P^X_{\mu\nu} \tag{2.5}$$

$$V^X_{\mu\nu} = \sum_{K\neq X}^{N} \left\{ \sum_{A\in K} \left\langle \mu \left| -\frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|} \right| \nu \right\rangle + \sum_{\rho\sigma\in K} D^K_{\rho\sigma} (\mu\nu|\rho\sigma) \right\} \tag{2.6}$$

$$P^X_{\mu\nu} = B \sum_{i\in X} \langle \mu | \varphi^h_i \rangle \langle \varphi^h_i | \nu \rangle \tag{2.7}$$

The electrostatic potential has a one-particle form and is a straightforward matrix element of the Coulomb operator, describing the attraction between the electron density of X and nuclei of fragments $K \neq X$, and the corresponding electron density–density repulsion. The nucleus–nucleus repulsion energy is added directly to E_X. μ, ν, ρ and σ run over atomic orbitals in X, K runs over N-n fragments not included in X. Z_A and $\mathbf{R}_A$ are atomic charges and coordinates, respectively. **D** is the density matrix.

The projection operator matrix **P** is built upon the hybridized orbitals φ^h_i. (B is a universal constant set to 10^6 a.u.) The reason why **P** is necessary is as follows. The bond detached atoms are redundant in the calculations (introduced in the form of quasi-atoms on the other side of the detached bonds). If no restriction is imposed, the fragment containing the BDA (e.g., fragment I in Figure 2.2) would have a tendency to occupy the detached bond region (assigned to the fragment on the other side

with BDA represented as a quasi-atom). For the latter fragment (e.g., fragment $I + 1$ in Figure 2.2), the electron density describing the bond without further restrictions would fall into the 1s state of the quasi-atom, or else occupy the sp^3 orbitals other than the detached bond.

The solution to impose restrictions for proper electronic states of the divided fragments is to project out the corresponding orbitals. For the system in Figure 2.2, the sp^3 orbital on the BDA describing the detached bond is projected out from fragment I, leaving the other orbitals in the variational SCF space (so that they form three sp^3 and one 1s orbitals on the BDA shown in Figure 2.2 by their physical nature, not by a restriction). Similarly, three sp^3 and one 1s orbitals on the BDA are projected from the variational space of fragment $I + 1$, so that the remaining freedom allows for the sp^3 orbital describing the detached bond (shown for fragment $I + 1$). The working of $\mathbf{P}$ in Equation 2.7 is similar to lifting the corresponding orbital energies to $B = 10^6$, effectively removing them from the variational space in SCF.

Thus, to perform FMO calculations, one needs to specify the fragmentation points (pairs of atoms defining the detached bonds) and the hybrid orbitals φ_i^h. Fortunately, φ_i^h are very easy to construct. By taking a small model system (CH_4 for sp^3 C), one computes its localized SCF orbitals and takes the coefficients for the carbon basis functions, defining the expansion of φ_i^h over carbon atomic orbitals. It is also necessary to point one C–H bond in the model system along the z-direction, because φ_i^h are automatically rotated so that the internal z-axis for the prestored coefficients matches the actual bond orientation. Thus, only one basis set dependent bundle of sp^3 orbitals is needed to do the FMO calculation of a large system with the same basis set. (In multilayer FMO, one set for each layer is specified, see Section 2.2.7.)

2.2.3 Alternative Bond Detachment Scheme in FMO

Above we described the traditional bond detachment scheme, which was used in all FMO publications so far, except one.[35] Recently, an alternative fragmentation scheme was developed,[35] which differs from the traditional one in the way of detaching covalent bonds. The traditional scheme relies on the hybrid orbital projection operator (Equation 2.7), abbreviated as HOP. The other scheme uses adaptive frozen orbitals (AFOs).

The electron and proton assignment of the detached bonds is identical in the two schemes, AFO differs from HOP in not adding the projection operator to the Hamiltonian in Equation 2.5, and instead, two operations are performed. First, the molecular orbital describing the detached bond is computed for a small model system and is frozen during FMO calculations. Second, redundant orbitals of the BDAs are projected out analogously to HOP, but using Fock operator transformations. The main difference between the HOP and AFO approaches lies in freezing the detached bond molecular orbital and its electron density, and there are also some lesser differences in the details of the orbital projection of the BDA orbitals. Full details can be found in Reference [36].

What is the need to use AFO in FMO? It was found that detaching several bonds located adjacent to each other strongly disturbs the electron density distribution

around the detached bonds. This does not happen in proteins, as the fragmentation points are far from each other, but it occurs ubiquitously in solids and surfaces, when atoms do not form one chain but are connected three dimensionally. Thus, to apply FMO to surfaces and solids, the AFO scheme is needed for better accuracy.

The disadvantage of the AFO scheme is the reduction of the total polarization of the system (because of the frozen density of the detached bonds), and it was found that the original (HOP) scheme has smaller errors when applied to highly polar and charged systems, such as proteins, provided that small or medium basis sets are used (double-ζ with polarization or smaller). For larger basis sets, detached bonds seem to be overpolarized in HOP, without the complementing charge transfer (which is not allowed at the monomer level), and the AFO scheme delivered better accuracy (6-311G*). In addition, the AFO scheme reproduced accurately the relative stabilities of the α-helix and β-strands of polyalanine, whereas the HOP scheme tends to overstabilize the α-helices.

From the user point of view, AFO is easier to use, as the model systems are automatically constructed, and the frozen orbitals are automatically calculated, eliminating the need to provide basis set dependent hybrid orbitals in the input file. In order to run FMO calculations with AFO, one has to specify only the pairs of atoms (BDA and BAA), describing the bonds to be detached, and a table of assignment of each atom to a fragment. User-friendly modeling software Facio generates this table automatically, so the user needs only to provide a set of Cartesian coordinates and choose fragmentation points using GUI.

2.2.4 Scaling of FMO

By looking at Equation 2.3, two questions naturally arise. First, is there a double counting of the Coulomb interaction, included both in ESP and in the *ab initio* Hamiltonian **H** (Equation 2.5)? It was shown diagrammatically[54] that the balance is proper. Second, is there any time savings relative to doing the full calculation? Equation 2.3 involves all possible pairs and triples of fragments, whose number is $N(N-1)/2$ and $N(N-1)(N-2)/6$, respectively. (N is the number of fragments.)

To answer the last question, let us for simplicity consider restricted Hartree–Fock (RHF) of a cluster of N water molecules, divided into N fragments, and let us assume the N^3 scaling of SCF. Formally, SCF scales as N^4 (the number of two-electron integrals), however, due to the integral screening, it practically reduces to about N^3 for systems with several dozens of atoms or more (it cannot fall below N^3 because of the $N \times N$ matrix operations). Define the time to do the *ab initio* N water molecule calculation as t_N. Then,

$$t_N \approx t_1 N^3 \tag{2.8}$$

In FMO2, two steps are performed. In the first, SCF calculations of monomers are done, which take

$$t_N^1 \approx 10 \times 2 t_1 N \tag{2.9}$$

The factor 2 is approximate, and it appears so because ESP calculations very roughly take the same time as SCF itself. The factor of 10 appears because monomer calculations are performed self-consistently, and 10 iterations are taken here as their number.

Consequently, pair calculations are performed, which take

$$t_N^2 \approx 2 \times 2^3 t_1 N^2/2 \tag{2.10}$$

where we took the number of dimers to be for simplicity equal to $N(N-1)/2 \approx N^2/2$, and 2^3 is the extra factor, because dimers are twice as large as monomers, and take eight times more time (assuming the cubic scaling of RHF).

Therefore, the total FMO timing without approximations is given by

$$t_N^{\mathrm{FMO2}} = t_N^1 + t_N^2 \approx 10 \times 2t_1 N + 2 \times 2^3 t_1 N^2/2 = (20N + 8N^2)t_1 \tag{2.11}$$

One can see roughly one order of magnitude time savings (cf. Equation 2.8) for FMO2 without approximations, which formally scales as N^2. This already is a tremendous saving, for instance if a cluster of $N = 1000$ water molecules is computed, the reduction of the scaling from N^3 to N^2 implies the speed-up by a factor of 1000. Now, efficient approximations were introduced,[26,54,66] and they can be briefly described as follows, denoted by their keywords in GAMESS:

1. *RESPAP* — Two-electron terms in ESP in Equation 2.6 are computed using Mulliken atomic orbital populations, reducing $O(L^4)$ integrals in ESP into $O(L^3)$ (where L is the number of basis functions in a fragment).
2. *RESPPC* — Two-electron terms in ESP in Equation 2.6 are replaced by the matrix elements of the Coulomb interaction with the atomic Mulliken charges, reducing formally $O(L^4)$ integrals in ESP into $O(L^2)$.
3. *RESDIM* — The internal dimer energy E'_{IJ} is represented as the electrostatic interaction between I and J monomer densities and nuclei. This approximation replaces $N(N-1)/2$ SCF calculations by aN SCF and $N(N-1)/2 - a$ electrostatic (ES) calculations. a is structure dependent, roughly equal to 3 to 6.
4. *RCORSD* — This is similar to RESDIM but applies to correlated wavefunction. If RCORSD is smaller than RESDIM, then some pairs are computed at the SCF level without electron correlation.
5. *RITRIM* — This is an array of four elements applied to trimers analogously to RESDIM (the first three) and RCORSD (the fourth) for dimers. In trimers, there are three interfragment distances, and thus more than one element is needed. RESDIM replaces SCF calculation by an ES one, but RITRIM completely avoids trimer calculations for far separated trimers. This is possible because ES calculations are pair additive, and all ES corrections for a trimer vanish.

To apply these approximations, a definition of the interfragment distance is needed. The unitless distance R_{XL} between n-mer X and monomer L is defined as the

closest distance between all pairs of atoms in X and L, divided by the sum of their van der Waals radii W:

$$R_{XL} = \min_{i \in X, j \in L} \{R_{ij}\}, \quad R_{ij} = \frac{|\mathbf{R}_i - \mathbf{R}_j|}{W_i + W_j} \tag{2.12}$$

where $\mathbf{R}_i$ are atomic coordinates. $R = 0.5$ roughly corresponds to a covalent bond, and if the two contact atoms are O and H, then $R = 1$ is equivalent to 2.6 Å. The reason for introducing the above definition of R is dictated by physics: two contact hydrogen atoms are not the same as two oxygen atoms, so the plain $|\mathbf{R}_i - \mathbf{R}_j|$ distance is not appropriate for considering the separation between fragments.

The real scaling with judicious approximations thus becomes (taking a to be 5 on the average and assuming $a << N^2$)

$$t_N^{\mathrm{FMO2}} \approx 10 \times 2t_1 N + 2 \times 2^3 t_1 5N + t_1^{\mathrm{ES}} N^2/2 = 100 N t_1 + t_1^{\mathrm{ES}} N^2/2 \tag{2.13}$$

where t_1^{ES} is the time for an ES dimer calculation per one fragment ($t_1^{\mathrm{ES}} << t_1$). According to our experience, all ES dimer calculations take a fraction of the time necessary to do SCC; in other words, $t_1^{\mathrm{ES}} N^2/2 \approx 10 \times 2t_1 N b$, where b is a small constant. (b seems to be typically about 0.1 to 0.3, although it depends on molecular geometry and the sparcity of fragment pairs.) This assumption appears to hold for systems containing up to several thousand fragments. In principle, if the number of fragments is even larger, the N^2 dependence might overgrow the linear term, which has a large multiplicative constant. The practical scaling becomes

$$t_N^{\mathrm{FMO2}} \approx (100 + 20b) N t_1 \tag{2.14}$$

This is the basis for the nearly linear scaling of FMO2, discussed earlier[2] and proved by computational timings in many FMO publications.[1,26,29,34] Interestingly, FMO3 timings are also nearly linear but with a larger constant (which is because the number of SCF trimers is proportional to N, due to using the RITRIM approximation).

Obviously, 100 is larger than N^2 for small N. Thus, for a small number of fragments, FMO is slower than full *ab initio* RHF calculations. Omitting b, one can expect that to happen for $N < 10$, which is of course a rough estimate, but not too far off: for polyalanine chains with 10 residues, FMO2 calculations are about twice faster (6-31G*) than *ab initio* RHF. Now, when correlated wavefunctions are employed, the time savings are even larger, as the RHF scaling of N^3 becomes N^5 and N^7 for MP2 and CCSD(T), respectively.

2.2.5 Solvent Description

There are three usual ways to include solvent in FMO calculations: (1) explicit QM solvent, (2) explicit solvent treated with molecular mechanics (MM), and (3) continuum solvent (PCM). In addition, some attempts[67] were made to combine the gas phase FMO calculations with another continuum model, the Poisson–Boltzmann surface area method.[68]

In the explicit QM case, one simply adds some water molecules to the system (solute) and assigns them as FMO fragments along with the solute. This approach has been quite extensively used in FMO-based molecular dynamics,[49] and it can also be used for other purposes, such as single-point calculations or geometry optimizations. It has some very important advantages as well as some serious disadvantages. Because solvent is treated quantum-mechanically, the solute–solvent interaction is properly described, including the charge transfer and solvent polarization. In addition, no parameters are needed — most importantly, no atomic radii (see below). The downside is the need to do a long time dynamics or at least a very careful configurational sampling, without which the solute–solvent binding is largely overestimated (as most solvent molecules move around and spend only some part of the time bound to the solute); also, the solvent energy loss usually termed the cavitation energy should be evaluated to define the solvation energy.

The explicit MM treatment of water was used to conduct geometry optimizations[59] of solvated polypeptides with FMO, when TINKER was combined with a local version of GAMESS to handle the MM calculation. This approach is not included in the production version of GAMESS and is still under development. It has a drawback of overstabilizing the solute–solvent hydrogen bonding by considering just one configuration. Also, due to using force fields, compared to explicit QM, the solvent effect upon the solute in this approach is reduced to mechanistic gradient corrections, neglecting the solvent-induced polarization and solute–solvent charge transfer, while gaining some computational efficiency. Nevertheless, some solvent effects upon the structure are considered, preventing, for example, inappropriately close clustering of groups of the opposite charge in the solute.

The PCM[69] has been very extensively used to describe the solvation of small molecules. It has considerable advantages of properly describing the solute polarization, based on some averaged solvent distribution. The disadvantage of PCM is the neglect of the solute–solvent charge transfer. Typically, cations and neutral solute molecules work best with PCM (the experimental solvation free energy reproduced within 1 kcal/mol for many small systems), but anions often have larger errors (several kcal/mol).

It was found[70] that adding very few explicit QM solvent molecules can considerably improve the continuum models. Perhaps it is especially important to do so for the excited states of the solvent-exposed chromophores, as then charge transfer and the QM-driven polarization may play an important role; such addition brings up the question of some configurational sampling for the explicit water molecules, as well as a few other issues, such as the entropy contribution for the explicit water (normally, entropy loss of the solvent is included in the PCM cavitation energy, see Equation 2.15).

Another important issue of PCM is the need to define atomic radii, which can be viewed as a disadvantage in the sense of the difficulty of their definition and the remaining doubts of the appropriateness of their parametrization; or as an advantage, because the radii can be used to bring in various factors not explicitly treated in PCM in a simple form, and due to the experiment-based parametrization on a test set, direct agreement with the experiment is often observed.

At present, three sets of atomic radii may be considered for FMO-PCM applications: van der Waals radii, united atom model for Hartree–Fock[71] (UAHF), which depend upon the hybridization and the charge state of an atom) and their

simplified set (SUAHF, one radii per atom). The former and the latter are built in GAMESS, and to use the UAHF set, one should generate them elsewhere and input manually. Although it is often assumed that the UAHF set works best (for small molecules), it may not be so for biological systems, and we often use the other two choices as well. Clearly, a well-fitted set of atomic radii for biological systems is needed.

A detailed description of FMO-PCM can be found elsewhere,[72] and here we provide only a brief practical outline. The solute is placed in a solvent cavity, made as a union of atomic spheres. Each sphere is divided into pieces called tesserae (usually, 240 pieces per sphere), and each tessera has a point charge on it. These apparent surface charges (ASCs) representing averaged configuration of the solvent are added to quantum-mechanical calculations, and they are determined self-consistently, mutually polarized by the solute.

The self-consistent loop to determine ASCs is built into SCC for FMO-PCM[1], which expands the PCM potential (exerted by the ASCs upon the solute) in the sum of monomer contributions, or it is put around FMO2 calculations for FMO-PCM[2], where the PCM potential is computed from both monomer and dimer densities similar to in Equation 2.3. An inexpensive and efficient version of the latter method is FMO-PCM[1(2)], when the PCM loop around FMO2 calculations is limited to two iterations, which corresponds to a noniterative version of FMO-PCM[2]. Two iterations are needed, because during the first one, the two-body corrected PCM potential is obtained, and the ASCs derived from it are used in the fragment calculations on the second iteration. The many-body expansion of the PCM potential can be varied independently of the degree of the energy expansion in Equation 2.3.

One of the basic assumptions in PCM is that the solvent charge is fully contained in the cavity, and it is not exactly satisfied when the solvent is described by quantum mechanics with a density distribution. Several charge compensation schemes[69] were developed to renormalize the charge in PCM equations, one of them (ICOMP=2) is implemented for FMO.

These PCM calculations coupled with QM give the polarized solute, whose internal energy can be easily separated from the electrostatic solute–solvent interaction, and the polarized solvent, represented by the polarizable cavity. The total PCM energy G_{PCM} is usually defined as

$$G_{\mathrm{PCM}} = G'_{\mathrm{PCM}} + \Delta G_{\mathrm{es}} + \Delta G_{\mathrm{cav}} + \Delta G_{\mathrm{disp}} + \Delta G_{\mathrm{rep}} \tag{2.15}$$

where G'_{PCM} is the internal, ΔG_{es} solute–solvent electrostatic, ΔG_{cav} cavitation, ΔG_{disp} solute–solvent dispersion, and ΔG_{rep} solute–solvent repulsion energies, respectively. These energies are often called free energies, due to the included effect of solvent motion (usually at 298K); they are only partial free energies if a single point QM calculation is done; the free energy contributions to the solute should be added to obtain full free energies. The cavitation energy ΔG_{cav} describes the energy loss of the solvent (including reduced entropy) to create the cavity. ΔG_{disp} relates quite well to the MP2 solute–solvent dispersion interaction (averaged over the motion of solvent molecules). ΔG_{rep} corresponds to the exchange-repulsion in RHF (i.e., the nonelectrostatic part of the RHF solute–solvent interactions).

Because PCM is designed with a good correspondence to QM, it is suitable to study the protein–ligand binding, because the latter is driven by the balance of protein–solvent, ligand–solvent, and protein–ligand interactions. ΔG_{cav}, ΔG_{disp}, and ΔG_{rep} are fitted to some simple form (such as cavity surface and pair interatomic potentials) and are independent of the electronic density (and of any QM property of the actual solute). PCM can be used for any solvent, not just water, and the parameters defining it can be manually specified.

2.2.6 Electron Correlation

Four ways to account for the electron correlation in FMO exist:

1. Density functional theory (DFT)
2. Møller–Plesset second-order perturbation theory (MP2) or coupled cluster (CC)
3. MCSCF and TDDFT
4. CI

DFT is a computationally cheap method that can include electron correlation in the density functional form. From the computational point of view, DFT is similar to RHF: it is a single reference method, and the additional cost to describe the electron correlation scales less than RHF. The drawback of DFT is its inability to describe the dispersion interaction, although the addition of dispersion to DFT is an active area of the present research,[73] and some solution may be available. From the point of view of FMO, DFT is more problematic than RHF, because it is more sensitive to fragmentation, and FMO-DFT has a several times larger error than FMO-RHF. Thus, to obtain better accuracy in FMO-DFT, a larger fragment size may be recommended.

MP2 and CC are also single-reference methods that, however, describe the dispersion properly, making them suitable for studying biological molecules. MP2, especially in its MP2(IMS) implementation[26] in GAMESS, is very efficient and takes little more time compared to RHF, in the case of single-point energy; computing the MP2 gradient, however, is more costly, and it may be about three times more expensive than RHF. MP2(IMS) relies on a heavy use of I/O, so parallel clusters without a local disk may need to use the alternative MP2(DDI) implementation, which requires a very large memory, almost no disk, and fast networking. The FMO version of MP2 or CC is a straightforward extension of Equation 2.3 for correlated energies.

CC in its CCSD(T) form is considered the gold standard of QM, and it delivers accurate energetics, frequently agreeing with experiment within 1 kcal/mol. The cost of CC is huge, and large memory and disk are required. At present, FMO-CC is practically feasible for molecular clusters, and, possibly, for multilayer FMO of general systems, when only the important part is treated with CC. Recent CC development in GAMESS provides new exciting ways to describe chemical reactions.[74]

MCSCF allows real multireference calculations, which can be considered indispensable in many systems involving transition metals. TDDFT is a single excitation scheme applied to DFT. In MCSCF only the static electron correlation is accounted

for (within the active space), whereas in TDDFT explicitly the dynamic correlation is computed. (The electron correlation of both static and dynamic type is included to some extent in the ground state of DFT, which provides the basis for computing the excitations in TDDFT.)

From the FMO point of view, these two methods are similar in the following regard. In both cases, one specifies the main fragment (MCSCF or TDDFT), and other fragments are RHF or DFT, respectively. The pair corrections are done at the MCSCF (TDDFT) level only for those dimers, which include the main fragment. All other pairs are computed at the RHF (DFT) level. The difference between FMO-based MCSCF and TDDFT is that in MCSCF, other fragment densities in SCC are determined self-consistently with the MCSCF fragment density, whose electronic state can be excited (e.g., a triplet) or ground; in TDDFT the fragment densities in SCC are always computed self-consistently for the ground DFT state (singlet). Thus, the density relaxation is better accounted for in MCSCF. In order to add pair corrections, the size-extensive complete active space (CAS)–type of MCSCF should be used.

When performing FMO-MCSCF calculations, the same problems are faced as in *ab initio* MCSCF — difficulties in defining the active space, preparing the good initial orbitals, and converging. In addition, MCSCF without dynamic correlation may not deliver sufficiently good energetics, although it is usually believed to give good geometries.

For CI, the present method development is limited to multilayer FMO-RHF:CI, with just one CI fragment assigned to the upper layer. The lack of the size extensivity in most CI methods prohibits adding pair corrections.

2.2.7 Multilayer and Multibasis Calculations

It is often the case that some part of the system is of particular interest, such as an active center for a chemical reaction or a chromophore for excited states. In FMO, one can address this by using multilayer FMO (MFMO),[75] when several fragments are assigned to a higher layer. Wavefunctions and basis sets can be defined separately for each layer. Such a setup may be ideal for using highly correlated wavefunction (CC) or a large basis set for a small part of the system, avoiding large computational costs, convergence difficulties, and considerable FMO errors for large basis sets. Force fields at present cannot be used as a layer. Because FMO-MP2 optimizations are still expensive, one can use multilayer FMO to reduce their cost.

The computational scheme of MFMO was already described in detail.[75] Summarizing it briefly for the case of two layers, first SCC calculations are performed at the level of the first layer for *all* fragments (both layers 1 and 2). Then dimer calculations are conducted for dimers between the two layers, and for dimers within layer 1. Consequently, SCC is done for fragments in layer 2, while fixing the electron densities in layer 1. That is, ESP for fragments in layer 2 is made of ESP for fragments in layer 1 (fixed at this point), and the variationally determined ESP for fragments in layer 2. Finally, dimers in layer 2 are computed. The final energy

expansion of MFMO is identical to Equation 2.3, with the provision that monomer and dimer energies are used from the highest layer at which they were computed. MFMO follows the *e pluribus unum* spirit of FMO, by computing fragments and dimers in all layers in the presence of all other fragments, always considering the electronic state of pieces as a part of the whole.

Multibasis runs are not the same as multilayer FMO with several basis sets. In a multibasis setup, some atoms may be allowed to have a basis set different from other atoms of the same kind. A typical example is given by carboxyl groups, where diffuse functions are frequently added. Therefore, in multibasis calculations, fragment densities are computed always with the same predefined basis set. On the contrary, in MFMO, layer 1 fragment densities are variationally determined with the layer 1 basis set. Typically, very few atoms may be assigned a different basis set in the multibasis case, or otherwise, an artificial polarization may be promoted on the border of the two basis sets, which is largely avoided in MFMO because the layer 1 densities are optimized at the lower layer in the presence of all fragments at the same level.

2.3 PAIR AND TRIPLE INTERACTIONS

The FMO expansion in Equation 2.3 is based on the monomer and dimer energies with the electrostatic field included, thus $E_{IJ} - E_I - E_J$ are not very convenient to discuss the pair interactions between fragments. To separate the ESP term, Equation 2.3 can be rewritten as follows[66]:

$$E^{\mathrm{FMO2}} = \sum_{I}^{N} E'_I + \sum_{I>J}^{N} (E'_{IJ} - E'_I - E'_J) + \sum_{I>J}^{N} Tr(\Delta \mathbf{D}^{IJ} \mathbf{V}^{IJ}) \tag{2.16}$$

where E'_I and E'_{IJ} are the internal energies of monomers and dimers, respectively, obtained by subtracting the electrostatic interaction of the electron density and nuclei in n-mers with ESP, for example, for $n = 2$, $E'_{IJ} = E_{IJ} - Tr(\mathbf{D}^{IJ}\mathbf{V}^{IJ})$. $\Delta \mathbf{D}^{IJ} = \mathbf{D}^{IJ} - \mathbf{D}^I \oplus \mathbf{D}^J$ is the density matrix difference of dimer IJ and the sum of monomer I and J electron densities, and $\mathbf{V}^{IJ}$ is the electrostatic potential due to the external fragments acting upon dimer IJ. For ES dimers, $\Delta \mathbf{D}^{IJ} = 0$ (pair additive dimer density), because no SCF is performed.

The meaning of terms in Equation 2.16 is as follows: E'_I are the internal energies of the self-consistently polarized fragments; E'_{IJ} is the internal energy of the polarized dimer IJ; $\Delta E'_{IJ} = E'_{IJ} - E'_I - E'_J$ is thus the internal pair interaction energy for polarized fragments; and $Tr(\Delta \mathbf{D}^{IJ} \mathbf{V}^{IJ})$ is the interaction of the relaxed density change with the external field (ESP). For correlated wavefunctions (e.g., MP2), $\Delta E'_{IJ}$ can be decomposed into RHF and correlation components $\Delta E'_{IJ} = \Delta E'^{\mathrm{RHF}}_{IJ} + \Delta E^{\mathrm{corr}}_{IJ}$, where $\Delta E^{\mathrm{corr}}_{IJ}$ is given by the difference of the correlation energies of dimer IJ and monomers I and J. $\Delta E^{\mathrm{corr}}_{IJ}$ basically describes the dispersion interaction between fragments.

The pair interaction energy of a pair of fragments is given by

$$\Delta E_{IJ}^{\text{int}} = \Delta E'_{IJ} + Tr(\Delta \mathbf{D}^{IJ} \mathbf{V}^{IJ}) \tag{2.17}$$

In the pair interaction energy decomposition analysis (PIEDA),[63] the interaction energy $\Delta E_{IJ}^{\text{int}}$ is decomposed into the same contributions as in the EDA[4]: electrostatic (ES), exchange-repulsion (EX), charge transfer plus higher-order mixed terms (CT+mix), and dispersion (DI) contributions. In contrast to EDA, PIEDA can be applied not only to molecular clusters and complexes,[76] but also to covalently bound systems.

In CAFI,[62] the CT and the stabilization component of PL can be defined by performing the appropriate CIS calculations within each dimer. In addition to the total PL and CT values, it is possible to analyze the individual orbital contributions, which may be of special interest to some applications. If PIEDA or CAFI are not applicable (e.g., for FMO-DFT), then one can use the two components in Equation 2.17 for the interaction analysis.

When PIEDA is applied to the usual electronic state of fully polarized fragments in FMO, the components are as follows (this analysis is chosen with IPIEDA=1):

$$\Delta E_{IJ}^{\text{int}} = \Delta E_{IJ}^{\text{ES}} + \Delta E_{IJ}^{\text{EX}} + \Delta E_{IJ}^{\text{CT+mix}} + \Delta E_{IJ}^{\text{DI}} \tag{2.18}$$

In the present implementation, PIEDA applies to RHF, MP2, or CC. The DI component is defined from the correlation energy of the latter two methods. Polarization (PL) in PIEDA is divided into two contributions: the destabilization PL (from monomer energies E'_I) and the stabilization PL, from a part of the electrostatic energy $\Delta E_{IJ}^{\text{ES}}$.

To define the polarization and also the coupling terms (polarization-exchange, polarization-dispersion, polarization-charge transfer, and the many-body polarization terms), one has to apply the full version of PIEDA (chosen with IPIEDA=2). This, however, requires a definition of the *free state* of fragments. For molecular clusters, the free state is naturally available as the stand-alone molecules. With the fragmentation of covalent bonds, the definition of the free state is rather arbitrary. We defined it for fragments with minimally possible caps, which for C–C bonds means methyl caps. For example, considering fragment 2 in Figure 2.1, its free state will be to take fragment 2, capped with methyl groups on both ends (e.g., the upper $-CH_3$ is similar to the group of C9,H11,H10,N7 atoms).

Consequently, the free electronic state of fragment 2 is computed in the FMO calculation ($N = 3$, two caps form separate fragments). Also, to separate the interatomic interaction for the BDAs, one has to compute CH_3–CH_3 (for C–C bonds) in a two-fragment FMO run. All of these calculations are performed automatically for a given system, and one has to do four separate runs to execute a full IPIEDA=2 calculation: (1) free state, (2) BDA corrections, (3) free state PIEDA (the PL0 state), and (d) the fully polarized state (the PL state). In the case of molecular clusters, BDA corrections are not needed.

A convenient way to consider the interactions is to divide the two dimer sums in Equation 2.16 into the connected and unconnected dimer contributions. The sum of

the former with the monomer energies is called the backbone energy, to which the latter terms add the nonbonding interaction of the molecular cluster type.

$$E = E^{\mathrm{BB}} + \Delta E^{\mathrm{U,int}}$$

$$E^{\mathrm{BB}} = \sum_{I}^{N} E'_I + \sum_{\substack{I>J \\ R_{IJ}=0}}^{N} \Delta E^{\mathrm{int}}_{IJ} \tag{2.19}$$

$$\Delta E^{U,\mathrm{int}} = \sum_{\substack{I>J \\ R_{IJ}\neq 0}}^{N} \Delta E^{\mathrm{int}}_{IJ}$$

where the distance between two fragments R_{IJ} is zero if they are connected. The backbone energy E^{BB} thus represents the stability of the polarized chain of fragments without any other interactions (except for the joint points holding the chain together, the latter type is included in E^{BB}). $\Delta E^{\mathrm{U,int}}$ contains the pair interaction energies for unconnected fragments, and the values of its components are similar in magnitude to intermolecular interactions. The meaning of various terms in PIEDA and the way to use them in applications can be best understood from the discussion of the application of PIEDA to the relative stabilities of polyalanine isomers.[63]

An extra benefit of Equation 2.16 is the explicit absence of monomer potentials $\mathbf{V}^I$, improving the accuracy of the ESP approximations. Their application based on the distance definition in Equation 2.12 is known as the uniform scheme (MODESP=0), and it results in a contribution from some fragment to the *IJ* dimer ESP computed at a higher level (e.g., without approximations), compared to the same fragment exerting Coulomb field upon the separate fragments *I* or *J*. A better balance of ESP approximations is by applying them to the fragment blocks in the ESP matrix, in accord with the approximations for those fragments individually. This method is known as the *n*-mer consistent scheme (MODESP=1), and its use is thought to be essential for FMO3 calculations with ESP approximations.

Finally, we address the triple corrections. Explicit three-body corrections provide food for thought about the functioning of biological and other large systems, and FMO3 delivers accurate energetics, often within 1 kcal/mol to *ab initio*.

$$E^{\mathrm{FMO3}} = E^{\mathrm{FMO2}} + \sum_{I>J>K}^{N} \Delta E^{\mathrm{int}}_{IJK} \tag{2.20}$$

$$\Delta E^{\mathrm{int}}_{IJK} = \Delta E'_{IJK} + \Delta E^{\mathrm{D}}_{IJK} \tag{2.21}$$

$$\Delta E'_{IJK} = E'_{IJK} - E'_I - E'_J - E'_K - \Delta E'_{IJ} - \Delta E'_{IK} - \Delta E'_{JK} \tag{2.22}$$

$$\Delta E^{\mathrm{D}}_{IJK} = Tr(\Delta \mathbf{D}^{IJK}\mathbf{V}^{IJK}) - Tr(\Delta \mathbf{D}^{IJ}\mathbf{V}^{IJ}) - Tr(\Delta \mathbf{D}^{IK}\mathbf{V}^{IK}) - Tr(\Delta \mathbf{D}^{JK}\mathbf{V}^{JK}) \tag{2.23}$$

where $\Delta\mathbf{D}^{IJK} \equiv \mathbf{D}^{IJK} - (\mathbf{D}^{I} \oplus \mathbf{D}^{J} \oplus \mathbf{D}^{K})$ contains the density relaxation within trimer *IJK* relative to separate fragments (combining both pair and triple corrections; the former are subtracted in Equation 2.23). Similarly to pair interactions, for correlated wavefunctions (e.g., MP2), $\Delta E'_{IJK}$ can be decomposed into RHF and correlation components $\Delta E'_{IJK} = \Delta E'^{\mathrm{RHF}}_{IJK} + \Delta E^{\mathrm{corr}}_{IJK}$.

$\Delta E^{\mathrm{corr}}_{IJK}$ is the correction to the interfragment dispersion interaction $\Delta E^{\mathrm{corr}}_{IJ} + \Delta E^{\mathrm{corr}}_{IK} + \Delta E^{\mathrm{corr}}_{JK}$ due to the charge transfer and other quantum effects involving the third fragment (for all three pairs *IJ*, *IK,* and *JK*, e.g., fragment *J* acting upon dimer *IK*). Similarly, $\Delta E'^{\mathrm{RHF}}_{IJK}$ is a correction to the sum of the three uncorrelated pair interaction energies $\Delta E'^{\mathrm{RHF}}_{IJ} + \Delta E'^{\mathrm{RHF}}_{IK} + \Delta E'^{\mathrm{RHF}}_{JK}$. The values of $\Delta E'_{IJK}$ and $\Delta E^{\mathrm{D}}_{IJK}$ are printed in GAMESS and can be used for analysis.

2.4 ACCURACY AND THE CHOICE OF THE LEVEL OF CALCULATIONS

The basic rule determining the accuracy of FMO versus *ab initio* is that larger fragments have better accuracy. Another way to improve it is to go from FMO2 to FMO3. Also, an important factor is the proper fragmentation reducing the interfragment charge transfer. After a calculation is done, one can check the amount of charge transfer with the ΔQ and $\Delta\bar{Q}$ criteria described above. In addition, larger basis sets have a larger error, which is related to the coupling of the polarization and charge transfer, as revealed by PIEDA.

When choosing the level of calculation, one should consider the following factors:

1. Is the description of the whole system important? A typical example is when one compares relative stabilities of very different conformers. If so, a higher level should be used (frequently, FMO3). Otherwise, if one is interested in an energy difference of two similar structures as in chemical reactions or protein–ligand binding, a lower level usually suffices (FMO2).
2. For geometry optimizations, fairly low levels are often appropriate (FMO2 with 1 residues per fragment). FMO-optimized structures of polypeptides agree with *ab initio* within about 0.1 to 0.2 Å (RMSD).[59]
3. Is a large basis set (3ζ with polarization or better) to be used? If so, a higher level should be used (FMO3) for better accuracy, or one can consider multilayer FMO2 with the large basis set in the higher layer. Diffuse functions in particular are difficult to treat in FMO, and multilayer or multibasis setup may be the only way in many cases.
4. Is the primary purpose of the calculation to perform the pair interaction analysis? If so, one residue per fragment is usually convenient. One can also add to it the total sum of the three-body corrections for better accuracy of the total properties (from a separate FMO3 run). For instance, if one studies protein–ligand binding, and wishes to analyze residue–ligand interactions, one can do PIEDA (corresponding to FMO2) and to that add three-body corrections from FMO3, in order to get better binding energies, as well as the pair interaction analysis.

5. Are the computational resources sufficient? FMO3 at the medium level of accuracy as a very rough estimate takes about six times more (for RHF) than FMO2 at the same fragmentation level. One can also consider choosing between FMO2 with two residues per fragment and FMO3 with one residue per fragment. The latter is in general more accurate and somewhat more expensive, although not very much more.
6. The wavefunction also has some effect upon the computational level choice. DFT in particular appears to require large fragments in FMO (at least two residues per fragment). TDDFT frequently works well at a modest level (FMO2), but one should be careful to consider if the excitation of interest is strongly delocalized, in which case a higher level is needed.
7. The solvation energies in PCM seem to be very well described at the FMO2 level (for 6-31G*).[72] PCM in general appears to stabilize the electronic state in terms of convergence and the accuracy in FMO.
8. The many-body expansion of the PCM potential can be varied independently from the energy expansion. It was found that the lowest level of FMO-PCM[1] works well, although adding dimer corrections to it (FMO-PCM[1(2)]) is recommended for reliable energetics.

2.5 PRACTICAL ASPECTS OF PERFORMING FMO CALCULATIONS IN GAMESS

2.5.1 Parallelization

FMO in GAMESS is parallelized with the two-level hierarchical method, generalized distributed data interface (GDDI).[77] The underlying basic DDI is an upper-level interface that stands between GAMESS and the lower-level communication libraries: socket or MPI. DDI eliminates the need to add library-dependent calls to the quantum-chemistry code (GAMESS), and it also provides means to distribute matrices over computer nodes and perform operations on them.

At present, GDDI fully works with the socket version of DDI, although in some environments GDDI can be used with MPI.[52] In general, the socket version is considered to be more efficient for the parallelization of GAMESS; however, some parallel clusters do not support those parallel applications that do not use MPI. In typical FMO applications, the distributed matrix mechanism of DDI is not used, and the socket library provides the functionality to handle the two-level parallelization of GDDI.

Good parallel scalability of FMO was reported (70% to 90%) on 128 Pentium III nodes (90% means the speed-up of 0.9*128 = 115), using slow Fast Ethernet communication.[77] So far, the largest reported parallel FMO run involved 900 Opteron nodes[52] on the AIST SuperCluster (Japan) and 4096 vector processor units[33] on the Earth Simulator (Japan).

The grouping mechanism of GDDI is shown in Figure 2.3. Physical nodes Ni ($i = 0$–2) are typically Unix computer nodes ("boxes") that have several CPU cores in them, and in the socket case they are required to have an Internet Protocol (IP), usually associated with a hostname denoted by node$i$. The RUNGMS script is used to

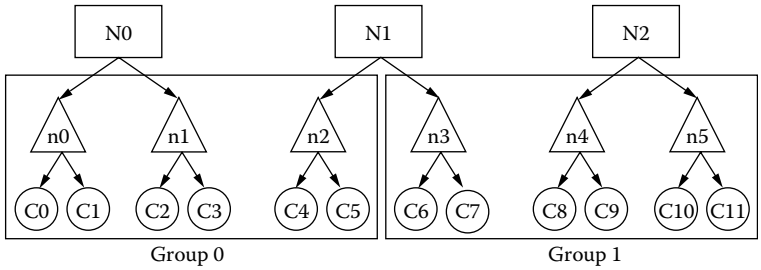


FIGURE 2.3 Physical (N*i*) and logical (n*i*) nodes and central processing unit (CPU) cores (C*i*), divided into groups in generalized distributed data interface (GDDI).

execute GAMESS, and the details of the available nodes are given in the HOSTLIST variable. One can define logical nodes in HOSTLIST by repeating the same hostname several times, and correspondingly reducing the number of CPU cores associated with it. For example, to define a logical node to be the same as a physical node, one would have

HOSTLIST="node0:cpus=4 node1:cpus=4 node2:cpus=4"

Here we assume that physical nodes node*i* have four CPU cores (denoted by C*i* in Figure 2.3, such as C0), in practice there are either two dual-core CPUs or a single quad-core CPU per node, and the two cases are identical as far as GAMESS setup is concerned. To define a logical node to have two CPU cores, one would have

HOSTLIST="node0:cpus=2 node0:cpus=2 node1:cpus=2 node1:cpus=2
node2:cpus=2 node2:cpus=2"

The latter case is visualized in Figure 2.3, and logical nodes are denoted by n*i*, i = 0–5. The need to define logical nodes in RUNGMS is driven by the GDDI implementation that cannot subdivide logical nodes into smaller units during the group definition.

With the above setup in RUNGMS, GAMESS starts off with six logical nodes, each having two CPU cores. These six nodes can be arbitrarily divided into groups. There are two reasons to vary the group division: a better parallel efficiency and to accumulate sufficient resources (memory and disk) per group. The group division can be automatic (with the user specifying the number of nodes per group) or manual (with each group having a desired number of nodes). The former is most commonly used; the latter is needed rarely, in two cases: semidynamic load balancing (see below) and when internode communications are inefficient, in which case it may be better to avoid having several physical nodes in the same group. (In Figure 2.3, group0 has two nodes N0 and N1.)

The workload in GDDI is distributed in two layers. In the upper layer, monomer and dimer SCF calculations are assigned to groups. In the lower layer, work is given to CPU cores, such as computing a portion of two-electron integrals for a given n-mer calculation. The load balancing can be of three types: *dynamic*,[77] *static*, or *semidynamic*,[52] and it can be chosen for the two layers independently.

(The semidynamic load balancing works only for the upper level.) In all cases at the upper level, larger jobs (i.e., those that have more basis functions) are computed first for better efficiency.

The dynamic load balancing at both levels is in most cases the best strategy. Its basic principle is that each compute unit (group or core) asks for a next portion of work upon finishing its present load. In the static case, work loads are assigned simply with a fixed schedule, in the order of loads given to CPU cores based on their ID.

The semidynamic load balancing is needed to get good parallel efficiency in the rather special case, when there are few very large fragments among many others, and many nodes are used in the run. In this case, the static load balancing is used for those large fragments, using several large groups in GDDI defined for this purpose. When these large groups finish doing large fragments with the static load balancing, they join other groups to do the remaining work in the dynamic fashion. This strategy was applied to the photosynthetic reaction center,[52] where large chlorophyll units were not fragmented for physical reasons.

Now we discuss the strategy to choose the optimum group division. Three cases are considered in Figure 2.4. In the first case, only one group ($N_G = 1$) is defined including all available CPU cores. In this case there is no synchronization penalty at the upper (intergroup) level, but on the other hand, parallel efficiency is lost because many nodes do the same SCF calculation, which has sequential parts (matrix diagonalization, etc.), and a large amount of communications, as well as due to the intragroup synchronization (when CPU cores that did their work wait for others to finish).

The second strategy is to have one node per group ($N_G = 6$). It is assumed that six fragments are computed, which have different sizes (e.g., amino acid residues are of various sizes). In this case, the intergroup synchronization (when groups having no more work wait for others to finish) is the bottleneck. It can be seen that roughly half of the time is wasted by nodes doing nothing. On the other hand, the

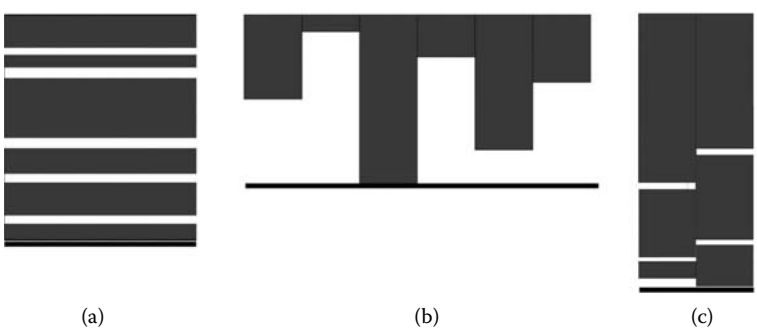


FIGURE 2.4 Comparison of node division in the FMO method: (a) one group, (b) six groups, and (c) two groups. (Note that the size of the tasks is represented by the height of the bars [=tasks, such as fragment SCF]. The empty white space before the finish thick line on the bottom and the lowest bar in a column [representing GDDI groups] shows the time wasted at the intergroup synchronization point at the end. The empty space between the bars is added to separate them.)

intragroup parallelization is efficient, as only one node with two cores works on a single fragment.

In the third case, two groups are defined. This is an intermediate case between the first two, and usually the most efficient. The synchronization penalty at the intergroup level is small because the smallest tasks are done at the end.

In FMO, the number of SCF tasks for the monomer, dimer, and trimer steps is N, aN, and cN, respectively (a and c depend on the compactness of the structure, typically a = 3–5 and c = 10–20 for globular polypeptides, assuming the default thresholds for the medium accuracy). The ES dimers are computed at the end of SCF dimers and their parallelization is much easier. Thus, one should determine the optimum number of GDDI groups among which to distribute N, 4N, and 15N tasks (taking a = 4 and c = 15). These three group numbers are specified in the NGRFMO keyword in GAMESS, as NGRFMO(1)=N1,N2,N3 where N1, N2, and N3 are the numbers of groups for monomer, dimers, and trimers, respectively.

Finally, we try to give simple advice for the best values of NGRFMO. The choice between the three above scenarios is determined by the balance of the penalties due to the synchronization and intranode parallelization efficiency. The simplest rule implemented in the modeling and input file processing programs like FMOutil[59] or Facio,[61] is to set the number of groups equal to M/3, where M is the number of tasks (i.e., N, 4N, and 15N, and N is the number of fragments). A little more elaborate choice is as follows:

1. For several nodes (N_{nod} is about 1…3), many fragments: NGRFMO(1) = 1, N_{nod}, N_{nod}.
2. For many nodes, few fragments: NGRFMO(1) = 1,1,1.
3. For many nodes, many fragments: NGRFMO(1) = N/3, 4*N/3, 5*N.

In the latter case, the number of groups cannot exceed N_{nod}. In addition, we assumed a single-layer FMO calculation. For multilayer computations, NGRFMO is defined separately for each layer, because layers usually involve a different number of SCF calculations.

Finally, we briefly discuss the memory aspect. FMO does not consume much memory, and typically less than 100 megabytes are taken. It is important to realize that in GAMESS/FMO there are three memory chunks that, with some exceptions, can only be used for specific purposes without an ability to return them and reuse them for something else. These chunks are (1) general memory, (2) in-core integral memory, and (3) DDI memory (MEMDDI).

In the beginning of a run, the MEMDDI amount is allocated on data servers for distributed matrices and cannot be used for anything else. The other two chunks are allocated from the MEMORY pool by compute processes. The in-core integral amount is given in NINTIC, and at present it can only be used for storing (1) two-electron AO integrals in memory, (2) MP2 partially transformed integrals (MP2(IMS) only), and (3) the Hessian matrix in OPTFMO. It is a common problem to allocate too much space in NINTIC, leaving not enough for SCF, because the two chunks cannot be used interchangeably (in this case, the calculation stops complaining of not having enough memory to run SCF).

For DFT, RHF, and MP2(IMS), no distributed arrays are used, thus most memory can be given to store two-electron integrals. MP2(DDI) requires very large distributed memory (MEMDDI) to store MP2 integrals. When dividing nodes into groups, one should consider the largest required amount and put enough CPU cores with their memory, so that their combined memory suffices. Similarly, parallel MCSCF with the FULLNR converger requires MEMDDI memory to store transformed integrals, and the same concerns apply. The default SOSCF converger in MCSCF does not use MEMDDI. The parallel CCSD(T) in GAMESS also needs MEMDDI memory, and one should accumulate enough CPU cores into one group to do a parallel CCSD(T) calculation.

Special attention should be paid to setting up parallel FMO-MCSCF and FMO-TDDFT calculations. The peculiarity of these methods is that they require very few correlated monomer (one) and dimer (several) calculations, involving the main MCSCF (or TDDFT) fragment; there are many uncorrelated dimer SCF calculations. Thus, the above grouping strategy supposing the same scaling behavior within monomers or dimers does not apply; some additional group setup in NGRFMO can be used to improve the parallel efficiency.

2.5.2 Analyzing FMO Results

As pointed out above, calculations of monomers and dimers are performed by individual groups in GDDI. The corresponding output goes into a separate output file produced by each group's master node. It is usually not necessary to look into individual output files, as the results are summarized in the main output file (made by CPU core0, the master node of group0). However, if the memory assignment was inappropriate (too much devoted to two-electron integrals), or if there was a convergence problem, one has to look into all output files to identify the problem.

In our opinion, the last majority of convergence difficulties are caused by problems in the input, either in the calculation setup, or by a rough structure with very close fragments. The necessity for good structure modeling cannot be overstressed. In some cases, convergence is difficult for physical reasons, especially for DFT, and it is necessary to converge the system as is. To assist in this, a number of options were added to the GUI Facio, and the user only needs to find out from the output at what stage the problem occurs.

Several ways are provided to visualize the FMO results. First, one can use the scripts supplied with GAMESS to convert raw output into the CSV format, which can be read directly by many programs. The CSV files contain either plain or decomposed (PIEDA) pair interaction energies, which can be plotted for a desired fragment (for an example, see the cover art for Reference [1]). Alternatively, one can use Facio as an interactive interface to visualize the results, including two-dimensional maps of pair interactions.

2.6 CONCLUSION

The development of the fragment molecular orbital method has enabled practical applications of *ab initio* methods, such as RHF, DFT, or MP2 to the molecules of biological size: systems with several thousands of atoms are routinely computed, and

with sufficient computational resources, several dozens of thousands of atoms are also tractable today. The many-body expansion in FMO appears to robustly converge to the full *ab initio* properties, as shown for two-body and three-body expansions, giving means for a systematic improvement of results.

The difficulty of the biological systems is not just in their large size, but also in their great flexibility, purporting the importance of the configurational sampling, entropy, and the free energy. The present applications of FMO have been performed either for a single structure or involved averaging of a very few conformations (sampled with force fields). In order to compete on par with well-conducted applications of force fields to predict the energetics comparable with experiment, some way of doing configurational sampling with FMO is desirable.

The pair interaction analysis in FMO can be very helpful in explaining the mechanism of various phenomena. Also, it is often useful to find correlation between computed and measured properties, such as in the quantitative structure–affinity relationship studies, and FMO can be valuable in this regard.

The applications of FMO are not limited to biochemical systems. Considerable success of FMO-based molecular dynamics of explicitly solvated small molecules and chemical reactions is one other area, another is the new area of solid state (catalysis,[35] surface chemistry, excited states[36]), and one can think of many more possible applications of FMO to inorganic and organic systems, some of which have already been explored.[76] Aided by easy-to-use graphical interfaces, we expect FMO to become a conventional tool in studying chemical and physical properties of molecular systems.

ACKNOWLEDGMENTS

We are greatly indebted to Professor M. Suenaga of Kyushu University for adding FMO functionality to his modeling software Facio. Our work was partially supported by the Grant-in-Aid for Scientific Research (JSPS, Japan) and the Next Generation SuperComputing Project, Nanoscience Program (MEXT, Japan).

REFERENCES

1. Fedorov, D.G., Kitaura, K. 2007. Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. *J. Phys. Chem. A* 111:6904–6914.
2. Fedorov, D.G., Kitaura, K. 2006. Theoretical development of the fragment molecular orbital (FMO) method. In *Modern methods for theoretical physical chemistry of biopolymers*, eds. E.B. Starikov, J.P. Lewis, and S. Tanaka, 3–38. Amsterdam: Elsevier.
3. Kitaura, K., Ikeo, E., Asada, T., Nakano, T., Uebayasi, M. 1999. Fragment molecular orbital method: an approximate computational method for large molecules. *Chem. Phys. Lett.* 313:701–706.
4. Kitaura, K., Morokuma, K. 1976. A new energy decomposition scheme for molecular interactions within the Hartree–Fock approximations. *Int. J. Quant. Chem.* 10:325–340.
5. Ohnishi, S.I., Orimoto, Y., Gu, F.L., Imamura, A., Aoki, Y. 2007. Nonlinear optical properties of polydiacetylene with donor-acceptor substitution block. *J. Chem. Phys.* 127:084702.

6. Orimoto, Y., Gu, F.L., Imamura, A., Aoki, Y. 2007. Efficient and accurate calculations on the electronic structure of B-type poly(dG)·poly(dC) DNA by elongation method: first step toward the understanding of the biological properties of aperiodic DNA. *J. Chem. Phys.* 126:215104.
7. Zhang, R.J., Tian, W.Q., Gu, F.L., Aoki, Y. 2007. Theoretical studies on the adsorption of Si and C chains onto unfaulted and faulted Si(111) surfaces. *J. Phys. Chem. C* 111:6350–6356.
8. Friedrich, J., Hanrath, M., Dolg, M. 2007. Energy screening for the incremental scheme: application to intermolecular interactions. *J. Phys. Chem. A* 111:9830–9837.
9. Friedrich, J., Hanrath, M., Dolg, M. 2008. Using symmetry in the framework of the incremental scheme: molecular applications. *Chem. Phys.* 346:266–274.
10. Jacob, C.R., Visscher, L. 2008. A subsystem density-functional theory approach for the quantum chemical treatment of proteins. *J. Chem. Phys.* 128:155102.
11. Li, W., Li, S.H., Jiang, Y.S. 2007. Generalized energy-based fragmentation approach for computing the ground-state energies and properties of large molecules. *J. Phys. Chem. A* 111:2193–2199.
12. Gadre, S.R., Ganesh, V. 2006. Molecular tailoring approach: towards PC-based *ab initi*o treatment of large molecules. *J. Theor. Comput. Chem.* 5:835–855.
13. Kobayashi, M., Imamura, Y., Nakai, H. 2007. Alternative linear-scaling methodology for the second-order Møller–Plesset perturbation calculation based on the divide-and-conquer method. *J. Chem. Phys.* 127:074103.
14. Collins, M.A. 2007. Molecular potential energy surfaces constructed from interpolation of systematic fragment surfaces. *J. Chem. Phys.* 127:024104.
15. Kamiya, M., Hirata, S., Valiev, M. 2008. Fast electron correlation methods for molecular clusters without basis set superposition errors. *J. Chem. Phys.* 128:074103.
16. Dahlke, E.E., Truhlar, D.G. 2008. Electrostatically embedded many-body expansion for simulations. *J. Chem. Theor. Comput.* 4:1–6.
17. Dahlke, E.E., Leverentz, H.R., Truhlar, D.G. 2008. Evaluation of the electrostatically embedded many-body expansion and the electrostatically embedded many-body expansion of the correlation energy by application to low-lying water hexamers. *J. Chem. Theor. Comput.* 4:33–41.
18. Huang, L., Massa, L., Karle, J. 2005. Kernel energy method illustrated with peptides. *Int. J. Quant. Chem.* 103:808–817.
19. Huang, L., Massa, L., Karle, J. 2005. Kernel energy method: application to insulin. *Proc. Natl. Acad. Sci. USA* 102:12690–12693.
20. Huang, L., Massa, L., Karle, J. 2005. Kernel energy method: application to DNA. *Biochem.* 44:16747–16752.
21. Huang, L., Massa, L., Karle, J. 2006. Kernel energy method: basis functions and quantum methods. *Int. J. Quant. Chem.* 106:447–457.
22. Huang, L., Massa, L., Karle, J. 2006. The kernel energy method: application to a tRNA. *Proc. Natl. Acad. Sci. USA* 103:1233–1237.
23. Huang, L., Massa, L., Karle, J. 2007. Drug target interaction energies by the kernel energy method in aminoglycoside drugs and ribosomal A site RNA targets. *Proc. Natl. Acad. Sci. USA* 104:4261–4266.
24. Huang, L., Massa, L., Karle, J. 2007. Kernel energy method: the interaction energy of the collagen triple helix. *J. Chem. Theory Comput.* 3:1337–1341.
25. Huang, L., Massa, L., Karle, J. 2008. The kernel energy method of quantum mechanical approximation carried to fourth-order terms. *Proc. Natl. Acad. Sci. USA*, 105:1849–1854.
26. Fedorov, D.G., Ishimura, K., Ishida, T., Kitaura, K., Pulay, P., Nagase, S. 2007. Accuracy of the three-body fragment molecular orbital method applied to Møller–Plesset perturbation theory. *J. Comput. Chem.* 28:1476–1484.

27. Chiba, M., Fedorov, D.G., Kitaura, K. 2007. Time-dependent density functional theory with the multilayer fragment molecular orbital method. *Chem. Phys. Lett.* 444:346–350.
28. Gao, Q., Yokojima, S., Kohno, T. et al. 2007. *Ab initio* NMR chemical shift calculations on proteins using fragment molecular orbitals with electrostatic environment. *Chem. Phys. Lett.* 445:331–339.
29. Chiba, M., Fedorov, D.G., Kitaura, K. 2007. Time-dependent density functional theory based upon the fragment molecular orbital method. *J. Chem. Phys.* 127:104108.
30. Ishikawa, T., Mochizuki, Y., Amari, S., et al. 2007. Fragment interaction analysis based on local MP2. *Theor. Chem. Acc.* 118:937–945.
31. Okiyama, Y., Watanabe, H., Fukuzawa, K. et al. 2007. Application of the fragment molecular orbital method for determination of atomic charges on polypeptides. *Chem. Phys. Lett.* 449:329–335.
32. Maezono, R., Watanabe, H., Tanaka, S., Towler, M.D., Needs, R.J. 2007. Fragmentation method combined with quantum Monte Carlo calculations. *J. Phys. Soc. Jap.* 76:064301.
33. Mochizuki, Y., Yamashita, K., Murase, T. et al. 2008. Large scale FMO-MP2 calculations on a massively parallel-vector computer. *Chem. Phys. Lett.* 457:396–403.
34. Chiba, M., Fedorov, D.G., Kitaura, K. 2008. Polarizable continuum model with the fragment molecular orbital based time-dependent density functional theory. *J. Comput. Chem.* 29:2667–2676.
35. Fedorov, D.G., Jensen, J.H., Deka, R.C., Kitaura, K. 2008. Covalent bond fragmentation suitable to describe solids in the fragment molecular orbital method. *J. Phys. Chem. A* 112:11808–11816.
36. Fukunaga, H., Fedorov, D.G., Chiba, M., Nii, K., Kitaura, K. 2008. Theoretical analysis of the intermolecular interaction effects on the excitation energy of organic pigments: solid state quinacridone. *J. Phys. Chem. A* 112:10887–10894.
37. Watanabe, T., Inadomi, Y., Fukuzawa, K., et al. 2007. DNA and estrogen receptor interaction revealed by fragment molecular orbital calculations. *J. Phys. Chem. B* 111:9621–9627.
38. Ozawa, T., Okazaki, K. 2008. CH/π hydrogen bonds determine the selectivity of the Src homology 2 domain to tyrosine phosphotyrosyl peptides: an *ab initio* fragment molecular orbital study. *J. Comput. Chem.* 29:2656–2666.
39. Watanabe, H., Enomoto, T., Tanaka, S. 2007. *Ab initio* study of molecular interactions in higher plant and *Galdieria partita* Rubiscos with the fragment molecular orbital method. *Biochem. Biophys. Res. Comm.* 361:367–372.
40. Ito, M., Fukuzawa, K., Mochizuki, Y., Nakano, T., Tanaka, S. 2008. *Ab initio* fragment molecular orbital study of molecular interactions between liganded retinoid X receptor and its coactivator; part II: influence of mutations in transcriptional activation function 2 activating domain core on the molecular interactions. *J. Phys. Chem. A* 112:1986–1998.
41. Iwata, T., Fukuzawa, K., Nakajima, K., et al. 2008. Theoretical analysis of binding specificity of influenza viral hemagglutinin to avian and human receptors based on the fragment molecular orbital method. *Comput. Biol. Chem.* 32:198–211.
42. Harada, T., Yamagishi, K., Nakano, T., Kitaura, K., Tokiwa, H. 2008. *Ab initio* fragment molecular orbital study of ligand binding to human progesterone receptor ligand-binding domain. *Naunyn-Schmiedeberg's Arch. Pharmac.* 377:607–615.
43. Sawada, T., Hashimoto, T., Tokiwa, H., et al. 2008. *Ab initio* fragment molecular orbital studies of influenza virus hemagglutinin-sialosaccharide complexes toward chemical clarification about the virus host range determination. *Glycoconj. J.* 25:805–815.

44. Ito, M., Fukuzawa, K., Ishikawa, T., Mochizuki, Y., Nakano, T., Tanaka, S. 2008. *Ab initi*o fragment molecular orbital study of molecular interactions in liganded retinoid X receptor: specification of residues associated with ligand inducible information transmission. *J. Phys. Chem. B* 112:12081–12094.
45. Kurisaki, I., Fukuzawa, K., Komeiji, Y. et al. 2007. Visualization analysis of interfragment interaction energies of CRP-cAMP-DNA complex based on the fragment molecular orbital method. *Biophys. Chem.* 130:1–9.
46. Fischer, B., Fukuzawa, K., Wenzel, W. 2008. Receptor-specific scoring functions derived from quantum chemical models improve affinity estimates for in-silico drug discovery. *Proteins: Struct., Funct., Bioinf.* 70:1264–1273.
47. Sato, M., Yamataka, H., Komeiji, Y., Mochizuki, Y., Ishikawa, T., Nakano, T. 2008. How does an S_N2 reaction take place in solution? Full *ab initio* MD simulations for the hydrolysis of the methyl diazonium ion. *J. Am. Chem. Soc.* 130:2396–2397.
48. Komeiji, Y., Ishikawa, T., Mochizuki, Y., Yamataka, H., Nakano, T. 2009. Fragment molecular orbital method-based molecular dynamics (FMO-MD) as a simulator for chemical reactions in explicit solvation. *J. Comput. Chem.* 30:40–50.
49. Komeiji, Y., Mochizuki, Y., Nakano, T., Fedorov, D.G. 2009. Fragment molecular orbital-based molecular dynamics (FMO-MD), a quantum simulation tool for large molecular systems. *J. Mol. Str. (THEOCHEM)*, 898:2–7.
50. Yoshida, T., Yamagishi, K., Chuman, H. 2008. QSAR study of cyclic urea type HIV-1 PR inhibitors using *ab initio* MO calculation of their complex structures with HIV-1. *QSAR Comb. Sci.* 27:694–703.
51. Zhang, Q., Yang, J., Liang, K. et al. 2008. Binding interaction analysis of the active site and its inhibitors for neuraminidase (N1 subtype) of human influenza virus by the integration of molecular docking, FMO calculation and 3D-QSAR CoMFA modeling. *J. Chem. Inf. Model.*, in press.
52. Ikegami, T., Ishida, T., Fedorov, D.G. et al. 2005. Full electron calculation beyond 20,000 atoms: ground electronic state of photosynthetic proteins. *Proc. of Supercomputing 2005*, Seattle, WA: IEEE Computer Society.
53. Schmidt, M.W., Baldridge, K.K., Boatz, J.A. et al. 1993. General atomic and molecular structure system. *J. Comput. Chem.* 14:1347–1363. GAMESS can be downloaded from: www.msg.chem.iastate.edu/gamess/gamess.html.
54. Fedorov, D.G., Kitaura, K. 2004. The importance of three-body terms in the fragment molecular orbital method. *J. Chem. Phys.* 120:6832–6840.
55. Nakano, T., Mochizuki, Y., Fukuzawa, K., Amari, S., Tanaka, S. 2006. Developments and applications of ABINIT-MP software based on the fragment molecular orbital method. In *Modern methods for theoretical physical chemistry of biopolymers,* eds. E.B. Starikov, J.P. Lewis, and S. Tanaka, 39–52. Amsterdam: Elsevier.
56. Sekino, H., Matsumura, N., Sengoku, Y. 2007. Evaluation of NMR chemical shift by fragment molecular orbital method. *Comput. Lett.* 3:423–430.
57. GridFMO home page: http://goc.pragma-grid.net/applications/FMO/FMO.html.
58. OpenFMO home page: www.openfmo.org/OpenFMO/.
59. FMO home page: http://staff.aist.go.jp/d.g.fedorov/.
60. Fedorov, D.G., Ishida, T., Uebayasi, M., Kitaura, K. 2007. The fragment molecular orbital method for geometry optimizations of polypeptides and proteins, *J. Phys. Chem. A* 111:2722–2732.
61. Suenaga, M. 2008. Development of GUI for GAMESS/FMO calculation. *J. Comput. Chem. Jpn.* 7:33–54 (in Japanese). Facio 11.8.1, www1.bbiq.jp/zzzfelis/Facio.html.
62. Mochizuki, Y., Fukuzawa, K., Kato, A., Tanaka, S., Kitaura, K., Nakano, T. 2005. A configuration analysis for fragment interaction. *Chem. Phys. Lett.* 410:247–253.

63. Fedorov, D.G., Kitaura, K. 2007. Pair interaction energy decomposition analysis. *J. Comput. Chem.* 28:222–237.
64. Inadomi, Y., Nakano, T., Kitaura, K., Nagashima, U. 2002. Definition of molecular orbitals in fragment molecular orbital method. *Chem. Phys. Lett.* 364:139–143.
65. Yasuda, K., Yamaki, D. 2006. The extension of the fragment molecular orbital method with the many-particle Green's function. *J. Chem. Phys.* 12:154101.
66. Nakano, T., Kaminuma, T., Sato, T., et al. 2002. Fragment molecular orbital method: use of approximate electrostatic potential. *Chem. Phys. Lett.* 351:475–480.
67. Nakanishi, I., Fedorov, D.G., Kitaura, K. 2007. Molecular recognition mechanism of FK506 binding protein: an all-electron fragment molecular orbital study. *Proteins: Struct., Funct., Bioinf.* 68:145–158.
68. Weis, A., Katebzadeh, K., Söderhjelm, P., Nilsson, I., Ryde, U. 2006. Ligand affinities predicted with the MM/PBSA method: dependence on the simulation method and the force field. *J. Med. Chem.* 49:6596–6606.
69. Tomasi, J., Mennucci, B., Cammi, R. 2005. Quantum mechanical continuum solvation models. *Chem. Rev.* (Washington, DC) 105:2999–3093.
70. Kelly, C.P., Cramer, C.J., Truhlar, D.G. 2006. Adding explicit solvent molecules to continuum solvent calculations for the calculation of aqueous acid dissociation constants. *J. Phys. Chem. A* 110:2493–2499.
71. Barone, V., Cossi, M., Tomasi, J. 1997. A new definition of cavities for the computation of solvation free energies by the polarizable continuum model. *J. Chem. Phys.* 107:3210–3221.
72. Fedorov, D.G., Kitaura, K., Li, H., Jensen, J.H., Gordon, M.S. 2006. The polarizable continuum model (PCM) interfaced with the fragment molecular orbital method (FMO). *J. Comput. Chem.* 27:976–985.
73. Schwabe, T., Grimme, S. 2007. Double-hybrid density functionals with long-range dispersion corrections: higher accuracy and extended applicability, *Phys. Chem. Chem. Phys.* 9 3397–3406.
74. Piecuch, P., Włoch, M., Gour, J.R., Kinal, A. 2006. Single-reference, size-extensive, non-iterative coupled-cluster approaches to bond breaking and biradicals. *Chem. Phys. Lett.* 418:467–474.
75. Fedorov, D.G., Ishida, T., Kitaura, K. 2005. Multilayer formulation of the fragment molecular orbital method (FMO). *J. Phys. Chem. A* 109:2638–2646.
76. Dem'yanov, P., Polestshuk, P., Gschwind, R. 2008. Hydrogen-bonding interactions in monomeric dimethylcuprates. A theoretical study. *J. Mol. Str. (THEOCHEM)* 861:85–96.
77. Fedorov, D.G., Olson, R.M., Kitaura, K., Gordon, M.S., Koseki, S. 2004. A new hierarchical parallelization scheme: generalized distributed data interface (GDDI), and an application to the fragment molecular orbital method (FMO). *J. Comput. Chem.* 25:872–880.

3 Developments of FMO Methodology and Graphical User Interface in ABINIT-MP

Tatsuya Nakano, Yuji Mochizuki, Akifumi Kato, Kaori Fukuzawa, Takeshi Ishikawa, Shinji Amari, Ikuo Kurisaki, and Shigenori Tanaka

CONTENTS

3.1 INTRODUCTION

The application of quantum chemical methods to large molecules such as proteins and DNAs remains a great challenge in computational chemistry. The calculations of electronic structures for biomolecules are usually difficult due to their huge size, and some models are needed to realize quantum mechanical calculations on such systems.

Even though an idea of the fragmentation of the whole system is straightforward and appealing for the reduction in computational time, the difficulties always take place concerning inaccuracies due to the artificial breakage of chemical bonds and the reconstruction of observable properties. Among a number of these approaches, the fragment molecular orbital (FMO) method proposed by Kitaura et al.[1–5] provides a very promising way to treat huge biopolymers with chemical accuracies in energy and other properties.

In the FMO method,[1–5] a molecule or a molecular cluster is divided into N_f fragments, and the molecular orbital (MO) calculations for the fragments (monomers), the fragment pairs (dimers), and trimers are performed to obtain the total energy and other molecular properties. The total energy of a system can then be calculated as

$$\begin{aligned} E^{\mathrm{FMO3}} &= \sum_I E_I + \sum_{I>J} (E_{IJ} - E_I - E_J) + \sum_{I>J>K} [(E_{IJK} - E_I - E_J - E_K) - (E_{IJ} - E_I - E_J) \\ &\quad - (E_{IK} - E_I - E_K) - (E_{JK} - E_J - E_K)] \\ &= \sum_{I>J>K} E_{IJK} - (N_f - 3) \sum_{I>J} E_{IJ} + \frac{(N_f - 2)(N_f - 3)}{2} \sum_I E_I \end{aligned} \tag{3.1}$$

where E_I, E_{IJ}, and E_{IJK} represent the energies of fragment monomer, dimer, and trimer, respectively, calculated under the electrostatic potentials from other surrounding fragments (environmental electrostatic potentials [ESPs]). The FMO expansion including trimers (Equation 3.1) is denoted by FMO3 (the three-body expansion), and Equation 3.1 without the three-body term defines FMO2 (the two-body expansion). The FMO method avoids the MO calculation of the whole system of a large molecule and reduces the computational time remarkably. Another advantage of this method is its ease in utilizing parallel processing, because the dimers and trimers can be treated independently.

A number of calculations have already been performed[1–5] concerning the accuracy and efficiency of the FMO method, providing extremely promising results for the applications to biomolecules. Actual implementations of the FMO method have also been carried out in such packages as GAMESS,[3,5] NWChem,[6–9] and ABINIT-MP.[4] In this chapter, we illustrate recent developments of the ABINIT-MP program and analysis and visualization tools for the results of FMO calculations, such as analysis of interfragment interaction energies (IFIEs) that are also called pair interaction energies (PIEs) in other chapters,[3,5,8,10–14] configuration analysis for fragment interactions (CAFI),[13,15] fragment interaction based on local MP2 (FILM),[16,17] IFIE map,[18] and visualized cluster analysis of protein–ligand interactions (VISCANA).[14] The FMO calculations in the restricted Hartree–Fock (RHF), the second-order Møller–Plesset perturbation (MP2),[19–21] and the configuration interaction singles with perturbative doubles correction (CIS(D)) are now available in an efficiently

parallelized fashion.[22–26] The details of excited state calculations with the CIS(D) method are documented by Mochizuki et al. in Chapter 4.

3.2 METHODOLOGY AND IMPLEMENTATION

3.2.1 ABINIT-MP Program: Outline of Implementation

In the FMO calculations at the two-body treatment (FMO2),[1–5] the total energy of the target is evaluated by the contributions of monomers and dimers:

$$E^{\mathrm{FMO2}} = \sum_I E_I + \sum_{I>J} (E_{IJ} - E_I - E_J) = \sum_{I>J} E_{IJ} - (N_f - 2) \sum_I E_I \tag{3.2}$$

The total electron density matrix of system, $\mathbf{D}^{\mathrm{FMO2}}$, can be constructed in a similar way:

$$\mathbf{D}^{\mathrm{FMO2}} = \sum_I {}^{\oplus}\mathbf{D}^I + \sum_{I>J} {}^{\oplus}\Delta\mathbf{D}^{IJ} = \sum_{I>J} {}^{\oplus}\mathbf{D}^{IJ} - (N_f - 2) \sum_I {}^{\oplus}\mathbf{D}^I \tag{3.3}$$

where $\mathbf{D}^I$ and $\mathbf{D}^{IJ}$ are monomer and dimer density matrices, respectively. The difference density matrix, $\Delta\mathbf{D}^{IJ}$, is defined as follows:

$$\Delta\mathbf{D}^{IJ} = \mathbf{D}^{IJ} - \mathbf{D}^I \oplus \mathbf{D}^J \tag{3.4}$$

$$\begin{aligned} D^I_{pq} \oplus D^J_{pq} &= D^I_{pq} + D^J_{pq} && \text{for } p,q \in I,J, \\ &= D^I_{pq} && \text{for } p,q \in I \text{ and } p,q \notin J, \\ &= D^J_{pq} && \text{for } p,q \in J \text{ and } p,q \notin I, \\ &= 0 && \text{for all other cases.} \end{aligned} \tag{3.5}$$

In the FMO3 method, total electron density matrix, $\mathbf{D}^{\mathrm{FMO3}}$, is also calculated by the following equation:

$$\begin{aligned} \mathbf{D}^{\mathrm{FMO3}} &= \sum_I {}^{\oplus}\mathbf{D}^I + \sum_{I>J} {}^{\oplus}\Delta\mathbf{D}^{IJ} + \sum_{I>J>K} {}^{\oplus}(\Delta\mathbf{D}^{IJK} - \Delta\mathbf{D}^{IJ} \oplus \Delta\mathbf{D}^{IK} \oplus \Delta\mathbf{D}^{JK}) \\ &= \sum_{I>J>K} {}^{\oplus}\mathbf{D}^{IJK} - (N_f - 3) \sum_{I>J} {}^{\oplus}\mathbf{D}^{IJ} + \frac{(N_f - 2)(N_f - 3)}{2} \sum_I {}^{\oplus}\mathbf{D}^I \end{aligned} \tag{3.6}$$

where

$$\Delta\mathbf{D}^{IJK} = \mathbf{D}^{IJK} - \mathbf{D}^I \oplus \mathbf{D}^J \oplus \mathbf{D}^K \tag{3.7}$$

In Equation 3.3 and Equation 3.6, the total density matrix expressions without difference density matrices are useful for implementation.

The total energy of molecules, Equation 3.2, is rewritten as follows:

$$E^{\mathrm{FMO2}} = \sum_{I} E'_I + \sum_{I>J} \Delta\tilde{E}_{IJ} \tag{3.8}$$

$$E'_I = E_I - \mathrm{Tr}(\mathbf{D}^I \mathbf{V}^I) \tag{3.9}$$

where E'_I is the monomer energy without environmental electrostatic energy, and $\mathbf{V}^I$ is the environmental electrostatic potential for monomer *I*. We define IFIE in RHF level of theory as follows[3,5,8,10–14]:

$$\Delta\tilde{E}_{IJ} = E'_{IJ} - E'_I - E'_J + \mathrm{Tr}(\Delta\mathbf{D}^{IJ}\mathbf{V}^{IJ}) \tag{3.10}$$

$$E'_{IJ} = E_{IJ} - \mathrm{Tr}(\mathbf{D}^{IJ}\mathbf{V}^{IJ}) \tag{3.11}$$

where E'_{IJ} is the dimer energy without environmental electrostatic energy, and $\mathbf{V}^{IJ}$ is the environmental electrostatic potential for dimer *IJ*. For proteins, the monomer corresponds typically to one or two amino acid residues, where the former setting is convenient for the analysis purpose with IFIE.

One of the most time-consuming parts in the FMO method is the calculations of ESPs,[12] $\mathbf{V}^x$:

$$V^x_{pq} = \sum_{K\neq x} \left(u^K_{pq} + v^K_{pq} \right) \tag{3.12}$$

$$u^K_{pq} = \sum_{A\in K} \langle p|(-Z_A/|\mathbf{r}-\mathbf{A}|)|q\rangle \tag{3.13}$$

$$v^K_{pq} = \sum_{rs\in K} D^K_{rs}(pq,rs) \tag{3.14}$$

where x means monomer *I*, dimer *IJ*, or trimer *IJK*, and $\mathbf{u}^K$ and $\mathbf{v}^K$ are the nuclear attraction and two-electron term of ESPs, respectively. It is accelerated by employing the Mulliken approximation for two-electron integrals for intermediately separated monomers (esp-aoc):

$$v^K_{pq} \cong \sum_{r\in K} (\mathbf{D}^K\mathbf{S}^K)_{rr}(pq,rr) \quad \text{for } R_{\min}(x,K) \geq L_{\mathrm{aoc}} \tag{3.15}$$

and a fractional point charge approximation using Mulliken population, Q, for largely separated monomers (esp-ptc):

$$v^K_{pq} \cong \sum_{A\in K} \langle p|(Q_A/|\mathbf{r}-\mathbf{A}|)|q\rangle \quad \text{for } R_{\min}(x,K) \geq L_{\mathrm{ptc}} \tag{3.16}$$

where $R_{\min}(x,K)$ is the shortest interatomic distance between the atoms in the fragments x and the monomer K. L_{aoc}, and L_{ptc} are threshold distances for the esp-aoc and the esp-ptc approximations, respectively. The threshold values are given in units of the van der Waals radius: $L = x$ meant that the threshold value was x times the sum of the van der Waals radii of the shortest contact atoms. In the ABINIT-MP program, default values of L_{aoc} and L_{ptc} are 0.0 and 2.0, respectively.

Another time-consuming part is calculation of the dimer energy without environmental electrostatic energy E'_{IJ}. The electrostatic interaction approximation for the dimer composed of largely separated monomers (dimer-es) is

$$E'_{IJ} \cong E'_I + E'_J + \mathrm{Tr}(\mathbf{D}^I\mathbf{u}^J) + \mathrm{Tr}(\mathbf{D}^J\mathbf{u}^I) + \sum_{pq\in\boldsymbol{I}}\sum_{rs\in J} D^I_{pq}D^J_{rs}(pq,rs) + \sum_{A\in I}\sum_{B\in J}\frac{Z_A Z_B}{|\mathbf{A}-\mathbf{B}|} \tag{3.17}$$

which reduces computational cost from $O(N_f^3)$ to $O(N_f^2)$.

Similar to the FMO2 method, the FMO3 total energy,[27–30] Equation 3.1, is rewritten as follows:

$$E^{\text{FMO3}} = \sum_I E'_I + \sum_{I>J}\Delta\tilde{E}_{IJ} + \sum_{I>J>K}\Delta\tilde{E}_{IJK} \tag{3.18}$$

where $\Delta\tilde{E}_{IJK}$ is a three-body interaction energy:

$$\Delta\tilde{E}_{IJK} = E'_{IJK} - E'_I - E'_J - E'_K + \mathrm{Tr}(\Delta\mathbf{D}^{IJK}\mathbf{V}^{IJK}) - \Delta\tilde{E}_{IJ} - \Delta\tilde{E}_{IK} - \Delta\tilde{E}_{JK} \tag{3.19}$$

Using the dimer-es approximation, if one or two monomers are largely separated from another, then three-body term, Equation 3.19, is equal to zero. Therefore, we could perform FMO3 calculations with only $O(N_f^2)$ operations. As shown in Reference [29], the consistency of ESPs between dimer and trimer is crucial for accurate FMO3 calculations. Thus, we use only esp-aoc approximation in FMO3 calculations at the present implementation of the ABINIT-MP program. The errors arising from the FMO2-RHF calculations are suppressed within a few kcal/mol, even for the medium-size proteins (see Chapter 2).

The index list of fragment is distributed over the groups of processors as the upper-level control of parallelization. Within each group of processors for a given x, the integral-direct parallelization is driven at the lower level, leading to the dual layer parallelism of the FMO. The actual controls of parallelization are done by using the generalized distributed data interface (GDDI) in GAMESS[31] and message passing interface (MPI)[32] in ABINIT-MP, respectively. The RHF calculations for monomers are iterated until the self-consistent charge (SCC) condition is satisfied under the ESP, and this is essential in taking a polarization effect in the target system into account. In contrast, the dimer or trimer RHF procedure is performed once for each dimer or trimer composed of two or three neighboring monomers, by which some electron delocalizations among fragments are introduced. Once the RHF stage is finished, the MP2 calculations for the respective monomers, dimers, or trimers

start immediately.[19–21] The MP2 correlation energies correct the IFIE values, and the contribution from dispersion, $\Delta E_{IJ}^{\mathrm{MP2}}$, could then be incorporated[33]:

$$\Delta E_{IJ}^{\mathrm{MP2}} = E_{IJ}^{\mathrm{MP2}} - E_I^{\mathrm{MP2}} - E_J^{\mathrm{MP2}} \tag{3.20}$$

$$\Delta \tilde{E}_{IJ}^{\mathrm{HF+MP2}} = \Delta \tilde{E}_{IJ}^{\mathrm{HF}} + \Delta E_{IJ}^{\mathrm{MP2}} \tag{3.21}$$

where E_I^{MP2} and E_{IJ}^{MP2} are the monomer and dimer correlation energies with MP2 method, respectively. In the FMO3 method, MP2 correlation correction of three-body interaction energy is expressed as follows[30]:

$$\Delta E_{IJK}^{\mathrm{MP2}} = E_{IJK}^{\mathrm{MP2}} - E_I^{\mathrm{MP2}} - E_J^{\mathrm{MP2}} - E_K^{\mathrm{MP2}} \tag{3.22}$$

$$\Delta \tilde{E}_{IJK}^{\mathrm{HF+MP2}} = \Delta \tilde{E}_{IJK}^{\mathrm{HF}} + \Delta E_{IJK}^{\mathrm{MP2}} - \Delta E_{IJ}^{\mathrm{MP2}} - \Delta E_{IK}^{\mathrm{MP2}} - \Delta E_{JK}^{\mathrm{MP2}} \tag{3.23}$$

Using the dimer-es approximation, computational costs of post-HF calculations are reduced to $O(N_f)$. To obtain qualitatively correct descriptions of proteins, the inclusion of MP2 corrections is essential even at a semiquantitative level with double-zeta basis sets.[34,35] This situation is illustrated by Fukuzawa et al. (Chapter 7) and Ozawa et al. (Chapter 10).

The output files of the calculations can also be employed for the analysis of the molecular characteristics in terms of the IFIEs and their configurations. All the pre- and postprocessors for the ABINIT-MP program would be integrated on a specialized graphical user interface called BioStation Viewer.[36]

3.2.2 MP2: Integral-Direct Parallelism

Once the converged RHF wavefunction is obtained for the fragment monomer, dimer, or trimer, x, the MP2 calculation starts immediately. Note that the MP2 calculation is skipped for a dimer or trimer that is applied the dimer-es approximation. The MP2 correlation energy is calculated by the contraction of the two-electron MO integral list of (ia, jb) (Mulliken notation with i, j for doubly occupied MOs and a, b for virtual MOs) with the orbital energies of ε_i, ε_j, ε_a, and ε_b[34,37]:

$$E^{\mathrm{MP2}} = \sum_{ijab} \frac{(ia, jb)[2(ia, jb) - (ib, ja)]}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{3.24}$$

The (ia, jb) list is generated by the series of quarter transformations from the two-electron AO integral list of (pq, rs):

$$(ia, jb) = \sum_{pqrs} C_{pi} C_{qa} C_{rj} C_{sb} (pq, rs) \tag{3.25}$$

where **C** is the MO coefficients matrix of the RHF solution. The (ia, jb) list is actually formed by the series of quarter transformations[19,20,38,39]:

$$(iq, rs) = \sum_p C_{pi}(pq, rs) \tag{3.26}$$

$$(ia, rs) = \sum_q C_{qa}(iq, rs) \tag{3.27}$$

$$(ia, js) = \sum_r C_{rj}(ia, rs) \tag{3.28}$$

$$(ia, jb) = \sum_s C_{sb}(ia, js) \tag{3.29}$$

and these processings substantially constitute the body of MP2 calculations. Mochizuki et al. has developed a new parallelized integral-direct MP2 algorithm[19,20] that has the following features:

1. The entire MP2 computation is divided into batches according to the MO index *i*. In other words, *i*-batch[38] is used to adjust the memory requirement.
2. The parallelization is primarily driven by the atomic orbital (AO) index *s* associated with the MO index *b*, under a replicated-data scheme.
3. No communication across the worker processes is needed. Alternatively, the single all-reduce operation is finally necessary.
4. Multiple screening is incorporated to reduce the transformation costs effectively.
5. The DAXPY or DDOT routines of vector–vector style are used for each quarter step.

Parallelized MP2 calculations are independently carried out for the respective fragments. The above fourth and fifth features have been effective for the production runs of FMO-MP2 on the PC clusters.[20,40,41]

The introduction of MP2 correlation should change the electron density from that of RHF. If the MP2 density is available for proteins and molecular clusters, such information would be valuable. We thus realized the evaluation of the MP2 density matrix in the ABINIT-MP program.[19] In the present implementation, the response elements to describe the occupied-virtual orbital rotations were approximated by omitting orbital Hessian.[19] Our final expressions were equivalent to those of Jensen et al.[42] The influence of the approximation of response elements was already checked to be small as long as the unperturbed HF wavefunction was good enough.[19]

3.2.3 Vectorization on the Earth Simulator

As pointed out by Almlöf et al.,[39,43] the generation of two-electron integrals with Gaussian AOs should dominate the total cost of integral direct algorithm for electronic structure calculations. We thus first address the vectorization of integral generation. The recurrence algorithm developed by Obara and Saika[44] has been employed to generate all the integrals in the ABINIT-MP program. In Obara's method,[44] the computation of two-electron integrals over s-type functions (symbolized as (SS,SS) in the Mulliken notation[34]) or the evaluation of $F_m(T)$.

$$F_m(T) = \int_0^1 t^{2m} \exp(-Tt^2)\, dt$$
$$m = 0, 1, 2, \ldots \text{and } T \geq 0 \tag{3.30}$$

is a rate-determining step because of the fourth-order dependence of contracted shells. In the ABINIT-MP program, the functions $F_m(T)$ are evaluated using two formulas depending on the value *T*. For values from zero to $2m+36$,[45] we employ the four-term Taylor expansion:

$$F_m(T) \cong \sum_{k=0}^{3} F_{m+k}(T^*)(T^* - T)^k / k! \tag{3.31}$$

where $F_{m+k}(T^*)$ has been evaluated by the power series[46]:

$$F_m(T^*) = \exp(-T^*) \sum_{i=0}^{\infty} \frac{(2T^*)^i}{(2m+1)(2m+3)\cdots(2m+2i+1)} \tag{3.32}$$

for T^* at intervals of $1/2^9$ and tabulated. Another formula, used for *T* greater than $2m+36$, is an asymptotic formula of $F_0(T)$ and upward recursion[45]:

$$F_0(T) \cong \frac{1}{2}\pi^{1/2}/T^{1/2} \tag{3.33}$$

$$F_{m+1}(T) \cong (2m+1)F_m(T)/(2T) \tag{3.34}$$

The proposed algorithm has one IF-branching and one square-root operation and could be vectorized by the compiler efficiently. A loop expansion technique has been straightforwardly introduced to vectorize a series of integral generations with higher angular-momenta through the recurrence relations.[44]

The computation of Fock matrix is another target to be vectorized in the direct SCF procedure[43] of FMO-RHF calculations. With the current density matrix, **D**,

a two-electron integral formally associates with the Fock matrix,[47] **F**, in Coulomb contributions:

$$F_{pq} \leftarrow F_{pq} + 2D_{rs}(pq,rs) \tag{3.35}$$

$$F_{rs} \leftarrow F_{rs} + 2D_{pq}(pq,rs) \tag{3.36}$$

and also in exchange contributions:

$$F_{pr} \leftarrow F_{pr} - \frac{1}{2}D_{qs}(pq,rs) \tag{3.37}$$

$$F_{ps} \leftarrow F_{ps} - \frac{1}{2}D_{qr}(pq,rs) \tag{3.38}$$

$$F_{qr} \leftarrow F_{qr} - \frac{1}{2}D_{ps}(pq,rs) \tag{3.39}$$

$$F_{qs} \leftarrow F_{qs} - \frac{1}{2}D_{pr}(pq,rs) \tag{3.40}$$

A simple vectorization with respect to a given block of index quartets and integral values can be in error because the conflicted addressing takes place for the exchange terms. Based on a buffered index algorithm originally proposed in the plasma simulations,[48] Mochizuki et al.[49] reported a vectorized Fock matrix construction appearing in relativistic RHF calculations. This algorithm is again adopted for the vectorized FMO-RHF calculations in the ABINIT-MP. Namely, the working Fock matrix, defined as **F** with a buffer index *w*, was used for the vectorization with indirect addressing, where its buffer length should be determined depending on the number of available registers of the vector processing unit (VPU). As the lower-level parallelization in FMO-RHF, the AO-index quartets are distributed over the VPUs associated with a certain fragment, and each VPU partially performs the Fock matrix construction according to a replicated data fashion.

By keeping the fundamental strategy of parallelization of References [19] and [20], we thus have created a new MP2 route in which all the quarter transformations are performed in a matrix–matrix style with DGEMM under the *ij*-batch control.[21] Figure 3.1 shows the schematic loop structure of the new MP2 processing. It is notable that the third and fourth steps by DGEMM are done in a manner of direct product written symbolically as

$$\mathbf{A}_{IJ} = \mathbf{B}_I \times \mathbf{C}_J \tag{3.41}$$

where *IJ* represents the combinations of $iajb$.

```
Loop over ij-batch ! size depending on available memory
  Loop over s ! to be parallelized
    Loop over r
      Preparing (pq, rs) ! for canonical pq-pair
      Forming (iq, rs) ! DGEMM, fixed rs, running over p
      Forming (ia, rs) ! DGEMM, fixed rs, running over q
      Forming (ia, js) ! DGEMM, fixed s, direct-product for fixed r
    End of loop over r
    Forming (ia, jb) ! DGEMM, direct-product for fixed s
  End of loop over s ! all reduce operation as barrier
  Calculate partial MP2 energy with respect to ij-batch
End of loop over ij-batch
```

FIGURE 3.1 Schematic loop structure of MP2 correlation energy calculation based on DGEMM.

As the platform used for benchmark FMO-MP2 calculations, the Earth Simulator system[50] has had the following massively parallel vector facilities:

- A theoretical peak speed per VPU is 8 GFLOPS.
- Eight VPUs form a single node equipping 16 GB shared memory.
- The nodes are equivalently connected through the highly efficient networks.
- Up to 512 nodes or 4096 VPUs are usable for a single job.

The BLAS libraries optimized for the Earth Simulator were statically linked. The buffer length of vectorized Fock matrix construction[48,49] was 256 in matching to the number of vector register. The 6-31G basis set[34,35] was used throughout, and the frozen-core restriction was imposed at the MP2 stage.[35] In the dual-layer parallelism in the ABINIT-MP with MPI, the number of processors for the integral-driven parallelization per fragment, denoted as NP, has been a pre-fixed parameter. The setting of NP could affect the performance depending on the target molecules.

By inserting extra trace routines into the MP2 kernel processes, breakdown analyses were first made for the FMO-MP2/6-31G calculations of Gly–Phe–Gly–Phe–Gly.[21] The single residue was treated as a monomer. The highest value is 5.72 GFLOPS (efficiency 71.5%) for the second and third steps with parallel execution on 8 VPUs (NP = 8). Meanwhile, the acceleration by parallelization with 8 VPUs is promising with more than a factor of 7 for each step.

$(Trp)_{127}$–His was employed as a larger model target,[21] as well as the realistic 17β-estradiol (EST)–estrogen receptor ligand-binding domain (ER) complex[21,41] consisting of 243 fragments: 241 residues of the receptor protein, EST as the ligand, and a water molecule in pharmacophore. The total numbers of atoms and AO functions in $(Trp)_{127}$–His are 3,068 and 18,659, respectively, whereas the corresponding numbers in the EST–ER complex are 3,946 and 21,583. Obviously, the ER protein contains a natural variety of amino acids from Gly to Trp.

TABLE 3.1
Timings and Performances of FMO-MP2/6-31G Calculations for $(Trp)_{127}$-His and 17β-Estradiol (EST)–Estrogen Receptor Ligand-Binding Domain (ER) Complex on the Earth Simulator

VPUs	N_P	Time (s)	Acceleration Factor	VPU (GFLOPS)	Total (TFLOPS)	Peak Ratio (%)
$(Trp)_{127}$His	16					
1024		1309.2	1.00	1.61	1.65	20.2
2048		672.3	1.95	1.57	3.22	19.7
EST–ER	32					
1024		1254.6	1.00	1.09	1.11	13.6
2048		793.5	1.58	0.86	1.76	10.7
3072		677.4	1.85	0.67	2.06	8.4
4096		570.4	2.20	0.60	2.45	7.5

The respective numbers of AOs in fragment residue thus have a range of 41 to 155, and the estradiol molecule has 228 AOs, meaning that there is a considerable variation in fragment sizes. Such differences are reflected in the timing data and performance presented in Table 3.1. As expected, the FMO-MP2 calculations provide good scalability and overall speed for the $(Trp)_{127}$–His. The observed gross speed of 1.65 TFLOPS for the FMO-MP2 job corresponds to an efficiency of 20.2% relative to a theoretical speed of 1,024 VPUs, where the averaged ratio of vector operation per VPU is 96.3%. These efficiencies would be reasonable, because the entire FMO-MP2 job involves not only initialization/finalization overheads but also nonvectorizable/parallelizable parts, besides the kernel processings. The acceleration of 1.95 (97.5%) against the VPU increase to 2048 is remarkable.

For the EST–ER complex, the degradation in efficiency is apparent relative to the $(Trp)_{127}$–His model, although it would be practically attractive that the FMO-MP2 job of such a realistic protein can be completed in about 20 minutes with 1024 VPUs using NP = 32. The reason for this situation could be attributed to the variation in fragment sizes. Actually, profiling analyses suggested that the computational time involving the estrogen as the largest fragment causes a substantial imbalance in granularity for the case of 2048 VPUs with NP = 16. Tests of the EST–ER complex imply the need of adaptive task distribution by varying NP like the GDDI control of GAMESS,[31] as a future issue to be implemented into the ABINIT-MP program.

