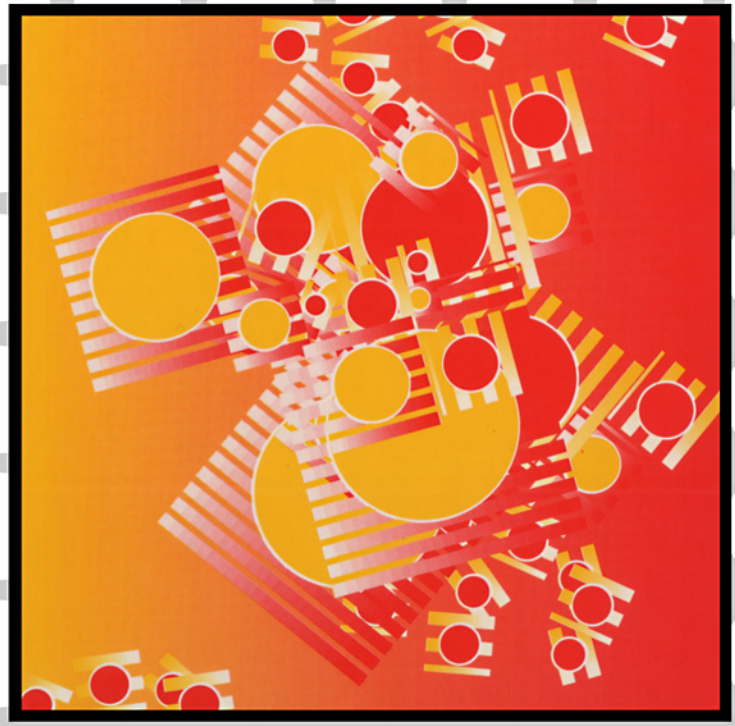
ELLIS HORWOOD SERIES IN
PHARMACEUTICAL TECHNOLOGY

ADVANCED DRUG DESIGN AND DEVELOPMENT

a medicinal chemistry approach

editors
P. N. KOUROUNAKIS and E. REKKA



ADVANCED DRUG DESIGN AND DEVELOPMENT

ELLIS HORWOOD SERIES IN PHARMACEUTICAL TECHNOLOGY

incorporating Pharmacological Sciences

Series Editor: Michael H.Rubinstein, Professor of Pharmaceutical Technology, School of Pharmacy, Liverpool John Moores University

Series Consultant: Professor C.G.Wilson, J.P.Todd Professor of Pharmaceutics, Strathclyde University

Titles published (still in print) and in press

Armstrong & James
UNDERSTANDING EXPERIMENTAL DESIGN AND INTERPRETATION IN PHARMACEUTICS

Cartwright & Matthews
PHARMACEUTICAL PRODUCT LICENSING: Requirements for Europe

Cartwright & Matthews
INTERNATIONAL PHARMACEUTICAL PRODUCT REGISTRATION: Quality, Safety, Efficacy

Clark & Moos
DRUG DISCOVERY TECHNOLOGIES

Cole
PHARMACEUTICAL PRODUCTION FACILITIES: Design and Application

Cole, Hogan & Aulton
PHARMACEUTICAL COATING TECHNOLOGY

Cook
POTASSIUM CHANNELS: Structure, Classification, Function and Therapeutic Potential

Damaini
SULPHUR-CONTAINING DRUGS AND RELATED ORGANIC COMPOUNDS: Volumes 1–3, Parts A and B

D'Arcy & McElnay
PHARMACY AND PHARMACOTHERAPY OF ASTHMA

Denyer & Baird
GUIDE TO MICROBIOLOGICAL CONTROL IN PHARMACEUTICS

Doods & Van Meel
RECEPTOR DATA FOR BIOLOGICAL EXPERIMENTS: A Guide to Drug Selectivity

Ford & Timmins
PHARMACEUTICAL THERMAL ANALYSIS: Techniques and Applications

Glasby
DICTIONARY OF ANTIBIOTIC-PRODUCING ORGANISMS

Hardy et al.
DRUG DELIVERY TO THE GASTROINTESTINAL TRACT

Harvey
DRUGS FROM NATURAL PRODUCTS: Pharmaceuticals and Agrochemicals

Hider & Barlow
POLYPEPTIDE AND PROTEIN DRUGS: Production, Characterization, Formulation

Ioannides & Lewis
DRUGS, DIET AND DISEASE, Volume 1: Mechanistic Approaches to Cancer

Izquierdo & Medina
NATURALLY OCCURRING BENZODIAZEPINES: Structure, Distribution and Function

Junginger
DRUG TARGETING AND DELIVERY: Concepts in Dosage Form Design

Krogsgaard-Larsen & Hansen
EXCITATORY AMINO ACID RECEPTORS: Design of Agonists and Antagonists

Kourounakis & Rekka
ADVANCED DRUG DESIGN AND DEVELOPMENT

Labaune
HANDBOOK OF PHARMACOKINETICS: The Toxicity Asssessment of Chemicals

Macheras, Reppas & Dressman
BIOPHARMACEUTICS OF ORALLY ADMINISTERED DRUGS

Martinez
PEPTIDE HORMONES AS PROHORMONES

Rainsford
ANTI-RHEUMATIC DRUGS: Actions and Side Effects

Ramabhadran

PHARMACEUTICAL DESIGN AND DEVELOPMENT: A Molecular Biological Approach

Ridgway Watt
TABLET MACHINE INSTRUMENTATION IN PHARMACEUTICS: Principles and Practice

Roth et al.
PHARMACEUTICAL CHEMISTRY, Volume 1: Drug Synthesis

Roth et al.
PHARMACEUTICAL CHEMISTRY, Volume 2: Drug Analysis

Rubinstein
PHARMACEUTICAL TECHNOLOGY, Volume 1: Controlled Drug Release

Russell & Chopra
UNDERSTANDING ANTIBACTERIAL ACTION AND RESISTANCE

Taylor & Kennewe
MODERN MEDICINAL CHEMISTRY

Theobald
RADIOPHARMACEUTICALS: Using Radioactive Compounds in Pharmaceutics and Medicine

Vergnaud
CONTROLLED DRUG RELEASE OF ORAL DOSAGE FORMS

Washington
PARTICLE SIZE ANALYSIS IN PHARMACEUTICS AND OTHER INDUSTRIES

Washington et al.
PHARMACOKINETIC MODELLING USING STELLA ON THE APPLE MACINTOSH (TM)

Wells
PHARMACEUTICAL PREFORMULATION

Wells & Rubinstein
PHARMACEUTICAL TECHNOLOGY Controlled Drug Release, Volume 2

Wells & Rubinstein
PHARMACEUTICAL TECHNOLOGY Tableting Technology, Volume 2

Wilson & Washington
PHYSIOLOGICAL PHARMACEUTICS

ADVANCED DRUG DESIGN AND DEVELOPMENT

A Medicinal Chemistry Approach

Edited by

P.N.Kourounakis and E.Rekka

ELLIS HORWOOD
NEW YORK LONDON TORONTO SYDNEY TOKYO
SINGAPORE

First published 1994 by
Ellis Horwood Limited
Campus 400, Maylands Avenue
Hemel Hempstead
Hertfordshire, HP2 7EZ

A division of
Simon & Schuster International Group

This edition published in the Taylor & Francis e-Library, 2005.

Library of Congress Cataloging-in-Publication Data
Available from the publisher

British Library Cataloguing in Publication Data
A catalogue record for this book is available from
the British Library

ISBN 0-203-99988-6 Master e-book ISBN

ISBN 0-13-336793-2 (Print Edition)

1 2 3 4 5 98 97 96 95 94

DEDICATION

This volume is dedicated to our graduate students, worthy and motivated research workers in the science we all serve.

Table of Contents

Preface

The Department of Pharmaceutical-Medicinal Chemistry, School of Pharmacy, Aristotelian University of Thessaloniki, organizes, every other year, a graduate seminar in advanced medicinal chemistry, with international participation, in Thessaloniki. The purpose of this seminar is to help young and ambitious pharmaceutical scientists in this field by updating their knowledge and informing them about the new trends in our science, through the presentations of renowned invited speakers. It is also aimed to create a forum for exchange of ideas in medicinal chemistry, in a different informal environment, bringing internationally recognized medicinal chemists closer to their Greek colleagues. Some of the invited speakers in this series of seminars are: A.H.Beckett, A.Benakis, N.Bodor, D.D.Breimer, J.Caldwell, R.Ganellin, J.Gorrod, R.Franke, H.Kappus, P.Krogsgaard-Larsen, A.Makriyiannis, J.Seydel, H.Sies, J.Tollenaere and W.Wiegrebe.

This volume, entitled *Advanced Drug Design and Development: A Medicinal Chemistry Approach,* is a collection of lectures by most of the invited speakers in our second seminar.

We hope that this book will be a source of inspiration to all young medicinal chemists, providing examples in drug design from various pharmacochemical approaches.

Before closing, we would like to thank all our invited speakers presented in the text, and, last but not least, all our colleagues and staff at the Department of Pharmaceutical-Medicinal Chemistry, without whose hard work these seminars could not have materialized: V.Demopoulos, A.Geronikaki, D.Hadjipavlou-Litina, E.Sotiropoulou-Dimitriou, Ek. Tani, D.Spiriounis, I.Andreadou, A.Gavalas, G.Rekatas, R.Mgonzo, A.Alexidis, M.Tomi-Hadji, A.Polymerou-Kapoula, M.Diamanti and M.Triandaphyllidou.

P.N.Kourounakis
E.A.Rekka
Department of Pharmaceutical Chemistry
School of Pharmacy
Aristotelian University of Thessaloniki
Thessaloniki 540 06, Greece

1
Introduction

P.N.Kourounakis and E.A.Rekka

Department of Pharmaceutical Chemistry, School of Pharmacy, Aristotelian University of Thessaloniki, Thessaloniki 540 06, Greece

The discovery of drugs and drug molecules has always been the aim of pharmaceutical sciences and, in particular, of medicinal chemistry, which evolved from pharmaceutical chemistry. Half a century ago, pharmacochemistry, the modern expression of pharmaceutical chemistry, as a science whose main interest is the design and development of new pharmacomolecules, was at the beginning of its evolution. Drug design in its broad sense and structure-activity relationship studies are essential and at the heart of medicinal chemistry, and it is the progress and development of this field of research that has made medicinal chemistry the modern and enormously productive science it has become in recent decades [1]. Today, studies on structure-activity relationships and their influence on the design of new drugs have rendered them one of the most useful and thus important activities of pharmacochemistry, a modern component science in the group of pharmaceutical sciences [2].

Despite the advances in medical and pharmaceutical sciences, there are still many diseases which are incurable or can only be treated symptomatically, and at a great economic and social cost owing to only moderately effective or even to the lack of appropriate therapeutic agents. Of the 30000 or so diseases or disorders currently known, only one-third can somehow be treated with drugs. Furthermore, there are incurable maladies, like viral diseases (influenza, AIDS), CNS disorders (Alzheimer's disease), cancer and autoimmune disorders, which can be fatal or cause great suffering and disability [3]. Therefore, there is still a great need for more and better drugs—more active and selective, drugs with fewer undesired or toxic side-effects, agents useful in prophylaxis and drugs which will cause as little as possible harmful contamination in the already polluted environment.

In a systematically planned programme of drug discovery, several questions have to be answered:

— Is the research for the discovery of a certain drug justified by the medical expectation?
— How will the expected drug contribute to health?
— What would be the economic or other, more noble benefit that is expected from the drug?

— Is the state of the art of medicinal chemistry at a satisfactory level so that the risk of investing in the project should be taken? That is, have the coordinated attempts a favourable possibility for solving the problem in a reasonable time period?
— Does the specific disease affect sufficient people for the economic attempt to be justified? It is tragic that serious diseases, mainly in developing countries, are sometimes uncontrolled because of a lack of effective therapeutic agents due to the non-existent financial profit [3].

Because of the strict prerequisites of national drug authorities, which are becoming ever more demanding, the cost of drug discovery is steadily increasing. Thus, rational drug design becomes the main objective of medicinal chemistry today. Based on rational design, new structures can be developed with a high probability of possessing the required properties. The setting of clear rules to help in the access to information hidden in accumulated experimental data is necessary, and this requires studies on the quantitative relationships between (physicochemical) properties and (biological) activity.

We are thus led to the selection of a subsystem of compounds originating from an initial structure, the *lead compound,* the discovery of which is the most decisive step in the process of drug discovery. Methods used in lead compound discovery include: folk/ethno-pharmacy and therapeutics; massive pharmacological screening; fortuitous discovery; modification of bioactive natural products; exploitation of secondary or side-effects of drugs; study of the basic processes of life; body biochemistry and the use of metabolic analogues; study and exploitation of differences in molecular biology, differential cytology, biochemistry and endocrinology; study of the biochemistry of diseases; an approach through the molecular mechanism of drug action; analysis of the mechanism of action of multipotent compounds; drug metabolism (hard, soft, pro-drugs); and chemical delivery systems [2, 5–7].

The pharmacochemical manipulations following the discovery of the lead compound include: attempts aimed at the development of substitutes of existing biologically active molecules; attempts aimed at the alteration of the activity spectrum of biologically active molecules; attempts aimed at the modification of the pharmacokinetics of the compounds used as drugs or as lead compounds; structural changes in natural products; molecular transformations performed by microbes; and other chemical processes that follow the molecular manipulations on the lead compounds (for example, construction of homologous series, application of the rule of bioisosteric groups, resolution of stereoisomers) [2, 5].

It is evident that in the process of drug development the molecular structure is the main feature that determines the molecular properties, and thus whether the particular molecule finally reaches the patient [8].

Since, in the majority of drugs, action appears after the interaction of the pharmacomolecule with its receptor, it seems reasonable to study the drug structure in relation to its receptor site (the dynamic aspect). For a productive

drug-receptor interaction a good fit, determined by physicochemical properties, is required. Besides solving the basic synthetic problems, studies on the geometry and shape, conformational analysis and investigation of the influence of electronic and hydrophobic effects on drug-receptor interaction are performed. In quantitative structure-activity relationship studies attempts have been made, with success, to correlate quantitatively biological activity with molecular properties (electronic, hydrophobic, steric). This relationship has been based on the assumption that the relative importance of physicochemical properties for biological activity can be described numerically, for an objective evaluation of drug-receptor interactions. Numerous methods have been invented for the quantification of electronic, hydrophobic and steric effects of functional groups [9]. Statistical methods, mainly Hansch or extrathermodynamic analysis [10], as well as those of Free and Wilson [11], pattern recognition/principal components analysis and cluster analysis, can lead to the prediction and optimization of activity, and ultimately to the design of better drugs. The development of powerful, interactive computers and molecular graphics systems helps in the analysis and visualization of biologically active compounds and in a better understanding of drug-receptor interactions. Techniques have been developed for the determination and visual presentation of pharmacophores (receptor mapping), as well as techniques for drug design based on a knowledge of receptor structure (receptor fitting) [2, 12].

The pharmacomolecule, before interacting with its receptor (this interaction being direct or indirect, a simple binding—affinity—or a productive interaction yielding a biological effect—efficacy), must reach, intact and in satisfactory concentration, the immediate environment of the receptor site. Access to the receptor is also determined by the physicochemical properties of the molecule. Thus, structure plays a decisive role not only in the dynamics, but also in the kinetics of the drug molecule.

Molecular structure is usually altered by the body. Drug metabolism, basically an adaptive process, is a rather useful property of the (liver) cell, as a whole. Drug biotransformation usually leads to more polar compounds, and thus to faster elimination, and to substances with lower or no activity. Only rarely is an increase of activity observed after biotransformation. However, in certain cases, very dangerous highly (chemically) reactive metabolic intermediates are formed. During drug metabolism, and through the catalytic activity of enzymes like the cytochrome P-450 family, prostaglandin synthase and xanthine oxidase, free radicals may be formed, which participate in the initiation and propagation of chain reactions. Oxygen is activated, and the presence of active oxygen species ($O_2^{-}\cdot$, H_2O_2, $HO\cdot$) may lead, via lipid peroxidation or other cellular structure damage, to cell injury and necrosis [12]. Numerous pathophysiological conditions are probably due to radical attack and oxidative damage [13]. A knowledge of the pathophysiology of diseases constitutes a decisive step towards the discovery of lead compounds. This could be conducted in various ways, for example by the study either of free radical scavenging activity or of free radical formation.

Therefore, drug metabolism and, in particular, relationships between the structure of the drug molecule and the enzyme systems responsible for drug biotransformation, resulting in detoxification, but also in biotoxification, are currently subjects of active pharmacochemical investigation. The ever-increasing number of modern, improved drug molecules, the discovery of which is based upon a knowledge of drug biotransformations and oxygen activation, supports the argument for the prominent position held by drug metabolism and free radical pharmacochemistry in currently used rational drug design techniques.

This volume covers topics such as drug discovery and physicochemical properties, structure-activity relationships, structure-interaction with specific receptor subtypes, and in combating serious diseases that cause great financial and social problems, for example Alzheimer's disease and gastric ulceration. Also discussed is the dependence of the biological properties of a compound on chemical structure, in terms of quantitative structure-activity relationships, the merits and shortcomings of computational chemistry and the techniques applied to gaining insight into the complex molecular phenomena in innovative drug research, the characterization and prediction of drug metabolism in humans and the importance of labelling of bioactive compounds in the study of the dynamics but mainly the kinetics of a prospective drug.

These topics are presented by contributors, each one a specialist in his or her own field within the greater subject of pharmacochemistry. We are certain that the following chapters, tackling the subject of drug design from different viewpoints, will stimulate the creativity of those involved or interested in innovative drug research. Young medicinal chemistry investigators could be helped and inspired in their attempts to find new and better drug molecules among the structures waiting to be discovered.

REFERENCES

1 [] Testa, B. (1992) *Chimia,* **46,** 297.
2 [] Kourounakis, P.N. and Rekka, E. (1992) in *Drug Design* (eds Simonis, A. and Hadjipantou, P.), Thessaloniki, Greece (in Greek) and Mediset, Geneva, in press (in English).
3 [] Büchel, K.H. (1987) in *Trends in Medicinal Chemistry* (eds Mutschler, E. and Winterfeldt, E.), VCH, Weinheim, p. 3.
4 [] Tute, M.S. (1990) in *Comprehensive Medicinal Chemistry,* vol. 4 (ed. Hansch, C.), Pergamon Press, Oxford.
5 [] Kourounakis, P.N. and Rekka, E. (1993) *Iatriki,* **63,** 128.
6 [] Bodor, N. (1982) in *Strategies in Drug Research* (ed. Keverling Bruissman, J.A.), Elsevier, Amsterdam, p. 137.
7 [] Bodor, N. (1987) in *Trends in Medicinal Chemistry* (eds Mutschler, E. and Winterfeldt, E.), VCH, Weinheim, p. 195.
8 [] Albert, A. (1985) in *Selective Toxicity: The Physicochemical Basis of Therapy,* Chapman & Hall, London.