In Chapter 7, it is shown that the FMO-MP2 based analyses on an influenza antigen–antibody system as the largest realistic target using the Earth Simulator.[21] In summary, the fundamental performance of MP2 kernel processing is acceptable, suggesting that proteins with a few hundred residues would be tractable under the intrinsic parallelism of FMO.[19–21,30,31,33]

3.3 ANALYSIS AND VISUALIZATION OF FMO CALCULATIONS

3.3.1 BioStation Viewer: A Program for Analyzing and Visualizing the Results of FMO Calculations

BioStation Viewer is designed to visualize the results obtained by the ABINIT-MP program by which the FMO calculations are performed for large molecules such as proteins. The BioStation Viewer provides a variety of methods in rendering molecules with colors; drawing isosurfaces of electron densities, electrostatic potentials, and molecular orbitals; and overlaying the group of molecules. The BioStation Viewer is written in Java and Java3D and, therefore, users can utilize the BioStation Viewer on any machine in which Java and Java3D are installed.

A main window of the BioStation Viewer is shown in Figure 3.2. The Hierarchy Window shows all objects in the molecular model on the left side of the window and a molecular structure in the three-dimensional (3D) window on the right. The BioStation Viewer can read a check point file that is the output file of the ABINIT-MP, PDB,[51] Tripos MOL2,[52] and Symyx CTfile (MDL MOL)[53] as the file format of molecule structure, and the ABINIT-MP grid file, and Gaussian cube file for visualizing the isosurface. In addition, it is possible to display a rhombohedron grid. Users can specify options on GUI, which is the edit function for the ABINIT-MP input file.

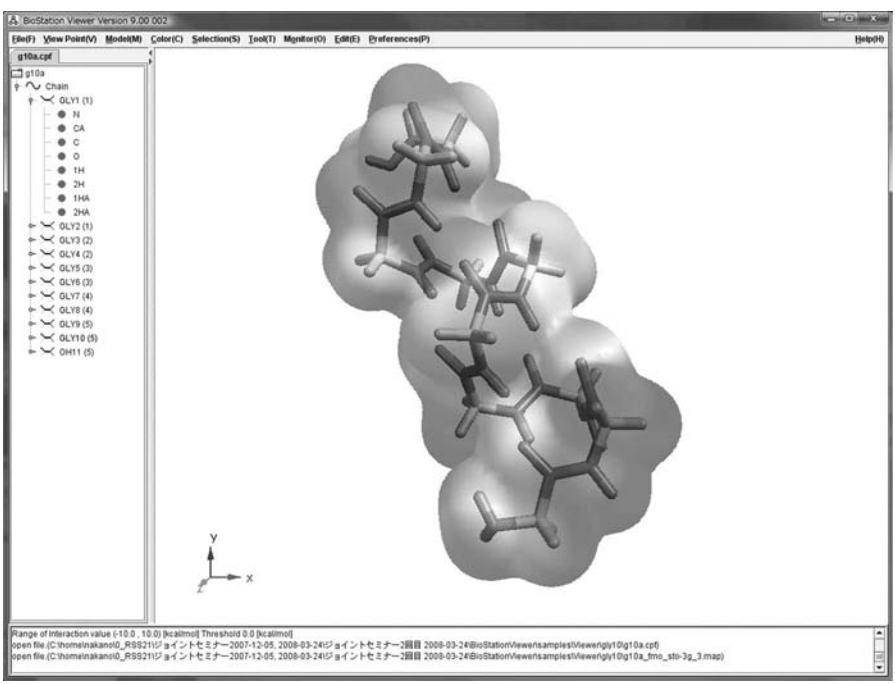


FIGURE 3.2 Main window of the BioStation Viewer with $(Gly)_{10}$ colored by the electrostatic potential value on the isosurface of electron density. **(See color insert following page 117.)**

The BioStation Viewer has three kinds of visualization methods for IFIEs — 1:1, n:1, and n:n:

1. *1:1* — When users select a reference fragment at the 3D window or Hierarchy Window, a molecular structure colored by the interaction energy value between the reference fragment and the other fragments is displayed. The color of the reference fragment is yellow and that of other fragments is changed from red to white to blue according to the changing IFIE value from low to high. In the default setting, the fragment with the attractive interaction is colored in red, and the fragment with the repulsive interaction is colored in blue. An example of the estrogen receptor–ligand complex is shown in Figure 3.3. In this case, the reference fragment is the ligand, and the other fragments around the ligand are shaded by interaction energy. The numerical values of IFIE could be shown in the list window of interaction energy.
2. *n:1* — By selecting multifragments as the reference fragment, the interaction energy between the reference fragment and the other fragments are displayed by color scale.
3. *n:n* — By selecting two multifragments, the interaction energy between two multifragments is displayed at the message area. This makes it possible to calculate the interaction energy of a complex between protein and protein, or DNA and protein.

FIGURE 3.3 An example of interfragment interaction energy (IFIE) analysis using the BioStation Viewer. Estrogen receptor–ligand complex shaded by interaction energy. The list of numerical values of IFIEs is displayed in the right window.

Advanced analysis methods based on the FMO methods are shown in the following subsections, and practical applications of IFIE analysis by using the BioStation Viewer are shown in Chapter 7.

3.3.2 Configuration Analysis for Fragment Interactions (CAFI)

Mochizuki et al. proposed a modified version of configuration analysis (CA) for the fragment interaction in conjunction with the FMO scheme.[13] The proposal is abbreviated as CAFI. The MO sets of fragments are merged and then orthonormalized by the use of a weighted Löwdin (WLO) orthonormalization.[54] The energy calculation is performed with the concurrent electron relaxation functional (CERF).[15] The relaxation energy is obtained in an orbital-wise fashion and is distinguished as the charge-transfer (CT) and the polarization (POL). The CAFI is designed to investigate the hydrogen-bonding interactions with CT primarily. The CT can be identified by the interfragment single excitations. Similarly, the intrafragment excitations provide the contribution of POL which may be incrementally induced by CTs. Figure 3.4 illustrates the conceptual scheme of the CAFI.

First, the monomer SCC calculations are completed, and the occupied MOs monomers would be localized with a localized MO method, such as Pipek and Mezey and Boys. Second, the MOs of monomers are merged and orthonormalized to form the MO-basis for the CERF energy calculation,[15] where the way of orthonormalization is a critical issue. We employ the WLO method[54] and set weight factor, ω_I, to unity for the occupied MOs and 0.07 for the virtual MOs. This setting yields the reasonable energy for the CT in hydrogen-bonding systems.

The CAFI wavefunction based on the CERF method[15] is written as an intermediately normalized style:

$$\Psi = \Psi_0 + \sum_{ia} C_{ia}\Psi_{ia} = \Psi_0 + \sum_{\substack{ia \\ i\in I,\, a\in J\neq I}}^{(\mathrm{CT})} C_{ia}\Psi_{ia} + \sum_{\substack{ia \\ i,\, a\in I}}^{(\mathrm{POL})} C_{ia}\Psi_{ia} \tag{3.42}$$

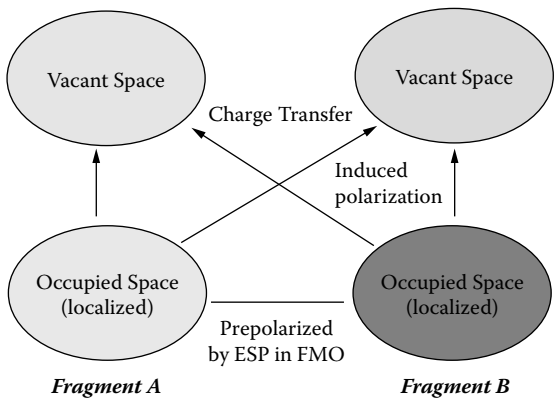


FIGURE 3.4 Conceptual scheme of configuration analysis for fragment interactions (CAFI).

where indices i and a specify the occupied MOs and virtual MOs of monomers, respectively. I and J are assigned as monomer indices. The RHF reference is given as Ψ_0 to be relaxed with the mixing of singly excited configuration functions, Ψ_{ia}. The CT and POL contributions could be approximately rewritten as a perturbative form:

$$e_i^{(\mathrm{CT})} = \sum_{\substack{a \\ a \in J \neq I}}^{(\mathrm{CT})} C_{ia} \langle \Psi_0 | H - E_0 | \Psi_{ia} \rangle \tag{3.43}$$

$$e_i^{(\mathrm{POL})} = \sum_{\substack{a \\ a \in I}}^{(\mathrm{POL})} C_{ia} \langle \Psi_0 | H - E_0 | \Psi_{ia} \rangle \tag{3.44}$$

Finally, the representation of the virtual MO space is considered. If the virtual-density elements about the single occupied MO labeled with i

$$D_{ab} = C_{ia} C_{ib} / N_i \tag{3.45}$$

are diagonalized, a unique natural orbital (NO) with a finite occupation number would be obtained. The single NO corresponds to the "particle orbital" to which electrons are transferred from the "hole orbital." The occupation number shows the transferred electrons. The computer graphics presentation by the BioStation Viewer for the pair of donor and acceptor orbitals would be useful in grasping the orbital-wise nature of hydrogen-bonding interactions.

The CAFI should be usable even if the target molecular system is divided with the bond-detached atom (BDA) in the fragmentation.[12,55] The glycine pentamer was employed for the test calculation, where the single glycine residue was treated as a fragment monomer. Table 3.2 lists the CAFI energies for the hydrogen bonding between Gly3 and Gly5 at the RHF/6-31G** level of theory. Two pairs of donor and acceptor orbitals associated with the "3–5" CT are visualized in Figure 3.5. The acceptor oribitals are σ^*_{NH}. It is notable that the π-type donor orbitals have the amplitude not only on the oxygen atom but also on the carbon atom at the carbonyl

TABLE 3.2
CAFI Energies (in kcal/mol) for Gly3-Gly5 Hydrogen-Bonding in Glycine Pentamer with FMO-RHF/6-31G Level of Theory**

Number	Energy (CT)	Energy (POL)	Donor Orbital	Acceptor Orbital	Occupation Number
1	–1.92	0.01	σ lone-pair	σ^*_{NH}	0.0025
2	–1.42	0.04	π lone-pair	σ^*_{NH}	0.0033

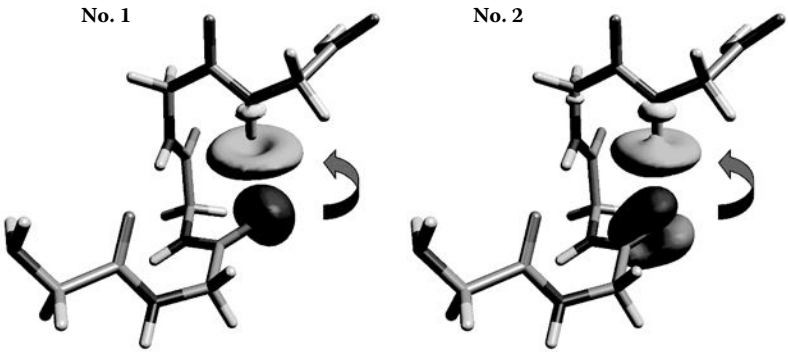


FIGURE 3.5 Two pairs of hole (donor) orbital and particle (acceptor) orbital concerning CT of Gly3–Gly5 hydrogen bonding in glycine pentamer. Arrows indicate the direction of electron transfer.

group in Gly3. Practical applications of CAFI are illustrated by Komeiji (Chapter 6) and Fukuzawa et al. (Chapter 7).

3.3.3 Fragment Interaction Analysis Based on Local MP2 (FILM)

Recently, Ishikawa et al. developed a method[16,17] named "fragment interaction analysis based on local MP2 (FILM)" by combining the FMO method and local MP2 (LMP2).[56–59] The FILM is a tool for analyzing site-specific interfragment interaction originated from dispersion interactions such as van der Waals, CH/π, and π/π interactions. The FILM enables us to decompose the dispersion interaction into the pair correlation energies between localized occupied orbitals. In the FILM analysis, correlation correction of IFIE, $\Delta E_{IJ}^{\mathrm{LMP2}}$, is calculated by summation of pair correlation energies, e_{ij}, between localized molecular orbitals of fragment I and J:

$$\Delta E_{IJ}^{\mathrm{LMP2}} = \sum_{i \in I} \sum_{j \in J} e_{ij} \tag{3.46}$$

A contribution of dispersion interaction energy between specific functional groups can be evaluated by summing up the pair correlation energies belonging to the specific site.

A demonstrative calculation of the human immunodeficiency virus type 1 protease (HIV-1 PR)–lopinavir (LPV) complex revealed that the FILM could provide useful information.[16,17] Because it was shown that the ratios of the pair correlation energies to the total interaction energy were acceptable even if relatively small basis sets were used, we employed a 6-31G* basis set in this calculation. We selected the three functional groups of LPV, $-CH_3$, $-C_6H_5$, and –CH, which are spatially close and directed to $I84_B$ of HIV-1 PR. By summing up the pair correlation energies

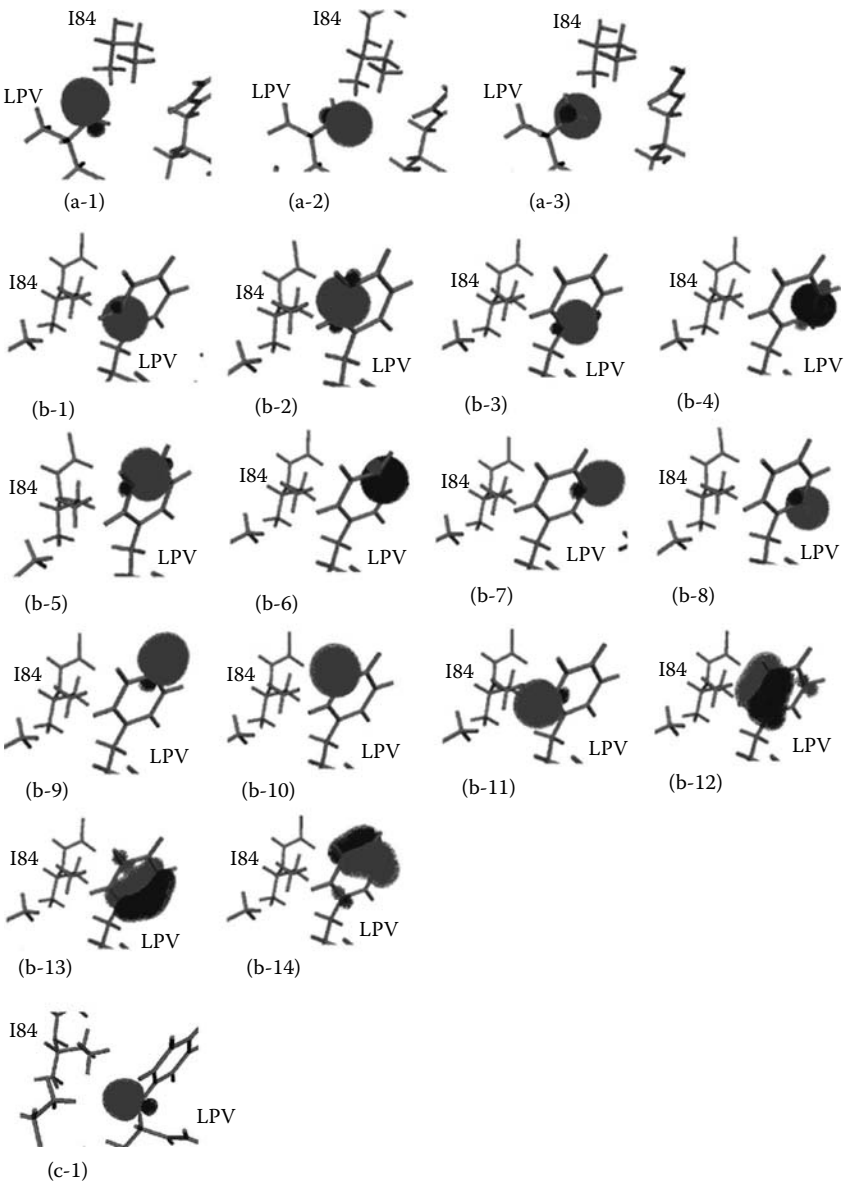


FIGURE 3.6 Localized orbitals located on the selected three functional groups of lopinavir (LPV).

between orbitals located at each site of LPV and all orbitals belonging to $I84_B$, we could know the degree of importance of the site with respect to binding with $I84_B$. The three orbitals of σ_{CH} in the methyl group, (a-1), (a-2), and (a-3) in Figure 3.6, contributed –1.28 kcal/mol to stabilization. This is 29.6% of the total stabilization

energy between $I84_B$ and LPV. Fourteen orbitals of the phenyl group, which correspond to six σ_{CC}, five σ_{CH}, and three π_{CC} orbitals, are shown in (b-1) through (b-14) in Figure 3.6. The sum of pair correlation energies between the fourteen orbitals and all orbitals in $I84_B$ was –1.26 kcal/mol, which is 29.2% of the total stabilization. The –CH site has just one σ_{CH} orbital (c-1) as shown in Figure 3.6. The sum of pair correlation energies between this orbital and all the orbitals in $I84_B$ was –0.51 kcal/mol, which was 11.8% of the total stabilization energy. The sum of stabilization energies of these three functional groups in LPV covers 70.6% of the total stabilization energy originating from dispersion interaction in the binding with $I84_B$. The remaining interaction energy (29.6%) is assigned to the other sites of LPV. This FILM analysis will be easily applied to interactions with other amino acid residues.

3.3.4 IFIE Map: Two-Dimensional Visual Representation of an IFIE Matrix

The IFIEs illustrate the information about the interaction energies between the fragments consisting of amino acids, nucleotides, and other molecules. Entire IFIEs are represented as a matrix form called an IFIE matrix. However, the number of IFIE values obtained substantially increases in proportion to the square of the number of fragments (e.g., 4950 IFIE values for a system containing 100 fragments). Thus, we chose some fragments that were expected to be important for biomacromolecular functions and treated only a limited set of IFIEs for the analysis, while a complete set of IFIEs would contain more information concerning the correlation with structural properties of biomacromolecules, such as positions of secondary structure and complex stability. Recently, Kurisaki et al. proposed a visual representation of the IFIE matrix, an IFIE map, for a comprehensive analysis of a whole set of IFIEs.[18]

Figure 3.7 shows the IFIE map with the data obtained from the FMO calculation at the MP2/6-31G level for the CRP–cAMP–DNA complex system.[18,40] In the intra-CRP IFIE (Figure 3.7), we find three kinds of patterns: (1) parallel patterns to abscissa or ordinate on the map, (2) a band pattern near and parallel to the diagonal line, and (3) band patterns roughly perpendicular to the diagonal line. Pattern 1 is associated with Coulomb interactions, because charged residues can interact strongly with other distant amino acid residues. Patterns 2 and 3 are explained by their correspondence to protein secondary structures. Comparing these patterns with the positions of secondary structure, we find that pattern 2 corresponds to an α-helix (Figure 3.8a), and pattern 3 corresponds to an antiparallel β-sheet (Figure 3.8b).

3.3.5 Visualized Cluster Analysis of Protein–Ligand Interactions (VISCANA)

Amari et al. proposed the visualized cluster analysis of protein–ligand interaction (VISCANA) for virtual ligand screening[14] based on the FMO method, by using the dissimilarity between the interaction energy patterns of two ligands and by representing each data point with a color that quantitatively and qualitatively reflects the

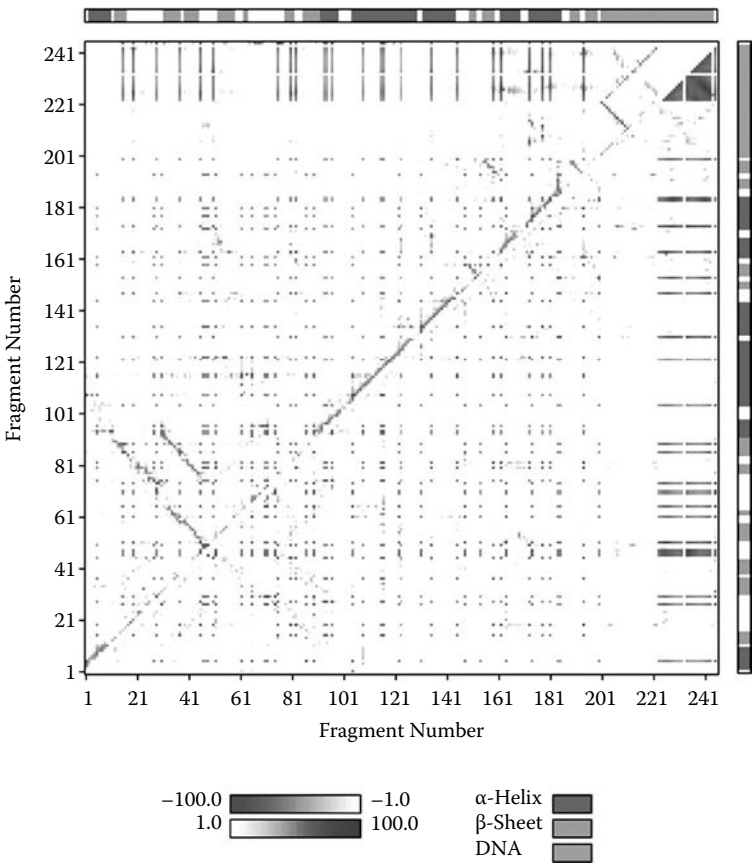
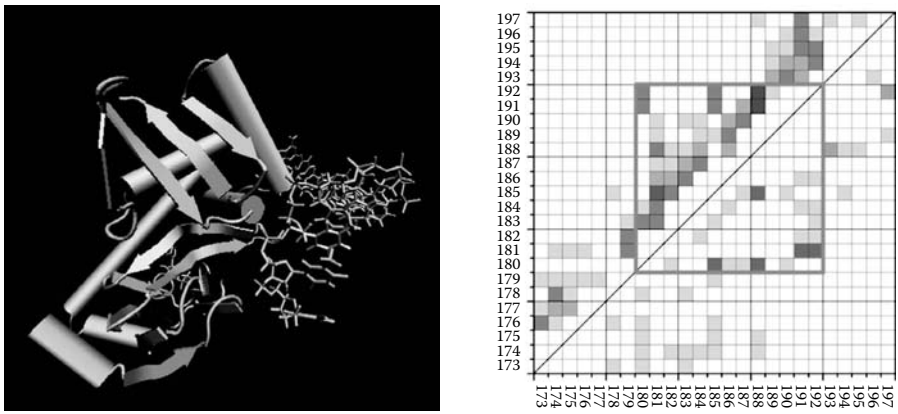


FIGURE 3.7 An interfragment interaction energy (IFIE) map for the CRP–cAMP–DNA complex system. The upper triangle is used to show the plots of negative IFIE values, and the lower triangle plots positive IFIE values. Numbers 1 to 200 and 201 to 244 correspond to amino acid residues and DNA, respectively, whereas number 245 is used for cAMP. (See also the color panel of Figure 7.12 in Chapter 7.)

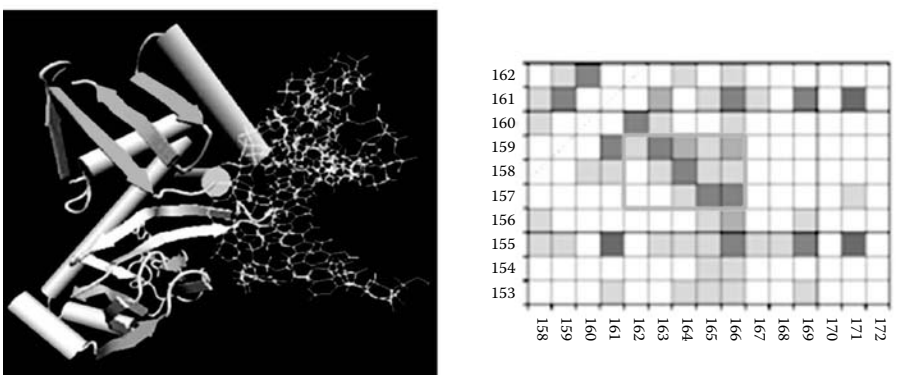
interaction energy. The dissimilarity of IFIEs between the two compounds I and J, d_{IJ}, is expressed by the following formula:

$$d_{IJ} = \sum_{K=1}^{N} (\Delta\tilde{E}_{IK} - \Delta\tilde{E}_{JK})^2 \tag{3.47}$$

where N is the number of amino acid residues of the target protein, and K is the fragment index of the residues. The VISCANA could be applied to not only the FMO method but also any molecular interaction analysis that can provide interaction energies or other properties of interest such as charge distribution. For example, we could also define $\Delta\tilde{E}_{IK}$ as the ligand–amino acid residue interaction energy by using the



(a)



(b)

FIGURE 3.8 Interfragment interaction energy (IFIE) maps focusing on regions forming secondary structures. (a) IFIE map corresponding to regions forming α-helix. (b) IFIE map corresponding to regions forming antiparallel β-sheet. **(See color insert.)**

energy of decomposition of nonbonding interaction energies (electrostatic and van der Waals) with the classical force field.

The receptor–ligand interaction energy matrix is defined as follows:

$$\begin{pmatrix} \Delta\tilde{E}_{\text{ligand}\,1,\,1} & \Delta\tilde{E}_{\text{ligand}\,1,\,2} & \cdots & \Delta\tilde{E}_{\text{ligand}\,1,\,N} \\ \Delta\tilde{E}_{\text{ligand}\,2,\,1} & \Delta\tilde{E}_{\text{ligand}\,2,\,2} & \cdots & \Delta\tilde{E}_{\text{ligand}\,2,\,N} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta\tilde{E}_{\text{ligand}\,L,\,1} & \Delta\tilde{E}_{\text{ligand}\,L,\,2} & \cdots & \Delta\tilde{E}_{\text{ligand}\,L,\,N} \end{pmatrix} \tag{3.48}$$

where L is the number of ligands. The IFIE matrix is visualized with a color scale from red ($\Delta\tilde{E}_{IK} < 0$; attractive interaction) to blue ($\Delta\tilde{E}_{IK} > 0$; repulsive interaction). The order of ligand compounds was sorted by a hierarchical clustering procedure. The farthest-neighbor method is used as default for the process of forming a hierarchical cluster and dendrogram. The VISCANA makes it possible to classify structurally similar ligand molecules through the interaction pattern of a ligand and amino acid residues of the receptor protein as well as the interaction energy of the ligands and the protein.

In addition to electrostatic interactions and hydrogen bonds, dispersion (van der Waals) interactions are very important in biomolecules. Because dispersion is an electron correlation effect, a level of at least the MP2 method has to be applied. We applied the VISCANA to a docking study of the pharmacophore of the human estrogen receptor (hER) ligand-binding domain (LBD) that consists of 57 amino acid residues. The FMO calculations were performed for 38 compounds from the *Ki*Bank,[60,61] with the hER LBD at the MP2/6-31G level of theory.

As shown in Figure 3.9, the VISCANA classified the docked compounds into three clusters: steroidal agonists such as 17β-estradiol, estrone, and ethynilestradiol (cluster A); nonsteroidal agonists such as bisphenol A, hexestrol, and diethylstilbestrol (cluster B); and other compounds including testosterone, progesterone, and methyltrienolone (cluster C). The IFIEs between each fragment — Glu353, Arg394, and His524 — and a ligand in cluster A were attractive, and these results correspond to the hydrogen bond of steroidal agonists with Glu353, Arg394, and His524.[62] In the case of nonsteroidal ligands in cluster B, the interaction between a ligand and Met522 is crucial, instead of His524.

3.4 CONCLUSION

The ABINIT-MP program provides a powerful and useful tool to enable the quantum chemical calculations for huge biomolecules based on the FMO method. The "Revolutionary Simulation Software for the 21st Century (RSS21)" project edition of the ABINIT-MP that supports FMO-RHF and FMO-MP2 methods is freely available at www.ciss.iis.u-tokyo.ac.jp/rss21/.

We have been developing an experimental FMO program, named ABINIT-MPX, in various ways, such as including more accurate accounts for electron correlation effects, optimizing geometry with the electron correlation method, and implementing the periodic boundary condition (PBC)-FMO to make more efficient *ab initio* calculations for biomolecules feasible. Additionally, developing a high-accuracy and high-speed algorithm for computation of the environmental electrostatic potential and the dimer-es approximation is necessary to put the FMO3 method into practice. We would note that the development of parallelized integral-direct module for MP3 calculations[34,35,39] was recently completed. Grimme's spin-component scaling[63] might be used if necessary.

The visualization software for molecule and molecular-interaction analyses based on the FMO calculations with the ABINIT-MP, named BioStation Viewer, was

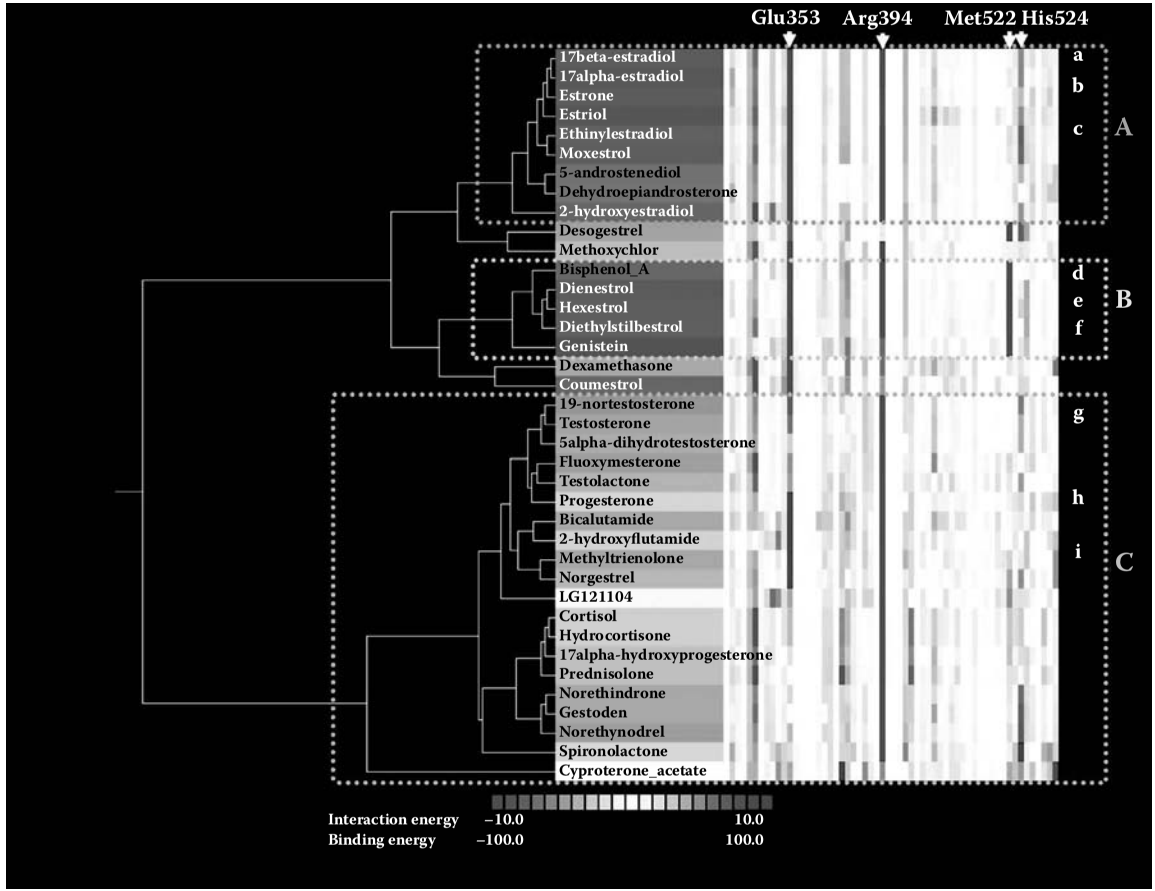


FIGURE 3.9 Hierarchical clustering by the interaction patterns of the human estrogen receptor ligand-binding domain and 38 compounds from the *Ki*Bank. Parts a through i denote the IFIEs of 17β-estradiol, estrone, ethynilestradiol, bisphenol A, hexestrol, diethylstilbestrol, testosterone, progesterone, and methyltrienolone, respectively. **(See color insert.)**

developed by our research group and is freely available at the Web site of the RSS21 project. The released version of the BioStation Viewer includes the VISCANA function with a supporting CSV format of the receptor–ligand interaction energy matrix. The developer version of the BioStation Viewer supporting CAFI, FILM, and IFIE map functions will be released later.

ACKNOWLEDGMENTS

The work reported here was supported by the "Revolutionary Simulation Software for the 21st Century (RSS21)" project operated by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and by the CREST project operated by the Japan Science and Technology Agency (JST). TN and YM should acknowledge the CPU resource supply of the Earth Simulator operated by the Japan Agency for Marine-Earth Science and Technology (JAMSTEC) and professional technical supports by Katsumi Yamashita and Tadashi Murase.

REFERENCES

1. Kitaura, K., Sawai, T., Asada, T., Nakano, T., Uebayasi, M. 1999. Pair interaction molecular orbital method: an approximate computational method for molecular interactions. *Chem. Phys. Lett.* 312:319–324.
2. Kitaura, K., Ikeo, E., Asada, T., Nakano, T., Uebayasi, M. 1999. Fragment molecular orbital method: an approximate computational method for large molecules. *Chem. Phys. Lett.* 313:701–706.
3. Fedorov, D.G., Kitaura, K. 2006. Theoretical development of the fragment molecular orbital (FMO) method. In *Modern methods for theoretical physical chemistry of biopolymers,* eds. E.B. Starikov, J.P. Lewis, S. Tanaka, 3–38. Amsterdam: Elsevier.
4. Nakano, T., Mochizuki, Y., Fukuzawa, K., Amari, S., Tanaka, S. 2006. Developments and applications of ABINIT-MP software based on the fragment molecular orbital method. In *Modern methods for theoretical physical chemistry of biopolymers,* eds., E.B. Starikov, J.P. Lewis, S. Tanaka, 39–52. Amsterdam: Elsevier.
5. Fedorov, D.G., Kitaura, K. 2007. Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. *J. Phys. Chem. A* 111:6904–6914.
6. Sekino, H., Sengoku, Y., Sugiki, S., Kurita, N. 2003. Molecular orbital analysis based on fragment molecular orbital scheme. *Chem. Phys. Lett.* 378:589–597.
7. Sugiki, S., Kurita, N., Sengoku, Y., Sekino, H. 2003. Fragment molecular orbital method with density functional theory and DIIS convergence acceleration. *Chem. Phys. Lett.* 382:611–617.
8. Sugiki, S., Matsuoka, M., Usuki, R. et al. 2005. Density functional calculations on the interaction between catabolite activator protein and cyclic AMP using the fragment molecular orbital method. *J. Theor. Comput. Chem.* 4:183–195.
9. Sekino, H., Matsumura, N., Sengoku, Y. 2007. Evaluation of NMR chemical shift by fragment molecular orbital method. *Comput. Lett.* 3:423–430.
10. Nemoto, T., Fedorov, D.G., Uebayasi, M., et al. 2005. *Ab initio* fragment molecular orbital (FMO) method applied to analysis of the ligand–protein interaction in a pheromone-binding protein. *Comput. Biol. Chem.* 29:434–439.

11. Fedorov, D.G., Kitaura, K. 2007. Pair interaction energy decomposition analysis. *J. Comput. Chem.* 28:222–237.
12. Nakano, T., Kaminuma, T., Sato, T. et al. 2002. Fragment molecular orbital method: use of approximate electrostatic potential. *Chem. Phys. Lett.* 351:475–480.
13. Mochizuki, Y., Fukuzawa, K., Kato, A. et al. 2005. A configuration analysis for fragment interaction. *Chem. Phys. Lett.* 410:247–253.
14. Amari, S., Aizawa, M., Zhang, J. et al. 2006. VISCANA: Visualized cluster analysis of protein–ligand interaction based on the *ab initio* fragment molecular orbital method for virtual ligand screening. *J. Chem. Inf. Model.* 46:221–230.
15. Mochizuki, Y. 2005. A size-extensive modification of super-CI for orbital relaxation. *Chem. Phys. Lett.* 410:165–171.
16. Ishikawa, T., Mochizuki, Y., Amari, S. et al. 2007. Fragment interaction analysis based on local MP2. *Theor. Chem. Acc.* 118:937–945.
17. Ishikawa, T., Mochizuki, Y., Amari, S. et al. 2008. An application of fragment interaction analysis based on local MP2. *Chem. Phys. Lett.* 463:189–194.
18. Kurisaki, I., Fukuzawa, K., Komeiji, Y. et al. 2007. Visualization analysis of Interfragment interaction energies of CRP-cAMP-DNA complex based on the fragment molecular orbital method. *Biophys. Chem.* 130:1–9.
19. Mochizuki, Y., Nakano, T., Koikegami, S. et al. 2004. A parallelized integral-direct second-order Møller-Plesset perturbation theory method with a fragment molecular orbital scheme. *Theor. Chem. Acc.* 112:442–452.
20. Mochizuki, Y., Koikegami, S., Nakano, T., Amari, S., Kitaura, K. 2004. Large scale MP2 calculations with fragment molecular orbital scheme. *Chem. Phys. Lett.* 396:473–479.
21. Mochizuki, Y., Yamashita, K., Murase, T. et al. 2008. Large scale FMO-MP2 calculations on a massively parallel-vector computer. *Chem. Phys. Lett.* 457:396–403.
22. Mochizuki, Y., Koikegami, S., Amari, S. et al. 2005. Configuration interaction singles method with multilayer fragment molecular orbital scheme. *Chem. Phys. Lett.* 406:283–288.
23. Mochizuki, Y., Tanaka, K., Yamashita, K. et al. 2007. Parallelized integral-direct CIS(D) calculations with multilayer fragment molecular orbital scheme. *Theor. Chem. Acc.* 117:541–553.
24. Mochizuki, Y., Nakano, T., Amari, S. et al. 2007. Fragment molecular orbital calculations on red fluorescent protein (DsRed). *Chem. Phys. Lett.* 433:360–367.
25. Mochizuki, Y., Tanaka, K. 2007. Modification for spin-adapted version of configuration interaction singles with perturbative doubles. *Chem. Phys. Lett.* 443:389–397.
26. Mochizuki, Y. 2008. A practical use of self-energy shift for the description of orbital relaxation. *Chem. Phys. Lett.* 453:109–116.
27. Fedorov, D.G., Kitaura, K. 2004. The importance of three-body terms in the fragment molecular orbital method. *J. Chem. Phys.* 120:6832–6840.
28. Fedorov, D.G., Kitaura, K. 2004. On the accuracy of the 3-body fragment molecular orbital method (FMO) applied to density functional theory. *Chem. Phys. Lett.* 389:129–134.
29. Fedorov, D.G., Kitaura, K. 2006. The three-body fragment molecular orbital method for accurate calculations of large systems. *Chem. Phys. Lett.* 433:182–187.
30. Fedorov, D.G., Ishimura, K., Ishida, T. et al. 2007. Accuracy of the three-body fragment molecular orbital method applied to Møller–Plesset perturbation theory. *J. Comput. Chem.* 28:1476–1484.
31. Fedorov, D.G., Olson, R.M., Kitaura, K., Gordon, M.S., Koseki, S. 2004. A new hierarchical parallelization scheme: generalized distributed data interface (GDDI), and an application to the fragment molecular orbital method (FMO). *J. Comput. Chem.* 25:872–880.

32. www.mpi-forum.org/.
33. Fedorov, D.G., Kitaura, K. 2004. Second order Møller–Plesset perturbation theory based upon the fragment molecular orbital method. *J. Chem. Phys.* 121:2483–2490.
34. Szabo, A., Ostlund, N.S. 1989. *Modern quantum chemistry.* New York: McGraw-Hill.
35. Foresman, J.B., Frisch, A.E. 1996. *Exploring chemistry with electronic structure methods,* 2nd ed. Pittsburgh: Gaussian Inc., and references therein.
36. Revolutionary Simulation Software for the 21st Century (RSS21) project edition of ABINIT-MP and BioStation Viewer are available from www.ciss.iis.u-tokyo.ac.jp/rss21/.
37. Helgaker, T., Jørgensen, P., Olsen, J. 2000. *Molecular electronic-structure theory.* Chichester: Wiley.
38. Head-Gordon, M., Pople, J.A., Frisch, M.J. 1988. MP2 energy evaluation by direct metods. *Chem. Phys. Lett.* 153:503–506.
39. Sæbø, S., Almlöf, J. 1989. Avoiding the integral storage bottleneck in LCAO calculations of electron correlation. *Chem. Phys. Lett.* 154:83–89.
40. Fukuzawa, K., Komeiji, Y., Mochizuki, Y. et al. 2006. Intra- and intermolecular interactions between cyclic-AMP receptor protein and DNA: *ab initio* fragment molecular orbital study. *J. Comput. Chem.* 27:948–960; 2007; ibid. 28:2237–2239.
41. Fukuzawa, K., Mochizuki, Y., Tanaka, S., Kitaura, K., Nakano, T. 2006. Molecular interactions between estrogen receptor and its ligand studied by the *ab initio* fragment molecular orbital method. *J. Phys. Chem. B* 110:16102–16110; ibid. 110:24276.
42. Jensen, H.J.A., Jørgensen, P., Ågren, H., Olsen, J. 1988. Second-order Møller–Plesset perturbation theory as a configuration and orbital generator in multiconfiguration self-consistent field calculations. *J. Chem. Phys.* 88:3834–3839; ibid. 89:5354 (1988).
43. Almlöf, J., Faegri, K., Korsell, K. 1982. Principles for a direct SCF approach to LCAO-MO *ab-initio* calculations. *J. Comput. Chem.* 3:385–399.
44. Obara, S., Saika, A. 1986. Efficient recursive computation of molecular integrals over Cartesian Gaussian functions. *J. Chem. Phys.* 84:3963–3974.
45. McMurchie, L.E., Davidson, E.R. 1978. One- and two-electron integrals over Cartesian Gaussian functions. *J. Comput. Phys.* 26:218–231.
46. Saunders, V.R. 1975. An introduction to molecular integral evaluation. In *Computational techniques in quantum chemistry and molecular physics,* eds. G.H.F. Diercksen, B.T. Sutcliffe, A. Veillard. D. Reidel, Dordrecht, 347–424.
47. Veillard, A. 1975. The logic of SCF procedures. In *Computational techniques in quantum chemistry and molecular physics*, eds. G.H.F. Diercksen, B.T. Sutcliffe, A. Veillard. D. Reidel, Dordrecht, 201–250.
48. Nishiguchi, A., Orii, S., Yabe, T. 1985. Vector calculation of particle code. *J. Comput. Phys.* 61:519–522.
49. Mochizuki, Y., Matsumura, M., Yokura, T., Hirahara, Y., Imamura, T. 2002. Vectorization of direct Fock matrix construction in Dirac-DHF calculations. *J. Nucl. Sci. Technol.* 39:195–199.
50. www.jamstec.go.jp/es/en/index.html.
51. www.wwpdb.org/docs.html.
52. www.tripos.com/tripos_resources/fileroot/mol2_format_Dec07.pdf.
53. www.symxl.com/downloads/public/ctfile/ctfile.pdf.
54. Maki, J., Yamagishi, H., Noro, T., Sasaki, F., Yamamoto, Y. 1996. New "localized orbitals" appropriate for post-Hartree–Fock calculations. *Int. J. Quant. Chem.* 60:731–742.
55. Kitaura, K., Sugiki, S., Nakano, T., Komeiji, Y., Uebayasi, M. 2001. Fragment molecular orbital method: analytical energy gradients. *Chem. Phys. Lett.* 336:163–170.
56. Pulay, P., Saebø, S. 1986. Oribital-invariant formulation and second-order gradient evaluation in Møller–Plesset perturbation theory. *Theor. Chim. Acta.* 69:357–368.

57. Rauhut, G., Pulay, P., Werner, H.J. 1998. Integral transformation with low-order scaling for large local second-order Møller–Plesset calculations. *J. Comput. Chem.* 19:1241–1254.
58. Schütz, M., Hetzer, G., Werner, H.J. 1999. Low-order scaling local electron correlation methods. I. Linear scaling local MP2. *J. Chem. Phys.* 111:5691–5705.
59. Saebø, S., Pulay, P. 2001. A low-scaling method for second order Møller–Plesset calculations. *J. Chem. Phys.* 115:3975–3983.
60. Aizawa, M., Onodera, K., Zhang, J. et al. 2004. *Ki*Bank: a database for computer-aided drug design based on protein-chemical interaction analysis. *Yakugaku Zasshi* 124:613–619.
61. Zhang, J., Aizawa, M., Amari, S. et al. 2004. Development of *Ki*Bank, a database supporting structure-based drug design. *Comput. Biol. Chem.* 28:401–407.
62. Brzozowski, A.M., Pike, A.C.W., Dauter, Z. et al. 1997. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* 389:753–758.
63. Grimme, S. 2003. Improved third-order Møller–Plesset perturbation theory. *J. Comput. Chem.* 24:1529–1537.

4 Excited States of Photoactive Proteins by Configuration Interaction Studies

Yuji Mochizuki, Tatsuya Nakano, Naoki Taguchi, and Shigenori Tanaka

CONTENTS

4.1 INTRODUCTION

A number of biochemical systems are photoactive in the visible region through electronic transitions. The rhodopsin protein in vertebrate eyes is a representative case of such photoabsorptions. As for emissions, the green fluorescent protein (GFP)[1] and related derivatives are useful marker labels in the field of biotechnology. The central region in photoactive proteins is known as the chromophore, and it consists of the crucial pigment part and some neighboring residues that should provide the electrostatic or hydrogen-bonding interactions with the pigment. Surrounding residues also put some electrostatic potentials on the chromophore system. These environmental

effects in proteins should cause the redshifts or blueshifts in transition energies of the pigment moieties, where the opsin shift is an archetype. Similar phenomena are known well for the solvated molecules as the solvatochromic shifts.[2] A sort of hybrid theoretical approach of quantum mechanics (QM) and classical molecular mechanics (MM), denoted QM/MM, has been frequently used to treat such environmental shifts. Namely, QM or molecular orbital (MO) calculations are applied only to the photoactive molecule and some neighboring areas when needed. Surrounding parts are treated by the MM methods. Although the MM technique makes the inclusion of environmental effects feasible, there exist many prefixed parameters that are empirically adjusted to reproduce the experimental values. The MM description for hydrogen bonding, which should play a crucial role for water molecules and amino acids, has not yet been well established. The treatment of polarization has been in a similar state, and the electron delocalization is hardly described in MM ways. These difficulties lead to an inherent ambiguity in QM/MM calculations. Thus, a full QM treatment is desirable if possible.

The fragment molecular orbital (FMO) method proposed by Kitaura et al.[3,4] is one of the promising recipes to calculate proteins and solvation clusters in a fully QM manner. In the two-body FMO scheme for the ground state of a given system, a series of Hartree–Fock (HF) calculations are performed for the fragment monomers and dimers under the environmental electrostatic potential (ESP), which is essential to ensure chemical reliability. The introduction of a bond-detachment atom (BDA) is another key point, by which no artificial hydrogen-capping is needed for the fragmentations. The electron correlation energies are corrected for the respective monomers and dimers through the second-order Møller–Plesset perturbation (MP2) theory, maintaining the size consistency. These FMO-HF and FMO-MP2 calculations can be efficiently parallelized by combining the (upper) fragment indices and the (lower) integral indices, and this combination parallelization provides reasonable costs of computation for realistic proteins. Additionally, the well-defined fragment-wise interaction energies are naturally obtained, which are useful to grasp the insightful picture of proteins. The multilayer FMO framework[5] is usable to treat chemically important regions of the target system at post-HF levels. For example, the region of photochemical interest (i.e., chromophore) is subjected to the excited states calculations by keeping the remaining environmental parts at the HF level. A variety of FMO calculations have been available with the GAMESS program[6] extensively modified by Fedorov et al.[7] and also with the ABINIT-MP program originally developed by Nakano et al.[8] for proteins. For more details on FMO methodologies and associated applications, refer to other chapters in this book or to the review article written by Fedorov and Kitaura.[9]

Hereafter, we focus on the excited states calculations of configuration interaction singles (CIS)[10] and CIS with second-order perturbative doubles (CIS(D))[11] in the multilayer FMO treatment.[5] The reason why these singles-based methods have been adopted relies on the fact that most of the intense visible absorptions or emissions of photoactive proteins are typically characterized by the single electron transitions among a few MOs of occupied–unoccupied boundaries (e.g., HOMO-LUMO). The remaining parts of this chapter are configured as follows. Section 4.2 summarizes the theoretical aspects of CIS and CIS(D) calculations[10,11] and the corresponding

parallelized integral-direct implementations[12,13] in a developer version of ABINIT-MP, currently distinguished as ABINIT-MPX. Higher-order modifications[14,15] of CIS(D) are also introduced in this section. The modeling of target systems is addressed in Section 4.3. A series of applications to photoactive proteins[12,13,16] is presented in Section 4.4, where it is shown that accuracy of 0.1 eV is obtainable in evaluating transition energies at the CIS(D) level as far as the modeled molecular structure is reliable enough. Finally, Section 4.5 addresses other works and also future directions of applications and related concerns. It should be noted that "multilayer FMO" was originally abbreviated as "MFMO" by Fedorov and Kitaura[5] (see Chapter 2) but is abbreviated as "MLFMO" by us[12,13,16] with equivalence. We will use the term "MLFMO" hereafter in this chapter.

4.2 METHODS OF CALCULATION

4.2.1 Configuration Interaction Singles (CIS) and Doubles CIS(D)

The CIS method[10] has been known as a semiquantitative tool to treat low-lying states being characterized by single excitations from the HF ground state. By letting the indices i and a specify the occupied and unoccupied orbitals, respectively, the CIS wavefunction can be written in the spin-orbital notation[17] as

$$\Psi_{\mathrm{CIS}} = U_1\Phi_0 = \Sigma_{ia} b_i^a \Phi_i^a \tag{4.1}$$

Operator U_1 acts the HF determinant Φ_0 to generate the CIS state with an excitation energy ω and amplitude vectors b_i^a of singly excited determinants Φ_i^a. The CIS energy and amplitudes are obtained by diagonalizing the Hamiltonian matrix of

$$\mathrm{H}_{ia,jb} = \delta_{ij}\delta_{ab}(\varepsilon_a - \varepsilon_i) + \langle aj||ib\rangle \tag{4.2}$$

where ε_a and ε_i are the orbital energies and Dirac's notation is used for two-electron integrals (TEIs). It is notable that the size consistency is satisfied in CIS calculations to estimate excitation energy, in contrast to the violation in CIS and doubles [CIS(D)] as a correlated treatment for the ground state.[17] A typical overestimation in CIS excitation energies of singlet states has been known to be 1 to 2 eV relative to experimental values for most of the closed-shell molecules. Foresman et al.[10] proposed a perturbative correction recipe for a certain CIS state in analogy with the MP2 correction for the ground state.[17] In this CIS-MP2 method, the doubles and triples determinants (Φ_{ij}^{ab} and Φ_{ijk}^{abc}, respectively) are employed in the perturbation expansion, where the former contribution commits the orbital relaxation energy and the latter contribution concerns the correlation energy. The size consistency is, however, violated in the latter contribution, leading to no significant remediation for the overestimated energies by CIS.[10] Head-Gordon et al.[11] devised an alternative correction recipe in which the triples are factorized by CIS-singles (operator U_1) and MP2-doubles (operator T_2) by keeping the doubles (operator U_2) preserved, as the CIS(D) method: please do not confuse CIS(D) and CISD. By the

factorization of triples, the size consistency is maintained in CIS(D), and the correlation energy contribution is given as a differential energy from the ground state correlated at the MP2 level. Head-Gordon et al.[11] discussed that the CIS(D) method can be considered a noniterative second-order approximation of coupled-cluster singles and doubles (CCSD) response theories, and they then showed that the correction is particularly effective in reducing overestimation errors for valence excited states.

For later convenience, the CIS(D) energy correction is symbolically denoted by using operators:

$$\omega^{\mathrm{CIS(D)}} = \langle \Psi_{\mathrm{CIS}} | V | U_2 \Phi_0 \rangle + \langle \Psi_{\mathrm{CIS}} | V | T_2 U_1 \Phi_0 \rangle_{\mathrm{c}} \tag{4.3}$$

where V is the fluctuation potential of the Mϕller–Plesset perturbation.[17] The first and second terms on the right-hand side of this equation correspond to the relaxation energy and the differential correlation energy, respectively. Operator U_2 can be triplet as U_1, but T_2 is always the spin singlet manifestly. The subscript "c" in the second term in Equation 4.3 means the "connected part" of operator contractions, whereas the "disconnected part" is just the MP2 correlation energy of the ground state. The actual expressions of $\omega^{\mathrm{CIS(D)}}$ have the form of tensor contractions among two-electron integrals, orbital energies, and CIS amplitudes. Because they are somewhat lengthy even in the spin-orbital notation, we would omit here the listing of contraction equations: please see Head-Gordon's original paper[11] if interested in more information. As in the MP2 case, the formal cost of CIS(D) computation scales is N^5 (where N is the number of basis functions as a symbolic index of molecular size). The relaxation energy contribution is clearly negative (or energy lowering), whereas the differential correlation energy is expected to be positive because a hole is created in the occupied orbital space by the single excitation. The gross amount of energy correction is thus determined by the balance of these two contributions. The CIS(D) method works for most low-lying valence excited states.[11] In contrast, it has been known that Rydberg excitation energies are rather underestimated through too large corrections from U_2, relative to the reference values of experiments and also of CCSD response calculations.

Although we pursued the CIS and CIS(D) methods to date, it may be fair to address the time-dependent density functional theory (TDDFT). TDDFT calculations have become the *de facto* standard as a quantitative method for excited states,[18] where the electron correlation is effectively incorporated in the set of potentials containing an amount of adjustable parameters. However, the TDDFT description could collapse for the charge transfer (CT) excitations with erratically low energies, as exemplified by Dreuw et al.[19] at the first time. This situation should cause a difficulty in treating the chromophore consisting of the pigment and neighboring residues (or solvents) because the spurious CT states could easily appear to interfere with the target valence excited state. To overcome the issues related to CT, modified DFT treatments with the long-range correction (LC) have been extensively developed in recent years.[18] Chiba et al.[20,21] enabled the TDDFT(LC) method for FMO calculations in the GAMESS program: see the corresponding documentation by Chiba et al. in Chapter 5.

4.2.2 Implementation of CIS and CIS(D) in ABINIT-MPX

As the first step, the CIS method[10] was implemented in a spin-adapted fashion for the separate treatment of singlet and triplet.[12] For the spin singlet, the Hamiltonian matrix is given, using Mulliken's notation for TEIs, as

$$H_{ia,jb} = \delta_{ij}\delta_{ab}(\varepsilon_a - \varepsilon_i) + 2(ia,jb) - (ij,ab) \tag{4.4}$$

Because at most a few low-lying solutions are of interest, the iterative diagonalization is usually carried out by computing the sigma vectors as the products of Hamiltonian elements and trial CIS vectors. Both the integral transformation of TEI and the explicit construction of Hamiltonian elements should be avoided for large molecules such as pigment moieties in photoactive proteins. According to Foresman et al.,[10] the sigma vectors can be directly computed from the TEI list of atomic basis indices by using the Fock-like contraction technique. This way of CIS computation was adopted in our ABINIT-MPX program, where the Fock-like processing of N^4 scaling was parallelized in a distributed integral-direct way.[8,12] The oscillator strength and (nonrelaxed) density matrix are calculated for analysis purpose after the convergence of iterative diagonalization.

In the dual-layer scheme of MFMO,[5] the HF calculations are first performed for the monomers and dimers belonging to both layer 1 (denoted as L_1) and layer 2 (L_2), as in the case of regular single-layer FMO. The environmental potential from layer 1 to layer 2 is thus incorporated without any empirical parameters, unlike MM-based methods. The FMO-HF calculations are, of course, done with an integral-direct parallelism.[8] If layer 2 as the chromophore, to which the CIS method is applied, corresponds to the single fragment, the MLFMO-CIS energy would simply be given by

$$E^{\text{MLFMO-CIS}} = E^{\text{FMO-HF}}(L_1,L_2) + \omega^{\text{CIS}}(L_2) \tag{4.5}$$

where the second term on the right-hand side is the excitation energy of layer 2. This is the present MLFMO-CIS implementation.[12] It is noted that layer 1 is kept frozen for the CIS excited states of layer 2: this treatment may be termed as FMO-HF:CIS, alternatively.[5] Layer 2 composition would be defined as having a "margin" around the pigment. As a result, the size of the layer 2 fragment can be potentially large relative to all the other fragments in layer 1.

The CIS(D) method[11] was successively implemented in the ABINIT-MPX program[13] by extending the parallelized integral-direct (or on-the-fly) algorithm for MP2 calculations.[17,22,23] The CIS(D) correction is additive, and then the MLFMO-CIS(D) energy becomes[13]

$$E^{\text{MLFMO-CIS(D)}} = E^{\text{FMO-HF}}(L_1,L_2) + \omega^{\text{CIS}}(L_2) + \omega^{\text{CIS(D)}}(L_2) \tag{4.6}$$

The orbital relaxation could be properly described through the U_2 contribution of CIS(D) if layer 2 is set with an appropriate "margin." This is essential to obtain the gross energy lowering from the CIS energy, as will be demonstrated by the MLFMO-CIS(D) calculations in Section 4.4.

Although Head-Gordon et al.[11,24] made the spin-orbital formulation and implementation, we used a spin-adapted CIS(D) method through reformulations using a generator algebra for the implementation[13] in relation with the precedent CIS implementation.[12] The resulting spin-adapted expressions[13] were more complicated than the spin-orbital expressions.[11] We will again refrain from listing the tedious equations. Alternatively, our parallel implementations[13] in ABINIT-MPX are outlined as follows. As in the MP2 case,[22,23] a series of quarter transformation from the TEIs with indices of atomic basis functions (pq,rs),

$$(iq,rs) = \Sigma_p C_{pi}(pq,rs) \tag{4.7}$$

$$(ia,rs) = \Sigma_q C_{qa}(iq,rs) \tag{4.8}$$

$$(ia,js) = \Sigma_r C_{rj}(ia,js) \tag{4.9}$$

$$(ia,jb) = \Sigma_s C_{sb}(ia,js) \tag{4.10}$$

drive the T_2U_1 processing in an i-batch mode, where the basis-orbital coefficients are determined by the HF calculations. The parallelization of Equations 4.7 through 4.10 is performed with an index s, requiring the single all-reduce operation for the incompletely transformed list (ia, jb_s). In other words, no worker-to-worker communication is necessary during the parallelization. Multiple integral screening is employed to reduce the operation cost effectively, and the innermost processing is done with DAXPY and DDOT routines.[22,23] When necessary, additional parallelization is also possible by i-batch in a combination with the s-index. From the prepared TEI list, the reference MP2 energy[17]

$$E^{MP2} = \Sigma_{ijab}\,(ia,jb)[2(ia,jb) - (ib,ja)]/(\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b) \tag{4.11}$$

is evaluated, and a couple of intermediate arrays (w_{ij} and w_{ab}) usable for all the CIS states[11,24] are computed in parallel at the cost of N^5. The Fock-like technique is utilized to construct the state-dependent intermediate array (w_{ia}).[24] The U_2 processing is completely dominated by the parallelized preparation of ($\underline{ia}$, jb), ($i\underline{a}$, jb), (ia, $\underline{jb}$), and (ia, $\underline{jb}$) in an ij-batch mode, where the underline means that the basis-orbital coefficients are transformed by CIS amplitudes[24] as

$$\underline{C_{pa}} = \Sigma_i C_{pi} b_i^a \tag{4.12}$$

$$\underline{C_{pi}} = \Sigma_a C_{pa} b_i^a \tag{4.13}$$

The cost to compute the U_2 contribution is considerably higher than for the T_2U_1 contribution because of the four types of TEI and also the state dependence.

The group of processers for parallelization is redefined for the excited states treatments of layer 2 after the FMO-HF stage is completed. For example, if 16 processors are available, the number of processors for each monomer (or dimer) might be defined as four, providing a four-by-four parallel usage in the FMO-HF calculations. Then 16 processors are gathered for the parallelized CIS and CIS(D) calculations. With such a utilization of computational resources, the MLFMO-CIS(D) job with basis set

of double-zeta plus polarization (DZP) quality[17] is enabled for photoactive proteins consisting of a few hundred residues, even on in-house clusters (e.g., cluster of Intel's multicore Xeon processors) within a few days.[16] Our fully direct implementations without I/O for any working quantities[8,12,13,22,23] could also be suitable for massively parallel supercomputers whose scratch I/O facilities should be considerably poor relative to their enhanced power of processors and networks. A preliminary trial was reported by using the Earth Simulator system with vectorized execution.[16]

4.2.3 Modification of CIS(D)

From the original CIS(D) proposal by Head-Gordon et al.,[11] various related developments have been derived until now, as briefly described below. Higher-order perturbation versions were proposed by Head-Gordon and Lee[25] and by Hirata,[26] where the computational scaling was increased to N^6. For the issue of near-degeneracy in excited states, iterative (or "perturb-then-diagonalize") variants were devised,[24,27] and they had some connection with the approximated second-order coupled cluster (CC2)[28] and the algebraic-diagrammatic construction through second-order (ADC(2)).[29] Laikov and Matsika[30] presented the CIS(2) method to simultaneously handle the ground and excited states with analytic energy gradients for geometry optimizations. Several flavors with spin-component scaling (SCS) for the spin-orbital expressions of CIS(D) have been developed by several groups.[31–34] These SCS versions have been shown to be useful, but care might be taken for the fact that they are based on empirically adjusted parameters.

On the basis of spin-adapted CIS(D) formulation,[13] we independently made a couple of modifications in which a partial renormalization (PR) of MP2 doubles amplitudes[35] and an extra inclusion of MP2 singles contribution is involved, keeping noniterative N^5 scaling of computation.[14] The CIS(Ds) with the latter correction was given with operator T_1 (defined through operator T_2 in a second-order wavefunction[25]) as

$$\omega^{\mathrm{CIS(Ds)}} = \langle \Psi_{\mathrm{CIS}} | V | U_2 \Phi_0 \rangle + \langle \Psi_{\mathrm{CIS}} | V | T_2 U_1 \Phi_0 \rangle_{\mathrm{c}} + \langle \Psi_{\mathrm{CIS}} | V | T_1 U_1 \Phi_0 \rangle_{\mathrm{c}} \qquad (4.14)$$

Recently, the self-energy[17] shift (SS) was included for the U_2 contribution (or relaxation energy), providing $\mathrm{CIS(D)_{SS}}$.[15] In that paper, it was notable that the $\mathrm{CIS(D)_{SS}}$ modification provides reasonable results for Rydberg states with keeping a balance of correction for valence states. The modified CIS(D) with PR and SS could incorporate some higher-order correlation effects, and these modifications would be favorable for the systems with near-degeneracy, as discussed on the dressed perturbation theories for the ground state.[35,36] The MLFMO calculations of CIS(Ds) and $\mathrm{CIS(D)_{SS}}$ with or without PR[14,15] have now been available in the ABINIT-MPX program with an integral-direct parallelism as well as nonmodified CIS(D).[13]

4.3 MODELING OF A TARGET SYSTEM

Total accuracy of MLFMO-CIS(D) type calculations[13–15] should depend on the geometrical adequacy of the photoactive proteins. Especially, if the crucial pigment geometry is deformed, the calculated results would be affected, even employing

highly correlated methods. Because the visible lights roughly correspond to a rather narrow range from 1.8 eV to 3.3 eV in transition energies, the errors in calculations should be maintained to (at least) less than 0.3 eV for quantitative discussion. It is thus important to make a proper computational model so as to obtain reliable excited states calculations. We will note a typical way of modeling, which we used in applications to photoactive proteins, in the following paragraph. The actual examples will be documented in Section 4.4.

The usual route to set up the target protein model starts with the downloading of a corresponding data set from the site of "Protein Data Bank" (PDB)[37]. There can, however, be many missing residues and (heavy) atoms in the naive data set archived as a certain PDB-ID, even by selecting the highest resolution case. By using some available modeling tools such as the "Molecular Operating Environment" (MOE)[38], the missing parts are first complemented. The linkages among residues (including Cys-Cys disulfide bond) are also checked and corrected. Next, the hydrogen atoms are attached to cap dangling bonds. For convenience in performing calculations, unimportant or redundant water molecules are deleted here. Then, the positions of attached hydrogen atoms are optimized in MM ways with appropriate force-field (FF) sets, where the geometries of heavy atoms might be fixed. We used the "Assisted Model Building with Energy Refinement" (AMBER)[39], version-99, as a default FF. After the fundamental modeling described above is finished, the pigment, neighbored residues, and water molecules as the chromophore components are subjected to detailed examination. Their charge states and networks of hydrogen bonding should be adjusted in a consistent way with experimental knowledge and chemical discussion. With overlay techniques, the PDB-based structure of a pigment part may be compared with that of an isolated pigment to which usual MO-based optimizations can be applied by using the standard programs, GAUSSIAN[40] or GAMESS.[6] If an unacceptable deformation is found in the fundamental PDB structure, the geometry of pigment would be replaced by inserting the optimized structure. In such a case, additional FF optimizations might be carefully performed for the chromophore. When needed, the geometries of neighbored residues and important water molecules would also be refined similarly. In summary, the actual preparation of target proteins has a sort of know-how, and the proper procedures should be taken on a case-by-case basis.

4.4 ILLUSTRATIVE APPLICATIONS

4.4.1 Photoactive Yellow Protein (PYP)

The PYP, which was isolated from the light sensor of *Ectothiorhodospira halophila*,[41] is water soluble and consists of 125 residues. Imamoto et al.[42] systematically revealed the PYP light-driven cycle. In that paper, Imamoto et al. reported that the pigment part of PYP is the deprotonated *p*-coumaric acid with the *trans*-form is linked by a thioester bond with Cys69 and also that the ground-state pigment absorbs 446 nm (2.78 eV) blue light and then starts the cycle of photoisomerization. The intense absorption in blue is responsible for the vivid yellow. The excited states of PYP had attracted theoretical studies (not cited here for simplicity), but only the modeled

pigments were frequently used without environmental effects in the protein. On the other hand, Yoda et al.[43] showed the importance of the positively charged Arg52 as the counterion to the phenolate part of pigment, though the calculations were of semi-empirical MO. They also pointed out the presence of a hydrogen-bonding network from the pigment to Glu46, Tyr42, and Thr50 (the farthest one).

The implemented MLFMO-CIS method[12] was applied to the PYP with the 6-31G basis set,[44] where the core orbitals were kept frozen at the CIS level: this FC restriction was imposed for all the post-HF calculations to be described in this section. The "3PYP" data set in PDB[37] was used with some MM refinements, but the examination of pigment geometry was unfortunately omitted. Layer 2 consisted of Arg52, Glu46, and Tyr42, beside the crucial pigment, *p*-coumarate linked with Cys69, according to Yoda's finding.[43] The lowest singlet MLFMO-CIS value was as high as 4.28 eV, although the state characterization was successfully made as the HOMO-LUMO transition of π-π^* in the coumarate moiety with a high oscillator strength of 1.52. The overestimation was 1.5 eV from the experimental value of 2.78 eV,[42] which would be remedied by the CIS(D) correction.[11,13] This MLFMO-CIS application, to the author's knowledge, was the first attempt to calculate the excited state in a realistic photoactive protein in a full QM manner in that regard.[12,45]

As expected, the MLFMO-CIS(D)/6-31G procedure[13] on the PYP presented a remedied value of 3.29 eV. The relaxation (U_2) energy and differential correlation (T_2U_1) energy were –0.1430 au and 0.1066 au, respectively, and thus the former was shown to be effective in gross energy lowering. A supplemental calculation with the 6-31G* basis set[44] provided almost the same value of 3.30 eV, meaning that the balance of cancellation is not much affected although the amounts of both relaxation energy and differential correlation energy are increased by the improvement from 6-31G to 6-31G* with polarization d-functions.

The remaining error of 0.5 eV in comparison with the experimental value[42] could be attributed primarily to the issue of "3PYP" geometry being used,[13] through the following examination. Nielsen et al.[46] measured the absorption spectra of the deprotonated *trans*-thiophenyl-*p*-coumarate, pCT^-, which is an experimental model chromophore of PYP, by a heavy-ion storage technique named electrostatic ion storage ring at aarhus (ELISA). They reported the absorption maximum at 460 nm or 2.70 eV, being slightly lower than the PYP value of 2.78 eV[42]. We directly checked the validity of CIS(D) with this free pCT^- molecule.[13] Table 4.1 shows the corresponding results. The CIS(D) value at the MP2 geometry is in reasonable agreement with the observed value of 2.70 eV,[46] relative to the HF case. This is consistent with the nature that the CIS(D) method relies on the MP2 theory.[11,17] The use of polarized 6-31G* basis[44] yields slightly higher energy than that of 6-31G. A critical dependence of excitation energies on the geometry has just been demonstrated by using the pCT^- model of PYP chromophore, indicating an inherent accuracy of 0.3 eV in CIS(D) calculations. We next compared the bond lengths of the phenolate part between the MP2/6-31G geometry of pCT^- and the "3PYP" data set of PDB[37] and found an average difference of 0.05 Å for the lengths between heavy atoms,[13] where a representative Ph-O^- length was optimized by the MO calculation as 1.295 Å but the corresponding value in PDB was 1.345 Å. The geometrical inadequacy of the "3PYP" pigment was obvious, and this PYP case became a good lesson.[13] Chiba et al.[20]

TABLE 4.1
Dependence of Excitation Energies (eV) for pCT⁻ Molecule

Basis	Geometry	CIS	CIS(D)
6-31G	HF	4.03	2.95
	MP2	3.87	2.75
6-31G*	HF	4.13	3.08
	MP2	4.01	2.94
Expt.			2.70

**Note:* The experimental value was observed by ELISA by Nielsen et al. (Nielsen, I.B., Boyé-Peronne, S., El Ghazaly, M.O.A. et al., *Biophys. J.* 89:2597–2604, 2005). Geometry was optimized by GAUSSIAN (see Foresman, J.B., Frisch, A., *Exploring Chemistry with Electrononic Structure Methods,* 2nd ed. Pittsburgh, PA: Gaussian Inc., 1996).

inserted the DFT-optimized pigment part into the chromophore of PYP and then made the TDDFT(LC) calculations in the FMO scheme with and without a polarizable continuum model (PCM) to mimic the hydrated situation (refer to Chapter 5 for details). As a final remark, the geometry assessment and refinement are highly desired, particularly for the photoactive pigment.

4.4.2 DsRed

From Tsien's pioneer work on the GFP derived from *Aequorea victoria* jellyfish, GFP-like photo-functional proteins with different colors have been developed in diverse ways. One such protein is red fluorescent protein (RFP), which was first isolated by Matz et al.[47] from *Discosoma* coral. Its characteristic peaks of excitation energy and emission energy are 558 nm (or 2.22 eV) and 583 nm (2.13 eV), respectively. This RFP has thus been called drFP583 or DsRed (trademark by CLONTECH). It was revealed that DsRed is actually a tetramer, and each monomer has a form of β-barrel in which the pigment is located at the center of barrel.[47–50] The crucial pigment of the DsRed chromophore is now considered as the deprotonated *p*-hydroxybenzilidene-imidazolinone moiety with an elongated π-conjugation tail, which is autocatalytically formed from three residues of Gln66-Tyr67-Gly68 through maturation.[48,49]

Figure 4.1 illustrates the chemical structure of pigment. Gross et al.[48] discussed the nature of electron delocalization involving an oxidized peptide bond toward Phe65 and also reported the TDDFT excitation energy of 2.33 eV for a model pigment. An importance of electrostatic interaction between the anionic pigment and charged residues was pointed out by Yarbrough et al.,[49] by analyzing their X-ray structure. Meanwhile, Nielsen's group[51] observed the gas-phase absorption spectra for several model pigments gradually elongating the conjugation tail, with the ELISA apparatus as in the case of pCT⁻ for PYP.[46] They discussed that the anionic state of pigment is responsible for a long wavelength in the red region, based on the observed spectra and the estimation of HOMO-LUMO gap.

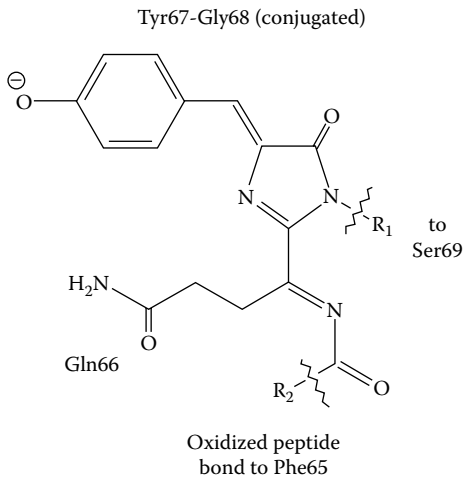


FIGURE 4.1 Chemical structure of a DsRed pigment part. (See text and Gross, L.A., Baird, G.S., Hoffman, R.C., Baldridge, K.K., Tsien, R.Y., *Proc. Natl. Acad. Sci. USA* 97:11990–11995, 2000; Yarbrough, D., Wachter, R.M., Kallio, K., Matz, M.V., Remington, S.J., *Proc. Natl. Acad. Sci. USA* 98:462–467, 2001.)

In our investigation of DsRed[16] by the MLFMO-CIS(D) method,[13] the anionic model pigments,[51] some of whose spectra were available as the reference data, were first examined by separate CIS(D) calculations. The observed excitation energies of models called as "*GFP,*" "*RFP(1),*" and "*RFP(2)*" were 2.59 eV, 2.38 eV, and 2.26 eV, respectively, where the conjugation was elongated in this order: see the chemical structures in the original paper of ELISA experiment.[51] The geometry optimization with GAUSSIAN[40] provided the planar structure of the photoactive π-conjugation part throughout "*GFP,*" "*RFP(1),*" and "*RFP(2).*" The excitation of interest was characterized as the HOMO-LUMO transition of π-π* type with a large oscillator strength of 1.6 in the xy-plane at the CIS level. On the MP2/6-31G* geometries,[40,44] the CIS(D)/6-31G* procedure yielded 2.87 eV, 2.62 eV, and 2.42 eV for these three pigment molecules.[16] Good correspondence between Boyé's experimental values[51] and the CIS(D) estimates have just been verified, expecting a similar accuracy for more realistic "*Model A*" and "*Model B*" with the oxidized peptide bond.[51]

The fundamental structure "1ZGO" (high resolution of 1.4 Å) available from the PDB server[37] was adopted to evaluate the excitation energy of DsRed, where the first monomer chain was extracted from the tetramer set. It has been known that the affect of tetramerization in excitation energy is as small as at most 0.01 eV,[50] being consistent with the central location of pigment inside a β-barrel. The usage of monomer could thus be justified. The standard modeling protocols (described in the previous section) were taken, where two water molecules nearest the pigment were retained. The total number of atoms was 3553.

Figure 4.2 is a computer-graphics (CG) presentation of the target protein in solid-ribbon fashion for this monomer, illustrating the barrel shape and the central location

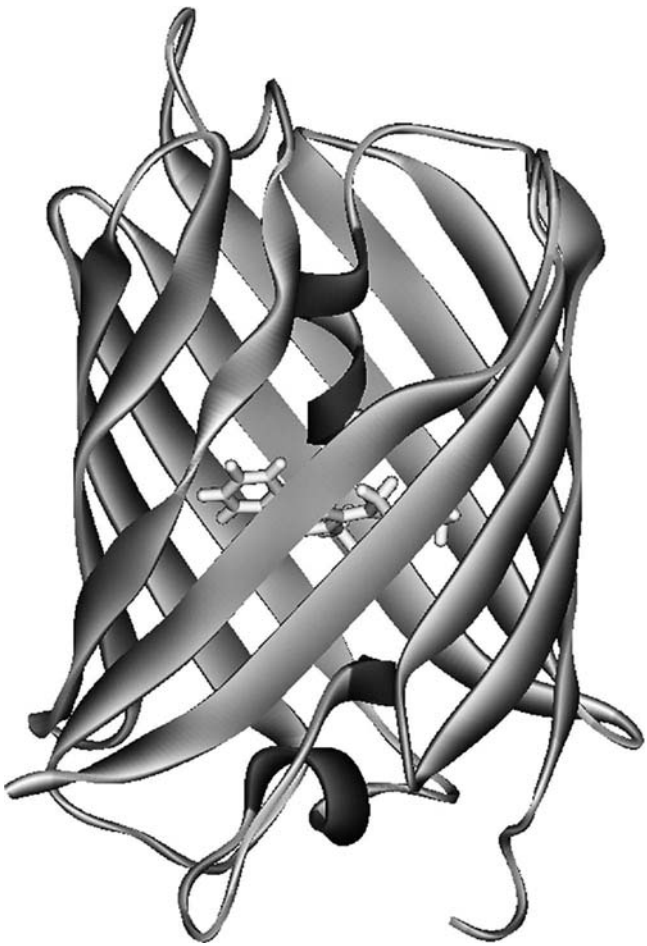


FIGURE 4.2 β-Barrel structure of a DsRed monomer. The pigment moiety is located at the center of the barrel.

of the pigment.[1,47,49,50] To check the adequacy of geometry for MLFMO-CIS(D) calculations,[13] we superimposed the MP2/6-31G* geometry[40,44] of "*Model B*"[51] on the pigment frame, termed "CRQ" originally in the "1ZGO" set in PDB,[37] through CG-based overlay techniques. We then found the structural coincidence for the crucial conjugation part. A representative bond length of Ph-O^- was obtained to be 1.253 Å by the MP2/6-31G* optimization on "*Model B,*" and it was in agreement with the corresponding X-ray value of 1.244 Å. The geometrical adequacy of pigment has just been shown, implying that the reliability of CIS(D) calculations for the model pigments[51] is transferable for DsRed.

In order to estimate the emission energy of DsRed, the chromophore geometry should be modified for the excited state, especially about the pigment. We thus optimized the geometry of "*Model B*"[51] by the CIS/6-31G* procedure[10,44] with GAUSSIAN[40] for the

lowest HOMO-LUMO state, by adopting a simple recipe proposed by Das et al.[52] for the GFP pigment models. It was observed that the conjugation part of "*Model B*" is slightly elongated in the excited state. Then, the naive pigment geometry of "1ZGO" was replaced by the CIS-optimized geometry, based on superimposing techniques. This import was followed by relaxation operations with FF under freezing the pigment geometry. The resulting new structure set was termed "$1ZGO_{(Ex)}$."

The following three settings were employed for layer 2 of DsRed: (a) only the pigment moiety or "CRQ"; (b) adding Phe65 ("F+CRQ"); and (c) adding also Ser69 ("F+CRQ+S"). A series of MLFMO-CIS(D) calculations[13] were performed by using the 6-31G and 6-31G* basis sets.[44] The total number of fragments was 220. Although the number of 6-31G* functions was as many as 30,443, such large-scale MLFMO computations were easily performed by our ABINIT-MPX program on an in-house Intel Xeon cluster with 20 cores, as addressed later.

For the excitation energy, we would start our discussion with the MLFMO-CIS(D)/6-31G results,[16] compiled in Table 4.2. At the CIS level, the calculated energies of "CRQ," "F+CRQ," and "F+CRQ+S" settings are 3.43 eV, 3.35 eV, and 3.33 eV, respectively. These CIS values are far from quantitative in comparison with the experimental value of 2.22 eV.[47] Certainly, they are remedied through the CIS(D) correction. It is notable here that the setting of pigment-only or "CRQ" looks insufficient relative to that of "F+CRQ" case providing a good estimate (2.52 eV versus 2.30 eV). The large difference of 0.22 eV in CIS(D) energies between "CRQ" and "F+CRQ" could be attributed to the contribution from orbital relaxation energy, indicating an importance of "margin" in layer 2: recall that Phe65 involves the oxidized peptide bond of C=N-C=O.[48,49] The effect of including Ser69, which does not commit the conjugation, is consistently small.

Figure 4.3 illustrates the MLFMO-CIS natural orbital (NO) pair of hole and particle[12] with the setting of "F+CRQ." Their natural occupations are 1.08 for hole and 0.92 for particle. Immediately, the π-π* characters of these NOs can be seen, where the extension of lobes to Phe65 is also visible. In Figure 4.2 and Figure 4.3, the conjugation system of pigment is essentially placed on the xy-plane and the

TABLE 4.2
Excitation Energies (eV) for DsRed

Layer 2	CIS	CIS(D)	Ratio
CRQ	3.43	2.52	0.73
F+CRQ	3.35	2.30	0.69
F+CRQ+S	3.33	2.29	0.69
Expt.		2.22	

Note: 6-31G basis set (see Foresman, J.B., Frisch, A., *Exploring Chemistry with Electronic Structure Methods,* 2nd ed. Pittsburgh, PA: Gaussian Inc., 1996) was used for MLFMO-CIS(D) calculations. Experimental energy was observed by Matz et al. (Matz, M.V., Fradkov, A.F., Labas, Y.A. et al. *Nature Biotech.* 17:969–973, 1999).

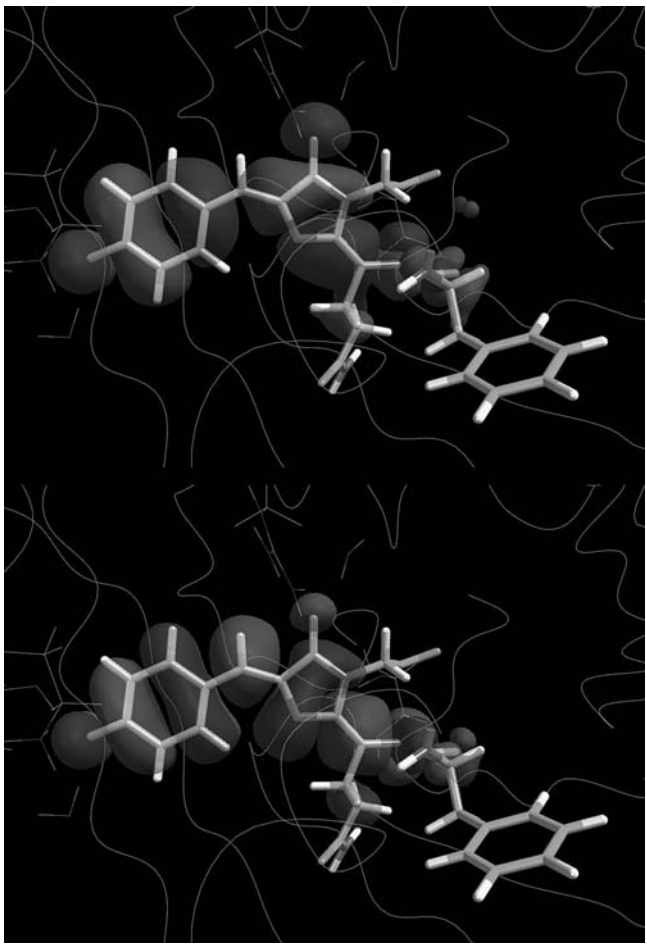


FIGURE 4.3 Hole (upper) and particle (lower) natural orbitals of MLFMO-CIS/6-31G calculation for DsRed. **(See color insert following page 117.)**

aperture of β-barrel is oriented along the z-axis in our processed "1ZGO" data set. As already denoted, the oscillator strength is large for the in-plane electronic transition. Yarbrough et al.[49] pointed out an antenna-like rectangular array of tetramerized structures of DsRed proteins to efficiently absorb/emit the light. Our MLFMO investigation has supported this speculation from the experimental side.

Table 4.3 lists the MLFMO-CIS(D)/6-31G* results of excitation energy. Except for the smallest "CRQ" setting of layer 2, "F+CRQ" and "F+CRQ+S" settings provide the excitation energies of 2.30 eV and 2.28 eV, respectively, corresponding well to the experimental value of 2.22 eV.[47] The 6-31G* values are almost the same as the 6-31G values shown in Table 4.2. As in the previous case of PYP,[13] we checked the amounts of both relaxation energy and differential MP2 correlation energy and found that the balance of cancellation is not much affected by the basis set enlargement from 6-31G

TABLE 4.3
Excitation Energy and Emission Energies (eV) for DsRed

Layer 2	CIS	CIS(D)	Ratio
Excitation			
CRQ	3.35	2.49	0.74
F+CRQ	3.27	2.30	0.70
F+CRQ+S	3.26	2.28	0.70
Expt.		2.22	
Emission			
CRQ	3.25	2.41	0.74
F+CRQ	3.20	2.21	0.69
F+CRQ+S	3.18	2.21	0.70
Expt.		2.13	

Note: 6-31G* basis (see Foresman, J.B., Frisch, A., *Exploring Chemistry with Electronic Structure Methods,* 2nd ed. Pittsburgh, PA: Gaussian Inc., 1996) was used for MLFMO-CIS(D) calculations. Structure of "$1ZGO_{(EX)}$" was employed for emission energy. Experimental energy was observed by Matz et al. (Matz, M.V., Fradkov, A.F., Labas, Y.A. et al. *Nature Biotech.* 17:969–973, 1999).

to 6-31G* favorably. The use of 6-31G basis set could be a convenient option for a survey of various photoactive proteins, as long as such exploratory calculations are followed with polarized basis sets[17] (at least 6-31G*) for final discussion.

The emission energies obtained with the "$1ZGO_{(EX)}$" structure are also given in Table 4.3. The MLFMO-CIS(D)/6-31G* energy is 2.21 eV for both "F+CRQ" and "F+CRQ+S" settings, and this value is again compared well with the experimental energy of 2.13 eV.[47] It is an encouraging fact that the redshift from the absorption (excitation) energy to emission energy (or Stokes shift) is reproduced by our calculations for DsRed.[16] The import technique by adopting the CIS-optimized geometry[10,44,52] of pigment could be a practical recipe to evaluate the emission energy of photoactive proteins. The DsRed case has demonstrated that the electronic transition energies could be quantitatively calculated by the MLFMO-CIS(D) method[13] as far as the appropriate molecular structure is prepared for the target protein.

Here, we would note the actual timings for DsRed calculations. The MLFMO-CIS(D)/6-31G job with "F+CRQ" setting required 20.2 hours on a cluster with 20 Xeon cores (2.8 GHz clock-rate and 2 GB memory per core), including the preceding FMO-HF stage. The same job took 34.5 minutes on 1024 vector processors of the Earth Simulator system. This was a promising example to show a potential of our ABINIT-MPX for massively parallel-vector machines: that timing[16] should, however, be rather preliminary because systematic tunings of the vectorization have been recently underway.[53] For the 6-31G* case, the timings to complete "F+CRQ" and "F+CRQ+S" settings were 76.0 hours (3.2 days) and 130.7 hours (5.4 days), respectively, on the Xeon cluster. These timings have proved a practical applicability of our

MLFMO-CIS(D) approach[13] to realistic photoactive proteins, even by using cluster computers. When memory space per core is enlarged, the computational time could be reduced because of fewer batches of the integral-direct processing in CIS(D).[13]

In the paper of DsRed,[16] we presented the analyses of interfragment interaction energy (IFIE)[8,45,54,55], which is closely related to the pair interaction energy (PIE)[56] in GAMESS[6] (refer to Chapter 2), since Yarbrough et al.[49] speculated the important residues having electrostatic interactions with the anionic pigment moiety. Our IFIE results (at the HF level) concurred with their speculation, illuminating that the positively charged Lys163, Lys70, and Arg52 are outstanding in electrostatic stabilization, whereas Glu148 and Glu215 have considerable destabilizations. Tsien's group[1,48,57] reported several X-ray structures[58] of monomeric mutants from DsRed, where these mutants were called as mFruits according to a variety of colorful emitting lights named after edible fruits. They also experimentally investigated that the color tunings could be controlled in a delicate way by buried charges and pH.[58] The mFruits is the topic in the next subsection.

4.4.3 mFruits

Recently, we completed a series of MLFMO-CIS(D)/6-31G* calculations,[59] including a couple of higher extensions,[14] for the excitation energies of three mFruits proteins whose structures were determined by X-ray experiments.[58] They are mCherry, mStrawberry, and mOrange, and the corresponding PDB-IDs[37] are "2H5Q," "2H5P," and "2H5O," respectively. The mutations concern not only the pigment part but also the neighbored residues, as summarized below. The Tyr67-Gly68 part that forms the π-conjugation system is retained as in DsRed, but the terminal Gln66 is replaced by Met in mCherry and by Thr in mStrawberry and mOrange.[57,58] In mOrange, the oxidized peptide bond is altered by making a five-membered ring toward Phe65 during an additional maturation, and some apparent structural modification by covalent bond is thus caused.[58] Whereas Lys163 in DsRed should play a primary role in stabilizing the negative charge of pigment moiety (more specifically the phenolate oxygen atom),[16,49] it is replaced by Gln in mCherry and by Met in mStrawberry and mOrange.[58] Similarly, Lys83 is changed in these mFruits. The deprotonated Glu215 is a main source of destabilization in DsRed.[16,49] In contrast, it is protonated (or neutralized) in mCherry and mStrawberry, where the attached proton position directed to the nitrogen atom in the imidazolinone ring is of special interest.[58] The biochemical differences in these three mFruits from DsRed yield the shifted excitation (emission) energies. The peak values in spectra are as follows: 2.11 (2.03) eV for mCherry, 2.16 (2.08) eV for mStrawberry, and 2.26 (2.21) eV for mOrange.[57,58]

Prior to treating mCherry, mStrawberry, and mOrange,[57,58] the excitation energy and emission energy of DsRed[47] were reevaluated by using the PR-CIS(Ds) extension.[14] Table 4.4 shows the results, where the CIS and CIS(D) values[16] are included for comparison. The inclusion of higher-order correlations precisely improves the estimated values for DsRed. That is to say, our best estimates are coincident with the experimental peak energies of both excitation and emission.[47] For the cross check, the model pigments of "*RFP(1)*" and "*RFP(2)*"[51] were also recalculated at the PR-CIS(Ds)/6-31G* level. The excitation energies were then obtained as 2.52 eV

TABLE 4.4
Excitation Energy and Emission Energies (eV) for DsRed

Layer 2	CIS	CIS(D)	CIS(Ds)	PR-CIS(D)	PR-CIS(Ds)
Excitation					
CRQ	3.35	2.49	2.54	2.36	2.41
F+CRQ	3.27	2.30	2.37	2.18	2.24
F+CRQ+S	3.26	2.28	2.34	2.16	2.22
Expt.					2.22
Emission					
CRQ	3.25	2.41	2.45	2.28	2.32
F+CRQ	3.20	2.21	2.27	2.09	2.15
F+CRQ+S	3.18	2.21	2.26	2.09	2.14
Expt.					2.13

Note: 6-31G* set (see Foresman, J.B., Frisch, A., *Exploring Chemistry with Electronic Structure Methods,* 2nd ed. Pittsburgh, PA: Gaussian Inc., 1996) was used for MLFMO-CIS(D) calculations with higher-order contributions (see Mochizuki, Y., Tanaka, K., *Chem. Phys. Lett.* 443:389–397, 2007). The structure of "$1ZGO_{(EX)}$" was employed for emission energy. Experimental energy was observed by Matz et al. (Matz M.V., Fradkov, A.F., Labas, Y.A. et al. *Nature Biotech.* 17:969–973, 1999).

for the former and 2.29 eV for the latter. These PR-CIS(Ds) results agree with the experimental values[51] of 2.38 eV of "*RFP(1)*" and 2.26 eV of "*RFP(2)*" satisfactorily: refer again to the corresponding CIS(D) results in the previous subsection. Similar accuracy would be expected for three mFruits,[57,58] because the crucial conjugation part of pigment (formed from Tyr67-Gly68) is common.

We were interested in the color tuning situation in mCherry, mStrawberry, and mOrange,[58] by comparing with the DsRed case.[16,47,49] Thus, the excitation energies of these mFruits were evaluated as the first step.[59] The IFIE analyses[8,45,54,55] were employed to reveal the differences in pigment-residue interactions of chromophore. The PDB data sets of "2H5Q" (mCherry), "2H5P" (mStrawberry), and "2H5O" (mOrange)[37,58] were used as the fundamental structures of molecular modeling for the MLFMO-CIS(D)/6-31G* calculations with higher-order contributions.[14] Shu et al.[58] noted a possibility of some deformation in the X-ray structure for mOrange because of unprecedented covalent bond modifications. The networks of hydrogen bonding in chromophore should be more important in mFruits[57,58] than in DsRed. All water molecules within 5.5 Å from the pigment moiety were thus preserved, where the numbers of corresponding water molecules were 11 for mCherry, 8 for mStrawberry, and 10 for mOrange, respectively. Standard protocols such as the hydrogen attachments and relaxations were then performed. In the modeling of mCherry, there was a special manipulation that the geometries of Glu215 and some important water molecules were carefully optimized by GAUSSIAN[40] at the MP2/6-31G* level[44] and

TABLE 4.5
Excitation Energies (eV) for DsRed and Three mFruits

Protein	CIS	CIS(D)	CIS(Ds)	PR-CIS(D)	PR-CIS(Ds)	Expt.
DsRed	3.26	2.28	2.34	2.16	2.22	2.22
mCherry	3.24	2.27	2.31	2.13	2.18	2.11
mStrawberry	3.04	2.22	2.27	2.07	2.12	2.16
mOrange	3.48	2.56	2.58	2.43	2.45	2.26

Note: 6-31G* basis (see Foresman, J.B., Frisch, A., *Exploring Chemistry with Electronic Structure Methods,* 2nd ed. Pittsburgh, PA: Gaussian Inc., 1996) was used for MLFMO-CIS(D) calculations with higher-order contributions (see Mochizuki, Y., Tanaka, K., *Chem. Phys. Lett.* 443:389–397, 2007). Experimental values were taken from Matz et al. (Matz, M.V., Fradkov, A.F., Labas, Y.A. et al., *Nature Biotech.* 17:969–973, 1999) for DsRed and Sharner et al. and Shu et al. (Sharner, N.C., Cambell, R.E., Steinbach, P.A. et al., *Nature Biotech.* 22:1567–1572, 2004; Shu, X., Shaner, N.C., Yarbrough, C.A., Tsien, R.Y., Remington, S., *Biochem.* 45:9639–9647, 2006) for mFruits.

then they were imported. This is because the attached proton of Glu215 should have a special importance in the color tuning through the hydrogen bonding to the nitrogen atom in the imidazolinone ring of the conjugation part.[58]

For simplicity, the excitation energies obtained with "F+X+S" setting ("X" means the pigment consisting of three residues) are compiled in Table 4.5, though other calculations of "X" and "F+X" have been performed for three mFruits proteins, of course.[59] The PR-CIS(Ds) treatment provides good overall agreement with the experimental excitation energies,[57,58] as expected. For mCherry, we made one computational trial that the proton position on Glu215 is directed away from the imidazolinone ring, and then the estimated value was worsened to be 2.27 eV. Another trial for mCherry was the deprotonation of Glu215 as in DsRed,[49] yielding poorer 2.33 eV. These results just support a speculative discussion on the protonated situation of Glu215 in mCherry from the experimental observations.[58] Similar examinations were performed also for mStrawberry and mOrange consistently. Errors observed for mOrange are large relative to those for mCherry and mStrawberry. A possible reason might be that a deformation remained in the X-ray structure of mOrange, as denoted above.[58]

Figure 4.4 is an illustration of the IFIE energies obtained by both HF and MP2 calculations (with 6-31G* basis[44]) for three representative residue positions: "X" setting was adopted for analysis purpose. Experimental discussions associated with the mutations in mFruits[57,58] are clearly justified by our FMO calculations.[59] The electrostatic interaction should be dominant in color tuning for DsRed.[16,49] On the contrary, the role of electrostatic interaction is suppressed by the missing of Lys163 and the protonation of Glu215 for mCherry and mStrawberry, except for Lys70, Arg95, and Glu148, which were kept unchanged even in mFruits.[57,58] Finally, we would note a recent timing that the MLFMO-CIS(D)/6-31G* job (with "F+X+S" setting and higher-order contributions[14]) for a mFruits protein was completed with 3.0 days on

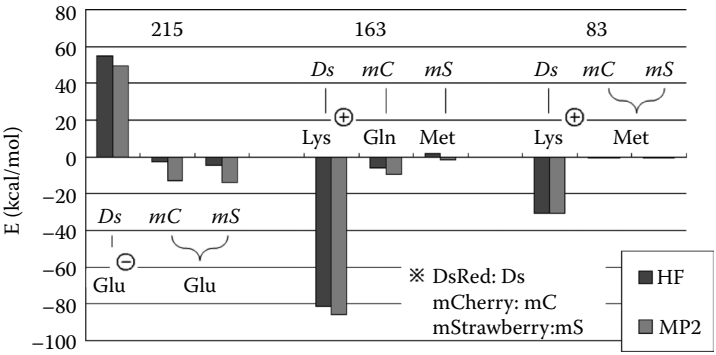


FIGURE 4.4 Interfragment interaction energy (IFIE) results of three representative residues for DsRed, mCherry, and mStrawberry. Calculations were performed at both HF and MP2 levels with 6-31G* basis. Setting of "X" (or pigment only) was used for analysis purposes. (See Foresman, J.B., Frisch, A., *Exploring Chemistry with Electronic Structure Methods,* 2nd ed. Pittsburgh, PA: Gaussian Inc., 1996.)

a commodity cluster computer of Intel Core 2 Duo (2.93 GHz clock-rate and 4 GB memory per core) 16 cores.

4.5 OTHER WORKS AND FUTURE DIRECTIONS

Up to this point, we have described the MLFMO-CIS(D) approach[12–15] and also associated applications to realistic photoactive proteins.[13,16,59] The feasibility and reliability of our FMO-based method in a computational biochemistry of excited states have thus been demonstrated. However, for the reader's convenience, it may be valuable to address the pioneering works by Nakatsuji's group in that field, by utilizing their original cluster expansion theory, called SAC-CI,[60] some of which are introduced here. The photosynthetic reaction center was investigated through a careful modeling to make the tough problem manageable,[61,62] a decade ago, providing an outstanding theoretical work, to date. Recently, Nakatsuji's group reported extensive studies on the bioluminescence of firefly luciferin[63] and the color-tuning in retinal protein,[64] both of which should be helpful in understanding the residue-specific interactions of chromophore in a quantitative fashion. A sort of QM/MM scheme was employed in these works,[63,64] and thus the FMO calculations could provide the cross-reference data for such important biochemical targets. The emission energy of firefly bioluminescence has just been successfully estimated by the MLFMO-CIS(D) calculations on the whole luciferase containing more than five hundred residues.[65,66] For the problem of retinal, the excitation energies of Schiff-base models in the gas phase[67] were evaluated at the PR-CIS(D)$_{SS}$/6-31G* level,[14,15] and the observed energy of 2.03 eV was almost reproduced with the calculated value of 2.05 eV for the "*11-cis*" model of rhodopsin. We have a plan to do a systematic investigation of retinal proteins, in which the binary correction recipe, proposed initially by Hirata et al.[68] for molecular clusters and followed by Chiba et al.[20] for TDDFT(LC) under the FMO scheme, may be employed to take a number of contributions from surrounding residues into account.

Next, we should consider the configurational fluctuation of protein or DNA in realistic conditions under finite temperature. The method of molecular dynamics (MD) simulation is a straightforward option to sample a series of fluctuating configurations, where the classical (or MM) forces are usually used from a viewpoint of computational cost. The hydration effect could also be easily incorporated in MD. There were two papers to be cited along this line. Kawaguchi and Yamato[69] reported a reasonable estimate (with multireference perturbation calculations) of 2.88 eV for the excitation energy of PYP (2.78 eV),[42] by averaging 10 MD-generated configurations of the chromophore geometry. Meanwhile, Valiev and Kowalski[70] obtained the lowest π-π and n-π* energies of cytosine moiety in the hydrated model DNA through a hybrid approach of CCSD excited states calculation and MD sampling. We also employed a multiple structure sampling of nuclear receptor protein, though the main interest was focused on the residue–ligand interactions in the ground state.[71] MD simulations with full QM forces would be ideal to exclude empirical features, but their executions should be too demanding for realistic proteins consisting of a few hundred residues, even if the FMO-HF gradient technique[72] is utilized as FMO-MD[73] with the parallelized executions on currently available resources of computation. We have, however, been promoting the FMO-MD simulations as preparation for such applications in future. In recent years, Komeiji et al. reimplemented the FMO-MD method with various improvments[74,75] and applied it to two realistic problems of condensed phase[76,77] (refer also to Chapter 6). One was the statistical estimation of blueshift in the lowest n-π* energy of hydrated formaldehyde.[76] Another was the investigation of diversity in reaction paths for hydrolysis of methyl diazonium ion.[77] Because the MLFMO-CIS(D) calculations[13] were used in the former application, the work will be summarized in the paragraphs below.

It has been well known that the excitation energies of n-π* states in carbonyl compounds are blueshifted as solvatochromism.[2] The hydrated formaldehyde molecule has been a pet system for theoretical calculations including statistical treatment, but sets of empirical MM parameters were usually involved in these studies, unfortunately. We thus intended to perform the first full QM simulation.[76] A droplet model with 128 water molecules was used for FMO-MD runs at the HF/6-31G level[44] (300 K). Four hundred samples were selected from about two thousand generated configurations and subjected to the MLFMO-CIS(D)/6-31G* calculations.[13] Figure 4.5 shows the histograms of excitation energies obtained by the CIS and CIS(D) treatments where six nearest water molecules were set in layer 2. A Gaussian-like feature can be seen for both histograms, although the overestimation of CIS energies is also obvious. The MLFMO-CIS(D) energies of hydrated condition were averaged to be 4.22 eV (standard deviation 0.15 eV), and it was compared with the corresponding gas-phase value of 4.08 eV (0.16 eV). As the result, the blueshift of 0.14 eV was estimated without any empirical parameters: see the comparative discussion in the paper of hydrated formaldehyde.[76]

Absolutely, MD simulations, regardless of with MM forces or QM forces, would be preferable for statistical discussion of electronic properties which should be obtained in QM ways. In fact, we expect that FMO-MD simulations for proteins and other large molecular systems will be tractable on the next-generation (peta-class)

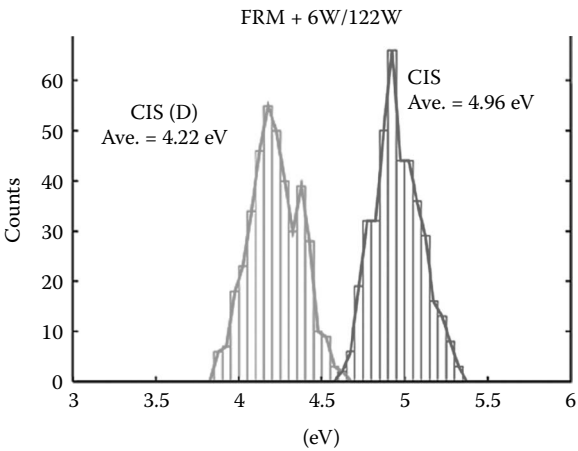


FIGURE 4.5 Histograms of configuration interaction singles (CIS) and CIS with second-order perturbative double [CIS(D)] energies (eV) for the lowest n-π* excitation of hydrated formaldehyde molecule mimicked by a droplet model. Configurations were generated by FMO-MD simulation, and excitation energies were obtained by MLFMO-CIS(D)/6-31G* calculations (see text).

supercomputers equipping several tens of thousands processor cores after several years.[75] Nowadays, however, detailed discussion based on the single "optimal structure" should still be valuable in quantitatively estimating the peak energy positions of observed spectra of target systems, as documented in Section 4.4. In other words, the geometry optimization at least for the essential chromophore should be important currently. Nakano et al. already implemented full or partial geometry optimizations with the FMO-HF gradient[72] in the ABINIT-MPX program. This optimization ability could make the modeling without MM/FF parameters possible, but the resulting structure could not be necessarily better than the empirically optimized one because the dispersion interactions are not incorporated at the HF level. Preliminary tests of FMO-HF geometry optimization showed that the results are reasonable for hydrogen-bonding systems (thus, HF forces are usable for water droplets in FMO-MD),[74–77] whereas the procedure fails for dispersion-dominant systems with such CH/π and π/π interactions. Manifestly, our FMO-MP2 calculations[22,23,53] should be extended for the geometry optimizations of proteins. The corresponding implementation with parallelism has been underway (since Summer 2008), by modifying the direct MP2 gradient method.[78] The geometry refinement for chromophore will soon be enabled at the FMO-MP2 level.

The gradient calculation of the excited states should be useful as well as that of the ground state. Actually, the excited state geometry is needed to evaluate an emission energy.[16,52] Foresman et al.[10] originally derived the set of analytical equations for CIS gradient calculations and implemented it to optimize the excited states geometries of several molecules. The qualitative accuracy of CIS geometries had been accepted. Ishikawa and Head-Gordon[79] presented the CIS(D) gradient formula.

Unlike the success of CIS(D) energy correction,[11] the CIS(D) gradient was shown to be unreliable relative to the CCSD response gradient and even, unfortunately, to the CIS gradient.[80] Hättig[81] discussed the usability of single-reference correlated methods in optimizing the excited states geometries. He also pointed out that the region near conical intersections could not be properly handled unless the correlated method with symmetric energy matrix is employed. The conical intersection has attracted great interest in the research of excited states dynamics.[82] For example, it should involve the radiationless decay to prevent the formation of thymine dimer in UV-irradiated DNA. Laikov and Matsika[30] devised a promising method, CIS(2), whose second-order correlated matrix is designed to be symmetric. Furthermore, they implemented the CIS(2) gradient and optimized the geometries of cytosine and uracil molecules, even at the conical intersection point. The CIS gradient[10] and also the correlated CIS(2) gradient[30] will be implemented in the ABINIT-MPX program.

4.6 CONCLUSION

In the present chapter, we reviewed our method[12,13] of configuration interaction singles (CIS[10]) with second-order perturbative doubles [CIS(D)][11] in conjunction with the multilayer version of fragment molecular orbital (MLFMO) scheme[5] and also a variety of actual applications to photoactive biomolecular systems such as fluorescent proteins (DsRed)[16] and (mFruits).[59] An integral-direct parallelism has been adopted in the CIS and CIS(D) implementations in our experimental FMO program ABINIT-MPX, and this adoption enabled the practical calculations of realistic proteins with a few hundred residues, even on in-house class cluster systems. Several modified CIS(D) treatments incorporating higher-order effects[14,15] have been available in the ABINIT-MPX. By using our MLFMO-CIS(D) approach, the electronic transition energies would be evaluated with accuracy of 0.1 eV, if reliable molecular structures are prepared.[16,59] Furthermore, the color tuning mechanism in the chromophore could be analyzed in terms of the interfragment interaction energy (IFIE).[8,45,54,55] Future directions of applications and related methodological developments were also addressed.

ACKNOWLEDGMENTS

The authors thank Kiyoshi Tanaka, Kaori Fukuzawa, Shinji Amari, Takeshi Ishikawa, and Yuto Komeiji for collaborations. YM and TN would be grateful to Katsumi Yamashita and Tadashi Murase for professional technical support. Minoru Sakurai is also acknowledged for discussion and encouragement. The works reported here were supported primarily by the CREST project operated by the Japan Science and Technology Agency (JST) and partially by the "Revolutionary Simulation Software for the 21st Century" (RSS21) project operated by the Ministry of Education, Culture, Sports, Science and Technology (MEXT). Additionally, YM would like to acknowledge a Grant-in-Aid for Scientific Research on Priority Area "Molecular Theory for Real Systems."

NOTE REGARDING THE AVAILABILITY OF ABINIT-MPX

The ABINIT-MPX program has not yet been opened, because both the new developments of theoretical methods and the code improvements for efficiency have been made continuously. Namely, ABINIT-MPX is an ongoing version currently. If, however, you are interested in the trial usage, please contact us via e-mail. Binary versions for Intel processor–based cluster machines might be available upon request.

REFERENCES

1. Tsien R.J. 1998. The green fluorescent protein. *Annu. Rev. Biochem.* 67: 509–544.
2. Reichardt C. 1994. Solvatochromic dyes as solvent polarity indicators. *Chem. Rev.* 4: 2319–2358.
3. Kitaura K., Sawai T., Asada T., Nakano T., Uebayasi M. 1999. Pair interaction molecular orbital method: an approximate computational method for molecular interactions. *Chem. Phys. Lett.* 312: 319–324.
4. Kitaura K., Ikeo E., Asada T., Nakano T., Uebayasi M. 1999. Fragment molecular orbital method: an approximate computational method for large molecules. *Chem. Phys. Lett.* 313: 701–706.
5. Fedorov D.G., Kitaura K. 2005. Multiconfiguration self-consistent-field theory based upon the fragment molecular orbital method. *J. Chem. Phys.* 122: 054108-1-10.
6. Schmidt M.W., Baldridge K.K., Boatz J.A., Elbert S.T., Gordon M.S., Jensen J.H. et al. 1993. The general atomic and molecular electronic structure system. *J. Comp. Chem.* 14: 1347–1363, www.msg.ameslab.gov/GAMESS/.
7. Fedorov D.G., Olson R.M., Kitaura K., Gordon M.S., Koseki S. 2004. A new hierarchical parallelization scheme: generalized distributed data interface (GDDI), and an application to the fragment molecular orbital method (FMO). *J. Comp. Chem.* 25: 872–880.
8. Nakano T., Kaminuma T., Sato T. et al. 2002. Fragment molecular orbital method: use of approximate electrostatic potential. *Chem. Phys. Lett.* 351: 475–480.
9. Fedorov D.G., Kitaura K. 2007. Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. *J. Phys. Chem.* A111: 6904–6914.
10. Foresman J.B., Head-Gordon M., Pople J.A., Frisch M.J. 1992. Toward a systematic molecular orbital theory for excited states. *J. Phys. Chem.* 96: 135–149.
11. Head-Gordon M., Rico R.J., Oumi M., Lee T.J. 1994. A doubles correction to electronic excited states from configuration interaction in the space of single substitutions. *Chem. Phys. Lett.* 219: 21–29.
12. Mochizuki Y., Koikegami S., Amari S. et al. 2005. Configuration interaction singles method with multilayer fragment molecular orbital scheme. *Chem. Phys. Lett.* 406: 283–288.
13. Mochizuki Y., Tanaka K., Yamashita K. et al. 2007. Parallelized integral-direct CIS(D) calculations with multilayer fragment molecular orbital scheme. *Theor. Chem. Acc.* 117: 541–553.
14. Mochizuki Y., Tanaka K. 2007. Modification for spin-adapted version of configuration interaction singles with perturbative doubles. *Chem. Phys. Lett.* 443: 389–397.
15. Mochizuki Y. 2008. A practical use of self-energy shift for the description of orbital relaxation. *Chem. Phys. Lett.* 453: 109–116.

16. Mochizuki Y., Nakano T., Amari S. et al. 2007. Fragment molecular orbital calculations on red fluorescent protein (DsRed). *Chem. Phys. Lett.* 433: 360–367.
17. Szabo A., Ostlund N.S. 1982. *Modern Quantum Chemistry,* New York: MacMillan.
18. Burke K., Werschnik J., Gross E.K.U. 2005. Time-dependent density functional theory: past, present, and future. *J. Chem. Phys.* 123: 062206-1-9.
19. Dreuw A., Weisman J.L., Head-Gordon M. 2003. Long-range charge-transfer excited states in time-dependent density functional theory require non-local exchange. *J. Chem. Phys.* 119: 2943–2946.
20. Chiba M., Fedorov D.G., Kitaura K. 2007. Time-dependent density functional theory based upon the fragment molecular orbital method. *J. Chem. Phys.* 127: 104108-1-11.
21. Chiba M., Fedorov D.G., Kitaura K. 2007. Time-dependent density functional theory with the multilayer fragment molecular orbital method. *Chem. Phys. Lett.* 444: 346–350.
22. Mochizuki Y., Nakano T., Koikegami S. et al. 2004. A parallelized integral-direct second-order Mϕller–Plesset perturbation theory method with a fragment molecular orbital scheme. *Theor. Chem. Acc.* 112: 442–452.
23. Mochizuki Y., Koikegami S., Nakano T., Amari S., Kitaura K. 2004b. Large scale MP2 calculations with fragment molecular orbital scheme. *Chem. Phys. Lett.* 396: 473–479.
24. Head-Gordon M., Oumi M., Maurice D. 1999. Quasidegenerate second-order perturbation corrections to single excitation configuration interaction. *Mole. Phys.* 96: 593–602.
25. Head-Gordon M., Lee T.J. 1997. Single reference coupled cluster and perturbation theories of electronic excitation energies. In *Recent Advances in Coupled Cluster Methods*, ed. Bartlett R.J., 221–253. Singapore: World Scientific.
26. Hirata S. 2005. Third- and fourth-order perturbation corrections to excitation energies from configuration interaction singles. *J. Chem. Phys.* 122: 094105-1-10.
27. Oumi M., Maurice D., Lee T.J., Head-Gordon M. 1997. A diagnostic for the applicability of the CIS and CIS(D) excitation energy methods. *Chem. Phys. Lett.* 279: 151–157.
28. Christiansen O., Koch H., Jϕrgensen P. 1995. The second-order approximate coupled cluster singles and doubles model CC2. *Chem. Phys. Lett.* 243: 409–418.
29. Schirmer J. 1982. Beyond the random-phase approximation: a new approximation scheme for polarization propagator. *Phys. Rev.* A26: 2395–2416.
30. Laikov D., Matsika S. 2007. Inclusion of second-order correlation effects for the ground and singly-excited states suitable for the study of conical intersections: the CIS(2) model. *Chem. Phys. Lett.* 448: 132–137.
31. Grimme S., Izgorodina E.I. 2004. Calculation of 0–0 excitation energies of organic molecules by CIS(D) quantum chemical methods. *Chem. Phys.* 305: 223–230.
32. Rhee Y.M., Head-Gordon M. 2007. Scaled second-order perturbation corrections to configuration interaction singles: efficient and reliable excitation energy methods. *J. Phys. Chem.* A111: 5314–5326.
33. Casanova D., Rhee Y.M., Head-Gordon M. 2007. Quasidegenerate scaled opposite spin second order perturbation corrections to single excitation configuration interaction. *J. Chem. Phys.* 128: 164106-1-11.
34. Hirata S., Fan P.-D., Shiozaki T., Shigeta Y. 2008. Single-reference methods for excited states in molecules and polymers. In *Radiation Induced Molecular Phenomena in Nucleic Acid: A Comprehensive Theoretical and Experimental Analysis,* in the book series *Challenges and Advances in Computational Chemistry and Physics,* Vol. 5., eds. Leszczynski J., and Shukla M., 15–64. Berlin: Springer.

35. Dykstra C.E., Davidson E.R. 2000. Enhanced second-order treatment of electron pair correlation. *Inter. J. Quant. Chem.* 78: 226–236.
36. Surján P., Szabados Á. 1998. Dyson-corrected orbital energies for the perturbative treatment of electron correlation. *Inter. J. Quant. Chem.* 69: 713–719.
37. www.rcsb.org/pdb/.
38. www.compchem.com/.
39. http://amber.scripps.edu/.
40. www.gaussian.com/.
41. Meyer T.E. 1985. Isolation and characterization of soluble cytochromes, ferredoxins and other chromophoric proteins from the halophilic phototrophic bacterium *Ectothiorhodospira halophila*. *Biochim. Biophys. Acta* 23: 175–183.
42. Imamoto Y., Kataoka M., Tokunaga F. 1996. Photoreaction cycle of photoactive yellow protein from *Ectothiorhodospira halophila* studied by low-temperature spectroscopy. *Biochemistry* 12: 14047–14053.
43. Yoda M., Houjou H., Inoue Y., Sakurai M. 2001. Spectral turning of photoactive yellow protein. Theoretical and experimental analysis of medium effects on the absorption spectrum of the chromophore. *J. Phys. Chem.* B 105: 9887–9895.
44. Foresman J.B., Frisch A. 1996. *Exploring Chemistry with Electronic Structure Methods,* 2nd ed. Pittsburgh, PA: Gaussian Inc.
45. Nakano T., Mochizuki Y., Fukuzawa K., Amari S., Tanaka S. 2006. Developments and applications of ABINIT-MP software based on the fragment molecular orbital method. In *Modern Methods for Theoretical Physical Chemistry of Biopolymers*, eds. Starikov E.B., Lewis J.P., and Tanaka S., 39–52. Amsterdam: Elsevier.
46. Nielsen I.B., Boyé-Peronne S., El Ghazaly M.O.A. et al. 2005. Absorption spectra of photoactive yellow protein chromophores in vacuum. *Biophys. J.* 89: 2597–2604.
47. Matz M.V., Fradkov A.F., Labas Y.A. et al. 1999. Fluorescent proteins from nonbioluminescent *Anthozoa* species. *Nature Biotech.* 17: 969–973.
48. Gross L.A., Baird G.S., Hoffman R.C., Baldridge K.K., Tsien R.Y. 2000. The structure of the chromophore within DsRed, a red fluorescent protein form coral. *Proc. Natl. Acad. Sci. USA* 97: 11990–11995.
49. Yarbrough D., Wachter R.M., Kallio K., Matz M.V., Remington S.J. 2001. Refined crystal structure of DsRed, a red fluorescent protein from coral, at 2.0-Å resolution. *Proc. Natl. Acad. Sci. USA* 98: 462–467.
50. Shrestha S., Deo S.K. 2006. Anthozoa red fluorescent protein in biosensing. *Anal. Bioanal. Chem.* 386: 515–524.
51. Boyé S., Nielsen S.B., Krogh H. et al. 2003. Gas-phase absorption properties of DsRed model chromophores. *Phys. Chem. Chem. Phys.* 5: 3021–3026.
52. Das A.K., Hasegawa J., Miyahara T., Ehara M., Nakatsuji H. 2003. Electronic excitations of the Gree fluorescent protein chromophore in its protonation states: SAC/SAC-CI study. *J. Comp. Chem.* 24: 1421–1431.
53. Mochizuki Y., Yamashita K., Murase T. et al. 2008. Large scale FMO-MP2 calculations on a massively parallel-vector computer. *Chem. Phys. Lett.* 457: 396–403.
54. Mochizuki Y., Fukuzawa K., Kato A. et al. 2005. A configuration analysis for fragment interaction. *Chem. Phys. Lett.* 410: 247–253.
55. Amari S., Aizawa M., Zhang J. et al. 2006. VISCANA: visualized cluster analysis of protein-ligand interaction based on the *ab initio* fragment molecular orbital method for virtual ligand screening. *J. Chem. Inf. Comput. Sci.* 46: 221–230.
56. Fedorov D.G., Kitaura K. 2007. Pair interaction energy decomposition analysis. *J. Comp. Chem.* 28: 222–237.
57. Sharner N.C., Cambell R.E., Steinbach P.A. et al. 2004. Improved monomeric red, orange, yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nature Biotech.* 22: 1567–1572.

58. Shu X., Shaner N.C., Yarbrough C.A., Tsien R.Y., Remington S. 2006. Novel chromophores and buried charges control color in mFruits. *Biochem.* 45: 9639–9647.
59. Taguchi N., Mochizuki Y., Nakano T. et al., *J. Phys. Chem. B,* in press.
60. Nakatuji H. 1978. Cluster expansion of the wavefunction. Excited states. *Chem. Phys. Lett.* 59: 362–364.
61. Hasegawa J., Ohkawa K., Nakatsuji H. 1998. Excited states of the photosynthetic reaction center of *Rhodopseudomonas viridis*: SAC-CI study. *J. Phys. Chem.* B102: 10410–10419.
62. Hasegawa J., Nakatsuji H. 1998. Mechanism and unidirectionality of the electron transfer in the photosynthetic reaction center of *Rhodopseudomonas viridis*: SAC-CI theoretical study. *J. Phys. Chem.* B102: 10420–10430.
63. Nakatani N., Hasegawa J., Nakatsuji H. 2007. On the yellow-green bioluminescence of the firefly, *Photinus pyralis*: SAC-CI study. *J. Amer. Chem. Soc.* 129: 8756–8765.
64. Fujimoto K., Hayashi S., Hasegawa J., Nakatsuji H. 2007. Theoretical studies on the color-tuning mechanism in retinal proteins. *J. Chem. Theory Comp.* 3: 605–618.
65. Tagami A., Ishibashi N., Kato D. et al. 2008. Theoretical study on emission spectra of bioluminescent luciferases by fragment molecular orbital method. *J. Comp. Aided Chem.* (Japanese) 9: 47–54.
66. Tagami A., Ishibashi N., Kato D. et al. Submitted to *Chem. Phys. Lett.*
67. Nielsen I.B., Lammich L., and Andersen L.H. 2006. S_1 and S_2 excited states of gas-phase Schiff-base retinal chromophores. *Phys. Rev. Lett.* 96: 018304-1-4.
68. Hirata S., Valiev M., Dupuis M. et al. 2005. Fast electron correlation methods for molecular clusters in the ground and excited states. *Mole. Phys.* 103: 2255–2265.
69. Kawaguchi K., Yamato T. 2006. Theoretical prediction of optical absorption maxima for photosensory receptor mutants. *Chem. Phys. Lett.* 430: 386–390.
70. Valiev M., Kowalski K. 2006. Hybrid coupled cluster and molecular dynamics approach: application to the excitation spectrum of cytosine in the native DNA environment. *J. Chem. Phys.* 125: 211101-1-3.
71. Ito M., Fukuzawa K., Mochizuki Y., Nakano T., Tanaka S. 2008. *Ab initio* fragment molecular orbital study of molecular interactions between liganded retinoid X receptor and its coactivator; Part II: influence of mutations in transcriptional activation function 2 activating domain core on the molecular interactions. *J. Phys. Chem.* A112: 1986–1998.
72. Kitaura K., Sugiki S., Nakano T., Komeiji Y., Uebayasi M. 2001. Fragment molecular orbital method: analytical energy gradients. *Chem. Phys. Lett.* 336: 163–170.
73. Komeiji Y., Nakano T., Fukuzawa K. et al. 2003. Fragment molecular orbital method: application to molecular dynamics simulation, "*ab initio* FMO-MD," *Chem. Phys. Lett.* 372: 342–347.
74. Komeiji Y., Ishikawa T., Mochizuki Y., Yamataka H., Nakano T. 2009. Fragment molecular orbital method-based molecular dynamics (FMO-MD) as a simulator for chemical reactions in explicit solvation. *J. Comp. Chem.* 30: 40–50. (DOI: 10.1002/jcc.21025).
75. Komeiji Y., Mochizuki Y., Nakano T., Fedorov D.G. 2008. Fragment molecular orbital-based molecular dynamics (FMO-MD), a quantum simulation tool for large molecular systems. *J. Mole. Struct. (THEOCHEM)*, 898: 2–7.
76. Mochizuki Y., Komeiji Y., Ishikawa T., Nakano T., Yamataka H. 2007. A fully quantum mechanical simulation study on the lowest n-π^* state of hydrated formaldehyde. *Chem. Phys. Lett.* 437: 66–72.
77. Sato M., Yamataka Y., Komeiji Y., Mochizuki Y., Ishikawa T., Nakano T. 2008. How Does an S_N2 reaction take place in solution? Full *ab initio* MD simulations for the hydrolysis of the methyl diazonium ion. *J. Am. Chem. Soc.* 130: 2396–2397.

78. Frisch M.J., Head-Gordon M., Pople J.A. 1990. A direct MP2 gradient method. *Chem. Phys. Lett.* 166: 275–280.
79. Ishikawa N., Head-Gordon M. 1995. Analytical gradient of the CIS(D) perturbative correction to single-excitation configuration interaction excited states. *Intern. J. Quant. Chem. Symp.* 29: 421–427.
80. Stanton J.F., Gauss J., Ishikawa N., Head-Gordon M. 1995. A comparison of single reference methods for characterizing stationary points of excited state potential energy surfaces. *J. Chem. Phys.* 103: 4160–4174.
81. Hättig C. 2005. Structure optimizations for excited states with correlated second-order methods: CC2 and ADC(2). *Adv. Quant. Chem.* 50: 37–60.
82. Matsika S. 2007. Conical intersections in molecular systems. *Rev. Comp. Chem.* 23: 83–124.

5 The Fragment Molecular Orbital–Based Time-Dependent Density Functional Theory for Excited States in Large Systems

Mahito Chiba, Dmitri G. Fedorov, and Kazuo Kitaura

CONTENTS

5.1 INTRODUCTION

Time-dependent density functional theory (TDDFT)[1–5] is widely used for excited state calculations, owing to its reasonable accuracy and low computational cost. TDDFT delivers excitation energies and oscillator strengths, and it requires much smaller computational resources than high-level *ab initio* methods such as symmetry-adapted cluster configuration interaction (SAC-CI).[6,7] The computational cost of TDDFT is similar to the configuration interaction with single excitations (CIS)[8,9] or time-dependent Hartree–Fock (TDHF),[10] but in general, TDDFT is known to give more accurate excitation energy than CIS or TDHF, because of the inclusion of the electron correlation in the density functional.[3–5] Also, Furche et al.[5] developed the analytic energy gradient of TDDFT and revealed that TDDFT gives better excited state geometries than CIS or TDHF.

However, TDDFT is typically applicable to single electron excitations. To take into account double excitations, one has to use multireference-based calculations such as complete active space second-order perturbation theory (CASPT2).[11–13]

In order to calculate excited states of molecules in solution, Cossi et al. combined TDDFT with the polarizable continuum model (PCM),[14,15] and the developed TDDFT/PCM method has been widely used for studying photochemical systems in solution.[16–18]

TDDFT calculations are usually carried out with the Davidson-like subspace algorithm,[4] in which case the cost is on the order of $O(N^3)$, where N is the size of the system. This order is much smaller than those of correlated *ab initio* methods (e.g., $O(N^6)$ for SAC-CI), but the scaling of $O(N^3)$ for large systems is still high. Alternatively, there are semi-empirical approaches such as the intermediate neglect of differential overlap with configuration interaction singles (INDO/CIS)[19] method. The semi-empirical approaches have the advantage of requiring moderate computational resources, which, however, increase more than $O(N)$, and usually matrix operations imply at least an $O(N^3)$ scaling. In addition, parameters are often unavailable or need to be reoptimized. Alternatively, one can also describe a part of the system with molecular mechanics (MM), and the excitation region with quantum mechanics (QM), resulting in the QM/MM method, which inherits the advantages and disadvantages of force fields.[20]

The fragment molecular orbital (FMO) method developed by Kitaura et al.[21,22] is one of many fragmentation approaches.[23] In FMO, the system is divided into small fragments, and the total properties, estimated from those of fragment monomers and dimers, are in close agreement with *ab initio* methods. For the excited state treatment in FMO, Mochizuki et al. combined multilayer FMO[24] with CIS and CIS with the second-order perturbative doubles correction [CIS(D)].[25,26] In their method, CI

calculation is performed on a single fragment of interest after determining the electronic densities of all fragments self-consistently (more details are given in Chapter 4). Among other methods to describe excitations in large systems, Hirata et al. developed an FMO-like scheme for the description of molecular clusters by treating the effect of the environment by fragment dipoles,[27] and the giant SAC-CI scheme was proposed for polymer-like systems.[28]

Recently, we developed FMO-based time-dependent density functional theory (FMO-TDDFT)[29–31] to calculate excited states of large systems. In this chapter, we introduce the basic theory of FMO-TDDFT and describe the accuracy and efficiency of FMO-TDDFT by calculating several systems by FMO-TDDFT. In addition, we provide guidelines on how to run FMO-TDDFT calculations practically, in particular, how to choose the TDDFT fragment. Finally, the application of FMO-TDDFT with and without PCM is described for the lowest singlet excitation of the photoactive yellow protein (PYP), as an example of an excited state calculation in proteins.

5.2 THEORY

5.2.1 Time-Dependent Density Functional Theory (TDDFT)

First, we briefly introduce TDDFT, a theory for calculating excitation energies and oscillator strengths. The TDDFT excitation energy ω and the corresponding excitation vectors **X** and **Y** are generally obtained by solving a non-Hermitian eigenvalue equation[3,4]:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = \omega \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & -\mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \tag{5.1}$$

where **X** and **Y** are vectors, which represent excitation and deexcitation components of the electronic density change, respectively; ω is the excitation energy. The elements of matrices **A** and **B** are

$$A_{ai\sigma,bj\sigma'} = \delta_{ab}\delta_{ij}\delta_{\sigma\sigma'}(\varepsilon_{a\sigma} - \varepsilon_{i\sigma'}) + K_{ai\sigma,bj\sigma'} \tag{5.2}$$

and

$$B_{ai\sigma,bj\sigma'} = K_{ai\sigma,jb\sigma'} \tag{5.3}$$

where σ and σ′ are spin indices, and $\varepsilon_{p\sigma}$ is the p-th Kohn–Sham molecular orbital (MO) energy. As usual, indices $i, j, \ldots$ and $a, b, \ldots$ are label occupied and virtual orbitals, respectively. Matrix element $K_{ai\sigma,bj\sigma'}$ in Equation 5.2 and Equation 5.3 is given by

$$K_{pq\sigma,rs\sigma'} = (pq\sigma|rs\sigma') - c_x\delta_{\sigma\sigma'}(pr\sigma|qs\sigma') + f^{xc}_{pq\sigma rs\sigma'} \tag{5.4}$$

where $p, q, \ldots$ indicate general MOs. $(pq\sigma|rs\sigma')$ is a two-electron repulsion integral,

$$(pq\sigma \mid rs\sigma') = \iint \psi^*_{p\sigma}(r_1)\psi_{q\sigma}(r_1)\frac{1}{r_{12}}\psi^*_{r\sigma'}(r_2)\psi_{s\sigma'}(r_2)d^3r_1 d^3r_2 \tag{5.5}$$

and c_x is the mixing parameter of the HF exchange integral in hybrid functionals. In Equation 5.4, $f^{xc}_{pq\sigma rs\sigma'}$ is the matrix element of the second derivative of the exchange-correlation energy functional, E_{xc}, with respect to the electron density in the adiabatic approximation, whose operator form is

$$f^{xc}_{\sigma\sigma'} = \frac{\delta^2 E_{xc}}{\delta\rho_\sigma(r_1)\delta\rho_{\sigma'}(r_2)} \tag{5.6}$$

5.2.2 Time-Dependent Density Functional Theory Based upon the Fragment Molecular Orbital Method

The FMO-based TDDFT calculations in general are based on the ground state FMO-DFT (see Chapter 2). At the end of DFT calculations, TDDFT equations are solved for one fragment and, optionally, for some fragment pairs (dimers). The two-body expansion of the TDDFT excited state energy E^* in FMO is analogous to the ground state expression and similar to the dipole field method for solvated systems.[27] The excitation is assumed to be mostly local to fragment M, and the other fragments add two-body corrections to it:

$$E^* = E^*_M + \sum_{I\neq M} E^0_I + \sum_{I\neq M}\left(E^*_{MI} - E^*_M - E^0_I\right) + \sum_{\substack{I>J \\ I,J\neq M}}\left(E^0_{IJ} - E^0_I - E^0_J\right) \tag{5.7}$$

where the superscripts 0 and * denote the ground state and excited states, respectively. Then, the excitation energy ω is given in FMO2-TDDFT[30] by

$$\omega \equiv E^* - E^0 = \omega_M + \sum_{I\neq M}(\omega_{MI} - \omega_M) \tag{5.8}$$

where the single fragment excitation energy $\omega_M \equiv E^*_M - E^0_M$ is corrected by pair contributions from dimer terms $\omega_{MI} \equiv E^*_{MI} - E^0_{MI}$.

If we ignore the second term of the right-hand side of Equation 5.8, we obtain the excitation energy in the one-body case (FMO1-TDDFT[29]):

$$\omega = \omega_M \tag{5.9}$$

In FMO1-TDDFT, the excitation energies are calculated only for the TDDFT fragment M with the surrounding Coulomb field exerted by the remaining fragments. The electronic densities of all fragments are converged self-consistently, for the ground state FMO-DFT, which corresponds to the ground state DFT orbitals used in the conventional TDDFT. Thus, all fragments are optimized in the electrostatic field of the ground state (DFT). This is different from FMO-MCSCF,[32] where the field contribution from the MCSCF fragment can come from either the ground or an excited state. By carrying out an FMO2-TDDFT calculation, we get the FMO1-TDDFT excitation energy ω_M, and the excitation energy analysis given by

the $\Delta\omega_{MI} \equiv \omega_{MI} - \omega_M$ contributions of the other fragments. In the current implementation, oscillator strengths are calculated only for FMO1-TDDFT.

If we concentrate on the excitation energy, only a very small number of pair calculations has to be performed. This can be understood from Equation 5.8: we only need to compute ω_{MI} dimers, further limited by an approximation threshold (see below). Practically, the number of dimers to be computed is about eight to 20, depending on the system.

In order to make the FMO2-TDDFT calculations efficient, we introduce a threshold that is applied to the interfragment distance R_{IJ}, defined in FMO as follows:

$$R_{IJ} = \min_{i \in I,\, j \in J} \left\{ \frac{|\vec{r}_i - \vec{r}_j|}{W_i + W_j} \right\} \tag{5.10}$$

where $\vec{r}_i$ and W_i are atomic coordinates and van der Waals radii of atom i, respectively. That is, the distance is given by the shortest atomic pair distance relative to the atomic size, and $R = 1$ means that the two closest atoms, represented by spheres with the van der Walls radii, touch each other. Using this separation information, we do not perform those dimer calculations involving fragment I, for which R_{MI} is greater than a chosen threshold value (usually 2.0).[30]

The necessary condition for a successful FMO2-TDDFT calculation is that the TDDFT fragment be large enough to describe the excitation, to which the two-body terms will add a correction. One can fairly easy determine if the assumption about the fragment size was appropriate: a large correction value $\omega_{MI} - \omega_M$ indicates that some part of fragment I may have to be included in fragment M. How large is "large" depends upon the desired accuracy and the magnitude of the explicit many-body effects not included in FMO2, and as a simple rule the pair corrections should be typically smaller than the total level of accuracy, so that some error accumulated in summing them may be hoped to be within the desired accuracy.

It should also be understood that FMO2-TDDFT is not intended[30] to be used in the "hopping" case, when the excitation can be thought of as from fragment M to fragment I, the proper scenario is to have an excitation mostly local to fragment M. The fragmentation scheme employed in FMO does not impose any spatial restrictions, so that several regions in space may be combined into the same fragment.

5.2.3 Multilayer Treatment of FMO-TDDFT

The multilayer FMO scheme[24] can be used for TDDFT. The reasons for doing it are (a) to reduce the computational cost and (b) to use a better basis set for the excitation energies. Typically, one has two layers. In the lower layer, one can use RHF or DFT to obtain the fragment densities. In the higher layer, one has DFT and TDDFT for the ground and excited states, respectively. The two layers can have different basis sets. The TDDFT calculations as described above will be limited to the higher layer, so that only the TDDFT fragment (FMO1-TDDFT) or some TDDFT dimers (FMO2-TDDFT) will be computed for the fragments in the higher layer.

Most fragments exert their influence upon the excitation energies only through the electrostatic field (that is, all dimers IJ where neither I nor J is M in Equation 5.8).

Moreover, the effect of the field frequently is somewhat similar to both the ground and the excited states, thus it is often sufficient to use RHF densities to describe the environment in this FMO-RHF:TDDFT method, which was found to reproduce the results by the standard FMO-TDDFT with negligible errors.[29] In addition, one can use diffuse functions for the higher layer, which is in general not possible at present for all fragments in FMO.

5.2.4 FMO-TDDFT with the Polarizable Continuum Model (PCM)

To describe the effect of solvent, we interfaced[31] FMO-TDDFT with the polarizable continuum model (PCM).[14,15] The general scheme for this method is to perform ground state FMO-DFT calculations in the presence of solvent, which is described by the polarizable cavity made of atomic spheres, followed by the TDDFT calculations in the same cavity. Two schemes were suggested[16] for TDDFT/PCM: the equilibrium and nonequilibrium methods. The former is suitable when the excited state is allowed to relax and equilibrate in solution, and the latter is best for vertical excitations. The practical difference is in the choice of the dielectric constant in the equations below: the equilibrium and nonequilibrium methods use the standard and optical values, respectively. So far, we only computed vertical excitations and thus used the nonequilibrium method.

In TDDFT/PCM, **K** in Equation 5.2 and Equation 5.3 is written as:

$$K_{pq\sigma,rs\sigma'} = (pq\sigma \mid rs\sigma') - c_x\delta_{\sigma\sigma'}(pr\sigma \mid qs\sigma') + f^{xc}_{pq\sigma rs\sigma'} + \nu^{PCM}_{ai\sigma,bj\sigma'} \qquad (5.11)$$

where $\nu^{PCM}_{ai\sigma,bj\sigma'}$ is a response contribution from PCM.[16] There are several levels of doing FMO/PCM calculations, depending on the many-body PCM potential expansion.[33] Because the excitation energies are less sensitive to the choice of this level than the solvation energies, we normally use the cheaper level of FMO/PCM[1], where the PCM potential (the electrostatic field from the point charges on the cavity acting upon the solute) is represented by the sum of fragment contributions without explicit two-body corrections. A more expensive method FMO/PCM[1(2)], where these corrections are included without full self-consistency, was used for some accuracy tests below.

5.3 SAMPLE CALCULATIONS BY FMO-TDDFT

5.3.1 Phenol+$(H_2O)_N$ with FMO-TDDFT

First, we briefly introduce the accuracy of FMO-TDDFT by comparing its results to those of full TDDFT. The test systems are solvated phenol and a polypeptide. The former is an example of a molecular cluster system, and the latter is a prototype of a large single molecule (protein).

As a first test of FMO-TDDFT, we introduce the computational results for several low lying singlet and triplet excitation energies of solvated phenol, PhOH + $(H_2O)_n$, n = 16,32,64.[30] All geometries were constructed in the following way. First, PhOH was optimized at the RHF/6-31G* level. Then 113 water molecules were added and

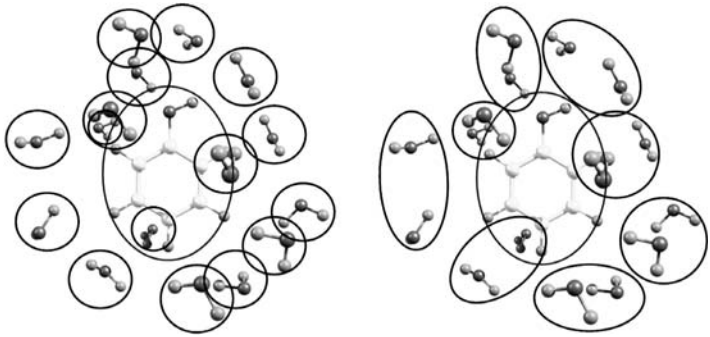


FIGURE 5.1 Fragmentation of PhOH+$(H_2O)_{16}$, denoted as (a) A and (b) B, for one and two water molecules per fragment, respectively.

their geometry was optimized using molecular mechanics. From that, 16, 32, or 64 water molecules closest to PhOH were taken.[30] PhOH formed the TDDFT fragment. To observe the fragment size effect upon the accuracy, two types of fragmentation were defined: fragmentation A, one water molecule per fragment ($n + 1$ fragments), see Figure 5.1a, and fragmentation B: two water molecules per fragment ($n/2 + 1$ fragments), see Figure 5.1b.

To establish the accuracy of FMO2-TDDFT, conventional TDDFT calculations were also done. All the excitation energy calculations by FMO-TDDFT and conventional TDDFT were performed by using BLYP[34,35] with long-range correction (LC-BLYP)[36,37] and 6-31G* basis set.

The results are summarized in Table 5.1. The stand-alone PhOH excitation energy for S_1 is at the same level of theory[29] equal to 5.374 eV, and by looking at the FMO1-TDDFT results, we can see that the effect of the electrostatic field due to the environment (water) is nearly negligible (about 0.001 to 0.004 eV, which is perhaps on the same order as the grid accuracy in DFT). Because in water clusters the polarization effects are significant, this indicates that both the ground and excited states are shifted in the same direction, so that the effect cancels out.

Increasing the number of water molecules shifts the FMO1-TDDFT excitation energies by 0.027 (T_2) or less (higher excitations in phenol appear to be influenced more by the electrostatic field of water molecules). The full TDDFT predicts a similar value of 0.034 eV for T_2. However, for other states, FMO1-TDDFT is not very accurate, the largest error being for the S_2 state. (It predicts a shift of 0.013 eV, whereas the TDDFT result is 0.044 eV.) FMO1-TDDFT can be used to separate the electrostatic contribution of the environment to the excitation energy shift (in which case scheme B should yield more reliable predictions), and one can see that quantum effects beyond the simple electrostatics play a considerable role.

The FMO1-TDDFT oscillator strengths reflect also just this water electrostatic field; however, the values agree with full TDDFT quite well, with the typical error

TABLE 5.1
LC-BLYP/6-31G* Excitation Energies (in eV) of PhOH+$(H_2O)_n$ (n = 16, 32, 64)†

		Fragmentation A			Fragmentation B				
System	State	FMO1-TDDFT		FMO2-TDDFT	FMO1-TDDFT		FMO2-TDDFT	TDDFT	
PhOH+$(H_2O)_{16}$	S_1	5.375	(0.030)	5.284	5.376	(0.030)	5.290	5.290	(0.036)
	S_2	6.379	(0.044)	6.204	6.380	(0.044)	6.211	6.213	(0.054)
	S_3	—	—	—	—	—	—	6.728	(0.022)
	T_1	3.757	(0.000)	3.737	3.758	(0.000)	3.740	3.739	(0.000)
	T_2	4.417	(0.000)	4.391	4.418	(0.000)	4.396	4.396	(0.000)
	T_3	4.852	(0.000)	4.802	4.853	(0.000)	4.806	4.805	(0.000)
PhOH+$(H_2O)_{32}$	S_1	5.373	(0.031)	5.228	5.375	(0.031)	5.240	5.268	(0.029)
	S_2	6.367	(0.049)	6.078	6.367	(0.050)	6.094	6.169	(0.052)
	T_1	3.786	(0.000)	3.750	3.788	(0.000)	3.752	3.749	(0.000)
	T_2	4.400	(0.000)	4.346	4.403	(0.000)	4.357	4.371	(0.000)
PhOH+$(H_2O)_{64}$	S_1	5.372	(0.031)	5.223	5.370	(0.031)	5.223	5.264	(0.033)
	S_2	6.371	(0.047)	6.083	6.366	(0.048)	6.083	6.174	(0.048)
	T_1	3.780	(0.000)	3.742	3.782	(0.000)	3.741	3.740	(0.000)
	T_2	4.393	(0.000)	4.337	4.391	(0.000)	4.338	4.362	(0.000)

†*Note:* Oscillator strengths are shown in parentheses.

of about 0.003 (in two cases up to 0.10). Neither the excitation energies nor the oscillator strengths are significantly improved by doubling the fragment size. (This is because in FMO1-TDDFT such doubling only improves the description of the electrostatic field, whose effect is small.) The FMO1-TDDFT excitation energies can be seen to have a considerable deviation from the full TDDFT values, and the two-body corrections are needed for an accurate description.

FMO2-TDDFT, on the other hand, reproduces the full TDDFT values quite accurately (Table 5.1). The errors for the S_1, S_2, T_1, and T_2 states are on the order of a few milli-eV for $n = 16$, increasing for $n = 32$ and $n = 64$ to the values typically on the order of 0.01 to 0.03 eV, although for the S_2 state the error was about 0.09 eV. For FMO2-TDDFT, we notice that doubling the fragment size had a favorable effect upon the accuracy, because differently from FMO1-TDDFT, larger non-TDDFT fragments can have a direct effect upon the excitation energy through explicit pair corrections. FMO2-TDDFT is a distinct improvement over FMO1-TDDFT, the error for the latter being as large as about 0.2 eV. This suggests that the orbital delocalization and charge transfer effects absent in FMO1-TDDFT play a noticeable role in the excitations.

It is interesting to observe that essentially without exceptions, all FMO2-TDDFT excitation energies underestimate the full TDDFT values for this system; whereas FMO1-TDDFT displayed both signs of the difference to TDDFT. The former observation gives further hints to the error origin: lower excitation energies stem from negative pair corrections, which are connected to the smaller highest occupied and the lowest unoccupied molecular orbital gap in dimers, and that may be due to the orbital interaction in the increased orbital space (for dimers versus monomers). Thus, the underestimation in FMO2-TDDFT implies the absence of higher-order many-body corrections of the same type, which further lowers the excitation energy. We note, however, that it should be possible to observe both signs of dimer corrections in principle, as there are several competing effects that determine them, such as the detailed electronic structure.

To reduce the error for the excited state of the primary interest, one can change the fragment size to incorporate the important part of the environment. The errors of up to 0.09 eV for the S_2 state suggest that one may have to increase the TDDFT fragment if that state is of interest. We also note that the fragment size in solvated phenol is very small, and thus, the errors are observed to be fairly large. In practical applications to biological systems, one has much larger fragments and smaller errors as exemplified below.

The errors of FMO2-TDDFT in the excitation energies for S_1, S_2, T_1, and T_2 states are summarized in Figure 5.2a and Figure 5.2b for fragmentations A and B, respectively. The error shows a distinct increase when the system size grows from $n = 16$ to $n = 32$, and with the further size doubling there is a small increase in error, which indicates that the extent to which excitations are affected by the surroundings is fairly limited. We also found that the error in the S_2 (or T_2) state is larger than that of S_1 (or T_1) state. Because higher states depend upon the lower states during the diagonalization process, the error in the latter promotes a larger error in the former. We also found that the triplet excited states gave less error than the singlet excited states, because for the triplet excitations, the Coulomb interaction term in the $\mathbf{A} + \mathbf{B}$

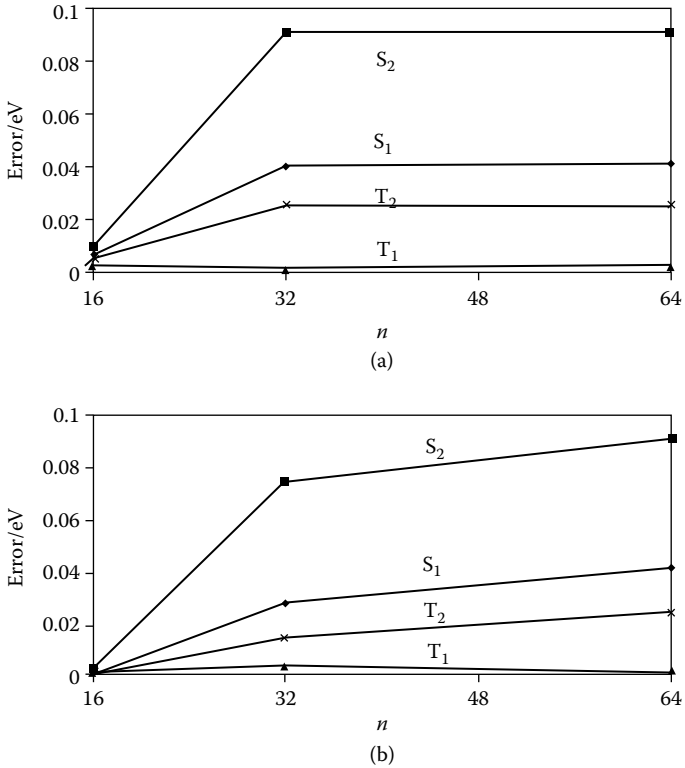
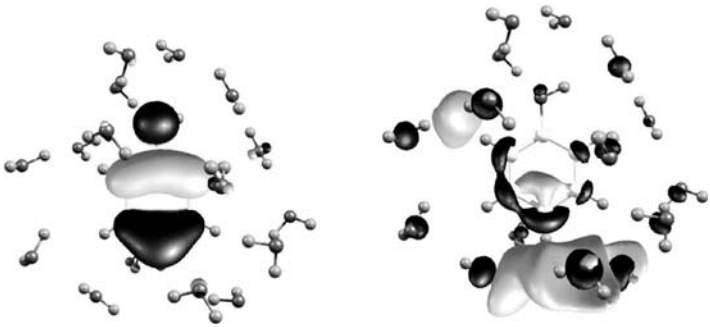


FIGURE 5.2 The errors (in eV) in the LC-BLYP/6-31G* excitation energies for the S_1, S_2, T_1, T_2 states (FMO2-TDDFT relative to full TDDFT), for fragmentation (a) A and (b) B (one and two water molecules per fragment, respectively).

matrix (see Equation 5.2 and Equation 5.3) vanishes, and they have a more localized character than the singlet excitations.

FMO2-TDDFT reproduced the singlet S_1, S_2 and the triplet T_1, T_2, and T_3 excited states with quite small errors relative to the full TDDFT calculations. The character of these excitations is a π-π^* excitation on the phenyl coupled with n-π^* involving lone pairs (n) of water. However, both FMO1- and FMO2-TDDFT failed to reproduce the excitation energy for the S_3 state found in the full TDDFT for PhOH+$(H_2O)_{16}$. This S_3 state corresponds to the transition from the occupied orbital shown in Figure 5.3a into the virtual orbital shown in Figure 5.3b. It can be seen that this transition is a charge transfer one, from PhOH into the surrounding water molecules. In FMO1-TDDFT, an excited state is determined as a local excitation in the TDDFT fragment (PhOH), and the two-body terms in FMO2-TDDFT add a correction to it. Thus, if a state does not appear in the TDDFT fragment, it cannot be corrected by FMO2-TDDFT. This is the reason for the absence of the S_3 results in FMO-TDDFT calculations. To treat this kind of state, one has to increase the fragment size to include some water molecules.



(a) (b)

FIGURE 5.3 The main transition of the S_3 excitation in PhOH+$(H_2O)_{16}$: (a) the occupied orbital and (b) the virtual orbital.

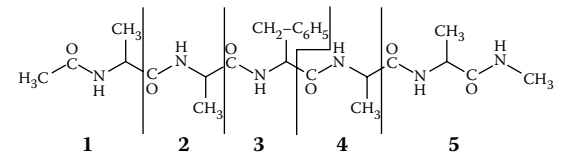
5.3.2 Polyalanine with FMO-TDDFT

The lowest singlet π-π* excitation in the phenyl group was calculated for the linear *n*-residue polyalanine ($n = 4, 8, 16$), with one phenylalanine residue inserted in the middle of this residue chain. This system is denoted by Phe-$(Ala)_n$, and it represents a simple model polypeptide used to establish the validity of our approach for the case when some covalent bonds are fragmented. We tested both the α helix and β strand isomers, and the geometries were capped and constructed with HYPERCHEM modeling software.[30] LC-BLYP functional with the 6-31G* basis set was used in all calculations. In the FMO-TDDFT calculations, the system was divided into $n + 1$ fragments (see Figure 5.4).

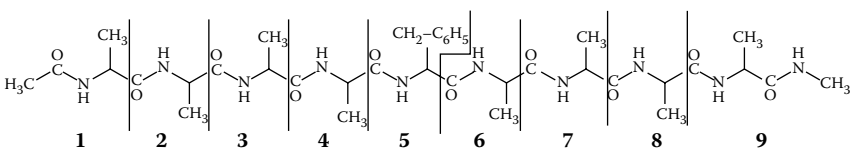
Since the excited state of interest is the π-π* excitation of the phenyl group, the TDDFT fragment was number 3, 5, and 9 for $n = 4, 8$, and 16, respectively (see Figure 5.4). In FMO2-TDDFT, the fragment dimers whose distances are greater than 2.0 were not calculated, because far separated fragment dimers do not contribute to the excitation energy as discussed above. For Phe-$(Ala)_{16}$, we failed to converge the full TDDFT calculations for the desired excited state. In this case, the full TDDFT calculations are difficult to perform, due to the large size and the convergence problem of the iterative solver for the desired excited state, which shows a large degeneracy.

The results are shown in Table 5.2. As can be seen, FMO1-TDDFT, which includes only the electrostatic field from the environment, shows a consistent deviation from the full TDDFT of about 0.05 eV and 0.014 eV for the β and α isomers, respectively. Interestingly, despite the large dipole moment of the α helix, and the number of hydrogen bonds (which involve a considerable charge transfer, whose coupling is one of the main sources of the FMO error for the ground state properties), a larger error is observed for the β strand. The origin of this can be thought to be in the larger field effect cancellation for the α helix (between the ground and the excited states).

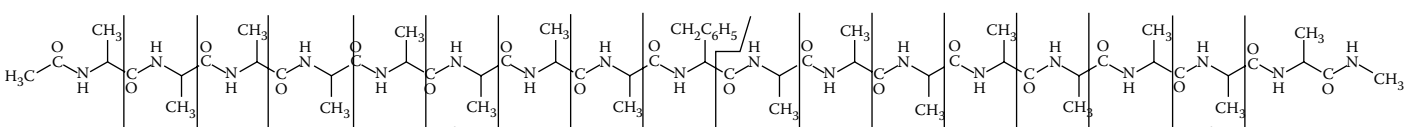
FMO2-TDDFT reproduced the excitation energies of full TDDFT calculations with virtually negligible errors (0.004 to 0.008 eV) for all *n*, and the errors for both



(a)



(b)



(c)

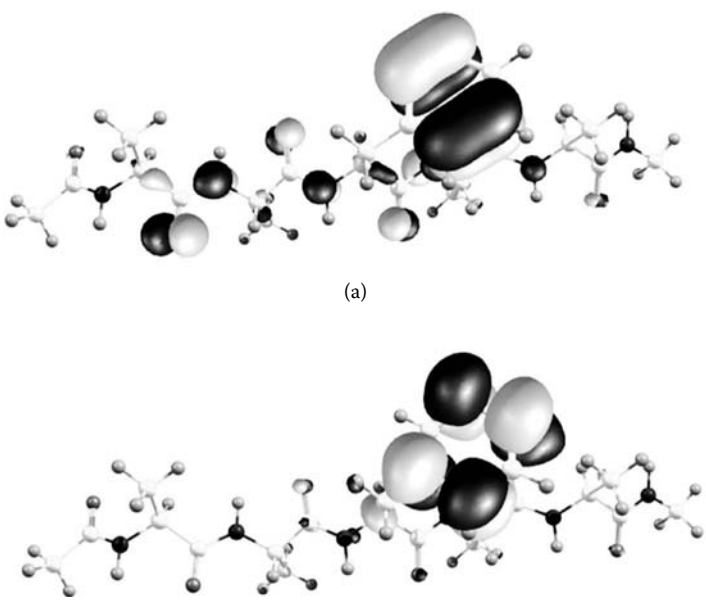
FIGURE 5.4 The fragment division of Phe-(Ala)$_n$: (a) $n = 4$, (b) $n = 8$, and (c) $n = 16$.

TABLE 5.2
LC-BLYP/6-31G* Excitation Energies (in eV) for the Lowest Singlet Excitation on the Benzene Ring of Phe-$(Ala)_n$ (n = 4, 8, 16)†

System	FMO1-TDDFT		FMO2-TDDFT	TDDFT	
β-Phe-$(Ala)_4$	5.415	(0.001)	5.363	5.367	(0.001)
β-Phe-$(Ala)_8$	5.415	(0.001)	5.362	5.366	(0.001)
β-Phe-$(Ala)_{16}$	5.414	(0.001)	5.360	5.365	(0.001)
α-Phe-$(Ala)_4$	5.413	(0.000)	5.396	5.408	(0.003)
α-Phe-$(Ala)_8$	5.416	(0.000)	5.402	5.402	(0.000)
α-Phe-$(Ala)_{16}$	5.415	(0.000)	5.401	—	—

†*Note:* Oscillator strengths are shown in parentheses.

isomers are similar and smaller than those of PhOH + $(H_2O)_n$. Also, the excitation energies do not noticeably change with the length of the peptide chain. The occupied and the virtual orbitals of the main transition in the S_1 excitation of Phe-$(Ala)_4$ are shown in Figure 5.5a and Figure 5.5b, respectively. As can be seen from the figure, this excitation is quite localized upon the phenyl group, which explains the small error seen in the excitation energies.



(a)

(b)

FIGURE 5.5 The main transition of the S_1 excitation in Phe-$(Ala)_4$: (a) the occupied orbital and (b) the virtual orbital.

TABLE 5.3
LC-BLYP/6-31G* Excitation Energy Decomposition (in eV) for the β Strand of Phe-$(Ala)_n$ (n = 4, 8, 16)†

System	ω_M	I	$\Delta\omega_{MI}$
Phe-$(Ala)_4$	5.415	4	–0.035
		5	–0.017
Phe-$(Ala)_8$	5.415	6	–0.036
		7	–0.017
Phe-$(Ala)_{16}$	5.414	10	–0.036
		11	–0.017

†*Note:* FMO1-TDDFT values ω_M are complemented by pair corrections $\Delta\omega_{MI}$ from fragment dimers *MI*. See Figure 5.4 for numbering scheme.

5.3.3 Excitation Energy Decomposition Analysis for Polyalanine by FMO-TDDFT

The excitation energy expression of FMO2-TDDFT (Equation 5.8) is given by a sum of a single monomer fragment plus dimer corrections, which straightforwardly provides the excitation energy decomposition analysis. We introduce this excitation energy decomposition analysis by using the results of β-Phe-$(Ala)_n$ (n = 4, 8, 16). In the output file, the dominant fragment-pair contributions are given just under the excitation energy of FMO2-TDDDFT.

The main pair contributions are shown in Table 5.3 for β-Phe-$(Ala)_n$ (n = 4, 8, 16). Only two fragment dimers have substantial contributions to the excitation energy for all systems, and other contributions are smaller than 0.001 eV. These two important fragments occupy the same position for all n of β-Phe-$(Ala)_n$, on the right side of the TDDFT fragment (see Figure 5.4). If the peptide chain grows longer, no more fragments contribute significantly to the excitation energy. This leads to the same excitation energy for all n of β-Phe-$(Ala)_n$ (see Table 5.2).

5.3.4 Excited States in Solution: FMO-TDDFT/PCM Calculations

The solvent effects can be included[33] in FMO-TDDFT by using PCM. We briefly introduce the accuracy of this approach at the FMO/PCM[1] and FMO/PCM[1(2)] levels (see Section 5.2.4). We calculated the lowest singlet π-π* excitation of Phe-$(Ala)_n$ (n = 4, 8, 16) in water, at the LC-BLYP/6-31G* level. The solvent effects were included by the conductor-like PCM (C-PCM[38,39]). The geometries and the fragmentation scheme are the same as those of Section 5.3.2. In addition to FMO, standard TDDFT/PCM calculations were performed for comparison.

The results are shown in Table 5.4. FMO2-TDDFT/PCM reproduced the excitation energies by TDDFT/PCM with the largest error of 0.005 eV. The two FMO/PCM[1] and FMO/PCM[1(2)] levels gave almost identical errors, which suggests that for polypeptides it may be possible to use the former method, which requires

TABLE 5.4
Excitation Energies and Solvent Shifts of Phe-(Ala)$_n$ (n = 4, 8, 16), Both in eV*

	TDDFT			FMO2-TDDFT				
	Gas Phase	PCM		Gas Phase	FMO/PCM[1]		FMO/PCM[1(2)]	
System	Energy	Energy	Shift	Energy	Energy	Shift	Energy	Shift
β-Phe-(Ala)$_4$	5.367	5.376	0.009	5.363	5.371	0.008	5.372	0.009
β-Phe-(Ala)$_8$	5.366	5.374	0.008	5.362	5.370	0.008	5.370	0.008
β-Phe-(Ala)$_{16}$	5.365	5.372	0.007	5.360	5.368	0.008	5.368	0.008
α-Phe-(Ala)$_4$	5.408	5.411	0.003	5.396	5.411	0.015	5.412	0.016
α-Phe-(Ala)$_8$	5.402	5.416	0.014	5.402	5.415	0.013	5.416	0.014
α-Phe-(Ala)$_{16}$	—	5.415	—	5.401	5.414	0.013	5.413	0.012

**Note:* LC-BLYP/6-31G* is used.

fewer computations. One can also conclude that the solvent effects upon the excitation energy are well described by the PCM potential at the monomer level. It is also interesting that for both polypeptides and phenols in water, FMO2-TDDFT/PCM had negative errors (underestimated the excitation energies) if pair corrections were included in PCM (which has a more systematic behavior). The errors for the α helices were smaller than those for the β strands. This is partially attributed to the excitations in the α structures being more localized.[30]

Next, we determined the solvation shifts in the excitation energy comparing the solvated and gas phase results. The shifts are shown in Table 5.4. It can be seen that they are positive (i.e., blue shifts). For the α-Phe-(Ala)$_{16}$, the regular TDDFT calculation in gas phase failed to converge, as we previously mentioned. The α-helices have about 1.5 to two times larger solvent shifts than β-strands, which may be connected to the former having a larger dipole moment. Because the shift values are in general quite insensitive to the chain length, the dipole moment effects are local, as felt by the excited state residue (phenylalanine). FMO2-TDDFT accurately reproduced the regular full TDDFT solvation shifts. The errors in the solvation shifts are practically zero (0.001 eV or less), and these errors are smaller than those in the excitation energies, with one exception of α-Phe-(Ala)$_4$, which is considered in detail below.

The main excitations for several Phe-(Ala)$_n$ systems are shown in Figure 5.6. For α-Phe-(Ala)$_4$, the error in the solvation shift is relatively large (~0.013 eV). The reason for this is quite interesting. There happens to be two nearly degenerate excitations in this system in gas phase, which are represented by the two occupied-virtual orbital pairs shown in Figure 5.6. One excitation (two lower pairs) is localized on phenol and is the desired π-π* transition. The other excitation is delocalized over several neighboring residues. We picked the second singlet state that has the largest π-π* weight in the full TDDFT calculations as the reference against which we compare FMO2-TDDFT. Note that the TDDFT fragment in FMO-TDDFT is not large enough to describe both excited states, and it only contains the phenol ring excitation. It was found that by adding one alanine residue to the TDDFT fragment (which thus has

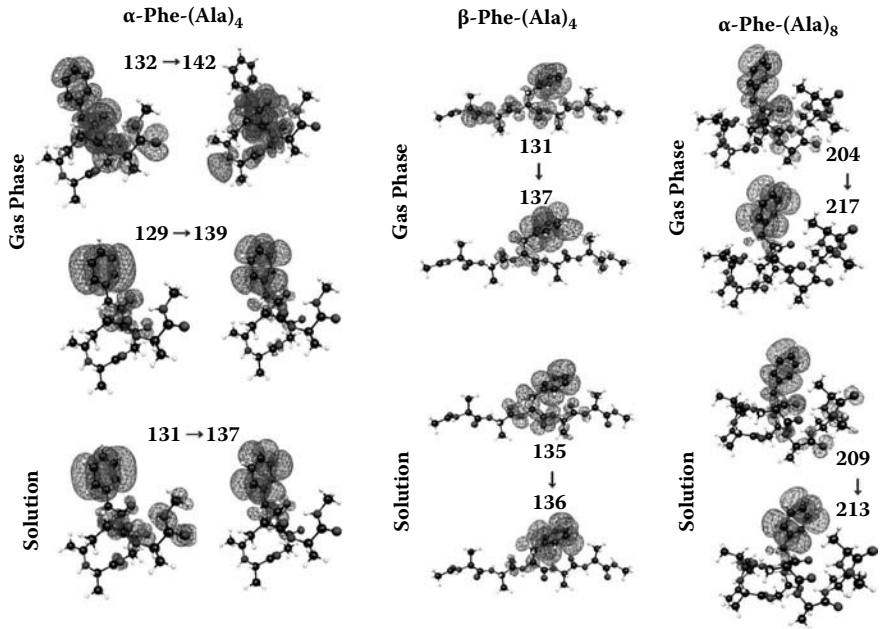


FIGURE 5.6 Molecular orbitals defining the major contributions to the lowest singlet excited state in Phe-$(Ala)_n$ systems (TDDFT in gas phase and in solution [PCM], LC-BLYP/6-31G*). The numbers indicate the serial orbital numbers, and the orbitals are listed in the occupied-virtual pairs. The two orbital phases are shown in red and blue. For Phe-$(Ala)_n$, there are two major contributions to the excited state of interest of nearly equal weight. (Reproduced from Chiba, M., Fedorov, D.G., Kitaura, K. 2008. *J. Comput. Chem.* 26: 2667–2676. With permission.) **(See color insert following page 117.)**

phenylalanine and alanine residues), the lowest singlet state is reproduced by FMO2-TDDFT with 0.004 eV error.

Except for α-Phe-$(Ala)_4$, other systems had simple transitions with the dominant weights of π-π*, as is illustrated in Figure 5.6. When PCM is included, all systems including α-Phe-$(Ala)_4$ had similar dominantly π-π* transitions. A general trend was observed to have more localized orbitals when solvent is present, which has a positive effect upon the FMO accuracy. The stronger localization is well pronounced in β-Phe-$(Ala)_4$, as can be seen in Figure 5.6.

The case of α-Phe-$(Ala)_4$ should serve as a warning against having very small TDDFT fragments, as one may not be able to obtain the proper excitations. It can be suggested that FMO2-TDDFT may well capture the pair corrections for the unconnected (not covalently bound) residues, but if the excitation of interest has the tendency for some delocalization within one or two covalently bound residues, all of them should be included in the TDDFT fragment. A simple solution may be to include both left and right neighbors of a residue of main interest, if the electronic structure details are unknown. One can also first perform an FMO-DFT calculation

TABLE 5.5
Central Processing Unit (CPU) Timings (in Hours) for the Calculations of the S_1 State of PhOH+$(H_2O)_n$ (n = 16, 32, 64)*

n	FMO1-TDDFT	FMO2-TDDFT	TDDFT
16	0.4	1.8	2.3
32	0.7	3.5	13.1
64	1.3	6.1	68.9

**Note:* Using LC-BLYP and 6-31G* with a single 3.2 GHz pentium4 node.

of this kind and plot the orbitals to see if the orbitals are sufficiently localized so that the TDDFT fragment size can be reduced.

5.3.5 Efficiency of FMO-TDDFT

We briefly introduce the timings for FMO-TDDFT. The CPU timings used in calculations of PhOH + $(H_2O)_n$ (n = 16, 32, 64) are shown in Table 5.5, and they are also plotted in Figure 5.7, together with the full TDDFT timings. It is obvious that FMO-TDDFT is by far faster than the full TDDFT calculations, and the scaling of both FMO1- and FMO2-TDDFT is nearly linear. The basis for the general nearly linear scaling of FMO is discussed in general in Fedorov and Kitaura.[40]

In some sense, the FMO-TDDFT scaling may be thought to be sublinear. That is, FMO-TDDFT requires a linear scaling load of DFT calculations, and the TDDFT cost is practically independent of the system size, if the above approximation is used to limit the number of the TDDFT dimer calculations from formally linear to a value independent of the system size. This sublinear scaling should become even more

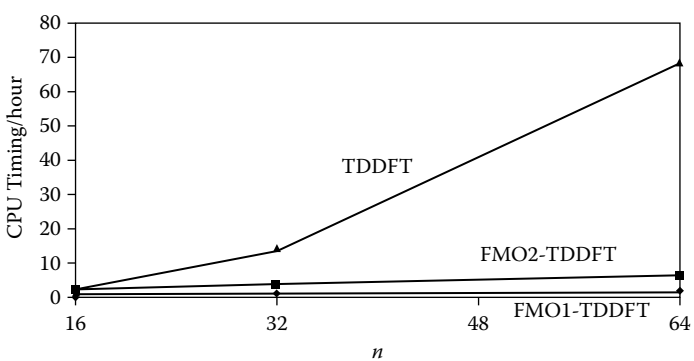


FIGURE 5.7 The CPU timings (hour) for PhOH+$(H_2O)_n$ (n = 16, 32, 64), computed with FMO1-TDDFT (diamonds), FMO2-TDDFT (squares), and TDDFT (triangles).

obvious if one uses the multilayer formalism,[29] and describes the electron densities of the environment with RHF, in which case the total cost of an FMO-RHF:TDDFT calculation is given by a nearly linear scaling of the FMO1-RHF part plus a constant for the TDDFT work.

5.4 HOW TO RUN FMO-TDDFT CALCULATIONS

5.4.1 How to Choose TDDFT Fragments

In FMO-TDDFT calculations, we need to define the TDDFT fragment (the fragment *M* in Equation 5.8 and Equation 5.9), which should be chosen carefully because the excitation is supposed to occur mostly on this fragment in FMO-TDDFT.

In the low-lying excitations of phenol+$(H_2O)_n$ (see Section 5.3.1), we chose TDDFT fragment as phenol, because the excitations of interest occur mainly around phenol. From Table 5.1, we found this choice of TDDFT fragment leads to a good accuracy in FMO2-TDDFT, but this choice produces a relatively large error at the FMO1-TDDFT level. This means that pair corrections to FMO1-TDDFT energy recover the missing many-body effects. If we add four waters, which are near the OH group of phenol, into the TDDFT fragment (i.e., the TDDFT fragment becomes phenol+$(H_2O)_4$), the errors in excitation energy at the FMO1-TDDFT level become less than 0.01 eV.[29]

Thus, by including four water molecules, one can stop at the FMO1-TDDFT level, whose demerit is the need to compute a larger (11 + 12 = 23 atoms) TDDFT fragment, whereas for FMO2-TDDFT with PhOH taken as the TDDFT fragment, one only computes TDDFT of at most 11 + 3 = 14 atoms. The merit of FMO1-TDDFT is the full many-body effect inclusion within the TDDFT fragment.

In the lowest singlet excitations of phenylalanine-(alanine)*n* case (see Section 5.3.2), we chose the TDDFT fragment as phenylalanine, because the excitations of interest occur mainly around it. From Table 5.2, we found this choice of TDDFT fragment leads to accuracy (errors in the excitation energies are less than 0.01 eV) in FMO2-TDDFT and this choice produces a relatively large error at the FMO1-TDDFT level, just like for phenol+$(H_2O)_n$. If we add two alanines near phenylalanine into the TDDFT fragment (i.e., the TDDFT fragment becomes phenylalanine-$(alanine)_2$), the errors in excitation energy at the FMO1-TDDFT level become less than 0.01 eV.[29]

In FMO1-TDDFT, the choice of the TDDFT fragment is crucial, because only the TDDFT fragment is treated by TDDFT (see Equation 5.9). In FMO2-TDDFT, the choice of the TDDFT fragment is important, but to a lesser degree. In addition to the TDDFT fragment, FMO2-TDDFT considers contributions from fragment dimers, including the TDDFT fragment, in order to include the many-body effects in excitations.

5.4.2 How to Choose the Level of FMO-TDDFT

At present, there are two levels to choose from: FMO1-TDDFT or FMO2-TDDFT. The former includes one TDDFT calculation on a single fragment of interest, in the electrostatic field determined by the monomer SCF process (see Chapter 2). The latter adds to the former dimer corrections, which bring in the many-body effects to the excitations (see Equation 5.8).

FMO2-TDDFT is a safer choice, because the selection of the TDDFT fragment can be a serious problem at the FMO1-TDDFT level as mentioned above. The number of fragment dimers that include the TDDFT fragment is not large, because we use a threshold to avoid unnecessary calculations of far separated pairs (which contribute negligibly). Overall, FMO2-TDDFT is only marginally more expensive than FMO1-TDDFT, especially considering the need to use a larger fragment of the latter.

5.4.3 Choosing the Functional Basis Set and Other Parameters

FMO-TDDFT calculation is available in the distribution version of GAMESS[41] parallelized by GDDI.[42] It is known that a hybrid functional or a long-range corrected (LC) functional usually have good accuracy in the excitation energies by TDDFT, while pure functionals often perform poorly.[4,37] B3LYP[43] (hybrid functional), and BOP[34,44] or BLYP with the long-range correction (LC-BOP, LC-BLYP)[36,37] are available for TDDFT calculation on GAMESS. B3LYP, which is a conventional hybrid functional, usually gives good excitation energies for most of valence excitations. However, for charge transfer excitations and Rydberg excitations, B3LYP typically underestimates the excitation energies.[37,45] On the other hand, the long-range corrected functionals, LC-BOP and LC-BLYP, give better excitation energies for these types of excitations.[37]

For TDDFT excitation energy calculation, at least a 2ζ + polarization level basis set is required. Furthermore, to treat a spreading excited state (e.g., a Rydberg state), diffuse basis functions are needed, which are difficult to treat in FMO calculations in general, but can be used in the multilayer scheme (see Section 5.2.3).[24]

In FMO-TDDFT/PCM, one can use C-PCM[8,39] or IEF-PCM with the iterative solver.[46] For the reliable estimation of the solvation energies, we used 240 tesserae per atom,[33] and it may be possible to use a smaller density of 60 tesserae to obtain excited states. This would reduce the extra FMO-TDDFT/PCM cost, which is quite considerable. In addition, one should choose the PCM radii, which affect the solvation energies. However, because the electrostatic effect upon the excitation energies is fairly small (though not negligible), one can expect that the effect of the particular PCM radii choice may be not very important in TDDFT/PCM. Some explicit water molecules may need to be added to describe the solute–solvent charge transfer effect upon the excitations.[16]

5.5 A Practical Application of FMO-TDDFT: The Lowest Singlet Excitation of the Photoactive Yellow Protein in Gas Phase/Aqueous Solution

We applied FMO-TDDFT to the lowest singlet π-π^* excitation of the photoactive yellow protein (PYP, PDB:2PHY) in gas/water phase. PYP is a small protein made of 1931 atoms and 125 amino residues (see Figure 5.8a) and is found in the light sensor of *Ectothiorhodospira halophila*. Its X-ray structure was determined in 1995.[47] The chromophore of PYP is *p*-coumaric (Figure 5.9a) acid linked by a thioester bond with Cys69 and absorbs 446 nm (2.78 eV) blue light.[48] PYP is a water-soluble

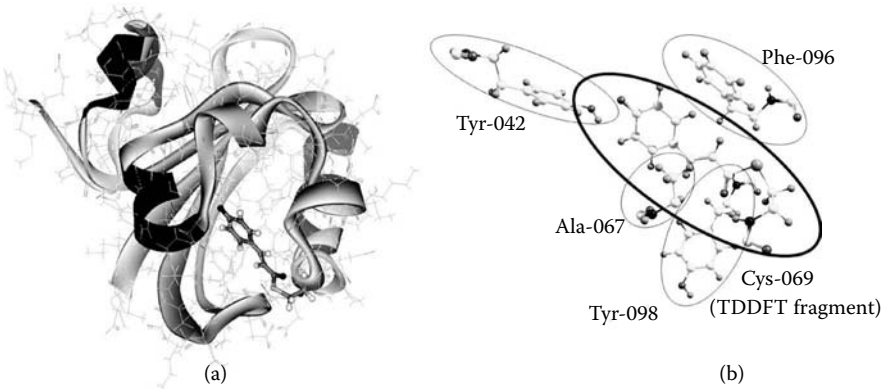


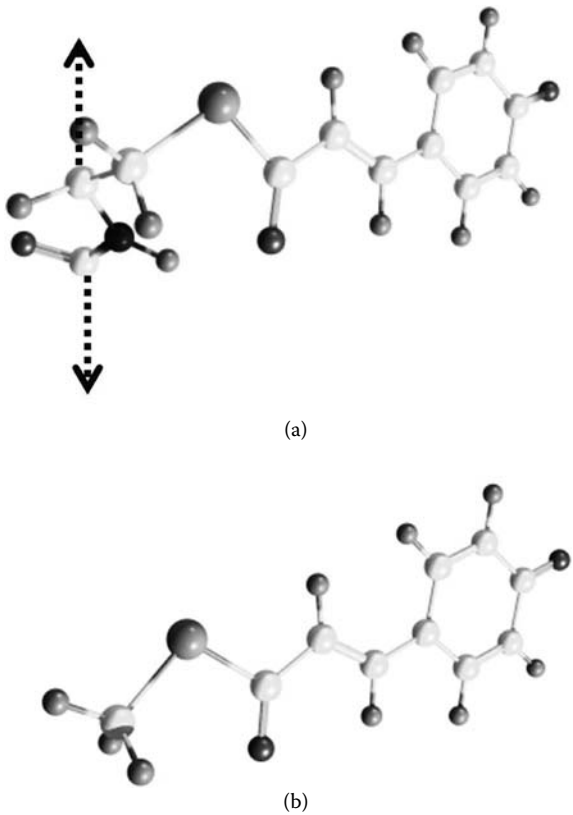
FIGURE 5.8 (a) The structure of the photoactive yellow protein (PDB: 2PHY). The chromophore part is shown as thick sticks, and the main residue factors lowering the excitation energy of the chromophore are shown with sticks. (b) The central part around the chromophore Cys69.

protein and its photoisomerization process occurs in the water phase. Mochizuki et al. calculated this system in gas phase employing the FMO-based CI, and obtained the excitation energy of PYP as 4.28 eV by CIS, and as 3.29 eV by CIS(D) using the 6-31G basis set.[25,26]

The PYP structure was obtained[30] by optimizing the chromophore geometry with B3LYP/6-31(+)G* and inserting it to the experimental geometry of PYP. PYP was fragmented as one residue per fragment, except that small glycine residues were appended to their neighbors. The total number of fragments was 112, and the chromophore was chosen as the TDDFT fragment. To introduce the solvent effects, we employed FMO2-TDDFT/PCM[1] and LC-BLYP/6-31G*, which took 90 hours using 24 nodes of the Soroban cluster (3.2 GHz Pentium4 nodes), while the gas phase FMO2-TDDFT calculation took about 26 hours. Thus, the computational cost increased about 64 hours by adding the PCM terms to FMO-DFT/TDDFT.

The results of both gas phase and PCM calculations are given in Table 5.6 as (1) ~ (5). (1) is the excitation energy of the model chromophore (see Figure 5.9b) calculated by TDDFT. The observed excitation energy of PYP is also shown in Table 5.6 as (6) for comparison. The solvent shifts induced by PCM are –0.211 eV for FMO1-TDDFT level (4) to (2) in Table 5.6, and –0.113 eV for FMO2-TDDFT level [(5) to (3) in Table 5.6]. It can be seen that the electrostatic effect of the residues in gas phase (model chromophore versus protein FMO1-TDDFT) lowers the excitation energy by 0.084 eV [(2) to (1) in Table 5.6]. The solvent also has a considerable effect upon the excitation energy.

Pair corrections add explicit quantum-mechanical contributions to the excitation energy (Equation 5.8), and this gives the excitation energy decomposition analysis (see Section 3.3). Their cumulative effect in the gas phase and solution is further lowering by 0.309 eV [(3) to (2)] and 0.211 eV [(5) to (4)], respectively. It is rather interesting that pair corrections became considerably smaller in solution. In the neutral



(a)

(b)

FIGURE 5.9 (a) The TDDFT fragment in the FMO-based TDDFT calculations of PYP, which corresponds to the chromophore. The two dashed lines represent the covalent bonds connecting to the neighboring fragments. (b) The structure of the chromophore model compound of PYP ($C_{10}H_9O_2S$).

model polypeptides, this trend was not observed: pair corrections were similar in gas phase and solution. Explicit dimer contributions to the excitation energy are listed in Table 5.7, and the most contributing four fragments and chromophore TDDFT fragment (Cys-69) are shown in Figure 5.8b. The addition of solvent lowers many of them, which suggests that the excitation becomes more localized in solution.

The anionic part of the chromophore in PYP is buried inside the protein, and the solvent interacts more with other residues, which transmit the solvent influence into the excitation energy shift, obtained to be –0.113 eV [(5) to (3) in Table 5.6, a red shift]. The role of the explicit solvent upon the solvation energies is in general fairly small, as it is made of the energetic difference of solvent–solvent hydrogen bonds replaced by those between the solute and solvent. However, in general, if the active part of the chromophore is exposed to the solvent, it may be necessary to consider some explicit solvent molecules, which by forming hydrogen bonds alter the electronic structure of the chromophore through charge transfer, and can considerably

TABLE 5.6
Excitation Energy *E* (in eV) of the Lowest Singlet Excitation of the Photoactive Yellow Protein*

Method	System	Level	*E*
(1)	Chromophore[a]	TDDFT	3.516
(2)	Protein[a]	FMO1-TDDFT	3.433
(3)	Protein[a]	FMO2-TDDFT	3.124
(4)	Protein[b]	FMO1-TDDFT/PCM[1]	3.222
(5)	Protein[b]	FMO2-TDDFT/PCM[1]	3.011
(6)	Protein[c]	Experiment	2.78

[a] Chromophore denotes the chromophore part of the protein capped with hydrogens. (See Chiba, M., Fedorov, D.G., Kitaura, K. 2007. *J. Chem. Phys.* 127: 104108/1-11.)

[b] See Chiba, M., Fedorov, D.G., Kitaura, K. 2008. *J. Comput. Chem.* 29: 2667–2676.

[c] Imamoto, Y., Kataoka, M., Tokunaga, F. 1996. *Biochemistry* 35: 14047–14053.

[*] *Note:* LC-BLYP/6-31G* is used.

TABLE 5.7
Residues with the Largest Effect upon the Chromophore Excitation Energy in the Photoactive Yellow Protein*

I	R_{MI}	$\Delta\omega_{MI}^{\text{gas phase}}$	$\Delta\omega_{MI}^{\text{PCM}}$
Ala-067	0.88	–0.050	–0.040
Phe-096	0.98	–0.044	–0.026
Tyr-098	0.86	–0.034	–0.027
Tyr-042	0.84	–0.026	–0.019
Thr-070	0	–0.024	–0.020
Glu-046	0.81	–0.022	–0.016
Thr-050+Gly-051[a]	0.86	–0.017	–0.011
Phe-062	1.11	–0.015	–0.008

[a] All glycine residues were appended to their neighbors.

[*] *Note:* FMO2-TDDFT with LC-BLYP/6-31G* is used; gas phase and solvated (PCM) pair corrections $\Delta\omega_{MI}$ to the excitation energy (eV) are listed for fragment dimers *MI* including the chromophore *M* and another fragment *I*, along with the unitless interfragment distances R_{MI}.

change the excitation energy in the way that a continuum model like PCM cannot take into account.

A number of theoretical studies of PYP were conducted,[49] which typically proceed by considering the chromophore alone and with a few neighboring residues. For instance, Gromov et al.[49] studied the chromophore and up to seven amino residues from PYP with correlated *ab initio* methods. When comparing such studies to our approach, it is helpful to understand the difference in the analysis. The fragment definition in FMO is shifted by one carbonyl group with respect to the conventional residue division, which makes a direct comparison more difficult. (We distinguish fragments from residues by adding a dash, so that Ala-67 is the fragment residue -CONHC H(CH_3)- in FMO, shifted to the conventional Ala67 residue -NHCH(CH_3) CO- by one CO group.)

The pair corrections in FMO, computed for this fully polarized (electrostatically relaxed) state of fragments, are made of the electrostatic interaction between such polarized densities, charge transfer, exchange correlation and dispersion (if included), as can be decomposed in gas phase using the pair interaction energy decomposition analysis.[50] It appears as a result of our study that some residues are important contributors to the excitation energy lowering, which were not considered in some other studies, such as Ala-67, Phe-62, and Phe-96. We also note that the active part of the chromophore is not facing the surface directly, and the solvent influence upon the excitations is largely mediated by other residues, because of both electrostatic quantum-mechanical effects (e.g., charge transfer).

On the other hand, Arg52 was reported[51] to have a large stabilizing effect owing to charge transfer. It appears that such influence is considerable for the ground state, and the pair correction to the excitation energy due to Arg-52 is not significant (–0.008 and –0.003 eV, in gas phase and solution, respectively), in agreement with the other prediction made by *ab initio* studies,[49] which concluded that the large apparent affect of Arg52 upon the excitation energy is an artifact of the small models, and it disappears when larger parts of the protein are considered.

5.6 CONCLUSION

TDDFT is widely used in excited state calculations due to its efficiency and accuracy. However, applications of TDDFT to gigantic systems are still impossible because the computational order of TDDFT is $O(N^3)$, where N is the number of basis set. In this chapter, we introduced the fragment molecular orbital–based TDDFT (FMO-TDDFT) at the two levels of the many-body expansion. In FMO1-TDDFT, one performs a TDDFT calculation on a specific fragment, in the electrostatic field produced by other fragments. FMO2-TDDFT adds explicit many-body corrections to the excitation energies of FMO1-TDDFT. Computationally, both levels are efficient, as they involve only converging fragment monomer densities, and one or a small number of TDDFT calculations (if only the excitation energy is of interest). Using this FMO-based scheme, we can apply TDDFT to gigantic systems such as proteins.

To demonstrate the accuracy of FMO-TDDFT, we calculated several low-lying excited states in the solvated phenol and polypeptides and established the accuracy in a detailed comparison to full TDDFT calculations. Typically, the errors for the

two-body method were below 0.1 eV, although for one excited state in the solvated phenol an error of about 0.2 eV was observed. In order to treat excitations in solvent, we also established FMO-TDDFT/PCM[31] and a similar accuracy was obtained.

The FMO-based TDDFT method requires some knowledge regarding the spatial location where the excitation takes places. In the one-body case (FMO1-TDDFT), this choice of the TDDFT fragment is crucial to successful calculation, and for the two-body method (FMO2-TDDFT) it is important, but to a lesser degree, and one can verify if the choice was appropriate by considering the magnitude of the pair corrections. These corrections constitute the excitation energy decomposition analysis, which can be valuable in establishing the contribution of the environment upon the excitation energy. By using this analysis, which complements the ground state energy analysis in FMO,[50] we found why the lowest excitation energy of the polypeptide chain does not significantly change when the residue chain grows beyond two residues in length. The computational scaling was demonstrated to be nearly linear, and the timings were drastically smaller than those for the conventional TDDFT.

As an example of a protein calculation, we calculated the lowest singlet excitation energy of the photoactive yellow protein (1931 atoms) at the LC-BLYP/ 6-31G* level with and without the solvent effects of water. The obtained values of 3.12 eV (gas phase) and 3.01 eV (water phase) are in agreement with the experimental absorption energy 2.78 eV. Also, the excitation energy decomposition analysis by FMO2-TDDFT provided the energy contributions of each residue to the excitation energy in gas/water phase.

Recently, Fukunaga et al.[52] calculated the excitation energies of several crystalline phases of quinacridone, which are important organic pigments. The solid-state environment was modeled by a large cluster (>10,000 atoms) that was represented by point charges, obtained from DFT calculations with periodic boundary conditions. The electrostatic field in FMO was also frozen, using these charges, to avoid an artificial introduction of edge effects. The inner part of the cluster was treated by both regular TDDFT (5 molecules) and FMO2-TDDFT (up to 20 molecules). It was found that due to orbital degeneracy, the accuracy of FMO-TDDFT for a cluster representing the solid state is somewhat lower than may be expected for a similarly sized molecule, and relative to regular TDDFT, the errors were 0.03 to 0.15 eV. Nevertheless, FMO-TDDFT provided much information on the role of molecular interactions in determining the excitation energies of the three crystalline phases.

As a future work, we are currently developing the excitation energy gradient method of FMO-TDDFT. By using the energy gradient of FMO-TDDFT, excited state geometry optimizations by FMO-TDDFT will become possible. This method can be very useful in order to investigate the fluorescence of large systems, where excited state geometry relaxation is important. We expect that FMO-TDDFT will become a powerful tool for the study of excited states in large systems.

ACKNOWLEDGMENTS

This work was supported by the Next Generation Super Computing Project (MEXT, Japan), and a Grant-in-Aid for Scientific Research (JSPS, Japan).