9 [] Franke, R. (1984) in *Theoretical Drug Design Methods* (eds Nauta, Th. and Rekker, R.F.), Pharmacochemistry Library, Vol. 7, Elsevier, Amsterdam.
10 [] Hansch, C. and Leo, A. (1979) in *Substituent Constants for Correlation Analysis in Chemistry and Biology,* Wiley, New York.
11 [] Free, S.M. and Wilson, J.N. (1964) *J. Med. Chem.,* **7,** 395.
12 [] Goodford, P.J. (1984) *J. Med. Chem.,* **27,** 557.
13 [] Sies, H. (1985) in *Oxidative Stress,* Academic Press, London.

2
Structure-activity studies and drug design with antagonists at histamine H_2 receptors

C.R.Ganellin

Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, UK

SUMMARY

H_2 receptor histamine antagonists now have a considerable diversity of chemical structure but incorporate at least two planar π-electron systems which are polar and have a propensity towards hydrogen bonding.

The activities of a series of antagonists structurally related to cimetidine have been compared to investigate the effect of replacing the cyanoguanidine moiety by other neutral, dipolar groups. Antagonist activity (measured *in vitro* on the histamine-stimulated right atrium of the guinea-pig) is very sensitive to relatively minor structural changes. Differences in activity are accounted for by dipole moment orientation and lipophilicity for 15 different structural groups according to the equation $-\log K_B = 9.1 \cos\theta + 0.60 \log P - 2.7$.

It is suggested that drug molecules in aqueous solution are in a water-solvent shell and have to undergo desolvation (log *P* octanol: water, representing a hydrophobic effect) and align themselves at a polar receptor; they then undergo hydrogen bonding and dipolar interactions with the receptor.

H_2 receptor histamine antagonists do not readily enter the brain and it appears that this may largely be due to their high polarity coupled with a relatively low lipophilicity. A physicochemical model relating to the difference between octanol/water and cyclohexane/water log *P* values, which is regarded as an approximate measure of hydrogen-bonding ability, correlates with brain/blood concentration ratios. A series of 3-(3-(l-piperidinylmethyl)phenoxy)propyl derivatives was sythesized with the aim of reducing hydrogen bonding ability whilst retaining H_2 antagonist potency, and several novel compounds were obtained which readily crossed the blood-brain barrier. In particular, zolantidine (SK&F 95282), an aminobenzo-thiazole derivative having an atrial pA_2 of 7.65 and a brain/blood concentration ratio of 1.4, has been identified for use in studying histaminergic H_2 receptor mechanisms in the brain.

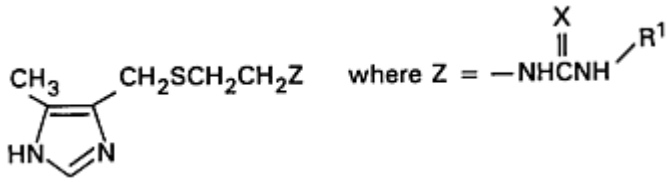


Fig. 1. General structure of a series of H_2 receptor histamine antagonists (cimetidine analogues).

INTRODUCTION

Three types of histamine receptor have now been characterized pharmacologically according to the antagonists and agonists which interact with them. They are designated H_1 [1], H_2 [2] and H_3 [3, 4]. This chapter describes some structure-activity studies concerned with designing antagonists at H_2 receptors as potential therapeutic drugs.

The first clinically valuable H_2 receptor histamine antagonist was cimetidine (Fig. 1, X=NCN, R^1=CH_3), which was introduced into many countries in 1977 and revolutionized the treatment of peptic ulcer disease [5–7]. In seeking alternative chemical structures as H_2 receptor antagonists various hydrogen-bonding polar groups were examined as alternatives to the cyanoguanidine group in cimetidine [8], including the isocytosine group (2-amino-pyrimidin-4-one).

It became of interest to study the chemical properties of these compounds to gain some insights into how they may interact with the H_2 histamine receptor. The analysis was simplified by concentrating on a series of antagonists depicted in Fig. 1 whose structures were closely related to cimetidine. The studies emphasized the importance of such physicochemical properties as geometry, acidity (pK_a), hydrophilicity (octanol-water partition, P) and dipole moment (electron density distribution).

In cimetidine, the cyanoguanidine moiety is regarded as a 'polar hydrogenbonding group' and initially a small chemically homogeneous series of 12 cimetidine analogues was studied (Fig. 1, X=O, S, N–CN, N–NO_2, N–CO– or $CHNO_2$).

The rules used of selection of structure in the first correlation analysis are as follows:

(1) The 'aromatic group' and chain are kept constant, that is, 2-[(5-methylimidazol-4-yl)methylthio]ethyl.
(2) Z is a planar, 1, 3-amidine NH system which is formally uncharged at pH 7. 4, that is, pK_a (proton gained) <5 and (proton lost) >9.

(3) In the 'hydrogen-bonding group', Z, alkyl substitution on C is excluded to avoid introducing new steric effects, but replacement of $NHCH_3$ by NH_2 is

permitted since this is of obvious importance to consideration of hydrophilicity; cyclic forms of acyl guanidine such as isocytosine are also included.

BIOLOGICAL ACTIVITY

In this study, the activity of H_2 receptor histamine antagonists was determined *in vitro* against the histamine-stimulated increase in the rate of beating of the guineapig right atrium at 34°C, by the method described by Parsons *et al*. [9]. Dose ratios *(A)* were calculated as the ratio of histamine concentrations required to produce half-maximal responses in the presence and absence of different concentrations *(B)* of antagonist, and dissociation constants *(K_B)* were derived from the equation $K_B=B(A-1)$. The compounds appeared to behave as simple competitive antagonists.

CORRELATION WITH OCTANOL-WATER PARTITION

For most of the compounds there appeared to be a good correlation between activity and a lipophilicity parameter (octanol–water partition for the model compounds HZ) in which a 10-fold increase in *P* brought about a 100-fold increase in activity. There were, however, two notable exceptions, namely the thiocytosine and the 1,1-diamino-2-nitroethene (see Ganellin [8]).

The thiocytosine was much less active (about one-hundredth) than predicted and the 1,1-diamino-2-nitroethene (Fig. 1, X=$CHNO_2$, R^1=CH_3) was much more active than predicted (approximately 30-fold), and it was concluded that some property other than the lipophilicity parameter was making a marked contribution to activity.

Partition coefficients were measured by a conventional shake-flask technique at 37°C [10]. The concentrations of the compound in the aqueous phase before and after partitioning were determined spectrophotometrically. Buffer salts were used to control the pH of the aqueous phase. Log *P* is the partition coefficient of the neutral form of the compound. It is assumed that only the neutral form of the compound partitions into the organic phase, whence log *P* is related to the apparent partition coefficient (P_a) of the compound, measured at a pH at which the neutral and monoprotonated forms are in equilibrium in the aqueous phase, and to the dissociation constant (K_a) of a proton from the most basic site in the molecule, by equation (1):

$$\log P = \log P_a + \log (1 + 10^{pK_a - pH}) \qquad (1)$$

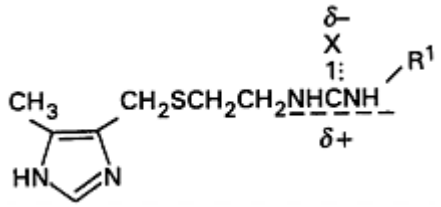


Fig. 2. General structures showing the partial charge separation in the 'polar H-bonding group'.

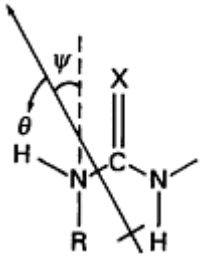


Fig. 3. Definition of the orientation parameter, that is, the deviation, from the optimal angle, of the dipole orientation with respect to the side-chain N-R bond. The optimum for ψ was found to be 30°, and the ideal orientation would have θ=0°.

CORRELATION WITH DIPOLE ORIENTATION

One property which cyanoguanidine has in common with other neutral moieties found in active analogues of cimetidine is high polarity. Simple derivatives of urea, thiourea, cyanoguanidine and 1,1-diamino-2-nitroethene have high dipole moments, especially in aqueous solution, where charge-separated canonical structures are believed to make important contributions to the overall electronic distribution (Fig. 2).

Consideration of the dipolar nature of the above groups prompted the measurement of the dipole moment of model compounds [11, 12] and led to the interesting proposition that it might be possible to account for activity by considering hydrophilicites and dipole orientation [13].

To investigate the importance of dipole orientation, CNDO/2 molecular orbital calculations were performed for each dipolar group, and estimates made of the dipole orientation with respect to the side-chain R–N bond, that is, angle ψ. The possible existence of an optimum value for ψ within the series was then examined by calculating the deviation θ (Fig. 3) from various arbitrary angles of ψ and assessing their biological significance.

In this study, only the *N*-methylated derivatives of the acyclic dipolar group were included (compounds **1**–**4**), together with four cyclic analogues **(5–8)**. For the eight compounds shown in Fig. 4, only a poor correlation was obtained using the dipole moment vector term, μ cos θ, but a surprisingly good correlation was obtained (correlation coefficient r=0.86) between antagonist activity and cos θ alone.

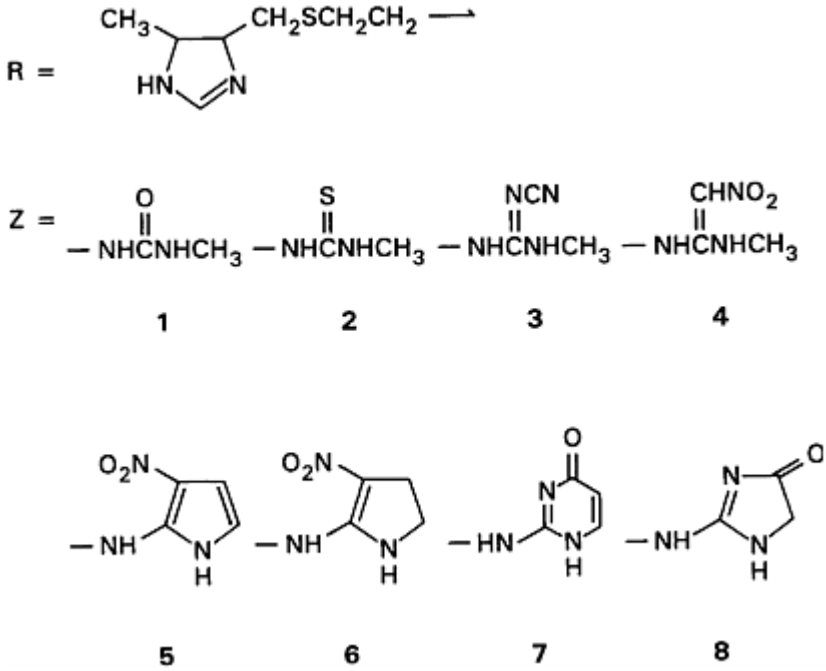


Fig. 4. Structures of the eight compounds (RZ) used in a structure-activity correlation between $-\log K_B$ (inhibition of histamine-stimulated increase in the rate of beating of the guinea-pig right atrium *in vitro*) and a function of log P (octanol-water partition of RZ) and dipole orientation, cos θ.

The correlation was improved by including a log P term (octanol-water partition for each compound RZ) and calculating the deviation from a value of 30°C for the angle ψ:

$$-\log K_B = 0.72 \log P + 10.3 \cos \theta - 3.9$$
$$n = 8,\ r = 0.94,\ s = 0.41 \qquad (2)$$

In searching for a more general relationship, the correlation was extended to include all the compounds for which there was adequate experimental data. A further five active structures (**11**–**15** in Fig. 5) were thereby accommodated; the expanded set was correlated by a very similar equation:

$$-\log K_B = 0.60 \log P + 9.1 \cos \theta - 2.7$$
$$n = 13,\ r = 0.91,\ s = 0.41 \qquad (3)$$

Equation (3) is highly significant and spans a range of H_2 antagonist activity of more than three orders of magnitude. Two other compounds (**9** and **10**) had no measurable antagonist activity and therefore could not be included in the correlation; their weak activity is, however, predicted by the equation.

Observed and predicted antagonist activities for the 13 compounds are compared graphically in Fig. 6. The work is described in detail by Young *et al.* [13].

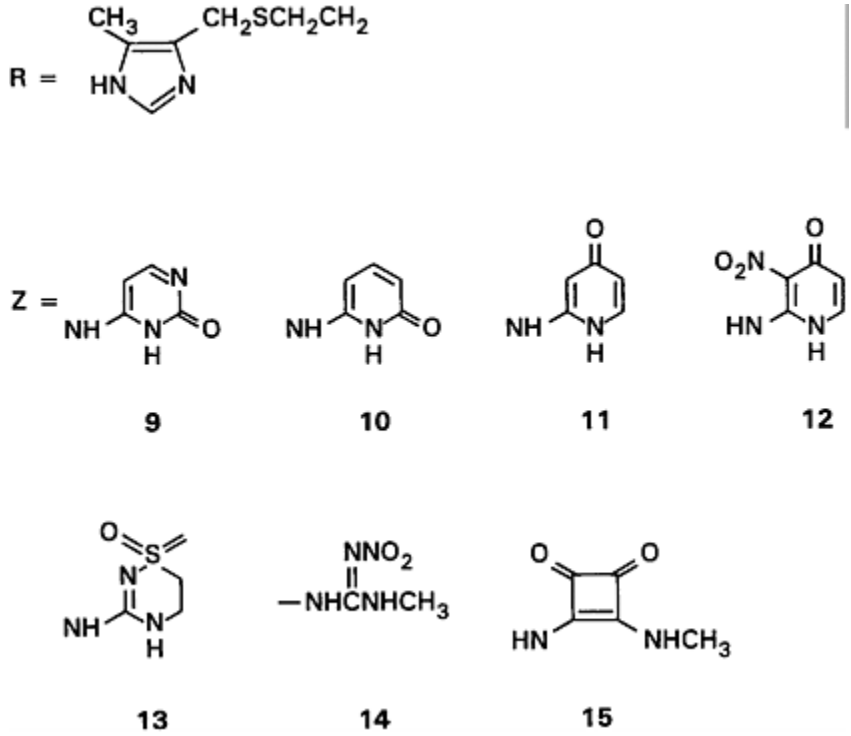


Fig. 5. Structures of the seven additional compounds (RZ) used to extend the correlation.

The high potency of the aminonitropyrrole derivative **5** as an antagonist may now be explained by its near-optimal dipole orientation and relatively high lipophilicity, while the weakly active aminoimidazolinone derivate **8** has an unfavourable dipole orientation and is hydrophilic. Likewise the thiocytosine mentioned above also has an unfavourable dipole orientation.

The diaminocyclobutenedione derivative **15,** which does not strictly conform to the structure in Fig. 1, is also included for comparison. Here, the geometry of the four-membered ring structure causes its dipole to be oriented in an unfavourable direction with respect to the **R–N** bond (θ=15°C). The weak H_2 antagonist activity of this compound is predicted by equation (3). The diaminocyclobutenedione moiety has recently been used in conjunction with different side-chains in compounds such as BMY 25368 (SK&F 94482) [14, 15] and, in contrast, was found to give very potent H_2 receptor antagonists. Presumably different structure-activity relationships apply in these cases.

At first sight it may seem surprising that it is the dipole orientation (cos θ) which correlates with activity rather than the dipole moment vector μ cos θ. It is probable, however, that for the set of structures selected all the compounds are sufficiently dipolar and all are effective hydrogen bond donors. Within this series of analogues, it therefore appears that variations in μ or hydrogen bonding ability

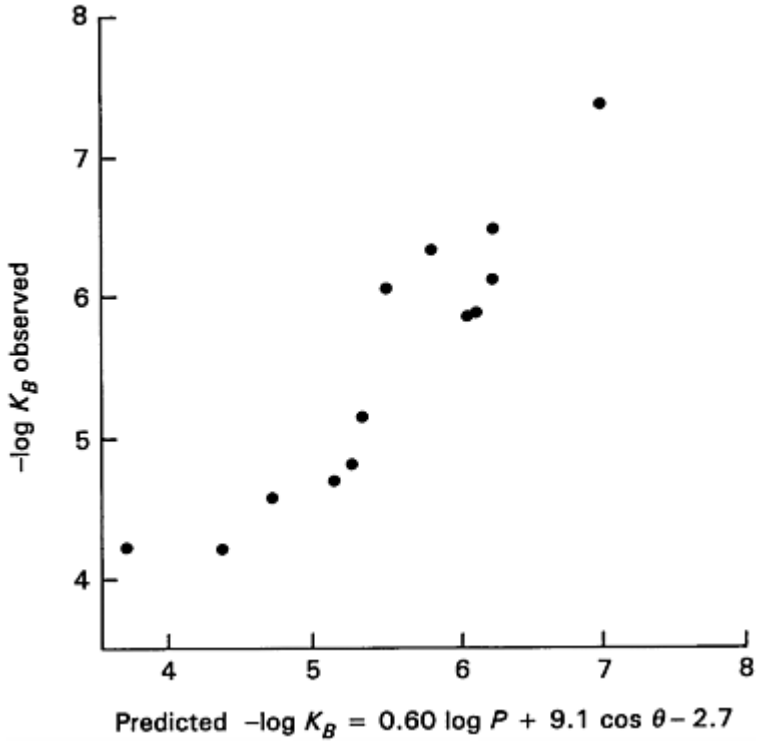


Fig. 6. Comparison of observed (vertical axis) and predicted (horizontal axis; predicted by equation (3): $-\log K_B = 0.60 \log P + 9.1 \cos\theta - 2.7$) H_2 receptor histamine antagonist activities ($-\log K_B$) for the series of antagonists (structures **1–15,** depicted in Figs 4 and 5). Compounds **9** and **10** are too weakly active to be included on the plot. Reprinted from Ganellin and Young [16] with permission from the Publisher, Ellis Horwood Limited.

are small in comparison to the variation in the cos θ term, and the biological effects of any changes are not observed.

Another study [17] of the same set of compounds using Modified Neglect of Differential Overlap (MNDO) optimized structures found a correlation with the dipole moment component μ_x directed along the N–H bond of the nitrogen atom bearing the (5-methylimidazol-4-yl)methylthioethyl chain, as given in equation (4). This correlation, however, has involved a correction to eight of the structures for the energy required to convert them into planar forms:

$$-\log K_B = 0.50\ \mu_x + 1.52 \log P + 3.92$$
$$n = 13,\ r = 0.89,\ s = 0.65,\ F = 18.8 \qquad (4)$$

INTERPRETATION OF THE CORRELATION

The value of such a correlation study is that it may provide a model for considering the mode of action of antagonists, and for suggesting alternative structures for synthesis and testing. The active H_2 receptor histamine antagonists are all very polar molecules containing groups which may enter into strong hydrogen-bonding interactions. Charge delocalization within the polar group results in a dipole moment and also acidifies the NH function, enhancing hydrogen-bonding ability, and it seems very likely that binding between antagonist molecules and the H_2 receptor involves powerful hydrogen bonding.

Within the limitations imposed by the selection of a series of compounds where close structural homogeneity is maintained, it appears from the initial correlation that activity depends on lipophilicity or is inversely related to hydrophilicity. Since these compounds were evaluated *in vitro,* partition is unlikely to represent an overall distribution property, although it could reflect access to the receptor. It is thought to be more probable, however, that partition represents a property or process concerned with receptor interaction, for example, involving desolvation; that is, these very polar molecules would be strongly solvated by hydrogen bonding to surrounding water molecules and would have to dissociate themselves from the surrounding water-solvent shell in order to participate in hydrogen-bonding interactions with the receptor. A reduced hydrophilicity should reflect a greater readiness to desolvate from the water solvent.