REFERENCES

1. Runge, E., Gross, E.K.U. 1984. Density-functional theory for time-dependent systems. *Phys. Rev. Lett.* 52: 997–1000.
2. Casida, M.E. 1995. In *Recent Advances in Density-Functional Methods,* ed. Chong, D.P., 155–192. World Scientific: Singapore.
3. Bauernschmitt, R., Ahlrichs, R. 1996. Treatment of electronic excitations within the adiabatic approximation of time dependent density functional theory. *Chem. Phys. Lett.* 256: 454–464.
4. Stratmann, R.E., Scuseria, G.E., Frisch, M.J. 1998. An efficient implementation of time-dependent density-functional theory for the calculation of excitation energies of large molecules. *J. Chem. Phys.* 109: 8218– 8224.
5. Furche, F., Ahlrichs, R. 2002. Adiabatic time-dependent density functional methods for excited state properties. *J. Chem. Phys.* 117: 7433–7477. (Erratum: 2004. *J. Chem. Phys.* 121: 12772.)
6. Nakatsuji, H. 1978. Cluster expansion of the wavefunction: excited states. *Chem. Phys. Lett.* 59: 362–364.
7. Nakatsuji, H., Hirao, K. 1978. Cluster expansion of the wavefunction: symmetry-adapted-cluster (SAC) expansion, its variational determination, and extension of open-shell orbital theory. *J. Chem. Phys.* 68: 2053–2065.
8. Foresman, J.B., Head-Gordon, M., Pople, J.A., Frisch, M.J. 1992. Toward a systematic molecular-orbital theory for excited-states. *J. Phys. Chem.* 96: 135–149.
9. Maurice, D.B., Head-Gordon, M. 1999. Analytical second derivatives for excited electronic states using the single excitation configuration interaction method: theory and application to b enzoapyrene and chalcone. *Mol. Phys.* 96: 1533–1541.
10. Oddershede, J. 1987. Propagator methods. *Adv. Chem. Phys.* 69: 201–239.
11. Roos, B.O. 1987. In Ab Initio *Methods in Quantum Chemistry,* ed. Lawley, K.P., 399–446. Wiley: New York.
12. Andersson, K., Malmquist, P.A., Roos, B.O. 1992. Second-order perturbation theory with a complete active space self-consistent field reference function. *J. Chem. Phys.* 96: 1218–1226.
13. Malmquist, P.A., Roos, B.O. 1989. The CASSCF state interaction method. *Chem. Phys. Lett.* 155: 189–194.
14. Tomasi, J., Persico, M. 1994. Molecular interactions in solution: an overview of methods based on continuous distributions of the solvent. *Chem. Rev.* 94: 2027–2094.
15. Tomasi, J., Mennucci, B., Cammi, R. 2005. Quantum mechanical continuum solvation models. *Chem. Rev.* 105: 2999–3093.
16. Cossi, M., Barone, V. 2001. Time-dependent density functional theory for molecules in liquid solutions. *J. Chem. Phys.* 115: 4708–4717.
17. Guillemoles, J.F., Barone, V., Joubert, L., Adamo, C. 2002. A theoretical investigation of the ground end excited states of selected Ru and Os polypyridyl molecular dyes. *J. Phys. Chem. A* 106: 11354–11360.
18. Yan, W.Z., Xue, Y., Zhu, H., Zeng, J., Xie, D.Q. 2004. A theoretical study of solvent effects on tautomerism and electronic absorption spectra of 3-hydroxy-2-mercaptopyridine and 2,3-dihydroxypyridine. *J. Comput. Chem.* 2: 1833–1839.
19. Ridley, J., Zerner, M.C. 1973. An intermediate neglect of differential overlap technique for spectroscopy: pyrrole and azines. *Theor. Chem. Acta.* 32: 111–134.
20. Fujimoto, K., Hasegawa, J., Hayashi, S., Kato, S., Nakatsuji, H. 2005. Mechanism of color-tuning in retinal proteins: SAC-CI and QM/MM study. *Chem. Phys. Lett.* 414: 239–242.

21. Kitaura, K., Ikeo, E., Asada, T., Nakano, T., Uebayasi, M. 1999. Fragment molecular orbital method: an approximation computational method for large molecules. *Chem. Phys. Lett.* 313: 701–706.
22. Fedorov, D.G., Kitaura, K. 2006. In *Modern Methods for Theoretical Physical Chemistry and Biopolymers,* eds. Starikov, E.B., Lewis, J.P., Tanaka, S., 3–38. Elsevier: Amsterdam.
23. Fedorov, D.G., Kitaura, K. 2007. Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. *J. Phys. Chem. A* 111: 6904–6914.
24. Fedorov, D.G., Ishida, T., Kitaura, K. 2005. Multilayer formulation of the fragment. molecular orbital method (FMO). *J. Phys. Chem. A* 109: 2638–2646.
25. Mochizuki, Y., Koikegami, S., Amari, S., Segawa, K., Kitaura, K., Nakano, T. 2005. Configuration interaction singles method with multilayer fragment molecular orbital scheme. *Chem. Phys. Lett.* 406: 283–288.
26. Mochizuki, Y., Tanaka, K., Yamashita, K. et al. 2007. Parallelized integral-direct CIS(D) calculations with multilayer fragment molecular orbital scheme. *Theor. Chem. Acc.* 117: 541–553.
27. Hirata, S., Valiev, M., Dupuis, M., Xantheas, S.S., Sugiki, S., Sekino, H. 2005. Fast electron correlation methods for molecular clusters in the ground and excited states. *Mol. Phys.* 103: 2255–2265.
28. Nakatsuji, H., Miyahara, T., Fukuda, R. 2007. SAC(symmetry adapted cluster)/SAC-CI(configuration interaction) methodology extended to giant molecular systems: ring molecular crystals. *J. Chem. Phys.* 126: 084104/1-18.
29. Chiba, M., Fedorov, D.G., Kitaura, K. 2007. Time-dependent density functional theory with the multilayer fragment molecular orbital method. *Chem. Phys. Lett.* 444: 346–350.
30. Chiba, M., Fedorov, D.G., Kitaura, K. 2007. Time-dependent density functional theory based upon the fragment molecular orbital method. *J. Chem. Phys.* 127: 104108/1-11.
31. Chiba, M., Fedorov, D.G., Kitaura, K. 2008. Polarizable continuum model with the fragment molecular orbital-based time-dependent density functional theory. *J. Comput. Chem.* 26: 2667–2676.
32. Fedorov, D.G., Kitaura, K. 2005. Multiconfiguration self-consistent-field theory based upon the fragment molecular orbital method. *J. Chem. Phys.* 122: 054108.
33. Fedorov, D.G., Kitaura, K., Li, H., Jensen, J.H., Gordon, M.S. 2006. The polarizable continuum model (PCM) interfaced with the fragment molecular orbital method (FMO). *J. Comput. Chem.* 27: 976–985.
34. Becke, A.D. 1988. Density-functional exchange-energy approximation with correct asymptotic-behavior. *Phys. Rev. A* 38: 3098–3100.
35. Lee, C., Yang, W., Parr, R.G. 1988. Development of the Colle–Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* 37: 785–789.
36. Iikura, H., Tsuneda, T., Yanai, T., Hirao, K. 2001. A long-range correction scheme for generalized-gradient-approximation exchange functionals. *J. Chem. Phys.* 115: 3540–3544.
37. Tawada, Y., Tsuneda, T., Yanagisawa, S., Yanai, T., Hirao, K. 2004. A long-range-corrected time-dependent density functional theory. *J. Chem. Phys.* 120: 8425–8433.
38. Barone, V., Cossi, M. 1998. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A* 102: 1995–2001.
39. Klamt, A., Schuurmann, G. 1993. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc. Perkin. Trans.* 2: 799–805.

40. Fedorov, D.G., Kitaura, K. 2005. Coupled-cluster theory based upon the fragment molecular-orbital method. *J. Chem. Phys.* 123: 134103/1-11.
41. Schmidt, M.W., Baldridge, K.K., Boatz, J.A. et al. 1993. General atomic and molecular electronic structure system. *J. Comput. Chem.* 14: 1347–1363.
42. Fedorov, D.G., Olson, R.M., Kitaura, K., Gordon, M.S., Koseki, S. 2004. A new hierarchical parallelization scheme: generalized distributed data interface (GDDI), and an application to the fragment molecular orbital method (FMO). *J. Comput. Chem.* 25: 872–880.
43. Becke, A.D. 1993. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* 98: 5648–5652.
44. Tsuneda, T., Suzumura, T., Hirao, K. 1999. A new one-parameter progressive Colle-Salvetti-type correlational functional. *J. Chem. Phys.* 110: 10664–10678.
45. Dreuw, A., Weisman, J.L., Head-Gordon, M. 2003. Long-range charge-transfer excited states in time-dependent density functional theory require non-local exchange. *J. Chem. Phys.* 119: 2943–2946.
46. Li, H., Jensen, J.H. 2004. Improving the efficiency and convergence of geometry optimization with the polarizable continuum model: new energy gradients and molecular surface tessellation. *J. Comput. Chem.* 25: 1449–1462.
47. Borgstahl, G.E.O., Williams, D.-W.R., Getzoff, E.D. 1995. 1.4 A structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore. *Biochemistry* 34: 6278–6287.
48. Imamoto, Y., Kataoka, M., Tokunaga, F. 1996. Photoreaction cycle of photoactive yellow protein from *Ectothiorhodospira halophila* studied by low-temperature spectroscopy. *Biochemistry* 35: 14047–14053.
49. Gromov, E.V., Burghardt, I., Koppel, H., Cederbaum, L.S. 2007. Electronic structure of the pyp chromophore in its native protein environment. *J. Am. Chem. Soc.* 129: 6798–6806, and references therein.
50. Fedorov, D.G., Kitaura, K. 2007. Pair interaction energy decomposition analysis. *J. Comput. Chem.* 28: 222–237.
51. Groenhof, G., Lensink, M.F., Berendsen, H.J.C., Snijders, J.G., Mark, A.E. 2002. Signal transduction in the photoactive yellow protein. I. Photon absorption and the isomerization of the chromophore. *Proteins: Struct. Funct. Genet.* 48: 202–211.
52. Fukunaga, H., Fedorov, D.G., Chiba, M., Nii, K., Kitaura, K. 2008. Theoretical analysis of the intermolecular interaction effects on the excitation energy of organic pigments: solid state quinacridone. *J. Phys. Chem. A* 112: 10887–10894.

6 FMO-MD *An* Ab Initio*-Based Molecular Dynamics of Large Systems*

Yuto Komeiji

CONTENTS

6.1 ALGORITHM AND IMPLEMENTATION OF FMO-MD

Fragment molecular orbital–based molecular dynamics (FMO-MD) is an *ab initio* molecular dynamics (MD) method that simulates the dynamics of a molecular system while considering the configuration-dependent change of the electronic structure by the FMO method.[1] By taking advantage of the highly parallelizable nature of the FMO method, FMO-MD is expected to play an important role in the simulation of chemical and biochemical reactions in explicit solvation. In this section, the algorithm and implementation of the FMO-MD method are presented in detail. Also proposed is a temporary standard simulation protocol.

6.1.1 Principles and Implementation

The FMO-MD method is based on the Born-Oppenheimer approximation,[2] which assumes that electronic motion is much faster than nuclear motion. This approximation allows for the electronic structure calculation of the force (energy gradient) for a fixed geometry of molecules. Hence, in FMO-MD, the force (**F**) exerted on the

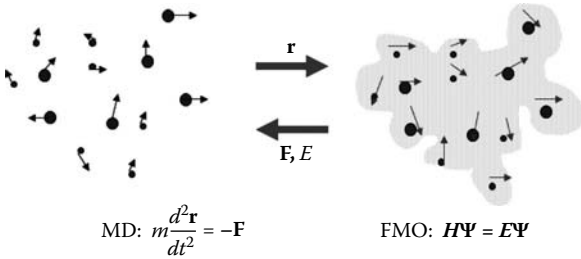


FIGURE 6.1 The FMO-MD simulation method. In FMO-MD, the energy gradients in terms of the atomic nuclei (**F**) and the potential energy (*E*) are calculated by FMO and the positions of the nuclei (**r**) are updated by MD, which are fed to FMO. (Adapted from Komeiji, Y., Mochizuki, Y., Nakano, T., Fedorov, D.G. 2009. *J. Mol. Struct.: THEOCHEM* 2009. With permission.)

nuclei along with the potential energy (*E*) is calculated by the FMO method, and the positions of the nuclei (**r**) are updated by an ordinary MD procedure (Figure 6.1). See Kitaura et al.[3] for the explicit presentation of the energy gradient. Thus, the principle and algorithm of FMO-MD are simple and straightforward.

The FMO-MD method was implemented for the first time in 2003[1,4] and renewed in 2006.[5,6] In both implementations, an MO software ABINIT-MP[7] and an MD software PEACH[8] were combined. The original system of 2003 was only capable of simulations of short polypeptides with or without a few surrounding solvent molecules, at STO-3G, and for 1 ps or so, but the system of 2006 is able to treat small molecules in explicit solvent with a larger basis set of 6-31G* as long as several picoseconds, owing to the improvement of both programs.

The software architecture of the new PEACH-ABINIT-MP system is illustrated in Figure 6.2. The programs are connected by a shell script named abinitrun.sh. The ABINIT-MP program is called from the PEACH program via the systemcall command and calculates the force and potential needed by PEACH to perform the time integration.

For the part of MD, many conventional algorithms are available with FMO-MD. For example, the SHAKE/RATTLE bond constraint,[9,10] Nosé-Hoover (chains) thermostat,[11,12] Gaussian thermostat,[13] blue moon ensemble,[14,15] and so on. For the part of FMO, because the calculation of the energy gradient is more difficult to implement than the energy, the current implementation of FMO-MD is limited to the FMO2-HF/6-31G(*) level. Another inevitable component of FMO-MD is the dynamic fragmentation implemented in the Auto-fragment program.

6.1.2 Dynamic Fragmentation

As already stated, the Auto-fragment program takes care of a procedure called "dynamic fragmentation" — that is, redefinition of the fragments of the molecular system during FMO-MD and FMO-energy minimization (EM) calculations.[5,16] In the FMO method, the molecular system of interest is divided into appropriate

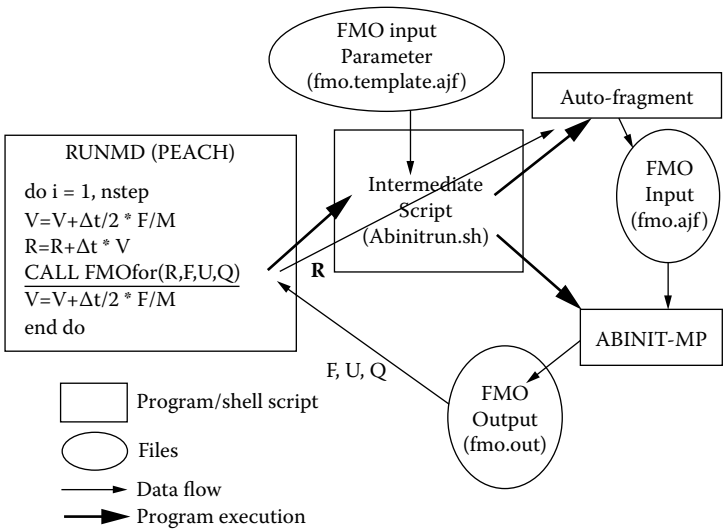


FIGURE 6.2 The software architecture of the new FMO-MD system (PEACH/ABINIT-MP). At each time step of MD, the "RUNMD" module of PEACH executes an intermediate shell script, "Abinitrun.sh," during MD simulation, which executes "Auto-fragment" to perform dynamic fragmentation and "ABINIT-MP" to perform a single-point FMO calculation. (Reproduced from Komeiji, Y., Ishikawa, T., Mochizuki, Y. et al. 2009. *J. Comput. Chem.* 30: 40–50. With permission.)

fragments so that each fragment can form a closed shell. No further fragmentation is necessary for a single-point energy calculation, but redefinition of the fragments sometimes becomes necessary in FMO-MD/EM because the molecules can undergo a large configuration change. We experienced several proton transfer reactions in the H_2CO simulation (see Section 6.2.1) during the heating-up process within the solvent.[16] We observed that upon such a transfer, the FMO-MD run, or more precisely the Self-Consistent-Field (SCF) stage of the FMO calculation, ended in failure. Therefore, we were urged to develop the algorithm of dynamic fragmentation.

Dynamic fragmentation is performed by the Auto-fragment program in three stages: (a) fragmentation of the solute molecules, (b) fragmentation of the solvent molecules, and, if necessary, (c) integration of the solute and solvent fragments. Each stage is conducted based on chemical intuition; that is, the chemical species that might appear are postulated *a priori*, and the molecular system is divided into these species. A practical example of this process is given for water.[5]

Currently, the Auto-fragment program must be modified depending on the simulated molecules. To avoid this tedious program modification, we are now trying to generalize the dynamic fragmentation by introducing a cluster analysis of the interatom connectivity, but such a generalized scheme is still in the developmental stage.

We should also mention that an accidental fragment rearrangement can often, if not always, be avoided by using a well-optimized initial structure. We observed frequent rearrangement of the fragments in the preparative stage of the H_2CO

simulation.[16] On the contrary, the FMO-MD simulations of pure water,[5] in which the initial configurations were optimized by classical MD/EM methods, showed no fragment rearrangement. Also, in the blue moon study of the Menschutkin reaction (see Section 6.2.3), we did not use the dynamic fragmentation scheme, treating each water molecule and the whole solute as independent fragments, but SCF of FMO always converged without any problem. In principle, a proton transfer, the origin of the fragment rearrangement, is an extremely rare event whose probability is as small as 10^{-7} in pure water. In the H_2CO simulation the initial water configuration was random, and no minimization or annealing was performed prior to the FMO-MD simulation. Thus, we now consider the frequent rearrangement seen in the H_2CO simulation to be an artifact caused by the crude simulation protocol. Based upon this experience, we are now using a standard simulation protocol.

6.1.3 Standard Simulation Protocol

Here, a practical, standard FMO-MD simulation protocol is presented. The advantage of FMO-MD is the fully *ab initio* (quantum) treatment of the simulated system, but the enormous computer resource needed for *ab initio* MD simulation necessitates a good initial configuration of the simulated molecules, with "good" meaning "close to the global minimum." Through the FMO-MD simulations of pure water,[5] we found that use of the classical MD/EM method in the preparative stage could reduce the computation time required for the subsequent FMO-MD simulation and that use of the bond constraint in the preparative stage saves computation time while avoiding undesirable fragment rearrangement caused by system instability (see Section 6.1.2). Thus, we are temporarily using the following protocol for simulation of a solvated molecule (solute) in explicit solvent:

1. Optimize the solute molecule in vacuum by *ab initio* MO calculation.
2. Generate solvent molecules around the solute.
3. Optimize the solvent by a classical MD/EM method while fixing the position and geometry of the solute.
4. Start FMO-MD by heating the solute/solvent system while fixing the equilibrium bond length using SHAKE/RATTLE.
5. Equilibrate the system for a while.
6. Start free FMO-MD simulation by removing the RATTLE constraint.

The above protocol is only a temporal standard and is not robust, but it serves as a reasonable starting point in FMO-MD simulation of solvated molecules.

6.2 PRACTICAL FMO-MD SIMULATIONS ON SOLVATED MOLECULES

In this section, three practical examples of FMO-MD simulations performed by the PEACH/ABINIT-MP system will be presented. All of them dealt with small molecules hydrated by explicit water solvent. The first example used FMO-MD as a conformation sampler, and the other two used FMO-MD as a reaction simulator.

6.2.1 Conformation Sampling for Excited State Calculation of Formaldehyde

The FMO-MD method was used to sample conformations of formaldehyde (H_2CO) in explicit solvent to prepare structures for the lowest n-π* state excitation energy calculations.[16] The solvent effect on the shift in excitation energy of a solute molecule is known as the solvatomichromism. In particular, the positive energy shift (blueshift) caused by hydration of the lowest n-π* state of carbonyl compounds such as formaldehyde has been extensively examined by theoretical methods.

The solvent-induced blueshift of the excitation energy was calculated as follows. In the preparative stage, a series of conformations of formaldehyde in the presence and absence of the explicit solvent were sampled by FMO-MD simulation at the HF/6-31G level. See Figure 6.3 for the initial configuration of the solvated formaldehyde. The molecular system was heated for 0.12 ps, followed by an FMO-MD run at 300 K for 2.5 ps. In the production stage, the excitation energy was calculated by the configuration interaction singles with perturbative doubles methods (CIS(D))[17] in the multilayer FMO framework at the HF/6-31G* level. A total of 400 sample configurations were chosen from the last 2 ps of the FMO-MD trajectories. As stated in Section 6.1.2, fragment rearrangements caused by a proton transfer occurred frequently during the preparative stage of the solvated simulation. Nonetheless, the rearrangement ended after the system was equilibrated, and hence we found no problem in sampling the equilibrated configurations. The configurations obtained in the gas phase were directly subjected to CIS(D) calculations, and the formaldehyde and six surrounding water molecules were chosen for the solvated configurations and were subjected to

FIGURE 6.3 The initial configuration of formaldehyde solvated by 128 water molecules. This configuration was subjected to FMO-MD at the HF/6-31G level, and configurations from the FMO-MD trajectory were subjected to CIS(D) calculation. See Section 6.2.1 for details. (Reproduced from Mochizuki, Y., Tanaka, K., Yamashita, K. et al. 2007. *Theor. Chem. Acc.* 117: 541–553. With permission.)

CIS(D) calculation in the presence of the environmental electrostatic potential from the remaining water molecules.

The calculated excitation energies were 4.22 ± 0.15 eV in solution and 4.08 ± 0.16 eV in vacuum, giving a blueshift value of 0.14 eV. This value was consistent with those given in several preceding computational studies. No experimental blueshift value is available for formaldehyde, but a value of 0.21 eV was determined for acetone, a similar carbonyl molecule. Thus, the blueshift calculated this time was comparable to the blueshift of the experiment. The special feature of this study was that all the computation procedures were conducted by full *ab initio* quantum mechanical methods.

6.2.2 Hydrolysis Simulation of a Methyl Diazonium Ion

The FMO-MD method was applied to the hydrolysis of methyl diazonium ion ($CH_3N_2^+$).[18] The hydrolysis is an S_N2-type substitution reaction that proceeds as follows:

$$H_2O + CH_3N_2^+ \rightarrow [H_2O\ldots CH_3^+\ldots N_2]\ddagger \rightarrow {}^+H_2OCH_3 + N_2$$

Traditionally, because the CH_3^+ cation is regarded as an unstable moiety that cannot form without interaction with H_2O, the trigger of the reaction has been ascribed to a nucleophilic attack by H_2O. This mechanism, in which the H_2O–CH_3^+ bond formation induces CH_3–N_2^+ bond cleavage, is known as the "enforced concerted mechanism," or "tight S_N2." However, a "step-wise mechanism," or "loose S_N2," is suggested in which the bond-cleavage and bond-formation steps occur successively.

The FMO-MD simulations of this reaction were performed at the HF/6-31G level in a droplet of water molecules confined within a sphere by a weak harmonic restraint (Figure 6.4). The temperature was raised to 1000 K to facilitate the reaction. From the resultant trajectory, 15 seed configurations were taken and subjected to further MD runs at 700 K. Ten trajectories were found to produce the substituted product ($^+H_2OCH_3$). These productive trajectories were analyzed to investigate the reaction path.

The productive trajectories showed different reaction paths, two extremes being the tight S_N2 and loose S_N2. A two-dimensional map of $R_{O\text{-}C}$ versus $R_{C\text{-}N}$ atomic distances (Figure 6.5) revealed that the ten productive trajectories fell within the two extremes, one in which the C–N bond cleavage was tightly coupled with O–C bond formation, and the other in which the coupling was relatively loose. The former (denoted as trajectory A, Figure 6.4a) and the latter (trajectory B, Figure 6.4b) were considered to present the tight S_N2 and loose S_N2, respectively. Furthermore, the Mulliken population analysis of the atomic charges and configuration analysis of fragment interaction (CAFI)[19] for the charge transfer (CT) revealed that CTs among atoms are concerted with the O–C bond formation in trajectory A but only loosely so in trajectory B (Figure 6.6).

The results of this study suggested the existence of a diversity of reaction paths, illustrating how the atoms in reacting molecules behave in solution at the molecular level from the reactant to the product. Such a simulation is only possible with

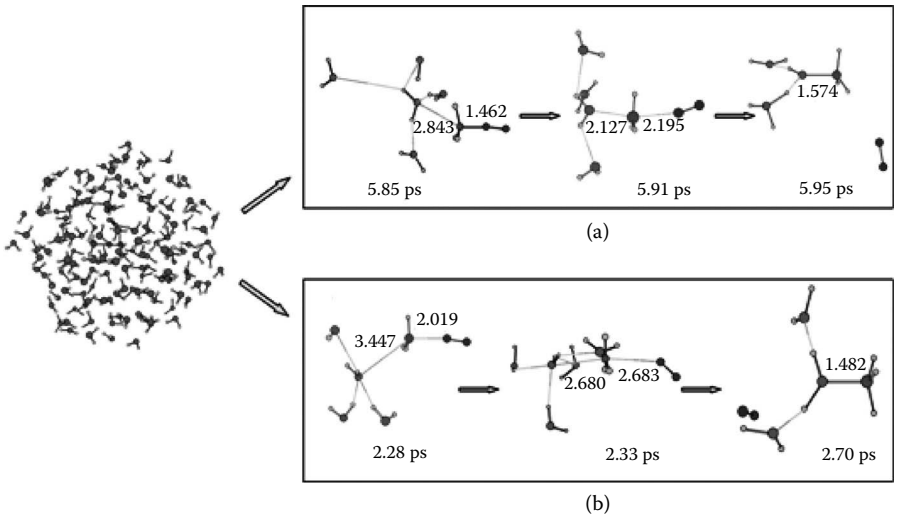


FIGURE 6.4 Initial droplet structure and structures of substrate and nearby water molecules along type A and B trajectories of the hydrolysis reaction of the $CH_3N_2^+$ substrate are described in Section 6.2.2. Numbers are atomic distances in Å. (Reproduced from Sato, M., Yamataka, H., Komeiji, Y. et al. 2008. *J. Am. Chem. Soc.* 130: 2396–2397. With permission.)

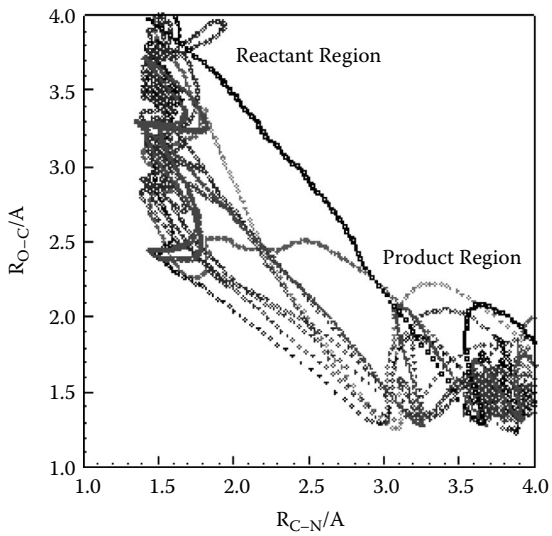


FIGURE 6.5 Ten trajectories on the two-dimensional surface. (Reproduced from Sato, M., Yamataka, H., Komeiji, Y. et al. 2008. *J. Am. Chem. Soc.* 130: 2396–2397. With permission.) **(See color insert following page 117.)**

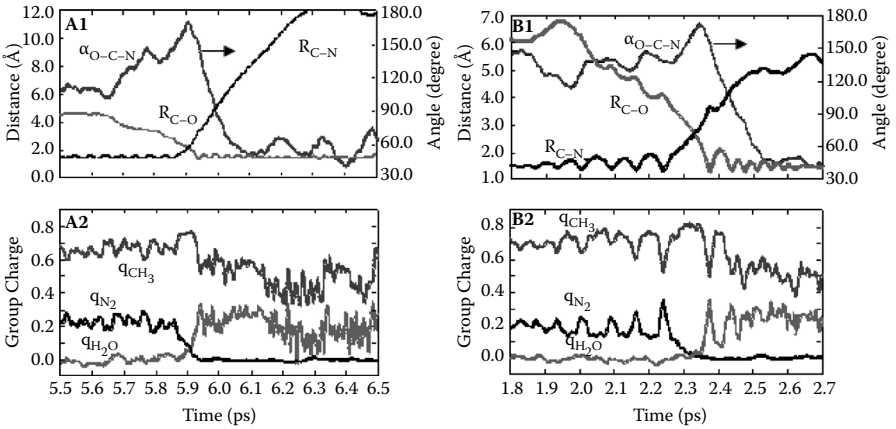


FIGURE 6.6 Variations of O-C and C-N atomic distances and the O-C-N angle (upper) and group charges (lower) along trajectory A (left) and B (right). (Reproduced from Sato, M., Yamataka, H., Komeiji, Y. et al. 2008. *J. Am. Chem. Soc.* 130: 2396–2397. With permission.)

a full quantum treatment of the explicit solvent, and thus showed the utility of the FMO-MD method.

6.2.3 Blue Moon Calculation of the Menschutkin Reaction

The blue moon method, a way to sample rare events by constraint MD, was combined with FMO-MD and applied to calculate the potential of mean force (PMF) (free energy profile) of the Menschutkin reaction in the presence and absence of the solvent.[5] The PMF curves showed a shift of transition state (TS) toward the reactant, consistent with the Hammond postulate.

The Menschutkin reaction is another S_N2-type reaction that proceeds in the following order[20]:

$$H_3N + CH_3Cl \rightarrow [H_3N\ldots CH_3\ldots Cl]‡ \rightarrow {}^{+}H_3NCH_3 + Cl^{-}$$

This reaction is endothermic in the gas phase but exothermic in aqueous solvent and hence serves as a benchmark test for any solvent model.

PMF curves of the reaction were drawn both in the presence and absence of the explicit solvent. All the calculations were performed at the HF/6-31G level as follows. The reaction coordinate ξ was the difference between two distances, namely,

$$\xi = R_{C\text{-}Cl} - R_{N\text{-}C}$$

The TS of the solute (H_3NCH_3Cl) was obtained by a conventional MO method in vacuum (Figure 6.7a) and used as the initial structure for the calculation in the gas phase. For the solvated calculation, a spherical water droplet was generated around

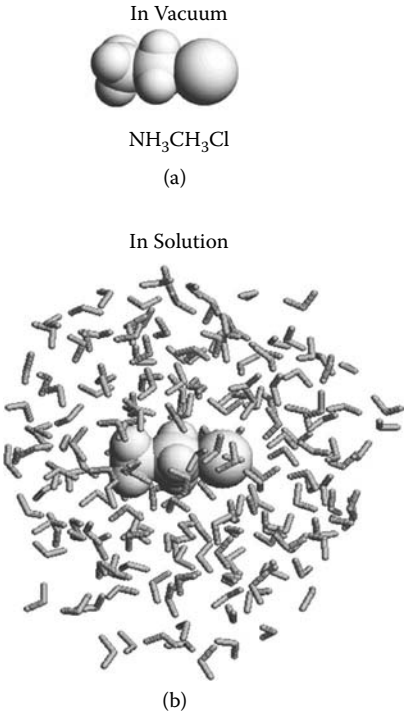


FIGURE 6.7 Initial configurations of NH_3CH_3Cl (a) in vacuum (the TS configuration obtained at the HF/6-31G level) and (b) in solution used in the Menschutkin reaction described in Section 6.2.3. The fragment molecular orbital–based molecular dynamics (FMO-MD) simulation was performed at the HF/6-31G level. (Reproduced from Komeiji, Y., Ishikawa, T., Mochizuki, Y. et al. 2009. *J. Comput. Chem.* 30: 40–50. With permission.)

the solute and was optimized by the classical MD/EM method prior to FMO-MD (Figure 6.7b). The initial structures, with and without solvent, were subjected to FMO-EM and heating FMO-MD (up to 300 K in 1 ps), according to the standard protocol proposed in Section 6.1.3, and then, the FMO-MD/blue moon calculations were performed for various values of ξ. The time steps were 1 fs with RATTLE and 0.5 fs without it. The equilibration and sampling durations were both 100 fs for each value of ξ.

In the calculated PMF curves of the Menschutkin reaction in the presence (in solution) and absence (in vacuum) of the explicit solvent molecules, we observed a clear TS shift toward the reactant (Figure 6.8), consistent with the Hammond postulate.[21,22] This observation agreed well with many preceding computational studies on this reaction. In addition to the Hammond shift, the two-dimensional map of $R_{N\text{-}C}$ versus $R_{C\text{-}Cl}$ (Figure 6.9) revealed a small anti-Hammond shift. As visualized within Figure 6.9, the Hammond shift refers to a shift parallel to the reaction coordinate ξ, whereas the anti-Hammond shift refers to a shift perpendicular to ξ. The anti-Hammond shift by the solvent was clear at the TS in solution

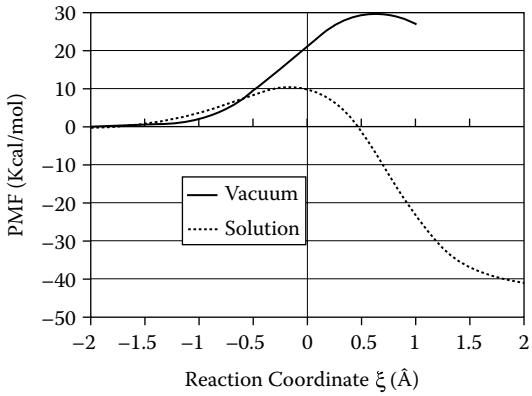


FIGURE 6.8 PMF (free energy change) curves of the Menschutkin reaction in vacuum and in solution as functions of ξ (= $R_{C\text{-}Cl} - R_{N\text{-}C}$). (Reproduced from Komeiji, Y., Ishikawa, T., Mochizuki, Y. et al. 2009. *J. Comput. Chem.* 30: 40–50. With permission.)

(ξ = –0.2 Å), where both $R_{N\text{-}C}$ and $R_{C\text{-}Cl}$ were longer in solution than in vacuum. This anti-Hammond shift should reflect a solvent-induced loosening of the whole structure of NH_3CH_3Cl at TS. Thus, the two-dimensional map indicated that solvation introduced a minor anti-Hammond shift in addition to the major Hammond shift of the TS.

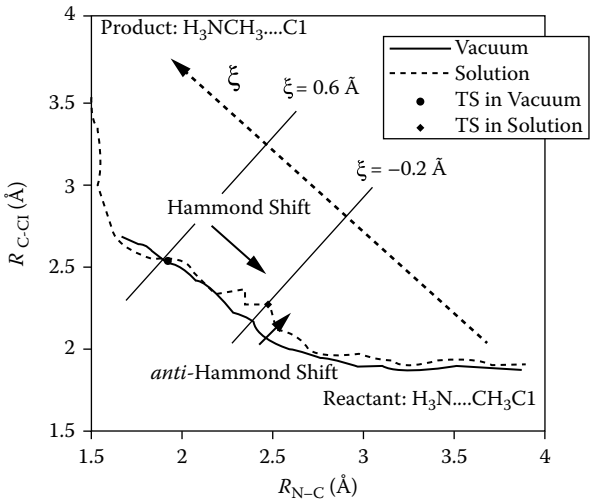


FIGURE 6.9 $R_{C\text{-}Cl}$ versus $R_{N\text{-}C}$ two-dimensional map. $R_{C\text{-}Cl}$ and $R_{N\text{-}C}$ were averaged over the 100 fs sampling period of each ξ and were plotted. The shift caused by the solvent can be divided into the Hammond shift, parallel to ξ, and the anti-Hammond shift, perpendicular to ξ. (Reproduced from Komeiji, Y., Ishikawa, T., Mochizuki, Y. et al. 2009. *J. Comput. Chem.* 30: 40–50. With permission.)

Our calculation on the Menschutkin reaction was consistent with the Hammond postulate. Though not discussed here, the results were in accordance with previous studies performed by different computational methods under different solvent models (see Komeiji et al.[5] for detailed comparison and discussion).

6.3 GENERAL DISCUSSION AND PERSPECTIVE

In spite of the successful examples of FMO-MD simulations so far presented, several problems need to be overcome to enhance the general applicability of the FMO-MD method. First, implementation of FMO3 is required to decrease the energy gap resulting from the dynamic fragmentation scheme.[23,24] Second, the electron correlation effect should be incorporated through MP2.[25–27] Third, the periodic boundary condition for the solvent box should be implemented to avoid the artificial surface effect.[28]

Another important issue concerning FMO-MD is the enhancement of conformation sampling. Like other *ab initio* MD methods, quite limited duration is available for FMO-MD, currently at most several tens of picoseconds on a typical laboratory PC cluster. Ishimoto et al.[29,30] succeeded in enhancing conformation sampling by employing a Hamiltonian algorithm (HA). In their FMO-HA studies, Gly polymers showed far more conformations than an ordinary FMO-MD could sample. Apart from the HA, quite a few algorithms were developed to enhance the conformation sampling in MD (for example, see Mitsutake et al.[31]). Hence, it will likely be worthwhile to combine FMO-MD with one of these algorithms to improve the sampling efficiency.

In conclusion, an improved FMO-MD methodology is needed, but even in its present form, the potential utility of FMO-MD is great as a general quantum simulation method for large molecular systems. Recently, a series of benchmark FMO-MP2 calculations were performed on a massively parallel computer, where proteins with a few hundred residues could be processed within a dozen minutes by using such high-performance computing resources.[32] This implies that the FMO-MD simulations for proteins and other large molecular systems will be tractable on peta-class supercomputers in the near future.

REFERENCES

1. Komeiji, Y., Nakano, T., Fukuzawa, K. et al. 2003. Fragment molecular orbital method: application to molecular dynamics simulation, "*ab initio* FMO-MD." *Chem. Phys. Lett.* 372: 342–347.
2. Born, M., and Oppenheimer, J.R. 1927. Zur Quantentheorie der Molekeln. *Ann. Phys.* 84: 457–484.
3. Kitaura, K., Sugiki, S., Nakano, T., Komeiji, Y., Uebayasi, M. 2001. Fragment molecular orbital method: analytical energy gradients. *Chem. Phys. Lett.* 336: 163–170.
4. Komeiji, Y., Inadomi, Y., Nakano, T. 2004. PEACH 4 with ABINIT-MP: a general platform for classical and quantum simulations of biological molecules. *Comput. Biol. Chem.* 28: 155–161.

5. Komeiji, Y., Ishikawa, T., Mochizuki, Y. et al. 2009. Fragment molecular orbital method–based molecular dynamics (FMO-MD) as a simulator for chemical reactions in explicit solvation. *J. Comput. Chem.* 30: 40–50.
6. Komeiji, Y., Mochizuki, Y., Nakano, T. Fedorov, D.G. 2009. Fragment molecular orbital–based molecular dynamics (FMO-MD), a quantum simulation tool for large molecular systems. *J. Mol. Struct.* (*THEOCHEM*), 898: 2–7.
7. Nakano, T., Kaminuma, T., Sato, T. et al. 2000. Fragment molecular orbital method: application to polypeptides. *Chem. Phys. Lett.* 318: 614–618.
8. Komeiji, Y., Uebayasi, M., Takata, R. et al. 1997. Fast and accurate molecular dynamics simulation of a protein using a special purpose computer. *J. Comput. Chem.* 18: 1546–1563.
9. Ryckaert, J., Ciccotti, G., Berendsen, H.J.C. 1977. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *n*-Alkanes. *J. Comput. Phys.* 23, 327–341.
10. Andersen, H.C. 1983. Rattle: a "velocity" version of the Shake algorithm for molecular dynamics calculations. *J. Comput. Phys.* 52: 24–34.
11. Nosé, S. 1991. Constant temperature molecular dynamics. *Prog. Theor. Phys.Suppl.* 103: 1–46.
12. Martyna, G.J., Klein, M.L., Tuckerman, M. 1992. Nosé-Hoover chains — the canonical ensemble via continuous dynamics. *J. Chem. Phys.* 97: 2635–2643.
13. Zhang, F. 1997. Operator-splitting integrators for constant-temperature molecular dynamics. *J. Chem. Phys.* 106: 6102–6106.
14. Sprik, M., Ciccotti, G. 1998. Free energy from constrained molecular dynamics. *J. Chem. Phys.* 109: 7737–7744.
15. Komeiji, Y. 2007. Implementation of the blue moon ensemble method. *CBI J.* 7, 12–23.
16. Mochizuki, Y., Komeiji, Y., Ishikawa, T., Nakano, T., Yamataka, H. 2007. A fully quantum mechanical simulation study on the lowest *n*-pi* state of hydrated formaldehyde. *Chem. Phys. Lett.* 437: 66–72.
17. Mochizuki, Y., Tanaka, K., Yamashita, K. et al. 2007. Parallelized integral-direct CIS(D) calculations with multilayer fragment molecular orbital scheme. *Theor. Chem. Acc.* 117: 541–553.
18. Sato, M., Yamataka, H., Komeiji, Y. et al. 2008. How does an S_N2 reaction take place in solution? Full *ab initio* MD simulations for the hydrolysis of the methyl diazonium ion. *J. Am. Chem. Soc.* 130: 2396–2397.
19. Mochizuki, Y., Fukuzawa, K., Kato, A. et al. 2005. A configuration analysis for fragment interaction. *Chem. Phys. Lett.* 410: 247–253.
20. Menschutkin, N. 1890. Beitrage zur Kenntnis der Affinitatskoeffizienten der Alkylhaloid und der organischen Amine. *Z. Physik. Chem.* 5: 589.
21. Leffler, J.E. 1952. Parameters for the description of transition states. *Science* 117: 340–341.
22. Hammond, G.S. 1955. A correlation of reaction rates. *J. Am. Chem. Soc.* 77: 334–338.
23. Fedorov, D.G., Kitaura, K. 2004a. The importance of three-body terms in the fragment molecular orbital method. *J. Chem. Phys.* 120: 6832–6840.
24. Fedorov, D.G., and Kitaura, K. 2004b. On the accuracy of the 3-body fragment molecular orbital method (FMO) applied to density functional theory. *Chem. Phys. Lett.* 389: 129–134.
25. Fedorov, D.G., Kitaura, K. 2004c. Second order Møller–Plesset perturbation theory based upon the fragment molecular orbital method. *J. Chem. Phys.* 121: 2483–2490.

26. Mochizuki, Y., Nakano, T., Koikegami, S. et al. 2004. A parallelized integral-direct second-order Møller–Plesset perturbation theory method with a fragment molecular orbital scheme. *Theor. Chem. Acc.* 112: 442–452.
27. Mochizuki, Y., Koikegami, S., Nakano, T., Amari, S., Kitaura, K. 2004. Large scale MP2 calculations with fragment molecular orbital scheme. *Chem. Phys. Lett.* 396: 473–479.
28. Komeiji, Y., Ishida, T., Fedorov, D.G., Kitaura, K. 2007. Change in a protein's electronic structure induced by an explicit solvent: an *ab initio* fragment molecular orbital study of ubiquitin. *J. Comput. Chem.* 28: 1750–1762.
29. Ishimoto, T., Tokiwa, H., Teramae, H., Nagashima, U. 2004. Development of an *ab initio* MO-MD program based on fragment MO method — an attempt to analyze the fluctuation of protein. *Chem. Phys. Lett.* 387: 460–465.
30. Ishimoto, T., Tokiwa, H., Teramae, H., Nagashima, U. 2005. Theoretical study of intramolecular interaction energies during dynamics simulations of oligopeptides by the fragment molecular orbital-Hamiltonian algorithm method. *J. Chem. Phys.* 122: 094905.
31. Mitsutake, A., Sugita, Y., Okamoto, Y. 2001. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* 60: 96–123.
32. Mochizuki, Y., Yamashita, K., Murase, T. et al. 2008. Large scale FMO-MP2 calculations on a massively parallel-vector computer. *Chem. Phys. Lett.* 457: 396–403.

7 Application of the FMO Method to Specific Molecular Recognition of Biomacromolecules

Kaori Fukuzawa, Yuji Mochizuki, Tatsuya Nakano, and Shigenori Tanaka

CONTENTS

7.1 INTRODUCTION

In structural biology, molecular simulations have played an important role in elucidating functions of the biological system. The understanding of biological phenomena at the molecular level is expected to lead the modeling of disease, drug discovery, and various applications. A variety of life phenomena occur through the combination of site-specific molecular recognition of biomacromolecules. Computer simulations thus provide a promising approach to elucidate these molecular interactions in detail. However, most calculations carried out to date have employed classical mechanical methods based on empirical force fields. Such methods remain limited for performing an accurate analysis of intermolecular interactions such as charge redistribution and charge-transfer (CT) interactions. In contrast to the limitations of classical approaches to molecular simulation, quantum mechanical simulations have been used to successfully characterize weak intermolecular interactions and CT processes. Because several different types of interactions are involved in the interactions of biomolecules, quantum mechanical treatment is necessary to obtain an accurate and systematic understanding of these interactions. The fragment molecular orbital (FMO) method[1–4] is one of the most reasonable tools with which to analyze the electronic structure of biomacromolecules.

The FMO method and FMO-based analysis can be widely used to understand the interactions between biomolecules such as protein–ligand, protein–protein, and DNA–protein interactions. In the FMO calculation, not only the consideration of total energy and electron density for a whole system but also the interactions analyses at the residue level and even at the orbital level are available. Details of FMO methodology and FMO-based analyses were presented in Chapter 3. In this chapter, several applications of the FMO method are illustrated concerning molecular recognition between biomacromolecules: receptor–ligand interaction and receptor–cofactor interaction of nuclear receptor, sequence-specific DNA–protein interaction of cyclic-AMP receptor protein, and antigen–antibody interaction and sialic acid recognition of influenza hemagglutinin. All the FMO calculations were performed with the ABINIT-MP program, and the visualization was carried out with the BioStation Viewer.[5]

7.2 MOLECULAR INTERACTION BETWEEN A NUCLEAR RECEPTOR AND ITS LIGAND

The nuclear receptor (NR) is a ligand-activated transcriptional factor and plays important roles in the regulation of growth, differentiation, and homeostasis in a variety of tissues. The NRs with ligands and their coactivators form complexes that bind to specific sequences called response elements in the promoter regions of the target genes, and then transcription of various target genes is mediated. The NR superfamily possesses common domain functions and structures — that is, variable N-terminal transactivation domains, conserved DNA-binding domains (DBD), variable hinge regions, conserved ligand-binding domains (LBD), and variable C-terminal regions. The characteristic ligand-induced motion of NR is the conformational change of the helix12 (H12) at the C-terminal of LBD. When agonists bind to NR LBD, the active

conformation of H12 is formed, which possesses the coactivator binding surface. Antithetically, when antagonists bind to LBD, the coactivator binding surface is not produced because H12 is prevented from reaching its correct position.

The estrogen receptor (ER) is a member of the NR superfamily. The functions of ER are induced by the binding of endogenous ligands estrogens to the ER LBD.[5,7] As well as the naturally occurring estrogens, it is known that a variety of compounds such as medical compounds and endocrine disrupters might bind to the ER and lead to estrogenic activities.[8,9] In particular, the ER is linked to diseases such as breast cancer, endometrial cancer, and osteoporosis, and the ER has therefore been one of the most important targets for the development of therapeutic agents and screening of environmental pollutants. One well-established methodology for the screening of seed compounds is a receptor-binding assay based on their binding affinities to the ER, and *in silico* screening has also recently been performed to predict their binding affinities. In order to perform efficient screening, it is necessary to understand the binding mechanism of the ER with ligands.

Figure 7.1 shows the crystal structures for the complex of human ER-LBD and xenoestrogen, 17β-estradiol (EST).[10,11] Some important charged/polarized residues, Glu353, Arg394, and His524, exist at the ligand-binding site of the ER and they construct hydrogen-bond networks (Figure 7.2). These hydrogen bonds are thought to play a key role in ER–ligand binding, but the details of their contributions to this binding remain to be elucidated. Furthermore, the roles of other surrounding residues and distant residues in the ER-LBD also remain unclear.

FIGURE 7.1 The ribbon display of the estrogen receptor ligand-binding domain complexed with 17β-estradiol. Fifty residues directly surrounding the ligand in the smaller model are shown as dark gray. The ligand and the water molecule are shown in the stick display. (See also text for details.)

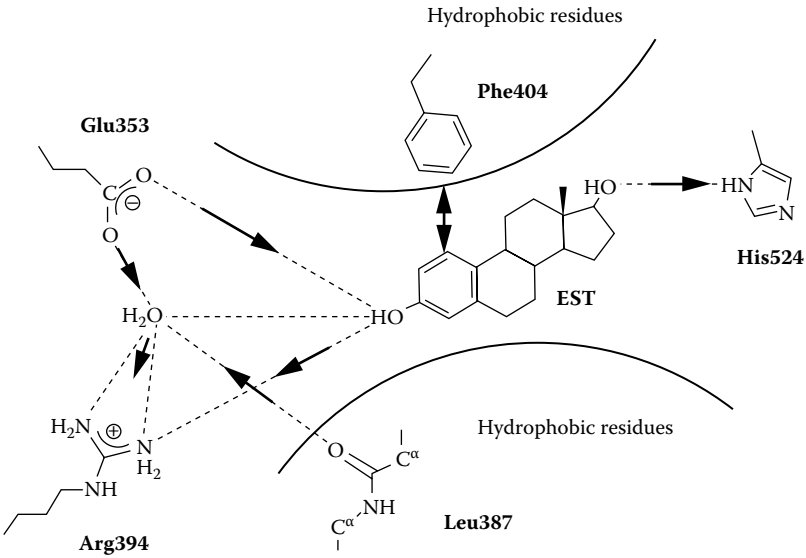


FIGURE 7.2 Hydrogen bond network at the ligand-binding site of the estrogen receptor complexed with 17β-estradiol (EST). Dotted lines, arrowed lines, and two-headed arrow line indicate hydrogen bonds, charge transfer interactions, and dispersion interaction, respectively. (See text for details of the interactions.)

A number of computational simulations have been carried out to clarify the mechanism of ER–ligand binding.[12,13] Although most calculations carried out to date have employed classical mechanical methods based on empirical force fields, such methods remain limited with respect to obtaining accurate analysis of intermolecular interactions such as charge redistribution and CT interactions. In contrast, quantum-mechanical simulations have been used to successfully characterize weak intermolecular interactions and CT processes. Because several different types of interactions are involved in ER–ligand interactions, quantum mechanical treatment is necessary to obtain an accurate and systematic understanding of these interactions. In this report, we apply the FMO method[1–4] to predict ligand-binding affinity of ER[14] and to understand detailed interaction between ER and ligand.[15] We also addressed the postbinding problem in addition to that of ligand binding; interaction between retinoid X receptor (RXR) and coactivator were studied, including the role of helix 12 in the transcriptional regulation of NR.[16–18]

7.2.1 Molecular Modeling

The 11 ligand molecules displayed in Figure 7.3 were examined for the ligand-binding energy calculations,[14] and the EST was especially used for detailed analysis, such as residue–ligand interactions and orbital interactions.[15] The initial atomic coordinates of the ER–ligand complexes were obtained from the Protein Data Bank (PDB)[10]; the entries 1ERE, 3ERD, 1ERR, and 3ERT for EST, diethylstilbestrol (DES), raloxifene (RAL), and 4-hydroxytamoxifen (OHT), respectively.[11,16] Based on these structures,

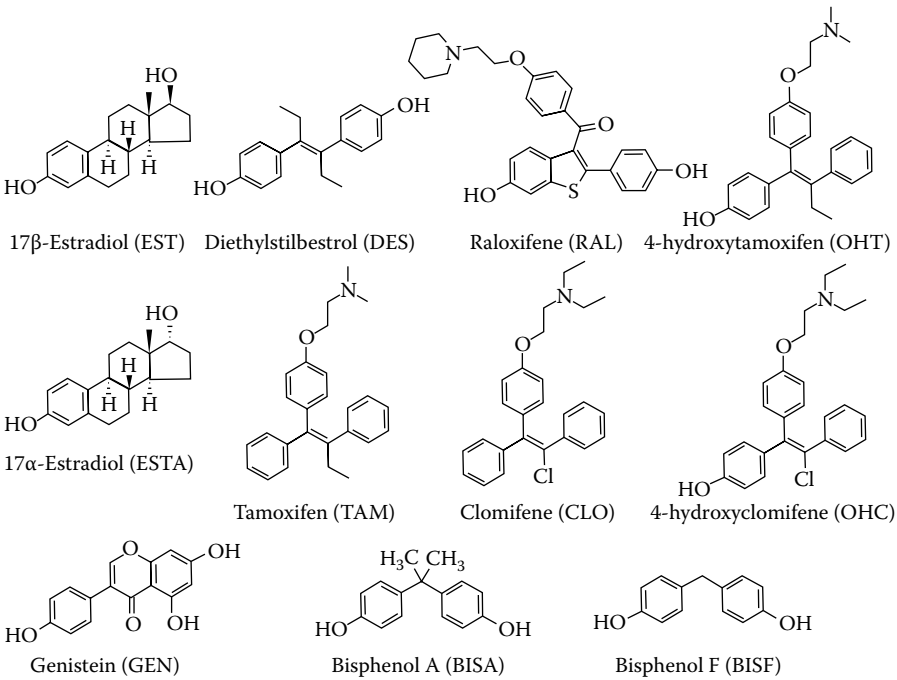


FIGURE 7.3 Estrogen-like compounds studied in binding energy calculations.

the entire LBD of the receptor protein (residues 307 through 547), ligand, and a water molecule were used for simulations. For the other ligands, binding geometries with receptor were modeled by superimposing ligands with PDB-available structures. Missing hydrogen atoms, side chains, and main chains in the PDB files were complemented manually using the molecular graphic software InsightII.[20] All the positions of hydrogen atoms and added chains were optimized by CHARMm force field calculations[21] with the other heavy atoms fixed at the positions given in the original PDB data. The geometries of the hydrogen bond network between ER and a ligand were optimized at the RHF/6-31G* level with Gaussian program package using small model complexes including ligand, single water molecule, and four residues (Glu353, Leu387, Arg394, and His524, see Figure 7.2). The binding energies were calculated using two model complexes with the optimized hydrogen-bond geometries. In most of the binding energy calculations, ER was modeled as 50 residues surrounding the ligand (smaller model, see Figure 7.1). Only these four ligands — EST, DES, RAL, and OHT — were entire LBDs and used for receptors (larger model).

7.2.2 Prediction of Binding Affinity between an Estrogen Receptor and Its Ligands

The binding energies were calculated at the FMO-RHF/STO-3G levels, and the CHARMm force field calculation was also used for comparison.[14] The energy of

each of the three systems — that is, the receptor ($E_{receptor}$), the ligand (E_{ligand}), and the ER–ligand complex ($E_{complex}$) — were calculated, where the hydrogen-bonded water molecule was included in the receptor. The binding energy for a given ligand, ΔE_{ligand}, can be obtained by the following supermolecule calculation:

$$\Delta E_{ligand} = E_{complex} - (E_{receptor} + E_{ligand}) \tag{7.1}$$

and the binding energies of a ligand relative to that of EST, $\Delta\Delta E_{ligand}$, are defined as

$$\Delta\Delta E_{ligand} = \Delta E_{EST} - \Delta E_{ligand} \tag{7.2}$$

Then, the correlation between $\Delta\Delta E_{ligand}$ and the experimental relative binding affinity (RBA) was examined. Note that the geometries of free ER and ligand were fixed at the complex geometries.

As shown in Figure 7.4, the correlation coefficient was drastically improved by the FMO calculation (R = 0.837) as compared with the CHARMm calculation (R = 0.035). Therefore, this FMO approach may provide a powerful tool for assessing the affinity of putative xenoestrogens *in silico* prior to biological experiments. An interesting finding was obtained from differential charge distributions.[14] The amount of charge transfer from ER to the ligands is highly related with the binding energy; ΔE becomes larger or the complex becomes more stable with the increase in the negative net charge of the ligand. This fact suggests that the CT between ER and the ligand plays an important role to characterize their binding. This finding cannot be obtained by fixed-charge approach such as empirical force field, and FMO calculations provide an advantage to discovering such a promising descriptor for *in silico* screening.

To elucidate the influence of model size of receptor protein, the ΔE values were compared to the differences between two models, larger and smaller models, for four

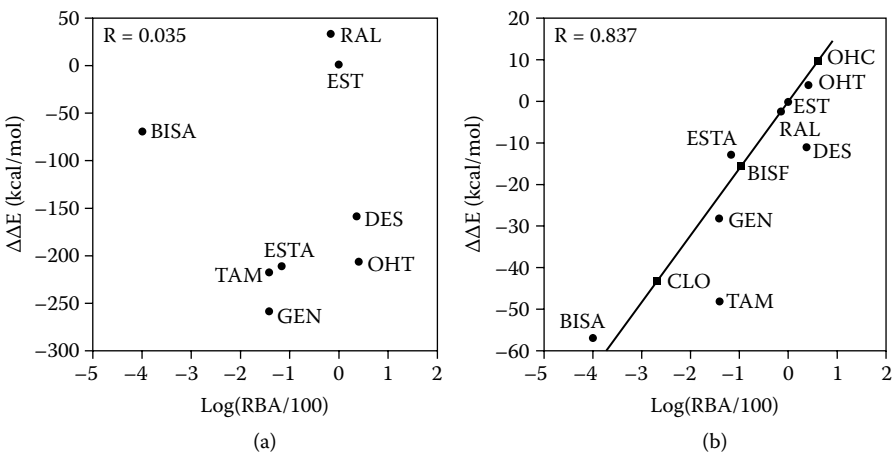


FIGURE 7.4 Relationship between theoretical calculation and experiment: (a) CHARMm and (b) FMO-RHF/STO-3G.

TABLE 7.1
Binding Energy (ΔE) and Interfragment Interaction Energy (IFIE) of Several Ligands with the Estrogen Receptor Calculated by the RHF/STO-3G Method[e]

	ΔE[d]		IFIE	
Ligand	Complete Receptor[a]	Model Receptor[b]	1 Residue Unit[c]	2 Residue Unit[d]
17β-estradiol (EST)	–37.65	–37.80	–42.46	–42.78
Diethylstilbestrol (DES)	–28.33	–26.70	–33.16	–32.31
Raloxifene (RAL)	–26.13	–35.30	–33.97	–34.12
4-hydroxytamoxifen (OHT)	–38.19	–41.73	–45.69	–45.13

[a] Entire ligand-binding domain and
[b] 50 residues surrounding the ligand were used as the estrogen receptor. Fragmentations were in
[c] 1 residue and
[d] 2 residue units.
[e] Energies are in kcal/mol.

ligands EST, DES, RAL, and OHT (Table 7.1).[15] No significant difference was found in the ΔE values between the receptor with the entire LBD and that with 50 residues surrounding only the ligand. Therefore, the ER–ligand interaction is localized in the binding region and is properly described by considering the amino acid residues in the first layered α-helices of the ligand. Such criteria are though to be efficient for *in silico* drug discovery to reduce computational time. Also, we verified the reliability of IFIE analysis for ER–ligand interactions. The sum of all the IFIEs between the ligand and each fragment in the ER could be assumed to be the interaction energy between the ligand and the ER (see Chapter 3 for details of the IFIE). Such IFIEs for four ligands were compared with ER–ligand binding energies (ΔE). The differences between the ΔE and IFIEs are due to the electron density deformation of the individual molecules in their complex. The corresponding deformation energies were of several kcal/mol, which were common among all the calculations. Therefore, we concluded that the relative IFIE values for each ligand were comparable to those of ΔE. In addition, we examined the errors of IFIEs obtained between two fragmentations of ER, one and two residues per fragment. The discrepancies of the sum of the IFIEs between the two fragmentations were less than 1 kcal/mol. Therefore, at least a qualitative understanding could be obtained by IFIE analysis with the smaller fragmentation, and residue–ligand and residue–residue interactions could then be estimated. These results suggest that the ligand-binding energy can be properly estimated not only by supermolecule calculation but also by IFIEs. The supermolecule calculation requires three single-point energies for receptor, ligand, and their complex (Equation 7.1), but only one calculation is needed for IFIE analysis. This hints at the potential to even further reduce computational times for *in silico* screening.

7.2.3 Molecular Interaction between a Ligand and Each Residue of an Estrogen Receptor

For a detailed analysis of the molecular interactions between ER and its ligand, the FMO calculations with several levels of theory and several FMO-based analyses were performed for binding complex between ER and natural ligand, EST.[15] The *ab initio* FMO calculations were carried out at the RHF level with STO-3G, 6-31G, 6-31G*, and 6-31G** basis sets. The energies and electron densities were further improved at the MP2 level[22,23] with 6-31G and 6-31G* basis sets. To analyze residue–ligand interactions, the fragmentation of the system was as follows: each amino acid residue of ER, the EST molecule, and the water molecule were treated as a single fragment.

Figure 7.5 illustrates the strength of the interactions of EST with each residue fragment of the ER according to color at the MP2/6-31G* level. Numerical representations of the stronger residue–ligand interactions are also listed in Table 7.2. All colored residues with strong interaction were geometrically located around the EST. These findings support previous discussions that the ΔE of the ER-ligand are

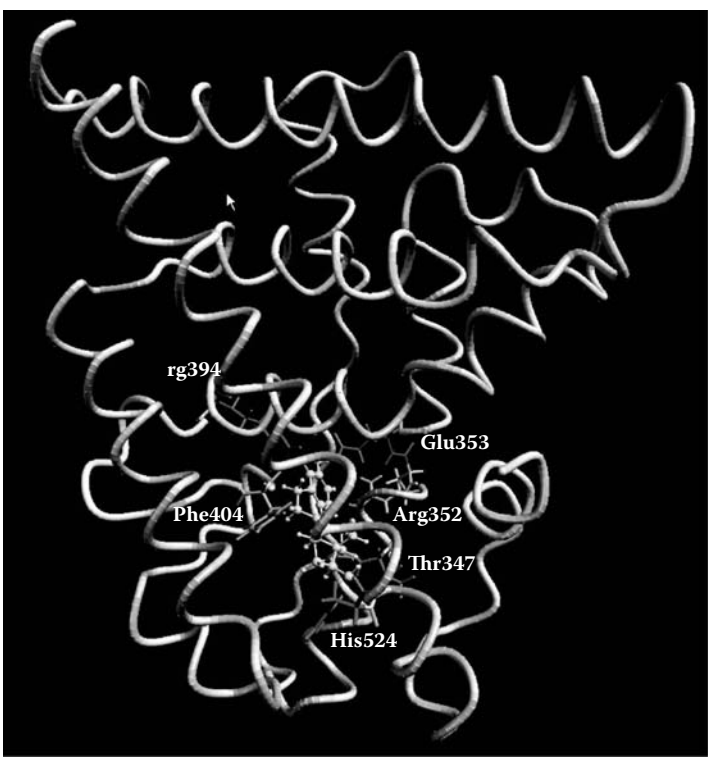


FIGURE 7.5 Visualization of interfragment interaction energies between each residue of an estrogen receptor and the ligand calculated at the MP2/6-31G* level. **(See color insert following page 117.)**

TABLE 7.2
Interfragment Interaction Energy between 17β-Estradiol and Each Amino Acid Residue of the Estrogen Receptor[a]

Method			RHF							MP2			
Basis Set	STO-3G		6-31G		6-31G*		6-31G**			6-31G		6-31G*	
Attractive Interactions	Glu353	–39.28	Glu353	–27.92	Glu353	–27.75	Glu353	–28.19		Glu353	–29.59	Glu353	–32.23
	His524	–5.98	Arg394	–8.75	Arg394	–7.10	Arg394	–7.00		Arg394	–10.48	Arg394	–9.11
	Arg394	–5.27	Thr347	–5.41	Thr347	–5.04	Thr347	–4.87		His524	–8.74	His524	–8.80
	Thr347	–1.92	His524	–5.26	His524	–3.89	His524	–3.61		Thr347	–8.11	Thr347	–7.98
	Water	–1.02	**Met522**	**–2.44**	**Met522**	**–2.21**	**Met522**	**–2.19**		**Phe404**	**–4.46**	**Phe404**	**–5.12**
	Glu330	–0.92	**Met343**	**–1.85**	**Met343**	**–1.58**	**Met343**	**–1.57**		**Met343**	**–3.66**	**Leu387**	**–3.71**
	Met522	**–0.82**	Water	–1.59	Glu330	–1.20	Glu330	–1.17		**Leu346**	**–3.59**	**Leu346**	**–3.60**
	Lys531	–0.74	Glu330	–1.27	Lys531	–0.92	Lys531	–0.87		**Met522**	**–3.29**	**Met343**	**–3.55**
	Asp351	–0.71	Lys531	–1.02	**Met528**	**–0.91**	**Met528**	**–0.83**		**Ala350**	**–3.07**	**Met522**	**–3.32**
	Val392	**–0.66**	**Met528**	**–1.02**	**Gly344**	**–0.84**	**Gly344**	**–0.79**		**Leu387**	**–3.04**	**Ala350**	**–3.28**
	Ala350	**–0.64**	**Gly344**	**–1.00**	**Val392**	**–0.78**	**Val392**	**–0.74**		Water	–2.90	**Met388**	**–2.52**
	Glu542	–0.51	**Val392**	**–0.89**	Asp538	–0.74	Asp538	–0.73		**Leu384**	**–2.00**	Water	–2.48
	Ala405	**–0.49**	Asp538	–0.77	Glu542	–0.69	Glu542	–0.68		**Met421**	**–1.94**	**Leu384**	**–2.12**
	Asp538	–0.47	**Ala350**	**–0.74**	**Ala350**	**–0.66**	**Ala350**	**–0.62**		**Met388**	**–1.75**	**Met421**	**–2.04**
	Glu380	–0.45	Glu542	–0.72	Arg412	–0.62	Arg412	–0.61		Glu330	–1.27	**Leu391**	**–1.63**
	Arg412	–0.44	**Leu346**	**–0.72**	**Leu346**	**–0.61**	Arg436	–0.60		**Leu391**	**–1.18**	**Ile424**	**–1.29**
	Lys529	–0.38	Glu380	–0.67	Glu380	–0.61	Glu380	–0.60		**Ile424**	**–1.17**	Glu330	–1.20
	Gly344	**–0.38**	Arg436	–0.62	Arg436	–0.60	Asp351	–0.57		**Met528**	**–1.16**	**Met528**	**–1.13**

(*Continued*)

TABLE 7.2 (CONTINUED)
Interfragment Interaction Energy between 17β-Estradiol and Each Amino Acid Residue of the Estrogen Receptor[a]

Method	RHF								MP2			
Basis Set	**STO-3G**		**6-31G**		**6-31G***		**6-31G****		**6-31G**		**6-31G***	
Repulsive Interactions	Arg352	1.83	**Leu391**	**1.79**	**Leu391**	**1.68**	**Leu391**	**1.70**	Leu354	0.96	Glu423	0.90
	Leu391	**2.35**	Arg352	2.69	**Met388**	**2.38**	**Met388**	**2.35**	Glu419	1.00	Glu419	0.95
	Met388	**2.51**	**Met388**	**2.76**	Arg352	2.52	Arg352	2.43	Glu339	1.25	Glu339	1.11
	Leu525	**5.15**	**Leu525**	**4.64**	**Leu525**	**4.41**	**Leu525**	**4.53**	Arg352	2.69	Arg352	2.52
[b]Charged/polarized	–54.48		–50.42		–45.42		–45.11		–62.66		–63.31	
[c]**Hydrophobic**	**12.02**		**4.95**		**5.15**		**5.77**		**–33.46**		**–37.77**	
[d]Total	–42.46		–45.48		–40.26		–39.34		–96.13		–101.07	

[a] Only selected values with stronger interactions are shown, and they are listed in order of the interaction energy values. Hydrophobic residues are indicated as bold characters. Energies are in kcal/mol.

[b] Sum of all IFIEs between EST and each charged or polarized residue in the ER.

[c] Sum of all IFIEs between EST and each hydrophobic residue in the ER.

[d] Sum of all IFIEs between EST and each residue in the ER.

properly described by considering model protein surrounding the ligand. The total IFIEs between the ligand and all residues of ER were –42.5, –45.5, –40.3, and –39.3 kcal/mol at the RHF/STO-3G, 6-31G, 6-31G*, and 6-31G** levels, respectively, and much larger in magnitude, –96.1 and –101.1 kcal/mol, at the MP2/6-31G and 6-31G* levels. The large energy difference between the HF and MP2 methods was due to the interaction with hydrophobic residues. The IFIEs of charged and polarized residues were qualitatively similar between both the HF and MP2 methods, but those of hydrophobic residues were different according to the methods. In the HF methods, the sums of IFIEs between the ligand and each hydrophobic residue were similar, about 5 kcal/mol, except with the STO-3G of larger value; all sums showed positive values (repulsive interactions). However, at the MP2 level, the sum of IFIEs were largely negative values (attractive interaction), under –30 kcal/mol. Many hydrophobic residues were found to be stabilized through electron correlations.

Figure 7.6 visualizes interactions between hydrophobic residue and the ligand at both the HF and MP2 levels. In the RHF/6-31G* method, some hydrophobic residues were visualized as either stabilized or destabilized. In contrast, in the MP2/6-31G* method, almost all hydrophobic residues surrounding the ligand appeared to be stabilized, and none were destabilized. These stabilizations were dominated by the van der Waals dispersion interactions that could be considered by the electron correlation methods. Weak, but many interactions with these hydrophobic residues significantly stabilized the ER–ligand binding in addition to the electrostatic interactions with charged and polarized residues.

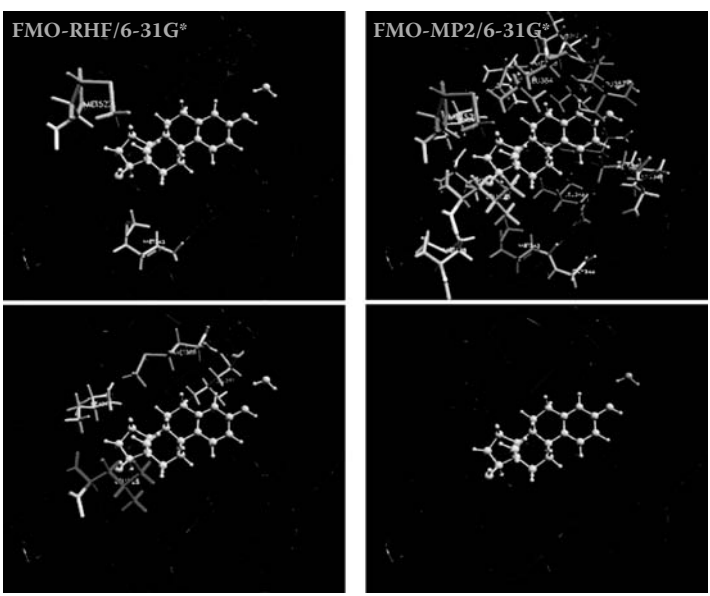


FIGURE 7.6 Interfragment interaction energies between the ligand and surrounding hydrophobic residues. Upper figure: attractive interaction less than –1 kcal/mol. Lower figure: repulsive interactions more than 1kcal/mol. Calculations were performed at the RHF/6-31G* (left) and MP2/6-31G* (right) levels. **(See color insert.)**

Receptor–ligand interactions were also analyzed in residue level based on the IFIE. Note that the fragment unit starting at Cα atom position is not exactly the same as the residue unit starting at the N atom position. We therefore add the "#" for the fragmented-residues name (e.g., Glu#353, compared to the usual residue name, Glu353). Table 7.2 shows the ligand interactions with each residue; Glu#353, His#524, Arg#394, Phe#404, and Thr#347 indicated stabilized interactions, and Arg#352 indicated destabilized interactions. The remarkably stabilized residue was Glu#353 throughout all levels of calculations, with IFIEs about –30 kcal/mol that account for a third of total interaction energy. These charged residues, Glu353, Arg394, and His524 are known to construct a hydrogen-bond network with EST and the side-chain phenyl group of Phe404 constructs T-shape π-π interactions with the ligand at the ligand-binding site of ER. The IFIE results indicate that strong interactions between ligand and these residues with the hydrogen bonds and hydrophobic moiety play a key role in the ER–ligand binding.

7.2.4 Orbital Interaction Analysis between Estrogen Receptors (ERs) and Ligands

As already mentioned, the negative charges transferred from ER to the ligands upon binding are related to their binding energies, and the ligand constructs a hydrogen-bond network with surrounding residues. In order to consider CT interactions between ER and EST at an orbital level, the configuration analysis for the fragment interaction (CAFI),[24,25] were carried out at the RHF/6-31G* level[15] (see Chapter 3 for details of the CAFI). These observations clearly explained the strength and the direction of hydrogen-bond networks in the ligand-binding site of the ER. The strong CT interactions between EST and ER were observed in Glu353 → EST, EST → Arg394, and EST → His524 in terms of the CAFI energy (Table 7.3). The pair

TABLE 7.3
The Charge Transfer and Polarization Energies[a] among Fragments Obtained in the Configuration Analysis for Fragment Interaction (CAFI) at the RHF/6-31G* Level[†]

	To					
From	**Glu353**	**His524**	**Leu387**	**Arg394**	**EST**	**water**
Glu353	–0.107	0.000	–0.001	–0.525	–19.792	–9.626
His524	0.000	–0.295	0.000	0.000	–1.045	0.000
Leu387	–0.001	0.000	–0.063	–0.067	–1.317	–4.281
Arg394	–0.031	0.000	–0.013	–2.750	–0.633	–0.373
EST	–0.870	–5.299	–0.432	–7.090	–3.168	–0.808
Water	–0.627	0.000	–1.685	–14.138	–0.355	–1.171

[a] Energies are in kcal/mol.

[†] *Note:* Interaction energies are shown with donor–acceptor directions.

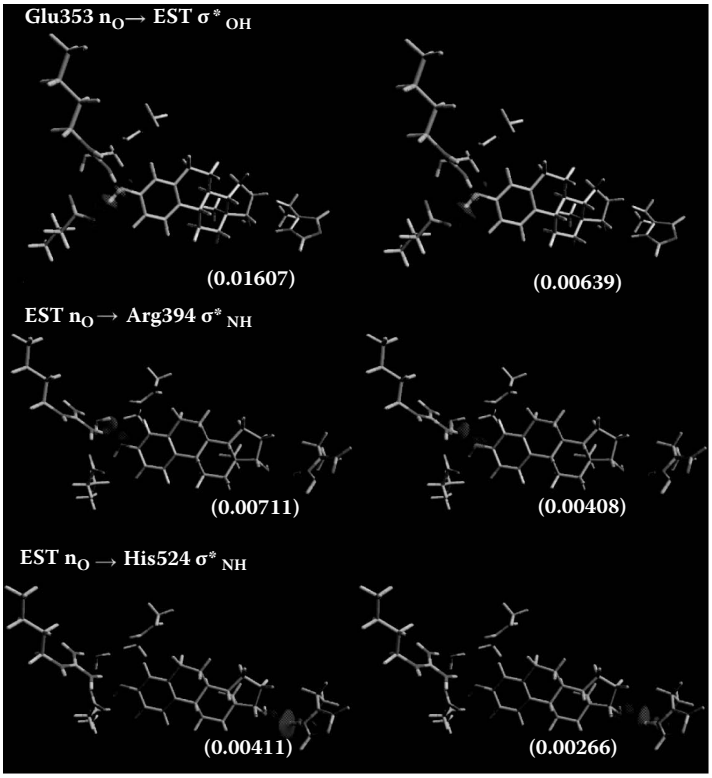


FIGURE 7.7 The pairs of "hole orbital" and "particle orbital" concerning charge transfer interactions between the estrogen receptor and the ligand obtained by CAFI. The phasing of the hole orbital is colored in red and blue. The corresponding colors of the particle orbital are yellow and green, respectively. Occupation numbers as transferred electrons for each pair are given in parentheses. Calculations were performed at the RHF/6-31G* level. **(See color insert.)**

of donor–acceptor orbitals is visualized in Figure 7.7. The strongest CT interactions were from the lone-pair orbital of carbonyl oxygen of Glu353 to the σ^*_{OH} orbital of the hydroxyl group of EST, "Glu353 $n_O \rightarrow$ EST σ^*_{OH}," and the occupation number of the transferred electron was 0.022. The second and third strongest CT interactions were EST $n_O \rightarrow$ Arg394 σ^*_{NH} and "EST $n_O \rightarrow$ His524 σ^*_{NH}" with occupation numbers of 0.011 and 0.007, respectively. These results clearly show that the major CT interactions were electron donation from Glu353 to EST, and that there were considerable CT interactions as the "back-donation" of electrons from EST to Arg394 and His524. Therefore, the "CT network" was found to occur through the "hydrogen-bond network" between the EST and strongly interacting charged/polarized residues of ER, which is displayed as arrows in Figure 7.2. The CAFI energy (or CT interaction energy) accounts for two-thirds of total IFIE, and therefore, CT interactions could be the major controlling factor of ER–ligand binding.

In addition to the residues of members of the hydrogen bond network, Phe404 strongly interacts with the ligand. From three-dimensional structure, the side-chain phenyl group of Phe404 is located perpendicular to the phenol ring of the ligand, and these moieties are thought to construct T-shape π–π interactions. As described above, the dispersion interaction between Phe#404 and the ligand was significant in ER–ligand binding. The fragment interaction based on local MP2 (FILM)[26] calculation was applicable to analyze weak dispersion interactions (see Chapter 3 for details of the FILM), and the CAFI was an efficient tool to consider hydrogen bond network and CT process. Here, FILM analysis was carried out with the local MP2 (LMP2) method and 6-31G basis set for the interaction between Phe#404 and the ligand. Figure 7.8 visualizes the orbital pair for the interactions between the π-orbitals of EST and moieties of Phe#404. The π orbital on the phenol group of EST interacts with the σ_{CH}, π and σ_{CC} orbital of Phe#404, and the CH–π, π–π, and CC–π interactions appear as stabilized interactions in order of interaction energy. These interactions

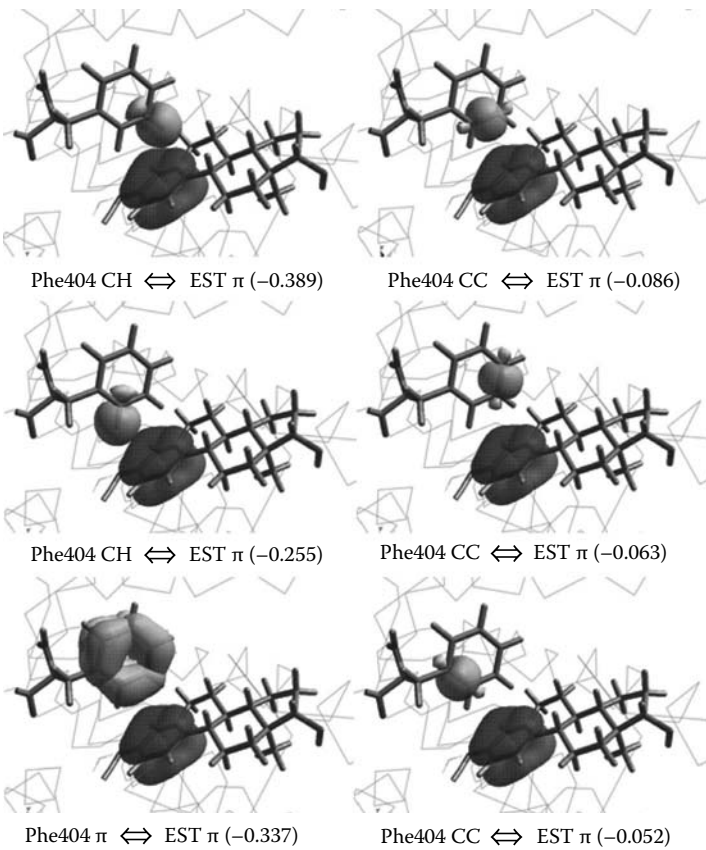


FIGURE 7.8 Visualizations of the orbital pairs for the interactions between the π-orbitals of Phe#404 and moieties of the ligand calculated at the LMP2/6-31G level. The values in parentheses are the pair correlation energies between displayed orbitals (in kcal/mol).

occupy 33% of IFIE of the Phe#404–ligand interaction. Here, both hydrogen bonding and dispersion interactions were characterized at the orbital level; the picture of these interactions at the ligand-binding site of ER is summarized in Figure 7.2.

As described above, molecular interactions between ER and its ligand were discussed in detail. The dispersion energies play an important role in structural stabilization in addition to the electrostatic energies, and the inclusion of electron correlation is essential to characterize ER–ligand interactions. By contrast, the agreement of relative binding energy with experimental binding affinity even at the RHF/STO-3G level would be attributed to the fact that CT interactions through hydrogen bond network were a major controlling factor of ER–ligand binding. Also, the cancellation of several effects, such as hydrophobic interaction and solvent effect, could occur by considering relative values. In such cases, semiquantitative results could be obtained even at the HF level with model receptor protein containing only the residues surrounding the ligand.

7.2.5 Molecular Interactions in the Liganded Retinoid X Receptor (RXR)

The RXR is also a member of the nuclear receptor (NR) superfamily that regulates the expression of many genes involved in various physiological actions of its ligands at the transcriptional level. RXR not only forms a homodimeric DNA complex, but also can form heterodimeric DNA complexes with various NRs.[27] Because RXR thus has diverse important biological roles associated with human life and diseases, it has been one of the primary targets of drug discovery. As well as the functions of other NRs, the functions of RXR are induced by the binding of ligands. A natural ligand of RXR is one of the carboxylic acid derivatives of vitamin A, 9-cis retinoic acid (9cRA), which controls morphogenesis, differentiation, and homeostasis during embryonal development and postnatal life. 9cRA is also an effective inhibitor of tumor cell growth, and this antitumor activity is useful in therapy and prevention of cancers such as human immunodeficiency virus (HIV)–associated Kaposi's sarcoma.[28,29]

On the basis of the FMO method, Ito et al.[16] addressed molecular interactions of liganded RXR with steroid receptor coactivating factor-1 (SRC1) coactivator to examine the contribution of helix 12 (H12), which contains the core of the transcriptional activation function 2 activating domain, to the coactivator binding of RXR (Figure 7.9). The interaction between H12 and SRC1 was thus proved to be the main cause for the stabilization of the coactivator binding. In particular, highly conserved charged (Glu#453) and hydrophobic (Phe#450) residues in H12 were found to have stronger electrostatic and dispersion interactions with SRC1 than the other charged and hydrophobic residues in H12, respectively. In addition, the CT from RXR to SRC1 was found to occur mainly by the changes in charges of H12 residues. Large positive and negative charge changes were observed especially for Glu#453 and for Lys#631 and Ile#632 in SRC1, respectively, indicating that Glu453 is an electron donor for Lys631 and Ile632 in this CT. They thus demonstrated quantitatively that H12 and its highly conserved residues significantly contribute to the coactivator binding not only by the Coulomb and dispersion interactions, but also by the CT described with the quantum-mechanical framework.

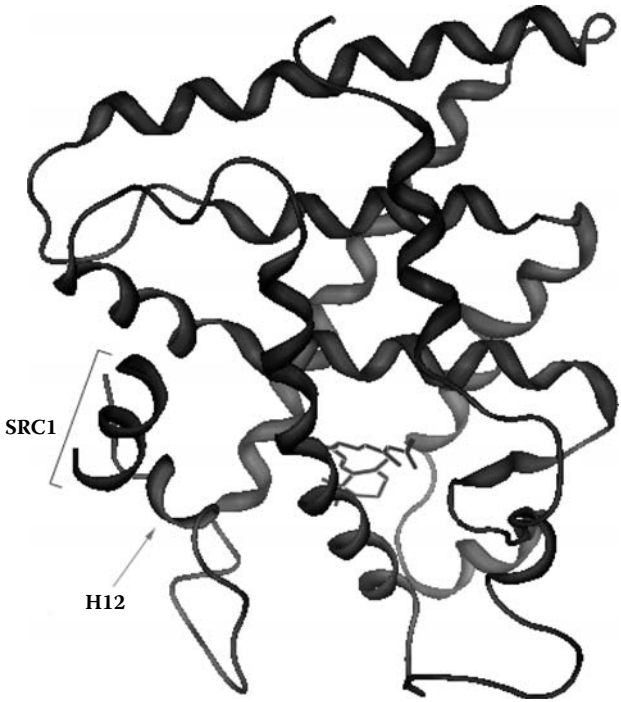


FIGURE 7.9 Ribbon display of the human retinoid X receptor (RXR) ligand-binding domain complexed with 9-cis retinoic acid (stick display) and steroid receptor coactivating factor-1 (SRC1) peptide. The position of helix 12 (H12) is also displayed.

Subsequently, Ito et al.[17] performed the FMO calculations for RXR complexes with its ligand 9cRA and coactivator SRC1 to examine the influence of mutations in transcriptional activation function 2 activating domain core (AF2C) of RXR on molecular interactions between 9cRA liganded RXR and SRC1 coactivator. The RXR–SRC1 interactions in three types of RXR–9cRA–SRC1 complexes, namely, a wild type (WT), a mutant whose Glu453 of AF2C was substituted by Lys (E453K), and another mutant whose Glu456 of AF2C was substituted by Lys (E456K), were compared. Through the comparison of WT, E453K, and E456K, possible causes for a marked decrease in the transcriptional activity of RXR by the mutation of Glu453, which is known as a highly conserved charged residue of AF2C, were discussed. It was quantitatively demonstrated that the strength of the RXR–SRC1 interaction correlates with the degree of the transcriptional activation (WT > E456K > E453K). In E453K, the RXR–SRC1 interaction was substantially reduced by the AF2C–SRC1 repulsive interaction, and the CT from RXR to SRC1 was also inhibited by the decreased electron donation from AF2C to SRC1. Thus, their findings suggested that the inhibitions of the local RXR–SRC1 interaction via AF2C and of the local CT from RXR to SRC1 via AF2C would be the possible causes for the marked decrease in the transcriptional activity of RXR.

Further, Ito et al.[18] carried out the *ab initio* FMO calculations for the R-subtype of the human retinoid X receptor (hRXR) complex with its natural ligand 9cRA to quantitatively specify the key residues with important roles for the ligand-inducible information transmission of RXR. In the RXR–9cRA complex, the transactivation H12 adopts a canonical agonist conformation, which corresponds to the transcriptional AF2C. Through the analyses of molecular interactions by the MP2 method, it was proven that Trp305 and Leu436 of the AF2C binding pocket would be important for the stabilization of the H12 canonical agonist conformation, and, at the same time, for the recognition of the 9cRA molecule. Besides, through the analyses of orbital interactions by the LMP2 method,[26] it was found that Trp305 and Leu436 would recognize the 9cRA molecule especially at its C19 methyl group, which has been most notably targeted to modify for agonist and antagonist design. Moreover, on the basis of the relationships of molecular interactions, it was suggested that the interactions of Trp305 and Leu436 with AF2C residues would be significantly influenced by the interactions of Trp305 and Leu436 with 9cRA. Thus, their findings quantitatively demonstrated that Trp305 and Leu436 would be the possible key residues for the information transmission in liganded RXR, accounting for their importance suggested by experiments. Altogether, these results substantiated that their approach is useful for the understanding of the detailed molecular mechanism underlying the transcriptional regulation of RXR and related nuclear receptors at the quantum-mechanical level.

7.3 BIOMOLECULAR SYSTEM INCLUDING DNA

The cyclic-AMP receptor protein (CRP) (also referred to as the "catabolite gene activator protein," or CAP) of *Escherichia coli* is among the most extensively studied sequence-specific DNA-binding proteins. Acting as a transcription factor, CRP regulates gene expression related to sugar metabolism. When cyclic-AMP (cAMP), acting as an intercellular signaling molecule, binds to CRP, following receptor dimerization, conformation changes and binding to specific DNA sequences near promoters are induced.[30–33] The CRP consists of two functional domains: a ligand-binding domain (135 residues) and a DNA-binding domain (74 residues). The CRP has the helix-turn-helix (HTH) DNA-binding motif and a recognition helix is the second helix of HTH. The consensus DNA sequences are 22 base pairs in length, and they exhibit twofold symmetry. Each half-site of the DNA has 11 base pairs ($5'$-$A_1A_2A_3T_4G_5T_6G_7A_8T_9C_{10}T_{11}$-$3'$) which bind to the CRP monomer, and the entire CRP–DNA complexed structure consists of a CRP dimer with cAMP and the 22 base pairs of DNA. There are highly conserved sequences in each half of the consensus DNA site (i.e., positions 4T:A to 8A:T), which play an important role in the sequence specificity for binding. Exhaustive experimental single base-pair substitutions were carried out in the conserved sequences in order to examine changes in the binding affinity between CRP and DNA; it was found that all types of the substitutions lead to lower affinities, and substitutions at positions 5G:C and 7G:C in particular gave the combinations with the lowest affinities.[33] Despite such experimental efforts to elucidate the sequence specificity, the mechanism of CRP–DNA binding remains unclear.

In this report, we applied the FMO method to examine the CRP–cAMP–DNA monomer complex system, primarily using the MP2 level of theory. In addition, we addressed the intermolecular interactions between CRP and DNA and the intramolecular interactions in the DNA in order to obtain insight into the sequence-specific binding between DNA and CRP and the effect of CRP binding on DNA duplex stability.

7.3.1 Molecular Modeling

The initial atomic coordinates of the CRP–cAMP–DNA monomer complex were obtained from the Protein Data Bank (PDB),[10] entry 1O3Q.[35] Based on this structure, one cAMP molecule, 200 amino acid residues of the CRP monomer, and 11 base pairs of the DNA (5′-$A_{-2}A_{-1}A_1A_2A_3T_4G_5T_6G_7A_8T_9$-3′) were selected for simulation. Classical molecular mechanics (MM) and molecular dynamics (MD) calculations, including the coordination and minimization of hydrogen atoms, were performed using the program PEACH.[36–38] The terminal amino acid and all charged amino acid residues were assumed to be in the charged state. The histidine residues were assigned as being in the τ-type configuration. Therefore, a total net charge of the CRP was zero. For DNA, the 5′- and 3′-terminals were capped with a hydroxyl group without a phosphate group and with a hydroxyl group, respectively. All phosphate groups in the DNA duplex were assumed to be in the charged state, and then the total charge of the DNA was –20e for 11 base pairs. The phosphate group in cAMP was also treated as charged, and the charge of cAMP was –1e. We explicitly included all atoms in the CRP–cAMP–DNA complex surrounded by a free 6Å-water shell (excluding those within 2.8 Å of the protein) and 21 Na^+ counterions. The solvated structure thus constructed was geometry-optimized successively by the Steepest Descent (SD) method (20 steps), by the Quenched Dynamics (QD) method (40 ps, 5 K), and then by the Conjugate Gradient (CG) method until the threshold energy decreased to less than 0.1 kcal/mol. The AMBER94 force field[39] was used for the calculations.

Using CG-minimized geometries, the *ab initio* FMO calculations were carried out at the RHF and the MP2 levels with 6-31G basis set.[34] The mechanism of molecular fragmentation is as follows: the cAMP molecule, each amino acid residue for CRP, and each base and backbone unit for DNA were individually treated as a single fragment.

7.3.2 Sequence-Specific DNA–CRP Binding

Figure 7.10 shows the IFIE between the CRP–cAMP complex and each DNA base or backbone fragment. The strongly stabilized fragments of DNA were backbone fragments around –1A:T ~ 4T:A, which is not at the highly conserved sequences. In the base–CRP interactions, however, the base pairs in the highly conserved sequence of the consensus sequence (4T:A ~ 8A:T) strongly interacted with CRP. In particular, 7G:C and 5G:C showed the strongest interactions; this result agreed with the experimental finding that the DNA base-pair substitutions at positions 5 and 7 gave the lowest binding affinity between CRP and DNA among an exhaustive number of possible single base-pair substitutions.[33] Although a similar tendency was observed by the quantum-mechanical analyses through several levels of theory, the AMBER94

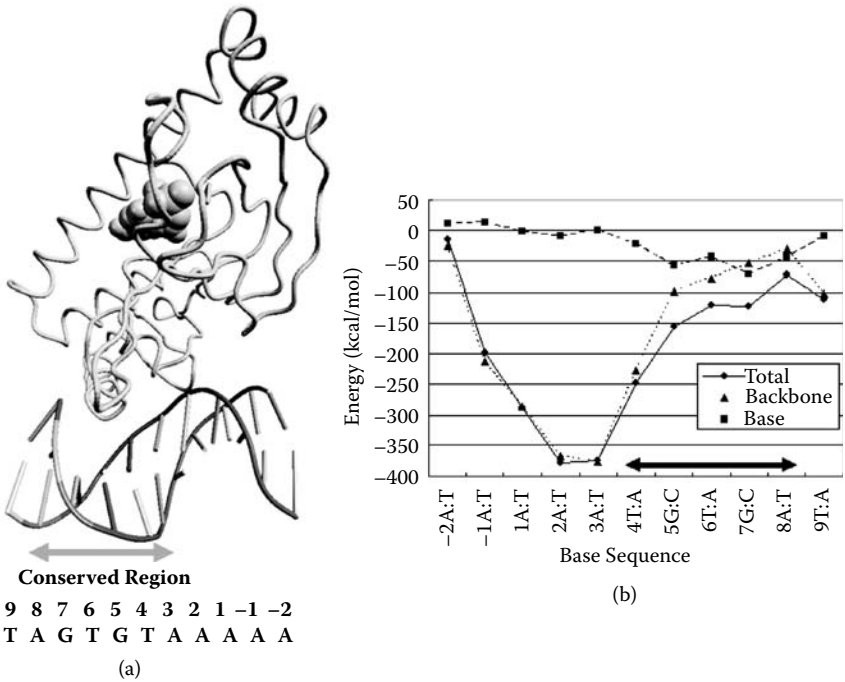


FIGURE 7.10 Interfragment interaction energy (IFIE) analysis for interactions between cyclic-AMP receptor protein (CRP) and DNA calculated at the MP2/6-31G level. (a) IFIE between the CRP-cAMP complex and each DNA base or backbone fragment. The strength of the interactions is represented according to darkness on DNA. (b) Numerical representation of IFIE between the CRP–cAMP complex and each base pair. Backbone, base, and total indicate backbone (sugar-phosphate) pair fragment, base-pair fragment (base portion only), and nucleotide pair fragment, respectively. **(See color insert.)**

analysis produced different features.[34] Remarkably stronger interactions were not observed at the highly conserved sequences with the use of the AMBER94, meaning that this "classical" approach is not enough to describe the site specificity. This finding suggests that descriptions of weak interactions between neutral species (e.g., base–protein interactions) require quantum-mechanical treatment, whereas strong electrostatic interactions between charged or well-polarized species can be represented by "classical" force field analysis. In the IFIE analysis between the DNA duplex and each amino acid residue or cAMP fragment, both stabilized and destabilized fragments were observed spread throughout the three-dimensional structure, and many strong interactions were found in the DNA recognition helix.[34]

7.3.3 Intra- and Interstrand Interactions of DNA

The IFIE analysis was also carried out for intra-DNA interactions. The inter- and intrastrand interactions of the DNA base were examined to identify the presence of

hydrogen bonding and stacking interactions.[34] In the DNA–CRP–cAMP complex, the interstrand hydrogen bonds between Watson–Crick base pairs showed stronger interactions than indicated by the simple base–base molecules in the gas phase. No qualitative difference was observed between the HF and MP2 methods. In contrast, stacking energies were different in both methods. The 1,2-stacking energies for the base pairs closest to each other such as 1A:T to 2A:T, can be described by the sum of IFIEs for intra- and interstrand interactions. The 1,2- and more separated stacking interactions included at positions 4T:A to 8A:T are displayed in Figure 7.11. The 1,2-stacking interactions were found to be repulsive at the HF level but attractive at the MP2 level. The HF-MP2 energy differences corresponding to the correlation energies were significant and comparable at each 1,2-stacking energy; the MP2 calculations revealed 14~21 kcal/mol more stable interactions than the HF calculations. In conjunction with the intrastrand stacking analysis, the dispersion term, which was not included in the HF approximation, was dominant in the stabilization of base stacking. For the second-neighbor interactions (1,3-stacking; e.g., 4T:A—6-T:A interactions), weak repulsive interactions were observed in both HF and MP2 levels. Stacking interaction energies beyond those of 1,3-stacking showed moderate decreases, but the interactions did not vanish entirely across the entire range of 11 base pairs.

The comparison of the results obtained with free and complexed DNA was performed for intra- and interstrand interactions.[34] The base-pairing interactions were stronger, and the stacking interactions were weaker in the complexed structure. In regard to the 1,2-stacking interactions, changes in the stabilization energy were significant in the highly conserved DNA sequence that consisted of positions 4T:A to 8A:T; these positions directly interacted with the DNA recognition helix of the CRP. Furthermore, the negative charges of the base portion were transferred to the

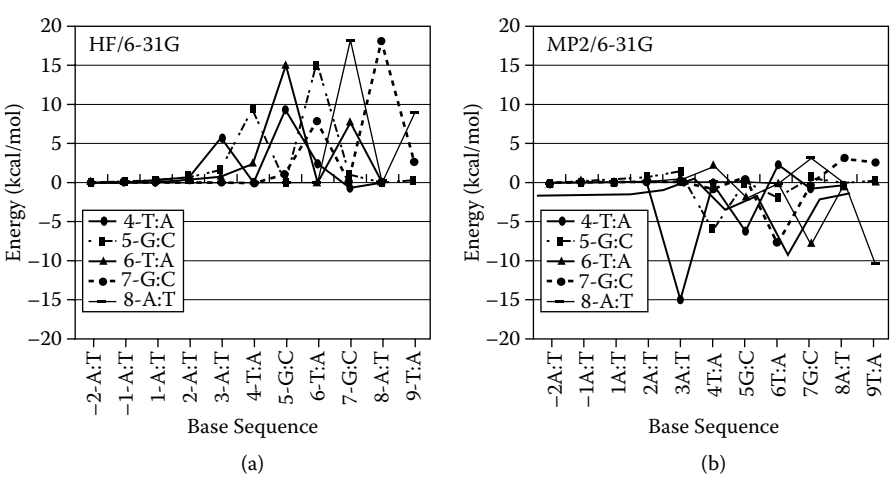


FIGURE 7.11 Stacking interactions of DNA calculated at the (a) RHF/6-31G and (b) MP2/6-31G levels.

CRP, and this reduction in charge in turn strengthened the base-pairing interactions. Therefore, the stability of the DNA was altered due to the electrostatic and CT interactions between DNA and CRP upon binding.

7.3.4 Interfragment Interaction Energy (IFIE) Map

In order to obtain a comprehensive picture for the interactions of the CRP–cAMP–DNA complex, an IFIE map was examined at the MP2/6-31G level[40] and visualized in Figure 7.12 (see Chapter 3 for details about the IFIE map). Upper (red) and lower (blue) triangles indicate the plots of negative and positive IFIE values, respectively. In the intra-CRP interactions (Fragment No. 1-200), secondary structures of protein were reflected as red bands based on the stable interactions of intramolecular

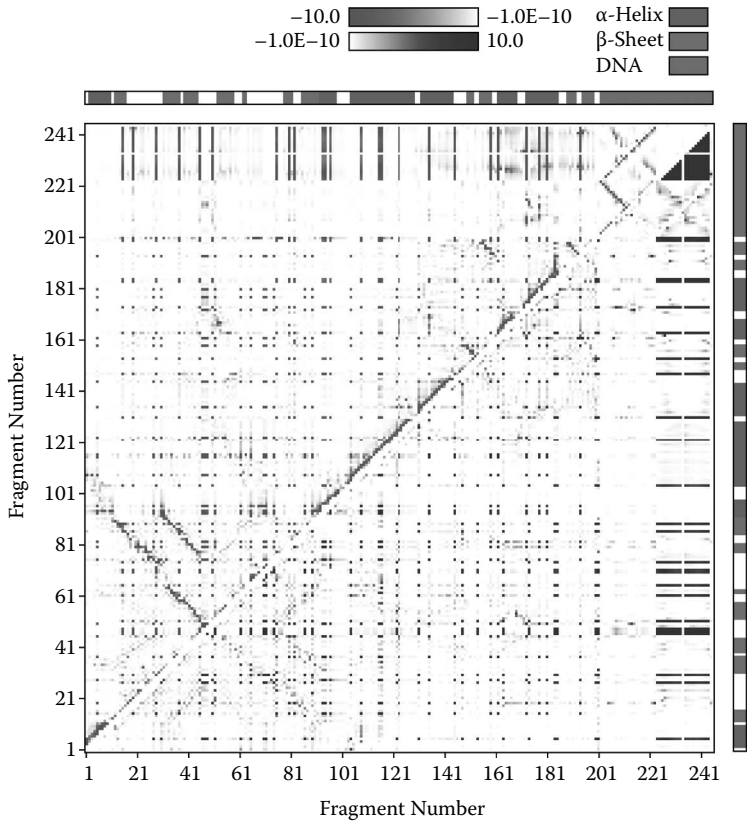


FIGURE 7.12 Two-dimensional map of the interaction energies (IFIE map) for interactions among DNA, cyclic-AMP receptor protein (CRP), and cyclic-AMP (cAMP) calculated at the MP2/6-31G level. Upper and lower triangles indicate the plots of negative and positive energy values, respectively. The Fragment Nos. 1 to 200, 201, 202 to 223, and 224 to 245 correspond to CRP, cAMP, DNA base, and DNA backbone, respectively. **(See color insert.)**

hydrogen bonds; α-helices appeared near or parallel to the diagonal line, and β-sheets appeared perpendicular to the diagonal line. The "cross motif" in the intra-DNA interactions of base portion (Fragment No. 202-223) indicates interactions of Watson–Crick base pair and stacking pair as perpendicular and parallel red bands, respectively. Also, the blue cross motif of backbone portion indicates electrostatic repulsion of the negatively charged phosphate group in the Watson–Crick and stacking pairs. As described in the former section, base-pair stacking interactions appeared to be attractive in the MP2 level and repulsive at the HF level. Therefore, in such pattern analysis, it is essential that dispersion interactions are correctly evaluated by using the electron correlation method.

In this section, the quantum-mechanical calculations for the whole DNA–CRP–cAMP system revealed both the presence of intermolecular electrostatic and CT interactions between DNA and CRP-cAMP, as well as the alteration of DNA duplex stability, including the associated intramolecular electrostatic and dispersion interactions, upon the binding of the DNA to CRP. However, we obtained only qualitative pictures from calculations carried out for the gas phase with the optimized geometry in an aqueous solution. Real water molecules and counterions should be included in future studies in order to obtain a more reliable understanding of these interactions.

7.4 PROTEIN MOLECULAR INTERACTIONS ASSOCIATED WITH INFLUENZA HEMAGGLUTININ

7.4.1 Sialic Acid Recognition of Influenza Hemagglutinin

Over the past century, the emergence of epidemic influenza has been a serious threat to human health. The origin of human influenza viruses is thought to be avian influenza virus, because all the subtypes are found in an avian host. Newly adapted avian influenza virus to human host or reassortant virus could be pandemic because we have no immunity for it, as shown in our history by the 1918 (H1N1), 1957 (H2N2), and 1968 (H3N2) pandemics. Recently, the first H5 avian influenza virus-infected patient was reported and the emergence of new pandemic influenza was alerted.

Influenza virus binds to receptors on the host cell surface by hemagglutinin (HA) protein. The HA exists as a homotrimer on the viral surface, and each monomer is composed of two subunits, HA1 and HA2. The receptor-binding domain (RBD) is located at membrane-distal tip of each monomer. Three secondary structure elements, Helix190 (residues 190 to 198), Loop130 (residues 135 to 138) and Loop220 (residues 221 to 228), and conserved residues Tyr98, Trp153, and His183 (H3 HA numbering) are involved in the RBD. Receptors contain glycans with terminal sialic acids and HAs primarily recognize the terminal part. Because the rich linkage type of sialic acid to vicinal galactose on the targeted cell differs according to species (α2–3 or α2–6), HA acquires the binding specificity for the linkage type on the host. Human influenza viruses preferentially recognize α2–6 (human receptor), whereas avian influenza viruses prefer α2–3 (avian receptor). Hence, it is believed that the change of receptor-binding preference of avian HA is the critical first step of adaptation to a human host.[41]

In the case of H1, H2, and H3 subtypes, it was shown that as few as two mutations at the RBD are responsible for adaptation to a human host by genetic and mutational

studies.[42] The substitutions Gln226Leu and Gly228Ser switch the binding specificity of H2 and H3 HAs from avian to human receptor, whereas the substitutions Glu190Asp and Gly225Asp change the binding specificity of H1 HA. To understand the structural basis of the receptor specificity of HAs, X-ray crystal structures of HA/receptor analogue complexes have been determined.[43,44] By the structural studies, a number of interaction sites involved in the receptor binding have been revealed, and several suggestions have been made as to the mechanism of how HAs recognize the different linkages and change the receptor specificity with only two residue substitutions. It was, however, difficult to evaluate inter- and intramolecular interaction energies of HA–receptor complexes quantitatively, and thus the roles of each residue in the receptor binding have not been understood well.

Iwata et al.[45] performed *ab initio* theoretical studies on the binding specificity of HAs (H3 avian HA, H1 human HA, H1 swine HA, and H5 avian HA) to avian and human receptors. They reported the interaction patterns of RBD of HAs estimated by *ab initio* FMO calculations, taking appropriate account of electron correlation effects. The coordinates used in their study were prepared from crystallographic structural data obtained from the PDB. The structures employed in the calculations and their experimental binding specificities are listed in Table 7.4. Stereo views of each HA complexed with avian and human receptors are shown in Figure 7.13. They optimized the location of unlocated oxygen atoms and hydrogen atoms of the complexes by molecular mechanics energy calculations based on the MMFF force field and cut out 82 residues of RBD of the optimized complex for the use in FMO calculations. Molecular mechanics and dynamics calculations were carried out by using the Molecular Operating Environment (MOE) software (Chemical Computing Group Inc.). Fragmentation of the complexes for the FMO calculations was performed as follows (Figure 7.14): Sialic acid was divided into two fragments and sugar was treated

TABLE 7.4
Influenza Hemagglutinin Employed in Iwata et al.[a]

Subtype	Origin	Preference	Ordered Receptor Analogue	PDB ID
H1N1	Human	α2-3, α2-6	α2-3 : Sia1-Gal2-GlcNAc3	1RVX
			α2-6 : Sia1-Gal2-GlcNAc3	1RVZ
H1N1	Swine	α2-3 < α2-6	α2-3 : Sia1	1RVT
			α2-6 : Sia1-Gal2-GlcNAc3-Gal4-GlcNAc5	1RVO
H3N2	Avian	α2-3	α2-3 : Sia1-Gal2-GlcNAc3	1MQM
			α2-6 : Sia1-Gal2	1MQN
H5N1	Avian	α2-3	α2-3 : Sia1-Gal2-GlcNAc3	1JSN
			α2-6 : Sia1	1JSO

[a] Receptor binding preference, ordered receptor analogues in the X-ray crystal structure analysis and Protein Data Bank (PDB) ID are listed.

Source: Iwata, T., Fukuzawa, K., Nakajima, K. et al. 2008. *Comput. Bio. Chem.* 32: 198–211.

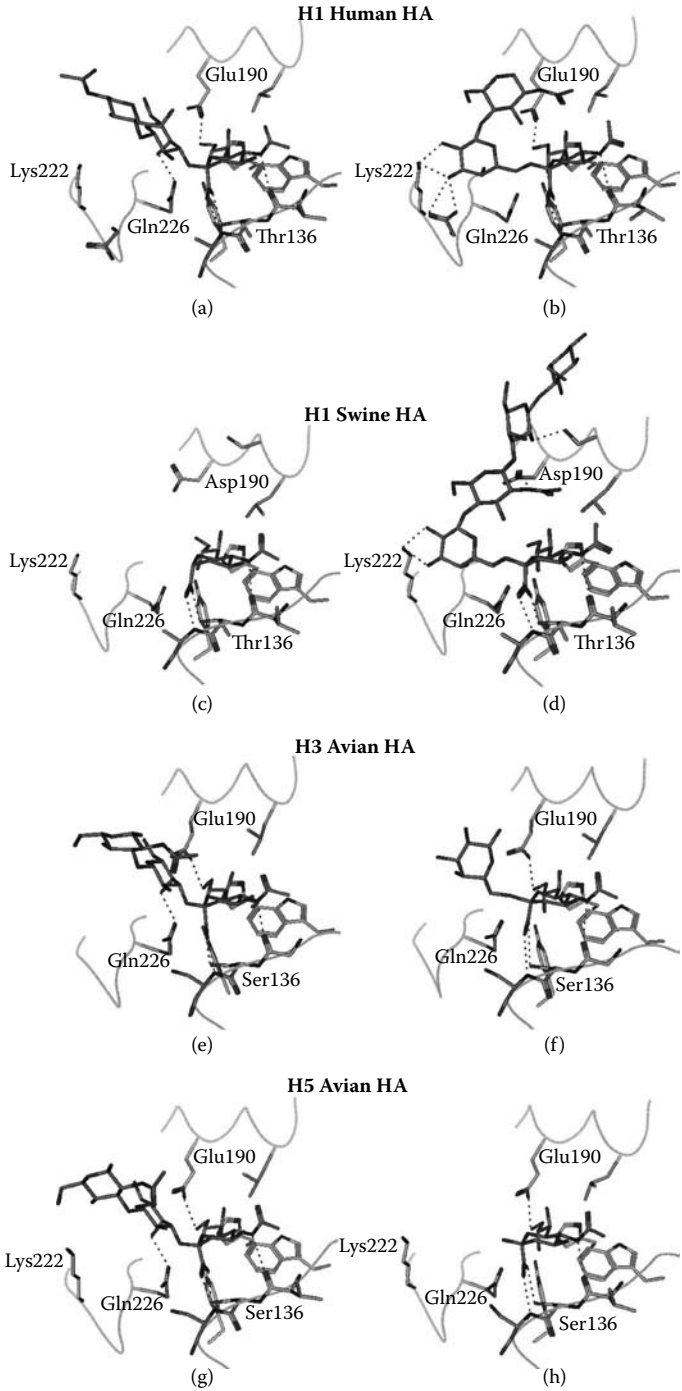

H1 Human HA
Glu190
Lys222
Gln226
Thr136
(a)
Glu190
Lys222
Gln226
Thr136
(b)
H1 Swine HA
Asp190
Lys222
Gln226
Thr136
(c)
Asp190
Lys222
Gln226
Thr136
(d)
H3 Avian HA
Glu190
Gln226
Ser136
(e)
Glu190
Gln226
Ser136
(f)
H5 Avian HA
Glu190
Lys222
Gln226
Ser136
(g)
Glu190
Lys222
Gln226
Ser136
(h)

FIGURE 7.13

FIGURE 7.13 Interactions of human H1, swine H1, avian H3, avian H5 hemagglutinin (HA) with avian receptor (left: a, c, e, g) and with human receptor (right: b, d, f, h). Possible hydrogen bonds are represented by broken lines. (a) H1 human HA with avian receptor and (b) H1 human HA with human receptor (PDB ID: 1RVX and 1RVZ). (c) H1 swine HA with avian receptor and (d) H1 swine HA with human receptor (PDB ID: 1RVT and 1RVO). (e) H3 avian HA with avian receptor and (f) H3 avian HA with human receptor (PDB ID: 1MQM and 1MQN). (g) H5 avian HA with avian receptor and (h) H5 avian HA with human receptor (PDB ID: 1JSN and 1JSO). (Reproduced from Iwata, T., Fukuzawa, K., Nakajima, K. et al. 2008. *Comput. Bio. Chem.* 32: 198–211. With permission.)

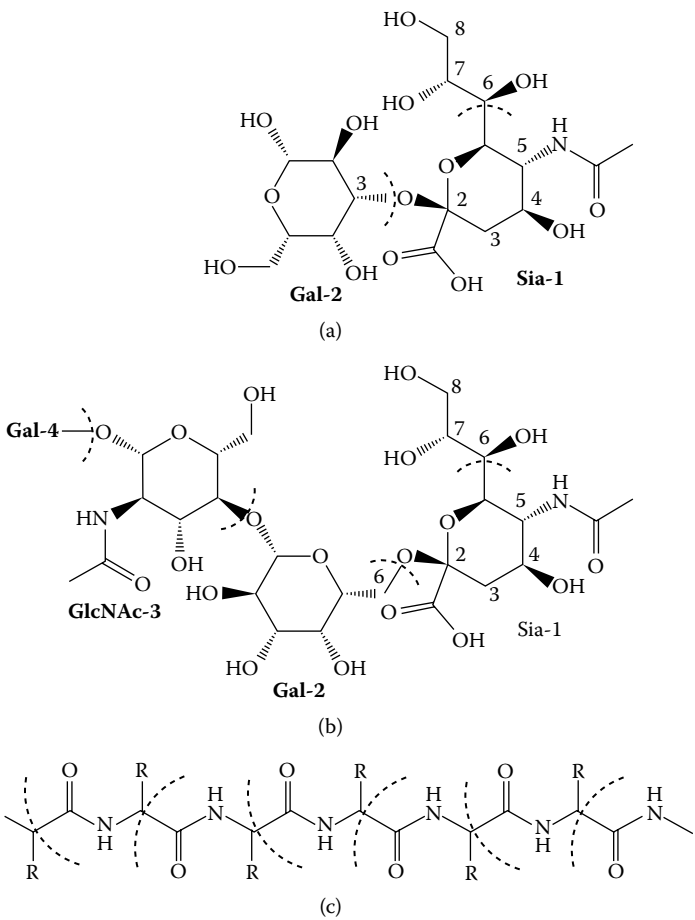


FIGURE 7.14 Fragmentation of molecules. Saccharides of (a) α2–3 and (b) α2–6 receptor analogues. (c) Hemagglutinin protein. Sia, Gal, and GlcNAc refer to sialic acid, galactose, and N-acetylglucosamine, respectively. (Reproduced from Iwata, T., Fukuzawa, K., Nakajima, K. et al. 2008. *Comput. Bio. Chem.* 32: 198–211. With permission.)

as a single fragment. Each amino acid residue at the RBD was assigned as a single fragment except cysteine pair forming S–S bond which was treated as a single fragment. *Ab initio* FMO calculations were carried out with the MP2/6-31G method.

The binding energy of HA and receptor is evaluated as

$$\Delta E = E_{complex} - (E_{HA} + E_{receptor}) \tag{7.3}$$

where E_{HA}, $E_{receptor}$, and $E_{complex}$ refer to the energies of HA, receptor, and their complex, respectively. In addition, the IFIE, ΔE_{IJ}, obtained in the FMO calculations provides useful information for specifying important residues. The IFIE values represent interaction energies of a ligand with an amino acid residue or between amino acid residues because each amino acid is assigned as a single fragment. The IFIEs were calculated to analyze the interaction pattern and to estimate the contributions of each residue to binding. In order to address the change of IFIEs between fragment I belonging to HA and fragment J on the receptor binding, it is also convenient to introduce

$$\Delta\Delta E_{IJ} = \Delta E_{IJ}(\text{HA-receptor complex}) - \Delta E_{IJ}(\text{uncomplexed HA}) \tag{7.4}$$

and their summation over the fragments J (not equal to I),

$$\Delta\Delta E_I^{total} = \Sigma_J \Delta\Delta E_{IJ} \tag{7.5}$$

The latter then refers to the contribution of each fragment I to the binding affinity between HA and receptor. It is noted here that $\Delta\Delta E_{IJ} = \Delta E_{IJ}$ (HA-receptor complex) when the fragment J belongs to the receptor.

For avian H3HA/avian receptor complex, three kinds of sialic acid-galactose receptors registered in PDB were employed (PDBID: 1MQM). Two of the receptors, R1 and R2, are of Sia1-Gal2-GlcNAc3 type, and the other, R3, is of Sia1-Gal2 type. Employing R3 as a stable binding structure, Iwata et al.[45] analyzed $\Delta\Delta E_I^{total}$ for all the residues in HA, where the binding energy was divided into the intramolecular (intra-HA) and intermolecular (HA-receptor) contributions. As shown in Figure 7.15a, $\Delta\Delta E_I^{total}$s for conserved residues such as Tyr#98, Ala#138, Trp#153, His#183, Glu#190, Leu#194, Gln#226, and Gly#228 were found to take totally negative values, thus stabilizing the receptor binding. In the case of Glu#190, the intermolecular interaction with the receptor was destabilized, and the intramolecular interaction was stabilized more strongly so that the total $\Delta\Delta E_I^{total}$ became negative. When $\Delta\Delta E_I^{total}$ for Glu#190 was decomposed into each contribution from the fragment J, the contributions from Tyr#98, Trp#153, and His#183 to the stabilized binding were found to be significant, as illustrated in Figure 7.15b. These results indicate that intramolecular (intra-HA) interactions between conserved residues play an important role for HA-receptor binding.

As for avian H3HA/human receptor complex, they performed the FMO calculations for two kinds of receptors, R1 and R2, both of which are of Sia1-Gal2 type (PDB ID:1MQN). They calculated the binding energies ΔE, and the calculated values are listed in Table 7.5 in comparison with the case of avian H3HA/avian receptor complex. Even taking account of the fluctuations between the various observed

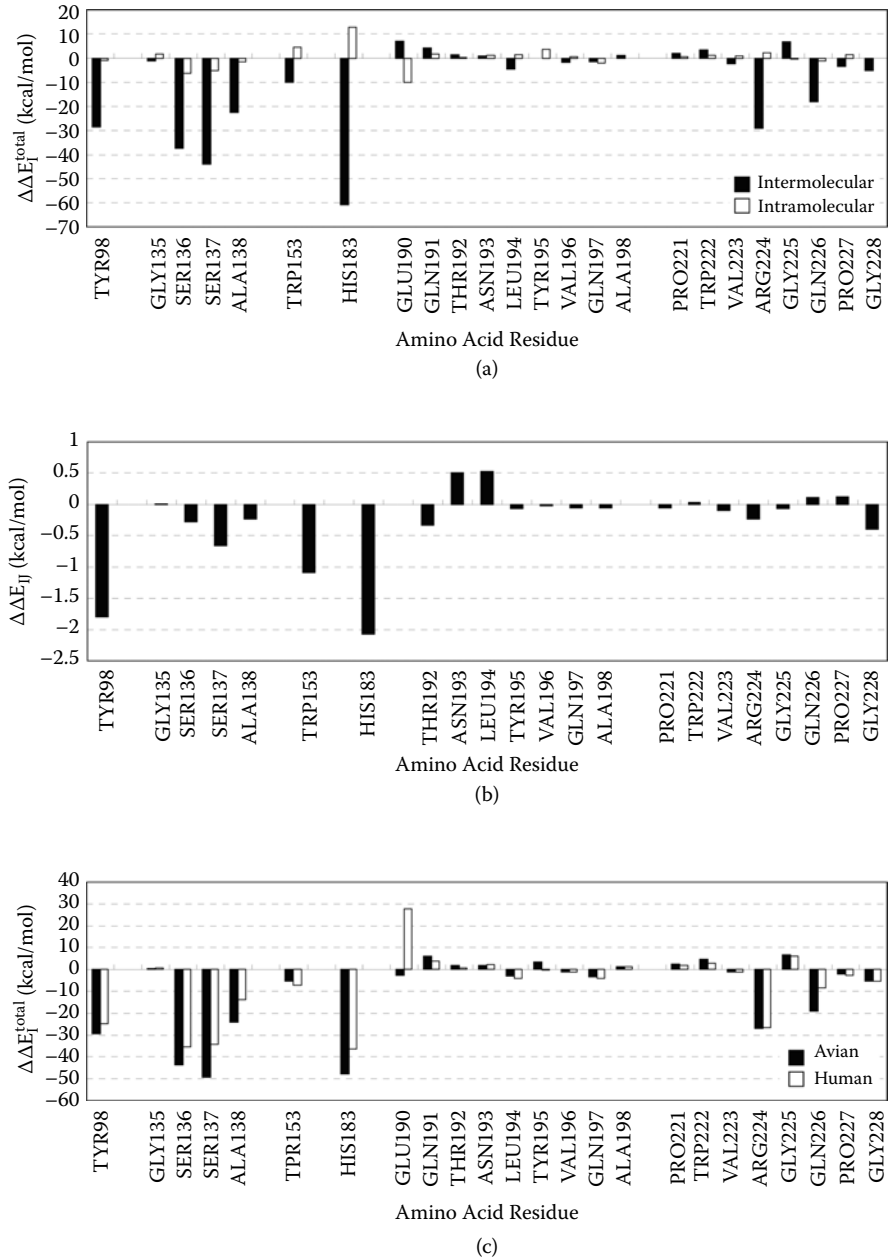


FIGURE 7.15 Structural variation and interaction patterns of H3 avian hemagglutinin–receptor complexes. (a) E_I^{total} of avian (R3) receptor complex is divided into inter- and intramolecular interactions. (b) E_I^{total} for Glu#190 is decomposed into each contribution from the fragment J. (c) Comparison of E_I^{total} of each amino acid residue between avian (R3) and human (R2) receptor complexes. (Reproduced from Iwata, T., Fukuzawa, K., Nakajima, K. et al. 2008. *Comput. Bio. Chem.* 32: 198–211. With permission.)

TABLE 7.5
The Binding Energies (ΔE)[a] between Influenza Hemagglutinin (HA) and Receptor[b] Calculated by the FMO Method at the MP2/6-31G Level

Receptor	H3 Avian HA	H1 Human HA	H1 Swine HA	H5 Avian HA
Avian	–352.9	–293.3	–363.3	–299.2
Human	–292.4	–335.9	–390.5	–283.9

[a] Energies are in kcal/mol.
[b] The receptor analogues used in the calculations differ among the complexes (see text for details).

structures, the magnitudes of binding energies for avian H3/human receptor complex are much smaller than those for avian H3/avian receptor complex, explaining the binding preference for the latter. Figure 7.15c illustrates a comparison of $\Delta\Delta E_I^{total}$ between avian H3/avian receptor (R3) and avian H3/human receptor (R2). As seen in Figure 7.15, the IFIEs associated with Tyr#98, Ser#136, Ser#137, Ala#138, His#183, Glu#190, and Gln#226 differ between the two complexes, suggesting that these residues would play essential roles for the binding specificity of avian H3 to avian receptor. In particular, Glu#190 shows a significant difference between avian and human receptors, indicating the importance for the binding specificity. The human receptor is complexed with avian H3 at a more distant location than the avian receptor. This shift in the binding location of human receptor has been ascribed to the avoidance of disadvantageous interactions between hydrophobic groups in the human receptor and the polar Gln226,[44] while the avian receptor can form hydrogen bonding with Gln226. However, the present quantitative analysis has shown that the human receptor loses favorable interactions with Loop130 and His#183 due to this shift, whose magnitudes seem to be greater than the gain associated with Gln#226.

Further, Iwata et al.[45] considered human H1HA/avian receptor (Sia1-Gal2-GlcNAc3) complex (PDB ID:1RVX) for which there are six X-ray crystal structures, R1 through R6. Because all the structures are seen to be similar regarding the location of receptor, they employed R1 as a representative structure for the IFIE analysis. Because the GlcNAc3 part is exposed to water solvent and has no contact with HA, they used the Sia1-Gal2 part for the FMO calculation. Through the calculations, they found that Glu#190 interacts repulsively with the avian receptor and makes the receptor binding unstable, even taking account of the change of intramolecular (intra-HA) interactions (Figure 7.16a). This result is somewhat strange in light of the experimental fact of mutations that Glu190 is indispensable for the avian receptor binding. The calculated binding energy of human H1 and avian receptor R1 was –293.3 kcal/mol. As for the humanH1HA/human receptor (Sia1-Gal2-GlcNAc3) complex, on the other hand, there are also six X-ray structures, R1 through R6 (PDB ID:1RVZ). The location of the receptors is similar, and thus, R1 was employed as a representative structure for the IFIE analysis. Because the Glc-NAc3 part has no contact with HA as in the avian

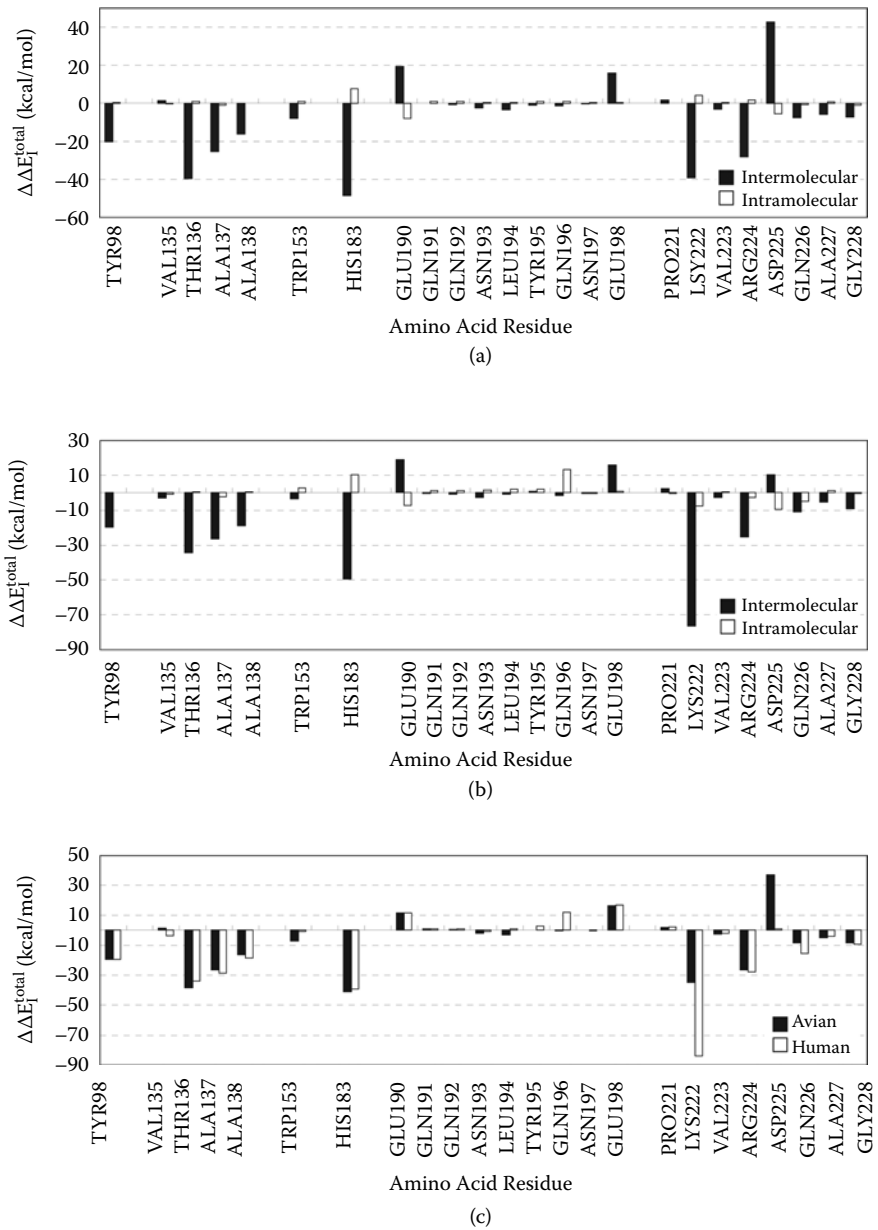


FIGURE 7.16 Interaction patterns of H1 human hemagglutinin–receptor complexes. The calculated $\Delta\Delta E_I^{total}$ values of each amino acid residue at the receptor binding site of (a) avian receptor complex and (b) human receptor complex are divided into inter- and intramolecular interactions. (c) Comparison of $\Delta\Delta E_I^{total}$ of each amino acid residue between avian and human receptor complexes. (Reproduced from Iwata, T., Fukuzawa, K., Nakajima, K. et al. 2008. *Comput. Bio. Chem.* 32: 198–211. With permission.)

receptor case, they used the Sia1-Gal2 part for the FMO calculation. The binding energy of human H1 and human receptor was then estimated to be –335.9 kcal/mol by the FMO calculation, which is much lower (more stable) than that for avian receptor. The calculated $\Delta\Delta E_I^{total}$ value for Gln#226 of human receptor complex is lower (more stable) than that of avian receptor complex (Figure 7.16c). This result explains the experimental results indicating that H1 avian HA acquires the binding ability for human receptor with retaining Gln226, in contrast to H2 and H3 HAs.[42] It was found in mutation experiments that Asp225 is important for the human H1/human receptor binding. However, as seen in Figure 7.16b, the present IFIE analysis has shown that Asp#225 has essentially no contribution to the receptor binding, because the intermolecular (HA-receptor) interaction gives unfavorable contribution compensating the favorable contribution due to the intramolecular (intra-HA) interaction.

Thus, in terms of the *ab initio* FMO method, Iwata et al.[45] established a computational scheme to quantitatively analyze and predict the binding affinity of HAs to avian and human receptors. Through the IFIE analysis based on the FMO calculations, they could also specify and characterize important residues that would play an essential role in the binding specificity between HA and receptor. These analyses would be useful for predicting possible shifts of the host range of influenza virus from birds to human, and some related theoretical approaches are underway.[46–49]

7.4.2 Antigen–Antibody Interaction of Influenza Hemagglutinin

Next, the FMO method was applied to the analysis of the influenza HA antigen–antibody system. Influenza virus has a remarkable ability to escape host defense mechanisms by altering its antigenic character, especially through changes of amino acid residues in the HA protein. This property is referred to as antigenic drift and has been thought to result from the accumulation of a series of amino acid changes in antigenically important regions of HA. It is thus essential to elucidate the molecular mechanisms by which viruses alter their antigenic character in order to find a way to control the epidemics of influenza.

Employing an HA-Fab antigen–antibody system of H3N2 A/Aichi/68 influenza virus, Mochizuki et al.[50] performed FMO-MP2/6-31G calculations and analyzed the IFIEs between the residues. The calculations were carried out both on PC clusters in laboratory and on the Earth Simulator in Yokohama using up to 4096 vector processors. The 1EO8 coordinate set of PDB (Figure 7.17a) was employed in a standard protocol of molecular modeling for FMO calculations. The antigen (HA) is composed of two domains, HA1 and HA2, and the antibody (Fab-fragment) also consists of two chains (see Figure 7.17a). The total number of residues is as many as 921 with 14,086 atoms. Then the actual number of fragments to be processed was 911 because of Cys-S-S-Cys treatments. As the result, the analyzed system had 78,390 AOs of 6-31G basis, and it has provided one of the world's largest correlated calculations for biomolecular systems. The benchmark results are given in Table 7.6, where the data of FMO-RHF jobs are also included for comparison. For the FMO-MP2 calculations, the computational performances of the influenza HA antigen–antibody system are superior to those of smaller E2020–AChE complex,[50] and the enlargement of computational task might be responsible for this improvement. The

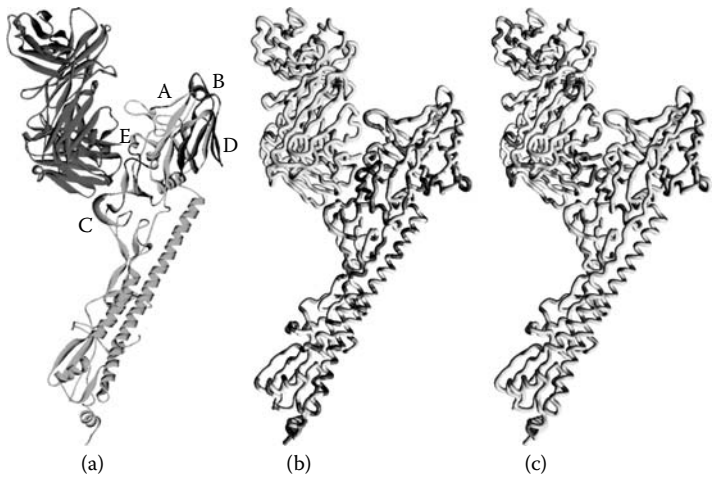


FIGURE 7.17 Graphic representation and visualized interfragment interaction energy (IFIE) results of influenza hemagglutinin (HA) antigen–antibody system. (a) Antigen and antibody (Fab-fragment) are drawn in green and purple colors, respectively. The crucial antigenic sites A through E in HA1 domain are indicated with special colors and labels (the PDB data 1EO8). (b) IFIEs of the residues in antigen with the whole Fab-fragment of antibody; (c) IFIEs of the residues in antibody with the HA antigen. MP2 correction was included. Red and blue colors mean the interaction energies of stabilization and destabilization, respectively. **(See color insert.)**

scalings of acceleration against the increments of VPUs are encouraging. It is noted here that the FMO-MP2 jobs could be completed in 91.4 minutes with 2048 VPUs and in 53.4 minutes with 4096 VPUs. The efficiency of about 12% in the gross speed is considered to be reasonable at the present time. In comparison with the FMO-HF jobs in which no electron correlation is incorporated, the relative cost of FMO-MP2 is confined to be only 2.4 to 2.7 times. Namely, a conventional discussion on the relative costs of HF and MP2 calculations is not valid in the FMO scheme, showing its superiority in large-scale calculations of proteins.

These FMO-MP2 calculations[50] have provided useful information concerning the specific molecular recognition in the HA antigen–antibody systems. Figure 7.17 visualizes the IFIE results of the residues in antigen with the whole Fab-fragment of antibody (Figure 7.17b) and those of the residues in antibody with the HA1-HA2 antigen (Figure 7.17c). This figure illuminates the situation that small but vital polarizations were induced through the antigen–antibody binding in a wide range of protein units, while the screening effects on the long-range electrostatic field would be important in actual situations. It is well known that the HA1 domain has five antigenic sites A through E (see again Figure 7.17a), which are remarkably recognized by the antibody in a specific way. The residue-summed values of IFIEs between these antigenic sites and the whole Fab-fragment of antibody are listed in Table 7.7. A large stabilization is obtained for site E being closest to the Fab-fragment, where two charged residues of Asp#63 (–201.5 kcal/mol) and Glu#82 (–104.4 kcal/mol) dominate it.

TABLE 7.6
Timings[a] and Performances of FMO-MP2 Job and FMO-RHF Job for Influenza Hemagglutinin Antigen–Antibody System[b] on the Earth Simulator

VPUs	Time (s)	Acceleration Factor[c]	Efficiency (%)[d]	GFLOPS[e]	TFLOPS[f]	Efficiency (%)[g]	Rel. Cost[h]
FMO-MP2							
1024	10,084.6			1.16	1.19	14.5	2.42
2048	5,486.2	1.84	91.9	1.07	2.19	13.3	2.57
3072	3,927.1	2.57	85.6	1.00	3.06	12.4	2.56
4096	3,204.4	3.15	78.7	0.92	3.75	11.5	2.66
FMO-HF							
1024	4,164.9			1.33	1.36	16.6	
2048	2,131.7	1.95	97.7	1.30	2.67	16.3	
3072	1,533.5	2.72	90.5	1.21	3.72	15.2	
4096	1,205.9	3.45	86.3	1.16	4.75	14.5	

[a] Turnaround time for FMO-MP2/6-31G or FMO-RHF/6-31G job in second. NP value was set to 16.
[b] Total numbers of atoms, AO functions and fragments were 14,086 (H atoms 6932), 78,390 and 911, respectively. The fundamental structure of 1EO8 in PDB was used for standard protocols of molecular modeling.
[c] Acceleration due to the increase of VPUs from 1024.
[d] Efficiency of acceleration against the increase of VPUs in percentage.
[e] Observed speed per VPU in average, by a job manger of the Earth Simulator.
[f] Observed gross speed with provided VPUs, by a job manger of the Earth Simulator.
[g] Efficiency of the observed gross speed relative to a theoretical peak speed of provided VPUs in percentage.
[h] Cost factor of MP2 job relative to RHF job.

TABLE 7.7
Sums of IFIE[a,b] between the Residues in Each Antigenic Site (A, B, C, D, and E) in Hemagglutinin HA1 Domain and the Whole Fab-Fragment of Antibody

Site	A	B	C	D	E	HA1 Total[c]
Residues[d]	121–147	155–198	53–57 275–278	200–215	62–65 78–83	1–327
IFIE sum	51.7	–30.6	50.8	63.6	–345.7	–258.9

[a] Energies are in kcal/mol.
[b] MP2 correction was included.
[c] HA2 total was –223.6 kcal/mol.
[d] Sequential number in HA1 domain.

Hydrophobic residues such as Ile#62 and Phe#79, however, contribute also to the gross stabilization, whose IFIE values are –15.1 kcal/mol and –17.7 kcal/mol, respectively, through the MP2 correction. An outstanding importance of site E was extensively discussed in the literature,[51–53] and the present numerical results coincide with these discussions.

Experimentally, Nakajima et al.[52] introduced single-point amino acid changes in the HA1 domain of HA proteins by using PCR-based random mutation or site-directed mutagenesis to clarify the effects of amino acid substitutions on the hemadsorption character of H3-HA. These substitutions were classified as positive or negative according to their effects on the hemadsorption activity. Combined with this hemadsorption experiment, the IFIE analysis in the framework of FMO method could provide useful information concerning a prediction for probable mutation sites in HA proteins. For example, among the ten hydrophobic residues in the antigenic region E, Ile#62 and Phe#79 show significant attractive interactions with the Fab-fragment of antibody, as mentioned above, reflecting a strong capture by the antibody. The Ile62 and Phe79 are positive and negative, respectively, in the hemadsorption experiment, and only the former has mutated in 1977 and 1997 after 1968 in which the X-ray crystal structure of unmutated antigen–antibody complex was taken. Thus, a hypothesis that amino acid residues with both positive hemadsorption activity and strongly attractive interaction with antibody have a tendency to mutate easily would be available.

Based on the FMO-MP2 calculations, a series of investigations on the HA antigen–antibody and receptor-binding systems have been performed, by which probable and important HA mutations may be predicted. Such computational insights could be helpful to prepare the effective influenza vaccines before epidemics or pandemics. If usual in-house PC clusters are employed as the computing platform, the elapsed time to complete the FMO-MP2 job for a single complex would be about a week or more. The utilization of massively parallel-vector computations[50] for the screening should thus be essential for saving time in practical applications.

7.5 CONCLUSION

In this chapter, we illustrated several applications of the FMO method performed by the ABINIT-MP program and visualized by the BioStation Viewer software. Receptor–ligand interaction of estrogen receptor, receptor–cofactor interaction of retinoid X receptor, sequence-specific DNA–protein interaction of cyclic-AMP receptor protein, and antigen–antibody interaction and sialic acid recognition of influenza hemagglutinin were examined. The electronic structure and the binding form of site-specific interactions for such biomacromolecules have been revealed by using FMO-based analyses, IFIE, CAFI, and FILM. It was shown that the quantum mechanical treatment with inclusion of electron correlation effect was essential to obtain an appropriate picture of the interaction. Through these analyses, the FMO method and the ABINIT-MP program system appear as powerful tools with which to understand interaction mechanism of biomacromolecules. More realistic simulations would become available through development of FMO-based methodologies, such as *ab initio* MD (Chapter 6) and order-made force fields based on FMO density,[54,55] in the near future.

ACKNOWLEDGMENTS

The authors thank Akifumi Kato, Yuto Komeiji, Mika Ito, Ikuo Kurisaki, Tatsunori Iwata, Kazutomo Takematsu, Katsuhisa Nakajima, Katsumi Omagari, Takeshi Ishikawa, and Hirofumi Watanabe for their collaborations. The works reported here were supported primarily by the CREST project operated by the Japan Science and Technology Agency (JST) and partially by the "Revolutionary Simulation Software for the 21st Century" (RSS21) project operated by the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

REFERENCES

1. Kitaura, K., Sawai, T., Asada, T., Nakano, T., Uebayasi, M. 1999. Pair interaction molecular orbital method: an approximate computational method for molecular interactions. *Chem. Phys. Lett.* 312: 319–324.
2. Kitaura, K., Ikeo, E., Asada, T., Nakano, T., Uebayasi, M. 1999. Fragment molecular orbital method: an approximate computational method for large molecules. *Chem. Phys. Lett.* 313: 701–706.
3. Nakano, T., Kaminuma, T., Sato, T., Akiyama, Y., Uebayasi, M., Kitaura, K. 2000. Fragment molecular orbital method: application to polypeptides. *Chem. Phys. Lett.* 318: 614–618.
4. Nakano, T., Kaminuma, T., Sato, T. et al. 2002. Fragment molecular orbital method: use of approximate electrostatic potential. *Chem. Phys. Lett.* 351: 475–480.
5. ABINIT-MP and BioStation Viewer are available at www.ciss.iis.u-tokyo.ac.jp/rss21/.
6. Renaud, J.P., Moras, D. 2000. Structural studies on nuclear receptors. *Cell Mol. Life. Sci.* 57: 1748–1769.
7. Steinmetz, A.C.U., Renaud, J.-P., Moras, D. 2001. Binding of ligands and activation of transcription by nuclear receptors. *Annu. Rev. Biophys. Biomol. Struct.* 30: 329–359.
8. Nilsson, S., Kuiper, G., Gustafsson, J.-Å. 1998. ERβ: a novel estrogen receptor offers the potential for new drug development. *Trends Endocrinol. Metab.* 9: 387–395.
9. Sonnenschein, C., Soto, A.M. 1998. An updated review of environmental estrogen and androgen mimics and antagonists. *J. Steroid. Biochem. Molec. Biol.* 65: 143–150.
10. The RCSB Protein Data Bank (www.rcsb.org/). Berman, H.M., Westbrook, J., Feng, Z. et al. 2000. The protein data bank. *Nucleic Acids Research* 28: 235–242.
11. Brzozowski, A.M., Pike, A.C.W., Dauter, Z., et al. 1997. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* 389: 753–758.
12. Oostenbrink, C., van Gunsteren, W.F. 2005. Physical sciences–chemistry–chemical theory and computation special feature. *Proc. Natl. Acad. Sci.* 102: 6750–6754.
13. van Lipzig, M.M., ter Laak, A.M., Jongejan, A. et al. 2004. Prediction of ligand binding affinity and orientation of xenoestrogens to the estrogen receptor by molecular dynamics simulations and the linear interaction energy method. *J. Med. Chem.* 47: 1018–1030.
14. Fukuzawa, K., Kitaura, K., Uebayasi, M., Nakata, K., Kaminuma, T., Nakano, T. 2005. *Ab initio* quantum mechanical study of the binding energies of human estrogen receptor α with its ligands: an application of fragment molecular orbital method. *J. Comp. Chem.* 26: 1–10.

15. Fukuzawa, K., Mochizuki, Y., Tanaka, S., Kitaura, K., Nakano, T. 2006. Molecular interactions between estrogen receptor and its ligand studied by the *ab initio* fragment molecular orbital method. *J. Phys. Chem. B* 110: 16102–16110: ibid. *J. Phys. Chem. B* 110: 24276.
16. Ito, M., Fukuzawa, K., Mochizuki, Y., Nakano, T., Tanaka, S. 2007. *Ab initio* fragment molecular orbital study of molecular interactions between liganded retinoid X receptor and its coactivator: roles of helix12 in the coactivator binding mechanism. *J. Phys. Chem. B* 111: 3525–3533.
17. Ito, M., Fukuzawa, K., Mochizuki, Y., Nakano, T., Tanaka, S. 2008. *Ab initio* fragment molecular orbital study of molecular interactions between liganded retinoid X receptor and its coactivator. Part II: influence of mutations in transcriptional activation function 2 activating domain core on the molecular interactions. *J. Phys. Chem. A* 112: 1986–1998.
18. Ito, M., Fukuzawa, K., Ishikawa, T., Mochizuki, Y., Nakano, T., Tanaka, S. 2008. *Ab initio* fragment molecular orbital study of molecular interactions in liganded retinoid X receptor: specification of residues with roles of ligand inducible information transmission. *J. Phys. Chem. B* 112: 12081–12094.
19. Shiau, A.K., Barstad, D., Loria, P.M., et al. 1998. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* 95: 927–937.
20. InsightII Version 98.0, Molecular Simulations Inc., San Diego, CA, 1998.
21. CHARMm: Chemistry at HARvard Macromolecular Mechanics (CHARMm), Version 25.2, Revision: 98.0731.; Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M. 1983. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.* 4: 187–217; MacKerell, Jr., A.D., Brooks, B., Brooks III, C.L. et al. 1998. *The Encyclopedia of Computational Chemistry*, eds. P.V.R. Schleyer et al. 1: 271–277. John Wiley & Sons: Chichester.
22. Mochizuki, Y., Nakano, T., Koikegami, S. et al. 2004. A parallelized integral-direct second-order Møller–Plesset perturbation theory method with a fragment molecular orbital scheme. *Theor. Chem. Acc.* 112: 442–452.
23. Mochizuki, Y., Koikegami, S., Nakano, T., Amari, S., Kitaura, K. 2004. Large scale MP2 calculations with fragment molecular orbital scheme. *Chem. Phys. Lett.* 396: 473–479.
24. Mochizuki, Y. 2005. A size-extensive modification of super-CI for orbital relaxation. *Chem. Phys. Lett.* 410: 165–171.
25. Mochizuki, Y., Fukuzawa, K., Kato, A., Tanala, S., Kitaura, K., Nakano, T. 2005. A configuration analysis for fragment interaction. *Chem. Phys. Lett.* 410: 247–253.
26. Ishikawa, T., Mochizuki, Y., Amari, S. et al. 2007. Fragment interaction analysis based on local MP2. *Theor. Chem. Account.* 118: 937–945.
27. Mangelsdorf, D.J., Thummel, C., Beato, M. et al. 1995. The nuclear receptor superfamily: the second decade. *Cell* 83: 835–839.
28. Ross, S.A., McCaffery, P.J., Drager, U.C., Luca, L.M.D. 2000. Retinoids in embryonal development. *Physiol. Rev.* 80: 1021–1054.
29. Miles, S.A., Dezube, B.J., Lee, J.Y. et al. 2002. Antitumor activity of oral 9-cis-retinoic acid in HIV-associated Kaposi's sarcoma. *AIDS* 16: 421–429.
30. Muller-Hill, B. 1996. *The lac OPERON*. Walter de Gruyter & Co: Berlin.
31. Branden, C., Tooze, J. 1999. *Introduction to Protein Structure,* 2nd ed. Garland: London.
32. Brown, T.A. 1999. *Genomes*; BIOS Scientific: London.

33. Gunasekera, A., Ebright, Y.W., Ebright, R.H. 1992. DNA sequence determinants for binding of the *Escherichia coli* catabolite gene activator protein. *J Biol. Chem.* 267: 14713–14720.
34. Fukuzawa, K., Komeiji, Y., Mochizuki, Y., Kato, A., Nakano, T., Tanaka, S. 2006. Intra- and inter-molecular interactions between cyclic-AMP receptor protein and DNA: *ab initio* fragment molecular orbital study. *J. Comp. Chem.* 27: 948-960: ibid. *J. Comp. Chem.* 28: 2237–2239.
35. Chen, S., Vojtechovsky, J., Parkinson, G.N., Ebright, R.H., Berman, H.M. 2001. Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: DNA binding specificity based on energetics of DNA kinking. *J Mol. Biol.* 314: 63–74.
36. Komeiji, Y., Uebayasi, M., Takata, R., Shimizu, A., Itsukashi, K., Taiji, M. 1997. Fast and accurate molecular dynamics simulation of a protein using a special-purpose computer. *J Comp. Chem.* 18: 1546–1563.
37. Komeiji, Y., Haraguchi, M., Nagashima, U. 2001, Parallel molecular dynamics simulation of a protein. *Parallel Comput.* 27: 977–987.
38. PEACH ver. 5.0, AIST, http://staff.aist.go.jp/y-komeiji/.
39. Cornell, W.D., Cieplak, P., Bayly, C.I. et al. 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am. Chem. Soc.* 117: 5179–5197.
40. Kurisaki, I., Fukuzawa, K., Komeiji, Y. et al. 2007. Visualization analysis of inter-fragment interaction energies of CRP-cAMP-DNA complex based on the fragment molecular orbital method. *Biophys. Chemist.* 130: 1–9.
41. Suzuki, Y., Ito, T., Suzuki, T. et al. 2000. Sialic acid species as a determinant of the host range of influenza A viruses. *J. Virol.* 74: 11825–11831.
42. Glaser, L., Stevens, J., Zamarin, D. et al. 2005. A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity. *J. Virol.* 79: 11533–11536.
43. Eisen, M.B., Sabesan, S., Skehel, J.J., Wiley, D.C. 1997. Binding of the influenza A virus to cell-surface receptors: structures of five hemagglutinin–sialyloligosaccharide complexes determined by X-ray crystallography. *Virology* 232: 19–31.
44. Ha, Y., Stevens, D.J., Skehel, J.J., Wiley, D.C. 2003. X-ray structure of the hemagglutinin of a potential H3 avian progenitor of the 1968 Hong Kong pandemic influenza virus. *Virology* 309: 209–218.
45. Iwata, T., Fukuzawa, K., Nakajima, K. et al. 2008. Theoretical analysis of binding specificity of influenza viral hemagglutinin to avian and human receptors based on the fragment molecular orbital method. *Comput. Bio. Chem.* 32: 198–211.
46. Li, M., Wang, B. 2006. Computational studies of H5N1 hemagglutinin binding with SA-α-2, 3-Gal and SA-α-2, 6-Gal. *Biochem. Biophys. Res. Commun.* 347: 662–668.
47. Sawada, T., Hashimoto, T., Nakano, H., Suzuki, T., Ishida, H., Kiso, M. 2006. Why does avian influenza A virus hemagglutinin bind to avian receptor stronger than to human receptor? *Ab initio* fragment molecular orbital studies. *Biochem. Biophys. Res. Commun.* 351: 40–43.
48. Sawada, T., Hashimoto, T., Nakano, H. et al. 2007. Influenza viral hemagglutinin complicated shape is advantageous to its binding affinity for sialosaccharide receptor. *Biochem. Biophys. Res. Commun.* 355: 6–9.
49. Sawada, T., Hashimoto, T., Tokiwa, H. et al. 2008. *Ab initio* fragment molecular orbital studies of influenza virus hemagglutinin–sialosaccharide complexes toward chemical clarification about the virus host range determination. *Glycoconj. J.* 25: 805–815.
50. Mochizuki, Y., Yamashita, K., Murase, T. et al. 2008. Large scale FMO-MP2 calculations on a massively parallel-vector computer. *Chem. Phys. Lett.* 457: 396–403.

51. Fleury, D., Daniels, R.S., Skehel, J.J., Knossow, M., Bizebard, T. 2000. Structural evidence for recognition of a single epitope by two distinct antibodies. *Proteins* 40: 572–578.
52. Nakajima, K., Nobusawa, E., Nagy, A., Nakajima, S. 2005. Accumulation of amino acid substitutions promotes irreversible structural changes in the hemagglutinin of human influenza AH3 virus during evolution. *J. Virol.* 79: 6472–6477.
53. Nakajima, S., Nakajima, K., Nobusawa, E., Zhao, J., Tanaka, S., Fukuzawa, K. 2007. Comparison of epitope structures of H3HAs through protein modeling of influenza A virus hemagglutinin: mechanism for selection of antigenic variants in the presence of a monoclonal antibody. *Microbiol. Immunol.* 51: 1179–1187.
54. Fischer, B., Fukuzawa, K. Wenzel, W. 2008. Receptor specific scoring functions derived from quantum chemical models improve affinity estimates for *in-silico* drug discovery. *Proteins: Structure, Function and Bioinformatics*, 70: 1264–1273.
55. Okiyama, Y., Watanabe, H., Fukuzawa, K. et al. 2007. Application of the fragment molecular orbital method for determination of atomic charges on polypeptides. *Chem. Phys. Lett.* 449: 329–335.

8 Detailed Electronic Structure Studies Revealing the Nature of Protein–Ligand Binding

Isao Nakanishi, Dmitri G. Fedorov, and Kazuo Kitaura

CONTENTS

8.1 INTRODUCTION

In the early 1990s, several successful structure-based drug designs (SBDDs) were reported,[1–4] and this approach has become common recently, implying that rational drug design based on the physical theory can compete with the conventional trial-and-error drug discovery. For example, interaction analysis is often used to identify

which amino acids are significant for ligand recognition and to obtain ideas for ligand modification. Docking simulations including the virtual in silico screening are increasingly used to find novel seed or lead compounds, because the number of real compounds suitable for the high-throughput screening is growing, and the screening costs rise whereas the hit rates decrease. Protein modeling is essential if experimental protein structures are unavailable. Affinity prediction is one of most important but the most difficult research areas. There are many methods for affinity prediction, for example, quantitative structure-activity relationship (QSAR) or more rigorous simulation-based methods such as the free energy perturbation (FEP)[5] and thermodynamic integration (TI)[6] methods.

Most of these computations rely on force fields, as large proteins can be treated within a reasonable computational time. There are some limitations of force fields: few potential parameters are established for synthetic ligands and complicated natural compounds. In addition, noncanonical interaction such as CH/π and a CH···O hydrogen bond cannot be described properly. Fixed charges are another serious problem because polarization and charge transfers are important factors for molecular interaction. A promising method for overcoming the limitations of force fields is the mixed quantum mechanical and molecular mechanical (QM/MM) method,[7–11] in which ligand atoms and possibly some protein atoms are treated by the QM method.

For a binding affinity calculation in the physiological condition, on the other hand, it is necessary to take into account the following: (1) the binding energy in the gas phase, (2) solvation energies of the complex components, (3) molecular motion, and (4) entropic effect on the complex formation. The binding energy should be calculated accurately considering the polarization and charge transfer energies in addition to the electrostatic and van der Waals energies. QM methods, which can be applied to the whole protein–ligand complex, are suitable for performing such calculations.

The desolvation energies on complex formation are frequently calculated using the implicit water models such as the Poisson-Boltzmann (PB) method,[12] the generalized Born (GB) method,[13] and the polarizable continuum model (PCM),[14] because the explicit treatment of water molecules, which requires a proper configurational sampling, results in a very time-consuming process. The PB or GB models in conjunction with the surface area (SA) method for the calculation of the nonpolar contribution of the solvation energy are often used with force fields,[15] and PCM is employed largely with quantum-mechanical methods. To take into account molecular motion, the statistical treatment using molecular dynamics (MD) or Monte Carlo (MC) simulation is necessary. Especially, if the systems under study are flexible and affect large conformational changes, the molecular fluctuations cannot be ignored. The entropic effect is very difficult to evaluate. There are modes to be considered on the complex formation such as translation, rotational, and vibration. Because protein molecules have a huge number of degrees of freedom, calculations of these changes with accuracy can be extremely difficult.

In this chapter, the fragment molecular orbital (FMO) method was applied to the binding energy calculation in order to analyze in detail the protein–ligand interaction energies. We stress the significant contribution of the dispersion and

polarization in the binding energy. In addition, the solvation energy is evaluated by the PB/SA method, and the binding free energy calculation was carried out combining gas phase FMO and PB/SA results. As a target protein–ligand complex system, we chose FK506 binding protein (FKBP) complexes with four high-affinity ligands. FKBP is a peptidyl-prolyl cis/trans isomerase (PPIase) consisting of 107 amino acids, and it is well known as an immunophilin that binds immunosuppressant FK506 (tacrolimus) or other ligand molecules at the substrate binding site for PPIase activity. Because the binding of FKBP to ligands has attracted much attention, a large amount of experimental data is available.[16–22] Structural studies[23,24] show a variety of interactions in the FKBP–ligand complexes, including a weak CH···O hydrogen bond, usual hydrogen bond, and van der Waals interaction. Therefore, FKBP complexes are suitable targets for a demonstrative analysis of protein–ligand interactions.

In this study, we performed FMO calculations of FKBP–ligand complexes using four ligands with distinct structures to obtain the binding energy. The residue–ligand interactions are discussed in detail. The solvent effect was estimated using the PB/SA solvation model for obtaining binding affinities in the condensed phase. We applied the normal mode analysis by the force field method for the calculation of the vibration free energy (including the entropic contribution).

8.2 MODELING OF PROTEIN–LIGAND COMPLEXES

8.2.1 Addition of Missing Hydrogen Atoms and Checkpoints of Experimental Structure Data

Crystal structures of FKBP complexes with four ligands were obtained from the Protein Data Bank (PDB) and used as the initial data in the following modeling procedure. The PDB codes of the four complexes are 1fkb, 1fkf, 1fkg, and 1fki. The chemical structures of the ligand molecules calculated are shown in Figure 8.1. Hereafter, the ligands are referred to by their PDB codes for convenience. Because the experimental structure of the 1fki complex contains two complex monomers in an asymmetric unit, the monomer with the lower temperature factor averaged over all atoms in a monomer was chosen for the calculations. All water molecules found in crystals were removed because they were thought not to mediate interactions between the protein and ligands. In general, water molecules having a significant effect on molecular interactions should be kept.

Because X-ray crystal structures do not include hydrogen atoms, they were added to both the protein and ligands. The addition of hydrogen atoms to proteins can be performed using the FMOUTIL program, which comes with the GAMESS software package,[25] or molecular modeling software for biomolecules. The positions of the hydrogen atoms involved in the hydrogen bond were oriented in the proper direction manually before optimization, because they are usually put by modeling software without properly considering local interactions. All acidic and basic amino acid residues as well as the N- and C-terminus are treated in their standard charged states. It should be noted that amino acid residues inside the protein sometimes exist in the neutral state. The charged states of these amino

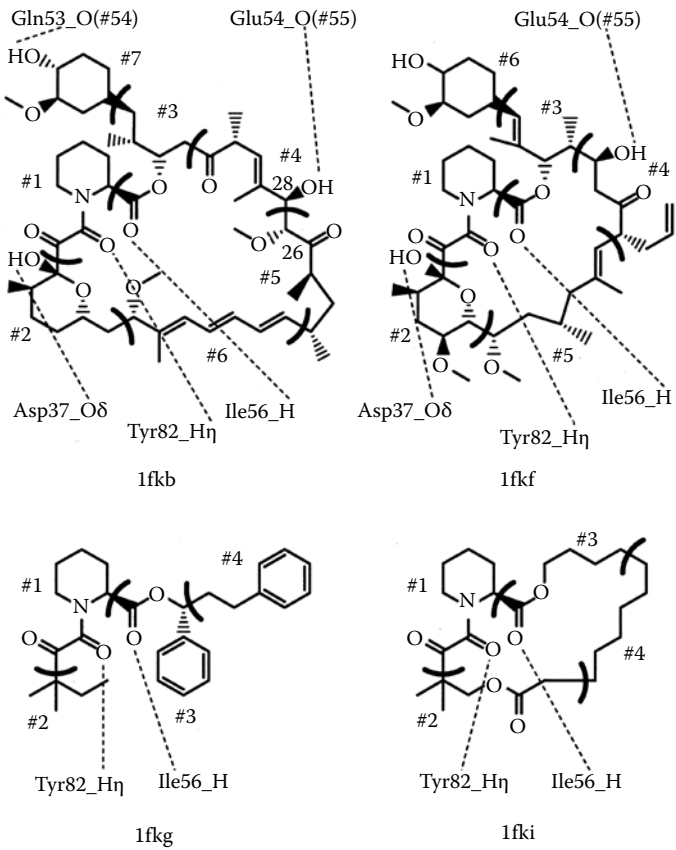


FIGURE 8.1 Chemical structures of the four FK506 binding protein (FKBP) ligands, denoted by their Protein Data Bank (PDB) codes. The fragmentation locations are shown with the arcs, and the fragment numbers of each ligand molecule are preceded by #. Atom numbers referred in the text are indicated. The hydrogen bonds between the ligands and FKBP are shown by the dashed lines together with the residue atoms (e.g., Gln53_O) in the FKBP complex. The corresponding residue fragment numbers of the hydrogen bonding sites are given in parentheses (#54 in this case).

acid residues are deduced from the overall conditions, such as the location, pH for crystallization, hydrogen bonding network with surrounding amino acid residues, and reaction mechanism. All histidine residues are also protonated because the crystal structure refinements were performed under protonated conditions, except for the 1fkf complex.

The modeling of histidine, asparagine, and glutamine residues requires special attention because the X-ray crystal structure analysis of proteins does not distinguish atomic species among C, N, and O. For example, labeling the N and O atoms of the carbamoyl moiety in the Asn and Gln residues is often ambiguous. Also, the imidazole ring orientation of histidine residues, and hence their tautomer states, are

difficult to determine. They are deduced from the hydrogen bonding network with adjacent moieties. Several software programs such as SYBYL[26] and WHATIF[27] can be used for the modeling process to determine the most favorable hydrogen bonding network automatically.

8.2.2 Structural Refinements by Geometry Optimization

Although resolutions of most protein crystal structure analyses are at the atomic level, the coordinates contain certain errors. Therefore, optimizing the structures is recommended, especially if the resolution is worse than 2.0 Å. Structural optimization by the quantum-mechanical method is desirable; however, such optimization of the entire protein structure is very time consuming, and thus, a truncated model is usually used. Amino acid residues within a 2.0 distance from the ligand were extracted from the complex structures to be used as a truncated model of the ligand-binding site. This unitless distance is defined as follows: the distance is relative to the van der Waals radii, for example, two atoms, A and B, separated by R have a distance equal to $R/(R_A + R_B)$, where R_A and R_B are the van der Waals radii of atoms A and B, respectively. The distance of 2.0 typically corresponds to 5 to 5.5 Å. If the total charges of the truncated models were not zero, amino acid residues with opposite charges located closest to the ligands were added until the systems became neutral. This neutralization is not always necessary, but it may help reduce the number of steps involved in geometry optimization.

The number of residues in the truncated models was 22, 18, 22, and 16 for 1fkb, 1fkf, 1fkg, and 1fki, respectively. Both ends of the extracted amino acid fragments were capped with hydrogen atoms, and their positions were optimized by the MM method. Note that the amino acid residues included in this small model are not suitable for optimization because they are located in the first layer around the ligand and miss important interactions with the remaining part of the protein. In order to optimize residues around the ligand, a larger model (e.g., ~8 Å from the ligand) should be extracted.

The structure of such model complexes was optimized by FMO. All ligand molecules and hydrogen atoms of FKBP involved in the hydrogen bonds with ligand molecules are optimized, and other protein atoms are frozen during optimization. The optimizations were performed by GAMESS at the FMO2-RHF/3-21G level using one residue per fragment division until the maximum gradient was less than 0.0001 Hartree/Bohr. After optimization, the capping hydrogen atoms were removed from the optimized model, and the atoms, whose positions were optimized, were put into the full complex.

The sizes of the model complexes (the number of atoms and basis functions) are (572, 3062), (471, 2503), (492, 2643), and (388, 2057) for 1fkb, 1fkf, 1fkg, and 1fki, respectively. All calculations were performed using the Gion cluster. It consisted of 40 nodes each equipped with dual single-core Xeon processors (3.0 GHz) and 2 GB memory, and the nodes were connected via a Gigabit Ethernet network. The details of the calculations are shown in Section 8.3.1. Single-point energy and gradient calculations took about 30 minutes using 20 CPUs in parallel, and the geometry optimization was completed in 200 to 350 steps (4 to 7 days).

8.2.3 Optimized Structure and Distortion Due to the Packing Effect in Crystal

Figure 8.2a shows the conformational changes of the four ligands before and after energy minimization, in addition to the root mean square (RMS) deviation values. Except for the 1fkg ligand, the RMS deviations of nonhydrogen atoms were within 0.5 Å. The diketopipecolate moieties, a common structural element to all ligands, showed very small deviations (~0.3 Å), because they are tightly bound to the bottom of the binding pocket forming hydrogen bonds and have strong van der Waals interactions. On the other hand, conformational changes are relatively large at the entrance of the binding pocket, especially for the 1fkg ligand with a noncyclic framework, which showed 0.8 Å RMS deviation due to the reorientation of the 1,3-diphenylpropyl group. Because this flexible region interacts with adjacent molecules in the crystalline state, the deviation becomes large when optimization is performed without packing constraints. In fact, when the replicated adjacent protein fragments are incorporated in the truncated model, the RMS displacement of the 1fkg ligand in the new model decreased radically to 0.45 Å after optimization (Figure 8.2b). There are no improper geometric parameters, and interactions observed in the optimized structures indicate that the FMO-optimized structures of the FKBP–ligand complexes in vacuo are reasonable.

However, it should be noted that the optimizations were performed at the Hartree–Fock level, and this study does not consider the electron correlation, which is needed to describe the van der Waals interaction. It is recommended that the geometry optimization is performed considering the electron correlation, because the dispersion is sometimes one of the main determining factors for the complex structure. In the case of the FKBP complexes, however, the RHF level optimizations are expected to produce reasonable complex structures, because there are two to five hydrogen bonds between ligands and FKBP, which is a predominant factor of the binding mode determination. One can also consider geometry optimization by FMO at the MP2 level (which describes the dispersion).

8.3 PROTEIN–LIGAND INTERACTIONS

8.3.1 FMO Calculations and Pair Interaction Analysis

In the FMO method,[28,29] the system is divided into N fragments, and the total energy E of the entire system is calculated using the following series expansion (truncated here at the two-body terms; the FMO2 method):

$$E = \sum_{I}^{N} E_I + \sum_{I>J}^{N} (E_{IJ} - E_I - E_J) \tag{8.1}$$

E_I and E_{IJ} are monomer and dimer energies, respectively, and are computed by *ab initio* RHF (or another wavefunction) considering the external electrostatic potential (ESP), which describes all remaining fragments (other than those composing monomer I or dimer IJ).

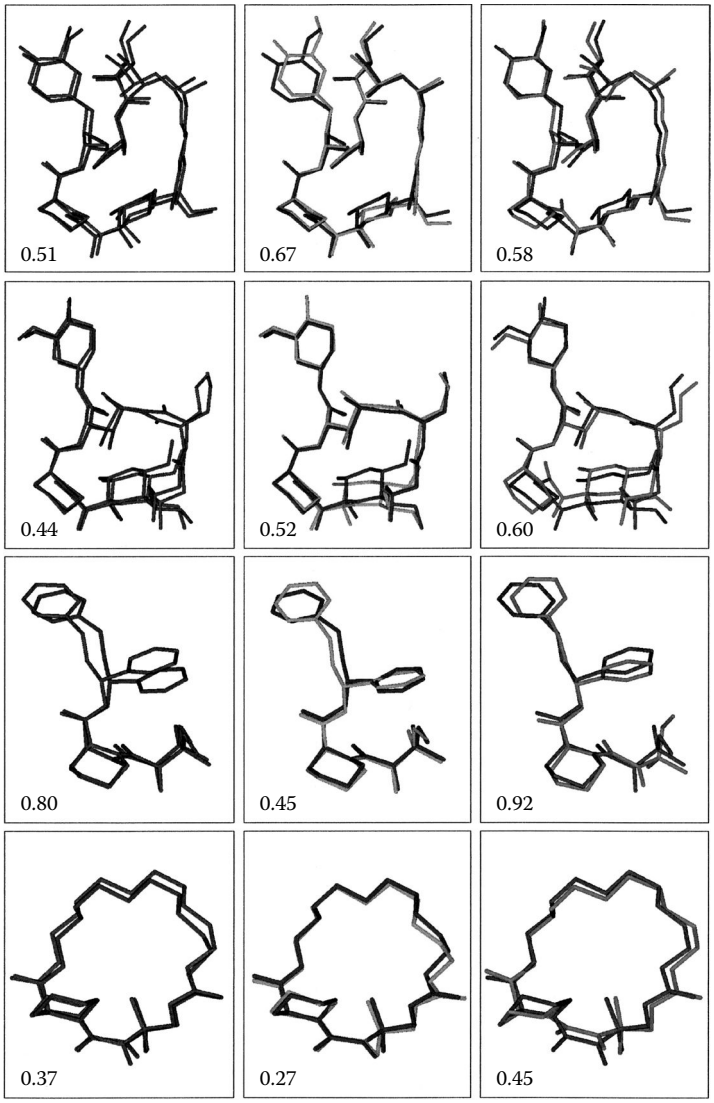


FIGURE 8.2 Comparison of the optimized ligand structures (hydrogen atoms are omitted for clarity). (a) PDB structures (red) and optimized structures in the complex (blue), (b) PDB structures (red) and optimized structures in the crystal environment (green), and (c) optimized structures in the complex (blue) and those in isolated state (magenta). The numerical values indicate the root mean square deviation (RMSD) between the two geometries (in Å). Each column is in the order of 1fkb, 1fkf, 1fkg, and 1fki ligands from the top. **(See color insert following page 117.)**

Equation 8.1 is usually expressed as the following form:

$$E = \sum_{I}^{N} E'_I + \sum_{I>J}^{N} \{(E'_{IJ} - E'_I - E'_J) + Tr(\Delta \mathbf{D}^{IJ} \mathbf{V}^{IJ})\}$$
$$= \sum_{I}^{N} E'_I + \sum_{I>J}^{N} \Delta \tilde{E}_{IJ} \tag{8.2}$$

where E'_I and E'_{IJ} are the internal monomer and dimer energies, respectively. They are obtained from monomer and dimer energies, E_I and E_{IJ}, by subtracting the electrostatic interaction energy with the surrounding monomers. They represent monomers and dimers polarized by the environment. $\Delta \mathbf{D}^{IJ}$ is the difference of density matrices of dimer IJ and monomers I and J, and V^{IJ} is the ESP for dimer IJ. $\Delta \tilde{E}_{IJ}$ represents the *pair interaction energy* (PIE) between fragments. PIEs are used for the analysis of interactions between protein residues and the ligand molecule.

We employed the FMO2 method throughout in this study and used PIE to analyze intermolecular interactions based on divided fragments. In FMO calculations, a protein was divided between the Cα atom and the adjacent carbonyl carbon atom of the main chain. Ligand molecules were partitioned at sp^3 carbon atoms such that each fragment contained about 10 nonhydrogen atoms, as shown in Figure 8.1.

GAMESS was used for FMO calculations, see Nakanishi et al.[30] for full computational details. A single point FMO2-MP2/6-31G* calculation of the entire FKBP complex (1,734 to 1,800 atoms and 13,844 to 14,400 basis functions) took about 1.5 and 12 days using 40 CPUs of the Gion cluster for one and two residues per fragment division, respectively.

8.3.2 Binding Energy

The binding energy ΔE_b between the protein and ligand molecule is calculated as follows:

$$\Delta E_b = E^C - E^P - E^L = \Delta E_{int} + \Delta E^L_{def} \tag{8.3}$$

$$\Delta E_{int} = E^C - E^P - E^{L(C)} \tag{8.4}$$

$$\Delta E^L_{def} = E^{L(C)} - E^L \tag{8.5}$$

where E^C, E^P, and E^L are the total energy of the complex, and the isolated states of protein and ligand, respectively. The internal energy of the ligand within the complex is denoted by $E^{L(C)}$. ΔE_{int} is the interaction energy between the protein and the ligand, and ΔE^L_{def} is the destabilizing deformation energy due to the geometric distortion of the ligand upon complexation. The negative energy means stabilization, and the energetics is discussed below at the FMO2-MP2/6-31G* level unless otherwise stated.

The structural and energetic aspects of the ligand deformation provide much insight for predictions of the binding affinities or the design of new drug molecules. The deformation energies ΔE^L_{def} of the ligands were calculated in vacuo. The

conformational changes and deformation energies of the ligand molecules are shown in Figure 8.2c and Table 8.1, respectively. The conformational changes of the diketopipecolate moieties were larger compared with other regions in the gas phase. The larger conformational change of the 1fkg ligand is derived from the torsion angle change for the diketo moiety. The torsion angle between two ketone groups, which tends to form a conjugated system giving planer conformation in the isolated state, is distorted to almost 90° in the complex. Consequently, the relative orientation of the tertialy pentyl group to the rest of the molecule was largely different in the complex. As the other three ligands have cyclic frameworks, such bond rotations were highly restricted. Accordingly, the RMSD values were small. The slightly larger deformation energy of the 1fkg ligand might be derived from the loss of the conjugation energy of the diketone moiety.

Considering the large deformation energy of Rapamycin (the 1fkb ligand), it can be deduced that an energetically unfavorable conformational change occurs on complexation with the protein. Rapamycin forms an intramolecular hydrogen bond between the 28-hydroxyl and 26-carbonyl groups in the free state. In the complex state, however, this hydrogen bond is eliminated by switching the hydrogen-bonding pair of the 28-hydroxyl group with a protein atom (the oxygen atom of the main chain carbonyl of Glu54). This means that the molecule has to lose the intramolecular hydrogen bond in the isolated state to form a complex with FKBP. This energy loss results in a large deformation energy of the molecule. Note that our recent study on structural optimization using the PCM solvation method[14] showed that this intramolecular hydrogen bond is not formed in water.

Next, molecular interactions are quantitatively described. The interaction energies between the ligand molecules and FKBP calculated at the FMO-MP2/6-31G* level are shown in Table 8.1. For the reference of the basis set dependence, the interaction energies were also calculated at the FMO-RHF/STO-3G and FMO-RHF/3-21G levels. In the FMO calculation, the fragment partitioning has some effect upon the accuracy. Two partition schemes were applied to the energy calculations: one with each amino acid residue put to one fragment (1res/frg), and the other with two consecutive amino acids put to one fragment (2res/frg). In general, a division with fewer fragments has a smaller error. Energy differences between the two partition schemes tend to increase as a larger basis set is used. The total interaction energies with the 2res/frg partition were 2.0 ~ 4.3 kcal/mol lower than for 1 res/frg.[30] Because this difference is not negligible, the energies from the 2 res/frg calculation are used in the following discussion.

The interaction energies that include the electron correlation effect are in the range of −77.5 to −120.6 kcal/mol. They resemble the experimental order of the binding affinities for the four ligands. But there is a large energy gap of ~30 kcal/mol between the high- (1fkb and 1fkf) and low-affinity groups (1fkg and 1fki). The reason behind the order of the binding affinities may be seen in the number of hydrogen bonds between the ligands and FKBP correlates with the binding energies. The pair interaction analysis in the next section explains this point clearly.

The most notable point about the FKBP–ligand interactions is the contribution of the electron correlation energies (about 60% to 70% of the total interaction), indicating that the dispersion interaction contributes considerably to the protein–ligand interactions

TABLE 8.1
The Binding ΔE_b, Ligand Deformation ΔE^L_{def}, and Interaction ΔE_{int} Energies (in kcal/mol) between the Ligand and FKBP and the Charge $\Delta Q^{P\to L}$ (in a.u.) Transferred from Protein to Ligand

		ΔE_b			ΔE^L_{def}			ΔE_{int}			
System	Level of Theory	Total	RHF	Disp[a]	Total	RHF	Disp[a]	Total	RHF	Disp[a]	$\Delta Q^{P\to L}$
1fkb	MP2/6-31G*	–103.9	–21.9	–82.0	16.7	21.0	–4.3	–120.6	–42.9	–77.7	–0.106
	RHF/3-21G		–76.3			25.4			–101.7		
	RHF/STO-3G		–15.8			21.4			–37.2		
1fkf	MP2/6-31G*	–102.2	–33.0	–69.2	8.2	7.8	0.4	–110.4	–40.8	–69.6	–0.073
	RHF/3-21G		–76.9			9.7			–86.6		
	RHF/STO-3G		–21.7			12.9			–34.6		
1fkg	MP2/6-31G*	–70.1	–12.4	–57.7	10.0	12.5	–2.5	–80.1	–24.9	–55.2	–0.015
	RHF/3-21G		–41.8			17.6			–59.4		
	RHF/STO-3G		7.5			18.6			–11.1		
1fki	MP2/6-31G*	–71.3	–16.0	–55.3	6.2	7.9	–1.7	–77.5	–23.9	–53.6	0.003
	RHF/3-21G		–46.5			9.7			–56.2		
	RHF/STO-3G		3.8			15.4			–11.6		

[a] MP2 correlation contribution.

in these systems. The dispersion energies are usually considered to increase proportionally with the contact area of the van der Waals surfaces. Therefore, it is necessary for large-scale molecular systems such as protein–protein and protein–ligand complexes to incorporate the electron correlation effects for evaluating intermolecular interaction energies properly by means of an *ab initio*–based MO calculation.

Finally, we show the correlation between the interaction energy and protein–ligand charge transfer in Table 8.1. The total charge of the ligand in the complex was calculated from Mulliken atomic populations. Except for the 1fki complex, the direction of the charge transfer is from the FKBP to the ligand (e.g., –0.106 corresponds to the charge of 0.106 electrons transferred). A negligible amount of charge transfer was found for the 1fki ligand. The charges are transferred mostly at the hydrogen-bonding sites. Interestingly, there is a high correlation between the amount of charge transfer and the protein–ligand interaction energies, with a correlation coefficient of 0.99. Such a correlation was also reported by Fukuzawa et al. for the FMO calculation of the estrogen receptor with its ligands.[31]

Before proceeding to the next section, the basis set dependence of the interaction energies is mentioned briefly. The energy values calculated by each basis set vary considerably. One of the reasons is due to no correction of the basis set superposition error (BSSE). The BSSE correction is a problem in the *ab initio*–based MO calculations. Particularly, in case of smaller basis sets such as RHF/STO-3G and RHF/3-21G, the error is considerable. However, the energy values of RHF/STO-3G (3-21G also) are closely related with those of MP2/6-31G* (for ΔE_{int} the correlation coefficients are larger than 0.97). Therefore, a lower theoretical level calculation without the BSSE correction could be used to compare just the tendencies of the interaction magnitude within a shorter calculation time. However, note that the RHF/STO-3G calculation gave positive (i.e., repulsive) binding energies for 1fkg and 1fki that have less hydrogen bonds than 1fkb and 1fkf.

8.3.3 Interaction Energy Decomposition

Figure 8.3 shows the pair interaction energies (PIEs) between the ligand molecule and each amino acid residue of FKBP for every complex. There are some strongly attractive interactions less than –10 kcal/mol, corresponding to interactions of the fragment pair that involve hydrogen bonds. Also, several moderately attractive interactions less than –5 kcal/mol are seen. These interactions are all from the fragments located within 5 Å of the ligand. In the geometric optimization of the complex, the truncated models are composed of about 20 amino acid residues located within 2.0 distances (~5 Å) from the ligands, and contain about 20% of all residues, because FKBP consists of 107 amino acid residues. However, the energy contributions from the truncated models account for 61% to 77% of the total interaction energies. Thus, 20% of the fragments are responsible for more than 60% of the total interaction energies. Therefore, we conclude that the cutoff criterion used for model building is a good balance of computation time and accuracy.

Other fragments beyond 5 Å from the ligands contribute less than ±5 kcal/mol to the binding energy. However, several of these interactions are not negligible, and most of these fragments are charged amino acid residues. When these energies are

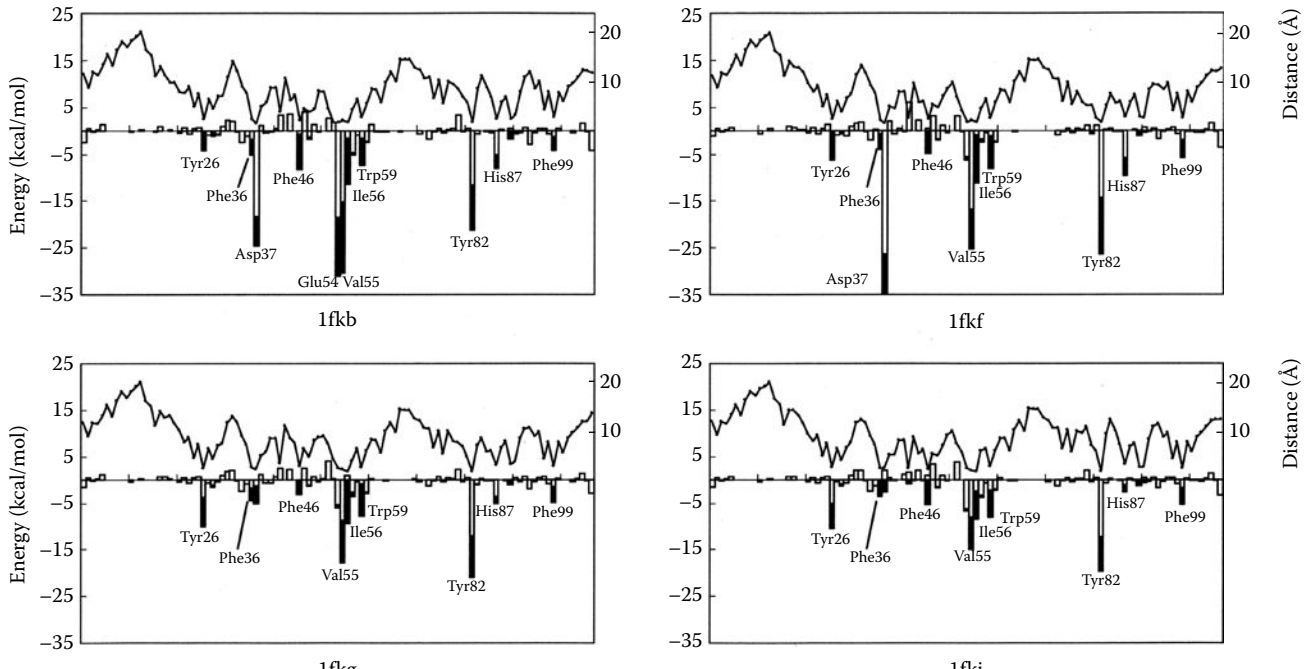


FIGURE 8.3 Diagrams of the pair interaction energies (PIEs) between ligands and each residue fragment of the protein (# is omitted in fragment labels) calculated by the FMO-MP2/6-31G* basis set. The white and black bars designate the RHF and correlated energy contribution, respectively. The distance between ligand and each residue fragment is shown with solid lines. The horizontal axis is the sequence number of the residue fragment. The left and right vertical axes indicate energies in kcal/mol and distances in Å, respectively.

added to the interaction energy in the truncated model, the ratio of the total interaction energies reaches 85% to 90%.

An interesting relationship was observed regarding the locations of these charged fragments and the interaction energies; if the acidic (basic) residue fragment is located on the same (opposite) side of the three carbonyl groups of the diketopipecolate moiety, the interactions are attractive. On the contrary, reverse positioning results in repulsive interactions. As these carbonyl groups face toward the outside of the ligand molecules in the bound states forming one terminal of the molecule, it is suggested that the interactions are between charged fragments and a molecular dipole, where the diketopipecolate group is regarded as the negative terminal and the opposite side of the molecule is the positive terminal.

There are two to five hydrogen bonds between the ligands and FKBP in the complexes. A quantitative discussion on the hydrogen bond energies and their geometries, especially bonding distances, is important for the analysis of the protein–ligand interactions. Note that residue fragments do not exactly correspond to amino acid residues, because the fragment partitioning in FMO calculations is usually performed between α carbon and the carbonyl carbon atoms of the main chain, and the main chain carbonyl group of the i-th residue is assigned to the $(i + 1)$-th residue fragment. In the following discussion, the residue fragment is referred to by its residue name, and the fragment number with #, for example Tyr#82, to distinguish it from the conventional notation Tyr82.

The structures of the diketopiperidino moieties (Lig#1) and the Phe36 and Tyr82 residues are shown in Figure 8.4, and the PIEs between the fragment Tyr#82 and

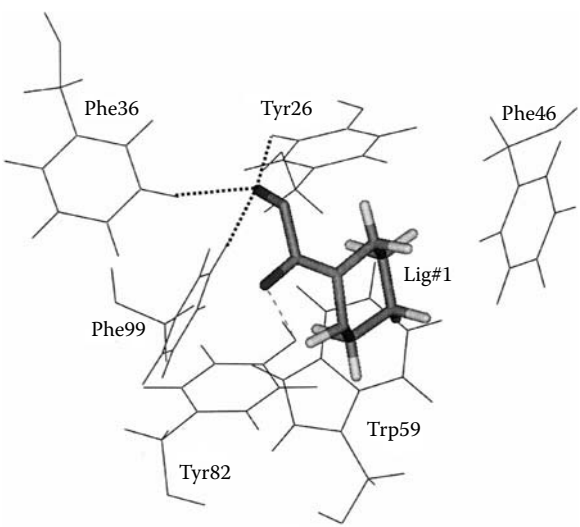


FIGURE 8.4 Orientation of aromatic amino acid residues around the diketopiperidino group (Lig#1) of the 1fkf complex. Lig#1 is shown as thick sticks. Only side-chain atoms are drawn for amino acid residues. Three CH···O and one hydrogen bond are drawn as dotted and dashed lines, respectively.

TABLE 8.2
Pair Interaction Energies $\Delta\tilde{E}^{C}$, Their Components[a] in kcal/mol and the Bonding Distances R in Å for Hydrogen and CH···O Bonds (FMO-MP2/6-31G*)

System	$\Delta\tilde{E}^{C}$	$\Delta\tilde{E}^{C}_{disp}$	$\Delta\tilde{E}^{C}_{ele}$	$\Delta\tilde{E}^{C}_{non-ele}$	*R*(O-H)
(1) OH…O hydrogen bond between Trp#82 and Lig#1					
1fkb	–20.7	–4.6	–23.4	7.3	1.73
1fkf	–20.1	–4.5	–20.9	5.3	1.78
1fkg	–19.6	–4.5	–23.0	7.9	1.74
1fki	–20.5	–4.8	–24.5	8.7	1.70
(2) CH…O bond between Phe#36 and Lig#1					
1fkb	–4.4	–2.2	–3.0	0.8	2.36
1fkf	–2.9	–2.2	–2.2	1.5	2.30
1fkg	–4.0	–1.8	–1.9	–0.3	2.65
1fki	–3.1	–2.0	–2.3	1.2	2.45

[a] Fragment pair index *IJ* is omitted. $\Delta\tilde{E}^{C} = \Delta\tilde{E}^{C}_{disp} + \Delta\tilde{E}^{C}_{ele} + \Delta\tilde{E}^{C}_{non-ele}$, where $\Delta\tilde{E}^{C}_{disp}$ is the dispersion energy contribution ($\Delta\tilde{E}^{C}_{disp} = \Delta\tilde{E}^{C} - \Delta\tilde{E}^{C}_{RHF}$), $\Delta\tilde{E}^{C}_{ele}$ is the electrostatic component of $\Delta\tilde{E}^{C}_{RHF}$, and $\Delta\tilde{E}^{C}_{non-ele}$ is the nonelectrostatic components of $\Delta\tilde{E}^{C}_{RHF}$: $\Delta\tilde{E}^{C}_{non-ele} = \Delta\tilde{E}^{C}_{RHF} - \Delta\tilde{E}^{C}_{ele}$.

the Lig#1 are listed in Table 8.2 with their components — that is, the electrostatic, nonelectrostatic (which correspond to the exchange-repulsion and charge transfer in PIEDA[32]), and MP2 electron correlation energies (also called the dispersion energies below). PIEs for all complexes are similar, in the range of –19.6 to –20.7 kcal/mol. However, attractive electrostatic energies are sensitive to small changes in the bonding distance, and highly correlate with it, with the correlation coefficient of 0.99. The difference between the shortest and longest pair is –3.6 kcal/mol. The mostly repulsive nonelectrostatic energies, however, vary by 3.4 kcal/mol as the bond distance becomes shorter and cancels the attractive electrostatic interaction.