In addition to lipophilicity (log P), differences in activity as H_2 receptor histamine antagonists are accounted for by dipole orientation with respect to the side-chain (rather than magnitude of the dipole). It seems probable, therefore, that the amidine-type moiety in this series of compounds has an orientational function rather than being involved in direct dipole-dipole or dipole-charge interactions at the receptor.

Since the orientation cos θ term in the equations was defined by reference to the R–N bond (Fig. 3), it can also be considered to relate to the N–H bond of the same nitrogen atom, suggesting a role for the (R)–NH group in hydrogen bonding to the H_2 receptor. This group is also a reference basis for equation (4).

Thus one may envisage that drug molecules in aqueous solution are in a water-solvent shell and have to undergo desolvation and align themselves through electrostatic interaction at a polar receptor. They then undergo hydrogen bonding and dipolar interactions with the receptor which must take place in a non-aqueous environment.

Dipole-dipole interactions must be of fundamental importance to drug—receptor interactions between polar molecules, yet there have been few structure—activity studies which take account of such contributions. Lien *et al*. [18] cite only 15 references in their review of 1982 on the use of dipole moment as a parameter in drug-receptor interaction and quantitative structure—activity relationship (QSAR) studies. Of particular interest is a more recent study [19] on

dopamine D_1 receptor antagonists in which dipole orientation rather than dipole moment was correlated with receptor binding affinity.

DESIGN OF ZOLANTIDINE, A BRAIN-PENETRATING H_2 RECEPTOR HISTAMINE ANTAGONIST

H_2 receptor histamine antagonists do not readily enter the central nervous system (CNS) and it appears that this may largely be due to their high polarity. It is generally accepted that the rate and extent of entry of a compound into the brain are related to the physicochemical properties of the compound, in particular, its proton dissociation constant (K_a), partition coefficient *(P)* and molecular size. No single partition solvent system has emerged as a universally applicable model for brain penetration, although the octanol—water system has received particular attention, and Hansch and coworkers have suggested [20, 21] that a log partition coefficient (log *P*) of about 2 in this system is optimal for entry into the CNS.

In view of the above it was surprising to find that icotidine (**16,** Fig. 7) (log *P*= 2.58) did not penetrate. Lupitidine (SK&F 93479, **18**), a relatively lipophilic H_2 antagonist [22, 23] which has log *P*=2.33, was also found hardly to penetrate. Neither did an increase in partition to log *P*=4.57, as in SK&F 93619 **(17)** lead to improvement (Table 1). These observations led to a search for a physicochemical model to serve as a guide for increasing the partitioning of H_2 antagonists into the brain [24],

Table 1. Brain penetration data, partition coefficients and histamine H_2 receptor antagonist activities for selected H_2 antagonists. Data from Young *et al*. [24].

Compound				Δ log P[d]	Atrium[e] −log K_B
1 Cimetidine	0.04	0.40	–	–	6.10
16 Icotidine	0.01	2.58	−2.60	5.18	7.49
17 SK&F 93619	0.03	4.57	−0.47	4.10	7.50
18 Lupitidine	0.09	2.33	−1.48	3.81	7.91
19 Roxatidine	–	–	–	–	7.57
20 NHAc analogue	0.35	2.15	0.22	1.93	7.79
21 Zolantidine	1.4	5.41	3.72	1.69	7.46
22 Pyridine analogue	4.9	4.29	3.23	1.06	6.09

[a]Concentration ratio of radiolabelled drug in brain and blood after 2–3 h intravenous infusion in anaesthetized male rats.
[b]Octanol-water partition of free base form.
[c]Cyclohexane-water partition of free base form.
[d]Δ log P=log P_{oct}−log P_{cyh} (equation [6]).
[e]Inhibition of histamine-stimulated beating of the guinea-pig right atrium *in vitro*.

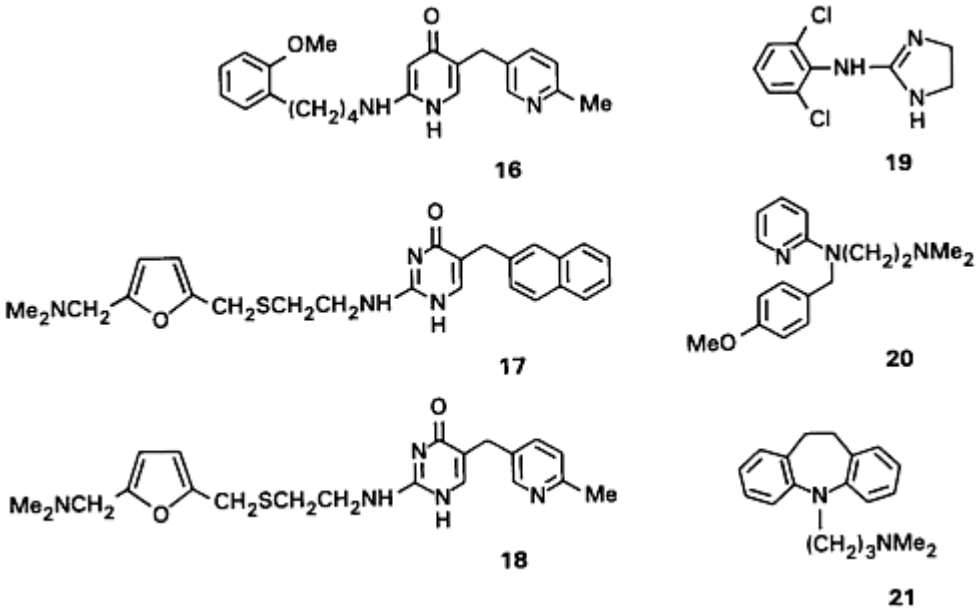


Fig. 7. Six compounds initially studied for brain penetration by measuring the equilibrium brain/blood concentration ratios of radiolabelled drug in the anaesthetized rat, that is, **16** (icotidine) **17** (SK&F 93619), **18** (lupitidine), **19** (clonidine), **20** (mepyramine) and **21** (imipramine).

The above three H_2 antagonists were compared with three standard compounds known readily to enter the brain (clonidine, mepyramine and imipramine) and a highly significant correlation was obtained with Δ log P (equation (5)) defined (equation (6)) as the difference between octanol-water and cyclohexane-water log P values. It appears that Δ log P is related to the hydrogen-bonding ability of a compound as shown in equation (7), where I_H denotes the additive increment to hydrogen bonding from molecular fragments, defined by Seiler [25], and b is a constant:

$$\log C_{brain}/C_{blood} = -0.60\,\Delta \log P + 1.23 \quad (5)$$

$$\Delta \log P = \log P_{oct} - \log P_{cyh} \quad (6)$$

$$\Delta \log P = \Sigma I_H - b \quad (7)$$

Thus, equation (5) suggests that brain penetration might be increased by decreasing the overall hydrogen-bonding ability of a compound. Using this model as a guide, the scope was explored for reducing the effective hydrogen-bonding abilities of different structural types of H_2 antagonists [24]. This posed a considerable problem since there appeared to be a conflicting situation. For H_2 antagonism strong hydrogen-bonding ability appears to be required for receptor

binding (see above), yet to encourage brain penetration hydrogen bonding has to be minimized.

Starting with the archetypal H_2 antagonist drugs cimetidine, tiotidine and ranitidine, structural modifications were made to reduce hydrogen-bonding ability. These led to substantial improvements in brain penetration but still did not give the desired combination of brain penetrability and antagonist potency.

A variant of the ranitidine-type side-chain which has been introduced in some of the more recent H_2 antagonists is the piperidinylmethylphenoxypropyl moiety. This side-chain is of particular interest as it occurs in some of the most potent H_2 receptor antagonists known, and the scope for substitution of polar groups to provide compounds with high *in vitro* antagonist activity appears to be wider than for any other type of aminomethylaryl derivative and includes groups possessing only a single NH function, for example roxatidine and the corresponding acetamide (Table 1) [26]. The acetamide **20,** reported [26] to be a very effective inhibitor of gastric acid secretion, has a relatively low polarity (Δ log P value of only 1.93) and showed a significant level of brain penetration (Table 1). The most polar group in **20** is the amide moiety, which has a hydrogen-bonding contribution (I_H) of 2.56.

A minimal structural requirement for H_2 antagonist activity in compounds of this type appears to be a hydrogen bond donor group and, in attempting to increase brain penetration, the overall hydrogen-bonding ability of analogues of **20** was further decreased by introducing replacements for the amide function having lower I_H values. Three such groups are the sulphonamide, alphatic hydroxyl and aromatic amino groups, reported by Seiler [25] to have I_H values of 1.93, 1.82 and 0.61, respectively; introduction of these in place of the carboxamide group in **20** gave a series of structures which included several potent novel compounds that readily cross the blood-brain barrier [24]. The 2-aminobenzothiazole zolantidine (SD&F 95282, **21**) exhibits a combination of brain penetration (brain/blood ratio=1.4) and selective H_2 antagonist activity (pA_2=7.46), which make it a very suitable candidate for studying histaminergic mechanisms in the brain [27]. Relative to cimetidine it represents over 1000-fold improvement in being likely to block H_2 receptors in the brain. Zolantidine is a lipophilic molecule, having log P_{oct}=5.41, but the Δ log P parameter is relatively low (1.69). The 2-aminopyridine analogue **22** had an even higher brain/blood ratio=4.9, which is consistent with its lower Δ log P=1.06 but, unfortunately, it was much less active as an H_2 antagonist.

The physicochemical model which guided the synthesis of novel H_2 antagonists throughout this study represents an important development in the rational design of brain-penetrating compounds. Its validity, reassessed at the end of the study when the brain penetration data for 20 compounds had been gathered [24], confirmed that a highly significant relationship still applied:

$$\log (C_{brain}/C_{blood}) = -0.48\ \Delta \log P + 0.89$$
$$n = 20,\ r = 0.83,\ s = 0.44,\ F = 40.2 \qquad (8)$$

A few words of caution are required, however. The estimate *in vitro* of brain/blood concentration ratios is subject to many errors and has to be regarded as a gross approximation. Although compounds are administered intravenously by infusion over an extended period (2–3 h) so that near steady state equilibrium is produced, they are subject to different rates of metabolism, elimination (excretion) and protein binding. All these factors will affect the concentration of drug available for brain penetration and will operate differently for each compound. In each case a check was made to determine the proportion of parent drug in the peripheral blood and in the brain by thin-layer chromatography. In most cases, the parent compound was still present at 90%; for a few compounds, the proportion of parent present in the blood had decreased to 50%, so that the ratios for these could be underestimated by a factor of up to twofold with respect to metabolism.

A precise interpretation of the above apparent correlation is, however, fraught with complications. In the first place it must be noted that it would be surprising if $\Delta \log P$ were the only determinant for brain penetration. One would certainly expect that overall lipophilicity should make some contribution. Secondly, the compounds under consideration are relatively complex structures in regard to their physicochemical properties, being strongly dipolar and having several hydrogen bond acceptor and hydrogen bond donor sites.

Kamlet, Taft, Abraham and coworkers have shown that many properties that depend on solute-solvent interactions are well correlated by a generalized linear solvation energy relationship (LSER) using solvatochromatic parameters. Octanol-water partition of a solute, for example, is suggested to be strongly dependent on dipolarity-polarizability and hydrogen-bond acceptor basicity, and to a lesser extent on hydrogen bond donor acidity [28, 29]. It is to be hoped that analytical approaches using their parameters will eventually provide a helpful basis for correlating biological activity with such important structural properties as hydrogen bond donor acidity and hydrogen bond acceptor basicity.

In view of the structural variation encompassed by equation (8) it seems likely that $\Delta \log P$ may have more general utility and prove useful for structure-activity correlations and the design of other types of drugs where hydrogen bonding plays an important role in brain penetration or drug-receptor interactions. It remains to be seen whether such an approach will also prove to be useful for other forms of drug access, for example skin penetration or ocular absorption.

ACKNOWLEDGEMENTS

The work described here was performed at Smith Kline & French Research Laboratories, Welwyn, UK, and involved a collaborative effort between many people, especially the authors of the cited publications. Particular mention, however, should be made of the excellent scientific contributions from Dr R.C. Mitchell (physicochemical measurements of partition, pK_a and hydrogen-

bonding studies) and Dr R.C.Young (molecular orbital calculations, correlation analysis and synthesis of novel compounds).

REFERENCES

1 [] Ash, A.S.F. and Schild, H.O., (1966) *Br. J. Pharmacol. Chemother.,* **27,** 427.

2 [] Black, J.W., Duncan, W.A.M., Durant, G.J., Ganellin, C.R. and Parsons, M.E. (1972) *Nature (London),* **236,** 385.

3 [] Arrang, J.M., Garbarg, M. and Schwartz, J-C. (1983) *Nature (London),* 302, 832.

4 [] Arrang, J.M., Garbarg, M., Lancelot, J.C., Lecomte, J.M., Pollard, H., Robba, M., Schunack, W. and Schwartz, J-C. (1987) *Nature (London),* **327,** 117.

5 [] Brimblecombe, R.W., Duncan, W.A. M., Durant, G.J., Emmett, J.C., Ganellin, C.R. and Parsons, M.E. (1973) *J. Int. Med. Res.,* **3,** 86.

6 [] Durant, G.J., Emmett, J.C., Ganellin, C.R., Miles, P.D., Prain, H.D., Parsons, M.E. and White, G.R. (1977) *J. Med. Chem.,* **20,** 901.

7 [] Wyllie, J.H., Clark, C.G., Alexader-Williams, J., Bell, P.R.F., Kennedy, T.L., Kirk, R.M. and MacKay, C. (1981) *Lancet,* 1307.

8 [] Ganellin, R. (1981) *J. Med. Chem.,* **24,** 913.

9 [] Parsons, M.E., Owen, D.A.A., Ganellin, C.R. and Durant, G.J. (1977) *Agents Actions,* **7,** 31.

10 [] Leo, A., Hansch, C. and Elkins, D. (1971) *Chem. Rev.,* **71,** 525.

11 [] Young, R.C., Ganellin, C.R., Graham, M.J. and Grant, E.H. (1982) *Tetrahedron,* **38,** 1493.

12 [] Young, R.C., Ganellin, C.R., Graham, M.J., Roantree, M.L. and Grant, E.H. (1985) *Tetrahedron.,* **26,** 1897.

13 [] Young, R.C., Durant, G.J., Emmett, J.C., Ganellin, C.R., Graham, M. J., Mitchell, R.C., Prain, H.D. and Roantree, M.L. (1986) *J. Med. Chem.,* **29,** 44.

14 [] Buyniski, J.P., Cavanagh, R.L., Pircio, A.W., Algieri, A.A. and Crenshaw, R.R. (1984) in *Highlights of Receptor Chemistry* (eds Melchiorre, C. and Gianella, M.), Elsevier, Amsterdam, p. 195.

15 [] Gavey, C.J., Smith, J.T. L., Nwokolo, C.U. and Pounder, R.E. (1989) *Aliment. Pharmacol. Ther.,* **3,** 557.

16 [] Ganellin, C.R. and Young, R.C. (1986) in *Neuropharmacology and Pesticide Action* (eds Ford, M.G., Lunt, G.G., Reay, R.C. and Usherwood, P.N.R.), Ellis Horwood, Chichester, p. 86.

17 [] Donné-Op den Kelder, G.M., Haaksma, E.E.J. and Timmerman, H. (1988) *Quant. Struct. Act. Relat.,* **7,** 7.

18 [] Lien, E.J., Guo, Z-R., Li, R-L. and Su, C-T., (1982) *J. Pharm. Sci.,* **71,** 641.

19 [] Charifson, P.S., Bowen, J.P., Wyrick, S.D., Hoffman, A.J., Cory, M., McPhail, A.T. and Mailman, R.B. (1989) *J. Med. Chem.,* **32,** 2050.

20 [] Glave, W.C. and Hansch, C. (1972) *J. Pharm. Sci.,* **61,** 589.

21 [] Hansch, C., Bjorkroth, J.P. and Leo, A. (1987) *J Pharm. Sci.,* **76,** 663.

22 [] Blakemore, R.C., Brown, T.H., Durant, G.J., Ganellin, C.R., Parsons, M.E., Rasmussen, A.C. and Rawlings, D.A. (1981) *Br. J. Pharmacol,* **74,** 200P.

23 [] Brown, T.H., Armitage, M.S., Blakemore, R.C., Blurton, P., Durant, G.J., Ganellin, C.R., Ife, R.J. Parsons, M.E., Rawlings, D.A. and Slingsby, B.P. (1990) *Eur. J. Med Chem.,* **25,** 217.

24 [] Young, R.C., Mitchell, R.C., Brown, T.H., Ganellin, C.R., Griffiths, R., Jones, M., Rana, K.K., Saunders, D., Smith, I.R., Sore, N.E. and Wilks, T.J. (1988) *J. Med. Chem.,* **31,** 656.

25 [] Seiler, P. (1974) *Eur. J. Med. Chem.,* **9,** 473.

26 [] Shibata, T. (1984) 104th Annual Meeting of the Japanese Pharmaceutical Association (Sendai), Abstr. 29N, 1–25.

27 [] Calcutt, C.R., Ganellin, C.R., Griffiths, R., Leigh, B.K., Maguire, J.P., Mitchell, R.C., Mylek, M.E., Parsons, M.E., Smith, I.R. and Young, R.C. (1988) *Br. J. Pharmacol.,* **93,** 69.

28 [] Taft, R.W., Abraham, M.H., Famini, G.R., Doherty, R.M., Abboud, J.L.M. and Kamlet, M.J. (1985) *J. Pharm. Sci.,* **74,** 807.

29 [] Kamlet, M.J., Abraham, M.H., Carr, P.W., Doherty, R.M. and Taft, R.W. (1988) *J. Chem. Soc. Perkin Trans. II,* 2087.

3
Acetylcholine, GABA and excitatory amino acid receptors in neurodegenerative disorders

Povl Krogsgaard-Larsen and Erik Falch

PharmaBiotec Research Centre, Department of Organic Chemistry, Royal Danish School of Pharmacy, Copenhagen, Denmark

SUMMARY

Neurodegenerative disorders are characterized by progressive loss of certain types of neurones in the central nervous system (CNS). In Parkinson's disease and Huntington's chorea dopamine and γ-aminobutyric acid (GABA) neurones, respectively, are selectively degenerating, whereas there is a loss of glutamic acid (Glu) and, in particular, acetylcholine neurones in Alzheimer's disease (AD). Although the primary causes of these diseases are unknown, a variety of factors such as excitotoxicity, free radical processes and autoimmunity may play crucial roles in the processes causing irreversible neuronal damage. In recent years, excitotoxicity, caused by overexcitation of different populations of neurones by Glu neurones, has been the subject of intensive studies.