The PIEs of Phe#36 with the diketopiperidino moiety were decomposed in the same manner as the hydrogen bond of the Tyr#82 fragment and are shown in Table 8.2. Although the absolute values are small, the contribution patterns of the CH···O hydrogen bonds are obviously different from those of the usual hydrogen bond (i.e., no correlation is observed between the bonding distances and the electrostatic or nonelectrostatic energies). Conversely, a high correlation is observed between distances and the dispersion energies (correlation coefficient of 0.97). A significant observation is the attractive nonelectrostatic interaction seen in the 1fkg complex with a longer bonding distance. Because all other complexes show a repulsive nonelectrostatic interaction, similar to normal hydrogen bonds, this behavior is uncharacteristic of hydrogen bonds, and this interaction should be categorized as nonpolar. Even excluding the 1fkg complex, no correlation typical for hydrogen

TABLE 8.3
Selected FMO-MP2/6-31G* Pair Interaction Energies (kcal/mol) between the 1fkf Ligand (FK506) and the FKBP Fragments in the Complex and Other Environments

Fragment Pair		Whole Complex Environment[a]	No Environment[b]	Whole Ligand Environment[c]
I	J	$\Delta\tilde{E}_{IJ}^{C}$	$\Delta\tilde{E}_{IJ}^{\text{no-env}}$	$\Delta\tilde{E}_{IJ}^{\text{lig-env}}$
Tyr#26	Lig#1	–6.9 (48.1)	–6.6	–6.7
Asp#37	Lig#2	–21.9 (25.1)	–31.9	–29.9
Glu#54	Lig#1	–3.2 (0.0)	–2.2	–3.0
Val#55	Lig#4	–13.5 (19.9)	–10.2	–9.6
Ile#56	Lig#3	–12.9 (44.9)	–14.0	–18.0
Trp#59	Lig#1	–7.4 (91.7)	–6.5	–6.7
Tyr#82	Lig#1	–20.1 (22.3)	–15.9	–18.8
His#87	Lig#2	–7.4 (39.9)	–5.7	–7.9
Phe#99	Lig#1	–4.9 (71.7)	–6.2	–6.4

[a] The percentage of dispersion energy in $\Delta\tilde{E}_{IJ}^{C}$ is shown in the parentheses.
[b] Just the pair of fragments with hydrogen caps.
[c] The protein residue with hydrogen caps and the whole ligand (divided into fragments).

bonds is observed. The high correlation with dispersion energy indicates that CH···O bonds in this system are a kind of van der Waals interaction.

Next, the environmental effects due to other fragments (from protein or ligands) upon the pair interactions are discussed for the 1fkf complex. We focus on the nine strong pair interactions shown in Table 8.3. To clarify the influence of the surrounding fragments on the interaction of a certain fragment pair, the fragment pair and some near fragments were extracted and the PIE was calculated for two reduced models: (a) no environment (only the hydrogen capped fragment pair) and (b) the entire ligand environment (all ligand fragments and one capped protein fragment). The FMO calculation was performed under the same conditions described in the previous section.

The results are shown in Table 8.3. In all fragment pairs considered, the PIEs were significantly affected by the surrounding fragments. The differences in the pair interaction energies are not very large between the "no environment" and the "whole ligand environment." However, a relatively large difference is found between the ligand environment and the whole complex environment. This means that a protein environment has the greatest effect on the PIE. The destabilization of 8.0 kcal/mol by the protein environment for the Asp#37-Lig#2 is the largest. This is explained by the polarization of Asp37, induced by the formation of the salt bridge and the hydrogen bond of the carboxyl anion of the Asp37 side chain with the guanidyl group of the Arg42 side chain and the hydroxyl group of the Tyr26 side chain, respectively.

8.3.4 Importance of the Polarization and the van der Waals Interactions

Using Equations 8.2 and 8.4, the interaction energy ΔE_{int} between a protein (P) and a ligand (L) in their complex (C) is expressed as follows:

$$\Delta E_{\text{int}} = \left\{ \left(\sum_{I \in \text{P}} E_I'^{\text{C}} + \sum_{\substack{I>J \\ I,J \in \text{P}}} \Delta \tilde{E}_{IJ}^{\text{C}} \right) - E^{\text{P}} \right) \right\} + \left\{ \left(\sum_{I \in \text{L}} E_I'^{\text{C}} + \sum_{\substack{I>J \\ I,J \in \text{L}}} \Delta \tilde{E}_{IJ}^{\text{C}} \right) - E^{\text{L(C)}} \right\} + \sum_{\substack{I>J \\ I \in \text{P}, J \in \text{L}}} \Delta \tilde{E}_{IJ}^{\text{C}}$$
$$= \Delta E_{\text{PLd}}^{\text{P}} + \Delta E_{\text{PLd}}^{\text{L}} + \Delta E_{\text{int}}^{\text{P-L}} \tag{8.6}$$

where superscript C denotes the values in the complex, and $\Delta E_{\text{PLd}}^{\text{P}}$ and $\Delta E_{\text{PLd}}^{\text{L}}$ are the destabilization portions of the polarization energy of the protein and ligand, respectively. $\Delta E_{\text{int}}^{\text{P-L}}$ is the protein–ligand interaction, which includes the stabilized part of the polarization interaction, protein–ligand charge transfer, and other types of interactions. The partial sum $\sum_{J \in \text{L}} \Delta \tilde{E}_{IJ}^{\text{C}}$ gives the interaction energy between residue I in the protein and the entire ligand.

The importance of the polarization in a protein–ligand interaction or a solvent–solute interaction was reported,[33–35] and it was revealed that the molecular polarization contributes from 10% to 40% of the total electrostatic interaction energy. However, these analyses were performed by the QM/MM method, where only a ligand molecule was treated as the QM part, and the polarization of proteins induced by ligand molecules was not considered. Our analysis estimated the polarization of both proteins and ligands, because we treat the entire complex system with the QM method.

The values of $\Delta E_{\text{PLd}}^{\text{P}}$, $\Delta E_{\text{PLd}}^{\text{L}}$, and $\Delta E_{\text{int}}^{\text{P-L}}$ are listed in Table 8.4. The values of $\Delta E_{\text{PLd}}^{\text{L}}$ are quite large: 22.5, 17.4, 13.1, and 12.3 kcal/mol for the 1fkb, 1fkf, 1fkg, and 1fki ligands, respectively. The corresponding protein values of $\Delta E_{\text{PLd}}^{\text{P}}$ are 16.2, 11.8,

TABLE 8.4
Destabilization Contributions to the Protein $\Delta E_{\text{PLd}}^{\text{P}}$ and Ligand $\Delta E_{\text{PLd}}^{\text{P}}$ Polarization Energies, and the Sum of the Intermolecular Pair Interaction Energies $\Delta \tilde{E}_{\text{int}}^{\text{P-L}}$ (All in kcal/mol), FMO-MP2/6-31G*

System	$\Delta E_{\text{PLd}}^{\text{P}}$	$\Delta E_{\text{PLd}}^{\text{L}}$	$\Delta \tilde{E}_{\text{int}}^{\text{P-L}}$
1fkb	16.2	22.5	–159.2
1fkf	11.8	17.4	–139.6
1fkg	4.4	13.1	–97.6
1fki	3.5	12.3	–93.3

4.4, and 3.5 kcal/mol, respectively. The polarization of the protein is much less than that of small ligands. It is interesting that the magnitude of the polarization energy is on the same order as the binding affinity, and it correlates with the number of hydrogen bonds between FKBP and ligands. Among the four complexes, the largest total destabilization energy of the polarization ($\Delta E^{P}_{PLd} + \Delta E^{L}_{PLd}$) is observed for 1fkb (38.7 kcal/mol). If the polarization of a molecule is assumed to be the dielectric linear response to the electric field exerted by the interacting partner molecules, the stabilization contribution of the polarization energy ($\Delta E^{P}_{PLs} + \Delta E^{L}_{PLs}$) is twice that of the destabilization; thus, the stabilization energies estimated from uncorrelated destabilization energies are about −59, −43, −24, and −22 kcal/mol for 1fkb, 1fkf, 1fkg, and 1fki, respectively. These energies account for 23% to 38% of $\Delta \tilde{E}^{P\text{-}L}_{int}$. The total polarization energies,[32] defined as $\Delta E^{Total}_{PLs} + \Delta E^{Total}_{PLd}$, are 14% to 20% of the total electrostatic interaction energy, in agreement with other studies.

The calculated dipole moments of the ligands at the geometry in the complexes are 13.1(11.6), 8.3 (5.1), 6.1 (2.3), and 6.8 (3.2) Debye for 1fkb, 1fkf, 1fkg, and 1fki, respectively, where the dipole moments of the isolated molecules are given in parentheses. The order of the dipole moment values in the series of ligands corresponds to that of ΔE^{L}_{PLd}, except for the 1fkg ligand. This suggests that the ligand dipole polarizes the protein, and the enhanced protein dipole is reflected in the polarization of the ligand. The exception of the 1fkg ligand suggests that both the local dipole of the ligand and the total dipole can be important. Interestingly, the dipole moments of the ligands are larger at the distorted structures in the complexes, compared to those of the isolated molecules. The increased dipole moments might enhance polarization stabilization and partially compensate for destabilization due to geometric distortion.

In the previous section, the significance of the electron correlation effect in FKBP–ligand binding was described. A more detailed analysis is provided here. It is notable that the pair interactions between Lig#1 and the hydrophobic aromatic residues surrounding it (i.e., Tyr26, Phe36, Phe46, Trp59, Tyr82, and Phe99) indicate a larger correlation contribution (Figure 8.4). In particular, the interatomic distances between ring carbon atoms suggest the CH/π interactions between the indole ring of Trp59 and the piperidine ring. The contribution of the electron correlation for this nonpolar interaction accounts for more than 90% of the total PIE for the 1fkf ligand (Table 8.3), and values for Phe#99 and Tyr#26 are considerable as well, 71.7% and 48.1%, respectively. Therefore, the van der Waals interaction is another major contributor in the Lig#1 fragment, as well as its hydrogen bond.

8.4 SOLVENT EFFECTS ON PROTEIN–LIGAND BINDING

8.4.1 Continuum Solvent Models

Thus far, we discussed molecular structures and interactions in vacuo. However, most physiological protein–ligand associations, reactions, and interactions take place in an aqueous medium. The solvent effect is important and cannot be ignored. There are two major methods of incorporating the solvent effects. One method treats water molecules explicitly, and the other is a continuum solvent model. Because the

most reliable explicit method needs an ensemble average of the water configuration from a statistical mechanics simulation of the system, a huge number of calculations are required for large biomolecular systems, such as proteins. On the other hand, several continuum solvation models, such as GB,[13] PB,[12] and the PCM,[14] have been proposed, and the GB and PB methods are frequently used for the binding analysis of proteins or nucleic acids, because of their reduced computational requirements. In particular, the MM-PB/SA method[15] has been applied to the binding affinity predictions of many protein–ligand complexes.

We estimated the solvation effect on the FKBP–ligand binding using the PB/SA model. Recently, Gräter et al. applied a combination of the QM/MM method and the PB/SA solvation model for calculating the protein–ligand binding affinity,[36] and the obtained results were in good agreement with experimental results for trypsin and FKBP complexes. They employed a semi-empirical MO method, which is less rigorous than the *ab initio* MO method, to describe molecular interactions. The same computational scheme was used as for MM-PB/SA,[37] except for ΔG_{int}, which was calculated as the sum of the FMO binding energy ΔE_{b} and the vibrational free energy contribution ΔG_{vib}:

$$\Delta G_{\mathrm{int}} = \Delta E_{\mathrm{b}} + \Delta G_{\mathrm{vib}} \tag{8.7}$$

ΔG_{vib} was calculated at 298 K by the MM normal mode analysis using the cff91 force field in the Discover program.[38] The electrostatic contribution $\Delta G_{\mathrm{sol}}^{\mathrm{ele}}$ was computed by the DelPhi program,[39,40] in which the dielectric constants of 1 and 80 are used for the interior and exterior of the solutes, respectively. The Mulliken charges of the FMO-RHF/6-31G* calculations were used for the atomic partial charges of the solute atoms. The nonpolar contribution $\Delta G_{\mathrm{sol}}^{\mathrm{nonpolar}}$ was calculated using the cff91 parameter set of the insightII program[38] — that is, 0.00682 × (solvent accessible surface area in Å^2) + 0.80 kcal/mol. The binding free energy ΔG_{b} was calculated as follows:

$$\Delta G_{\mathrm{b}} = (\Delta E_{\mathrm{b}} + \Delta G_{\mathrm{vib}}) + (\Delta G_{\mathrm{sol}}^{\mathrm{ele}} + \Delta G_{\mathrm{sol}}^{\mathrm{nonpol}}) \tag{8.8}$$

8.4.2 Desolvation Penalty for the Association of the Protein and Ligand

The calculated and experimental binding free energies are listed in Table 8.5. The solvation energy contribution ΔG_{sol} to the binding energy (which can be thought of as the desolvation penalty) accounts for 50 to 90 kcal/mol, and compensates the large binding energy calculated in vacuo by the FMO method. The calculated binding free energies reproduced experimental binding free energies with a maximum error of 7.2 kcal/mol and are in the same energy range as the experimental data. However, the relative order of affinities of the four ligands does not agree with experimental values. The negative vibration free energy change for the 1fkg and 1fki complexes is hard to explain. The harmonic approximation on the normal mode calculation may not be accurate enough. It is noteworthy that both the dispersion and solvation energy contributions are essential for the protein–ligand binding study. The FMO

TABLE 8.5
Calculated and Experimental Binding Affinities and the PB/SA Solvation Energies (ΔG_{sol}) with Components[a] (kcal/mol)

	ΔG_{int}		ΔG_{sol}			
System	ΔE_b [b]	ΔG_{vib}	ΔG_{sol}^{ele}	$\Delta G_{sol}^{nonpolar}$	ΔG_b**(calc)**	ΔG_b**(expt)**[c]
1fkb	–103.9	4.4	95.4	–7.3	–11.4	–13.2
1fkf	–102.2	11.2	81.6	–7.0	–16.4	–12.8
1fkg	–70.1	–8.1	69.9	–6.1	–14.4	–10.9
1fki	–71.3	–1.6	62.0	–5.7	–16.7	–9.5

[a] The calculated binding affinity is given by $\Delta G_b(\text{calc}) = \Delta E_b + \Delta G_{vib} + \Delta G_{sol}^{ele} + \Delta G_{sol}^{nonpolar}$.
[b] FMO2-MP2/6-31G*.
[c] The experimental binding free energies converted from the inhibition constants reported in Holt, D.A., Luengo, J.I., Yamashita, D.S. et al. 1993. *J. Am. Chem. Soc.* 115: 9925–9938.

interaction energies that include the electron correlation (dispersion) energy make it possible to compare the calculated ΔG_b with the experimental values at the same energy level. The RHF binding energies are far from the experimental values.

In the future, we plan to investigate the solvent effect with the PCM method, which was recently added to the FMO code in GAMESS. Although several issues remain, the preliminary data show that the binding free energies are also reproduced well for the FKBP complexes with the four ligands. However, the contribution of the electrostatic energy component in the PCM differs from that of the MM-PB/SA method.

8.5 CONCLUSION

In this chapter, we presented the interaction analysis of protein–ligand complexes. This includes geometry optimization, pair interaction analysis, interaction energy decomposition, and solvent effect analysis. Several important observations are as follows: (a) the large contribution of the dispersion energy to the protein–ligand interaction, (b) the polarization effect accounts for 10% to 40% of the total electrostatic energy, (c) large binding energies calculated in vacuo are made small by the solvent effects, and (d) the calculated binding free energies are on the same order as the experimental values.

Thus, the FMO method is useful for the interaction analysis between proteins and ligands. In particular, the contributions that are difficult to treat in the MM method, such as the polarization, charge transfers, and weak interactions like CH/π and CH···O, are highlighted. Our future directions for the accurate binding affinity calculations are as follows: (a) inclusion of the electron correlation in the structure optimization, (b) use of the larger basis set, (c) the basis set superposition error correction, (d) better treatment of the solvation energy, and (e) conformation sampling.

REFERENCES

1. Appelt, K., Bacquet, R.J., Bartlett, C.A. et al. 1991. Design of enzyme inhibitors using iterative protein crystallographic analysis. *J. Med. Chem.* 34:1925–1934.
2. Ealick, S.E., Babu, Y.S., Bugg, C.E. et al. 1991. Application of crystallographic and modeling methods in the design of purine nucleoside phosphorylase inhibitors. *Proc. Natl. Acad. Sci. USA* 88:11540–11544.
3. Lam, P.Y., Jadhav, P.K., Eyermann, C.J. et al. 1994. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science* 263:380–384.
4. Webber, S.E., Bleckman, T.M., Attard, J., Deal, J.D. et al. 1993. Design of thymidylate synthase inhibitors using protein crystal structures: the synthesis and biological evaluation of a novel class of 5-substituted quinazolinones. *J. Med. Chem.* 36:733–746.
5. Bash, P.A., Singh, U.C., Brown, F.K., Langridge, R., Kollman, P.A. 1987. Calculation of the relative change in binding free energy of a protein-inhibitor complex. *Science* 235:574–576.
6. Beveridge, D.L., Dicapua, F.M. 1989. Free-energy via molecular simulation: applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Bio.* 18:431–492.
7. Warshel, A., Levitt, M. 1976. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* 103:227–249.
8. Singh, U.C., Kollman, P.A. 1986. A combined *ab initio* quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: applications to the CH3Cl + Cl^- exchange reaction and gas phase protonation of polyethers. *J. Comput. Chem.* 7:718–730.
9. Field, M.J., Bash, P.A., Karplus, M. 1990. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.* 11:700–733.
10. Maseras, F., Morokuma, K. 1995. IMOMM: a new integrated *ab initio* + molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *J. Comput. Chem.* 16:1170–1179.
11. Saen-oon, S., Kuno, M., Hannongbua, S. 2005. Binding energy analysis for wild-type and Y181C mutant HIV-1 RT/8-Cl TIBO complex structures: quantum chemical calculations based on the ONIOM method. *Proteins* 61:859–869.
12. Honig, B., Nicholls, A. 1995. Classical electrostatics in biology and chemistry. *Science* 268:1144–1149.
13. Still, W.C., Tempczyk, A., Hawley, R.C., Hendrickson, T. 1990. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* 112:6127–6129.
14. Miertius, S., Scrocco, E., Tomasi, J. 1981. Electrostatic interaction of a solute with a continuum. A direct utilization of *ab initio* molecular potentials for the prevision of solvent effects. *Chem. Phys.* 55:117–129.
15. Srinivasan, J., Cheatham, T.E. III, Cieplak, P., Kollman, P.A., Case, D.A. 1998. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *J. Am. Chem. Soc.* 120:9401–9409.
16. Holt, D.A., Luengo, J.I., Yamashita, D.S. et al. 1993. Design, synthesis, and kinetic evaluation of high-affinity FKBP ligands and the x-ray crystal structures of their complexes with FKBP12. *J. Am. Chem. Soc.* 115:9925–9938.
17. Becker, J.W., Rotonda, J., McKeever, B.M. et al. 1993. FK-506-binding protein: three-dimensional structure of the complex with the antagonist L-685,818. *J. Biol. Chem.* 268:11335–11339.

18. Verkhivker, M.G., Rejto, A.P., Gehlhaar, K.D., Freer, T.S. 1996. Exploring the energy landscapes of molecular recognition by a genetic algorithm: analysis of the requirements for robust docking of HIV-1 protease and FKBP-12 complexes. *Proteins* 25:342–353.
19. Rejto, A.P., Verkhivker, M.G. 1997. Mean field analysis of FKBP12 complexes with FK506 and rapamycin: implications for a role of crystallographic water molecules in molecular recognition and specificity. *Proteins* 28:313–324.
20. Becker, J.W., Retonda, J., Cryan, J.G. et al. 1999. 32-Indolyl ether derivatives of ascomycin: three-dimensional structures of complexes with FK506-binding protein. *J. Med. Chem.* 42:2798–2804.
21. Dubowchik, G.M., Vrudhula, V.M., Dasgupta, B. et al. 2001. 2-Aryl-2,2-difluoroacetamde FKBP12 ligands: synthesis and X-ray structutal studies. *Org. Lett.* 3:3987–3990.
22. Zacharias, M. 2004. Rapid protein–ligand docking using soft modes from molecular dynamics simulations to account for protein deformability: binding of FK506 to FKBP. *Proteins* 54:759–767.
23. Duyne, G.D.V., Standaert, R.F., Karplus, P.A., Schreiber, S.L., Clardy, J. 1991. Atomic structure of FKBP-FK506, an immunophilin-immunosuppressant complex. *Science* 252:839–842.
24. Duyne, G.D.V., Standaert, R.F., Schreiber, S.L., Clardy, J. 1991. Atomic structure of the rapamycin human immunophilin FKBP12 complex. *J. Am. Chem. Soc.* 113:7433–7434.
25. Schmidt, M.W., Baldridge, K.K., Boatz, J.A. et al. 1993. General atomic and molecular electronic structure system. *J. Comput. Chem.* 14:1347–1363.
26. SYBYL Version 6.9.1. 2003. St. Louis, MO: Tripos, Inc.
27. Vriend, G. 1990. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* 8:52–56.
28. Fedorov, D.G., Kitaura, K. 2007. Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. *J. Phys. Chem. A* 111:6904–6914.
29. Fedorov, D.G., Kitaura, K. 2004. The importance of three-body terms in the fragment molecular orbital method. *J. Chem. Phys.* 120:6832–6840.
30. Nakanishi, I., Fedorov, D.G., Kitaura, K. 2007. Molecular recognition mechanism of FK506 binding protein: an all-electron fragment molecular orbital study. *Proteins* 68:145–158.
31. Fukuzawa, K., Kitaura, K., Uebayasi, M. et al. 2005. *Ab initio* quantum mechanical study of the binding energies of human estrogen receptor a with its ligands: an application of fragment molecular orbital method. *J. Comput. Chem.* 26:1–10.
32. Fedorov, D.G., Kitaura, K. 2007. Pair interaction energy decomposition analysis. *J. Comput. Chem.* 28:222–237.
33. Gao, J., Xia, X. 1992. A priori evaluation of aqueous polarization effect through Monte Carlo QM-MM simulations. *Science* 258:631–635.
34. Garcia-Viloca, M., Truhlar, D.G., Gao, J. 2003. Importance of substrate and cofactor polarization in the active site of dihydrofolate reductase. *J. Mol. Biol.* 327:549–560.
35. Hensen, C., Hermann, J.C., Nam, K., Ma, S., Gao, J., Holtje, H.-D. 2004. A combined QM/MM approach to protein-ligand interactions: polarization effects of the HIV-1 protease on selected high affinity inhibitors. *J. Med. Chem.* 47:6673–6680.
36. Gräter, F., Schwarzl, S.M., Dejaegere, A., Fischer, S., Smith, J.C. 2005. Protein/ligand binding free energies calculated with quantum mechanics/molecular mechanics. *J. Phys. Chem. B* 109:10474–10483.

37. Wang, W., Kollman, P.A. 2001. Computational study of protein specificity: the molecular basis of HIV-1 protease drug resistance. *Pro. Natl. Acad. Sci. USA* 98:14937–14942.
38. Discover and InsightII. 2005. San Diego, CA: Accelrys Software Inc.
39. Rocchia, W., Alexov, E., Honig, B. 2001. Extending the applicability of the nonlinear Poisson-Boltzmann equation: multiple dielectric constants and multivalent ions. *J. Phys. Chem. B* 105:6507–6514.
40. Rocchia, W., Sridharan, S., Nicholls, A. et al. 2002. Rapid grid-based construction of the molecular surface for both molecules and geometric objects: applications to the finite difference Poisson-Boltzmann method. *J. Comput. Chem.* 23:128–137.

9 How Does the FMO Method Help in Studying Viruses and Their Binding to Receptors?

Toshihiko Sawada, Tomohiro Hashimoto, Hiroaki Tokiwa, Tohru Suzuki, Hirofumi Nakano, Hideharu Ishida, Makoto Kiso, and Yasuo Suzuki

CONTENTS

9.1 INFLUENZA A VIRAL HEMAGGLUTININ AND ITS SIALOOLIGOSACCHARIDE RECEPTORS

A new type of influenza virus appears every few decades and causes pandemic outbreaks of fatal influenza; examples include Spanish influenza (H1N1, 1918), Asian flu (H2N2, 1957), and Hong Kong flu (H3N2, 1968). These three pandemic influenza viruses are classified type A virus on the basis of the internal protein antigens. New influenza viruses evolve in avian hosts and then quickly spread worldwide by means of bird migration. In the 1990s, some clinical cases of human infection with avian influenza A/H5N1 virus were reported,[1–5] suggesting that the H5N1 virus has mutated and acquired great potential to cause pandemic influenza in humans. The original H5N1 avian viruses can bind to avian-type receptors on the human lower respiratory tract[6,7]; however, this infection mechanism does not cause pandemic human influenza. We should pay attention to the binding affinity of H5N1 virus to human-type receptors. Recent studies revealed the molecular biological bases of virus infection as well as the mechanism of virus host range alteration from

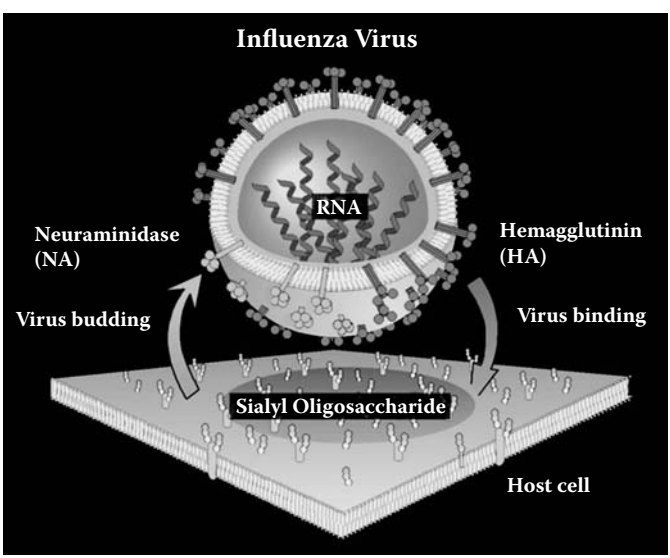


FIGURE 9.1 Influenza virus and host cell. Hemagglutinin (HA) works on the virus binding to host cell surface and on the membrane fusion of virus with the cell (see Figure 9.2). Neuraminidase (NA) hydrolyzes the neuraminic acid (sialic acid)–oligosaccharide linkage on the HA receptor sialosaccharides when the viruses go outside the host cell. (This illustration was drawn by Dr. Osamu Kanie, Mitsubishi Kagaku Institute of Life Sciences, Machida, Tokyo, Japan.) **(See color insert following page 117.)**

avian to human. Reverse genetics studies demonstrated that several two-point amino acid substitutions in H5N1 hemagglutinin (HA) markedly increased the virus's affinity for human-type receptors.[8,9] Steavens et al. and Yang et al. reported that single- or double-point mutations of H5 gave larger binding affinity to human-type receptors.[10,11]

Figure 9.1 illustrates the influenza virus and host cell surface. Two membrane glycoproteins hemagglutinin (HA) and neuraminidase (NA) work on the virus infection to host cell and on the virus budding from the host.[12–14] The host range of influenza A viruses has been mainly determined by HA binding specificity to host cell surface receptors sialoglycoproteins and sialoglycolipids.[15,16]

Several X-ray crystallographic structures of HAs complexed with sialooligosaccharide receptor analogues show their intermolecular interaction patterns.[8,17–22] HA forms a trimer and three sialosides can attach to the surface sialoside binding site on each HA1 monomer (Figure 9.2). HA specifically recognizes the nonreducing terminal of sialoside receptor and which is sensitive to differences in sialic acid (Sia) species and sialic acid-galactose (Gal) linkage.[15,23]

FIGURE 9.2 Energy minimum structure of avian influenza A viral H3 trimer in complex with avian-type receptor Neu5Acα(2-3)Gal disaccharide analogues. Left: avian H3 trimer-three Neu5Acα(2-3)Gal analogues complex. HA trimer; ribbon. Neu5Acα(2-3)Gal; CPK model. Right: avian H3 monomer-Neu5Acα(2-3)Gal complex. Sialoside binding domain HA1; colorful ribbon, membrane fusion domain HA2; yellow ribbon. (Figures were prepared by Discovery Studio Visualizer v2.0.) **(See color insert.)**

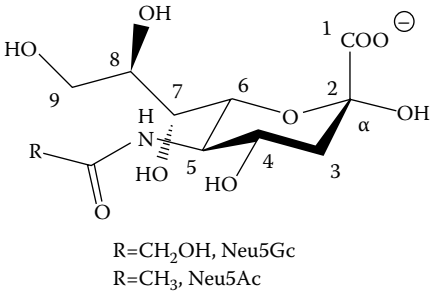


FIGURE 9.3 Chemical structure of sialic acids.

Two major Sia are 5-*N*-glycolylneuraminic acid (Neu5Gc) and 5-*N*-acetylneuraminic acid (Neu5Ac) (Figure 9.3). Neu5Gc-type glycoconjugates are widely found in many animals such as duck, swine, and horses, whereas Neu5Ac are expressed on the normal human tissues and chicken. Duck, a natural host of influenza viruses, also expresses a limited amount of Neu5Ac.[15,16] Figure 9.4 shows chemical structures of nonreducing terminal Sia-Gal disaccharides on sialoside receptors. Human HA more strongly binds to human-type receptor Neu5Acα(2-6)Gal than to avian-type receptor Neu5Ac and 5Gcα(2-3)Gal.

9.2 RELATIONSHIP BETWEEN HA BINDING SPECIFICITY TO SIALOSIDES AND VIRUS HOST RANGE

Unfortunately, people experienced the serious pandemic human influenza A/H3N2. We believe that it is important to investigate the mechanism by which the host ranges of the avian and human H3 viruses are determined, before another outbreak of the human virus occurs.

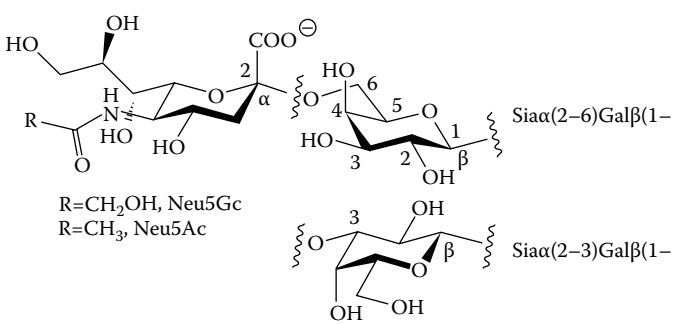


FIGURE 9.4 Chemical structures of nonreducing terminal sialylgalactose on the sialooligosaccahride receptors.

Human A viral HA H3 subtype has a Leu at position 226 on a sialoside binding site that binds to human-type receptor Neu5Acα(2-6)Gal more strongly than to avian-type receptor Neu5Acα(2-3)Gal.[24–31] Avian H3 has a Gln at position 226 instead of Leu that binds strongly to avian-type α2-3 sialosaccharide.[26,29] The Human H3 mutant Leu226Gln preferentially binds to avian α2-3 receptor, as indicated by a thin-layer chromatography (TLC)-virus binding assay (Figure 9.5).[32] These results suggest that amino acid substitution at the HA1 sialoside binding site shifts the virus host range between birds and humans.[15,16,24,39,40]

Figure 9.6A shows molecular structures of the sialoside binding site of avian and human influenza A virus H3 HA1 in complex with Neu5Acα(2-3 and 2-6)Gal analogues.[19] The sialoside binding site consists of 130-loop, 150-loop, 190-helix, and 220-loop to total 70 amino acids in the models whose tertiary structures are nearly the same. The corresponding amino acid sequences of the binding sites differ at 10 positions: two positions that are nonmatching residues, six positions that are weakly matching residues, and two positions that are strongly matching residues

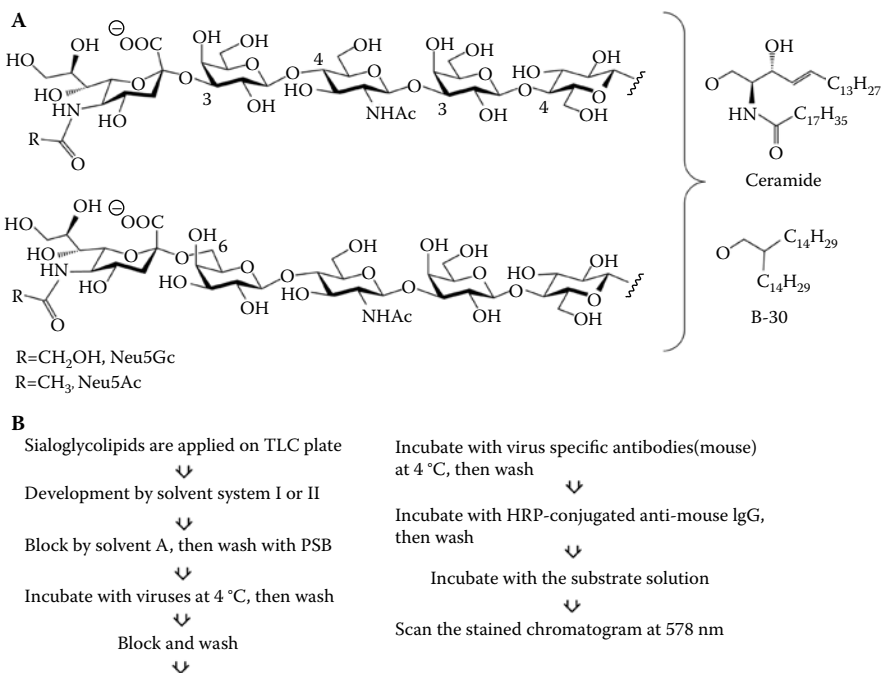


FIGURE 9.5 Sialylparaglobosides and corresponding B-30 derivatives for TLC-virus binding assay. (A) Chemical structures of sialylparaglobosides and corresponding B-30 derivatives.[33,34,35] (B) Protocol of TLC-virus binding assay.[36] Solvent system I, $CHCl_3$-CH_3OH-12 mM $MgCl_2$ = 5:4:1; solvent system II, $CHCl_3$:CH_3OH:2.5 N ammonia = 50:40:9; PBS, phosphate-buffered saline, solution A; PSB + [1% egg albumin] + [1% polyvinylpyrrolidone]; HRP, horseradishperoxidase, the substrate solution; [10 m/M Tris-HCl buffer, pH 7.2]:3% 4-chloro-1-naphthol in CH_3OH:3% H_2O_2 aq = 5:1:0.02. TLC-virus binding assays were reported in our previous works.[29,32,37,38]

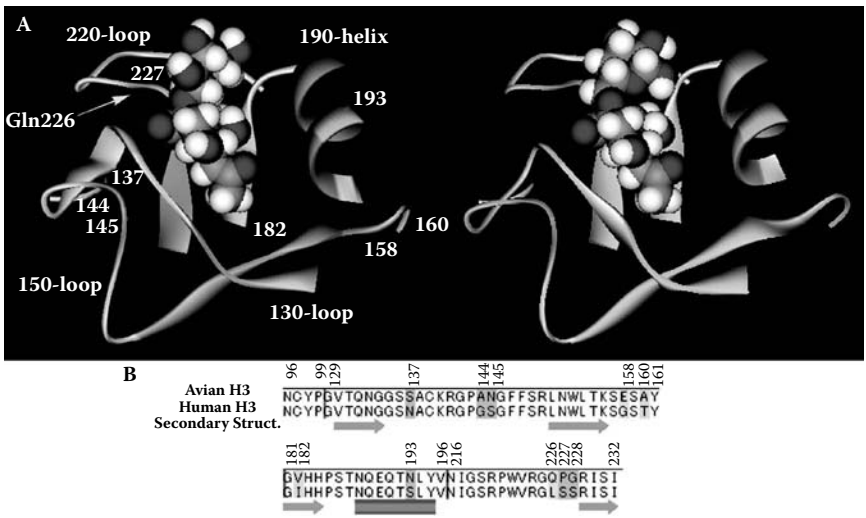


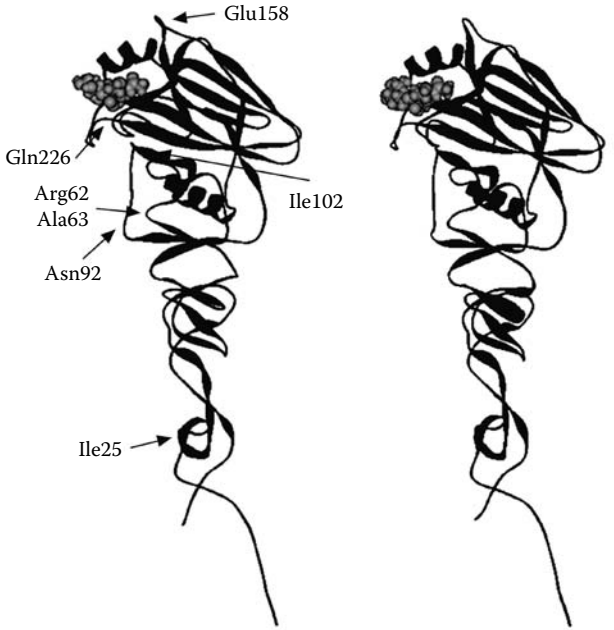
FIGURE 9.6 Sialoside binding sites on avian and human H3s HA1 in complex with avian Neu5Acα(2-3)Gal and human Neu5Acα(2-6)Gal. (A) Sialoside binding sites consist of Asn96-Pro99, Gly129-Tyr161, Gly181-Val196, and Asn216-Ile232 (total 70 amino acids). Left: avian H3-avian Neu5Acα(2-3)Gal complex. Right: human H3–human Neu5Acα(2-6)Gal complex. (B) Sequence alignments of avian and human H3s in the models; yellow, nonmatching residues; green, weak matching residues; light blue, strong matching residues. The substitution positions are shown in Figure 9.6A, left. (Reproduced from Sawada, T., Hashimoto, T., Tokiwa, H. et al. 2008. *Glycoconj. J.* 25:805–815. With permission.) **(See color insert.)**

(Figure 9.6B). The differences at positions 137, 145, 226, and 228 are located at the sites that interact directly with Neu5Ac-Gal disaccharides.

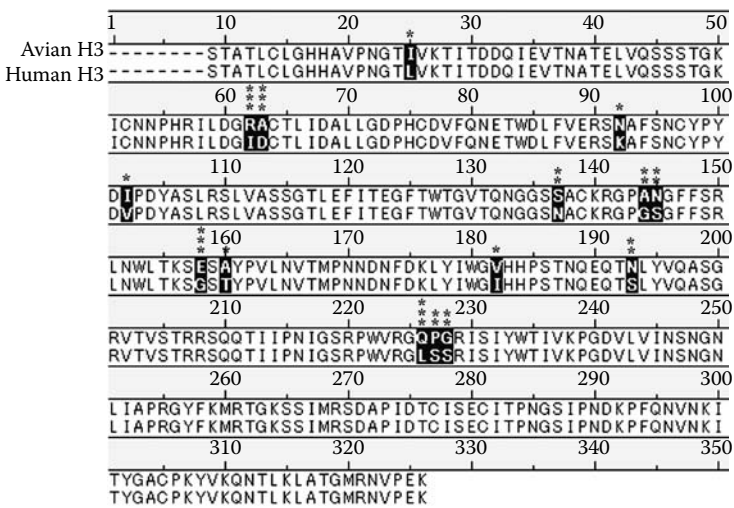
The avian and human H3 HA1 full domains (Ser9-Lys326, 318 amino acids) differ at 15 amino acid residues (Figure 9.7), indicating that most (10 out of 15) of the amino acid substitutions occur at the sialoside binding site. These substitutions sometimes dramatically change the HA–sialoside binding property, as mentioned above, point mutation at position 226 affects the virus host range.

9.3 FMO STUDIES OF HA-NEU5ACα(2-3 AND 2-6) GAL DISACCHARIDE COMPLEXES TO CLARIFY HOW A VIRUS HOST RANGE IS DETERMINED

If the host range of a new or mutant influenza virus could be predicted in advance, measures to prevent an outbreak of pandemic human influenza could be taken immediately after the new virus emerges. The interaction between viral HA and sialoside receptors is a target for in silico chemical prediction studies of the mechanism by which the virus host range is determined. Quantum chemical calculations can provide chemical information about HA–sialosaccharide interaction without the need to work with potentially dangerous influenza virus mutants. Using the H3 subtype system,



(a)



(b)

FIGURE 9.7 HA1 full domains of avian and human H3 in complex with Neu5Acα(2-3 and 2-6)Gal disaccharide. (a) Left: avian H3–avian Neu5Acα(2-3)Gal complex (Ser9-Lys326); Right: human H3–human Neu5Acα(2-6)Gal complex. (b) Amino acid sequence of avian/human H3 HA1 domains differ at 15 positions. ***: nonmatching residues; **: weak matching residues; *: strong matching residues.

we are attempting to construct HA–sialoside complexes and then subject them to *ab initio*–based fragment molecular orbital (FMO) calculations in order to explain the HA binding specificity.[41–43] We are currently focusing our attention on avian viral HA binding to Neu5Acα(2-3 and 2-6)Gal and on the comparison of avian HA mutant–human α2-6 interaction with the original human HA–human α2-6 interaction.

Note that avian H3 weakly recognizes human α2-6 in a TLC-virus binding assay using synthetic sialylparaglobosides or corresponding B30 derivatives (Figure 9.5)[32]; this result is supported by the X-ray crystallographic structure of the avian H3-human α2-6 sialooligosaccharide complex.[25] The conclusion raises two questions: Why does avian H3 bind avian α2-3 more strongly than it binds human α2-6? How does avian Gln226Leu H3 bind with moderate affinity to human α2-6?

To answer these questions, we used the FMO method to analyze the relationship between avian H3-sialoside binding specificity and the corresponding molecular interaction of avian H3 with the Neu5Acα(2-3 and 2-6)Gal receptors. This study was part of our in silico chemical prediction studies of virus host range determination. FMO method estimates the binding energies of avian and human H3s with Neu5Acα(2-3 and 2-6)Gal disaccharide analogues. This approach is based on the hypothesis that a larger binding energy indicates a stronger binding affinity. In addition to the FMO studies, molecular dynamics simulations,[9,44,45] docking simulation analysis,[46] and topological comparison of HA-sialoside complexes[47] are promising and essential approaches.

9.4 AVIAN H3 HA BINDS TO THE AVIAN-TYPE α2-3 RECEPTOR MORE STRONGLY THAN TO THE HUMAN-TYPE α2-6 RECEPTOR AT THE FMO-MP2/6-31G LEVEL

9.4.1 Computational Methods

Avian H3-Neu5Acα(2-3 and 2-6)Gal disaccharides complexes for FMO studies were generated from the crystal structures of the corresponding complexes of H3 trimer with α(2-3 and 2-6)sialooligosaccharide analogues.[19] First, we modified the sialooligosaccharides in the complexes to Neu5Acα(2-3 and 2-6)Gal disaccharides in silico, then we used Discovery Studio program (ver. 1.5.1, Accelrys Software Inc.) to optimize the geometries by means of molecular mechanics (MM) calculations with a consistent force field.[48] The optimum avian Gln226Leu H3-human α2-6 disaccharide complex was obtained by an *in silico* mutation of Gln226 to Leu in the avian H3-human α2-6 disaccharide complex; a change in the dihedral angle of the α2-6 bond to that of the human Leu226 H3-type orientation, with reference to the crystal structure of Neu5Acα(2-6)Galβ(1-4)GlcNAcβ(1-3)Galβ(1-4)Glc:LSTc complexed with human H3[49]; and optimization of the disaccharide complex geometry by MM calculations. Human H3-human α2-6 complex was obtained by modification of the human H3-Neu5Acα(2-3)Galβ(1-4)Glc complex.[50] We replaced the α2-3 sialotrisaccharide with human Neu5Acα(2-6)Gal disaccharide by superposing common Neu5Ac residue coordinates and then changing the dihedral angle of the α2-6 bond on Neu5Acα(2-6)Gal disaccharide with reference to the LSTc conformation

complexed with human H3.[49] Human Leu226Gln H3-avian α2-3 complex was prepared by *in silico* point mutation from Leu226 to Gln in the human H3-avian α2-3 complex,[50] changed α2-3 bond dihedral angle to avian Gln226 H3-type orientation,[19] and geometry optimization. MM calculations were carried out with structurally determined water molecules in the crystal structures.

FMO calculations applied to small H3-sialoside model complexes. We cut out the small H3-sialoside models (Figure 9.6A, 70 amino acids) from the geometry optimum structures. This approach includes the amino acid residues of the sialoside binding sites, especially position 226, beyond which the residues are too far away to play a role in binding. Peptides terminals in the models were treated as NH3+ and COO–. Because there are few water molecules around the complex between the sialoside binding site and the Neu5Acα(2-3 and 2-6)Gal analogues in the H3 crystal structures, we computed the model complexes without water molecules. Single-point energies of the model complexes were computed at the FMO-RHF level with STO-3G, 6-31G, 6-31G(d) basis sets and the FMO-MP2/6-31G level of theory using ABINIT-MP program.[51] The later correlated method evaluates various interaction stabilizations based on van der Waals dispersion force in the carbohydrate–amino acid interaction. Spiwok et al. calculated the stabilizations of CH–π interaction between tryptophan and carbohydrate in the β-galactosidase substrates or products complex to be 2.4 to 5.2 kcal/mol at the MP2/6-31+G(d) level.[52] Fernández-Alonso et al. estimated the stabilizing interaction energy of the fucose–benzene complex to be 3.0 kcal/mol at the MP2/6-31G(d,p) level of theory.[53]

The 70 amino acids on the H3-sialoside complexes were divided into one amino acid residue as a single fragment (with the exception of Cys S-S Cys) by means of the automatic fragmentation program in the ABINIT-MP package, and each Neu5Ac-Gal disaccharide was also treated as a single fragment. The receptors and sialoside binding sites were assigned charges –1 and +1, respectively. We calculated single-point energies of each complex ($E_{complex}$), Neu5Ac-Gal ($E_{receptor}$), and the binding site (E_{H3}) to estimate binding energies (ΔE) between the receptor and H3 by the following expression: $\Delta E = (E_{receptor} + E_{H3}) - E_{complex}$. We also analyzed interfragment interaction energies between Neu5Ac-Gal receptor and amino acids on the sialoside binding site.

9.4.2 The Binding Energy of the Avian H3–Avian α2-3 Complex Was Larger Than That of the Avian H3–Human α2-6 Complex at the FMO-MP2/6-31G Level

Table 9.1 gives ΔE values for the complexes between H3s and Neu5Acα(2-3 and 2-6) Gal disaccharide. Avian H3 bound to avian α2-3 receptor 11.4 kcal/mol stronger than to human α2-6 with intermolecular hydrophilic and lipophilic stabilizations at the FMO-MP2/6-31G level (Table 9.1, entries 1, 3).[43] In the avian H3 complexes, extension from minimal basis sets to valence double zeta basis sets at the HF level afforded a $\Delta E_{\alpha 2\text{-}3} - \Delta E_{\alpha 2\text{-}6}$ value of 13.1 kcal/mol. The addition of a polarization function to 6-31G at the RHF level decreased the $\Delta\Delta E$ value, which was close to the

TABLE 9.1
Δ*E*s of Avian/Human H3s with Neu5Acα(2-3 and 2-6)Gal Disaccharides

			RHF			MP2[c]
Entry	Δ*E*[a]	H3	STO-3G[b]	6-31G	6-31G(d)	6-31G
1	$\Delta E_{\alpha 2\text{-}3}$	Avian	136.9	149.4	135.3	180.4
2		Human Leu226Gln	137.6	148.8	134.8	179.7
3	$\Delta E_{\alpha 2\text{-}6}$	Avian	128.7	136.3	126.5	169.0
4		Avian Gln226Leu	118.2	125.2	115.8	157.6
5		Human	117.9	129.4	117.8	154.3

[a] Δ*E*s are given in kcal/mol.
[b] Sawada, T., Hashimoto, T., Nakano, H. et al. 2006. *Biochem. Biophys. Res. Commun.* 351:40–43.
[c] See Reference [43].

corresponding STO-3G energy. Δ*E*s at the MP2/6-31G level were 25 to 33 kcal/mol larger than the RHF/6-31G energies, and thus, intermolecular van der Waals interactions strongly stabilized the H3-Neu5Acα(2-3/6)Gal complexes to give $\Delta E_{\alpha 2\text{-}3}$–$\Delta E_{\alpha 2\text{-}6}$ 11.4 kcal/mol in the avian H3 complexes. Unexpectedly, $\Delta E_{\alpha 2\text{-}6}$ for avian Gln226Leu H3 and human α2-6 was almost the same with the value for human H3–human α2-6 complex with amino acid sequence differences at nine positions at the FMO-MP2/6-31G level (Table 9.1, entries 4, 5). The reverse mutation, Leu226Gln for human H3, gave $\Delta E_{\alpha 2\text{-}3}$ of 179.7 kcal/mol, which is similar to the binding energy of avian Gln226 H3 with avian α2-3 (Table 9.1, entries 1, 2). This result supported the result of previous virus–sialoside binding assay.[32] With more-accurate studies, our theoretical approach may predict the infectious level of new viruses and point out as-yet-unknown dangerous mutation positions.

9.4.3 Interfragment Interaction Stabilization between Avian H3 and Sialoside Receptor

Figure 9.8 shows intra- and intermolecular interactions of avian H3-Neu5Acα(2-3 or 2-6)Gal complexes, and the interaction energies are summarized in Table 9.2. In this section, we discuss the molecular interactions with the corresponding energies at the FMO-MP2/6-31G level.

In the avian H3–avian α2-3 complex, side-chain NH_2CO on Gln226 forms an intermolecular hydrogen bond network with 8-OH and 1-COO on Neu5Ac, with Gal 4-OH, and weakly with glycoside oxygen O3, to give the interaction energy of 23.9 kcal/mol, which is 6.5 kcal/mol larger than the corresponding energy in avian H3–human α2-6 complex (Figure 9.8A,D; Table 9.2, entry 3). Furthermore, the formation of hydrogen bonds between Neu 1-COO and Ser136, Ser137, and Asn145 strongly stabilizes the avian H3–avian α2-3 complex, with the interaction energies of 27.1, 33.9, and 13.2 kcal/mol (Figure 9.8B,E; Table 9.2, entries 8 to 10). The last of these three interactions consists of an intermolecular hydrogen bond 2.78 Å

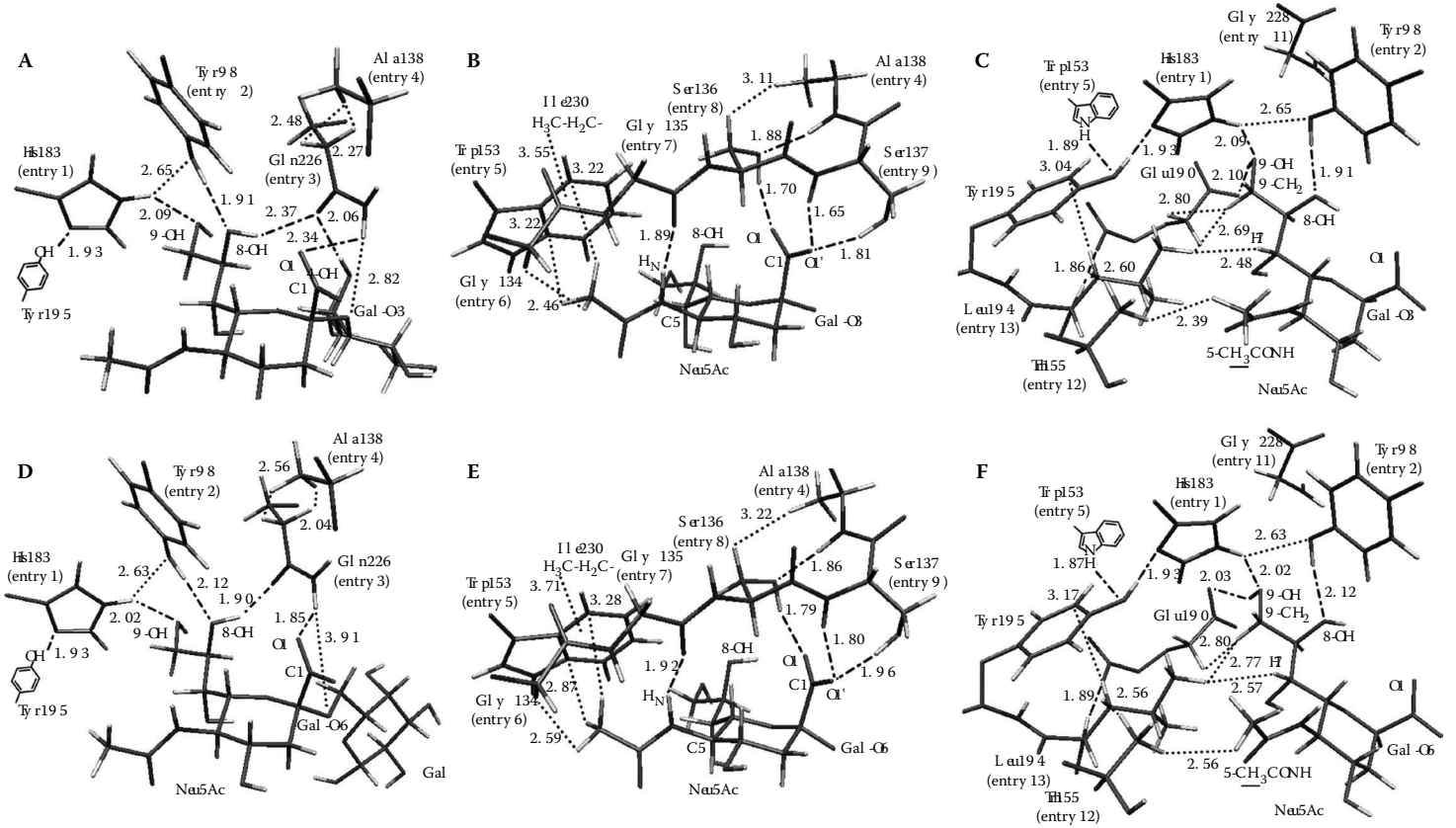


FIGURE 9.8 Intermolecular interactions of Neu5Acα(2-3 and 2-6)Gal with amino acid residues on the sialoside binding site in avian H3.[43] (A, B, C) Avian H3–avian α2-3 complex. (D, E, F) Avian H3–human α2-6 complex. The -·-·-· and ····· lines represent hydrogen bonds and long-range interactions whose distances are given in angstrom. (Reproduced from Sawada, T., Hashimoto, T., Tokiwa, H. et al. 2008. *Glycoconj. J.* 25:805–815. With permission.)

TABLE 9.2
Interaction Energies of Neu5Acα(2-3 and 2-6)Gal with Amino Acid Residues on the Sialoside Binding Site in Avian H3

			Interaction Energies: kcal/mol							
			Avian α2-3				Human α2-6			
	Avian H3	Interaction Site on Neu5Ac-Gal	RHF			MP2*[b]	RHF			MP2*[b]
Entry	Amino Acid	Avian α2-3 / Human α2-6	STO-3G*[1]	6-31G	6-31G(d)	6-31G	STO-3G	6-31G	6-31G(d)	6-31G
1	His183	Neu 8,9-OH	5.2	9.4	8.2	12.4	7.5	11.7	10.3	14.5
2	Tyr98	Neu 8-OH	6.7	11.2	9.9	14.6	5.1	10.6	9.9	14.2
3	Gln226	Neu 8-OH, 1-CO1O1′, Gal 4-OH / Neu 8-OH, 1-CO1O1′	8.8	18.1	15.5	23.9	10.5	12.7	11.6	17.4
4	Ala138	Neu 1-COlO1′	10.1	15.0	14.8	15.8	11.1	15.7	15.4	16.5
5	Trp153	Neu 5-$NHCOCH_3$	0.1	3.4	2.8	8.5	–0.4	2.7	2.1	8.6
6	Gly134	Neu 5-$NHCOCH_3$	–1.4	–1.8	–1.8	–1.3	–1.9	–2.3	–2.3	–1.8
7	Gly135	Neu 5-$NHCOCH_3$	1.9	1.5	1.3	2.6	2.7	2.6	2.3	3.9
8	Ser136	Neu 1-CO1O1′	20.5	21.1	16.6	27.1	16.2	17.3	12.9	23.2
9	Ser137	Neu 1-CO1O1′	36.5	29.1	30.1	33.9	27.7	23.6	25.8	28.7
10	Asn145	Neu 1-CO1O1′	6.6	12.1	11.5	13.2	10.7	16.1	14.9	18.2
11	Gly228	—	1.7	4.2	4.0	4.7	1.8	4.0	3.8	4.3
12	Thr155	Neu 5-$NHCOCH_3$	–1.8	–1.0	–0.9	0.1	–2.2	–1.5	–1.4	–0.3
13	Leu194	Neu 7-CH, 9-CH_2	–1.2	0.8	0.7	3.7	–1.8	–0.3	–0.2	2.5

[a] Sawada, T., Hashimoto, T., Nakano, H. et al. 2006. *Biochem. Biophys. Res. Commun.* 351:40–43.

[b] Sawada, T., Hashimoto, T., Tokiwa, H. et al. 2008. *Glycoconj. J.* 25:805–815.

between Asn145 side-chain $CONH_2$ and O1'O1C1 on Neu5Acα(2-3)Gal (not indicated in figures). Gly228 on influenza A virus H3 subtype correlates with the shift of the virus host range between avian and human. In our study, Gly228 interacts with Neu5Acα(2-3 or 2-6)Gal disaccharides with the interaction energies of 4.7 and 4.3 kcal/mol (Figure 9.8C,F; Table 9.2, entry 11).

In the X-ray crystal structure of avian H3 complexed with α(2-3 or 2-6)-pentasaccharide receptors, the sialoside binding site reliably recognizes the nonreducing terminal Neu5Ac-Gal[19]; this recognition serves as some justification of our theoretical approach.

9.4.4 Similarity in Manner of Intermolecular Interactions in the Human H3–Human α2-6 Complex and the Corresponding Avian Gln226Leu H3 Complex

$\Delta E_{\alpha 2\text{-}6}$ in the human H3–human α2-6 complex has a value similar to that in the avian Gln226Leu H3 complex (Table 9.1, entries 4 and 5), Here we compare their interaction manners and the corresponding interaction energies at the FMO-MP2/6-31G level.

Figure 9.9 shows molecular interactions of human Neu5Acα(2-6)Gal with amino acid residues on the sialoside binding site in human H3. Human H3–human α2-6 complex has an intramolecular hydrogen bond at Neu 8-OH···O1C1O1' instead of the intermolecular Gln226···Neu5Ac interaction observed in the avian H3 complex (Figure 9.8D and Figure 9.9A), and thus, $\Delta E_{\alpha 2\text{-}6}$ in the human H3 complex is smaller by 15 kcal/mol than $\Delta E_{\alpha 2\text{-}6}$ in the avian Gln226 H3 complex (Table 9.1, entries 3 and 5). Neu 1-COO interacts with Ser136 with the interaction energy of 29.3 kcal/mol, Asn137 with 45.5 kcal/mol, and Ser145 with 2.6 kcal/mol (Table 9.3, entries 8 to 10), the summation of these energies is larger by 7.3 kcal/mol than that of the avian H3–human α2-6 complex (Table 9.2, entries 8 to 10). Inter- and intramolecular lipophilic stabilization is significant for the human H3–human α2-6 binding. Leu226 interacts with Gal 6-CH_2 on human α2-6 by dispersion stabilization of 6.1 kcal/mol (Figure 9.9A; Table 9.3, entry 3). The Leu226···Gal 6-CH_2 association is supported by a lipophilic network beyond Leu226 consisting of Tyr98, Pro99, Ala138, $(CH_2)_3$ on Arg220 and Arg229, Ile230, and Trp153.

In the avian Gln226Leu H3–human α2-6 complex (Figure 9.10, left), Leu226 interacts with lipophilic Gal 6-CH_2 whose interaction energy is smaller by 1.3 kcal/mol than that of Leu226···Gal 6-CH_2 interaction in the human H3 complex (Table 9.3 and Table 9.4, entry 3). Neu 8-OH makes an intramolecular hydrogen bond with O1C1O1'-Neu in a similar manner to that in the human Leu226 H3 complex (Figure 9.9A and Figure 9.10, left). Thus, avian Gln226Leu H3 virus can infect humans at the same level as can human H3 virus.

In the human Leu226Gln H3–avian α2-3 complex, Gln226 makes intermolecular hydrogen bond network with 8-OH, 1-CO1O1' on Neu5Ac and Gal 4-OH to give the interaction energy 31.1 kcal/mol (Figure 9.10, right; Table 9.5, entry 3), and this interaction manner is similar to that of avian Gln226 H3–avian α2-3 complex (Figure 9.8A). We previously reported that A/Udorn Leu226Gln (human Leu226Gln

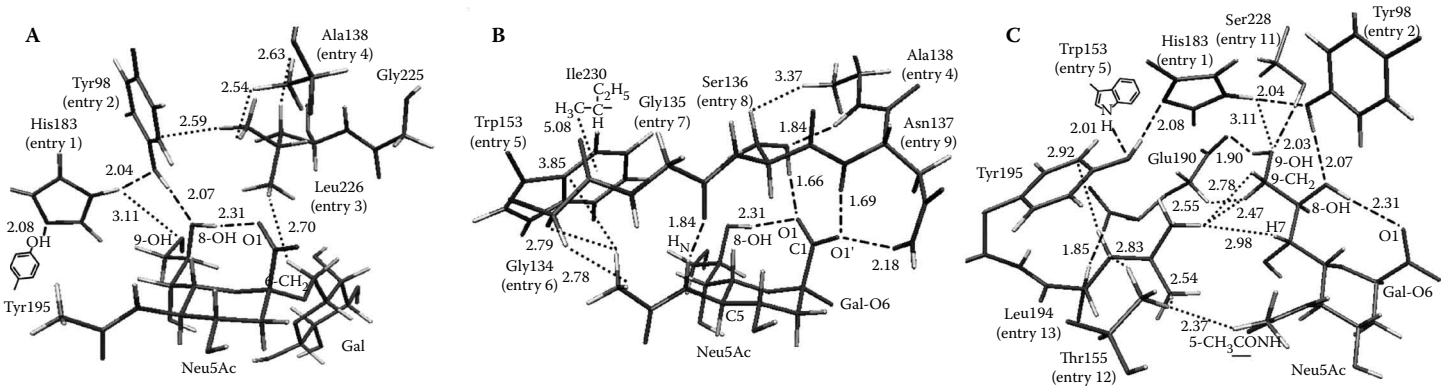


FIGURE 9.9 Intermolecular interactions of human Neu5Acα(2-6)Gal with amino acid residues on the sialoside binding site in human H3.[43] (Reproduced from Sawada, T., Hashimoto, T., Tokiwa, H. et al. 2008. *Glycoconj. J.* 25:805–815. With permission.)

TABLE 9.3
Interaction Energies of Human Neu5Acα(2-6)Gal with Amino Acid Residues on the Sialoside Binding Site in Human H3

			Interaction Energies: kcal/mol			
			RHF			MP2[a]
Entry	Human H3 Amino Acid	Interaction Sites on Human α2-6	STO-3G	6-31G	6-31G(d)	6-31G
1	His183	Neu 8,9-OH	3.5	7.6	7.2	8.9
2	Tyr98	Neu 8-OH	6.4	13.0	11.7	16.2
3	Leu226	Gal 6-CH_2	–1.3	2.0	1.5	6.1
4	Ala138	Neu 1-CO1O1′	8.6	14.8	14.7	15.7
5	Trp153	Neu 5-$NHCOCH_3$	–0.2	2.2	1.8	4.0
6	Gly134	Neu 5-$NHCOCH_3$	–1.2	–1.3	–1.4	–0.9
7	Gly135	Neu 5-$NHCOCH_3$	1.0	–0.2	–0.2	1.1
8	Ser136	Neu 1-CO1O1′	24.2	23.4	19.5	29.3
9	Asn137	Neu 1-CO1O1′	38.5	40.0	39.5	45.5
10	Ser145	Neu 1-CO1O1′	0.0	2.4	2.2	2.6
11	Ser228	Neu 9-OH	6.8	13.2	11.3	14.8
12	Thr155	Neu 5-$NHCOCH_3$	–1.6	–1.1	–1.0	–0.3
13	Leu194	Neu 7-CH, 9-CH_2	–2.0	–0.7	–0.8	2.4

[a] Sawada, T., Hashimoto, T., Tokiwa, H. et al. 2008. *Glycoconj. J.* 25:805–815.

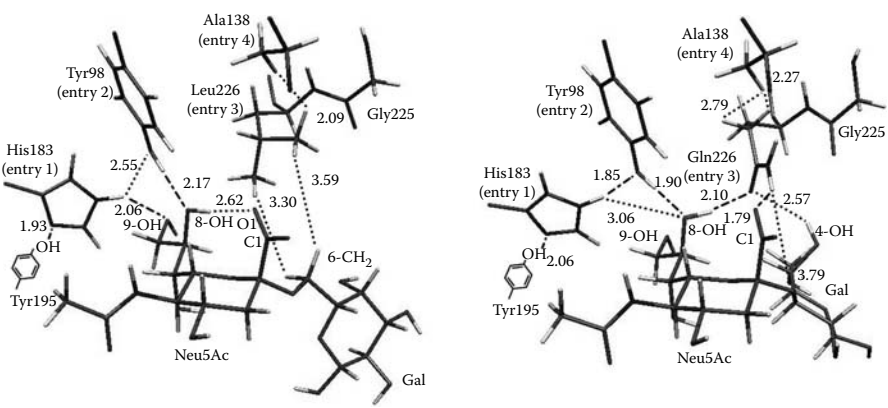


FIGURE 9.10 Intermolecular interactions of Neu5Acα(2-6 and 2-3)Gal with amino acid residues on the sialoside binding site in avian Gln226Leu and human Leu226Gln H3s.[43] Left: avian Gln226Leu H3-human α2-6 complex. Right: human Leu226Gln H3-avian α2-3 complex. (Reproduced from Sawada, T., Hashimoto, T., Tokiwa, H. et al. 2008. *Glycoconj. J.* 25:805–815. With permission.)

TABLE 9.4
Interaction Energies of Human Neu5Acα(2-6)Gal with Amino Acid Residues on the Sialoside Binding Site in Avian Gln226Leu H3

Entry	Avian Q226L H3 Amino Acid	Interaction Sites on Human α2-6	Interaction Energies: kcal/mol MP2/6-31G
1	His183	Neu 8,9-OH	14.3
2	Tyr98	Neu 8-OH	14.5
3	Leu226	Gal 6-CH_2	4.8
4	Ala138	Neu 1-CO1O1′	16.5
5	Trp153	Neu 5-$NHCOCH_3$	8.6
6	Gly134	Neu 5-$NHCOCH_3$	–1.6
7	Gly135	Neu 5-$NHCOCH_3$	3.6
8	Ser136	Neu 1-CO1O1′	25.0
9	Ser137	Neu 1-CO1O1′	26.2
10	Asn145	Neu 1-CO1O1′	18.1
11	Gly228	—	4.4
12	Thr155	Neu 5-$NHCOCH_3$	–0.2
13	Leu194	Neu 7-CH, 9-CH_2	2.7

Source: Reproduced from Sawada, T., Hashimoto, T., Tokiwa, H. et al. 2008. *Glycoconj. J.* 25:805–815. (With permission.)

TABLE 9.5
Interaction Energies of Avian Neu5Acα(2-3)Gal with Amino Acid Residues on the Sialoside Binding Site in Human Leu226Gln H3

Entry	Human L226Q H3 Amino Acid	Interaction Sites on Avian α2-3	Interaction Energies: kcal/mol MP2/6-31G
1	His183	Neu 8,9-OH	8.6
2	Tyr98	Neu 8-OH	15.6
3	Gln226	Neu 8-OH, 1-CO1O1′, Gal 4-OH	31.3
4	Ala138	Neu 1-CO1O1′	15.2
5	Trp153	Neu 5-$NHCOCH_3$	7.9
6	Gly134	Neu 5-$NHCOCH_3$	–1.3
7	Gly135	Neu 5-$NHCOCH_3$	2.6
8	Ser136	Neu 1-CO1O1′	27.4
9	Asn137	Neu 1-CO1O1′	49.1
10	Ser145	Neu 1-CO1O1′	3.1
11	Ser228	Neu 9-OH	14.3
12	Thr155	Neu 5-$NHCOCH_3$	–0.4
13	Leu194	Neu 7-CH, 9-CH_2	1.6

Source: Reproduced from Sawada, T., Hashimoto, T., Tokiwa, H. et al. 2008. *Glycoconj. J.* 25:805–815. (With permission.)

H3N2) bound to avian α2-3 receptor.[32] Our FMO studies confirmed that human Leu226Gln H3 moderately interacts with avian α2-3 analogue at the correlated FMO-MP2/6-31G level.

9.5 HOW MUCH OF THE SIZE OF THE HA1-SIALOSIDE COMPLEX MODEL SHOULD WE TREAT UNDER FMO STUDIES?

9.5.1 Influenza Viral HA1 Takes a Complicated Shape beyond the Sialoside Binding Site

We usually select limited atoms on the protein–ligand complex for quantum mechanics calculation with matching our aims. However, the adequacy of the clipping area with arbitrariness has not been theoretically understood.

Influenza HA1 takes a complicated shape beyond the sialoside binding site as well as other carbohydrate-binding proteins (Figure 9.7a). How much advantage does the shape have to the HA-sialoside binding? Dynamic fluctuations of protein shape govern its behavior.[54,55] Is the advantage of the backyard complexity only this one? Disgustingly, amino acid substitutions on influenza HA far away from the sialoside binding sites often change the sialoside binding specificities. Do the substitutions easily cause any structural changes to the binding site without conformational relaxation by backyard fluctuation?

The amino acid substitutions may alter not only the binding site conformations but also intra- and intermolecular stabilizations in the HA1-sialoside complex. FMO studies allow us to investigate the later topic. Recently, Iwata et al. compared the intra- and intermolecular stabilizations in the HA–sialosaccharide complexes (82 amino acids) at the FMO-MP2/6-31G level.[56] These stabilizations will be altered along with the difference of model sizes. How much size of HA1-sialoside complex model should we treat under FMO studies? Do we always require the full-size model of HA1 monomer–sialoside complex? In this section, we demonstrate that ΔE between human H3 HA1 and human Neu5Acα(2-6)Gal disaccharide substantially increases as the model complex becomes larger at the FMO-RHF/STO-3G level.

9.5.2 Computational Methods

Optimum human H3–Neu5Acα(2-6)Gal complex for FMO studies was prepared by molecular mechanics energy calculation described above. We clipped three different sizes of the sialoside-binding domains **A**-**C**, full-size HA1 domain **D**, and the modified shape **B-A** in order to compare their binding energies (Figure 9.11). We computed the single-point energies of the clipped complexes (E_{complex}), corresponding H3 receptor–binding domains (E_{H3}), Neu5Acα(2-6)Gal (E_{receptor}), and then estimated ΔE in the complexes at the FMO-RHF/STO-3G level with the same methodology using ABINIT-MP program.

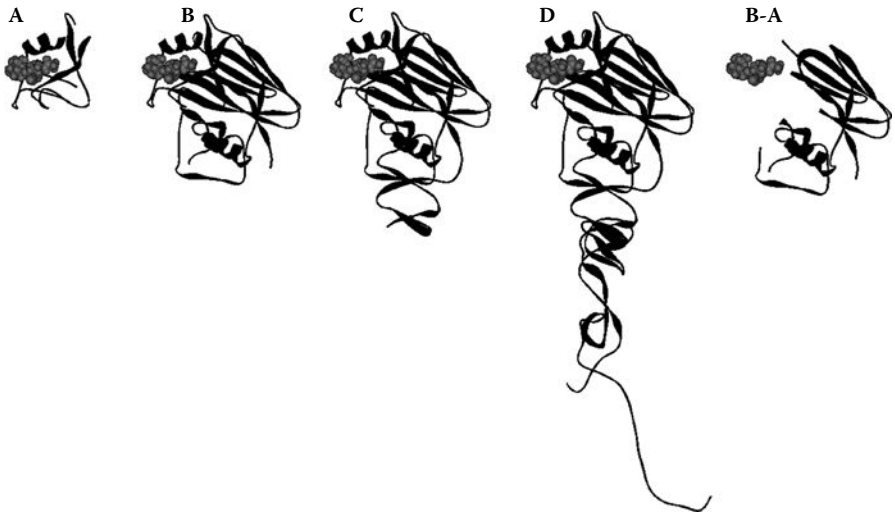


FIGURE 9.11 The complexes of human H3 sialoside binding domain HA1 with human Neu5Acα(2-6)Gal for the binding energy calculations.[42] **A:** The complex has the smallest receptor binding domain (70 amino acids, Figure 9.6A, right). **B:** Binding domain in the complex consists of Ile62-Gly263 (202 amino acids). **C:** The domain involves Gly49-Thr283 (235 amino acids). **D**: The complex has the full size of receptor binding domain (Gln1-Thr328). **B–A:** The modified complex is prepared by cutting out the binding domain in complex **A** from complex **B** (132 amino acids). The complex has no amino acid residues that directly interact with Neu5Acα(2-6)Gal. (Reproduced from Sawada, T., Hashimoto, T., Nakano, H. et al. 2007. *Biochem. Biophys. Res. Commun.* 355:6–9. With permission.)

9.5.3 Interaction between HA and Sialoside Increase as the Model Complexes Become Larger

The binding energies in the complexes are summarized in Table 9.6. The full shape of the H3 HA1 domain benefits the binding between H3 and the human α2-6 receptor. Full-size complex **D** had the highest binding energy $\Delta E_{\mathbf{D}}$ 170.3 kcal/mol, which

TABLE 9.6
Binding Energies between Human H3 and Human Neu5Acα(2-6) Gal in the Models at the FMO-RHF/STO-3G Level

	Human H3-Neu5Acα(2-6)Gal Complex				
	A	**B**	**C**	**D**	**B-A**
Binding energy ΔE; kcal/mol	118.0	158.0	144.6	170.3	−1.3

Source: Reproduced from Sawada, T., Hashimoto, T., Nakano, H. et al. 2007. *Biochem. Biophys. Res. Commun.* 355:6–9. (With permission.)

was 52.3 kcal/mol larger than that of the smallest domain complex **A**. $\Delta E_{\mathbf{B}}$ in the complex **B** was also 40.0 kcal/mol larger than that of complex **A**. There was hardly any binding energy in the modified complex **B-A** with luck of the direct interaction site to human α2-6. These results indicate that amino acid residues far away from the sialoside binding site on the HA1 domain are not interacting with human α2-6 disaccharide, but they can affect the binding energy via the direct interaction site. Comparison of complex **B** with **D**, the latter had more than 126 amino acids than the former that gave an advantage to sialoside binding by $\Delta E_{\mathbf{D}}$–$\Delta E_{\mathbf{B}}$ 12.3 kcal/mol. Since $\Delta E_{\mathbf{D}}$–$\Delta E_{\mathbf{B}}$ was 27.7 kcal/mol smaller than $\Delta E_{\mathbf{B}}$–$\Delta E_{\mathbf{A}}$ 40.0 kcal/mol, amino acids near the binding site strongly affected the binding energy. $\Delta E_{\mathbf{C}}$ in the model **C** was smaller than $\Delta E_{\mathbf{B}}$ by 13.4 kcal/mol, and thus, larger backyard bulk beyond the sialoside binding site does not increase monotonously the binding energy. The bulk effect plausibly causes intramolecular stabilization in the HA1 domain and will influence the difference of HA-Neu5Acα(2-3 or 2-6)Gal binding specificities.

This approach has been applied to the complexes of avian H3 HA1 full domain with Neu5Acα(2-3 and 2-6)Gal disaccharides in order to compare the corresponding $\Delta E_{\alpha 2\text{-}3}$, $\Delta E_{\alpha 2\text{-}6}$, and $\Delta E_{\alpha 2\text{-}3}$– $\Delta E_{\alpha 2\text{-}6}$ at the FMO-MP2/6-31G level, then we have confirmed the tendency that Δ*E*s increase as the model complexes become larger. Avian H3 binds to avian α2-3 16.5 kcal/mol more strongly than to human α2-6 in the HA1 full model (Ser9-Lys326; 318 residues, Figure 9.7). The details will soon be reported elsewhere.

9.6 CONCLUSION AND PERSPECTIVES

We applied *ab initio*–based FMO method to study influenza A virus hemagglutinin H3 subtype complexed with avian/human-type receptor at the correlated MP2/6-31G level. Avian H3 bound to avian α2-3 11.4 kcal/mol stronger than to human α2-6 in the model complexes with taking account of intermolecular dispersion interaction. Single-point substitution at position 226 on H3 subtype sialoside binding site HA1 changes its binding specificity between avian α2-3 and human α2-6. The binding energy of avian Gln226Leu H3 with human α2-6 was similar value to that of the human H3–human α2-6 complex at the FMO-MP2/6-31G level with amino acid differences at nine positions in our models. The reverse mutation Leu226Gln in the human H3 gave the moderate binding energy to avian α2-3 that supported our previous virus-sialoside binding assay. Because the performance of the computer has improved rapidly, we are able to calculate the binding energy of HA and sialoside receptor covering all the HA mutants. Our theoretical approach may predict the infectious level of new viruses and point out some unknown dangerous mutation positions on HA in advance against human pandemic influenza. In perspective, the solvated HA-sialoside system will be studied by molecular dynamics simulation, and then we will investigate how much model size of HA-sialoside complex we should treat with the FMO/PCM method.[57] We already calculated the complex of human H3 trimer with human-type Neu5Acα(2-6)Gal analogues at the FMO-RHF/6-31G level to find HA trimerization effects for the sialoside binding. You should select carefully a suitable model complex for your purpose. Mochizuki et al. reported large-scale FMO-MP2

calculations of influenza viral HA in complex with a specific antibody[58]; thus, it is not impossible to study the antigenic drift of influenza viral proteins.

ACKNOWLEDGMENTS

This work was supported in part by CREST, Japan Science and Technology Agency (JST), and a Grant-in-Aid (No.17101007 to M. Kiso) for Scientific Research from Japan Society for the Promotion of Science.

REFERENCES

1. de Jong, J.C., Claas, E.C.J., Oseterhaus, A.D.M.E., Webster, R.G., Lim, W.L. 1997. A pandemic warning? *Nature* 389:554.
2. Subbarao, K., Klimov, A., Katz, J. et al. 1998. Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness. *Science* 279:393–396.
3. Claas, E.C., Osterhaus, A.D., van Beek, R. et al. 1998. Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *Lancet* 351:472–477.
4. Bender, C., Hall, H., Huang, J. et al. 1999. Characterization of the surface proteins of influenza A (H5N1) viruses isolated from humans in 1997–1998. *Virology* 254:115–123.
5. Wong, S.S.Y., Yuen, K.-Y. 2006. Avian influenza virus infections in humans. *Chest* 129:156–168.
6. Shinya, K., Ebina, M., Yamada, S. et al. 2006. Avian flu: influenza virus receptors in the human airway. *Nature* 440:435–436.
7. van Riel, D., Munster, V.J., de Wit, E. et al. 2006. H5N1 virus attachment to lower respiratory tract. *Science* 312:399.
8. Yamada, S., Suzuki, Y., Suzuki, T. et al. 2006. Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human type receptors. *Nature* 444:378–382.
9. Auewarakul, P., Suptawiwat, O., Kongchanagul, A. et al. 2007. An avian influenza H5N1 virus that binds to a human-type receptor. *J. Virol.* 81:9950–9955.
10. Stevens, J., Blixt, O., Tumpey, T.M. et al. 2006. Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. *Science* 312:404–410.
11. Yang, Z.-Y., Wei, C.-J., Kong, W.-P. et al. 2007. Immunization by avian H5 influenza hemagglutinin mutants with altered receptor binding specificity. *Science* 317:825–828.
12. Bottcher, C., Ludwig, K., Herrmann, A., van Heel, M., Stark, H. 1999. Structure of influenza haemagglutinin at neutral and at fusogenic pH by electron cryo-microscopy. *FEBS Lett.* 463:255–259.
13. Horimoto, T., Kawaoka, Y. 2005. Influenza: lessons from past pandemics, warnings from current incidents. *Nat. Rev. Microbiol.* 3:591–600.
14. Noda, T., Sagara, H., Yen, A. et al. 2006. Architecture of ribonucleoprotein complexes in influenza A virus particles. *Nature* 439:490–492.
15. Suzuki, Y. 2005. Sialobiology of influenza: molecular mechanism of host range variation of influenza viruses. *Biol. Pharm. Bull.* 28:399–408.
16. Matrosovich, M.N., Klenk, H.-D., Kawaoka, Y. 2006. Receptor specificity, host range, and pathogenicity of influenza viruses. In *Influenza virology: current topics,* ed. Y. Kawaoka, 95–137. Caister Academic Press: Wymondham, UK.

17. Skehel, J.J., Wiley, D.C. 2000. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annu. Rev. Biochem.* 69:531–569.
18. Ha, Y., Stevens, D.J., Skehel, J.J., Wiley, D.C. 2001. X-ray structures of H5 avian and H9 swine influenza virus hemagglutinins bound to avian and human receptor analogs. *Proc. Natl. Acad. Sci. USA* 98:11181–11186.
19. Ha, Y., Stevens, D.J., Skehel, J.J., Wiley, D.C. 2003. X-ray structure of the hemagglutinin of a potential H3 avian progenitor of the 1968 Hong Kong pandemic influenza virus. *Virology* 309:209–218.
20. Gamblin, S.J., Haire, L.F., Russell, R.J. et al. 2004. The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science* 303:1838–1842.
21. Russell, R.J., Stevens, D.J., Haire, L.F., Gamblin, S.J., Skehel, J.J. 2006. Avian and human receptor binding by hemagglutinins of influenza A viruses. *Glycoconj. J.* 23:85–92.
22. Stevens, J., Blixt, O., Tumpey, T.M. et al. 2006. Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. *Science* 312:404–410.
23. Stevens, J., Blixt, O., Paulson, J.C., Wilson, I.A. 2006. Glycan microarray technologies: tools to survey host specificity of influenza viruses. *Nat. Rev. Microbiol.* 4:857–864.
24. Rogers, G.N., Paulson, J.C., Daniels, R.S. et al. 1983. Single amino acid substitutions in influenza haemagglutinin change receptor binding specificity. *Nature* 304:76–78.
25. Weis, W., Brown, J.H., Cusack, S. et al. 1988. Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. *Nature* 333:426–431.
26. Connor, R.J., Kawaoka, Y., Webster, R.G., Paulson, J.C. 1994. Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates. *Virology* 205:17–23.
27. Gambaryan, A.S., Piskarev, V.E., Yamskov, I.A. et al. 1995. Human influenza virus recognition of sialyloligosaccharides. *FEBS Lett.* 366:57–60.
28. Gambaryan, A.S., Tuzikov, A.B., Piskarev, V.E. et al. 1997. Specification of receptor-binding phenotypes of influenza virus isolates from different hosts using synthetic sialylglycopolymers: nonegg-adapted human H1 and H3 influenza A and influenza B viruses share a common high binding affinity for 6′-sialyl (N-acetyllactosamine). *Virology* 232:345–350.
29. Ito, T., Suzuki, Y., Mitnaul, L. et al. 1997. Receptor specificity of influenza A viruses correlates with the agglutination of erythrocytes from different animal species. *Virology* 227:493–499.
30. Nobusawa, E., Ishihara, H., Morishita, T., Sato, K., Nakajima, K. 2000. Change in receptor-binding specificity of recent human influenza A viruses (H3N2): a single amino acid change in hemagglutinin altered its recognition of sialyloligosaccharides. *Virology* 278:587–596.
31. Kumari, K., Gulati, S., Smith, D.F. et al. 2007. Receptor binding specificity of recent human H3N2 influenza viruses. *Virol. J.* 4:42.
32. Ito, T., Suzuki, Y., Suzuki, T. et al. 2000. Recognition of N-glycolylneuraminic acid linked to galactose by the α2,3 linkage is associated with intestinal replicaton of influenza A virus in ducks. *J. Virol.* 74:9300–9305.
33. Kameyama, Y., Ishida, H., Kiso, M., Hasegawa, A. 1990. Synthetic studies on sialylglycoconjugates 13. stereoselective synthesis of sialyl-lactotetraosylceramide and sialyl-neolactotetraosylceramide. *Carbohydr. Res.* 200:269–285.
34. Hasegawa, A., Hotta, K., Kameyama, A., Ishida, H., Kiso, M. 1991. Synthetic studies on sialoglycoconjugates 23. total synthesis of sialyl-alpha(2-6)-lactotetraosylceramide and sialyl-alpha(2-6)-neolactotetraosylceramide. *J. Carbohydr. Chem.* 10:439–459.
35. Fukunaga K., Toyoda T., Ishida H., Kiso M. 2003. Synthesis of lacto- and neolacto-series ganglioside analogs containing N-glycolylneuraminic acid: probes for investigation of specific receptor structures recognized by influenza A viruses. *J. Carbohydr. Chem.* 22:919–937.

36. Suzuki, Y., Nakano, T., Ito, T. et al. 1992. Structual determination of gangliosides that bind to influenza A, B, and C viruses by an improved binding assay: strain-specific receptor epitopes in sialo-sugar chains. *Virology* 189:121–131.
37. Masuda, H., Suzuki, T., Sugiyama, Y. et al. 1999. Substitution of amino acid residue in influenza A virus hemagglutinin affects recognition of sialyl-oligosaccharides containing N-glycolylneuraminic acid. *FEBS Lett.* 464:71–74.
38. Shinya, K., Hatta, M., Yamada, S. et al. 2005. Characterization of a human H5N1 influenza A virus isolated in 2003. *J. Virol.* 79, 9926–9932.
39. Rogers, G.N., Daniels, R.S., Skehel, J.J. et al. 1985. Host-mediated selection of influenza virus receptor variants. Sialic acid-alpha 2, 6 Gal-specific clones of A/duck/Ukraine/1/63 revert to sialic acid-alpha 2,3Gal-specific wild type in ovo. *J. Biol. Chem.* 260:7362–7367.
40. Naeve, C.W., Hinshaw, V.S., Webster, R.G. 1984. Mutations in the hemagglutinin receptor-binding site can change the biological properties of an influenza virus. *J. Virol.* 51:567–569.
41. Sawada, T., Hashimoto, T., Nakano, H. et al. 2006. Why does avian influenza A virus hemagglutinin bind to avian receptor stronger than to human receptor? *Ab initio* fragment molecular orbital studies. *Biochem. Biophys. Res. Commun.* 351:40–43.
42. Sawada, T., Hashimoto, T., Nakano, H., et al. 2007. Influenza viral hemagglutinin complicated shape is advantageous to its binding affinity for sialosaccharide receptor. *Biochem. Biophys. Res. Commun.* 355:6–9.
43. Sawada, T., Hashimoto, T., Tokiwa, H. et al. 2008. *Ab initio* fragment molecular orbital studies of influenza virus hemagglutinin-sialosaccharide complexes toward chemical clarification about the virus host range determination. *Glycoconj. J.* 25:805–815.
44. von der Lieth, C.-W., Kozar, T. 1996. Towards a better semiquantitative estimation of binding constants: molecular dynamics exploration of the conformational behavior of isolated sialyllactose and sialyllactose complexed with influenza A hemagglutinin. *J. Mol. Struct. THEOCHEM.* 368:213–222.
45. Frank, M., von der Lieth, C.-W. 1997. Comparison of the conformational behavior of sialyllactose complexed with the two viral attachment proteins influenza A hemagglutinin and the Murine Polyomavirus. *J. Mol. Model.* 3:408–414.
46. Li, M., Wang, B. 2006. Computational studies of H5N1 hemagglutinin binding with SA-α-2, 3-Gal and SA-α-2, 6-Gal. *Biochem. Biophys. Res. Commun.* 347:662–668.
47. Chandrasekaran, A., Srinivasan, A., Raman, R., et al. 2008. Glycan topology determines human adaptation of avian H5N1 virus hemagglutinin. *Nat. Biotechnol.* 26:107–113.
48. Maple, J.R., Hwang, M.J., Jalkanen, K.J., Stockfisch, T.P., Hagler, A.T. 1998. Derivation of class II force fields: quantum force field for amides, peptides, and related compounds. *J. Comp. Chem.* 19:430–458.
49. Eisen, M.B., Sabesan, S., Skehel, J.J., Wiley, D.C. 1997. Binding of the influenza A virus to cell-surface receptors: structures of five hemagglutinin-sialyloligosaccharide complexes determined by x-ray crystallography. *Virology* 232:19–31.
50. Sauter, N.K., Hanson, J.E., Glick, G.D. et al. 1992. Binding of influenza virus hemagglutinin to analogs of its cell-surface receptor, sialic acid: analysis by proton nuclear magnetic resonance spectroscopy and x-ray crystallography. *Biochemistry* 31:9609–9621.
51. Nakano, T., Mochizuki, Y., Fukuzawa, K., Amari, S., Tanaka, S. 2006. Developments and applications of ABINIT-MP software based on the fragment molecular orbital method. In *Modern methods for theoretical physical chemistry of biopolymers,* ed. E.B. Starikov, J.P. Lewis, and S. Tanaka, 39–52. Amsterdam: Elsevier.
52. Spiwok, V., Lipovová, P., Skálová, T. et al. 2004. Role of CH/π interactions in substrate binding by *Escherichia coli* β-galactosidase. *Carbohydr. Res.* 339:2275–2280.

53. Fernández-Alonso, M.D.C., Cañada, F.J., Jiménez-Barbero, J., Gabriel Cuevas, G. 2005. Molecular recognition of saccharides by proteins. Insights on the origin of the carbohydrate–aromatic interactions. *J. Am. Chem. Soc.* 127:7379–7386.
54. Boehr, D.B., McElheny, D., Dyson, H.J., Wright, P.E. 2006. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* 313:1638–1642.
55. Vendruscolo, M., Dobson, C.M. 2006. Dynamic visions of enzymatic reactions. *Science* 313:1586–1587.
56. Iwata, T., Fukuzawa, K., Nakajima, K. et al. 2008. Theoretical analysis of binding specificity of influenza viral hemagglutinin to avian and human receptors based on the fragment molecular orbital method. *Comput. Biol. Chem.* 32:198–211.
57. Fedorov, D.G., Kitaura, K., Li, H., Jensen, J.H., Gordon, M.S. 2006. The polarizable continuum model (PCM) interfaced with the fragment molecular orbital method (FMO). *J. Comput. Chem.* 27: 976–985.
58. Mochizuki, Y., Yamashita, K., Murase, T. et al. 2008. Large scale FMO-MP2 calculations on a massively parallel-vector computer. *Chem. Phys. Lett.* 457:396–403.

10 FMO as a Tool for Structure-Based Drug Design

Tomonaga Ozawa, Kosuke Okazaki, and Motohiro Nishio

CONTENTS

10.1 A BRIEF REVIEW OF WEAK MOLECULAR INTERACTION

10.1.1 Introduction

Weak molecular forces are important in controlling the three-dimensional structure of proteins and nucleic acids and play vital roles in regulating biochemical processes. Among noncovalent molecular forces, the hydrogen bond[1] is one of the most abundant, and its energy is within the range of 3 to 7 kcal/mol. In the conventional hydrogen bond, Coulomb energy is the most important energy source because this is the interaction between a hard acid (HA) and a hard base (HB) in the context of the Pearson hard soft acid base (HSAB) principle.

In the last half of the twentieth century, evidence has accumulated to show that weaker hydrogen bonds are also ubiquitous, including XH/π (X = O, N) and CH/*n* interactions (*n*: lone pair electrons in contrast to π). The former includes hydrogen bonds between HA and a soft base (SB) (2 to 4 kcal/mol), while the latter encompasses hydrogen bonds between a soft acid (SA) and a HB (2 to 4 kcal/mol). More recently, a still weaker attractive force, the CH/π hydrogen bond (0.5 to 2 kcal/mol), has been shown to play an important role in a variety of chemical and biological phenomena. This is a hydrogen bond occurring between an SA and an SB. Stabilization of the CH/π hydrogen bond comes mostly from the dispersion force, while contribution from the electrostatic and charge-transfer interaction is relatively negligible. The HA/SB and SA/HB hydrogen bonds fall between these two extremes. Table 10.1 compares the energy components of the four hydrogen bonds.

Note that the electrostatic contribution gradually decreases from the ordinary hydrogen bond to CH/O, NH/π, and then to the CH/π hydrogen bond. Interested readers are referred to the following monographs aimed at these unconventional hydrogen bonds: *The Weak Hydrogen Bond in Structural Chemistry and Biology* by Desiraju and Steiner[2] and *The CH/π Interaction: Evidence, Nature, and Consequences* by Nishio, Hirota, and Umezawa.[3] The former is a thorough treatise on the weak hydrogen bond, dealing largely with crystallographic evidence and the consequences of such bonding. The latter is aimed at the CH/π hydrogen bond and deals with methods of detection and aspects of this molecular force in chemistry and structural biology. Steiner wrote a review to provide a survey of the hydrogen

TABLE 10.1
Four Types of Hydrogen Bonds

Type of H-Bond		Example	E_{Total}	E_{ES}	E_{ER}	E_{Other}	E_{DISP}	E_{ES}/E_{Total}
Ordinary H-bond	HA/HB	H_2O/H_2O	–5.38	–9.12	7.43	–1.52	–1.90	1.70
CH/*n*	SA/HB	C_2H_2/H_2O	–3.54	–4.64	3.25	–0.74	–1.22	1.31
XH/π	HA/SA	NH_3/C_6H_6	–2.21	–1.00	1.14		–2.35	0.45
CH/π	SA/SB	CH_4/C_6H_6	–1.45	–0.25	1.10		–2.30	0.17

Notes: E_{Total}: Total interaction energy (in kcal/mol); E_{ES}: electrostatic energy (in kcal/mol); E_{ER}: exchange repulsion (in kcal/mol); E_{Other}: other energy (in kcal/mol); and E_{DISP}: dispersion energy (in kcal/mol).

bonds, with a focus on the structure of the solid on the grounds of modern theory and technology.[4] A more recent discourse dealing with the weak molecular interaction in organic chemistry and biological science recently appeared.[5] Another source of information is a literature list at one author's Web site (www.tim.hi-ho.ne.jp/dionisio). The list is focused on the CH/π hydrogen bond and is updated regularly. The topics are categorized into several parts such as host/guest chemistry, biochemistry, structural biology, theoretical calculations, database analyses, and drug design.

Weak molecular interactions have been studied by various methods. Database analyses and theoretical calculations are briefly described. The Cambridge Structural Database (CSD) (www.ccdc.cam.ac.uk/), the Protein Data Bank (PDB) (www.rcsb.org/pdb/home/home.do), and the Nucleic Acid Database (NDB) (http://ndbserver.rutgers.edu/) were used. The CSD is equipped with ConQuest for data retrieval, VISTA for statistical analyses, and Mercury for data analyses and presentation. The January 2008 release includes more than 410,000 crystal structures of organic, organometallic, and coordination compounds. The PDB and NDB are free for use and contain more than 44,000 protein and 3,500 nucleic acid coordinates, respectively, based on crystallographic and nuclear magnetic resonance (NMR) determinations.

Ab initio molecular orbital (MO) calculations were used.[6] Correction to BSSE (basis set superposition error) was necessary for supramolecules, and MP2 or higher-level approximations were employed. This was especially true for weaker hydrogen bonds such as CH/π or CH/N, because the contribution from the correlation energy is important. The density functional theory (DFT) method was used in calculating relatively large complexes. To analyze the contribution of energetic components, Kitaura–Morokuma partitioning[7] or alternatives such as the IMPT[8] (intermolecular perturbation theory) method were applied. Bader's AIM (atoms in molecules) method[9] was used to investigate the nature of the interaction.

10.1.2 XH/π Hydrogen Bond

The XH/π hydrogen bond includes OH/π and NH/π hydrogen bonds. These bonds were actively studied from the mid-1950s to the 1960s by infrared (IR) spectroscopy. For typical OH/π and NH/π hydrogen bonds, the hydrogen bond shift $\Delta\nu$ falls between 20 and 100 cm^{-1}. Perutz suggested the role of aromatic rings as hydrogen-bond acceptors,[10] and MO calculations support this conjecture.[11] Steiner and Koellner discussed the role of NH/π and OH/π hydrogen bonds in 592 high-resolution protein structures.[12] NH/π hydrogen bonds are more frequently found than OH/π hydrogen bonds, presumably because NH is a softer acid than OH. OH is likely harder as an acid and seeks a hard base as the partner in the physiological environment.

10.1.3 CH/*n* Hydrogen Bond

Thermochemical studies and measurements of CH stretching bands of the CH/O and CH/N complexed species suggest the capability of CH groups as hydrogen donors. The lower frequency shifts $\Delta\nu$ are on the order of 10 to 100 cm^{-1} depending on the system examined. The C–H absorption bands often shift to higher frequencies in CH/O and CH/π hydrogen bonds. In 1962, Sutor proposed the hydrogen-bond nature

TABLE 10.2
Comparison of the H···X Distance (Å) in CH/X Hydrogen Bonds

	O-Acceptor	N-Acceptor	Cl-Acceptor
$CHCl_3$	2.31	2.37	2.38
CH_2Cl_2	2.40	2.40	2.56
Sp-CH	2.49	2.53	2.59
CH_3CN	2.57	2.59	2.72
$(CH_3)_2SO$	2.56	2.65	2.94
$(CH_3)_2CO$	2.60	2.64	2.98

of the CH/O interaction based on her crystallographic data. In 1982, Taylor and Kennard presented unequivocal evidence for the CH/*n* hydrogen bond by surveying crystal structures in the CSD.[13] Thus, retrieval of neutron diffraction data and statistical analyses on the distance and angle parameters established unequivocally the hydrogen-bond nature of the CH/O, CH/N, and CH/halogen interactions. Desiraju, Steiner, and coworkers extensively studied CH/O interactions in crystals, and their results appear in several review articles.[14] Table 10.2 compares the H···X distance in CH/X hydrogen bonds examined by a CSD study.[15] Note that the distance depends on the strength of both the CH donor and acceptor in the interacting system. Houk, Stoddart, and their coworkers reported the role of CH/O interactions as a control element in supramolecular complexes[16] Database studies demonstrated the importance of CH/O hydrogen bonds in the protein structure,[17] and recent high-level *ab initio* MO calculations have also been performed.

Table 10.3 compares the strength of CH/O hydrogen bonds calculated for methane and its fluorinated derivatives in complex with water.[18] Note that the proportion of the electrostatic term is comparable to the ordinary hydrogen bond, whereas the contribution from the correlation energy is negligible. Acetylene, ethylene, and ethane

TABLE 10.3
Energy of CH/O Hydrogen Bond (in kcal/mol) Calculated for Methane and Its Fluorinated Derivatives in Complex with Water

	E_{Total}	E_{ES}	E_{ER}	E_{POL}	E_{CT}	E_{DISP}	E_{ES}/E_{Total}
CH_4/H_2O	–0.36	–0.42	0.38	–0.13	–0.11	–0.08	1.17
CH_3F/H_2O	–1.55	–1.96	1.17	–0.24	–0.32	–0.20	1.26
CH_2F_2/H_2O	–2.91	–3.83	2.06	–0.36	–0.53	–0.25	1.31
CHF_3/H_2O	–4.83	–7.06	4.14	–0.69	–0.97	–0.25	1.46
H_2O/H_2O	–5.28	–7.58	4.24	–0.71	–0.93	–0.30	1.44

Notes: MP2/6-31+G**.E_{Total}: total interaction energy; EES: electrostatic energy; EER: exchange repulsion; EPOL: polarization energy; E_{CT}: charge transfer; and E_{DISP}: dispersion energy.

have been shown to interact favorably with H_2O and NH_3.[19] Interaction of CHs with N or halogen is important[20] but is less common than the CH/O hydrogen bond.

10.1.4 CH/π Hydrogen Bond

In 1952, Tamres showed that benzene and its analogues dissolve in chloroform exothermically. Later, Huggins and Pimentel found by measuring IR spectra that the interaction of chloroform with benzene showed a behavior consistent with the criteria for the hydrogen bond. In 1978, Nishio et al. presented evidence that this is a hydrogen bond–like molecular force and termed this the CH/π interaction. Many studies have since demonstrated that the interaction of CH with π-bases was attractive.[21] For example, nuclear Overhauser effect (NOE) studies on a series of simple organic compounds gave evidence for the hydrogen bond nature of the CH/π interaction,[22] and IR experiments also provided firm evidence.[23] Moreover, crystallographic database analyses of organic[24] and organometallic compounds[25] gave results consistent with the above conclusion. Sakaki et al. first studied the CH/π interaction using MO calculations at the correlation level and reported that the energy comes mostly from the dispersion force.[26] Many theoretical studies supporting this concept followed.[27]

Table 10.4 compares the energy of CH/π bonds estimated by high-level *ab initio* calculations.[28] Note that the electrostatic contribution is less than 20% when a nonpolar CH group is involved. The proportion of the electrostatic energy increases from sp^3-CH to sp^2-CH to *sp*-CH. A similar trend occurs by substituting the hydrogens with halogen atoms. Interactions involving aromatic CHs are stronger than the aliphatic ones. The former is termed the aromatic CH/π interaction, but this is often referred to as the T-shape or edge-to-face π/π interaction.[29,30] A considerable portion of the interactions between aromatic side chains in protein[31] may originate from this aromatic CH/π interaction.

A unique feature of the CH/π bond is that several CH groups may simultaneously participate in the interaction with a π-base. This type of interaction is favorable

TABLE 10.4
Energy of CH/π Hydrogen Bond Calculated for Methane and Its Halogenated Derivatives in Complex with Benzene (in kcal/mol)

	E_{Total}	E_{ES}	E_{ER}	E_{DISP}	E_{ES}/E_{Total}
CH_4/C_6H_6	–1.45	–0.25	1.10	–2.30	0.17
C_2H_4/C_6H_6	–2.06	–0.65	1.82	–3.22	0.31
C_2H_2/C_6H_6	–2.83	–2.01	1.44	–2.26	0.71
CHF_3/C_6H_6	–4.20	–2.40	1.70	–3.40	0.57
CH_3Cl/C_6H_6	–3.00	–1.10	1.40	–3.40	0.37
CH_2Cl_2/C_6H_6	–4.50	–1.80	2.40	–5.10	0.40
$CHCl_3/C_6H_6$	–5.60	–2.40	4.60	–7.90	0.43

Notes: E_{Total}: total interaction energy; E_{ES}: electrostatic energy; E_{ER}: exchange repulsion; and E_{DISP}: dispersion energy.

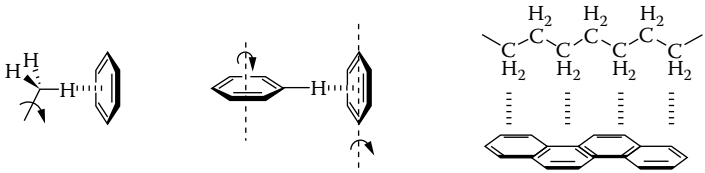


FIGURE 10.1 Features of the CH/π hydrogen bond.

in view of entropy, because assorted arrangements of CH groups are common in organic molecules. Every π-group has at least one plane of symmetry, and the total Gibbs energy of the interaction thus increases; this point is crucial in understanding the role of CH/π interactions (Figure 10.1).

Another characteristic of supreme importance when considering biochemical interactions is the role that the CH/π hydrogen bond plays in protic media such as water, as well as in nonpolar solvents. This is because the majority of its energy originates from dispersion interactions. The specificity of supramolecules comes from weak forces of enthalpic origin, including the van der Waals force and weak hydrogen bonds. The "hydrophobic effect" is by no means the cause[32] of the specific interactions in molecular recognition, including those of proteins.

The number of papers reporting a role for CH/π interactions is rapidly increasing; only recent key findings are cited here. Evidence for the role of the CH/π interaction in self-assembly has been presented,[33] and their significance in molecular capsules and lattice- and cavity-type clathrates is well documented.[34] Possibilities in designing useful drugs[35] and agents for enantiomer separation[36] and enantioselective catalysts[37] have also been suggested. The origin of a remarkable endo-stereoselectivity in Diels–Alder reactions has been attributed to the CH/π hydrogen bond.[38] Importance in the efficiency and stereoselectivity in solid-state reactions has also been demonstrated.[39]

Significance in the structure of proteins such as guanosine triphosphate (GTP)-binding proteins, proteins with Src homology 2 (SH2) motifs,[40] major histocampatability antigen complex (MHC) antigens,[41] and TATA-box-binding protein/DNA complexes[42] has been reported. Stacked arrangements of carbohydrates with aromatic residues have been interpreted as a consequence of the CH/π interaction.[43] A variety of inhibitors have been shown to bind effectively the substrate-binding cleft of acetylcholine esterase, which is lined with many aromatic residues, and this was attributed to the CH/π interaction.[44] An example is given in Figure 10.2 for a sequence of CH/π bonds observed in human growth hormone–binding protein. Note that the methylene hydrogens in Lys and Arg are involved in a stacked manner with the aromatic residues Phe, Tyr, and Trp. Spiwok et al.,[45] Sujatha et al.,[46] Harigai et al.,[47] and Fantini et al.[48] also reported significance of CH/π hydrogen bonds in protein biochemistry. By analyzing 19 adenine cofactor-specific enzymes, Chakrabarti and Samanta found that CH/π interactions play an essential role in the binding of substrate.[49] Weiss and coworkers analyzed 1154 PDB entries and found that CH/π interactions are ubiquitously present in the protein structure.[50] Further information is available in several reviews[51] and Chapter 11 of Steiner.[4] The importance of CH/π hydrogen bonds in the DNA structure has also been reported.[52]

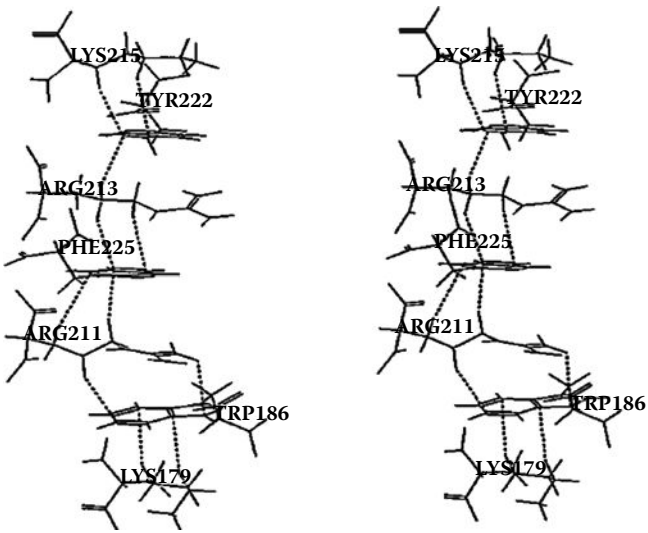


FIGURE 10.2 A sequence of CH/π hydrogen bonds observed in human growth hormone–binding protein. (Reproduced from Nishio, M. 1997. *J. Syn. Org. Chem. JAPAN* 55: 18–28, Figure 23. With permission.)

10.1.5 Summary and Prospects

The key property of the hydrogen bond is summarized as follows: the stronger the proton-donating ability of XH, the stronger the attractive nature of the hydrogen bonds. Thus, the order among acids is OH > NH > CH and *sp*-CH > > *sp*2-CH > *sp*3-CH, and the distance between H and the acceptor atom increases in the above order. As for the proton acceptor, O is stronger than N, S, and π-bases. Among the π-bases, the electron density of the H-acceptor (π-donor) is important. Directionality is requisite for hydrogen bonding and distinguishes it from the van der Waals force. The directionality follows the above order of strength in which the stronger the hydrogen bond is, the stronger the trend for linearity will be. Hence, the implication of unconventional hydrogen bonds in protein chemistry is immeasurable. The interaction between molecules is becoming the most fascinating target in biology, and the weak hydrogen bond undoubtedly plays a central part.

10.2 APPLICATION OF FRAGMENT MOLECULAR ORBITAL (FMO) TO DRUG DESIGN

10.2.1 Leukocyte-Specific Protein Tyrosine (LCK) Kinase

Structure-based drug design (SBDD) has become increasingly useful for drug discovery, and the interactions between protein and ligand are particularly important. Here, we show the application of the FMO-MP2 method to the design of LCK-kinase inhibitors, taking into account weak hydrogen bonds, especially the CH/π hydrogen bond.[53]

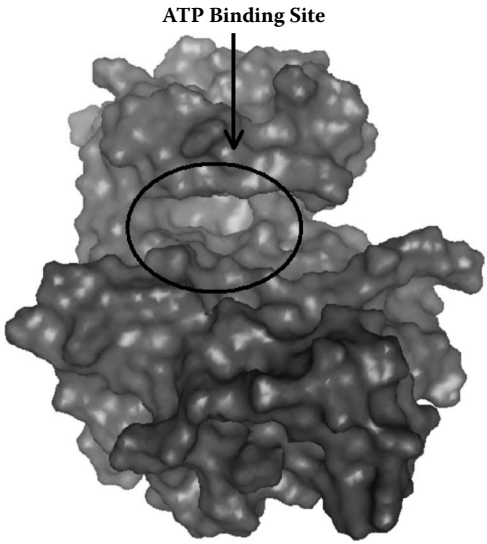


FIGURE 10.3 The ATP-binding site of the LCK protein kinase domain.

Src-family tyrosine kinases consist of eight highly homologous proteins that are expressed primarily in hematopoietic tissues,[54] two of which, LCK and FYN, are expressed in T cells, with LCK playing a critical role in the initial steps of T cell-receptor signaling.[55,56] Hence, an inhibitor of LCK has potential utility as an autoimmune agent. Several studies reported on the synthesis and characterization of LCK kinase inhibitors.[57–68] These compounds are ATP-competitive inhibitors, and the crystal structure of various LCK complexes has been reported.[69] ATP is known to be bound in the cleft formed between the two lobes of the protein kinase fold (Figure 10.3).[70]

Three sites common to all Src family kinases are critical for the binding of LCK inhibitors (Figure 10.4). Donor–acceptor pairs of hydrogen bonds are formed between the backbone atoms of the linker region and adenine. The adenine moiety is positioned at the adenine pocket, which is composed mainly of aliphatic amino acid residues. Another pocket, which is unoccupied by ATP, binds an aromatic group of the inhibitors.

Mukaiyama et al. proposed that CH/π hydrogen bonds play a role in the activity of the protein tyrosine kinase c-Src, and many aliphatic amino acid residues are found in the adenine and aromatic pockets of the enzyme.[71] Therefore, we hypothesized that electrostatic interactions and hydrogen bonds are relatively unimportant, while CH/π hydrogen bonds mainly contribute to recognition.

In view of the above information, an SBDD of LCK inhibitors was performed by crystal structure determinations and FMO calculations. Herein, we report the results of the FMO calculations of four complexes between LCK and inhibitors (Figure 10.5). Several CH/π, CH/O, and NH/π hydrogen bonds have been found to contribute to stabilizing the structure of LCK complexes.

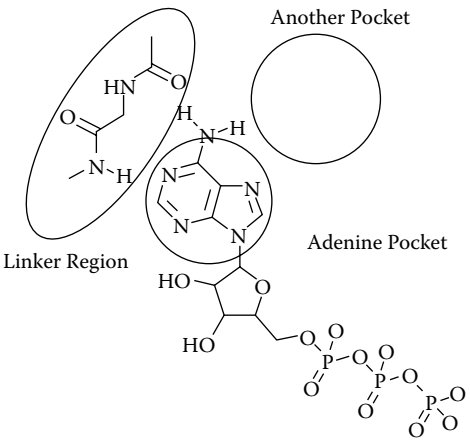


FIGURE 10.4 Features of the ATP-binding site.

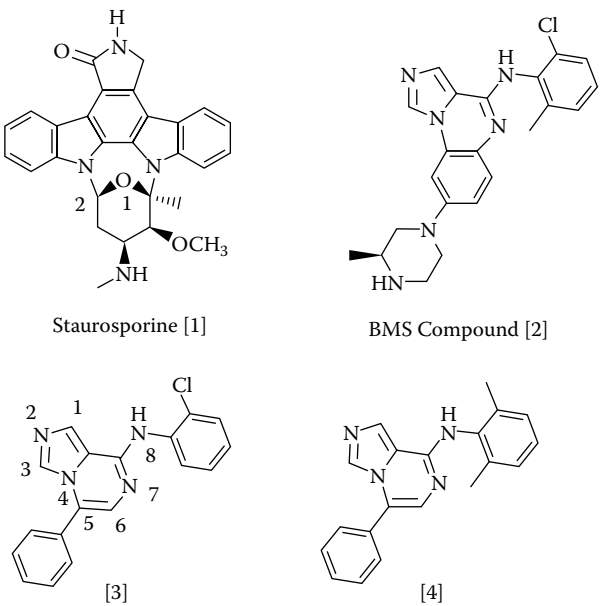


FIGURE 10.5 Structure of the four LCK inhibitors analyzed by the fragment molecular orbital (FMO) method.

10.2.2 Methods

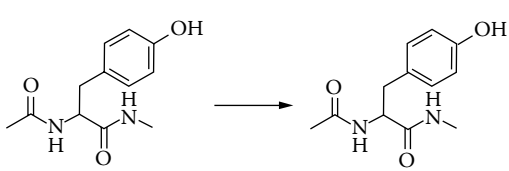
10.2.2.1 Molecular Modeling

The structure of LCK/**1** complex was retrieved from the Protein Data Bank (PDB); the PDB code is 1qpj.[69] The crystal structures of LCK complexes with inhibitors **2**, **3**, and **4** were determined in our laboratory (PDB codes 2zm4, 2zm1, and 2zob, respectively). The resolutions of the crystallographic determinations of the protein/ligand complexes were 2.2, 2.7, 2.1, and 2.6 Å for **1**, **2**, **3**, and **4**, respectively. Hydrogen atoms were generated by using the molecular graphic software Quanta 2000 (Accelrys, Inc., San Diego, California). We assumed that the N-termini of the lysine and arginine side chains were protonated, while the C-termini of aspartic and glutaminic side chains were deprotonated. The amino groups of inhibitors **1** and **2** were not protonated. The CHARMm force field implemented in Quanta 2000 was used at the minimization steps. The protein structures were optimized by the steepest descent (SD) method at dielectric constant $\varepsilon = 4R$. The optimization was performed stepwise. At the first step, the structures were minimized under the condition that the nonhydrogen atoms were constrained. Next, the protein backbone atoms were constrained. At the final step, all atoms were minimized with the harmonic atom constraint. The force constants of the harmonic atom constraint gradually decreased from 100 to 10, and then to 1.

10.2.2.2 Conditions for the FMO Calculations

In the present study, the proteins were divided at individual amino acid residues because our aim was to investigate the intermolecular interactions between the LCK protein and the inhibitors based on amino acid residue units. Note that fragmented residues do not exactly correspond to amino acid residues because the fragmentations in the FMO calculations were performed between the Cα atom and the main chain carbonyl group. Therefore, the main chain carbonyl group of the *i*th residue was assigned to the $(i + 1)$th residue fragment. In the following discussion, we refer to fragmented amino acid residues using the residue name and its position from the N-terminus (#). For example, Figure 10.6 shows that Tyr#318 consists of the carbonyl group of Glu317 through Tyr318 but does not include the carbonyl group of Tyr318.

Single-point energy calculations were performed by the RHF and MP2 methods using the 6-31G basis set (FMO-RHF/6-31G, FMO-MP2/6-31G). All of the FMO calculations were carried out using the ABINIT-MP program. The calculations were



Tyr318 **Tyr#318**

FIGURE 10.6 Amino acid residues renumbered based on the FMO fragment rule.

done on Pentium 4 3.4-GHz clusters (20 CPUs). The CPU time took about 52 h for the LCK/**4** complex (4,433 atoms, 24,524 basis functions).

10.2.3 Application of FMO to Drug Design

10.2.3.1 Staurosporine

Hydrogen bonds are formed in the linker region of the ATP-binding site of protein kinases, which is the key interaction in the recognition of inhibitors or substrates. Figure 10.7 illustrates the hydrogen bonds formed between staurosporine **1** and the enzyme. The crystal structure of the complex shows that the enzyme uses two hydrogen bonds.[69] One is found between the peptide oxygen of Glu317 and the lactam amide hydrogen of **1**. The other hydrogen bond is formed between the amide hydrogen of Met319 and the lactam carbonyl group of **1**. These two pairs of hydrogen bonds are important for the binding of the inhibitor to the protein kinase.

Our FMO calculations gave results consistent with these observations. That is, the interaction energies E_{MP2} between **1** and the LCK protein have been estimated to be –19.4 and –11.0 kcal/mol, respectively, for Tyr#318 and Met#319. The interaction of Tyr#318 is reflected in the hydrogen bond between the peptide oxygen of Glu317 and **1**, according to the FMO fragment rule described in the Methods section. Thus, the importance of the two hydrogen bonds between the linker region and **1** was confirmed by the FMO result.

The ATP-binding site also recognizes an aromatic group of the inhibitor. We investigated the mode of recognition by the FMO method. Table 10.5 shows that **1** interacts with Leu#251 (E_{MP2} –5.4 kcal/mol), Val#259 (–6.4 kcal/mol), and Leu#371 (–6.4 kcal/mol). No hydrogen bond was found between **1** and these three residues. The RHF interaction energies E_{RHF} were positive or slightly negative. These results imply that the contribution from the dispersion energy is important for the interaction involving these aliphatic residues: Leu#251 ($E_{MP2\text{-}RHF}$ –6.4 kcal/mol), Val#259 (–5.7 kcal/mol), and Leu#371 (–7.1 kcal/mol). Therefore, we consider that CH/π hydrogen bonds play a dominant role in the recognition of **1** by the LCK protein. In Figure 10.8

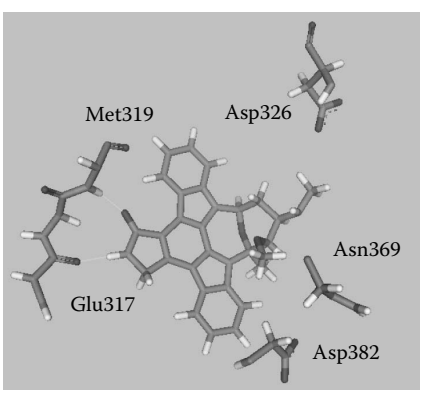


FIGURE 10.7 Interactions between **1** and Glu317 and Met319 of LCK. The lines indicate hydrogen bonds.

TABLE 10.5
Interaction Energies (in kcal/mol) between the LCK and Staurosporine (1)

	E_{MP2}[a]	E_{RHF}[b]	$\Delta E_{MP2\text{-}RHF}$[c]
Tyr318	–19.4	–13.6	–5.8
Asp382	–15.1	–11.2	–3.9
Gly252	–11.1	–6.9	–4.2
Met319	–11.0	–8.3	–2.7
Asn369	–10.9	–7.6	–3.2
Asp326	–8.3	–7.1	–1.1
Glu288	–7.4	–6.3	–1.1
Leu371	–6.4	0.7	–7.1
Val259	–6.4	–0.7	–5.7
Leu251	–5.4	1.0	–6.4
Lys269	–4.5	–4.5	0.0
Asp364	–4.2	–4.2	0.0
Gly254	–3.3	–2.8	–0.4
Thr316	–3.1	–1.6	–1.5
Val272	–3.0	–2.8	–0.2
Glu432	–2.1	–2.1	0.0
Asn321	–2.1	–1.8	–0.3
Gly322	–1.7	1.0	–2.8
Lys246	–1.7	–1.7	0.0
Gly262	–1.6	–1.6	0.0
Val372	–1.6	–1.4	–0.2
Asp422	–1.5	–1.5	0.0
His267	–1.1	–1.1	0.0
Lys379	–1.0	–1.0	0.0

[a] Interaction energy calculated at the MP2/6-31G level.
[b] Interaction energy calculated at the RHF/6-31G level.
[c] $E_{MP2} - E_{RHF}$.

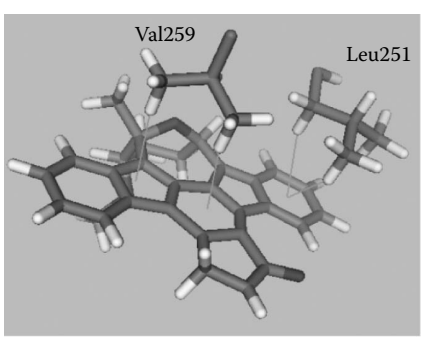


FIGURE 10.8 Interactions between **1** and Leu251 and Val259 of LCK. The lines indicate CH/π hydrogen bonds.

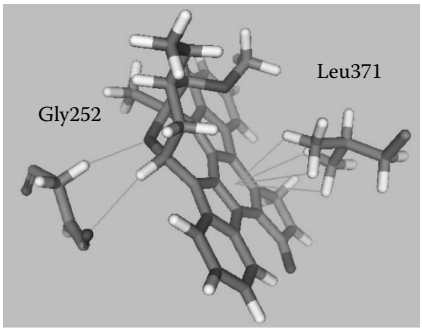


FIGURE 10.9 Interactions between 1 and Gly252 and Leu371 of LCK. The lines indicate CH/O and CH/π hydrogen bonds.

and Figure 10.9, four short individual CH/π contacts were observed between Leu251, Val259, and Leu371 and **1**. Consequently, the FMO calculations suggest that the CH/π hydrogen bond plays an important role in the ATP site of the LCK protein. Crystal structures of several protein kinases revealed that the ATP-binding sites are similar to each other and are occupied by an aromatic ring system of the inhibitor.

The glycosyl part of **1** also interacts with the LCK protein. The crystal structure of the complex shows that the methylamino group of the glycosyl part of **1** is involved in a hydrogen bond with the LCK protein. This group is positioned to allow hydrogen bonding to a carbonyl group of the LCK protein. The FMO results gave results consistent with this; **1** interacts with Asp#382, Asn#369, and Asp#326 through electrostatic interactions and hydrogen bonds. The interaction between Gly#252 and the LCK protein is difficult to explain, but it may be due to CH/O hydrogen bonds between Cα of Gly252 and 1-oxygen of the glycosyl group, a carbonyl oxygen of Gly252, and a hydrogen attached to carbon-2 (Figure 10.5).

To summarize this section, FMO calculations showed that the CH/π hydrogen bond, in addition to conventional hydrogen bonds and electrostatic interactions, is important in the recognition of the LCK protein to **1**. Therefore, the CH/π hydrogen bond, as well as the hydrogen bond, must be taken into account in the following drug design.

10.2.3.2 BMS Compound

The Lipinski's "rule-of-five" states that "to be a useful drug, at least five hydrogen bond donor- and ten acceptor-functionalities should be achieved in a candidate molecule."[73] Recently, a "rule-of-three," which limits the number of hydrogen bond donors to three, has been proposed.[74] In BMS compound **2**, the number of hydrogen bond donors is smaller than in other kinase inhibitors.[65] The BMS compound was chosen as the lead compound of our study because this shows one CH/O hydrogen bond in the linker region of the LCK protein. Chen et al. reported that analogs of compound **2** have two hydrogen bonds that are critical in the binding of protein: N^3-nitrogen of the imidazole group to NH of Met319, and NH of the aniline moiety to OH of Thr316.[65] They, however, did not comment on the role of CH/O hydrogen bonds.

TABLE 10.6
Interaction Energies (in kcal/mol) between the LCK and BMS Compound (2)

	E_{MP2}[a]	E_{RHF}[b]	$\Delta E_{MP2\text{-}RHF}$[c]
Met319	–8.4	–5.2	–3.3
Tyr318	–8.2	–5.7	–2.5
Thr316	–7.3	–2.3	–5.0
Asp326	–7.0	–5.8	–1.3
Lys273	–7.0	–1.4	–5.7
Glu288	–5.1	–2.9	–2.2
Gly252	–5.1	–2.8	–2.3
Leu371	–4.3	0.2	–4.5
Val272	–3.3	–2.7	–0.6
Ile315	–3.3	–2.2	–1.1
Val259	–3.1	1.0	–4.1
Met292	–1.9	–0.6	–1.3
Leu251	–1.9	0.7	–2.6
Lys269	–1.5	–1.5	0.0
Phe383	–1.3	–0.3	–1.0
Trp260	–1.2	–0.9	–0.3
Lys379	–1.2	–1.2	0.0
Ile314	–1.2	0.4	–1.5
Leu303	–1.0	–1.0	0.0

[a] Interaction energy calculated at the MP2/6-31G level.
[b] Interaction energy calculated at the RHF/6-31G level.
[c] $E_{MP2} - E_{RHF}$.

Table 10.6 shows that **2** interacts with Met#319, Tyr#318, and Thr#316 (see Figure 10.10).

The interactions involving Met#319 correspond to those between the amide hydrogen of Met319 and N_3 of the imidazole part of **2**. Tyr#318 consists of the carbonyl group of Glu317, while Tyr318 does not include its carbonyl group. The crystal structure of the complex shows that the side chain of Tyr#318 is more than 4.2 Å from **2** (Figure 10.11). We therefore consider that **2** interacts with Tyr#318 through a CH/O hydrogen bond.[75] The distance between the carbonyl group of Tyr#318 and the hydrogen attached to C4 of the imidazole ring was reported to be about 2.4 Å; this is compatible with the observation by Pierce et al., which mentioned the importance of CH/O hydrogen bonds in kinase inhibitors.[76,77]

The interactions involving Thr316 correspond to those between the hydroxyl group of Thr316 and the hydrogen of the aniline part. In the tricyclic aromatic part, several short CH/π distances have been disclosed (Figure 10.12). Thus, Table 10.6 shows that **2** interacts with Leu251 (ΔE_{MP2} –1.9 kcal/mol), Val259 (–3.1 kcal/mol), and Leu371 (–4.3 kcal/mol). The above three residues have significant dispersion interactions: Leu251 ($E_{MP2\text{-}RHF}$ –2.6 kcal/mol), Val259 (–4.1 kcal/mol), and Leu371 (–4.5 kcal/mol).

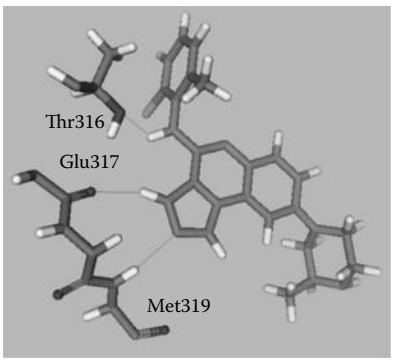


FIGURE 10.10 Interactions between 2 and Thr316, Glu317, and Met319 of LCK. The lines indicate hydrogen bonds.

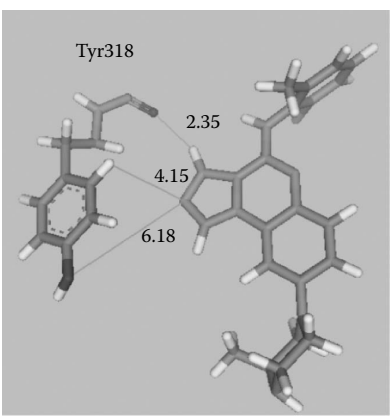


FIGURE 10.11 The distance between **2** and Tyr318.

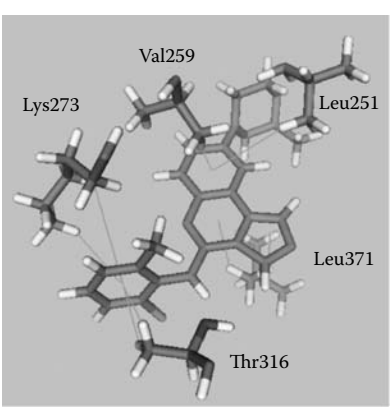


FIGURE 10.12 Interactions between **2** and Leu251, Val259, Lys273, Thr316, and Leu371 of LCK. The lines indicate NH/π and CH/π hydrogen bonds.

These interactions are weaker than those found in the **1**-complex. This is considered to be attributable to the difference in the number of CH/π hydrogen bonds.

We thought that the interaction energy involving the tricyclic aromatic portion would be adequate to be a candidate for the lead compound. In the 2-chloro-6-methyl-aniline part, short CH/π contacts were disclosed between the benzene ring and the side chain of Lys273 and Thr316. The interactions energies E_{MP2} were –7.0 kcal/mol for Lys273 and –7.3 kcal/mol for Thr316, and $\Delta E_{MP2\text{-}RHF}$ was –5.7 kcal/mol and –5.0 kcal/mol, respectively, for Lys273 and Thr316. The ε-amino group of Lys273 was located above the benzene ring of 2, which may have been due to a NH/π hydrogen bond.

To summarize this section, the FMO results revealed that BMS-279700 **2** interacts with the LCK protein through not only hydrogen bonds but also through CH/O, NH/π, and CH/π hydrogen bonds, and that this has enough interaction energy to be a candidate for the lead compound. We therefore tried to improve the binding ability of **2**, taking CH/π hydrogen bonds into further consideration.

10.2.3.3 Compound 3: N-(2-Chlorophenyl)-5-Phenylimidazo[1,5-A]pyrazin-8-Amine

In view of the results obtained by the FMO analysis of the above two complexes, we hypothesized that CH/O and CH/π hydrogen bonds play an important role in the binding of inhibitors with the LCK protein. Thus, *N*-(2-chlorophenyl)-5-phenylimidazo[1,5-a]pyrazin-8-amine (**3**) was designed and synthesized as a candidate inhibitor. First, we determined the crystal structure of **3** in complex with the LCK kinase domain. Next, FMO calculations were carried out to obtain information necessary for optimizing the structure of potential ligands.

Table 10.7 shows that **3** interacts with Thr#316 (–7.8 kcal/mol), Tyr#318 (–9.7 kcal/mol), and Met#319 (–9.7 kcal/mol). The binding mode of these three residues is similar to that of **2**.

In the crystal structure (Figure 10.13), short CH/π contacts are shown between the imidazo-pyrazine moiety of **3** and the side-chain groups of Val259 and Leu371. In agreement with this observation, the FMO calculations show that **3** interacts with Val#259 and Leu#371; their interaction energies are –2.0 and –4.3 kcal/mol, respectively. In contrast, the RHF interaction energies involving these two residues are positive, suggesting that hydrogen bonds do not play a role. Although Leu251 interacts with the core-ring part of **1** and **2** via CH/π hydrogen bonds, in **3**, Leu251 is remote from the imidazo-pyrazine ring. In contrast, short CH/π contacts have been noted between the side chain of Leu251 and the phenyl ring at position 5 of the imidazo-pyrazine ring. The calculated interaction energy (–2.5 kcal/mol) is consistent with the crystal structure. The tricyclic ring in the core region of the inhibitor **2** is converted to a bicyclic ring in **3**. Three amino acid residues (Val259, Leu371, and Leu251) interacting with **2** by CH/π hydrogen bonds are also involved in **3**.

Figure 10.14 shows short CH/π contacts between the benzene ring of **3** and the side chain of Lys273 and Thr316. In agreement with this observation, the interaction energies were calculated to be –6.7 kcal/mol for Lys273 and –7.8 kcal/mol for Thr316 (Table 10.7). The 2-chlorol-aniline part of **3** interacts with Lys273 through NH/π and CH/π hydrogen bonds; Thr316 interacts via hydrogen bonds and CH/π hydrogen bonds.

TABLE 10.7
Interaction Energies (in kcal/mol) between the LCK and Compound 3

	E_{MP2}[a]	E_{RHF}[b]	$\Delta E_{MP2\text{-}RHF}$[c]
Met319	–9.7	–6.0	–3.7
Tyr318	–9.7	–6.1	–3.6
Thr316	–7.8	–3.1	–4.7
Glu288	–6.9	–4.7	–2.2
Lys273	–6.7	–2.3	–4.5
Gly252	–4.7	–3.7	–1.1
Leu371	–4.3	0.9	–5.2
Leu251	–2.5	0.5	–3.0
Lys269	–2.2	–2.2	0.0
Val259	–2.0	1.6	–3.6
Asp326	–2.0	–1.9	–0.1
Met319	–1.9	–0.5	–1.3
Val272	–1.8	–1.2	–0.6
Trp260	–1.8	–1.7	–0.2
Ile314	–1.7	–0.3	–1.4
Val301	–1.6	0.7	–2.3
Ala381	–1.6	–0.9	–0.6
Gly322	–1.5	–0.6	–0.9
Ser274	–1.2	–1.1	–0.1

[a] Interaction energy calculated at the MP2/6-31G level.
[b] Interaction energy calculated at the RHF/6-31G level.
[c] $E_{MP2} - E_{RHF}$.

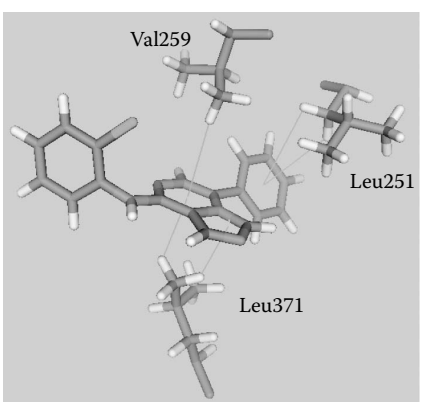


FIGURE 10.13 Interactions between **3** and Leu251, Val259, and Leu371 of LCK. The lines indicate CH/π hydrogen bonds. **(See color insert following page 117.)**

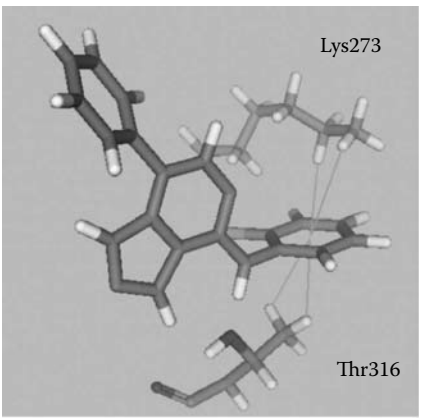


FIGURE 10.14 Interactions between **3** and Lys273 and Thr316 of LCK. The lines indicate NH/π and CH/π hydrogen bonds.

To summarize, the FMO calculation shows that **3** interacts with the LCK protein through CH/O, NH/π, and CH/π hydrogen bonds. However, the inhibition value of **3** to the LCK protein remained modest (IC_{50} = 220 nM).

10.2.3.4 Compound 4: N-(2,6-Dimethylphenyl)-5-Phenylimidazo[1,5-a]Pyrazin-8-Amine

In the previous section, we noted that the interaction of the LCK protein with the 2-chlorol-aniline part of **3** is stabilized by CH/π and NH/π hydrogen bonds. The substituent effects on the CH/π hydrogen bond have been measured in many studies using IR and NMR spectroscopy,[78,79] X-ray crystallography,[80] and *ab initio* calculations.[81] According to these papers, it was shown that an electron-donating substituent on the π-ring system favors the CH/π hydrogen bond. We therefore designed *N*-(2,6-dimethylphenyl)-5-phenylimidazo[1,5-a]pyrazin-8-amine (**4**) to improve the interactions. The present FMO results agree well with the structure-activity relationship (SAR) of **3** and **4** (Table 10.8).

The difference in the interaction energies of Lys273 between **3** and **4** were –6.0 kcal/mol, which was the largest value among the amino acid residues in the protein. Because the side chain of Lys273 was located above the 2,6-dimethylphenyl ring of **4**, the enhanced activity of **4** was attributable to the NH/π and CH/π hydrogen bonds. In contrast, the difference in the interaction energies in Thr316 was small (–0.02 kcal/mol). The RHF interaction energy decreased 1.2 kcal/mol from **3** to **4**, while the dispersion energy increased by –1.2 kcal/mol. This result may be attributable to the decrease in the acidity of NH in the aniline part of the candidate inhibitors and the increase of electron density in the benzene ring by the introduction of methyl groups at positions 2 and 6. The differences in the interaction energies involving other residues were insignificant in the cases of Met319 (0.13 kcal/mol), Tyr318 (0.89 kcal/mol), Leu371 (0.05 kcal/mol), Val259 (–0.6 kcal/mol), and Leu251 (0.02 kcal/mol). The interactions between **4** and the LCK protein were similar to **3** in the linker region

TABLE 10.8
Interaction Energies (in kcal/mol) between the LCK and Compound 4

	E_{MP2}[a]	E_{RHF}[b]	$\Delta E_{MP2\text{-}RHF}$[c]
Lys273	–12.7	–7.7	–5.0
Met319	–9.6	–5.9	–3.7
Tyr318	–8.8	–5.7	–3.1
Thr316	–7.8	–1.9	–5.9
Gly252	–5.8	–4.2	–1.6
Leu371	–4.3	0.5	–4.8
Glu288	–3.5	–0.7	–2.8
Ile315	–3.1	–1.9	–1.2
Val272	–2.9	–2.1	–0.8
Val259	–2.6	1.4	–4.0
Leu251	–2.4	0.9	–3.4
Val301	–1.8	0.7	–2.5
Met292	–1.7	–0.1	–1.6
Lys269	–1.7	–1.7	0.0
Ile314	–1.4	0.6	–2.0
Gly322	–1.4	–0.5	–0.8
Asp326	–1.3	–1.3	0.0
Phe383	–1.2	0.0	–1.2
Trp260	–1.1	–0.9	–0.3

[a] Interaction energy calculated at the MP2/6-31G level.
[b] Interaction energy calculated at the RHF/6-31G level.
[c] $E_{MP2} - E_{RHF}$.

and imidazo-pyrazine ring. In agreement with the statement by Snow et al., 2,6-disubstitutions might have stabilized the conformation of the inhibitor in the complex.[59] Consequently, **4** is highly potent as an inhibitor of the LCK-protein (IC_{50} = 20 nM).

10.2.4 Summary

The interaction energy was calculated by the *ab initio* FMO method for complexes between the LCK protein and four inhibitors (staurosporine **1**, BMS compound **2**, and compounds **3** and **4**). In the four complexes investigated, CH/π hydrogen bonds were found to play an important role in the binding of the protein to the inhibitors. In the complexes of every ligand, several CH/π hydrogen bonds were disclosed in the adenine pocket. In **2**, **3**, and **4**, CH/π and NH/π hydrogen bonds were also observed in another pocket, which was not occupied by ATP. Additionally, CH/O hydrogen bonds have been shown to play a role in the linker region. According to the above findings, the aniline ring in **3** was substituted by 2,6-dimethyl in **4** to increase the potency for LCK kinase. Thus, a tenfold increase in the potency was achieved for **4** (IC_{50} = 20 nM) over **3** (IC_{50} = 220 nM).

10.3 DEPENDENCE OF IFIES ON THE BASIS SETS

10.3.1 SH2 Domain of SAP

The dependence of the interfragment interaction energies (IFIEs) on the basis sets was investigated. The complexes between a peptide and the SH2 domain of SAP (SAP-SH2), which has 1,844 atoms and 115 amino acids, were calculated by STO-3G, 3-21G, 6-31G, 6-31G*, and 6-31G**, because we previously studied SAP-SH2 and showed that the weak interaction, especially the CH/π hydrogen bond, plays an important role in the recognition of protein and ligand peptides. SAP, a product of the gene mutated in X-linked lymphoproliferative syndrome (XLP), is a small protein consisting of 128 residues and has a single SH2 domain.[82] SAP binds the cytoplasmic tail of SLAM, a self-ligand glycoprotein expressed on the surface of activated B and T cells.[83] In general, the SH2 domain recognizes its specific ligand peptide by two key regions: one is a phosphorylated-tyrosine residue (pTyr)-binding site and the other is the "specificity-determining region" that interacts with ligand residues at the C-terminus of pTyr peptides.[84] The IFIEs between Ile(+3) and SAP were compared with different basis sets. Tyr54 and Phe87 were especially studied in terms of the CH/π hydrogen bond. We use the 1t4w (PDB code), the resolution of which is 1.1 Å, for this analysis.[85]

10.3.2 Interaction between the SAP-SH2 Domain and a Peptide

We analyzed the structure of the SAP-SH2 domain in complex with a 10-residue phosphopeptide (pdb code 1d4t). Figure 10.15 shows the interactions revealed between Val(+3) and SAP-SH2. Table 10.9 lists the interactions calculated between

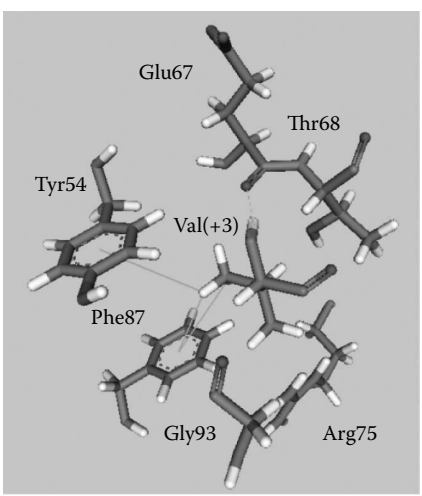


FIGURE 10.15 Interactions between Val(+3) and SAP-SH2. The line indicates the CH/π hydrogen bond.

TABLE 10.9
Basis Sets Dependence on Interaction Energies (in kcal/mol) between the Syk-SH2 and Leu(+3) of Phosphotyrosyl Peptide

6-31G**	E_{MP2}[a]	E_{RHF}[b]	$\Delta E_{MP2\text{-}RHF}$[c]
Thr68	–16.5	–10.2	–6.3
Glu67	–7.1	–6.5	–0.6
Tyr54	–5.6	–3.1	–2.6
Gly93	–2.4	–1.2	–1.3
Phe87	–1.2	1.1	–2.3
Val95	–0.8	–0.7	–0.1
6-31G*	E_{MP2}[a]	E_{RHF}[b]	$\Delta E_{MP2\text{-}RHF}$[c]
Thr68	–17.4	–11.0	–6.4
Glu67	–7.4	–6.8	–0.6
Tyr54	–5.8	–3.3	–2.5
Gly93	–2.3	–1.1	–1.2
Phe87	–1.1	1.1	–2.2
Val95	–0.9	–0.8	–0.1
6-31G	E_{MP2}[a]	E_{RHF}[b]	$\Delta E_{MP2\text{-}RHF}$[c]
Thr68	–17.0	–11.9	–5.1
Glu67	–7.1	–6.6	–0.5
Tyr54	–5.5	–3.4	–2.1
Gly93	–2.6	–1.5	–1.1
Val95	–0.9	–0.8	–0.1
Phe87	–0.8	1.1	–1.9
3-21G	E_{MP2}[a]	E_{RHF}[b]	$\Delta E_{MP2\text{-}RHF}$[c]
Thr68	–24.7	–18.6	–6.1
Tyr54	–6.8	–5.1	–1.7
Glu67	–6.3	–5.9	–0.4
Gly93	–3.2	–2.1	–1.1
Phe87	–1.2	0.6	–1.8
Val95	–0.8	–0.7	–0.1
STO-3G	E_{MP2}[a]	E_{RHF}[b]	$\Delta E_{MP2\text{-}RHF}$[c]
Thr68	–8.5	–6.6	–1.9
Glu67	–5.9	–5.8	–0.2
Tyr54	–1.9	–1.5	–0.4
Phe87	0.6	1.2	–0.6

[a] Interaction energy calculated at the MP2 level.
[b] Interaction energy calculated at the RHF level.
[c] $E_{MP2} - E_{RHF}$.

Val(+3) and SAP-SH2. Val(+3) interacts with Thr68, Glu67, Tyr54, Gly93, Arg75, Val95, and Phe87.

The crystal structure shows that the amide hydrogen of Val(+3) forms a hydrogen bond with the carbonyl oxygen of Glu67. This is reflected in the interaction energies between Val(+3) and Thr68, according to the FMO fragment rule. Three close contacts were found between hydrogens and the phenyl ring. One was revealed between the γ methyl hydrogens of Val(+3) and the phenyl ring of Tyr54. The other two are γ methyl hydrogens of Val(+3) and the phenyl ring of Phe87. Consistent with these observations, the difference in ($\Delta E_{\text{MP2-RHF}}$) was –2.6kcal/mol (6-31G**) for Val(+3) and Tyr54 and –2.3 kcal/mol (6-31G**) for Val(+3) and Phe87. The γ methyl group of Val(+3) is located at the middle of the two phenyl rings. These facts suggest that the CH/π hydrogen bond is an important factor in the recognition of SAP-SH2. From the FMO results and the relative orientation of the residues, we surmise that Val(+3) interacts with Gly93, Arg75, and Val95 via van der Waals interactions. The interaction between Val(+3) and Glu67 is shown. The magnitude of $\Delta E_{\text{MP2-RHF}}$ was calculated to be –0.6 kcal/mol, and thus this interaction may be electrostatic in nature.

10.3.3 Comparison of IFIEs

Table 10.9 shows a comparison of IFIEs with a variety of basis sets. The IFIEs of Tyr54 were –5.6, –5.8, –5.5, –6.8, and –1.9 kcal/mol for MP2/6-31G**, 6-31G*, 6-31G, 3-21G, and STO-3G levels, respectively, and –3.1, –3.3, –3.4, –5.1, and –1.5 kcal/mol, respectively, for RHF/6-31G**, 6-31G*, 6-31G, 3-21G, and STO-3G. The difference in energy between the MP2 and RHF methods, which represent the dispersion interactions, were –2.6, –2.5, –2.1, –1.7, and –0.4 kcal/mol. The IFIEs of the MP2/6-31G**, 6-31G*, and 6-31G basis sets had nearly the same values. The dispersion interaction of 6-31G was rather small compared to that of 6-31G** (–0.5 kcal/mol). Conversely, the IFIE of 3-21G was about 1 kcal/mol larger in comparison to that of 6-31G**, and the dispersion interaction of 3-21G was about 1 kcal/mol smaller compared to that of 6-31G**. STO-3G extremely underestimated the IFIE and dispersion interaction. For Phe87, the IFIEs of MP2/6-31G** and 6-31G* were similar, and that of 6-31G was slight (–0.4kcal/mol). The IFIE of MP2/3-21G was equal to that of 6-31G**, and the dispersion was 0.5 kcal/mol lower relative to that of 6-31G**. The IFIE of STO-3G was a positive value. The computation times using Pentium 4 3.4-GHz clusters (20 CPUs) were about 218, 53, 13, 10, and 2 h for MP2/6-31G**, 6-31G*, 6-31G, 3-21G, and STO-3G, respectively.

10.3.4 Summary

The MP2/6-31G** and 6-31G* basis sets provided identical results with respect to the CH/π hydrogen bond. MP2/6-31G also demonstrated similar results in qualitative terms; the FMO results of MP2/6-31G are available for structure-based drug design. Conversely, the FMO results for MP2/3-21G are not useful because the IFIEs and dispersion interaction values were not always consistent with those of MP2/6-31G** and tended to overestimate the hydrogen bond. The results of STO-3G were poor and thus should not be used in the evaluation of dispersion interactions.

10.4 CONCLUSION

We suggest that CH/π hydrogen bonds play an indispensable role in LCK protein-ligand recognition. The concept that weak hydrogen bonds such as CH/π, NH/π, and CH/O hydrogen bonds play a role in recognition is useful for rational design of autoimmune agents. We also showed that CH/π hydrogen bonds play an indispensable role in the recognition of the Src homology 2 (SH2) domains with their specific pTyr peptide[86] and that SH2 domains participate in the signal transduction system. On the basis of these results, we propose that CH/π hydrogen bonds are vital in the signal transduction system. We propose that the FMO method is useful in SBDD because the FMO can evaluate not only the conventional hydrogen bond but also the weak hydrogen bond.

ACKNOWLEDGMENTS

We thank Kazuo Kitaura (Kyoto University) for his valuable advice and suggestions in performing the FMO calculations, Tatsuya Nakano (National Institute of Health Sciences) for his advice in interpreting the FMO results, Eiich Tusji (Kissei Pharmaceutical Company, Ltd.) for his X-ray crystal analysis, Motoyasu Ozawa (Kissei Pharmaceutical Company, Ltd.) for his advice on molecular modeling, Chiaki Handa (Kissei Pharmaceutical Company, Ltd.) for his technical assistance in computations, Harunobu Mukaiyama, Toshihiro Nishimura, and Satoko Kobayashi (Kissei Pharmaceutical Company, Ltd.) for their medicinal study. TO would acknowledge the professional technical supports by the "Revolutionary Simulation Software" (RSS21) project of the Ministry of Education, Culture, Sports, Science, and Technology.

REFERENCES

1. Pimentel, G.C., McClellan, A.L. 1960. *The Hydrogen Bond*, W.H. Freeman, San Francisco, CA.
2. Desiraju, G.R., Steiner, T. 1999. *The Weak Hydrogen Bond in Structural Chemistry and Biology*, Oxford University Press, Oxford.
3. Nishio, M., Hirota, M., Umezawa, Y. 1998. *The CH/π Interaction. Evidence, Nature, and Consequences*, Wiley-VCH, New York.
4. Steiner, T. 2002. The hydrogen bond in the solid state. *Angew. Chem. Int. Ed.* 41: 48–76.
5. Nishio, M. 2008. *Introduction to Weak Molecular Forces in Organic Chemistry and Life Science*, Kodansha Scientific, Tokyo (in Japanese).
6. Stone, A.J. 1996. *The Theory of Intermolecular Forces*, Clarendon Press, Oxford; Scheiner, S. 1997. *Hydrogen Bonding, A Theoretical Perspective*, Oxford University Press, Oxford.
7. Kitaura, K., Morokuma, K. 1976. A new energy decomposition scheme for molecular interactions within the Hartree–Fock approximation. *Int. J. Comput. Chem.* 10: 325.
8. Hayes, I.S., Stone, A. 1984. An intermolecular perturbation theory for the region of moderate overlap. *J. Mol. Phys.* 53: 83–105.
9. Popelier, P. 1998. *Atoms in Molecule-An Introduction.* Prentice-Hall.
10. Perutz, M.F. 1993. The role of aromatic rings as hydrogen-bond acceptors in molecular recognition. *Phil. Trans. R. Soc. A* 345: 105–112.

11. Rozas, I., Alkorta, I.E. 1997. Unusual hydrogen bonds: H···π Interactions. *J. Phys. Chem. A* 101: 9457-9463; Samanta, U., Chakrabarti, P., Chandrasekhar, J. 1998. *Ab initio* study of energetics of X-H···π(X = N, O, and C) interactions involving a heteroaromatic ring. *J. Phys. Chem. A* 102: 8964–8969; Tsuzuki, S., Honda, K., Uchimaru, T. et al. 2000. The magnitude of the CH/π interaction between benzene and some model hydrocarbons. *J. Am. Chem. Soc.* 122: 11450–11458; Tarakeshwar, P., Kim, K.S. 2002. Comparison of the nature of π and conventional H-bonds: a theoretical investigation. *J. Molec. Struct.* 615: 227–238.
12. Steiner, T., Koellner, G. 2001. Hydrogen bonds with π-acceptors in proteins: frequencies and role in stabilizing local 3D structures. *J. Mol. Biol.* 305: 535–557.
13. Taylor, R., Kennard, O. 1982. Crystallographic evidence for the existence of CH/O, CH/N, and CH/Cl hydrogen bonds *J. Am. Chem. Soc.* 104: 5063–5070.
14. Desiraju, G.R. 1991. The CH/O Hydrogen Bond in Crystals: What is it? *Acc. Chem. Res.* 24: 290–296; Steiner, T. 1996. *Cryst. Rev.* 6: 1–57; Steiner, T. 1997. Unrolling the hydrogen bond properties of C–H···O interactions. *Chem. Commun.* 727–734.
15. T. Steiner. 1998. Donor and acceptor strengths in C–H···O hydrogen bonds quantified from crystallographic data of small solvent molecules. *New J. Chem.* 22: 1099–1103.
16. Houk, K.N., Menzer, S., Newton, S.P. et al. 1999. [C-H···O] interactions as a control element in supramolecular complexes: experimental and theoretical evaluation of receptor affinities for the binding of bipyridinium-based guests by catenated hosts. *J. Am. Chem. Soc.* 121: 1479–1487.
17. Derewenda, Z.S., Lee, L., Derewenda, U. 1995. The occurence of C–H · · · O hydrogen bonds in proteins. *J. Mol. Biol.* 252: 248–262; Chakrabarti, P., Chakrabarti, S. 1998. C—H· ·O hydrogen bond involving proline residues in α-helices. *J. Mol. Biol.* 284: 867–873; Fabiola, G.F., Krishnaswamy, S., Nagarajan, V. 1997. *Acta Cryst., Sect. A. 53*: 316–320.
18. Gu, Y., Kar, T., Scheiner, S. 1999. Fundamental properties of the CH···O interaction: is it a true hydrogen bond? *J. Am. Chem. Soc.* 121: 9411–9422.
19. Hartmann, M., Wetmore, S.D., Radom, L. 2001. C-H···X hydrogen bonds of acetylene, ethylene, and ethane with first- and second-row hydrides. *J. Phys. Chem. A.* 105: 4470–4479.
20. Steiner, T. 1998. Donor and acceptor strengths in C–H···O hydrogen bonds quantified from crystallographic data of small solvent molecules. *New J. Chem.* 22: 1099–1103.
21. Madhavi, N.N.L., Katz, A.M., Carrell, H.L. et al. 1997. Evidence for the characterisation of the C–H···π interaction as a weak hydrogen bond: toluene and chlorobenzene solvates of 2,3,7,8-tetraphenyl-1,9,10-anthyridine *Chem. Commun.* 1953–1954.
22. Suezawa, H., Hashimoto, T., Tsuchinaga, K. et al. 2000. Electronic substituent effect on intramolecular CH/π interaction as evidenced by NOE experiments. *J. Chem. Soc., Perkin Trans. 2.* 1243–1249.
23. Sundararajan, K., Sankaran, K.S., Viswanathan, K.S. et al. 2002. H-π complexes of acetylene-ethylene: a matrix isolation and computational study. *J. Phys. Chem. A* 106: 1504–1510; Fujii, A., Morita, S., Miyazaki, M. et al. 2004. A molecular cluster study on activated CH/π interactions: infrared spectroscopy of aromatic molecule-acetylene clusters. *J. Phys. Chem. A* 108: 2652–2658.
24. Umezawa, Y., Tsuboyama, S., Honda, K. et al. 1998. CH/π interaction in the crystal structure of organic compounds. A database study. *Bull. Chem. Soc. Jpn.* 71: 1207–1213; Suezawa, H., Yoshida, T., Hirota, M. et al. 2001. The CH/π interaction as an important factor in the crystal packing and in determining the structure of clathrates. *J. Chem. Soc., Perkin Trans. 2.* 2053–2058.
25. Suezawa, H., Yoshida, T., Umezawa, Y. et al. 2002. CH/π interactions implicated in the crystal structure of transition metal compounds. A database study. *Eur. J. Inorg. Chem.* 3148–3155.

26. Sakaki, S., Kato, K., Miyazaki, T. et al. 1993. Structures and binding energies of benzene-methane and benzene-benzene complexes. *J. Chem. Soc., Faraday Trans. 2*. 89; 659–664.
27. Novoa, J.J., Mota, F. 2000. The CH/π bonds: strength, identification, and hydrogen-bonded nature: a theoretical study. *Chem. Phys. Lett. 318*: 345–354; Tsuzuki, S., Honda, K., Uchimaru, T. 2000. The magnitude of the CH/π interaction between benzene and some model hydrocarbons. *J. Am. Chem. Soc.* 122: 3746–3753; Scheiner, S., Grabowski, S.J. 2002. Acetylene as potential hydrogen-bond proton acceptor. *J. Molec. Struct.* 615: 209–218; Bagno, A., Saielli, G., Scorrano, G. 2002. Through-space spin-spin coupling in van der Waals dimers and CH/π interacting systems. An *ab initio* and DFT study. *Chem., Eur. J. 8*: 2047–2056; Ugozzoli, F., Arduini, A., Massera, C. et al. 2002. CH/π interaction between benzene and model neutral organic molecules bearing acid CH groups. *New J. Chem.* 26: 1718–1723; Kobayashi, Y., Saigo. K. 2005. Periodic *ab initio* approach for the cooperative effect of CH/π interaction in crystals: relative energy of CH/π and hydrogen-bonding interactions. *J. Am. Chem. Soc.* 127: 15054–15060; Ran, J., Wong, M.W. 2006. Saturated hydrocarbon-benzene complexes: theoretical study of cooperative CH/π interactions. *J. Phys. Chem. A* 110: 9702–9729.
28. Tsuzuki, S., Honda, K., Uchimaru, T. 2002. The interaction of benzene with chloro- and fluoromethanes: effects of halogenation on CH/π Interaction. *J. Phys. Chem. A* 106: 4423–4428.
29. Hunter, C.A. 1994. Meldola Lecture. The role of aromatic interactions in molecular recognition. *Chem. Soc. Rev.* 101–109.
30. Carver, F.J., Hunter, C.A., Seward, E.M. 1998. Structure-activity relationship for quantifying aromatic interactions. *Chem. Commun.* 775–776.
31. Burley, S.K., Petsko, G.A. 1988. Weakly polar interactions in proteins. *Adv. Protein Chem.* 39: 125–189.
32. Hildebrand, J.H. 1979. Is there a "hydrophobic effect?" *Proc. Natl. Acad. Sci. USA* 76: 194.
33. Cantrill, S.J., Preece, J.A., Stoddart, J.F. et al. 2000. The idiosyncrasies of tetrabenzo[24]crown-8 in the solid state. *Tetrahedron* 56: 6675–6681.
34. O'Leary, B.M., Grotzfeld, R.M., Rebek, J. 1997. Ring inversion dynamics of encapsulated cyclohexane. *J. Am. Chem. Soc.* 119: 11701–11702; Chapman, R.G., Olovsson, G., Trotter, J. et al. 1998. Crystal structure and thermodynamics of reversible molecular capsules. *J. Am. Chem. Soc.* 120: 6252–6260; Biradha, K., Dennis, D., MacKinnonn, V.A. 1998. Supramolecular synthesis of organic laminates with affinity for aromatic guests: a new class of clay mimics. *J. Am. Chem. Soc.* 120: 11894–11903; Arduini, A., Nachtigall, F., Pochini, A. 2000. *Supramol. Chem. 12*: 273–291; Takahashi, H., Tsuboyama, S., Umezawa, Y. et al. 2000. CH/π interactions as demonstrated in the crystal structure of host/guest compounds. A database study. *Tetrahedron* 56: 6185–6191.
35. Umezawa, K., Kawakami, M., Watanabe, T. 2003. Molecular design and biological activities of protein-tyrosine phosphatase inhibitors. *Pharmacology & Therapeutics* 99: 15–24.
36. Kinbara, K., Oishi, K., Harada, Y. et al. 2000. Effect of a substituent on an aromatic group in diastereomeric resolution. *Tetrahedron* 56: 6651–6655; Fujii, I., Hirayama, N. 2002. Chiral space formed by (+)-(1S)-1,1′-binaphthalene-2,2′-diyl phosphate: recognition of aliphatic L-α-amino acids. Helv. Chim. *Acta 85*: 2946–2960; H. Suezawa et al. 2004. Aromatic CH/π hydrogen bond as an important factor deciding the relative stability of diastereomeric salts relevant to enantiomeric resolution. A crystallographic database study. *Eur. J. Org. Chem.* 23: 4816–4822.
37. Hashihayata, T., Punniyamurthy, T., Irie, R. et al. 1999. Conformational analysis of cationic (*R*,*S*)- and (*R*,*R*)-(salen)manganese complexes possessing axial chirality

as a chiral element based on x-ray crystallography: an explanation of the effect of apical ligand on enantioselection by (salen)manganese catalyst. *Tetrahedron* 55: 14599–14610; Yamakawa, M., Yamada, I., and Noyori, R. 2001. CH/π attraction: the origin of enantioselectivity in transfer hydrogenation of aromatic carbonyl compounds catalyzed by chiral 6-arene-ruthenium(II) complexes. *Angew. Chem. Int. Ed.* 40: 2818–2821.

38. M. Soduple et al. 1997. *J. Am. Chem. Soc.* 119: 4332–4238; Ujaque, G., Lee, P.S., Houk, K.N. et al. 2002. The origin of *endo* stereoselectivity in the hetero-Diels-Alder reactions of aldehydes with *ortho*-xylylenes: CH-π, π-π, and steric effects on stereoselectivity. *Chem. Eur. J.* 8: 3423–3430.
39. Obata, T., Shimo., T., Yasutake, M. et al. 2001. Remarkable interaction effects of molecular packing on site- and stereoselectivity in photocycloaddition of 2-pyrones with maleimide in the solid state. *Tetrahedron* 57: 1531–1541; Nagahama, S., Inoue, K., Sada, K. et al. 2003. Two-dimensional hydrogen bond networks supported by CH/π interaction leading to a molecular packing appropriate for topochemical polymerization of 1,3-diene monomers. *Crystal Growth & Design* 3: 247–256.
40. Umezawa, Y., Nishio, M. 1998. CH/π interactions in the crystal structure of guanine-nucleotide binding proteins, Src Homology-2 Domains and human growth hormone in complex with their specific ligands. *Bioorg. Med. Chem. 6*: 493–504.
41. Umezawa, Y., Nishio, M. 1998. CH/π interactions in the crystal structure of class I MHC antigens and their complexes with peptides. *Bioorg. Med. Chem.* 6: 2507–2515.
42. Umezawa, Y., Nishio, M. 2000. CH/π interactions in the crystal structure of TATA-box binding protein/DNA complexes. *Bioorg. Med. Chem.* 8: 2643–2650.
43. Muraki, M. 2002. The importance of CH/π in interactions to the function of carbohydrate binding proteins. *Protein and Peptide Lett.* 9: 195–209.
44. Y. Umezawa, M. Nishio. 2005. CH/π hydrogen bonds as evidenced in the substrate specificity of acetylcholine esterase. *Biopolymers* 79: 248–258.
45. Spiwok, V., Lipovova, P., Skalova, T. et al. 2004. Role of CH/π interactions in substrate binding by *Escherichia coli* beta-galactosidase. *Carbohydr. Res.* 339: 2275–2280.; Sujatha, M.S., Sasidhar, Y.U., Balaji, P.V. 2005. Insights into the role of the aromatic residue in galactose-binding sites: MP2/6-311G^{++}** Study on galactose- and glucose-aromatic residue analogue complexes. *Biochemistry.* 44: 8554–8562.
46. Sujatha, M.S., Sasidhar, Y.M., Balaji, P.V. 2007. MP2/6-311++G(d,p) study on galactose-aromatic residue analog complexes in different position-orientations of the saccharide relative to aromatic residue. *J. Mol. Struct (THEOCHEM)* 814: 11–24.
47. Harigai, M., Kataoka, M., Imamoto, Y. 2006. A single CH/π weak hydrogen bond governs stability and the photocycle of the photoactive yellow protein. *J. Am. Chem. Soc.*128: 10646–10647.
48. Fantini, J., Garmy, N., Yahi, N. 2006. Prediction of glycolipid-binding domains from the amino acid sequence of lipid raft-associated proteins: application to HpaA, a protein involved in the adhesion of *Helicobacter pylori* to gastrointestinal cells. *Biochemistry* 45: 10957–10962.
49. Chakrabarti, P., Samanta, U. 1995. CH/π interaction in the packing of the adenine ring in protein structures. *J. Mol. Biol.* 251: 9–14.
50. Brandi, M., Weiss, M.S., Jabs, A. et al. 2001. CH/π-interactions in proteins. *J. Mol. Biol.* 307: 357–377.
51. Nishio, M., Umezawa, Y., Hirota, M. et al. 1995. The CH/π interaction: significance in molecular recognition. *Tetrahedron* 51:8665–8701.
52. Nishinaka, T., Ito, Y., Yokoyama, S. et al. 1997. An extended DNA structure through deoxyribose-base stacking induced by RecA protein. *Proc. Natl. Acad. Sci. U.S.A.* 94: 6623–6628.; Chou, S.H, and Tseng, Y.Y. 1999. Cross-strandpurine-pyrimidine stack and sheared purine·pyrimidine pairing in the human HIV-1 reverse transcriptase

inhibitors. *J. Mol. Biol.* 285: 41–48; Umezawa, Y., Nishio, M. 2002. Thymine-methyl/pi interaction implicated in the sequence-dependent deformability of DNA. *Nucleic Acids Res.* 30: 2183–2192.
53. Ozawa, T., Tsuji, E., Ozawa, M. et al. 2008. The importance of CH/π hydrogen bonds in rational drug design: an *ab initio* fragment molecular orbital study to leukocyte-specific protein tyrosine (LCK) kinase. *Bioorg. Med. Chem.* 16: 10311–10318.
54. Susva, M., Missbach, M., Green, J. 2000. Src inhibitors: drugs for the treatment of osteoporosis, cancer or both? *Trends in Pharmacological Sciences* 21: 489–495.
55. Marth, J.D., Lewis, D.B., Cooke, M.P. et al. 1989. Lymphocyte activation provokes modification of a lymphocyte-specific protein tyrosine kinase (p56lck). *J. Immunol.* 142: 2430–2437.
56. Palacios, E.H., Weiss, A. 2004. Function of the Src-family kinases, Lck and Fyn, in T-cell development and activation. *Oncogene*. 23: 7990–8000.
57. Hanke, J.H., Gardner, J.P., Dow, R.L. et al. 1996. Discovery of a novel, potent, and Src family-selective tyrosine kinase inhibitor. *J. Biol. Chem.* 271: 695–701.
58. Trevillyan1, J.M., Chiou, G. Gr., Ballaron. S. et al. 1999. Inhibition of p56lck tyrosine kinase by isothiazolones. *Arch. Biochem. Biophys.* 364:19–29.
59. Snow, R.J., Cardozo, M.G., Morwick. et al. 2002. Discovery of 2-phenylamino-imidazo[4,5-h]isoquinolin-9-ones: a new class of inhibitors of lck kinase. *J. Med. Chem.* 45: 3394–3405.
60. Goldberg, D.R., Butz, T., Cardozo, M.G. et al. 2003. Optimization of 2-phenylaminoimidazo[4,5-h]isoquinolin-9-ones: orally active inhibitors of lck kinase. *J. Med. Chem.* 46: 1337–1349.
61. Chen, P., Norris, D., Iwanowicz, E.J. et al. 2002. Discovery and initial SAR of imidazoquinoxalines as inhibitors of the Src-family kinase p56Lck. *Bioorg. Med. Chem. Lett.* 12: 1361–1364.
62. Chen, P., Iwanowicz, E.J., Norris, D. et al. 2002. Synthesis and SAR of novel imidazoquinoxaline-based Lck inhibitors: improvement of cell potency. *Bioorg. Med. Chem. Lett.* 12: 3153–3156.
63. Das, J., Lin, J., Moquin, R.V. et al. 2003. Discovery of 2-amino-heteroaryl-benzothiazole-6-anilides as potent p56lck inhibitors. *Bioorg. Med. Chem. Lett.* 13: 2145–2149.
64. Chen, P., Norris, D., Das, J. et al. 2004. Discovery of novel 2-(aminoheteroaryl)-thiazole-5-carboxamides as potent and orally active Src-family kinase p56lck inhibitors. *Bioorg. Med. Chem. Lett.* 14: 6061–6066.
65. Chen, P., Doweyko, A.M., Norris, D. et al. 2004. Imidazoquinoxaline Src-family kinase p56Lck inhibitors: SAR, QSAR, and the Discovery of (S)-N-(2-Chloro-6-methylphenyl)-2-(3-methyl-1-piperazinyl)imidazo- [1,5-a]pyrido[3,2-e]pyrazin-6-amine (BMS-279700) as a potent and orally active inhibitor with excellent in vivo antiinflammatory activity. *J. Med. Chem.* 47: 4517–4529.
66. Arnold, L.D., Calderwood, D.J., Dixon, R.W. et al. 2000. Pyrrolo[2,3-d]pyrimidines containing an extended 5-substituent as potent and selective inhibitors of lck II. *Bioorg. Med. Chem. Lett.* 10: 2167–2170.
67. Burchat, A.F., Calderwood, D.J., Friedman, M.M. et al. 2002. Pyrazolo[3,4-d]pyrimidines containing an extended 3-substituent as potent inhibitors of Lck-a selectivity insight. *Bioorg. Med. Chem. Lett.* 12: 1687–1690.
68. Borhani, D.W., Calderwood, D.J., Friedman, M.M. et al. 2004. A-420983: a potent, orally active inhibitor of LCK with efficacy in a model of transplant rejection. *Bioorg Med. Chem. Lett.* 14: 2613–2616.
69. Zhu, X., Kim, J.L., Newcomb, J.R. et al. 1999. Structural analysis of the lymphocyte-specific kinase Lck in complex with non-selective and Src family selective kinase inhibitors. *Structure* 7: 651–661.

70. Chery, M., Williams, D.H. 2004. Recent kinase and kinase inhibitor X-ray structures: mechanisms of inhibition and selectivity insights. *Current Med. Chem.* 11: 663–673.
71. Mukaiyama, H., Nishimura, T., Kobayashi, S. et al. 2007. Synthesis and c-Src inhibitory activity of imidazo[1,5-a]pyrazine derivatives as an agent for treatment of acute ischemic stroke. *Bioorg Med Chem.*15: 868–885.
72. ABINIT-MP. Available at www.ciss.iis.u-tokyo.ac.jp/rss21/result/download/index.php#download_2.
73. Lipinski, C.A., Lombardo, F., Dominy, B.W. 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev.* 46: 3–26.
74. Congreve, M., Carr, R., Murray, C. 2003. A "Rule of Three" for fragment-based lead discovery? *Drug Discov. Today* 8: 876–877.
75. Sarkhel, S., Desiraju, G.R. 2004. NH/O, OH/O, and CH/O hydrogen bonds in protein-ligand complexes: strong and weak interactions in molecular recognition. *PROTEINS: Structure, Function, and Bioinformatics* 54: 247–259.
76. Pierce, A.C., Sandretto, K.L., Bemis, G.W. 2002. Kinase inhibitors and the case for CH/O hydrogen bonds in protein–ligand binding. *PROTEINS: Structure, Function, and Bioinformatics* 49: 567–576.
77. Pierce, A.C., Haar, E., Binch, H.M. 2005. CH···O and CH···N hydrogen bonds in ligand design: a novel quinazolin-4-ylthiazol-2-ylamine protein kinase inhibitor. *J. Med. Chem.* 48: 1278–1281.
78. Ehama, R., Tsushima, M., Yuzuri, T. et al. 1993. Substituent effect on the enthalpies of formation of CH/π complexes of aromatic π-bases. *Bull. Chem. Soc. Jpn.* 66: 814–818.
79. Suezawa, H., Mori, A., Sato, M. et al. 1993. Evidence for the presence of the CH/π interacted ap-conformers of benzyl formats. *J. Phys. Org. Chem.* 6: 399–406.
80. Chowdhury, S.K., Joshi, V.S., Samuel, A.G. 1994. Conformational preferences in molybdenum(II) π-allyl complexes: role of CH/π interaction. *Organometallics* 13: 4092–4096.
81. Lee, E.C., Hong, B.H., Lee, J.Y. 2005. Substituent effects on the edge-to-face aromatic interactions. *J. Am. Chem. Soc.,* 127: 4530–4537.
82. Coffey, A.J., Brooksbank, R.A., Brandau, O. et al. 1998. Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. *Nat. Genet.* 20: 129–135.
83. Sayos, J., Wu, C., Morra, M. et al. 1998. The X-linked lymphoproliferative-disease gene product SAP regulates signals induced through the co-receptor SLAM. *Nature* 395: 462–469.
84. Roussel, R.R., Brodeur, S.R., Shalloway, D. et al. 1991. Selective binding of activated pp60c-src by an immobilized synthetic phosphopeptide modeled on the carboxyl terminus of pp60c-src. *Proc. Natl. Acad. Sci. U.S.A.* 88: 10696–10700.
85. Poy, F., Yaffe, M.B., Sayos, J. et al. 1999. Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition. *Mol. Cell* 4: 555–561.
86. Ozawa, T. Okazaki, K. 2008. CH/π hydrogen bonds determine the selectivity of the Src homology 2 (SH2) domain to tyrosine phosphotyrosyl peptides: an *ab initio* fragment molecular orbital (FMO) study. *J. Comput. Chem.* 29: 2656–2666.

11 Modeling a Protein Environment in an Enzymatic Catalysis
A Case Study of the Chorismate Mutase Reaction

Toyokazu Ishida

CONTENTS

11.1 INTRODUCTION: ENZYME REACTION

Enzyme reactions are crucial for most biological processes.[1–4] The most important properties of enzymes are generally their catalytic power, substrate specificity, and capacity for regulation. Catalytic acceleration by enzymes is typically in the range of 3 to 9 orders of magnitude, and sometimes up to the order of 15, which is far greater

than the corresponding chemical reactions in an aqueous environment. Therefore, understanding the molecular origin of enzyme catalyses is one of the fundamental problems in modern biological science. Although a general catalytic principle was proposed over half a century ago in the simple statement that *enzymes bind the transition state more tightly than the ground state*, atomistic details of the transition state stabilization (TSS) have not yet been clarified.

To understand the mechanism of enzymatic reactions, it is best to consider an energetic relationship based on simple Michaelis–Menten kinetics. Within the framework of this kinetic equation, the following reaction scheme is assumed:

$$\mathrm{E} + \mathrm{S} \longleftrightarrow \mathrm{ES} \longrightarrow \mathrm{E} + \mathrm{P} \tag{11.1}$$

where E denotes the enzyme, S is the substrate, ES is the enzyme-substrate complex, and P is the product of the reaction. Under the steady-state assumption, the initial rate of reaction can be expressed as

$$v_0 = \frac{k_{cat}[\mathrm{S}][\mathrm{E}]_{total}}{K_M + [\mathrm{S}]} \tag{11.2}$$

where k_{cat} is the turnover number, defined as the number of substrates converted into product per enzyme per unit time when the enzyme is saturated with the substrate, and K_M is the Michaelis constant, which is proportional to the binding affinity of the substrate. When $K_M >> [\mathrm{S}]$ or the concentration of the free enzyme is approximately equal to $[\mathrm{E}]_{total}$, the reaction rate is expressed as

$$v_0 \approx \frac{k_{cat}}{K_M}[\mathrm{E}][\mathrm{S}] \tag{11.3}$$

Here, the factor k_{cat}/K_M is a pseudo second-order rate constant. While k_{cat} represents the kinetic efficiency of enzymes, k_{cat}/K_M is an index of the catalytic efficiency. It is generally believed that enzymes evolved by optimizing the efficiency of k_{cat}/K_M, rather than k_{cat}.[5]

To discuss the rate enhancement of enzyme reactions, we need to define a reference reaction; corresponding chemical reactions in the aqueous phase are considered ideal counterparts in most cases.[6]

Figure 11.1 summarizes the relationship between the kinetic constants and related energy differences. This figure illustrates free energy changes in enzyme reactions compared with reactions in a reference solvent cage. Because the binding effect is generally obvious in many cases, the major difficulties lie in determining the crucial factors responsible for the energetic difference between $\Delta g^{\neq}_{cat}$ and $\Delta g^{\neq}_{cage}$. Until now, many appealing hypotheses have been proposed to explain the catalytic power of enzymes, some of which are listed below[1–6]:

1. *Steric strain*: A strain force that fixes the substrate inside the enzyme can reduce the activation energy by destabilizing the distorted geometry of the ground state.

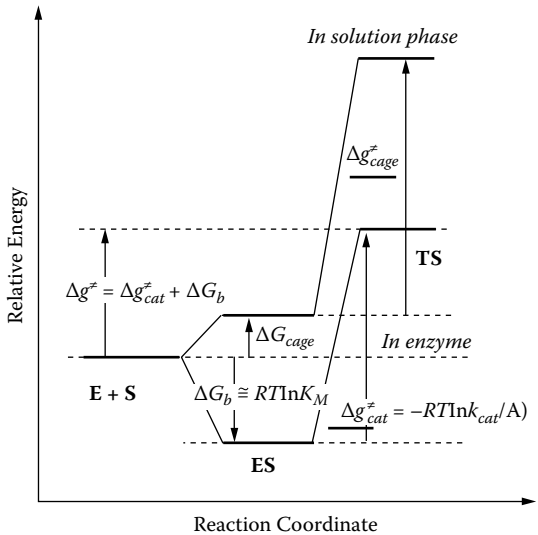


FIGURE 11.1 Free energy relationship between the enzyme catalysis and corresponding reference reaction.

2. *Desolvation*: A nonpolar, gas phase–like protein environment can reduce the activation energy by destabilizing the highly charged ground state.
3. *Entropic factor*: A restriction of a particular mode of reactive fragments at the enzyme active site can reduce the activation energy.
4. *Dynamic fluctuation*: A special kind of fluctuation induced by enzyme motions can decrease a transmission factor far below unity.
5. *Low-barrier hydrogen bond*: A short, strong hydrogen bond formed in the transition state will reduce the activation energy by stabilizing the catalytic transition state.

Although all of these proposals have certain appealing aspects, it is generally difficult to confirm the validity of these hypotheses from an atomistic viewpoint. For example, the ground state destabilization by steric strain or the desolvation effect cannot explain a large increase in the k_{cat}/K_M factor: this effect leaves k_{cat}/K_M constant while k_{cat} and K_M are changed.

Owing to the rapid progress of modern chemical theory and computational architecture, many types of theoretical calculations have been reported for various enzymatic reactions in the last few decades. Among many proposals, the combined quantum mechanical (QM)/molecular mechanical (MM) approach has gained wide popularity.[7–14] Although the validity and limitations of the QM/MM treatment are still unclear, it has been applied to many applications for studying various enzymes, and now, it is generally accepted that one key factor of enzyme reactions is the electrostatic stabilization for TS binding.[15] Considering these situations, in this chapter we summarize recent theoretical studies that demonstrate a systematic modeling

approach to investigating the polar protein environment in enzyme reactions.[16,17] By focusing on one of the well-known systems, *chorismate mutase*, as a typical enzymatic system, we carefully discuss the procedure of QM/MM modeling, accompanied by detailed analysis by FMO computations. However, the main focus of this chapter is to demonstrate and establish an *ab initio*–based modeling approach to investigate large enzymatic systems, not to discuss the entire catalytic event comprehensively and to pursue a reproduction of the experimental energetics.

11.2 EXAMPLE: *CHORISMATE MUTASE* REACTION

Chorismate mutase (CM) is an essential enzyme in the biosynthetic route of aromatic amino acids (e.g., phenylalanine and tyrosine) in bacteria, fungi, and plants. It catalyzes the skeletal rearrangement of chorismate into prephenate (see Figure 11.2), a rare example of biocatalyzed cycloaddition.[18]

In contrast to most enzymatic systems where reactions proceed along a complicated, multistep pathway, CM exhibits several unique features in its reaction: (1) a simple, one-step intramolecular isomerization occurs without forming chemical bonds with amino acid residues inside the protein; (2) a reference reaction in the solution phase occurs readily in a similar chemical manner; and (3) many kinetic parameters have already been measured for both enzyme and solution phase reactions.[19–22] Mainly due to these unique features and its advantage for modeling reaction processes, researchers performed several types of QM/MM and related calculations to explore the catalytic mechanism of CM.[23–42]

One standard approach of QM/MM calculations is to investigate an origin of transition state stabilization (TSS) based on reaction path modeling of *Bacillus subtilis* (Bs) CM reactions.[43] Lyne et al. applied the semiempirical QM/MM method and

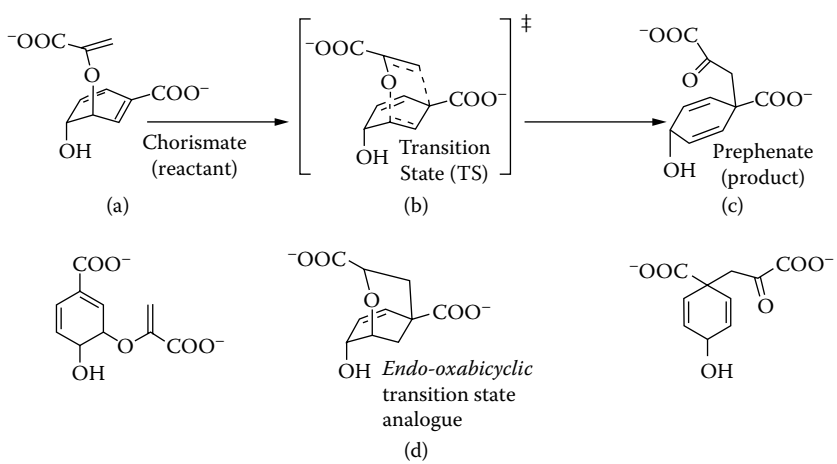


FIGURE 11.2 Reaction scheme of the interconversion of chorsimate (a) into prephenate (c), catalyzed by BsCM. The chemical structures of putative TS (b) and TSA (d) are also shown in this scheme.

showed that the enzymatic transition state (TS) binds strongly to Arg90 and Glu78 residues.[23] Worthington et al. identified Glu78 and Tyr108 as critical for catalysis, and Arg90 contributes equally to both the reactant and TS stabilization.[26] Lee et al. showed that Arg7, Glu78, and Arg90 play a catalytic role due to differential stabilization along the reaction path.[27] By employing a nonempirical variation–perturbation approach, Szefczyk et al. evaluated the interaction energy between the substrate and selected residues for the active site cluster.[29] In general, most of these calculations support the concept of electrostatic stabilization of the TS.[23,25,27–29,40,42]

At present, theoretical approaches for enzyme reactions are commonly limited to QM/MM calculations, wherein large protein structures are described by empirical force fields. To advance the current status of QM/MM treatment, we proposed a computational strategy that combines all-electron FMO analysis with *ab initio* QM/MM modeling and analyzed the interaction energy components during the BsCM reaction process.[17] The main discussion in the following sections is to clarify the role of the polar protein environment, which is considered the major origin of electrostatic stabilization, rather than to model the entire catalytic event comprehensively.

11.3 COMPUTATIONAL METHODS AND MODELING

Basically, we followed the systematic modeling procedures that follow. First, we constructed a reliable molecular model using *ab initio* QM/MM reaction path searches. To examine the influence of geometrical parameters on reaction energetics, we prepared two types of QM/MM models (small-QM- and large-QM-region) and compared the calculated properties. Second, to estimate the contribution of the free energy factor in the aqueous phase, we performed QM/MM combined with MD-FEP simulations and compared the free energy profile of the enzymatic reaction with that of the model reference reaction. (However, due to the space limitations of this chapter, we must skip the contents of free energy based modeling and the discussion of its results. For details, one may refer to the recent article by Ishida 2008). Finally, after confirming the reliability of two-body FMO calculations for the prepared QM/MM models, we performed detailed analyses of the energy components and compared the properties calculated by the QM/MM and FMO methods.

11.3.1 Reaction Path Modeling by *Ab Initio* QM/MM Calculations

In the standard QM/MM approach, the total energy of an entire enzyme complex is calculated based on the following definition:

$$\hat{H}_{tot} = \hat{H}_{QM} + \hat{H}_{QM/MM} + \hat{H}_{MM} \tag{11.4}$$

where $\hat{H}_{QM}$ is the Hamiltonian of the QM region, $\hat{H}_{MM}$ is that of the MM region, and $\hat{H}_{QM/MM}$ is the Hamiltonian that describes the interactions between the QM and MM regions. The active site region, which accompanies a change of electronic structure in the reactions, is determined by quantum mechanical calculations, and the surrounding environment, which does not require a description of the bond formation or breakdown, can be represented by molecular mechanics force fields. As for the

QM/MM interaction energy, we consider the following interaction energy terms in this study[44–46]:

$$\hat{H}_{QM/MM} = \hat{H}^{elec}_{QM/MM} + \hat{H}^{vdw}_{QM/MM} + \hat{H}^{strain}_{QM/MM} \tag{11.5}$$

where $\hat{H}^{elec}_{QM/MM}$ describes the electrostatic interaction between the QM and MM regions, $\hat{H}^{vdw}_{QM/MM}$ represents the van der Waals interaction, and $\hat{H}^{strain}_{QM/MM}$ is a strain term at the QM/MM boundaries. The electrostatic interaction is evaluated by the one-electron operator in the electronic structure calculations:

$$\hat{H}^{elec}_{QM/MM} = -\sum_{MM}\sum_{i}\frac{Z_{MM}}{\mathbf{r}_i - \mathbf{R}_{MM}} + \sum_{MM}\sum_{QM}\frac{Z_{QM}Z_{MM}}{\mathbf{R}_{QM} - \mathbf{R}_{MM}} \tag{11.6}$$

where $\mathbf{r}_i$ denotes the electron coordinate and Z_i means the nuclear (or MM atomic site) charge. The last two terms in Equation 11.5 are approximated by the classical force field. $\hat{H}^{strain}_{QM/MM}$ includes the bonded terms at the boundary (bond, bend, torsion, and improper torsion energy terms). We employed the AMBER parameter set (parm.96) to describe the MM region because of its *ab initio* MO-based parameterization nature for the charge model, which seems to be consistent with the present QM/MM framework.[47–50]

For the preparation of the protein model, we adopted the initial coordinates of the enzyme from the X-ray crystal structure of BsCM bound to the *endo-oxabicyclic* transition state analogue (TSA), PDB code 2CHT.[43] Among the four trimers observed in the X-ray structure, only the first one was adopted, and a single binding site (between domains 1 and 2) was considered. Considering the physiological conditions in which the enzyme works, we assumed all polar residues (Asp, Glu, Lys, and Arg) to be in their ionized states and all His to be in the protonated form, as assigned in the original PDB file. Hydrogen atoms were added to the enzyme–substrate (ES) complex in the standard manner. We placed the chorismate structure at the active site to retain maximum overlap with the original X-ray coordinates of the TSA. In the initial model, both counterions and crystal water were not considered for simplicity. Unfavorable steric contact was removed by initial rough MM energy minimization (steepest descent optimization with ~500 cycles). Then, we refined the structure of the ES complex by *ab initio* QM/MM geometry optimizations. To examine the geometrical and energetic effects on selecting the QM region, we prepared two QM/MM partition models. In the small-QM-region model, the QM region was limited only to the substrate for simplicity. In this case, the steric strain term that links the QM/MM boundaries in Equation 11.5 was not required. While in the large-QM-region model, the side chain atoms of Arg7, Glu78, and Arg90 were added to the QM region. In the latter case, linked hydrogen atoms saturated the boundaries between QM/MM regions ($C_\beta - C_\lambda$), and these capping atoms were allowed to move freely during the QM/MM optimizations.

In the QM/MM geometry optimizations (reaction path search and structural refinement), we used the restricted Hartree–Fock (RHF) method with the 6-31(+)G** basis set (diffuse functions were added only to the two carboxylic groups).

Note that this provides a consistent treatment of electrostatic interaction with atomic charges in the MM region, which are determined by the RESP method at the HF/6-31G* level.[47,48] First, we determined the reaction path using the small-QM model. In this case, the reaction coordinate is defined as a linear combination of two bond distances: one for breaking C–O and the other for forming C–C bonds (see Figure 11.2). The convergence criteria used in *ab initio* QM/MM optimizations were as follows: the maximum gradient of 1×10^{-4} au in the QM region and the RMS gradient less than 1×10^{-2} kcal/mol/Å in the MM region. After locating the TS position on the potential energy surface, we calculated the minimum energy reaction path (MEP) that connects the TS to the reactant and the product. Restrained geometry optimizations were performed in the internal coordinate space with two selected bond distances, which were kept constant. The MM region was fully relaxed in each step, and no cutoff was introduced between the QM and MM interactions. In the second stage, we refined the molecular structures by employing the large-QM model. In this process, we reoptimized the QM/MM models along the selected reaction coordinates. All intramolecular geometries for the entire complex, except the reaction coordinate, were optimized. Based on these two calculated models, we evaluated and compared the geometrical parameters and energetics. Considering the system size used in the enzymatic studies, generally MP2 and DFT are the first choices to account for electron correlation. Here, we performed MP2 and DFT (with B3LYP functional) calculations for the optimized geometries along the reaction coordinate.