Since excitotoxicity as well as loss of Glu neurones appear to be important aetiological factors in AD, the development on a rational basis of drugs for the treatment of this mortal disease is a particularly complex challenge to medicinal chemists. Whereas subtype-selective Glu receptor antagonists are potential neuroprotective agents in AD, there may also be a need for Glu replacement therapies based on Glu receptor agonists in this disease. Partial Glu agonists may be of particular therapeutic interest in AD.

Whereas AMPA and Br-HIBO (2-amino-3-(3-hydroxy-4-bromoisoxazol-5-yl) propanoic acid) are subtype-specific Glu receptor agonists, AMOA and AMNH are subtype-selective Glu antagonists. The design and molecular pharmacology of these and a number of other subtype-specific Glu receptor ligands will be discussed.

GABA neurones exert inhibitory control on central acetylcholine neurones mediated by $GABA_A$ receptors. Thus, administration of $GABA_A$ agonists, such as THIP, are unlikely to improve the symptoms of AD patients and may actually aggravate their symptomatic profile. On the other hand, $GABA_A$ antagonists, which are convulsants, are unlikely to be useful memory-improving drugs in AD. Low-efficacy partial $GABA_A$ agonists showing a dominating $GABA_A$ antagonist profile are potentially non-convulsive memory-improving drugs in AD.

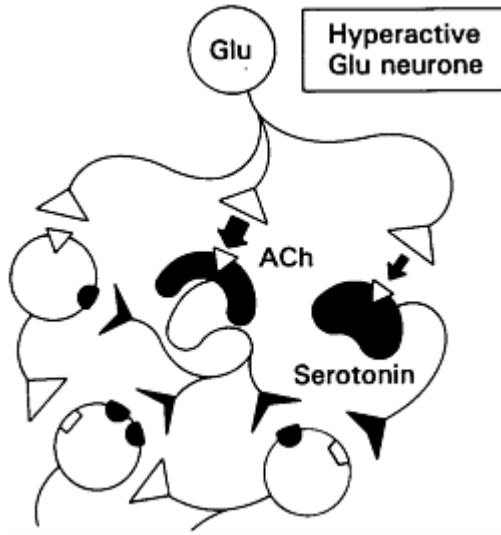


Fig. 1. Schematic illustration of the degeneration of acetylcholine (ACh) and serotonin neurones caused by hyperactive glutamic acid (Glu) neurones.

NEURODEGENERATIVE DISORDERS

Hyperactivity of central excitatory amino acid (EAA) neuronal pathways, where glutamic acid (Glu) is the major neurotransmitter, has been associated with the aetiology of certain neurodegenerative diseases, such as status epilepticus, Huntington's chorea and senile dementia of Alzheimer type (SDAT) [1, 2]. In Alzheimer patients a regional degeneration of neurones, notably cholinergic (see subsequent section) and serotonergic neurones (Fig. 1) is observed [3]. In addition, loss of glutamatergic neurones is seen in the progression of Alzheimer's disease (AD) [2] (Fig. 2). Thus, hyperactive as well as hypoactive EAA neuronal mechanisms may be operative in AD. The neuronal degeneration observed after ischaemia, including stroke, hypoxia and hypoglycaemia, may also be due to prolonged and excessive stimulation of EAA receptors [2]. There is some evidence that hypoactivity of the central EAA system(s) may be a key factor in schizophrenia [4].

Due to the accumulating evidence implicating hyperactivity as well as hypoactivity of the EAA neuronal systems in, for example, AD, drugs capable of protecting as well as activating EAA receptors may be of therapeutic interest. An obvious challenge is to design partial agonists with appropriately balanced agonist/antagonist profiles.

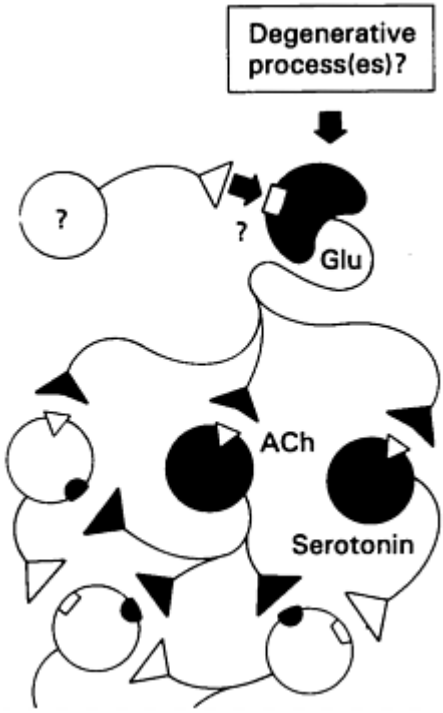


Fig. 2. Schematic illustration of the degeneration of glutamic acid (Glu) neurones resulting in hypostimulation of other neurones, notably acetylcholine (ACh) and serotonin neurones, which are normally receiving stimulatory inputs from Glu neurones. The mechanism(s) underlying this neurodegeneration are unknown.

CHOLINERGIC DYSFUNCTIONS AND SENILE DISORDERS

There is accumulating evidence of major impairments in the central cholinergic neurotransmission in patients with the pathology characteristic of AD and SDAT [5–7]. This cholinergic deficit may be of particular relevance to disturbances in learning and memory in AD/SDAT patients.

Neurochemical examination of biopsy and autopsy brain material from Alzheimer patients has revealed loss of the presynaptic marker enzymes acetyl coenzyme A (acetyl-CoA), choline *O*-acetyltransferase and acetylcholinesterase (AChE), and of muscarinic receptor sites of the M_2 subtype correlating with dementia score and severity of neurohistopathology [8]. These alterations do, to some extent, reflect the well-documented neuronal cell loss in the nucleus of Meynert in presenile dementia and AD. Lesions of this brain nucleus in rats, a limited model for the cholinergic deficit of AD, result in marked reductions in the levels of cholinergic enzyme markers in the neocortex accompanied by great decreases in the total muscarinic receptor levels [9, 10].

The M_2 muscarinic receptors appear to be located predominantly on cholinergic nerve terminals in different brain regions [11], and may function as autoreceptors regulating the release of acetylcholine (ACh). Accordingly, most neurochemical studies on brain material have demonstrated decreases in the density of M_2 receptor sites corresponding to the loss of presynaptic cholinergic enzyme markers. Postsynaptic muscarinic receptor sites which primarily are of the M_1 subtype do, however, to a large extent seem to survive the loss of cholinergic nerve terminals in different brain regions [5].

SYNAPTIC MECHANISMS AS THERAPEUTIC TARGETS IN ALZHEIMER'S DISEASE

In light of this well-documented loss of cholinergic nerve terminals in certain brain areas of AD/SDAT patients, notably in the cerebral cortex and the hippocampus, there is an obvious need for an ACh replacement therapy in these diseases. Ideally, such therapeutic approaches should be selectively targeted at cholinergic synapses in the brain areas containing the degenerating ACh systems.

In principle, cholinergic neurotransmission can be stimulated indirectly via the GABA system, which appears to exert inhibitory control of ACh neurones in different brain areas through $GABA_A$ receptors or at one of the modulatory sites of the $GABA_A$ receptor complex, $GABA_A$ antagonists should theoretically be applicable in AD/SDAT (see subsequent sections).

Most therapeutic interest in AD/SDAT has so far been focused on the processes and mechanisms at muscarinic cholinergic synapses [13]. The heterogeneity of muscarinic ACh receptors in the CNS may make it possible to identify a subtype of such receptors, which is of particular pharmacological relevance in AD/SDAT. On the basis of the neurochemical evidence so far available for muscarinic receptors in AD/SDAT patients and in relevant animal models, the postsynaptic M_1 receptors seem to be of primary therapeutic interest [5, 11, 14]. Partial agonists at M_1 receptors probably have less predisposition to cause receptor desensitization than full agonists, making partial agonists more interesting from a therapeutic point of view. Antagonists at presynaptic M_2 receptors, which may function as autoreceptors, might be useful drugs at the early stages of AD/SDAT [15]. Compounds with mixed M_1 agonist/M_2 antagonist profiles may prove to be of particular therapeutic interest [16].

Since hyperactivity of central Glu neurones may be one of the primary causes of the degeneration of ACh neurones in AD/SDAT, Glu receptor antagonists have interest as neuroprotective drugs in these diseases. So far, competitive [17] as well as non-competitive [18] antagonists at the N-methyl-D-aspartic acid (NMDA) subtype of Glu receptors (Fig. 3) have been shown to possess neuroprotective properties in animal models and cell culture systems (see later section). Studies along these lines may lead to the development of agents capable

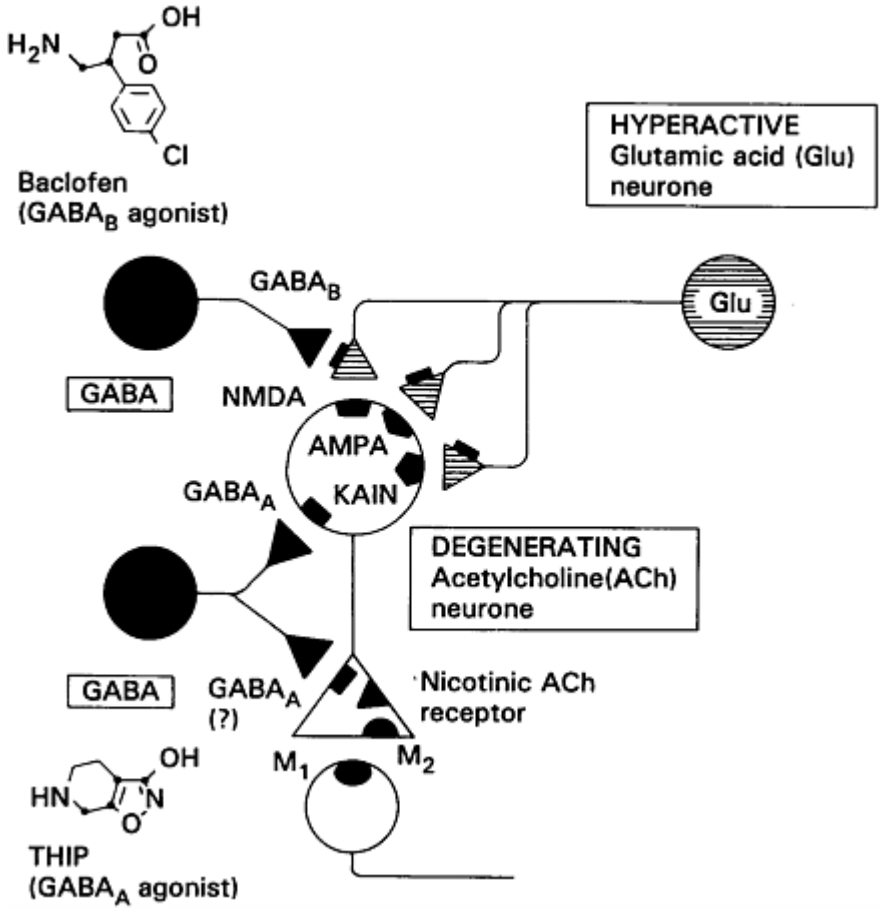


Fig. 3. Schematic illustration of synaptic contacts between central acetylcholine (ACh), glutamic acid (Glu) and γ-aminobutyric acid (GABA) neurones and their receptors (for details see subsequent sections). The structures of the specific $GABA_A$ agonists isoguvacine and (THIP) [72] and the low-efficacy partial $GABA_A$ agonist 5-(4-piperidyl) isoxazol-3-ol (4-PIOL) [30] are indicated.

of slowing down, or at least limiting, the severity of the neurodegenerative processes in AD/SDAT patients.

Normal function of central ACh neurones appears to be dependent on stimulation by Glu neurones (Figs 1 and 2), and therefore it may be difficult to prevent Glu receptor-mediated degeneration of ACh neurones by Glu antagonists without reducing or, perhaps, blocking the excitatory input from Glu terminals to ACh neurones. Consequently, future therapies in AD/SDAT based on Glu antagonists or partial agonists/antagonists probably have to be supplemented by concomitant treatment by M1 receptor agonists or M_2 antagonists in order to

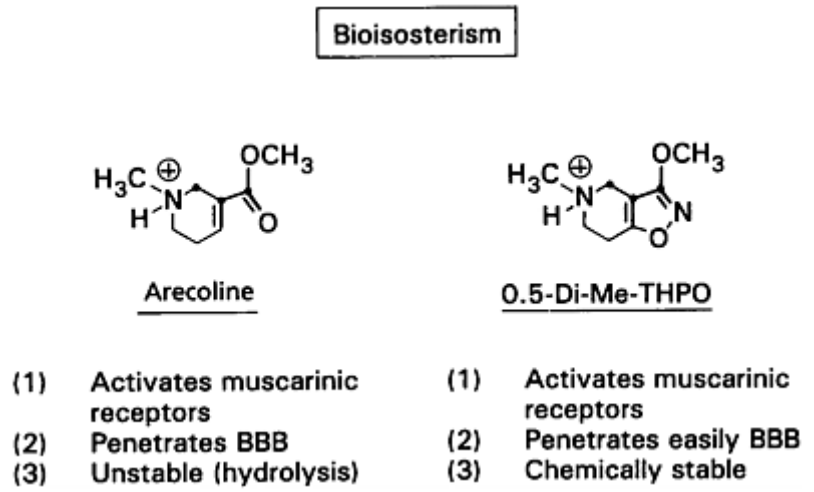


Fig. 4. Structures of arecoline and its 3-methoxyisoxazole bioisostere, O,5-di-Me-THPO, and a comparison of their effects and pharmacokinetic properties.

maintain the function of muscarinic synapses during the treatment with neuroprotective drugs. Thus, such muscarinic agents may be of major therapeutic interest, not only as drugs for symptomatic treatment of AD/SDAT patients but also as essential components of neuroprotective treatments of such patients.

ISOXAZOLE BIOISOSTERES OF THE MUSCARINIC AGONIST ARECOLINE

Bioisosteric replacements of the carboxyl group of GABA or the terminal carboxyl group of Glu analogues by the 3-isoxazolol group or structurally related heterocyclic units with protolytic properties almost identical with those of the carboxyl group have led to a number of specific and very potent GABA [19] or Glu agonists [20] (see subsequent sections). These findings prompted us to develop an ester bioisostere of arecoline containing the hydrolysis-resistant 3-methoxyisoxazole group, namely 3-methoxy-5-methyl-4,5,6,7,-tetrahydroisoxazolo[4.5-*c*]pyridine (*O*,5-di-Me-THPO) [21–23] (Fig. 4).

The *in vitro* pharmacological profiles of these and a number of structurally related bicyclic muscarinic agonists were evaluated on the basis of ligand receptor binding studies and studies in functional test systems. Quinuclidinyl benzilate (QNB) and pirenzepine (PZ) were used as non-selective M_2- and M_1-selective muscarinic antagonists, respectively. Oxotremorine-M (Oxo-M) was used as a muscarinic agonist ligand in order to estimate the muscarinic agonist character of the compounds. The ratio between the K_i values of a compound determined in QNB (brain) and Oxo-M (brain) binding experiments was used as a muscarinic agonist index (M-agonist index) of a compound. This method of

Compound	Structure	M2/M1-Index	M-agonist	Hydrolysis	p*K*a	log *P*	Penetration of BBB
Arecoline		0.07	890	Yes	7.8	−0.20	Yes
		0.3	204	No	6.6	0.40	Yes
		0.4	192	No	7.6	0.40	Yes
		0.2	140	No	–	–	No

Fig. 5. The relationship between structures, M_2/M_1- and M-agonist indices, pK_a values and pharmacokinetic properties of arecoline and a number of isoxazole bioisosteres.

estimating efficacy at muscarinic cholinergic receptors is analogous to that described recently [24]. M-agonist index values above 1500 reflect full agonism, whereas values of 20–200 and below predict partial agonism and antagonism, respectively, of muscarinic agents [24]:

$$\text{M-Agonist index} = \frac{K_i\ (\text{QNB, brain})}{K_i\ (\text{Oxo-M, brain})} = \frac{IC_{50}\ (\text{QNB, brain})}{IC_{50}\ (\text{Oxo-M, brain})} \times 0.162$$

The ratio between K_i values of a compound determined in QNB (heart) and PZ (brain) binding experiments was used as an index of M_1 selectivity (M_2/M_1 index), higher values of this index indicating higher degrees of M_1 selectivity (Fig. 5):

$$M_2/M_1 \text{ index} = \frac{K_i\ (\text{QNB, heart})}{K_i\ (\text{PZ, brain})} = \frac{IC_{50}\ (\text{QNB, heart})}{IC_{50}\ (\text{PZ, brain})} \times 0.125$$

Muscarinic agonist, partial agonist or antagonist effects estimated on the basis of these receptor binding studies were confirmed in experiments on isolated guineapig ileum. Effects on central muscarinic receptors and abilities of the compounds to penetrate the blood-brain barrier (BBB) were confirmed by the demonstration of central pharmacological (antinociceptive and anticonvulsant) effects after systemic administration to animals [21–23].

Arecoline and its isoxazole bioisostere, *O*,5-di-Me-THPO, show very similar *in vitro* profiles (Fig. 5). Like *O*,5-di-Me-THPO, *O*-propargyl-THPO shows higher M_1 selectivity than, for example, arecoline, as indicated by their higher M_2-M_1 indices.

The partial muscarinic agonists *O*,5-di-Me-THPO and *O*-propargyl-THPO (Fig. 5) have been tested pharmacologically in animal models. So far, vasodepressor, antinociceptive and anticonvulsant effects of arecoline and a series of isoxazole bioisosteres in anaesthetized rats gave parallel log dose-response curves, and the order of potency of the compounds in this test system was identical with that measured in the guinea-pig ileum preparation, arecoline being the most potent compound, followed by *O*,5-di-Me-THPO and *O*-propargyl-THPO. This order of potency was different from those of their antinociceptive and anticonvulsant effects. In these tests *O*-propargyl-THPO proved to be the most active compound.

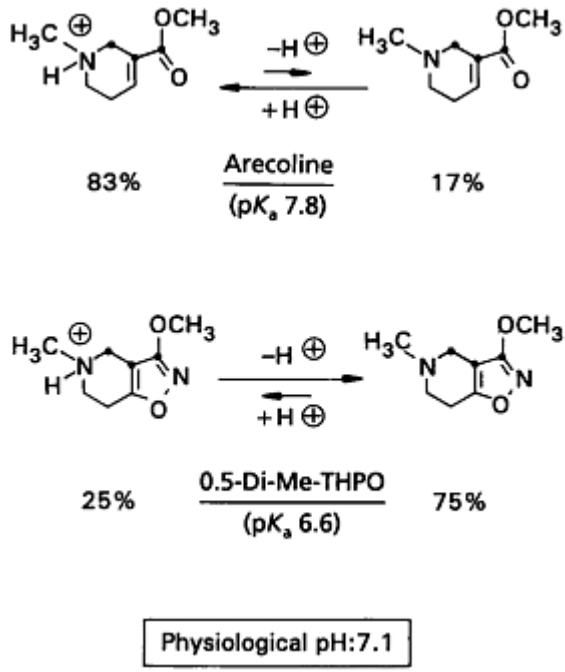


Fig. 6. Structures, pK_a values and degrees of ionization of arecoline and its isoxazole bioisostere, **O,5-di-Me-THPO**, at pH 7.1.

Arecoline is highly susceptible to hydrolysis *in vivo* and, furthermore, it seems to have a limited ability to penetrate the BBB in agreement with its low apparent partition coefficient *(P)* and relatively high pK_a value (Fig. 6). The positive log *P* values of *O*-propargyl-THPO (Fig. 5) indicate a higher degree of lipophilicity of these compounds and, thus, an increased ability to penetrate the BBB. The lower pK_a values of these bioisosteres, especially *O*,5-di-Me-THPO (Fig. 6), contribute to their improved bioavailability. Accordingly, these compounds show potent pharmacological effects after systemic administration in animals. Centrally mediated pharmacological effects could not be detected after systemic administration of the fully charged sulphonium arecoline bioisostere, *O*-propargyl-S-Me-DHTO, reflecting its inability to penetrate the BBB (Fig. 5).

Pharmacological studies in animal models relevant to memory and learning impairments are in progress. Furthermore, the design of new types of bioisosteres and structural modifications of the most active compounds so far synthesized are in rapid progress, with the object of developing muscarinic agents showing a pharmacological profile that is optimal for future treatment of AD/SDAT patients (Fig. 7).

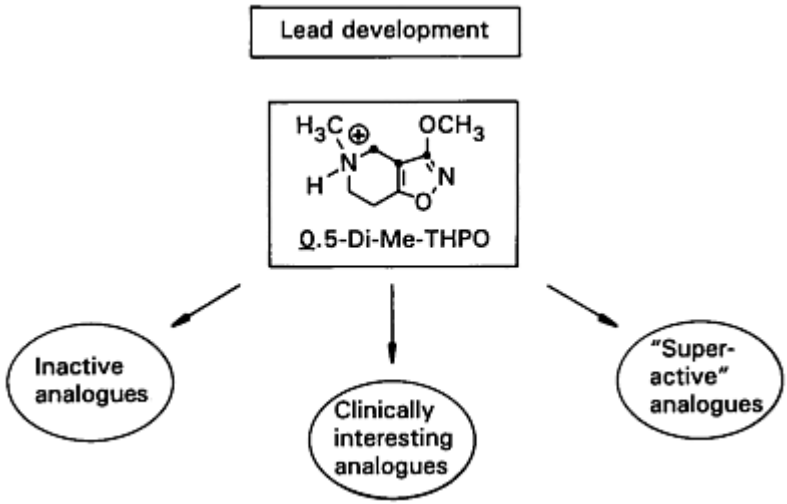


Fig. 7. Outline of the development in progress of heterocyclic muscarinic agonists.

$GABA_A$ RECEPTORS AS POTENTIAL THERAPEUTIC TARGETS IN ALZHEIMER'S DISEASE

As mentioned earlier, central cholinergic neurones appear to be under inhibitory GABAergic control and, consequently, the function of such neurones may be indirectly stimulated by blockade of the $GABA_A$ receptors involved in this regulation. These $GABA_A$ receptors may be located pre- or postsynaptically on ACh neurones (Fig. 3). Therapies based on agents with antagonist actions at $GABA_A$ receptors, or at one of the modulatory sites of the $GABA_A$ receptor antagonists, which in addition show low-efficacy $GABA_A$ agonist effects, might stimulate ACh release and, thus, improve learning and memory in Alzheimer patients without causing convulsions.

The results of studies on different $GABA_A$ receptor ligands in animal models relevant to learning and memory seem to support such GABAergic therapeutic approaches in AD. Thus, whereas administration of $GABA_A$ agonists impairs learning and memory in animals [25, 26] via modulation of cholinergic pathways, memory enhancement was observed after injection of the $GABA_A$ antagonist bicuculline methiodide. Similarly, agonists and inverse agonists at the benzodiazepine (BZD) site of the $GABA_A$ receptor complex impair and enhance, respectively, performance in learning and memory tasks [27]. Administration of THIP to Alzheimer patients failed to significantly improve cognitive performance [28].

In the subsequent section we describe the molecular and single-cell pharmacology of a novel low-efficacy partial $GABA_A$ agonist, 5-(4-piperidyl) isoxazol-3-ol (4-PIOL), showing dominating $GABA_A$ antagonist profile.

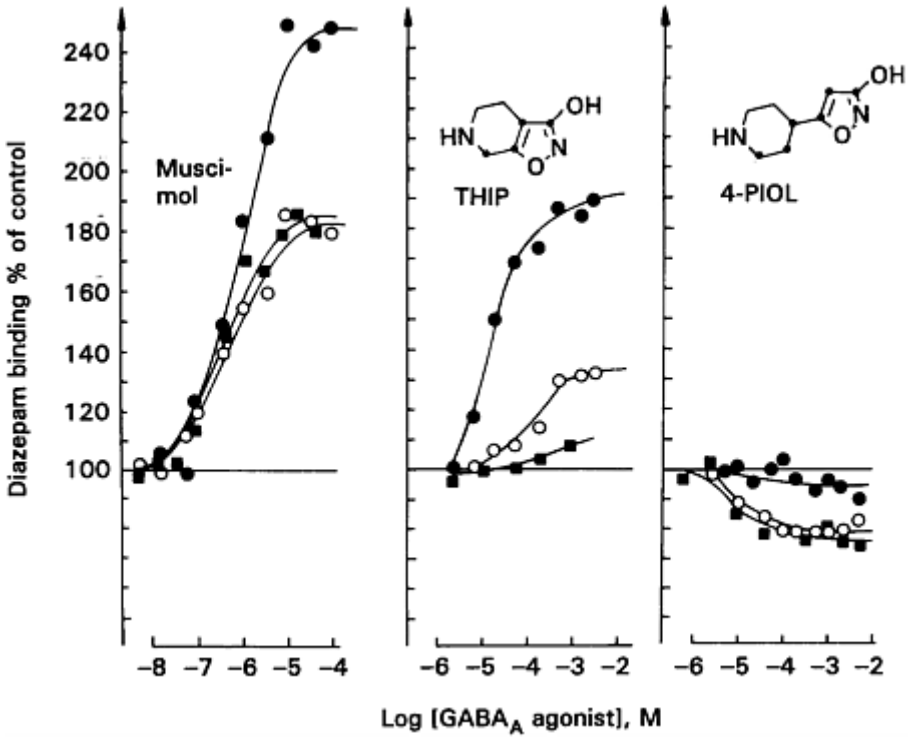


Fig. 8. Effects of the full $GABA_A$ agonist muscimol, the efficacious partial $GABA_A$ agonist THIP, and the low-efficacy partial $GABA_A$ agonist, 4-PIOL, on the binding of radioactive diazepam at 0°C and in the absence of chloride (■), or at 30°C in the presence (●) or absence (○) of 150mM sodium chloride.

4–PIOL, A LOW-EFFICACY PARTIAL $GABA_A$ AGONIST SHOWING DOMINATING $GABA_A$ ANTAGONIST EFFECTS

4-PIOL is a relatively weak agonist at $GABA_A$ receptors in the cat spinal cord [29]. The slow recovery following inhibitory responses to 4-PIOL compared with that of GABA probably reflects the fact that 4-PIOL does not inhibit GABA uptake. In spite of this $GABA_A$ agonist effect of 4-PIOL, it does not significantly affect the binding of diazepam to rat brain membranes under different experimental conditions (Fig. 8). The apparent deactivation of BZD binding by 4-PIOL seen at low temperature and in the absence of chloride ions probably reflects blockade of the stimulatory effects of the very low concentrations of residual GABA, which remains following preparation of the brain synaptic membranes. These effects of 4-PIOL actually are observed at surprisingly low concentrations as compared with the relatively weak effects of 4-PIOL as an inhibitor of $GABA_A$ receptor binding (IC_{50}=6μM).

The effects of 4-PIOL on muscimol-induced stimulation of diazepam binding are very similar to those of the $GABA_A$ antagonist bicuculline methochloride (BMC) [30] (Fig. 9). This unique pharmacological profile seems to suggest that 4-PIOL is acting as an agonist at spinal $GABA_A$ receptors, but as a $GABA_A$ antagonist in supraspinal tissue preparations.

Whole-cell patch-clamp recordings from cultured hippocampal neurones have been used to further characterize the action of 4-PIOL [31] (Fig. 10). The action of 4-PIOL was compared with that of the $GABA_A$ agonist isoguvacine.

The response to 4-PIOL was competitively antagonized by bicuculline methobromide (BMB) (not illustrated). 4-PIOL was about 200 times less potent as an agonist than isoguvacine. The maximum response to 4-PIOL was only a small fraction of that of submaximal concentrations of isoguvacine, and 4-PIOL antagonized the response to isoguvacine with a parallel shift to the right of the dose-response curve [31]. On the basis of these studies it is concluded that 4-PIOL is a low-efficacy partial $GABA_A$ agonist showing also a $GABA_A$ antagonist profile (Fig. 10), being about 30 times weaker than BMB as a $GABA_A$ antagonist.

Only unionized neutral amino acids are able to penetrate the BBB. At physiological pH the fraction which is unionized is small, and frequently negligible. The ratio between the concentrations of ionized (zwitterionic) and unionized molecules (I/U ratio, zwitterionic constant) is a function of the difference between the pK_a I and II values [19]. A large difference between the pK_a values is tantamount to high I/U ratios. Only very small concentrations of unionized compound will be present in the solution, and the ability to penetrate the BBB will be correspondingly low. Whereas GABA (I/U ratio=800000) does not penetrate the BBB to any significant extent, muscimol (I/U ratio=900) or THIP (I/U ratio=1000) are better suited for this purpose in view of their low I/U ratios. A relatively large difference between the pK_a values of 4-PIOL (5.15; 10.19) and, consequently, a high I/U ratio (30000) suggests that 4-PIOL would not penetrate the BBB and 4-PIOL actually does not show pharmacological effects after systemic administration.

EXCITATORY AMINO ACID RECEPTOR SUBTYPES AS THERAPEUTIC TARGETS

Until a few years ago, EAAs were thought to mediate their actions through three different classes of receptors [32, 33]. As the result of extensive neurochemical, pharmacological and, in recent years, molecular biological studies, central EAA receptors are now most conveniently subdivided into five main classes, some of which, if not all, are heterogeneous (Fig. 11) [34–36]:

(1) NMDA receptors at which NMDA and quinolinic acid (QUIN) are selective agonists. NMDA receptors are competitively and very effectively blocked by a number of phosphono amino acids, notably (*R*-2-amino-5-

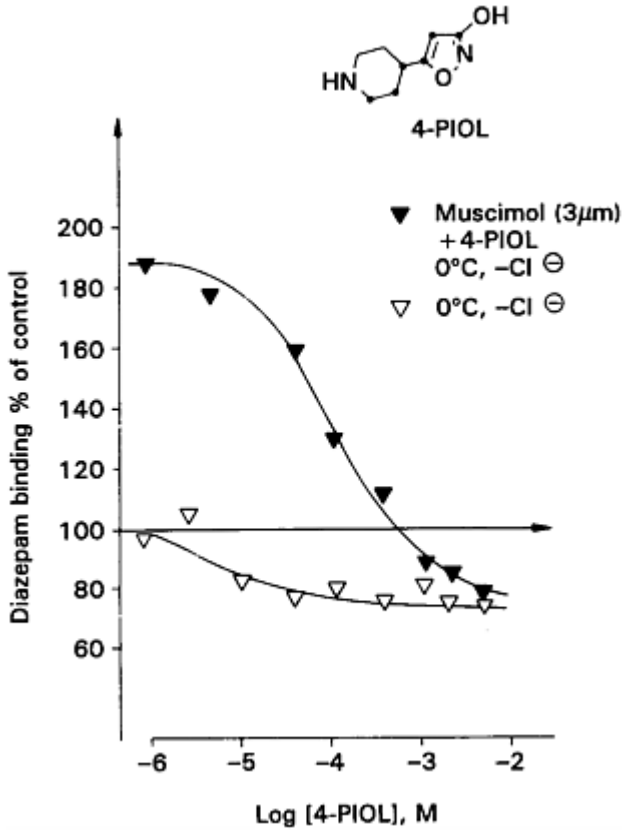


Fig. 9. Effect of the low-efficacy partial $GABA_A$ agonist 4-PIOL on the binding of radioactive diazepam at 0°C in the absence of chloride , or in the presence of a fixed concentration of the $GABA_A$ agonist muscimol (3 μM) at 0°C in the absence of chloride (▼).

phosphonopentanoic acid (D-AP5) (Fig. 12) [35, 37]. Evidence derived from electrophysio-logical and receptor binding studies strongly suggests heterogeneity of NMDA receptors [38, 39] (Fig. 11). Thus, NMDA receptors in the brain and the spinal cord show different agonist sensitivities [1]. Furthermore, NMDA receptors in different brain regions show dissimilar pharmacological profiles, consistent with the view that several isoforms of NMDA receptors may exist in the mammalian CNS [40, 41].

(2) AMPA receptors at which quisqualic acid (QUIS) is a non-selective and *(RS)*-2-amino-3-(3-hydroxy-5-methylisoxazol-4-yl)propionic acid (AMPA) a highly selective agonist [20, 42]. A number of quinoxalinediones are very potent but non-selective antagonists at AMPA receptors [43, 44]. More recently *(RS)*-2-amino-3-[3(carboxymethoxy)-5-methylisoxazol-4-yl] proprionic acid (AMOA) has been shown to antagonize AMPA-induced

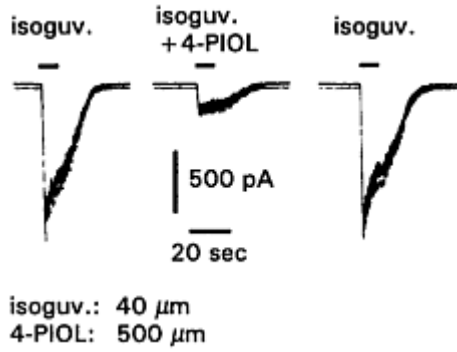


Fig. 10. Whole-cell patch-clamp recording from a hippocampal neurone. Holding potential was −60 mV with −10 mV command potentials superimposed to monitor membrane conductance. Drugs were applied in the vicinity of the neurone by a multi-barrel perfusion pipette. The response to isoguvacine was reduced by simultaneous application of 4-PIOL to a value slightly higher than the intrinsic agonist response to 4-PIOL alone. In contrast to the response to 20 μM isoguvacine alone the responses with 4-PIOL present did not show desensitization However, the responses to 10 μM isoguvacine, comparable in strength to the 1 mM 4-PIOL response, did not show desensitization either.

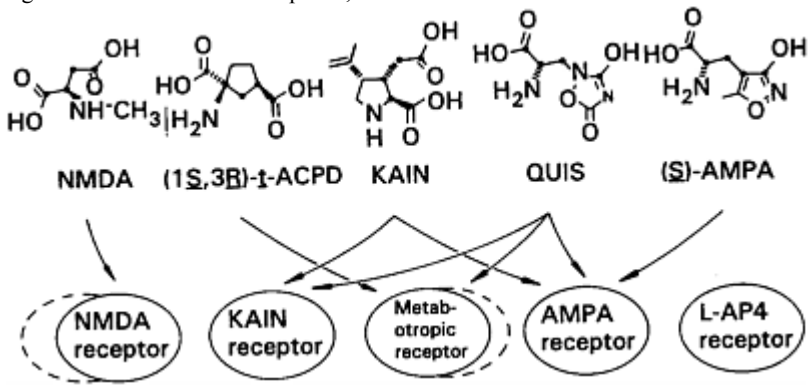


Fig. 11. Schematic illustration of the multiplicity of central EAA receptors and the sites of action of a number of selective or non-selective agonists.

excitation, showing much less effect on excitations by QUIS or kainic acid (KAIN) [45] (see Figs 14 and 15).

(3) KAIN receptors, which are selectively activated by KAIN [46]. A number of other naturally occurring amino acids show very potent KAIN agonist actions [47]. The 3-isoxazolol amino acid *(RS)*-2-amino-3-[2-(3-hydroxyl-5-methylisoxazol-4-yl)methyl-5-methyl-3-oxoisoxazolin-4-yl]propionic acid (AMNH), which is structurally related to AMOA, has recently been shown to be a selective but relatively weak KAIN receptor antagonist [45].

(4) Recent investigations by a number of research groups have disclosed that, in addition to ion channel-linked receptors mediating fast depolarizing

responses in the CNS, an EAA receptor named the metabotropic (Q_p) receptor exists that is coupled to inositol 1,4,5-triphosphate (IP_3) and diacylglycerate turnover [34, 48, 49]. Whereas QUIS is a non-selective agonist at this receptor, the *trans*-form of 1-aminocyclopentane-1,3-dicarboxylate (*t*-ACPD), notably the 1*S*, 3*R* enantiomer (Fig. 11) appears to be a selective agonist at the metabotropic receptor [50, 51]. There is some evidence to suggest that two metabotropic receptors exist [52].

(5) L-AP4 receptors, through which *(S)*-2-amino-4-phosphonobutanoic acid (L-AP4) inhibits synaptic excitation [53].

THE NMDA RECEPTOR COMPLEX

The NMDA receptor is a receptor complex coupled to ion channel(s) which, in contrast to those associated with AMPA or KAIN receptors, show a voltage-dependent blockade by magnesium [1, 37, 54, 55] (Fig. 12). Following membrane depolarization, for example via activation of AMPA or KAIN receptors in the cell membrane, the magnesium blockade is released, causing further depolarization. Thus, the NMDA receptor appears to function as an amplification system capable of generating long-lasting physiological changes, and it has been associated with the phenomenon of long-term potentiation (LTP), which is a sustained increase in synaptic sensitivity following high-frequency stimulation [56]. LTP can last for weeks and may play a key role in memory and learning processes. This component of synaptic plasticity mediated by NMDA receptors may, however, also be involved in the processes causing neuronal damage and epileptiform activity. These aspects have brought the NMDA receptor complex into focus as a potential target for therapeutic attack in certain neurological diseases, including AD [34].

The NMDA receptor complex also contains distinct binding sites for zinc ions, for polyamines and for glycine, which are closely associated with the ion channel of this complex [57]. The physiological roles of these additional sites are unknown, but activation of the glycine site, which is distinctly different from the strychninesensitive glycine receptor, strongly potentiates the responses to NMDA via allosteric interaction with the NMDA receptor site [57, 58].

The dissociative anaesthetics, such as ketamine and phencyclidine (PCP), selectively inhibit NMDA-associated excitatory effects without affecting excitation induced by agonists at AMPA, KAIN or metabotropic receptors [57, 59]. The binding site for the polycyclic compound MK-801 is identical to or closely associated with that for PCP. In agreement with the findings for the blockade by magnesium ions of the NMDA responses, PCP, MK-801 and other dissociative anaesthetics block NMDA responses in a voltage-dependent manner, indicating a close association between the binding sites concerned and the ion channel (Fig. 12) [18]. Furthermore, the blockade of the binding site(s) for PCP and MK-801 is use-dependent, in that a prerequisite for these antagonist

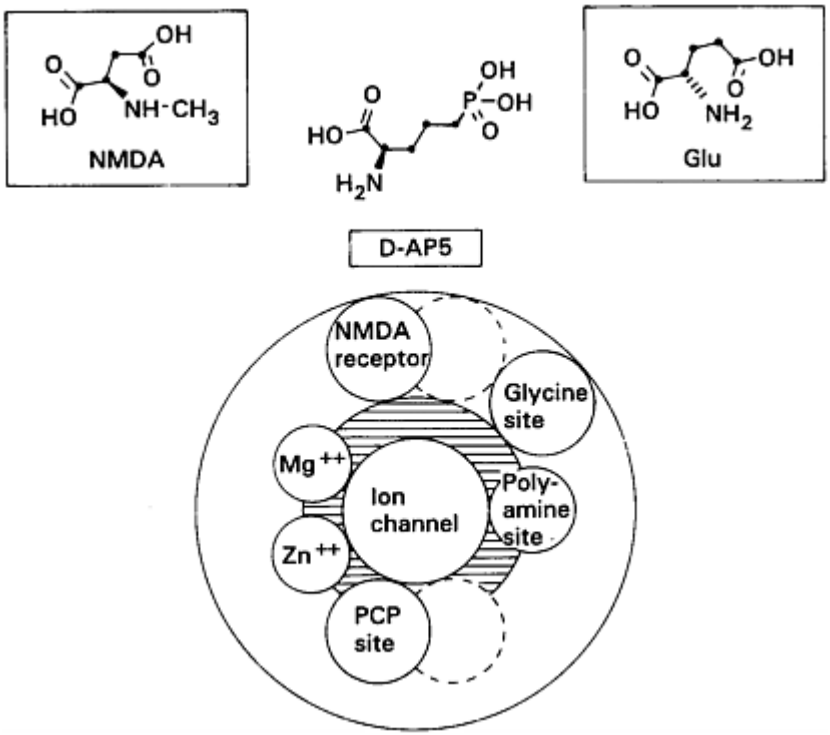


Fig. 12. Schematic illustration of the NMDA receptor complex and the structures of Glu, NMDA and the NMDA antagonist D-AP5.

effects is the presence of an NMDA receptor agonist, suggesting that opening of the ion channel uncovers these binding sites [18].

In principle, compounds like MK-801 would be almost ideal drugs for treatment of neurodegenerative diseases caused by hyperactivation of NMDA receptors (Fig. 3). The use dependence of the antagonist effects of such compounds suggests that, for example, MK-801 would show limited effect on EAA neurotransmission under normal conditions. Excessive release of Glu or other endogenous EAAs and, consequently, overstimulation of NMDA receptors would, however, be expected to disclose the binding site(s) for these non-competitive antagonists in the fully opened ion channels. Unfortunately, psychomimetic side-effects of this class of pharmacological agents may severely restrict their use in the human clinic [37].

AMPA AND KAINIC ACID RECEPTORS: COMPLEX INTERACTIONS

Receptor binding studies have revealed the existence of both high- and low-affinity AMPA and KAIN receptor sites [35, 44]. The equilibrium between the two stages of AMPA receptors is facilitated by chaotropic isothiocyanate ions. Calcium ions, on the other hand, somehow block high-affinity KAIN receptor sites [45, 60], but the underlying mechanism is not clear. The high-affinity sites of AMPA and KAIN receptors in the mammalian CNS show different ligand binding characteristics [20]. There is evidence to suggest that low-affinity stages of these receptors are functionally important and the AMPA receptors mediate at least some of the excitatory effects of KAIN [61].

The nature and degree of interactions between AMPA and KAIN receptors in different parts of the CNS, as detected under different experimental conditions, are still far from being fully elucidated. Based on the pieces of information so far available, it may be postulated that non-NMDA receptors are hetero-oligomeric complexes comprising subunits with preferential affinities for the EAAs, AMPA, KAIN or QUIS.

AMPA RECEPTOR AGONISTS AND PARTIAL AGONISTS

AMPA is a specific and very potent agonist at AMPA receptors, and radiolabelled AMPA has become the standard ligand for studies of this subtype of EAA receptors [62, 63].

Using AMPA as a lead structure a variety of structurally related 3-isoxazolol amino acids have been designed and tested *in vivo* and *in vitro* [36]. The *tert*-butyl analogue of AMPA, ATPA (Fig. 13), is an agonist at AMPA receptors somewhat weaker than AMPA [64]. Interestingly, replacement of the bulky and spherical *tert*-butyl group of ATPA by the planar phenyl group affords a compound, APPA, which shows markedly lower potency and efficacy than ATPA as an AMPA receptor agonist (Fig. 13). APPA actually is the first example of a partial agonist at AMPA receptors [65].

AMPA RECEPTOR ANTAGONISTS

The specificity of *(S)*-AMPA as an agonist at AMPA receptors is assumed to be associated with the rigid and planar structure of the acidic 3-isoxazolol unit of this compound (Fig. 14) as established by X-ray crystallography [66–69]. This heterocyclic nucleus is a bioisostere of the terminal carboxyl group of Glu, and the structural characteristics of AMPA are only to a limited extent shared by the structurally related heterocyclic amino acid QUIS [70]. The conformational flexibility and lack of coplanarity of the heterocyclic ring of QUIS (Fig. 14) may be of some importance for the ability of this amino acid to interact not only with

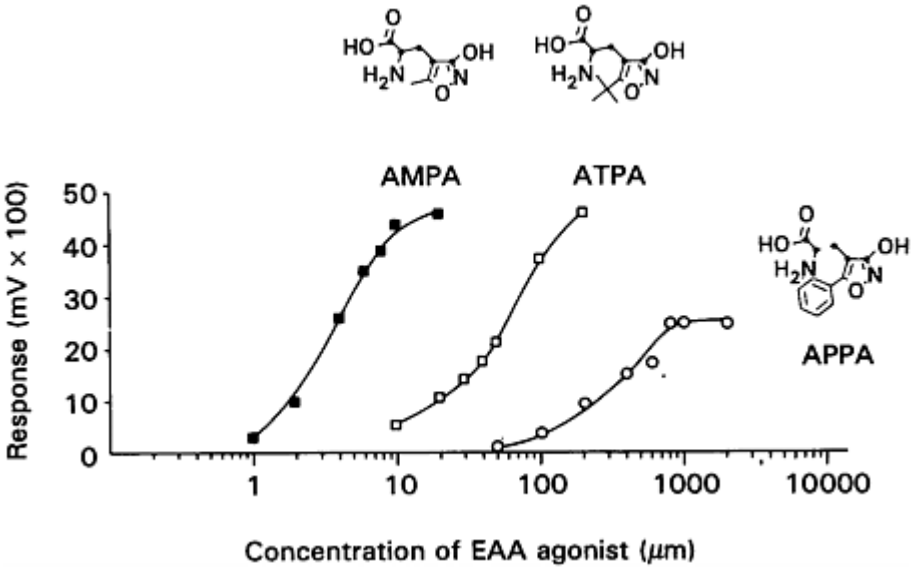


Fig. 13. Excitatory responses (normalized) to AMPA, (RS)-2-amino-3-(3-hydroxy-propionic acid (ATPA) and (RS)-2-amino-3-(3-hydroxy- propionic acid (APPA) in the rat cortical slice preparation. Data from Christensen *et al*. [65].

AMPA receptors, but also with KAIN and metabotropic receptors (Fig. 11) and with EAA-associated enzyme and transport sites [34–36].

Using AMPA as a lead structure in attempts to design antagonists selective for non-NMDA receptors, we have synthesized and tested the two compounds AMOA (Fig. 14) and AMNH as potential EAA antagonists [45].

AMOA and AMNH were shown to be stable in aqueous solution at pH values close physiological pH. Neither AMOA nor AMNH showed detectable affinities for the receptor, ion channel or modulatory sites of the NMDA receptor complex. Quantitative receptor autoradiographic and conventional binding techniques were used to study the affinities of AMOA and AMNH for non-NMDA receptor sites. Both compounds were inhibitors of the binding of [^{3}H] AMPA.

In the rat cortical slice preparation, AMOA was shown to antagonize, competitively, excitation induced by AMPA (Fig. 15) with some selectivity, whereas AMNH proved to be a weak but rather selective antagonist of KAIN-induced excitation (not illustrated). AMOA was a weak antagonist of excitation by QUIS, whereas AMNH did not affect excitation by this non-selective AMPA receptor agonist. On cat spinal neurones, both AMOA and AMNH reduced AMPA- and KAIN-induced excitations but, again, the excitatory effects of QUIS were much less sensitive [45].

AMNH and, in particular, AMOA effectively protected rat striatal neurones against the neurotoxic effects of KAIN, whereas the toxic effects of AMPA were only slightly reduced by AMOA (Fig. 15). Neither antagonist showed protection

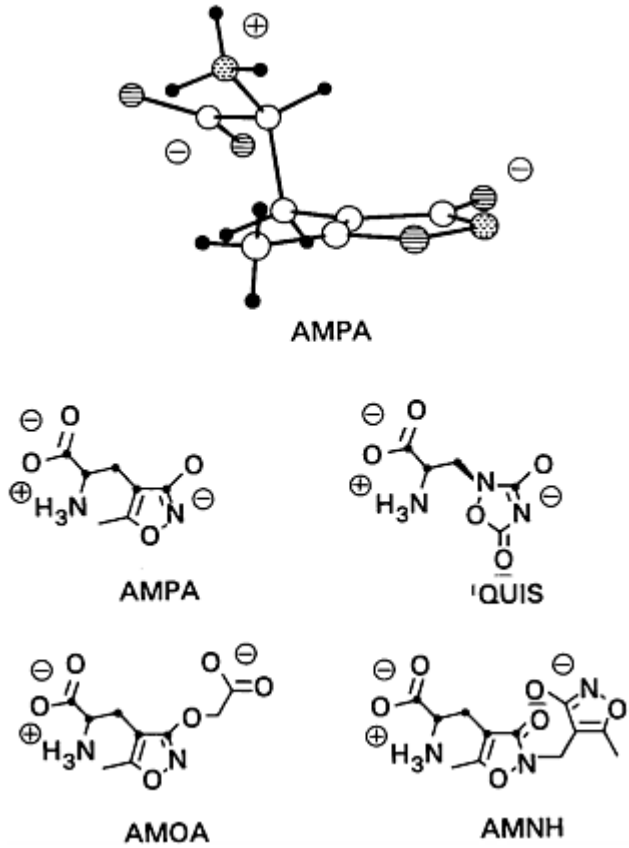


Fig. 14. Structure of *(S)*-AMPA, as established on the basis of an X-ray analysis, and of QUIS and the non-NMDA antagonists AMOA and AMNH.

against the cell damage caused by intrastriatal injection of the NMDA agonist QUIN [45]. In cultured cortical neurones, AMOA effectively protected against KAIN-induced toxicity, whereas no significant protective effect of AMOA could be demonstrated, when the cells were exposed to AMPA or NMDA [71] (Fig. 15).

These comparative studies on AMOA and AMNH as non-NMDA receptor antagonists and as neuroprotective agents strongly suggest that the mechanisms underlying neuroexcitation and neurotoxicity are not identical. Thus, although AMOA is a selective AMPA receptor antagonist, the neurotoxic effects induced by KAIN are much more effectively blocked by AMOA than those induced by AMPA. These observations suggest that it may be possible to develop therapeutic agents that block EAA-induced toxicity but not EAA excitatory mechanisms.

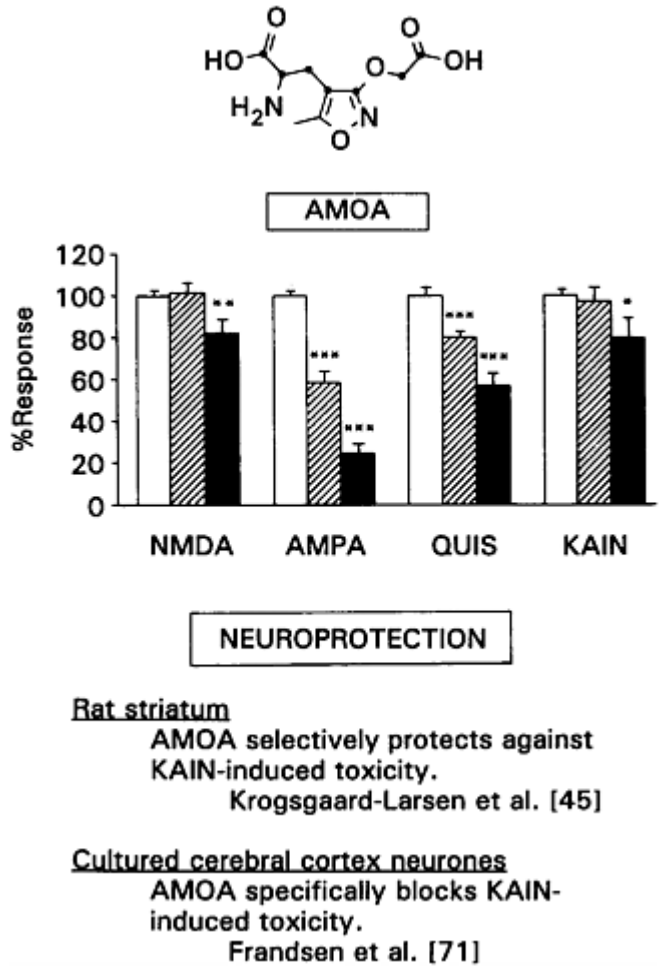


Fig. 15. Effects of AMOA on the neuronal excitations induced by NMDA, AMPA, QUIS or KAIN in the rat cortical slice preparation. Excitatory effects of NMDA (10 μM), AMPA (5 μM), QUIS (10 μM), or KAIN (5 μM) in the absence of antagonist. The effects of the same concentrations of agonists in the presence of AMOA (250 μM) or 1 mM ■. Percentage response (normalized) s.e.m. Data from Krogsgaard-Larsen *et al*. [45] and Frandsen *et al*. [71].

ACKNOWLEDGEMENTS

This work was supported by grants from the Danish Medical and Technical Research Councils and from the Lundbeck Foundation. The secretarial assistance of Mrs Birte Hare is gratefully acknowledged.

REFERENCES

1 [] Lodge, D. (ed.) (1988) in *Excitatory Amino Acids in Health and Disease,* Wiley, New York.

2 [] Greenamyre, J.T. (1988) *Arch. Neurol.,* **43,** 1058.

3 [] Narang, P.K. and Cutler, N.R. (1986) *Prog. Neuro-Psychopharmacol. Biol. Psychiatry,* **10,** 519.

4 [] Deutsch, S.I., Mastropaolo, J., Schwartz, B.L., Rosse, R.B. and Morihisa, J.M. (1989) *Clin. Neuropharmacol.,* **12,** 1.

5 [] Price, D.L., Whitehouse, P.J. and Struble, R.G. A. (1985) *Annu. Rev. Med.,* **36,** 349.

6 [] Perry, E. (1988) *Br. J. Psychiatry,* **152,** 737.

7 [] Smith, C.J., Perry, E.K., Perry, R.H., Candy, J.M., Johnson, M., Bonham, J.R., Dick, D.J., Fairbairn, A., Blessed, G. and Birdsall, N.J. M. (1988) *J. Neurochem.,* **50,** 847.

8 [] Palmer, A.M., Francis, P.T., Benton, J.S., Sims, N.R., Nann, D.M. A., Neary, D., Snowden, J.S. and Bowen D.M. (1987) *J. Neurochem.,* **48,** 8.

9 [] Whitehouse, P.J., Price, D.L., Struble, R.G., Clarke, A.W., Coyle, J.T. and DeLong, M.R. (1982) *Science,* **215,** 1237.

10 [] Watson, M., Vickroy, T.W., Fibiger, H.C., Roeske, W.R. and Yamamura, H.I. (1985) *Brain Res.,* **346,** 387.

11 [] Pomara, N., Bagne, C.A., Stanley, M. and Yarbrough, G.G. (1986) *Prog. Neuro-Psychopharmacol. Bio.. Psychiatry,* **10,** 553.

12 [] Kendall, D.A., Browner, M. and Emma, S.J. (1982) *Pharmacol. Exp. Ther.,* **220,** 482.

13 [] Hershenson, F.M. and Moos, W.H. (1986) *J. Med. Chem.,* **29,** 1125.

14 [] Mitchelson, F. (1988) *Pharmacol. Ther.,* **27,** 357.

15 [] Mash, D.C., Flynn, D.D. and Potter, L.T. (1985) *Science,* **228,** 1115.

16 [] Whitehouse, P.J. and Au, K.S. (1986) *Prog. Neuro-Psychopharmacol. Biol. Psychiatry,* **10,** 665.

17 [] Watkins, J.C. and Olverman, H.J. (1987) *Trends Neurosci.,* **10,** 265.

18 [] Kemp, J.A., Foster, A.C. and Wong, E.H.F. (1987) *Trends Neurosci.,* **10,** 294.

19 [] Krogsgaard-Larsen, P., Hjeds, H., Falch, E., Jorgensen, F.S. and Nielsen, L. (1988) *Adv. Drug Res.,* **17,** 281.

20 [] Krogsgaard-Larsen, P., Honore, T., Hansen, J.J., Curtis, D.R. and Lodge, D. (1980) *Nature,* **284,** 64.

21 [] Sauerbert, P., Larsen, J.J., Falch, E. and Krogsgaard-Larsen, P. (1986) *J. Med. Chem.,* **29,** 1004.

22 [] Sauerbert, P., Fjalland, B., Larsen, J.J., Bach-Lauritsen, T., Falch, E. and Krogsgaard-Larsen, P. (1986) *Eur. J. Pharmacol.,* **130,** 125.

23 [] Krogsgaard-Larsen, P., Falch, E., Sauerberg, P., Freedman, S.B., Lambol, H.L. and Meier, E. (1988) *Trends Pharmacol. Sci.,* Suppl. 69.

24 [] Freedman, S.B., Jarley, E.A. and Iversen, L.L. (1988) *Br. J. Pharmacol.,* **23,** 437.

25 [] Brioni, J.D., Nagahara, A.H. and McGaugh, J.L. (1989) *Brain Res.,* **487,** 105.

26 [] Brioni, J.D., Decker, M.W., Gamboa, L.P., Izquierdo, I. and McGaugh, J.L. (1990) *Brain Res.,* **552,** 227.

27 [] Venault, P., Chapouthier, G., Prado de Carvalho, L., Simiand, J., Morre, M., Dodd, R.H. and Rossier, J. (1986) *Nature (London),* **321,** 864.
28 [] Bruno, G., Forster, N.L., Fedio, P., Mohr, E., Cox, C., Gillespie, M.M. and Chase, T.N. (1984) *Neurology*, Suppl. **34,** 225.
29 [] Byberg, J.R., Labouta, I.M. Falch, E., Hjeds, H., Krogsgaard-Larsen, P., Curtis, D.R. and Gynther, B.D. (1987) *Drug Des. Del.,* **1,** 261.
30 [] Falch, E., Larsson, O.M., Schousboe, A. and Krogsgaard-Larsen, P. (1990) *Drug Dev. Res.,* **21,** 169.
31 [] Kristiansen, U., Lambert, J.D.C., Falch, E. and Krogsgaard-Larsen, P. (1991) *Br. J. Pharmacol.,* **104,** 85.
32 [] Watkins, J.C. and Evans, R.H. (1981) *Annu. Rev. Pharmacol. Toxicol.,* **21,** 165.
33 [] McLennan, H. (1983) *Prog. Neurobiol.,* **20,** 251.
34 [] Monaghan, D.T., Bridges, R.J. and Cotman, C.W. (1989) *Annu. Rev. Pharmacol. Toxicol.,* **29,** 365.
35 [] Watkins, J.C., Krogsgaard-Larsen, P. and Honore, T. (1990) *Trends Pharmacol. Sci.,* **11,** 25.
36 [] Hansen, J.J. and Krogsgaard-Larsen, P. (1990) *Med. Res. Rev.,* **10,** 55.
37 [] Watkins, J.C. and Collingridge, G.L. (eds) (1989) *The NMDA Receptor,* Oxford University Press, Oxford.
38 [] Monaghan, D.T., Olverman, H.J., Nguyen, L., Watkins, J.C. and Cotman, C.W. (1988) *Proc. Natl Acad. Sci. USA,* **85,** 9836.
39 [] Young, A.B. and Fagg, G.E. (1990) *Trends Pharmacol. Sci.* **11,** 126.
40 [] Stone, T.W. and Burton, N.R. (1988) *Prog. Neurobiol.* **30,** 333.
41 [] Monaghan, D.T. and Cotman, C.W. (1989) in *The NMDA Receptor* (eds Watkins, J.C. and Collingridge, G.L.), Oxford University Press, Oxford, p. 53.
42 [] Krogsgaard-Larsen, P., Johnston, G.A.R., Lodge, D. and Curtis, D.R. (1977) *Nature (London),* **268,** 53.
43 [] Honore, T., Davies, S.N., Drejer, J., Fletcher, E.J., Jacobsen, P., Lodge, D. and Neilsen, F.E. (1988) *Science,* **241,** 701.
44 [] Honore, T., Drejer, J., Nielsen, E.O. and Nielsen, M. (1989) *Biochem. Pharmacol.,* **38,** 3207.
45 [] Krogsgaard-Larsen, P., Ferkany, J.W., Nielsen, E.O., Madsen, U., Ebert, B., Johansen, J.S., Diemer, N.H., Bruhn, T., Beattie, D.T. and Curtis, D.R. (1991) *J. Med. Chem.,* **34,** 123.
46 [] Coyle, J.T. (1983) *J. Neurochem.,* **41,** 1.
47 [] Shinozaki, H. (1988) *Prog. Neurobiol.,* **30,** 399.
48 [] Sladeczek, F., Pin, J.-P., Recasemüns, M., Bockaert, J. and Weiss, S. (1985) *Nature,* **317,** 717.
49 [] Schoepp, D.D. and Johnson, B.G. (1989) *J. Neurochem.,* **53,** 273.
50 [] Watson, G.B., Monaghan, D.T. and Lanthorn, T.H. (1990) *Eur. J. Pharmacol.,* **179,** 479.
51 [] Schoepp, D.D., Bockaert, J. and Sladeczek, F. (1990) *Trends Pharmacol. Sci.,* **11,** 508.
52 [] Manzoni, O.J.J., Finiels-Marlier, F., Sassetti, I., Bockaert, J., le Peuch, C. and Sladeczek, F.A.J. (1990) *Neurosci. Lett.,* **109,** 146.
53 [] Johnson, R.L. and Koerner, J.F. (1988) *J. Med. Chem.,* **31,** 2057.
54 [] Barnard, E.A. and Costa, E. (eds) (1989) in *Allosteric Modulation of Amino Acid Recaptors: Therapeutic Implications,* Raven Press, New York.

55 [] Reynolds, I.J. (1990) *Life Sci.,* **47,** 1785.
56 [] Collingridge, G.L. and Singer, W. (1990) *Trends Pharmacol Sci.,* **11,** 290.
57 [] Lodge, D. and Johnson, K.M. (1990) *Trends Pharmacol Sci.,* **11,** 81.
58 [] Huettner, J.E. (1991) *Biochem. Pharmacol.,* **41,** 9.
59 [] Lodge, D. and Johnson, K.M. (1990) *Trends Pharmacol. Sci.,* **11,** 81.
60 [] Braitman, D.J. and Coyle, J.T. (1987) *Neuropharmacology,* **26,** 1247.
61 [] Honore, T. (1989) *Med. Res. Rev.,* **9,** 1.
62 [] Honore, T., Lauridsen, J. and Krogsgaard-Larsen, P. (1982) *J. Neurochem.,* **38,** 173.
63 [] Murphy, D.E., Snowhill, E.W. and Williams, M. (1987) *Neurochem. Res.,* **12,** 775.
64 [] Lauridsen, J., Honore, T. and Krogsgaard-Larsen, P. (1985) *J. Med. Chem.,* **28,** 668.
65 [] Christensen, I.T., Reinhardt, A., Nielsen, B., Ebert, B., Madsen, U., Nielsen, E.O., Brehm, L. and Krogsgaard-Larsen, P. (1989) *Drug Des. Del.,* **5,** 57.
66 [] Krogsgaard-Larsen, P., Brehm, L., Johansen, J.S., Vinzents, P., Lauridsen, J. and Curtis, D.R. (1985) *J. Med. Chem.,* **28,** 673.
67 [] Hansen, J.J., Nielsen, B., Krogsgaard-Larsen, P., Brehm, L., Nielsen, E.O. and Curtis, D.R. (1989) *J. Med. Chem.,* **32,** 2254.
68 [] Brehm, L., Jorgensen, F.S., Hansen, J.J. and Krogsgaard-Larsen, P. (1988) *Drug News Perspect.,* **1,** 138.
69 [] Brehm, L., Christensen, I.T., Nielsen, B., Reinhardt, A., Ebert, B., Madsen, U., Jorgensen, F.S., Hansen, J.J. and Krogsgaard-Larsen, P. (1990) in *Frontiers in Drug Research: Crystallographic and Computational Methods* (eds Jensen, B., Jorgensen, F.S. and Koford, H.), Munksgaard, Copenhagen, p. 92.
70 [] Jackson, D.E., Bycroft, B.W. and King, T.J. (1989) *J. Comput.—Aided Mol. Design,* **2,** 321.
71 [] Frandsen, Aa., Krogsgaard-Larsen, P. and Schousboe, A. (1990) J. *Neurochem.,* **55,** 1821.
72 [] Krogsgaard-Larsen, P., Johnston, G.A.R., Lodge, D. and Curtis, D.R. (1977) *Nature (London),* **268,** 53.

4
Physicochemical properties and drug action: alternative QSAR methods

R.Franke, S.Dove and A.Gruska

Institute of Drug Research, Alfred-Kowalke-Str. 4, 0–1136 Berlin, Germany

SUMMARY

Biological properties of chemical compounds depend in a very sensitive way on their chemical structure. The understanding of such dependencies in terms of socalled structure-activity relationships is one of the principal goals of medicinal chemistry and at the heart of modern computer-aided drug design (CADD). CADD methods have become an indispensable tool in the search for new drugs and may not only aid to guide experimental work effectively but also to elucidate mechanisms of action at the molecular level. Because of the complexity of biological matter and the huge number of chemical structures and possible variations, a large number of quite different CADD methods serving different purposes have been developed. These methods can be roughly categorized into (i) statistical approaches relating physicochemical parameters to biological potency, (ii) heuristic methods usually based on substructural consideration, and (iii) molecular modelling with powerful interactive computer graphics as key instrument. Although the recent development has very much stressed molecular modelling the other methods are still very useful in many cases. This presentation will concentrate on the first category, where so-called quantitative structure-activity relationships (QSARs) are derived with biological potency as dependent and physicochemical parameters as independent. These characterize hydrophobic, electronic and steric properties of drugs and are usually applied to characterize variations of substituents. The most important parameters and their physical meaning will be discussed, together with specific aspects of their application to biological problems. With typical examples from Hansch analysis important aspects of QSARs such as interpretability, predictive power, general QSARs know-how and limitations of QSARs approaches will be outlined. Alternative statistical methods such as classification methods or principal component analysis will also be briefly introduced.

INTRODUCTION

The most commonly used mathematical technique in classical quantitative structure-activity relationship (QSAR) work is multiple regression analysis. There are, however, cases where this method cannot be applied or does not represent the optimal choice. Some typical situations of this kind frequently encountered in drug design will be discussed in this chapter. These are: (i) biological potency is not expressed on a continuous scale but as a classification (activity score); (ii) several biological tests have been performed in parallel; (iii) the number of variables to be included in an analysis is large compared to the number of observations, and/or the variables are highly collinear; (iv) the molecular descriptors are not normally distributed (for example, logical quantities); (v) the mathematical model representing the relation of descriptors with biological potency is unknown and cannot be guessed.

A variety of methods partly similar with respect to the underlying mathematical principles can be applied in such situations. A complete discussion would be far beyond the scope of this chapter. We have therefore selected those methods which we have found to be most important in our own QSAR work (see also Franke [1]). This selection is, of course, influenced by our taste and research history, but we certainly hope that it is representative from a QSAR point of view.

Although mathematics itself is a rigorously defined and objective discipline, its application to practical drug design problems is not without a subjective component. For example, from a very strict mathematical point of view, simple multiple regression analysis is, for various reasons, not applicable to the majority of drug design problems. Yet, this method has proven to be of immense practical value and has yielded chemically valid information in a vast number of cases. It is not the beauty of a perfect mathematical solution that is important in QSAR work but rather being able to obtain some model which may be interpreted straightforwardly in chemical terms. Thus, it is chemistry which is in the foreground and most important, while mathematics is just a tool (among others) to systematize chemical (and biological) data. Taking a very pragmatic view it seems justified to apply certain mathematical methods even if not all mathematical requirements are strictly fulfilled, as long as one knows what one is doing and the results are interpreted with the necessary care. We admit that this view may provoke arguments and is by no means shared by all QSAR workers, but one must not forget that promising predictions are not pure results of completely computerized operations, but in the first place also of mechanistically based interpretations of QSARs. Having said that a warning is in order not to misuse mathematical methods. With all the computers and software around the temptation may be great, especially for a beginner in the QSAR field, just to throw the data in and see what happens.

Even if the procedure selected is not adequate results will come out in many cases which look just fine to the inexperienced eye but are completely worthless. What is important is to learn the limits of the various mathematical techniques in

their application to QSAR problems and how to select the most appropriate (which implies the most simple) variant according to the nature of the data and to a well-defined objective of the analysis, taking into account the possibilities offered from the experimental side to put theoretical conclusions to a practical test.

PRINCIPAL COMPONENT ANALYSIS AND RELATED METHODS

In this section we shall consider the rather general case where for a series of chemical compounds measurements are made in a number of parallel biological tests and where a set of descriptor variables is believed to be related to the biological potencies observed. In order to understand the data in their entirety and to deal adequately with the mathematical properties of such data, methods of multivariate statistics are required. A variety of such methods is available as, for example, multivariate regression, canonical correlation, principal component analysis, principal component regression, partial least squares analysis, and factor analysis, which have all been applied to biological or chemical problems (for reviews, see [1–11]). Which method to choose depends on the ultimate objective of an analysis and the property of the data. We have found principal component and factor analysis particularly useful. For this reason and also since many multivariate methods make use of 'components' for 'factors' we will start with these methods in some detail, while the discussion of other approaches will be less extensive.

Principal component analysis [1, 11]

If for n chemical compounds (j=1, …, n) biological potencies are measured in m biological tests (i=1, …, m) the results can be arranged in a matrix which we shall call the biological data matrix. If the tests are put in the rows and the compounds in the columns, the matrix has the form

$$\mathbf{Y} = (y_{ij})_{m,n} \tag{1}$$

where y_{ij} is the biological potency of the jth compound in the ith test. The row items of this matrix are usually called variables and the column items objects. We shall follow this convention although, from a purely mathematical point of view, there is no need for such a decision. We shall assume, furthermore, from now on that all variables, y_i, are standardized by autoscaling so that they have means of zero and unity variance according to

$$y_{ij} = (y_{ij}^{(0)} - \bar{y}_i^{(0)}) / s_i^{(0)} \tag{2}$$

In this equation, the are the activity values originally measured, is the mean value in the ith test, and s_i is the standard deviation of the measured values in the ith test. This standardization is necessary in order to give all variables the same

importance. Thus, if we talk of measured values or measurements in this chapter, we shall always tacitly assume that autoscaling according to equation (2) has been performed as part of the measurement.

If the biological tests considered are similar from a biological point of view the following simplifying assumptions can be made:

(i) The observed biological response in each test depends on a number of basic effects as, for example, transport through a membrane, binding to a biological target, etc.
(ii) These basic effects are present in all tests but to varying degrees.
(iii) The biological response, y_{ij}, may be expressed as linear combinations of these effects.
(iv) The measurements are free from experimental error (which, of course, is never true) and from other influences not due to the basic effects; that is the y_{ij} can be exactly reproduced by such linear combinations.

If there are p such effects E_k (k=1, …, p) we then obtain

$$y_{ij} = \sum_{k=1}^{p} w_{ik} E_{kj} \tag{3}$$

or, in matrix notation,

$$\mathbf{Y} = (w_{ik})_{m,p} (E_{kj})_{p,n} = \mathbf{W} \cdot \mathbf{E} \tag{4}$$

The value of E_{kj} reflects how strongly the kth basic effect is affected by the jth compound. Thus, the E_{kj} are characteristic of the compounds since the strength of the basic effects is determined by the chemical properties of the compounds. The weights w_{ik} are a measure of how important the corresponding E_{kj} are in each biological test. They are characteristic of the biological tests since it depends on the properties of the biological systems which basic effects operate there and to what extent. Let us construct a simple example where only two basic effects (p=2) operate in three biological tests (m=3). Then, the system of equations represented by equation (3) takes the form

$$\begin{aligned} Y_{1j} &= w_{11} E_{1j} + w_{12} E_{2j} \\ Y_{2j} &= w_{21} E_{1j} + w_{22} E_{2j} \\ Y_{3j} &= w_{31} E_{1j} + w_{32} E_{2j} \end{aligned} \tag{5}$$

and the matrices **W** and **E** become

$$\mathbf{W} = \begin{array}{c} \\ \text{t} \\ \text{e} \\ \text{s} \\ \text{t} \\ \text{s} \end{array} \begin{array}{c} \text{basic effects} \\ \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{bmatrix} \end{array} \tag{6}$$

$$\mathbf{E} = \begin{matrix} \\ \text{basic} \\ \text{effects} \end{matrix} \begin{matrix} \text{effects} \\ \text{compounds} \end{matrix} \begin{bmatrix} E_{11} \ldots E_{1j} \ldots E_{1n} \\ \\ E_{21} \ldots E_{2j} \ldots E_{2n} \end{bmatrix} \tag{7}$$

By the matrices **W** and **E,** the features of the chemical compounds are completely separated from those of the biological tests since **E** solely depends on the properties of the molecules and **W** solely on the characteristics of the tests. Such a separation may give a much deeper insight into the data structure and its underlying effects than can be obtained from the global response data. In addition, the dimensionality of the data space is reduced. While the original data matrix has three rows, the matrices **E** and **W** have only two rows or two columns, respectively. Geometrically, this means that the objects and test systems can now be represented in two-dimensional coordinate systems spanned by the row vectors of **E,** or the column vectors of **W,** respectively, while, originally, the respective coordinate systems would have had 3 (rows of **Y**) or n (columns of **Y**) axes. If, in addition, the two effects could indeed be labelled 'transport' and 'target binding' by some suitable procedure, we would have learned a lot more about the data and actually obtained information about the basic effects operating and underlying the entirety of data without having measured them.

A well-known mathematical method to treat data matrices in accordance with the model in equation (3) is principal component analysis (PCA) (for some reviews see, for example, [1, 11–13]).

The model underlying PCA exactly corresponds to equations (3) and (4): the elements of the standardized data matrix **Y** (see equation (1)) are described by a sum of product terms where, in each term, one factor is characteristic of the molecules and the other of the biological test. In the terminology of PCA this model reads:

$$y_{ij} = \sum_{k=1}^{p} a_{ik} \; P_{kj} + \sum_{k=p+1}^{m} a_{ik}^{(0)} \; P_{kj}^{(0)} \tag{8}$$

The P_{kj} in equation (8) are called principal components (PCs). They correspond to the E_{kj} in equation (3) and characterize the features of the compounds. Mathematically, they are orthogonal vectors which are so determined that the original data matrix is reproduced. Analogously, the a_{ik} in equation (8) correspond to the weights w_{ik} in equation (3) and characterize the test systems. The 'weights' a_{ik} are a measure of the contribution of the kth PC to the ith variable, y_i (ith row of **Y**), and represent the correlation coefficient between this PC and y_i. As a consequence, a high value of $|a_{ik}|$ points to a high importance of the kth PC for the ith variable. The ith variable is then said to be highly 'loaded' in the kth PC, and the a_{ik} are, therefore, also called 'loadings'.

Mathematically, the number of PCs which can be extracted from a data matrix is usually equal to *m,* the number of original variables, y_i. With this number of components the data matrix is exactly reproduced. This, however, is not a desired result since it would not lead to a reduction of dimensionality of the data space. From a model point of view, one wishes to find the minimum number, *p,* of components such that, in a space spanned by them, the original variables can be represented without loss of relevant information. Only then will the components truly reflect the basic effects 'behind' the data in the sense of equation (3). These *p* components are represented by the first term in equation (8) while the components with superscript '(0)' in the second term represent 'non-relevant' or 'residual' information comprising errors of measurement and, possibly, some model error. The objective of PCA is to evaluate only the first sum in equation (8); the resulting components will then reproduce the data matrix **Y** within residual error. The following relation, which will not be derived here, serves as a starting point to evaluate the loadings and components:

$$\mathbf{R} = \mathbf{A}\,\mathbf{A}^T \tag{9}$$

R denotes the correlation matrix of the data which, for standardized variables, is identical to the covariance matrix, and **A** is the loading matrix:

$$\mathbf{A} = (a_{ik})_{m,p} \tag{10}$$

To solve equation (9) the additional condition is introduced that the components are determined in sequence in such a way that the first component accounts for the largest and every next for the largest remaining part of the correlation in **R**. An eigenvalue problem then ensues according to

$$\mathbf{R}\,\alpha_k = \lambda_k\,\alpha_k \tag{11}$$

where the λ_k are eigenvalues and the α_k the corresponding ortho-normalized eigenvectors. It is solved by diagonalizing **R** using standard procedures. The *k*th loading vector (*k*th column vector of matrix **A**) is then obtained from

$$\mathbf{a}_k = \alpha_k^2 \sqrt{\lambda_k} \tag{12}$$

and the component matrix

$$\mathbf{P} = (P_{kj})_{p,n} \tag{13}$$

can be computed from

$$\mathbf{P} = \Lambda^{-1}\,\mathbf{A}^T\mathbf{Y} \tag{14}$$

where is the inverse of the diagonal matrix of the eigenvalues. The eigenvalues are the variance contributions from the component; the quotient λ_k/m represents the part of the overall variance of the standardized data extracted by the *k*th PC. As a consequence, the first component will have the largest and the last the smallest eigenvalues.

In order to find the minimum number of components, *p,* necessary for data reproduction within residual error, the components are added step by step to the model according to equation (8). After each step the data matrix is reproduced and the procedure continues until only non-systematic 'noise'

remains. Possibilities to recognize this point are the technique of cross-validation [14] or the indicator function introduced by Malinowski [11, 15], which we found quite useful. The indicator function can easily be computed after each step [1, 11, 15]. Its value passes through a minimum for the correct number of components ($k=p$). The presence of such minimum also proves that the data under investigation can be correctly expressed by a PCA model.

If the minimum number of PCs necessary to reproduce the data within experimental error has been found, the PCA as such is essentially complete. The practical evidence made so far is that the number of relevant components reflects the true complexity of the data in the sense of basic effects. Further interpretation of the result can be achieved at two levels:

(i) Considering matrix **A** will give information about the internal relatedness of the biological tests. As already stated the a_{ik} represent correlation coefficients between original variables and components. A grouping of variables (in our case, biological tests) can, for example, be demonstrated by a simple plot with the a_{ik} as axes. In such a plot each variable appears as a point, the points of related variables being close together. To obtain a numerically comprehensive picture of the relatedness of the tests, it is also possible to rotate the loading matrix **A**. This procedure corresponds, in principle, to a rotation of the axes of the original plot so that they pass through swarms of points representing closely related variables and, thus, basic effects. For mechanistical reasons the basic effects must not necessarily be uncorrelated; that is, in some cases oblique rotations can produce the simplest, most interpretable structure of the loadings.

(ii) Identification of the 'abstract' components with physically meaningful parameters will indicate what the physical nature of the basic effects underlying the components is and eventually lead to multivariate QSARs. To this end a special target rotation procedure introduced by Weiner and Malinowski (for details, see [11]) can be used. First, the parameters which are believed to be related to the components must be selected either from a theoretical model or from educated guesses (analogously as in Hansch analysis). The components are then rotated into vectors of these (also autoscaled) parameters (for example π, α) which are called test vectors; the elements of the test vectors are the values of the corresponding parameters for the objects (in our case, compounds) of the data matrix.

If the target rotation has been successfully performed the data can be reproduced by the test vectors t_k. Equation (8) then transforms into

$$y_{ij} = \sum_{k=1}^{p} a_{ik}^{R} t_{kj} + \epsilon_{ij} \qquad (15)$$

where t_{kj} is the value of the kth vector for the jth compound, is the residual including the second term of equation (8) plus the error of the target rotation, and the are the elements of the rotated loading matrix, $\mathbf{A}^R$, obtained from

$$\mathbf{A}^R = (a_{ik}^R)_{m,p} = \mathbf{A}\ \mathbf{Q}^4 \qquad (16)$$

The elements of the rotation matrix, **Q,** are results of least-squares fits as in multiple regression analysis, with the test vectors as dependent and the components $\mathbf{P}_k$ as independent variables. Equation (15) represents a system of 'multivariate' equations (QSARs) as in multivariate regression analysis, with the original variables, y_i (autoscaled measurements in the biological tests), as dependent and the test vectors as independent variables.

In many cases, the primary objective is not a complete replacement of components by test vectors but rather the interpretation or identification of individual components with physicochemical parameters. A much simpler approach can then be used. The components as dependent variables are correlated with molecular parameters in a standard regression procedure. As a result, regression equations for each of the components considered are obtained.

To demonstrate more clearly what PCA can do we shall discuss a simple example [16] concerning the inhibitory potency of a series of nine 4-hydroxyquinoline-3-carboxylic acids of structure **I** (Fig. 5, p. 82) against respiration of Ehrlich ascites tumour cell suspensions and three respiratory enzymes *in vitro* (muscle lactate dehydrogenase, M4-LDH; pig heart cytoplasmic malate dehydrogenase, s-MDH; pig heart mitochondrial malate dehydrogenase, m-MDH; for data see Dove *et al*. [16]). Two components (eigenvalues: λ_1=2.752, λ_2=0.873) are necessary to reproduce the data within experimental error as followed from the indicator function of Malinowski (see Table 2). Figure 1 shows a plot of the four test systems in a coordinate system with the $\mathbf{a}_k$ as axes. It follows that the ascites test is completely separate from the three enzyme tests, so that these do not model ascites respiratory inhibition and do not provide information about the latter. Thus, none of the enzymes can be the sole target for the inhibition of Ehrlich ascites respiration. The same picture is revealed by the loadings presented in Table 1. All enzymes have high positive loadings in (high correlations with) the first component, indicating that they are closely interrelated. Their loadings in the second component are practically zero. The ascites test, on the other hand, has only a small (negative) loading in the first but a high loading in the second component reflects enzyme inhibition *in vitro,* whereas the second one is afforded by the ascites system so that the first component, P_1, may be regarded as an average expression for the inhibition of all three enzymes, with only a small contribution from the ascites system. Correlating P_1 with molecule parameters (see Table 2) yielded the following relationships (the SO_3^- derivatives are not included because of difficulties in calculating μ):

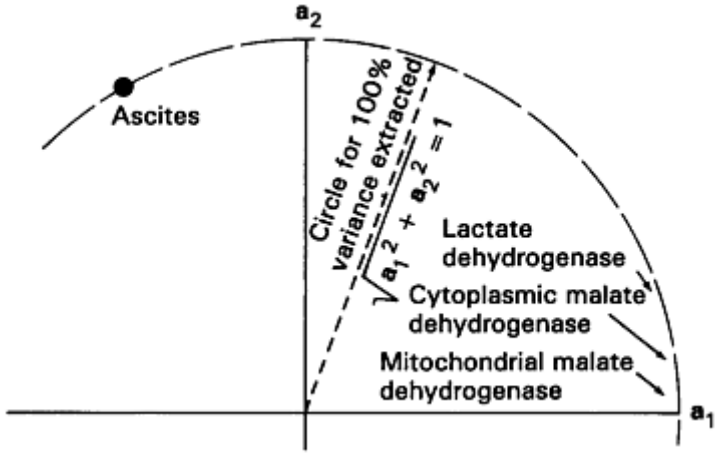


Fig. 1. Plot of the four test systems applied to investigate the inhibitory potency of structure I compounds in a coordinate system with the loading vectors a_1 and a_2 as axes.

$$P_1 = 0.34(\pm 0.27)\mu + 1.38(\pm 0.78)B_4 - 4.86(\pm 2.22)$$
$$n = 8 \quad r = 0.902 \quad s = 0.520 \tag{17}$$

Table 1. PCA of the data on enzyme and ascites inhibition. Loadings a_{iK} of the test systems with respect to the two relevant PCs, and part of the variance of the variables extracted by the PC model.

Test system	a_{i1}	a_{i2}	% variance extracted
Lactate dehydrogenase	0.87	0.30	84.7
Cytoplasmic malate dehydrogenase	0.97	0.12	95.5
Mitochondrial malate dehydrogenase	0.91	0.03	82.9
Ehrlich ascites tumour cells	−0.47	0.87	97.8

Table 2. PCA of the data on enzyme and ascites inhibition. PCs P_1 and P_2 (standardized according to equation (14)) of the 7-substituted 4-hydroxyquinoline-3-carboxylic acids, and descriptor variables used for the identification of the PCs

R	P_1	P_2	π^a		μ^c
SO_2CH_3	1.066	0.070	−1.39	3.15	4.123
OH	0.206	−0.064	0.06	1.93	5.371
Cl	−1.990	0.941	0.55	1.80	3.130
COOH	−0.061	−1.895	−2.80	2.66	4.011
SO_3^-	−0.803	−0.524	−4.76	2.70	(2.103)
SO_2NH_2	0.600	−0.829	−1.36	3.07	4.379
$COCH_3$	0.290	0.463	−0.39	2.93	3.653

R	P_1	P_2	π^a		μ^c
NO_2	−0.527	−0.345	−0.40	2.44	1.097
$N(CH_3)_2$	1.221	1.489	1.10	2.80	5.942

[a]Calculated from apparent log P values [17]. [b]Taken from [19]. [c]Calculated by CNDO/2 method [16].

In equation (17), μ is the dipole moment (calculated by CNDO/2 method [16], and B_4 is Verloop's width parameter (values taken from [19]). It thus follows that enzyme inhibition depends on both steric and electronic effects, inhibitory potency increasing with increasing dipole moment or electronic-donating power of the substituents and with increasing substituent width.

For the second component the following relation was obtained (π=hydrophobic substituent constant calculated from apparent partition coefficients [17]:

$$P_2 = 0.79(\pm 0.24)\,\pi + 0.51(\pm 0.33) \qquad n = 8 \quad r = 0.931 \quad s = 0.415 \tag{18}$$

No relationships with significant contributions of steric or electronic parameters could be found.

Equations (17) and (18) cannot be used directly for predictive purposes since they only describe and identify the components and not the original pI_{50} values directly. They clearly show, however, what the physical nature of the components is and in what direction structure I has to be varied in order to arrive at more potent derivatives. True equations in the sense of a QSAR could have been obtained by the already mentioned special target rotation of Weiner and Malinowski, which would lead to a direct replacement of the abstract P_{kj} by molecule parameters according to equation (15). This, however, would not add new information in the present case, at least not for the three enzyme systems. For the ascites test the situation is a little different since here a small contribution of the first component in addition to the second component is present. If a Hansch analysis is directly applied to the experimental data of ascites inhibition alone a relation with π very similar to equation (18) is again obtained which can, however, be improved *slightly* by adding an electronic term.

As a result of PCA we have thus quite straightforwardly obtained a clear picture of the data structure as well as of the physical nature of the two 'basic effects' reflected by the two components.

PCA has been applied to a variety of problems in the natural sciences. We shall briefly mention here only some applications in the QSAR field.

Multivariate data obtained from different strains of bacteria are especially suited for PCA; most examples thus concern antibacterials [1, 20–25, 181, 184, 187, 208]. Other applications deal with herbicides [26], opioids [4], structure-odour relationships [186], anti-inflammatory compounds [27, 182] and with the quantification of toxic compounds [207], respectively. Niemi *et al*. carried out a PCA of 90 molecular connectivity indices and needed eight PCs to explain more

than 93% of the variance of the data set. The authors were able to predict successfully the biodegradations of 340 compounds [180]. In some cases, data matrices with measures of a biological activity at different times were investigated by PCA to separate and identify pharmacodynamic and pharmacokinetic components [1, 3, 28]. With the same objective, Schaper and Seydel [29] successfully analysed multivariate sets of pharmacokinetic parameters of several drug classes.

A somewhat different application of PCA can be found in the work of Cammarata and Menon [30, 31]. They derived a data matrix for compounds with different biological actions by coding the presence or absence of groups at designated positions and weighing this coding by the molar refractivity of the groups present. A coordinate system with the resulting PCs as axes then served to plot the data, leading to a certain clustering of compounds with similar types of action.

It should be noted that the application of PCA in QSAR work often ends at the stage of such clustering; that is, the method serves as a cluster analysis approach. The objective is that of classification methods (pattern recognition), to be discussed later. If the classes of compounds are known before analysis, however, this approach works better with separate PC models for each class, as in the SIMCA method of Wold *et al*. [9]. Some examples of using simple PCA of the whole training series for cluster analysis are given in [32–35]; diverse pattern recognition techniques (linear learning machine, Bayes linear, *K*-nearest neighbour) were used in Henry *et al*. [179]. Another example of a mapping technique using principal components is the TMIC method [1, 36] for the design of representative training series (good spread of descriptor variables, no collinearities present) for Hansch analysis. In TMIC (two-dimensional mapping of intra-class correlation matrices) PCA is applied to an intra-class correlation matrix obtained from substituent constants for a set of compounds. The compounds are then plotted in the coordinate system spanned by the two first PCs, and representative training series can then be selected from the resulting map by simple inspection. Such and other mapping techniques are of great importance in QSAR work but cannot be discussed in more detail in this chapter. For reviews the reader is referred to Lewi [6, 7] and, with regard to modelling and prediction of toxicity, to [183].

Applications of PCA to chemical problems will not be mentioned in this chapter, with the exception of the following few examples which have direct importance for QSAR work. Applying PCA to a large data matrix containing various physicochemical constants for a great number of chemical compounds, Cramer [37] derived the new parameters B, C, D, E and F which seem to describe basic properties of compounds in terms of bulk and polarity, and they have successfully been applied in QSAR work [38]. Results of PCA applied to various data matrices of hydrophobicity parameters (π, log P) helped a better understanding of the intrinsic nature, dependence on chemical structure and solvent dependence of these quantities [1, 39–43, 185, 206]. Sjöström and Wold

[44] subjected 20 properties (molecular weight, p*K* values, steric, spectroscopic and chromatographic constants as well as thermodynamic parameters) of 20 amino acids to PCA, yielding three significant PCs. With these PCs as descriptor variables Z (Z_1 is mainly related to hydrophobicity, Z_2 is a combination of size, hydrophilicity and ^{1}H-nuclear magnetic resonance (^{1}H-NMR), Z_3 relates to p*K*, p*I* and ^{1}H-NMR shift data), peptides can be characterized in QSAR approaches [45, 177, 178, 188]. Several authors [189, 209, 219] extended the existing Z scales to additional non-coded amino acids. Skagerberg *et al*. [211] extracted three 'principal properties' from a 9×100 data set for aromatic substituents.

At the end of this section a relatively new method, correspondence analysis [46], should be mentioned, which is methodologically closely related to PCA and especially suited for investigating structure-affinity relationships (ligand-receptor association) [47]. Correspondence analysis differs from PCA mainly by the starting point. Instead of the correlation matrix (containing correlations of affinity constants for different receptors), a matrix of distances calculated from the relative frequencies of binding between ligands and receptors is subjected to a principal component approach. An advantageous property of correspondence analysis is that the values of the resulting 'loadings' and 'components' are independent of whether one starts from the distances between receptor or between object pairs; that is, the two fields of the data matrix (rows and columns) have the same importance, and 'loadings' as well as 'components' can be represented in the same plot. That allows objects to be clustered according to the receptor (test) of which they are characteristic ligands.

Factor analysis

Factor analysis (FA) is very similar to PCA (for reviews see, for example, [1, 12, 13]). The only, but essential, difference is that in FA only part of the data variance is considered to be common to all variables. The remaining part is attributed to unique properties of one variable at a time. With this in mind equation (8) may be written to give the model of FA as

$$y_{ij} = \sum_{k=1}^{p} a_{ik}\ f_{kj} + d_i\, q_{ij} \qquad (19)$$

In this equation, the so-called common factors, $\boldsymbol{f}_k$, which span the common factor space replace the components in equation (8), and the a_{ik} are again the loadings representing the correlations between common factors and the original variables. The $\mathbf{q}_i$ are called unique factors; their squared loadings, the uniquenesses d_i^2, comprise that part of the data variance which is attributable to unique variable properties not involved in the common correlation structure. FA is the method of choice in all cases where such unique properties of variables occur, and this is to be expected when a data matrix contains variables quite different in nature and meaning which are only loosely interrelated. In such matrices, error variances

differing in size are to be expected even for autoscaled data, which is another reason for applying FA instead of PCA.

The common features (the basic effects) searched for in the data space are represented by the factors f_k. Their extraction from given data is based on the general assumption that only a certain part of the variability of, for example, a biological test can be explained by effects also present in the other variable under consideration, that is, is due to the correlation structure. As a consequence only so much of variance is considered that a minimum of common factors results.

The first objective of FA is the evaluation of the loading matrix

$$\mathbf{A} = (a_{ik})_{m,p} \tag{20}$$

which is also called factor pattern in FA. The procedure is nearly the same as outlined in equations (11) and (12), the only difference being that instead of the correlation matrix **R** the so-called reduced correlation matrix $\mathbf{R}^+$ is being diagonal ized. $\mathbf{R}^+$ differs from **R** in that the 1's in the diagonal of the latter have been replaced by communalities h_i^2. As a result, the reduced correlation matrix only contains that part of the data variance which can be assigned to the common factor space. Representing this part, the communality for the ith variable is defined as

$$h_i^2 = \sum_{k=1}^{p} a_{ik}^2 \tag{21}$$

and the variance of the ith standardized variable then becomes where d_i^2 represents the uniqueness. The uniqueness consists of the error variance and the socalled specificity, the latter representing mechanistically meaningful, specific properties and systematic divergencies of the ith variable:

$$s_i^2 = h_i^2 + d_i^2 = 1 \tag{22}$$

Communalities must be estimated prior to the analysis. This can be accomplished in several ways which, however, will not be discussed here. During the computations, these estimates can be improved through iteration cycles. The number of relevant factors, *p,* is usually determined from the corresponding eigenvalues. Only those factors are considered significant whose eigenvalues exceed a given borderline value and which therefore account for more than a given minimum variance. The borderline value is usually taken as representing an eigenvalue of 1 or a value equal to 5% of the sum of all positive eigenvalues.

Before interpretation the factors are usually so rotated that the factor pattern becomes as simple as possible (Thurstone's simple structure). This structure is characterized by the property that a maximum of variables is close to the coordinate axes when presented in common factor space (axes=factors), so that the largest possible number of factor loadings becomes zero. Thus, in the presence of a simple structure the variables are divided into mutually exclusive groups with, in an ideal case, non-zero loadings only in one factor. Whereas the original factors are always orthogonal due to the method of their extraction, the rotation can be orthogonal or oblique. In some cases, a simple structure is

achieved only by oblique methods which are also justified by the fact that the 'basic effects' underlying the data do not necessarily have to be uncorrelated.

In many applications of FA the procedure is stopped at this stage. Many investigators are primarily interested only in how the variables are mutually interrelated and how many basic features are behind them, which can be learned completely from the factor pattern. Another reason is that the evaluation of the factors f_{kj} (factor scores) is not straightforward as in PCA. Since the rank of the factor matrix (common+unique factors) generally exceeds the number of variables, equation (14) cannot be applied, and the f_{kj} must be estimated in an indirect way. Two methods of doing that which will not, however, be discussed here are the regression estimation and the Bartlett estimation [13]. Although it is somewhat laborious to estimate the f_{kj} it is worth while since the factors characterize the features of the objects of the data matrix and can be handled in much the same way as the principal components (provided that the part of the data variance represented by the common factor space is large enough).

A data matrix where FA seems to be more effective than PCA because of the reasons mentioned above is provided by measurements performed for a series of diphenylaminopropanols (structure **II,** Fig. 5, p. 82) by Keasling and Moffett [48] in 11 biological tests. From a pharmacological point of view these tests represent quite different mechanisms of action, so that the results cannot be expected to be very closely interrelated. Since, in addition, it is to be assumed that the error of measurement differs from test to test this is a typical case for FA. It is therefore not surprising that no convincing results could be obtained in an attempt to apply PCA to these data [49]. Application of FA [27] yielded three relevant common factors which account for 79.8% of the data variance. Table 3 presents the factor loadings after orthogonal rotation of the factors to give a simple structure, together with the communalities and the values for the uniqueness, and Fig. 2 shows a plot of the variables in a coordinate system with the rotated factors as axes. A grouping of the biological tests becomes evident. The simple reflexes and, surprisingly, also the anticonvulsant action belong to the first factor, indicating that both types of action may be due to at least similar molecular interaction mechanisms. Note that the