11.3.2 All-Electron QM Analysis by Fragment Molecular Orbital Calculations

Because theoretical formulations of FMO methodologies are discussed in detail in Chapters 1 and 2[51–53]; in this chapter, we will comment briefly on the practical details for enzymatic applications. Considering the system size of BsCM structure (up to ~6,000 atoms), we employed the two-body expansion in all FMO calculations (denoted as FMO2). As mentioned in Section 11.1, one of the most important factors is the electrostatic interaction inside a polar protein environment. To clarify differences between the QM/MM and the FMO treatments, we focused on the interaction energy components defined in FMO model. For this purpose, we evaluated ΔE_{IJ} using the isolated and complex electron densities of the protein complex (enzyme + substrate), where the former indicates the electrostatic interaction of the unperturbed (isolated) density distribution and the latter indicates the density distribution mutually polarized inside the protein complex. In contrast to the standard QM/MM approach that uses a fixed point charge on each MM atomic site, in the FMO2 formulation, the electrostatic interaction component is evaluated in the following *ab initio* MO manner:

$$\Delta E_{IJ}^{(A)} = \sum_{\mu\nu\in I} \mathbf{D}_{\mu\nu}^{I,(A)} \sum_{\alpha\in J} \langle \mu | \frac{-Z_\alpha}{|\mathbf{r}-\mathbf{R}_\alpha|} | \nu \rangle + \sum_{\rho\sigma\in J} \mathbf{D}_{\rho\sigma}^{J,(A)} \sum_{\beta\in I} \langle \rho | \frac{-Z_\beta}{|\mathbf{r}-\mathbf{R}_\beta|} | \sigma \rangle + \sum_{\mu\nu\in I} \sum_{\rho\sigma\in J} \mathbf{D}_{\mu\nu}^{I,(A)} \mathbf{D}_{\rho\sigma}^{J,(A)} (\mu\nu|\rho\sigma) + \Delta E_{IJ}^{nuc} \tag{11.7}$$

where (A) designates the *isolated* or *complex* state of the system, and *I* and *J* indicate the substrate and amino acid residues, respectively.

As for the structural models used in the FMO computations, we mainly used the large-QM-region model for a detailed discussion of the reaction energy profile. To compare the geometric effects on the reaction energetics, we also evaluated the activation energy using the small-QM-region model. In general, in the fragment-based method, larger fragments have a smaller error compared with the reference *ab initio* MO methods. To confirm the numerical accuracy of the energetics in the BsCM reaction, we compared two partition methods (one or two amino acid residues per fragment) for the RHF/DFT calculations. Note that the protein was divided into fragments at the C_α position, keeping the peptide bonds intact. The hybrid sp^3 orbitals of the carbon atom were used to divide molecular orbital space at the bond fraction points appropriately. The substrate was treated as a single fragment in all FMO2 calculations. In such a case, we directly applied the interaction energy decomposition analysis within the FMO2 model, when one amino residue per fragment partition was employed.

We performed all FMO calculations using the GAMESS implemented version,[52–54] and prepared general input data for FMO calculations using the FMOutil modeling utility (included in the distributed version of GAMESS). Technical details of FMO2 computations are as follows. Both the atomic and molecular orbital accuracies were increased to 10^{-12} using ICUT = 12, ITOL = 24, and CUTOFF = 10^{-12}, and the SCF convergence was tightened to 10^{-7}. The same values were used during the monomer SCF cycle where monomer densities converge. Because using diffuse functions in the fragment-based methods often results in problems, we employed the 6-31G* basis set in all FMO2 calculations. The option to remove *s* contaminants from *d* functions was employed. As for the treatment of the surrounding electrostatic potential (ESP), we employed the following approximations: RESPPC = 2.0, RESDIM = 2.5, and RESPAP = 0 (we never used the atomic population approximation for the ESP). In FMO2-MP2/6-31G* calculations, we used the approximation of RCORSD = 2.0. The core orbitals were not correlated (1*s* orbitals for carbon, nitrogen, and oxygen atoms). In the DFT calculations, we employed the B3LYP functional based on the standard set of functional parameters in GAMESS and used a dense grid ($96 \times 20 \times 40$) to improve the numerical accuracy.

11.4 COMPUTATIONAL RESULTS

11.4.1 QM/MM Modeling: Structures and Energy Profiles

The main purpose of QM/MM modeling is to prepare a reliable reaction path model to enable all-electron energy components analysis within a moderate computational time. In such a case, the minimum requirement for QM/MM modeling is to qualitatively describe the energy profiles (e.g., the location of the TS, the relative energetics, etc.) along the selected reaction coordinate. To confirm this important point, we first compared the difference of calculated energy profiles and geometrical parameters between the small-QM and large-QM region models. Table 11.1 summarizes the calculated activation energies in the two models.

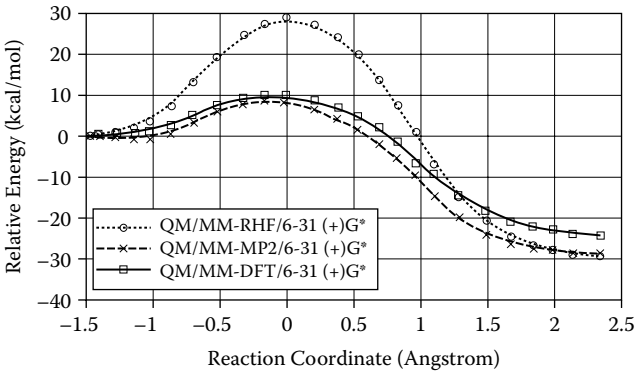
TABLE 11.1
Activation Energies by *Ab Initio* QM/MM and FMO Methods with Two Types of QM Region Definition (in kcal/mol)

	Small-QM Region Model	Large-QM Region Model
QM/MM RHF/6-31(+)G*	25.3	28.8
QM/MM MP2/6-31(+)G*	8.9	9.5
QM/MM B3LYP/6-31(+)G*	7.8	9.8
FMO2 RHF/6-31G*	29.5	28.3 (27.7)
FMO2 MP2/6-31G*	7.4	8.1 (---)
FMO2 B3LYP/6-31G*	9.6	9.5 (9.3)

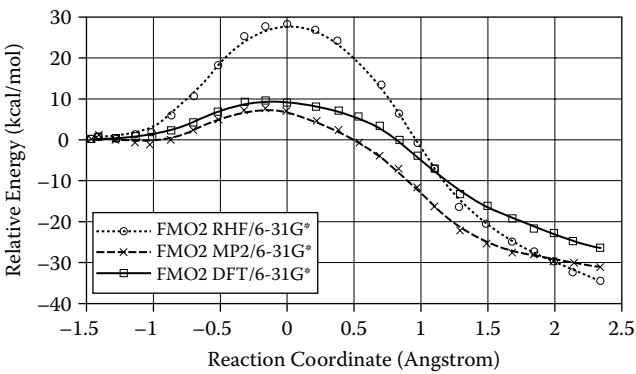
Notes: In the small-QM region model, the QM region contains the substrate only, while in the case of the large-QM region model, the substrate and the side chain atoms of Arg7, Glu78, and Arg90 (all located on the same domain) are included in the QM region. Numerical data of two residues per one fragment are listed in parentheses. In this case, MP2 calculations were not performed due to hardware limitations.

Although the calculated activation energies in the small-QM model are slightly smaller than those in the large-QM model, all the energy curves (RHF/MP2/B3LYP) reproduce qualitatively similar energy profiles along the distinguished reaction coordinate. As a representative example, Figure 11.3a shows the potential energy profiles of the large-QM-region model. For a discussion of the following FMO computations under equal conditions, a summary of three types of theoretical calculations is provided (RHF/MP2/B3LYP) in this figure. Comparing the MP2/DFT profiles with the RHF profile shows that the location of the highest energy point is shifted slightly toward the reactant side. This trend is precisely observed in the energy profiles using the small-QM model. In our discussion, we defined this highest MP2-energy geometry as the TS structure and employed this structure for all energy component analyses mentioned below.

As for geometrical parameters, Figure 11.4a includes the optimized structures of the hydrogen-bonding network around the active site, and Figure 11.4c shows the active site geometry of the TS. In this schematic picture, only the hydrogen-bonding distances between the TS geometry and the surrounding residues in both small-QM and large-QM models are summarized. Several previous calculations suggested that the effects of including some amino acid residues in the QM region are small.[27,30,41] In general, this is true in our QM/MM modeling results. By comparing the hydrogen-bonding geometries of the large-QM model with those of the small-QM model, however, we found an apparent difference in the Arg7-substrate and Arg90-substrate interactions. By including the side-chain atoms of Arg7/Arg90 (both on the domain 2 unit) in the large-QM model, the hydrogen-bonding distances were slightly extended, and we obtained more relaxed geometries compared with those obtained in the small-QM model. In contrast to these structural changes, the hydrogen-bonding alignment in Glu78-Arg90-substrate is constant despite adding the side-chain atoms of Glu78 to the QM region.



(a)



(b)

FIGURE 11.3 Potential energy profiles along the distinguished reaction coordinate in each theoretical method and model: (a) QM/MM reaction profiles with the large-QM region model, and (b) FMO2 reaction profiles with the large-QM region model. Each graph shows three types of theoretical calculations (RHF/MP2/B3LYP). The x-axis designates the reaction coordinate defined by the linear combination of C–O and C–C bond lengths (in angstrom). The y-axis shows the relative energy of the reaction (kcal/mol).

11.4.2 All-Electron QM Calculations 1: Reaction Energetics

Here, we discuss the reaction energetics and energy profiles obtained by all-electron QM calculations. As mentioned before, we used the QM/MM optimized geometries along the same reaction coordinate in all FMO2 calculations. For comparing the FMO results with QM/MM results under the same conditions, we performed three types of theoretical calculations clarifying the effect of the electron correlation: RHF/6-31G*, MP2/6-31G*, and B3LYP/6-31G*. For the analyses of energy curves and interaction energy components, we employed the large-QM region model. Figure 11.3b shows the calculated energy profiles, and Table 11.1 lists the activation energies of both the small- and large-QM models.

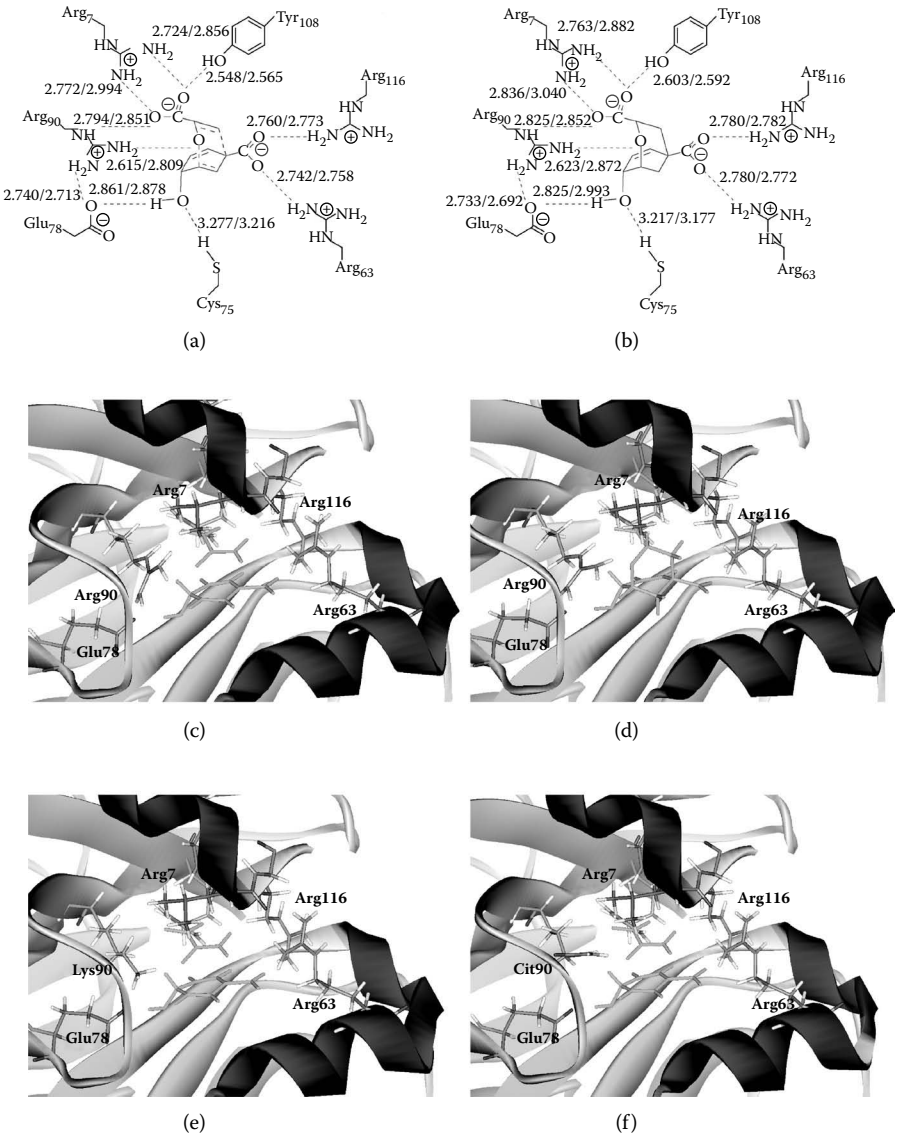


FIGURE 11.4 (Top panel) (a) The hydrogen-bonding network at the active site of wild-type BsCM. The left data show the optimized geometries in the small-QM region model, while the right data are those in the large-QM region model. (b) The same schematic picture of the enzyme-TSA complex (EI). (Middle panel) (c) QM/MM optimized structure of the enzymatic TS in the wild-type BsCM active site. Only important residues around the active site are drawn as sticks. The green stick designates the TS geometry of the substrate. (d) QM/MM optimized structure of the EI complex. (Bottom panel) (e) QM/MM optimized structure of the TS in Lys90 mutant reaction. (f) QM/MM optimized structure of the TS in Cit90 mutant reaction. **(See color insert following page 117.)**

Figure 11.3b shows the potential energy profiles of FMO2-RHF/MP2/B3LYP. As clearly shown in these results, the overall reaction profiles (a global shape of energy curves, the location of the TS geometry, and the calculated energetics) of both the QM/MM and FMO2 methods show similar behavior. When we compared the activation energies of the FMO2 results with the QM/MM results, the calculated values based on the large-QM model show similar results in all three types of theoretical methods, while those based on the small-QM model show a slight deviation. This result implies that for predicting the molecular properties of large protein complexes reliably, more accurate QM/MM modeling is required, especially for all-electron FMO computations. Compared with the QM/MM calculations, FMO2 energetics seem rather sensitive to the geometrical parameters, especially to the polar hydrogen-bonding alignment around the reactive site.

Table 11.1 summarizes the FMO2-RHF/6-31G* activation energies obtained by one and two residues per fragment partitions. In the case of fragment-based calculations, fewer fragments generally reproduce more accurate energetics. The activation energies in the two partition models are similar. From this result, we can see that the one residue per fragment partition, which is more suitable for the residue-based analysis, can be used reliably for the present purpose. Because of a large computational cost, we have not performed the two residues per fragment calculations for the MP2/6-31G* calculations. Comparing the FMO2-MP2 values with the other results, the activation energies show a slightly lower value. One possible reason is an overestimation of the electron correlation effect at the MP2 level. More electrons are spatially close at the TS geometry of this chemical reaction, and as a result, a significant correlation energy is observed relative to the reactant.[42] Table 11.1 also includes the B3LYP/6-31G* activation energies. Again, both partition models show similar energetics. In general, it is known that DFT cannot describe the dispersion interaction accurately. However, because the essential catalytic component of this reaction is the electrostatic interaction at the catalytic center and DFT implicitly includes the electron correlation effect, the calculated activation energies are closer to the experimental data than those from other methods. In this case, it is preferable to compare the activation enthalpy, which is measured to be ~12 kcal/mol.[22] The present FMO2-B3LYP result provides a reasonable estimation.

11.4.3 All-Electron QM Calculations 2: Interaction Residue Analysis

After confirming the validity of QM/MM modeling from both structural and energetic viewpoints, we analyzed the interaction energy between the substrate and the entire protein matrix. Although this type of analysis is popular in QM/MM treatment,[23,26–29] most of these are essentially empirical and inherit a limitation of the force fields. Considering these situations, we first compared the QM/MM result with the FMO results. For this purpose, we employed the small-QM-region model and redefined the amino acid residues in the QM/MM model in the FMO manner (the protein was divided into units at the C_α position). As described in the earlier chapters of this book, the pair interaction energies (ΔE_{IJ}) in FMO2 are evaluated under the converged electron density of the entire protein complex, and both the polarization and the charge transfer in the system are incorporated explicitly into the results.

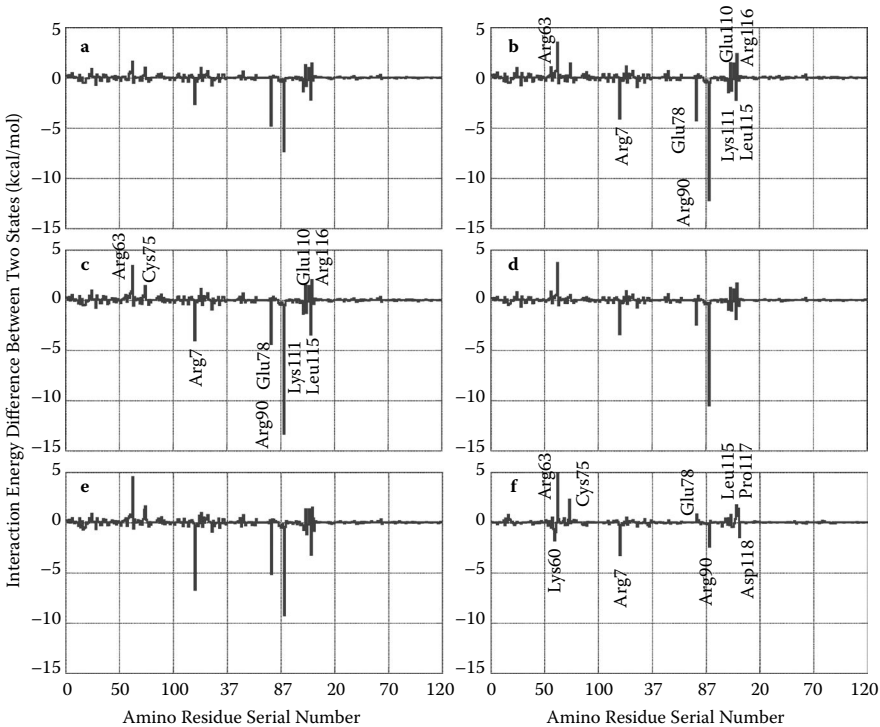


FIGURE 11.5 Interaction residue analysis of the wild-type BsCM reaction. The interaction energy difference between TS and ES of the wild-type reaction using the small-QM region model by (a) *ab initio* QM/MM-RHF/6-31(+)G**/AMBER level calculation, (b) FMO2-RHF/6-31G*, (c) FMO2-MP2/6-31G*, and (d) FMO2-B3LYP/6-31G*. (e) The interaction energy difference between the TS and ES of the wild-type reaction using the large-QM region model by FMO2-MP2/6-31G*. (f) The interaction energy difference between the EI and ES of the wild-type BsCM using the large QM-region model by FMO2-MP2/6-31G*. The x-axis indicates the amino acid residue serial number in this calculation, we used the first trimer structures (protein A+B+C) observed in the original X-ray data. The residue number of this figure reflects the serial number in each protein unit), and the y-axis designates the interaction energy difference between the two states (in kcal/mol).

Figure 11.5 shows the interaction energy differences in each theoretical method. First, we compare the energetic components by FMO2-RHF (see Figure 11.5b) with those by QM/MM-RHF (see Figure 11.5a). Although the contributions of each residue are generally similar in both calculations, FMO2-RHF shows stronger affinity for several residues. In particular, two polar residues, Arg90 and Arg7 (both located in the same domain), exhibit an apparent catalytic contribution. The difference between QM/MM and FMO is mainly due to the force field treatment of these polar residues around the active site. In contrast to the tight hydrogen bondings summarized in Figure 11.4a, no crucial TSS contribution is observed between the substrate and Arg63/Arg116 residues. In the MP2 case (see Figure 11.5c), the overall

profile is similar except the contribution from nonpolar Leu115 in the TS formation process. Because Leu115 is located at the lid position of the catalytic pocket, the result implies a nonnegligible effect of Leu115 on substrate binding. As for the B3LYP result (see Figure 11.5d), although each residual contribution is somewhat smaller than in RHF/MP2, the overall profile is similar to the other two FMO2 results. These comparative analyses clearly demonstrate that the catalytic contributions of Arg90, Glu78, and Arg7 are essentially of an electrostatic nature. Note that here we compared the energy *differences* between the two relative geometries and that both the QM/MM and FMO methods produced similar interaction energy diagrams. However, the absolute values of energy components are slightly different in several residues.

To examine the influence of geometrical parameters (on the selection of the QM region) on the energetics, we performed the same analyses using the large-QM region model. Although a slight difference between the small- and large-QM region models is observed in the interaction energy components of Arg7 and Arg90, the overall profile is similar to the result of the small-QM model. This result is consistent with the changes in the geometrical parameters summarized in Figure 11.4a. Here, Figure 11.5e shows the interaction energy diagram of FMO2-MP2 as a representative result of the large-QM model. Again, the present comparative analyses clearly revealed the electrostatic nature of the TSS in the BsCM reaction.

11.4.4 Probing Protein Environment 1: Transition State Analogue

One advantage of performing all-electron QM calculations is an explicit inclusion of the collective polarization inside the protein environment, which is difficult to achieve with other conventional QM/MM approaches. To further clarify the environmental effect of protein in the catalytic process, we also applied the same analysis for the enzyme-transition-state analogue (TSA) complex. Chemically stable analogues similar to the enzymatic transition state are widely used to probe the electronic environment of the active site. In the case of BsCM, X-ray structures bound to the *endo-oxabicyclic* TSA have been analyzed. Several kinetic measurements demonstrated that the analogue binds to the enzyme with a larger affinity, relative to the reactant.[55–59] Although several catalytic hypotheses (e.g., the chair-like TS is a polar chemical species stabilized by polar amino acid residues through the electrostatic effect) are based on structural studies of the enzyme–TSA complex (denoted as EI here), the molecular origin of a tight binding with the TSA has not yet been clearly elucidated.[60] Through a systematic modeling approach, we discuss a difference in the molecular recognition between the enzymatic TS and TSA in the BsCM reaction from a quantum chemical viewpoint. We prepared the molecular model of the EI complex by the same procedure used for the enzymatic reaction. In this calculation, we used the same definition of the large-QM model used in the enzymatic study.

Figure 11.4b summarizes the calculated geometrical parameters, and Figure 11.4d shows the active site geometry of the EI complex. In contrast to the enzymatic reaction shown in Figure 11.4a and Figure 11.4c, we observed a structural deviation of the hydrogen bond distance between the ether oxygen and the guanidium side chain

of Arg90, especially in the MD-sampled structures (see Ishida, 2008). As a result, a strong hydrogen bond between Glu78 and the hydroxyl group of the substrate was formed. Structural analysis revealed that the TSA is tightly bound to the enzyme active site through strong hydrogen bonds with Arg7 and Glu78, while it is weakly attracted by the guanidium side chain of Arg90. As for the electronic structure of the EI complex, Figure 11.5f shows the interaction residue analysis at the FMO2-MP2 level for the ES and EI complexes. Because the TSA is an inactive molecular species, here we compared the difference between the electronic character of enzymatic ES and that of EI complexes. Comparing the TSA profile with the enzyme profile, it is apparent that the interaction energies between the TSA and the protein are quite different from those of the enzymatic TS.

To understand the role of the protein environment from an atomistic perspective, we calculated the electrostatic contribution to the pair interaction energies based on Equation 11.7. By a comparative analysis of the electrostatic interaction between the isolated and complex states (where *isolated* means that the distance between the substrate and the active site is infinite, and *complex* means that the substrate is bound to the enzyme), we estimated the degree of polarization during the enzymatic process. Table 11.2 summarizes the energy components of the selected residues at the active site.

As clearly shown by this result, among the many energetic components of the ΔE_{IJ}^{RHF} term, the electrostatic contribution is the most dominant element. The difference between $\Delta E_{IJ}^{isolated}$ and $\Delta E_{IJ}^{complex}$ indicates how the enzyme is polarized by binding the reactive substrate. As for the residues that have important catalytic contributions on domain 2 unit, the degree of polarization in the enzymatic TS is higher than in the ES complex. In addition, this result clearly shows that the TSA is a chemical species with an electronic structure similar to the ES complex and not to the enzymatic TS. The amino acid residues listed in Table 11.2 are all aligned around the substrate at the active site. We could not observe an apparent energy difference between $\Delta E_{IJ}^{isolated}$ and $\Delta E_{IJ}^{complex}$ terms of other residues, indicating that the amino acid residues sensitive to the electrostatic perturbation are located near the substrate at the reactive center. These energy component analyses revealed that in the enzymatic TS, both the protein and the substrate are polarized under the influence of environmental effects and the strong electrostatic stabilization is stored in the molecular interactions between the protein and the substrate.

11.4.5 Probing Protein Environment 2: Site-Directed Mutation

To identify the role of particular amino acid residues experimentally, site-directed mutagenesis is one of the most powerful and established techniques in protein science.[1,3] Because introduction of a point mutation usually causes other side effects, it is generally difficult to identify the molecular origin of reduction in catalytic activity (or sometimes, an increase in binding affinity). Therefore, clarifying the effects of mutations in atomistic details by computer simulations is challenging. Although a dynamic effect induced by mutations may sometimes cause loss of the catalytic activity, by performing all-electron QM analyses, we focus only on the effect of protein environment that causes the TSS through the electrostatic interaction.

TABLE 11.2
Interaction Energy Components between the Substrate and the Selected Amino Acid Residues at the Active Site by FMO2-MP2/6-31G* Calculations (kcal/mol)

	Enzymatic Transition State (TS)					Enzyme-Substrate Complex (ES)					Enzyme–TSA complex (EI)				
Residue	$\Delta E_{IJ}^{isolated}$	$\Delta E_{IJ}^{complex}$	$\Delta\Delta E_{IJ}^{diff}$	ΔE_{IJ}^{RHF}	$\Delta\Delta E_{IJ}^{corr}$	$\Delta E_{IJ}^{isolated}$	$\Delta E_{IJ}^{complex}$	$\Delta\Delta E_{IJ}^{diff}$	ΔE_{IJ}^{RHF}	$\Delta\Delta E_{IJ}^{corr}$	$\Delta E_{IJ}^{isolated}$	$\Delta E_{IJ}^{complex}$	$\Delta\Delta E_{IJ}^{diff}$	ΔE_{IJ}^{RHF}	$\Delta\Delta E_{IJ}^{corr}$
Phe57	–2.38	–4.45	–2.07	–3.07	–4.27	–2.39	–4.63	–2.24	–4.26	–3.51	–3.24	–5.82	–2.58	–3.71	–4.57
Ala59	–5.85	–7.87	–2.02	–4.98	–4.27	–6.30	–8.79	–2.49	–5.42	–4.21	–6.00	–8.96	–2.96	–6.55	–3.75
Lys60	–50.25	–52.58	–2.33	–53.09	–1.49	–49.74	–52.90	–3.16	–53.52	–1.66	–50.87	–54.41	–3.54	–55.30	–1.67
Arg63	–133.93	–149.70	–15.77	–138.72	–7.83	–137.03	–153.80	–16.77	–143.24	–7.81	–131.82	–148.36	–16.54	–138.39	–7.78
Val73	–6.87	–9.11	–2.24	–9.24	–0.74	–6.93	–9.37	–2.44	–9.35	–0.99	–6.63	–9.24	–2.61	–9.37	–0.71
Thr74	17.05	16.00	–1.05	16.24	–3.00	16.13	14.74	–1.39	15.44	–3.45	17.11	16.07	–1.04	16.62	–2.37
Cys75	–20.23	–25.21	–4.98	–12.66	–6.38	–19.99	–26.10	–6.11	–14.24	–6.38	–18.86	–24.63	–5.77	–14.34	–6.35
Arg7	–144.57	–167.16	–22.59	–158.98	–8.07	–138.69	–157.85	–19.16	–152.60	–7.76	–141.88	–162.65	–20.77	–155.62	–7.99
Ala9	13.08	10.55	–2.53	10.03	–0.59	12.56	10.65	–1.91	10.21	–0.64	13.17	10.82	–2.55	10.26	–0.76
Glu78	73.41	63.86	–9.55	65.44	–3.05	77.28	70.62	–6.66	70.45	–2.93	78.65	70.95	–7.70	71.37	–3.04
Arg90	–131.82	–146.59	–14.77	–136.62	–11.64	–125.42	–136.92	–11.50	–128.33	–10.71	–128.19	–140.16	–11.97	–130.65	–10.83
Tyr108	–13.32	–34.14	–20.82	–18.48	–5.02	–12.75	–31.99	–19.24	–17.46	–5.28	–12.49	–31.18	–18.69	–17.11	–5.22
Leu115	3.32	–4.22	–7.54	–2.28	–5.81	2.58	–2.92	–5.50	0.47	–5.36	1.04	–2.00	–3.04	–0.22	–2.96
Arg116	–112.49	–130.82	–18.33	–121.84	–8.21	–112.43	–131.74	–19.31	–123.46	–8.09	–113.42	–131.17	–17.75	–121.97	–8.92

Notes: The electrostatic energy components in the FMO2 formulation are evaluated by Equation 11.7 both for the isolated (wherein the substrate and the enzyme are separated by an infinite distance) and complex states. The difference of electrostatic interaction energy between the isolated and complex state is defined by $\Delta\Delta E_{IJ}^{diff} = \Delta E_{IJ}^{complex} - \Delta E_{IJ}^{isolated}$. The correlation energy is given by $\Delta\Delta E_{IJ}^{corr} = \Delta E_{IJ}^{MP2} - \Delta E_{IJ}^{RHF}$. The upper seven amino acid residues (Phe57 – Cys75) are located on domain 1 of the enzyme, while the lower seven amino acid residues (Arg7 – Arg116) are located on domain 2 of the enzyme.

As demonstrated clearly in the previous sections, the most crucial residue in BsCM reaction is Arg90 located on the domain 2 region. Earlier experimental studies have shown that lack of Arg90 results in a large loss of the catalytic activity. However, it is unclear whether the critical catalytic component of Arg is the positive charge or the molecular shape of the side chain. Mutation experiments have shown that even in a charge-conserved substitution of Arg with Lys, substantial catalytic reduction is observed.[57–59] To eliminate the possibility of steric effects of the side chain, Hilvert et al. prepared a BsCM variant containing citrulline (isosteric/neutral arginine analogue) at Arg90 position and measured the catalytic activity. By conducting systematic experiments, they concluded that the electrostatic stabilization of the polarized TS is the paramount factor.[61] Here, to identify the effect of both molecular shape and charge on Arg90 from the QM viewpoint, we compared the electronic structure of the protein in the wild-type as well as Lys90 and Cit90 mutant reactions. Considering the computational costs to perform detailed QM/MM reaction path modeling, we employed a simple strategy for the analysis of mutant systems. The ES and TS geometries of Lys90/Cit90 mutants were prepared by employing wild-type structures, assuming that the reaction proceeds in a similar manner in both mutants. The original Arg90 position was replaced by Lys90 or Cit90, while retaining a maximum overlap of the heavy atom coordinates. In each mutant, entire enzyme–substrate structures were optimized by *ab initio* QM/MM calculations. Because the main purpose of this study is to clarify the effect of point mutations qualitatively, we adopted the small-QM region model and repeated the same systematic all-electron QM analyses. For all FMO2 calculations, the 6-31(+)G* basis set was used, and only diffuse functions were added to the carboxylic groups in the protein and substrate. Note that the procedure to add (or delete) diffuse functions is a technical matter of converging the SCF cycle in FMO calculations. This procedure has a small effect on the calculation results, and it does not influence the qualitative discussion below. Also, instead of performing full FMO2-MP2 calculations, we employed the multilayer FMO2 computations wherein the selected residues around the active site were treated using MP2-level calculations.[62] The details are described elsewhere.[16] Because of a slight difference in the calculation details and to clarify the differences among the previous discussions, Figure 11.6 shows the interaction residue analysis for three types of enzyme systems.

The TS geometries around the active site of both Lys90 and Cit90 variants are shown in Figure 11.4e and Figure 11.4f, respectively. In both mutants, no large conformational change is observed in the overall protein structure. However, changes were observed in the geometries around the mutation site. Due to a loss of the steric hindrance by the guanidium side chain of Arg90, most of the hydrogen bonds at the active site were slightly weakened in the ES complex. As a result, the substrate forms a strong hydrogen bond with Glu78, and compared with the wild-type reaction, the ES is further stabilized in the mutants. Although there were no mutations inside the protein structure, this situation is exactly the same as observed in the TSA binding. A slight modification of geometric parameters significantly influences the relative stability of the substrate. In particular, the steric factor of Arg90 that bridges Glu78 to the substrate through hydrogen bonds seems to be a key element in achieving effective catalysis.

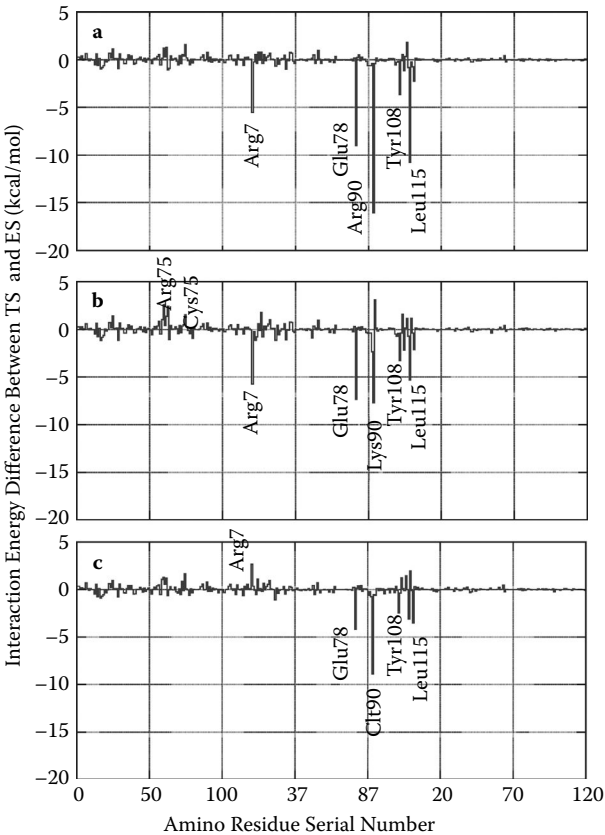


FIGURE 11.6 Interaction energy difference between the TS and ES of the BsCM reactions using the small-QM region model: (a) wild-type reaction, (b) Lys90-mutant reaction, and (c) Cit90-mutant reaction. All calculations were performed by FMO2-RHF:MP2/6-31(+) G* level. The x-axis indicates the amino acid residue serial number (the same definition in Figure 11.5), and the y-axis designates the interaction energy difference between the two states (in kcal/mol).

Figure 11.6 shows the comparative FMO analysis of interaction energy diagrams between the substrate and enzymes. In the Lys90 mutant reaction (see Figure 11.6b), although the positive charge that polarizes the substrate is similar to that of Arg90, we observed a catalytic reduction at the Lys90 position. The main reason is that ES is further stabilized due to loss of the steric factor that prevents Glu78 from forming a strong hydrogen bond with the substrate. In the case of the Cit90 mutant (see Figure 11.6c), structural changes destroy the detailed balance in the hydrogen bond of the Glu78-Arg90-substrate alignment observed in the wild-type enzyme. Compared with the other two cases, the charge distribution of the substrate is also different, and we observe an induced dipole even in the ES complex. As a result, the buried carboxylate part of the substrate forms strong hydrogen bonds with Arg7 even

in the ES complex. The catalytic factor of Cit90 is mainly due to the strong attraction to the buried carboxylate of the substrate, not the induced charge on the ether oxygen at the TS structure. As clearly shown in this study, Cit90 stabilizes the TS in a different manner than Arg90 in the wild-type reaction.

11.4.6 Are All-Electron QM Calculations Necessary?

In this chapter, we discussed the electronic structure of protein in the enzymatic process by performing all-electron QM calculations combined with *ab initio* QM/MM modeling. Although its validity and limitations are currently unclear, QM/MM is becoming an increasingly popular tool for studying enzymatic reactions. Also, FMO is only one of several developing methods for treating large molecules at the present stage. Therefore, it is meaningful to discuss the differences clarified in all-electron QM analyses by comparison with QM/MM modeling results.

To demonstrate a systematic modeling strategy of enzymatic reactions, we first examined and compared the effects of geometrical parameters (on the selection of the QM region) on the energetics using the small- and large-QM models. In the present QM/MM reaction path modeling, the selection of QM region had a small effect on the calculated properties. As for the geometric parameters, the hydrogen bonds between the substrate and Arg7/Arg90 changed only slightly, and as a result, the interaction energy components of these terms also changed. Concerning reaction energetics summarized in Table 11.1 and Figure 11.3, the global profiles of potential energy show relatively similar behavior, and a deviation of the activation energy using both QM models is within a few kcal/mol with each theoretical method (RHF/MP2/B3LYP). When we compare the energy profiles by the QM/MM with those by the FMO2 using the large-QM model, the overall profiles are similar, and these energy curves predict similar activation energies. Because both QM/MM and FMO are approximate methods that employ a division of covalent bonds and a selection of small regions extracted from large protein structures, it is difficult to judge which profiles are more reliable. As clarified in the energy component analysis (see Table 11.2), only the amino acid residues adjacent to the reactive center are influenced and polarized by the change in electronic structure of the substrate. The present analysis clearly confirms the validity of QM/MM treatment. If we select an appropriate QM region that contains all catalytically important residues around the substrate, QM/MM geometries could be relied on for the all-electron QM analysis. In other words, a reliable prediction by employing all-electron QM calculations requires a reasonable molecular structure, such as *ab initio* QM/MM optimized geometries.

On the basis of the systematic computational modeling and accompanying energy components analysis, several important findings about the role of protein environment are summarized as follows:

1. Comparing the residues located on domain 1 with those on domain 2, the degrees of induced polarization show different trends in the enzymatic process.
2. The residues on domain 1 are more polarized in the ES complex, while those on domain 2 are more polarized in the TS in the wild-type reaction.

3. In particular, Arg7, Glu78, and Arg90 are highly polarized in the TS in response to the delocalized electronic character of the reactive substrate.
4. In most cases, the contributions of polarization energy are not large compared with the absolute values of electrostatic interaction energy.
5. In the case of the EI complex with TSA, the degree of induced polarization is similar to the ES complex in most amino acid residues.
6. In the case of Arg90 mutant, the major reason for the catalytic reduction is loss of the electrostatic stabilization, although loss of the detailed balance of hydrogen bonding alignment in Glu78-Arg90-substrate and the resultant ES stabilization is a nonnegligible component. Also, the change in electronic character introduced by a point mutation is limited to the small active site region around the substrate.
7. The contribution of the electron correlation energy is relatively similar throughout the enzymatic process. This result is consistent with the structural analysis showing that the isomerization proceeds in a local active site region without a large conformational change both in the substrate and the protein. In the case of the EI complex, most correlation energy terms are similar to those in the enzymatic process, implying that the enzyme recognizes the molecular shape of the substrate and the TSA in a similar fashion.

11.5 CONCLUSION AND OUTLOOK

Finally, we summarize the major findings clarified in this *ab initio*–based modeling study. By introducing a systematic computational strategy combining all-electron QM calculations with *ab initio* QM/MM modeling, we analyzed the details of the electronic structure of the protein environment in the BsCM catalyzed interconversion of chorismate into prephenate. QM-based energy decomposition analysis reveals that amino acid residues located on different protein domains show different behavior of the electronic polarization in response to the change in the electronic character of the substrate. The residues on domain 2 are more polarized in the TS, and those on domain 1 show an opposite character in the reaction process. The delocalized charge in the TS of the substrate is highly polarized by the protein environment, and the residues at the active site are also polarized in response to the electron distribution of the TS. As a result, a large amount of electrostatic stabilization energy is stored in the molecular interaction between the enzyme and the substrate and supplied for the TSS. Among several polar amino acid residues located at the active site, Arg90 is crucial for efficient catalysis, and the major catalytic factor of this system is clarified to be the electrostatic interaction in the TS geometry. A comparative analysis of the interaction energy differences between the wild-type reaction and the Lys90/Cit90 mutants confirmed the electrostatic nature of Arg90 contribution. In the case of the EI complex, although a relative orientation of the TSA is similar to that of the enzymatic TS in the active site, the degree of polarization caused by the TSA binding is more similar to that of the ES complex.

Because a qualitative discussion of enzymatic reactions (including a comparative study of the mutant reactivity) usually requires (semi) quantitative estimation of the

reaction energetics within a few kcal/mol, it is difficult to reproduce relative energetics using all-electron FMO calculations. Moreover, major challenges of enzymatic reactions are to estimate the free energy profile in an aqueous environment, and to clarify some of the catalytic factors (including protein dynamics). For the time being, *ab initio* QM/MM will remain the major theoretical tool to investigate the reaction process, and all-electron QM calculations by methods such as the FMO technique may become useful methods to analyze the interaction energy components inside the protein environment. Applications of the present systematic modeling technique to another important class of enzymes are important steps to understand the role of the protein environment in the enzyme reactions.

ACKNOWLEDGMENTS

This work was supported by a Grant-in-Aid for Scientific Research from MEXT and the Next Generation Super Computing Project, Nanoscience Program, MEXT, Japan. Some numerical calculations were performed at the Computer Center of the Institute for Molecular Science (IMS). The author thanks Kazuo Kitaura for his continuous encouragement and is grateful to Dmitri G. Fedorov.

REFERENCES

1. Fersht, A. 1999. *Structure and Mechanism in Protein Science. A Guide to Enzyme Catalysis and Protein Folding,* 2nd ed. W. H. Freeman: New York.
2. Jencks, W.P. 1987. *Catalysis in Chemistry and Enzymology.* Dover Publications: New York.
3. Frey, P.A., Hegeman, A.D. 2007. *Enzymatic Reaction Mechanisms.* Oxford University Press: New York.
4. Whitford, D. 2005. *Proteins: Structure and Function.* John Wiley & Sons: England.
5. Wolfenden, R., Snider, M.J. 2001. The depth of chemical time and the power of enzymes as catalysts. *Acc. Chem. Res.* 34: 938–945.
6. Warshel, A. 1991. *Computer Modeling of Chemical Reactions in Enzymes and Solutions,* John Wiley & Sons: New York.
7. Field, M.J. 2002. Simulating enzyme reactions: challenges and perspectives. *J. Comput. Chem.* 23: 48–58.
8. Gao, J., Truhlar, D.G. 2002. Quantum mechanical methods for enzyme kinetics. *Annu. Rev. Phys. Chem.* 53: 467–505.
9. Warshel, A. 2003. Computer simulations of enzyme catalysis: methods, progress, and insights. *Annu. Rev. Biophys. Biomol. Struct.* 32: 425–443.
10. Ryde, U. 2003. Combined quantum and molecular mechanics calculations on metalloproteins. *Curr. Opin. Chem. Biol.* 7: 136–142.
11. Garcia-Viloca, M., Gao, J., Karplus, M., et al. 2004. How enzymes work: analysis by modern rate theory and computer simulations. *Science* 303: 186–195.
12. Friesner, R.A., Guallar, V. 2005. *Ab initio* quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis. *Annu. Rev. Phys. Chem.* 56: 389–427.
13. Shaik, S., Kumar, D., deVisser, S.P. et al. 2005. Theoretical perspective on the structure and mechanism of cytochrome p450 enzymes. *Chem. Rev.* 105: 2279–2328.

14. Riccardi, D., Schaefer, P., Yang, Y. et al. 2006. Development of effective quantum mechanical/molecular mechanical (QM/MM) methods for complex biological processes. *J. Phys. Chem. (B)* 110: 6458–6469.
15. Warshel, A., Sharma, P.K., Kato, M. et al. 2006. Electrostatic basis for enzyme catalysis. *Chem. Rev.* 106: 3210–3235.
16. Ishida, T., Fedorov, D.G., Kitaura, K. 2006. All electron quantum chemical calculation of the entire enzyme system confirms a collective catalytic device in the chorismate mutase reaction. *J. Phys. Chem. (B)* 110: 1457–1463.
17. Ishida, T. 2008. Probing protein environment in an enzymatic process: all-electron quantum chemical analysis combined with *ab initio* quantum mechanical/molecular mechanical modeling of chorismate mutase. *J. Chem. Phys.* 129: 125105–125119.
18. Ganem, B. 1996. The mechanism of the claisen rearrangement: déjà vu all over again. *Angew. Chem. Int. Ed.* 35: 936–945.
19. Andrews, P.R., Smith, G.D., Young, I.G. 1973. Transition-state stabilization and enzymatic catalysis: kinetic and molecular orbital studies of the rearrangement of chorismate to prephenate. *Biochemistry* 12: 3492–3498.
20. Copley, S.D., Knowles, J.R. 1987. The conformational equilibrium of chorismate in solution: implications for the mechanism of the non-enzymatic and the enzyme-catalyzed rearrangement of chorismate to prephenate. *J. Am. Chem. Soc.* 109: 5008–5013.
21. Guilford, W.J., Copley, S.D., Knowles, J.R. 1987. On the mechanism of the chorismate mutase reaction. *J. Am. Chem. Soc.* 109: 5013–5019.
22. Kast, P., Asif-Ullah, M., Hilvert, D. 1996. Is chorismate mutase a prototypic entropy trap? Activation parameters for the *Bacillus subtilis* enzyme. *Tetrahedron Lett.* 37: 2691–2694.
23. Lyne, P.D., Mulholland, A.J., Richards, W.G. 1995. Insights into chorismate mutase catalysis from a combined QM/MM simulation of the enzyme reaction. *J. Am. Chem. Soc.* 117: 11345–11350.
24. Martí, S., Andrés, J., Moliner, V. et al. 2001. A hybrid potential reaction path and free energy study of the chorismate mutase reaction. *J. Am. Chem. Soc.* 123: 1709–1712.
25. Martí, S., Andrés, J., Moliner, V. et al. 2004. A comparative study of claisen and cope rearrangements catalyzed by chorismate mutase. An insight into enzymatic efficiency: transition state stabilization or substrate preorganization? *J. Am. Chem. Soc.* 126: 311–319.
26. Worthington, S.E., Roitberg, A.E., Krauss, M. 2001. An MD/QM study of the chorismate mutase-catalyzed claisen rearrangement reaction. *J. Phys. Chem. (B)* 105: 7087–7095.
27. Lee, Y.S., Worthington, S.E., Krauss, M. et al. 2002. Reaction mechanism of chorismate mutase studied by the combined potentials of quantum mechanics and molecular mechanics. *J. Phys. Chem. (B)* 106: 12059–12065.
28. Ranaghan, K.E., Ridder, L., Szefczyk, B. et al. 2004. Transition state stabilization and substrate strain in enzyme catalysis: *ab initio* QM/MM modelling of the chorismate mutase reaction. *Org. Biomol. Chem.* 2: 968–980.
29. Szefczyk, B., Mulholland, A.J., Ranaghan, K.E. et al. 2004. Differential transition-state stabilization in enzyme catalysis: quantum chemical analysis of interactions in the chorismate mutase reaction and prediction of the optimal catalytic field. *J. Am. Chem. Soc.* 126: 16148–16159.
30. Crespo, A., Scherlis, D.A., Martí, M.A. et al. 2003. A DFT-based QM-MM approach designed for the treatment of large molecular systems: application to chorismate mutase. *J. Phys. Chem. (B)* 107: 13728–13736.
31. Crespo, A., Martí, M.A., Estrin, D.A. et al. 2005. Multiple-steering QM-MM calculation of the free energy profile in chorismate mutase. *J. Am. Chem. Soc.* 127: 6940–6941.

32. Woodcock, H.L., Hodoscek, M., Sherwood, P. et al. 2003. Exploring the quantum mechanical/molecular mechanical replica path method: a pathway optimization of the chorismate to prephenate claisen rearrangement catalyzed by chorismate mutase. *Theor. Chem. Acc.* 109: 140–148.
33. Hur, S., Bruice, T.C. 2003. Comparison of formation of reactive conformers (NACs) for the Claisen rearrangement of chorismate to prephenate in water and in the E-coli mutase: the efficiency of the enzyme catalysis. *J. Am. Chem. Soc.* 125: 5964–5972.
34. Hur, S., Bruice, T.C. 2003. Just a near attack conformer for catalysis (chorismate to prephenate rearrangements in water, antibody, enzymes, and their mutants). *J. Am. Chem. Soc.* 125: 10540–10542.
35. Hur, S., Bruice, T.C. 2003. The near attack conformation approach to the study of the chorismate to prephenate reaction. *Proc. Natl. Acad. Sci. USA* 100: 12015–12020.
36. Guimarães, C.R.W., Repasky, M.P., Chandrasekhar, J. et al. 2003. Contributions of conformational compression and preferential transition state stabilization to the rate enhancement by chorismate mutase. *J. Am. Chem. Soc.* 125: 6892–6899.
37. Guimarães, C.R.W., Udier-Blagovic, M., Tubert-Brohman, I. et al. 2005. Effects of Arg90 neutralization on the enzyme-catalyzed rearrangement of chorismate to prephenate. *J. Chem. Theory Comput.* 1: 617–625.
38. Guo, H., Cui, Q., Lipscomb, W.N. et al. 2001. Substrate conformational transitions in the active site of chorismate mutase: their role in the catalytic mechanism. *Proc. Natl. Acad. Sci. USA* 98: 9032–9037.
39. Guo, H., Cui, Q., Lipscomb, W.N. et al. 2003. Understanding the role of active-site residues in chorismate mutase catalysis from molecular dynamics simulations. *Angew. Chem. Int. Ed.* 42: 1508–1511.
40. Strajbl, M., Shurki, A., Kato, M. et al. 2003. Apparent NAC effect in chorismate mutase reflects electrostatic transition state stabilization. *J. Am. Chem. Soc.* 125: 10228–10237.
41. Ranaghan, K.E., Mulholland, A.J. 2004. Conformational effects in enzyme catalysis: QM/MM free energy calculation of the "NAC" contribution in chorismate mutase. *Chem. Commun.* 1238–1239.
42. Claeyssens, F., Harvey, J.N., Manby, F.R. et al. 2006. High-accuracy computation of reaction barriers in enzymes. *Angew. Chem. Int. Ed.* 45: 6856–6859.
43. Chook, Y.M., Ke, H., Lipscomb, W.N. 1993. Crystal structures of the monofunctional chorismate mutase from *Bacillus subtilis* and its complex with a transition-state analog. *Proc. Natl. Acad. Sci. USA* 90: 8600–8603.
44. Ishida, T., Kato, S. 2003. Theoretical perspectives on the reaction mechanism of serine proteases: the reaction free energy profiles of the acylation process. *J. Am. Chem. Soc.* 125: 12035–12048.
45. Ishida, T., Kato, S. 2004. Role of Asp102 in the catalytic relay system of serine proteases: a theoretical study. *J. Am. Chem. Soc.* 126: 7111–7118.
46. Ishida, T. 2006. Low-barrier hydrogen bond hypothesis in the catalytic triad residue of serine proteases: correlation between structural rearrangement and chemical shifts in the acylation process. *Biochemistry* 45: 5413–5420.
47. Bayly, C.I., Cieplak, P., Cornell, W.D. et al. 1993. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges — the RESP model. *J. Phys. Chem.* 97: 10269–10280.
48. Cieplak, P., Cornell, W.D., Bayly, C. et al. 1995. Application of the multimolecule and multiconformational RESP methodology to biopolymers — charge derivation for DNA, RNA, and proteins. *J. Comput. Chem.* 16: 1357–1377.
49. Cornell, W.D., Cieplak, P., Bayly, C.I. et al. 1995. A second generation force field for the simulation of proteins, nucleic acid, and organic molecules. *J. Am. Chem. Soc.* 117: 5179–5197.

50. Kollman, P., Dixon, R., Cornell, W. et al. 1997. In *Computer Simulation of Biomolecular Systems*, Vol. 3, eds. van Gunsteren, W.F., Weiner, P.K., Wilkinson, A.J., Kluwer Academic, Dordrecht.
51. Kitaura, K., Ikeo, E., Asada, T. et al. 1999. Fragment molecular orbital method: an approximate computational method for large molecules. *Chem. Phys. Lett.* 313: 701–706.
52. Fedorov, D.G., Kitaura, K. 2004. The importance of three-body terms in the fragment molecular orbital method. *J. Chem. Phys.* 120: 6832–6840.
53. Fedorov, D.G., Kitaura, K. 2004. Second order Møller–Plesset perturbation theory based upon the fragment molecular orbital method. *J. Chem. Phys.* 121: 2483–2490.
54. Schmidt, M.W., Baldridge, K.K., Boats, J.A. et al. 1993. General atomic and molecular electronic-structure system. *J. Comput. Chem.* 14: 1347–1363.
55. Kast, P., Asif-Ullah, M., Jiang, N. et al. 1996. Exploring the active site of chorismate mutase by combinatorial mutagenesis and selection: the importance of electrostatic catalysis. *Proc. Natl. Acad. Sci. USA* 93: 5043–5048.
56. Kast, P., Hartgerink, J.D., Asif-Ullah, M. et al. 1996. Electrostatic catalysis of the claisen rearrangement: probing the role of Glu78 in *Bacillus subtilis* chorismate mutase by genetic selection. *J. Am. Chem. Soc.* 118: 3069–3070.
57. Cload, S.T., Liu, D.R., Pastor, R.M. et al. 1996. Mutagenesis study of active site residues in chorismate mutase from *Bacillus subtilis*. *J. Am. Chem. Soc.* 118: 1787–1788.
58. Gustin, D.J., Mattei, P., Kast, P. et al. 1999. Heavy atom isotope effects reveal a highly polarized transition state for chorismate mutase. *J. Am. Chem. Soc.* 121: 1756–1757.
59. Kast, P., Grisostomi, C., Chen, I.A. et al. 2000. A strategically positioned cation is crucial for efficient catalysis by chorismate mutase. *J. Biol. Chem.* 275: 36832–36838.
60. Barbany, M., Gutiérrez-de-Terán, H., Sanz, F. et al. 2003. On the generation of catalytic antibodies by transition state analogues. *Chem. Bio. Chem.* 4: 277–285.
61. Kienhöfer, A., Kast, P., Hilvert, D. 2003. Selective stabilization of the chorismate mutase transition state by a positively charged hydrogen bond donor. *J. Am. Chem. Soc.* 125: 3206–3207.
62. Fedorov, D.G., Ishida, T., Kitaura, K. 2005. Multilayer formulation of the fragment molecular orbital method (FMO). *J. Phys. Chem. (A)* 109: 2638–2646.

Index

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

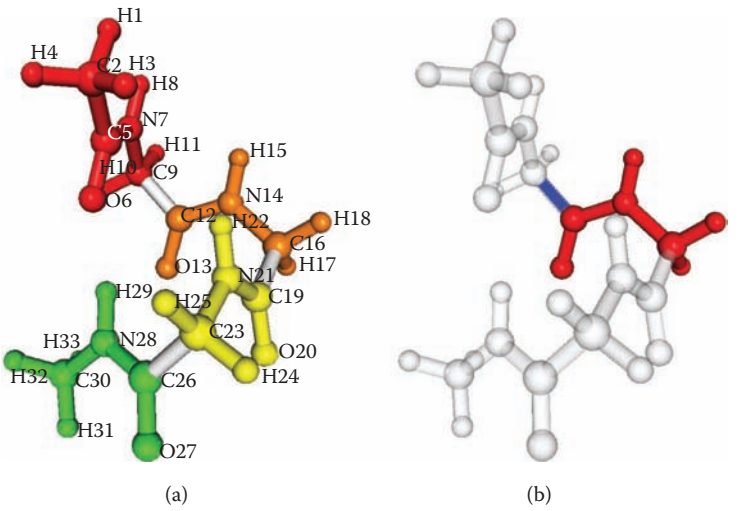
T

U

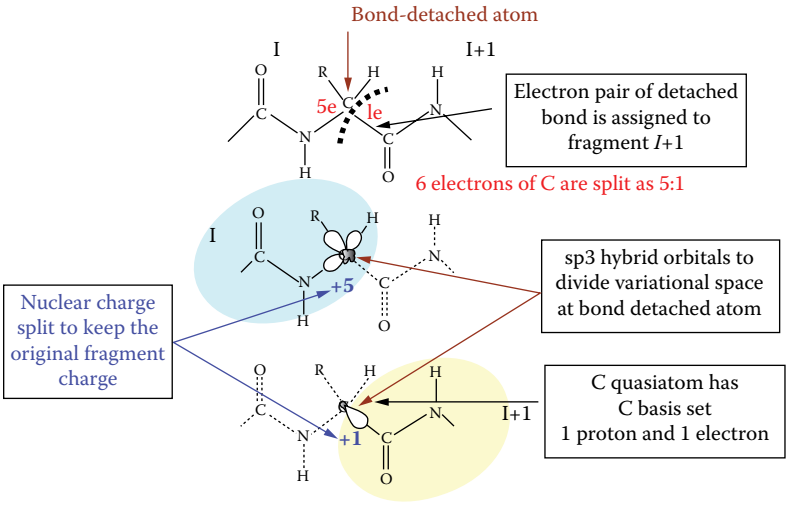
V

W

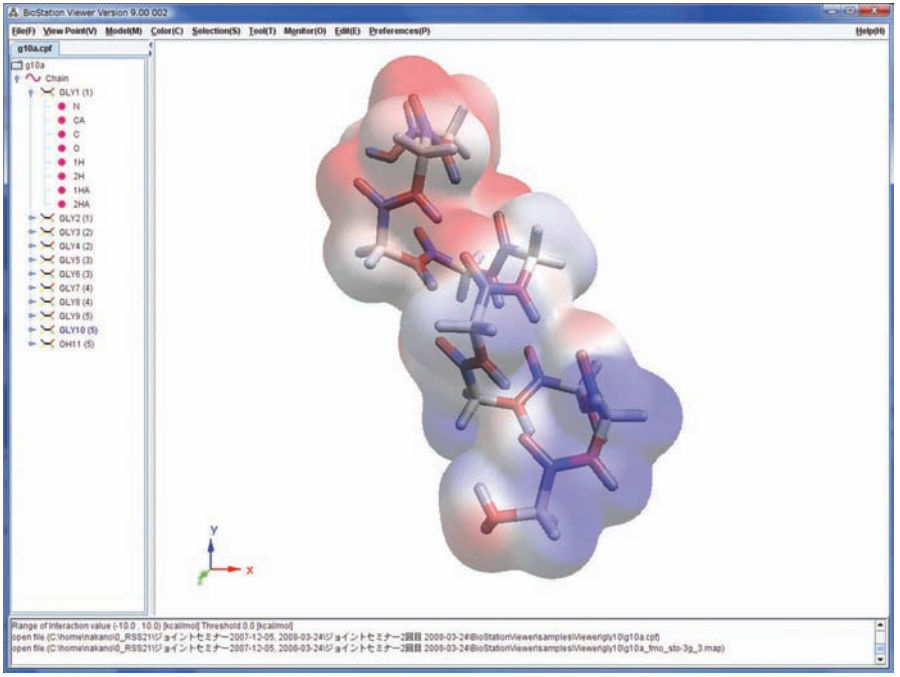
X



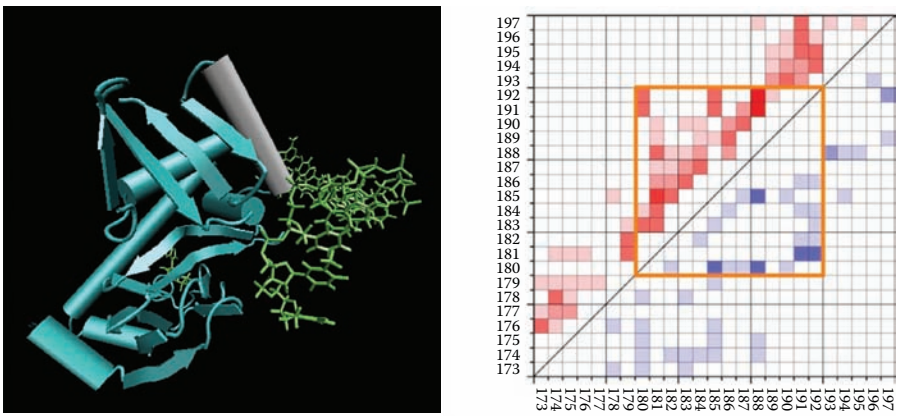
COLOR FIGURE 2.1 Fragmentation of capped $(GLY)_3$: (a) the whole system and the fragmentation points and (b) fragment 2.



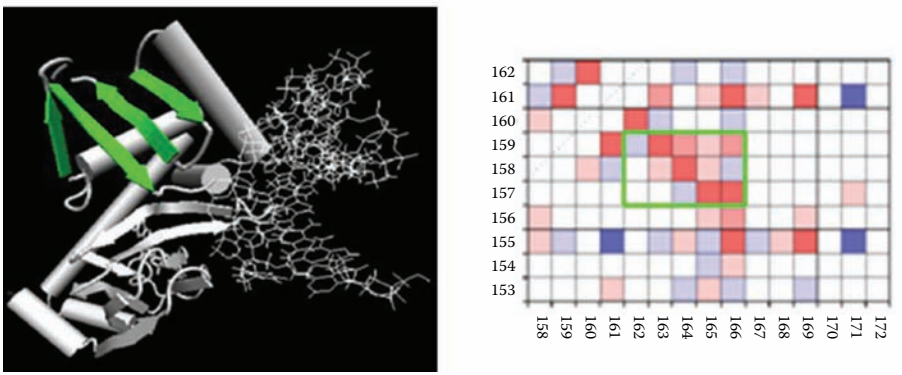
COLOR FIGURE 2.2 Details of bond detachment in fragment molecular orbital (FMO), showing the use of the hybrid orbitals and the formal proton reassignment at the bond detached atom, located on the border of fragments I and $I + 1$.



COLOR FIGURE 3.2 Main window of the BioStation Viewer with $(Gly)_{10}$ colored by the electrostatic potential value on the isosurface of electron density.

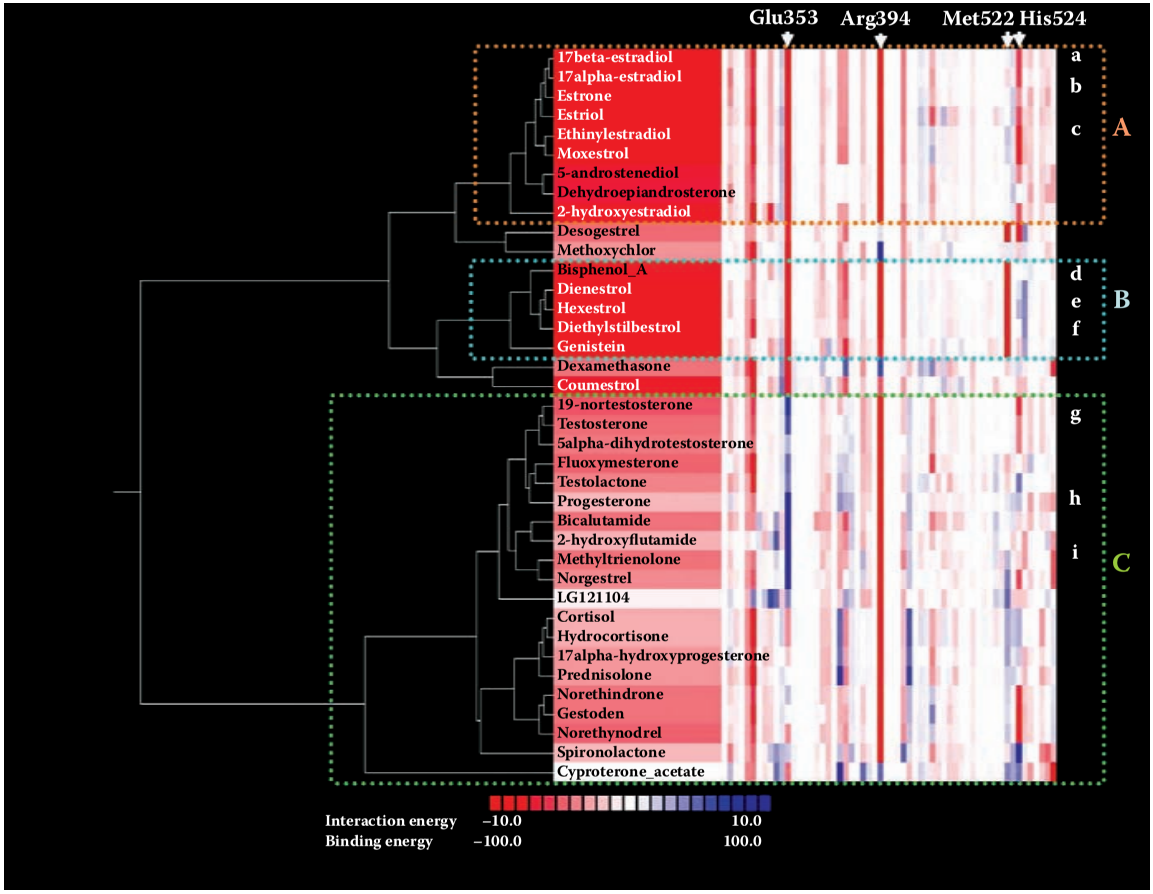


(a)

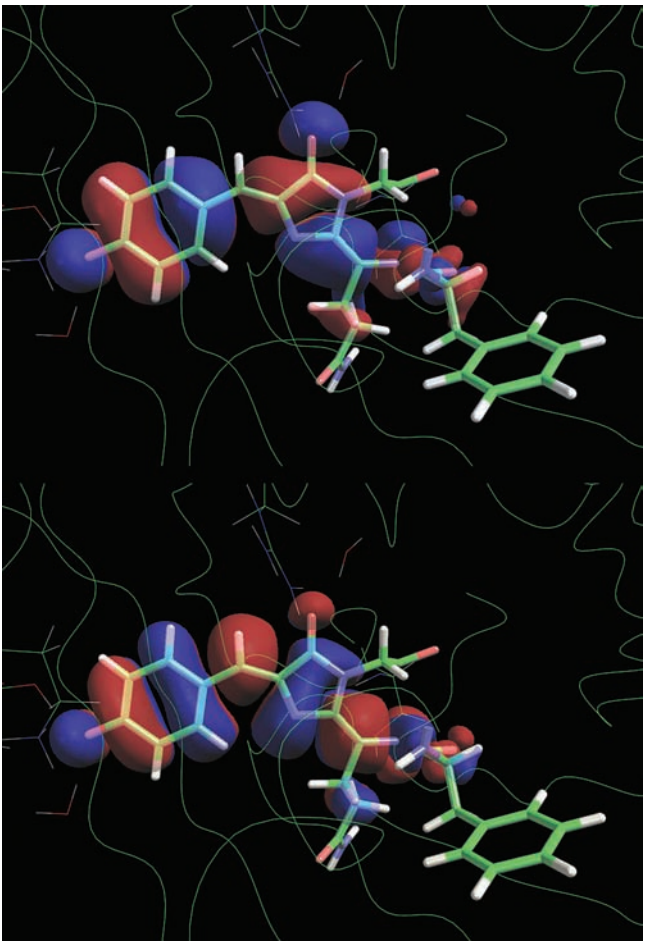


(b)

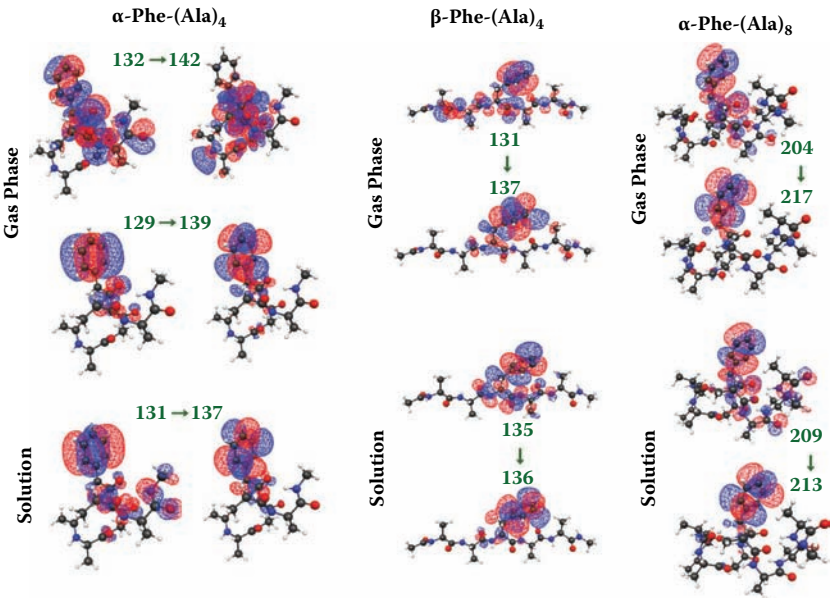
COLOR FIGURE 3.8 Interfragment interaction energy (IFIE) maps focusing on regions forming secondary structures. (a) IFIE map corresponding to regions forming α-helix. (b) IFIE map corresponding to regions forming antiparallel β-sheet.



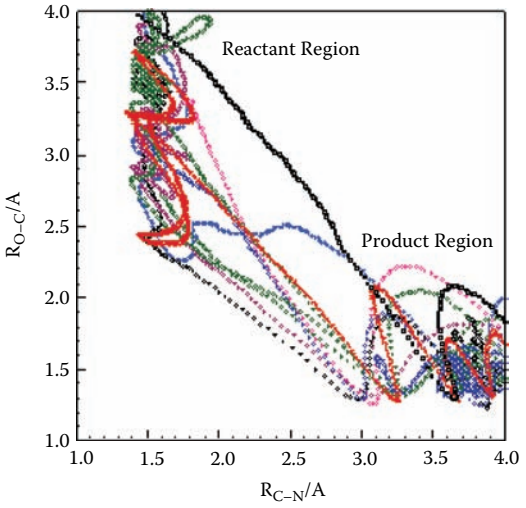
COLOR FIGURE 3.9 Hierarchical clustering by the interaction patterns of the human estrogen receptor ligand-binding domain and 38 compounds from the *Ki*Bank. Parts a through i denote the IFIEs of 17β-estradiol, estrone, ethynilestradiol, bisphenol A, hexestrol, diethylstilbestrol, testosterone, progesterone, and methyltrienolone, respectively.



COLOR FIGURE 4.3 Hole (upper) and particle (lower) natural orbitals of MLFMO-CIS/6-31G calculation for DsRed.

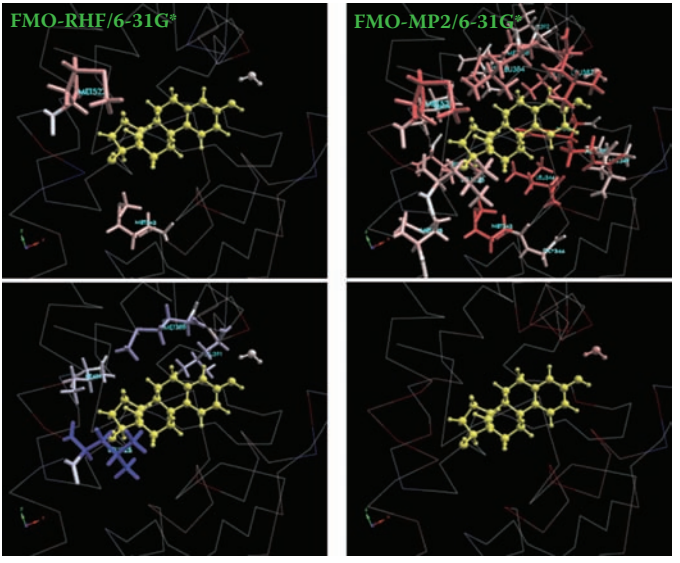


COLOR FIGURE 5.6 Molecular orbitals defining the major contributions to the lowest singlet excited state in Phe-$(Ala)_n$ systems (TDDFT in gas phase and in solution [PCM], LC-BLYP/6-31G*). The numbers indicate the serial orbital numbers, and the orbitals are listed in the occupied-virtual pairs. The two orbital phases are shown in red and blue. For Phe-$(Ala)_n$, there are two major contributions to the excited state of interest of nearly equal weight. (Reproduced from Chiba, M., Fedorov, D.G., Kitaura, K. 2008. *J. Comput. Chem.* 26: 2667–2676. With permission.)

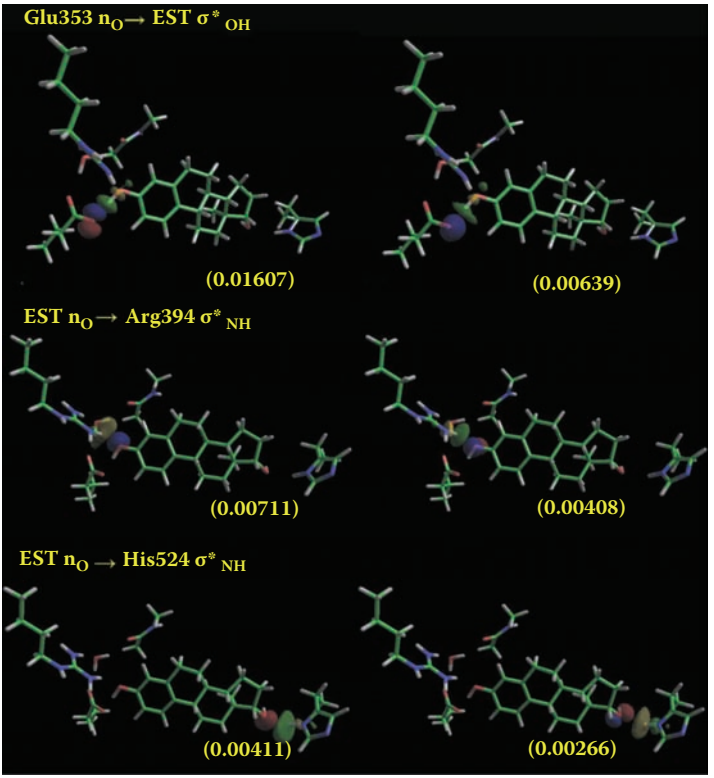


COLOR FIGURE 6.5 Ten trajectories on the two-dimensional surface. (Reproduced from Sato, M., Yamataka, H., Komeiji, Y., et al. 2008. *J. Am. Chem. Soc.* 130: 2396–2397. With permission.)

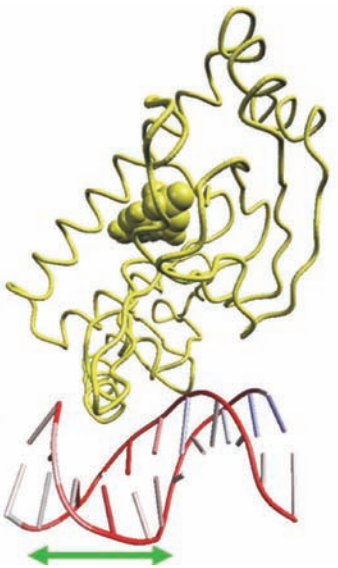
COLOR FIGURE 7.5 Visualization of interfragment interaction energies between each residue of estrogen receptor and the ligand calculated at the MP2/6-31G* level.



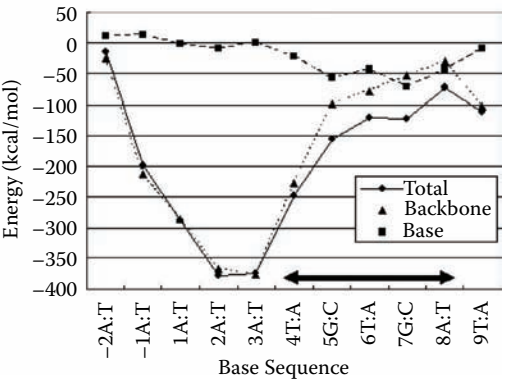
COLOR FIGURE 7.6 Interfragment interaction energies between the ligand and surrounding hydrophobic residues. Upper figure: attractive interaction less than –1 kcal/mol. Lower figure: repulsive interactions more than 1kcal/mol. Calculations were performed at the RHF/6-31G* (left) and MP2/6-31G* (right) levels.



COLOR FIGURE 7.7 The pairs of "hole orbital" and "particle orbital" concerning charge transfer interactions between the estrogen receptor and the ligand obtained by CAFI. The phasing of the hole orbital is colored in red and blue. The corresponding colors of the particle orbital are yellow and green, respectively. Occupation numbers as transferred electrons for each pair are given in parentheses. Calculations were performed at the RHF/6-31G* level.

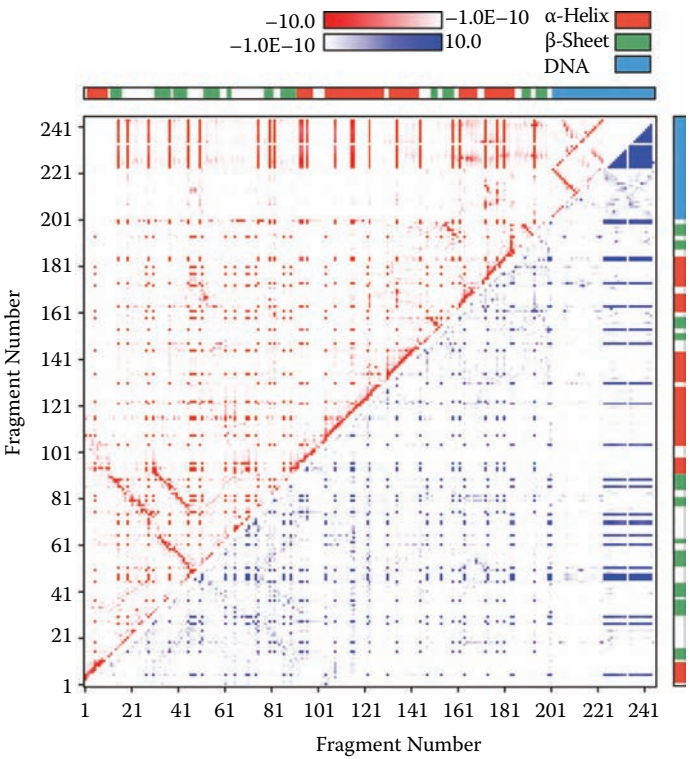


(a)

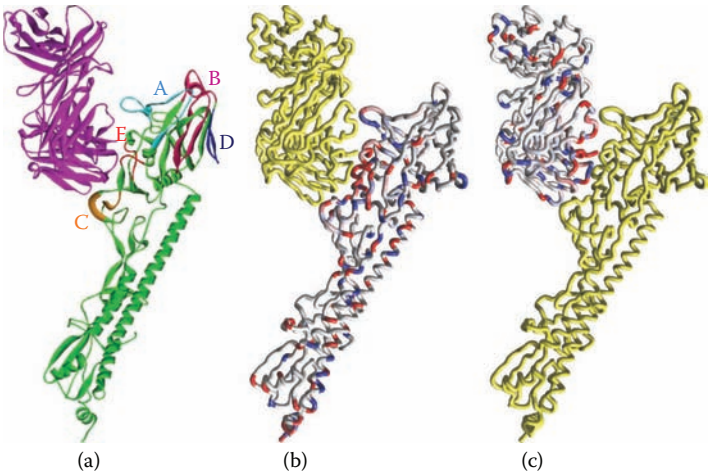


(b)

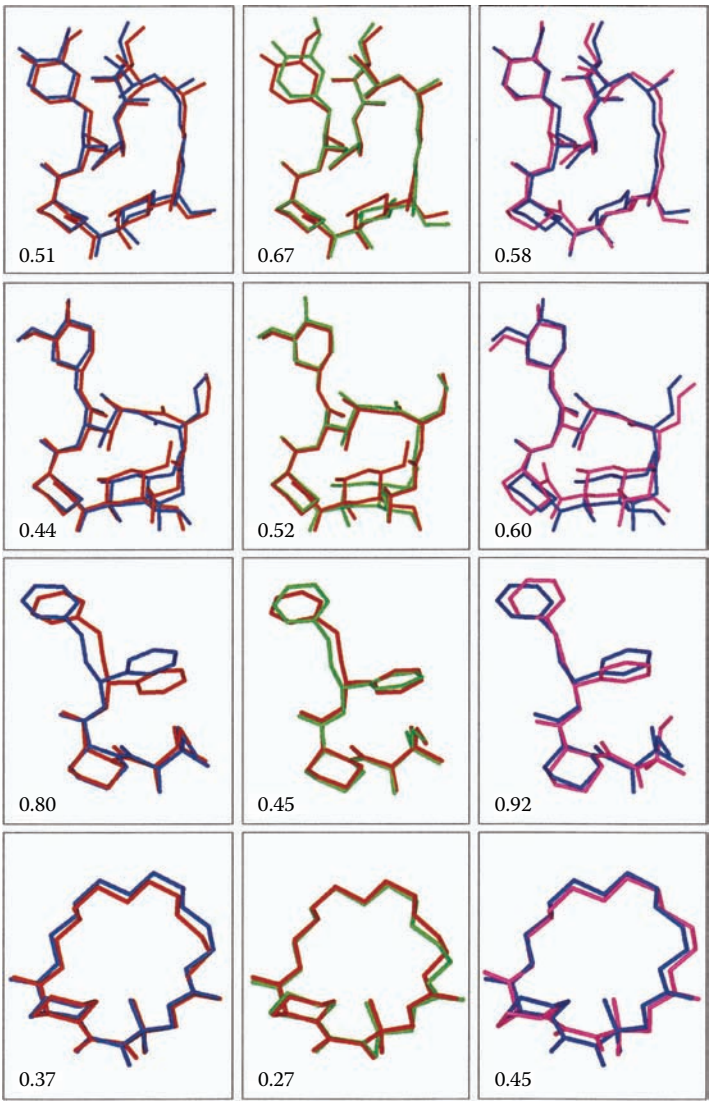
COLOR FIGURE 7.10 Interfragment interaction energy (IFIE) analysis for interactions between cyclic-AMP receptor protein (CRP) and DNA calculated at the MP2/6-31G level. (a) IFIE between the CRP-cAMP complex and each DNA base or backbone fragment. The strength of the interactions is represented according to darkness on DNA. (b) Numerical representation of IFIE between the CRP–cAMP complex and each base pair. Backbone, base, and total indicate backbone (sugar-phosphate) pair fragment, base-pair fragment (base portion only), and nucleotide pair fragment, respectively.



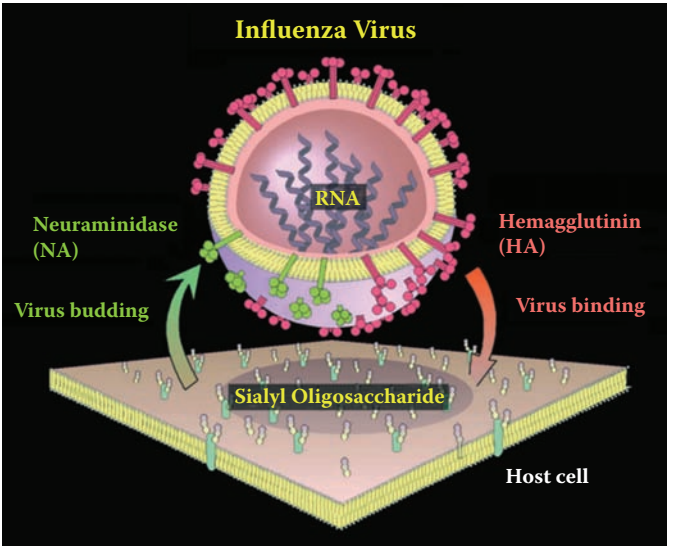
COLOR FIGURE 7.12 Two-dimensional map of the interaction energies (IFIE map) for interactions among DNA, cyclic-AMP receptor protein (CRP), and cyclic-AMP (cAMP) calculated at the MP2/6-31G level. Upper and lower triangles indicate the plots of negative and positive energy values, respectively. The Fragment Nos. 1 to 200, 201, 202 to 223, and 224 to 245 correspond to CRP, cAMP, DNA base, and DNA backbone, respectively.



COLOR FIGURE 7.17 Graphic representation and visualized interfragment interaction energy (IFIE) results of influenza hemagglutinin (HA) antigen–antibody system. (a) Antigen and antibody (Fab-fragment) are drawn in green and purple colors, respectively. The crucial antigenic sites A through E in HA1 domain are indicated with special colors and labels (the PDB data 1EO8). (b) IFIEs of the residues in antigen with the whole Fab-fragment of antibody; (c) IFIEs of the residues in antibody with the HA antigen. MP2 correction was included. Red and blue colors mean the interaction energies of stabilization and destabilization, respectively.

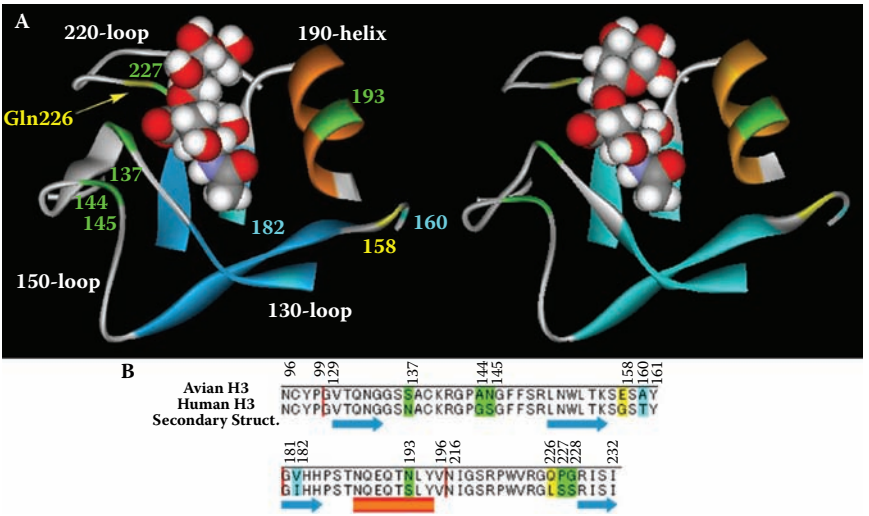


COLOR FIGURE 8.2 Comparison of the optimized ligand structures (hydrogen atoms are omitted for clarity). (a) PDB structures (red) and optimized structures in the complex (blue), (b) PDB structures (red) and optimized structures in the crystal environment (green), and (c) optimized structures in the complex (blue) and those in isolated state (magenta). The numerical values indicate the root mean square deviation (RMSD) between the two geometries (in Å). Each column is in the order of 1fkb, 1fkf, 1fkg, and 1fki ligands from the top.

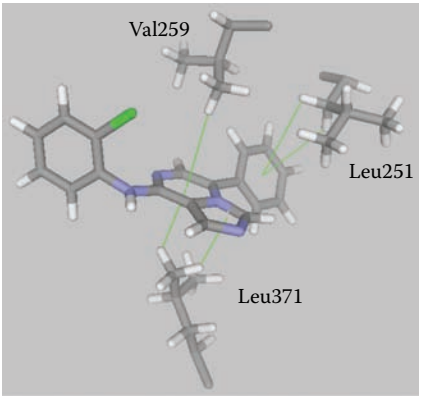


COLOR FIGURE 9.1 Influenza virus and host cell. Hemagglutinin (HA) works on the virus binding to host cell surface and on the membrane fusion of virus with the cell (see Figure 9.2). Neuraminidase (NA) hydrolyzes the neuraminic acid (sialic acid)–oligosaccharide linkage on the HA receptor sialosaccharides when the viruses go outside the host cell. (This illustration was drawn by Dr. Osamu Kanie, Mitsubishi Kagaku Institute of Life Sciences, Machida, Tokyo, Japan.)

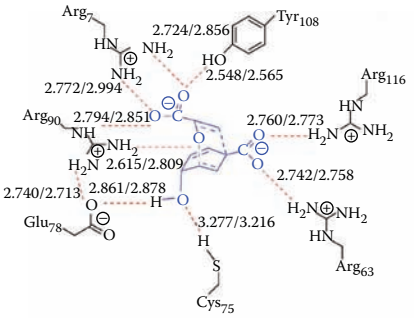
COLOR FIGURE 9.2 Energy minimum structure of avian influenza A viral H3 trimer in complex with avian-type receptor Neu5Acα(2-3)Gal disaccharide analogues. Left: avian H3 trimer-three Neu5Acα(2-3)Gal analogues complex. HA trimer; ribbon. Neu5Acα(2-3) Gal; CPK model. Right: avian H3 monomer-Neu5Acα(2-3)Gal complex. Sialoside binding domain HA1; colorful ribbon, membrane fusion domain HA2; yellow ribbon. (Figures were prepared by Discovery Studio Visualizer v2.0.)



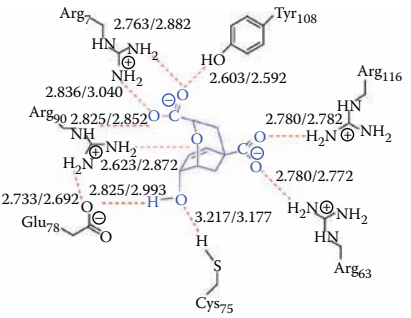
COLOR FIGURE 9.6 Sialoside binding sites on avian and human H3s HA1 in complex with avian Neu5Acα(2-3)Gal and human Neu5Acα(2-6)Gal. (A) Sialoside binding sites consist of Asn96-Pro99, Gly129-Tyr161, Gly181-Val196, and Asn216-Ile232 (total 70 amino acids). Left: avian H3-avian Neu5Acα(2-3)Gal complex. Right: human H3–human Neu5Acα(2-6)Gal complex. (B) Sequence alignments of avian and human H3s in the models; yellow, nonmatching residues; green, weak matching residues; light blue, strong matching residues. The substitution positions are shown in Figure 9.6A, left. (Reproduced from Sawada, T., Hashimoto, T., Tokiwa, H. et al. 2008. *Glycoconj. J.* 25: 805–815. With permission.)



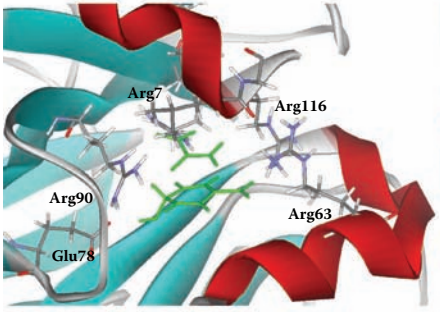
COLOR FIGURE 10.13 Interactions between **3** and Leu251, Val259, and Leu371 of LCK. The lines indicate CH/π hydrogen bonds.



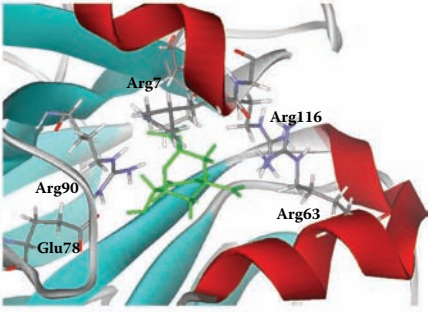
(a)



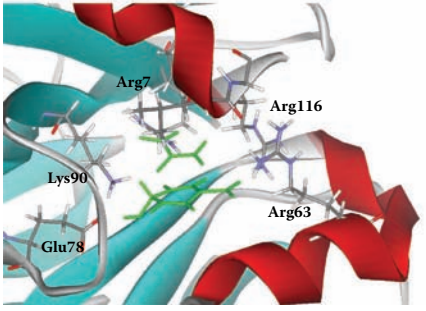
(b)



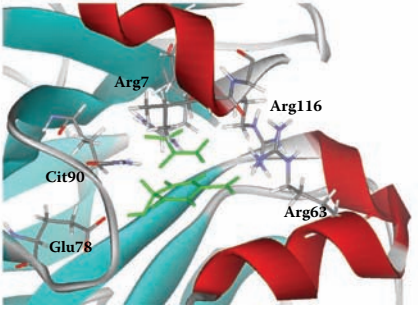
(c)



(d)



(e)



(f)

COLOR FIGURE 11.4 (Top panel) (a) The hydrogen-bonding network at the active site of wild-type BsCM. In parentheses, the left data show the optimized geometries in the small-QM region model, while the right data are those in the large-QM region model. (b) The same schematic picture of the enzyme-TSA complex (EI). (Middle panel) (c) QM/MM optimized structure of the enzymatic TS in the wild-type BsCM active site. Only important residues around the active site are drawn as sticks. The green stick designates the TS geometry of the substrate. (d) QM/MM optimized structure of the EI complex. (Bottom panel) (e) QM/MM optimized structure of the TS in Lys90 mutant reaction. (f) QM/MM optimized structure of the TS in Cit90 mutant reaction.

General Chemistry/Computational Chemistry

The FRAGMENT MOLECULAR ORBITAL METHOD

PRACTICAL APPLICATIONS TO LARGE MOLECULAR SYSTEMS

EDITED BY
Dmitri G. Fedorov and Kazuo Kitaura

Answering the need to facilitate quantum-chemical calculations of systems with thousands of atoms, Kazuo Kitaura and his coworkers developed the Fragment Molecular Orbital (FMO) method in 1999. Today, the FMO method can be applied to the study of whole proteins and protein–ligand interactions, and is extremely effective in calculating the properties of biological systems and molecular clusters.

Providing a unique and accessible approach, **The Fragment Molecular Orbital Method: Practical Applications to Large Molecular Systems** is for those researchers eager to obtain useful information from electronic structure calculations of large systems, and for those who wish to know what can be elucidated with the calculations at present and in the near future. The text emphasizes the practical aspects, with as little mathematical detail as possible and in language that is easy to understand.

The free modeling software Facio, in which FMO-related functions are implemented, is provided on the accompanying CD-ROM, which also provides input file samples, usage hints, annotated output from typical calculations, easy-to-follow tutorial material, and AppliGuide movies that show the sequence of mouse operations for data processing.

The book encourages readers to perform their own calculations — describing the features of the freely available FMO programs (GAMESS and ABINIT-MP) and reviewing many successful applications of the FMO method to practical problems. Filled with practical advice from the inventors of the method and from world-renowned contributors, this reference provides general scientists with the foundation required to use FMO computational methods in a wide range of biomolecular applications, including drug design, protein–ligand binding, enzyme reactivity, and light-driven processes.

Developers interested in extending FMO capabilities or in advancing their own methods will find sufficient information and mathematical detail to encourage method development.

78488
ISBN: 978-1-4200-7848-0
90000
9 781420 078480
www.crcpress.com

CRC Press
Taylor & Francis Group
an informa business

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK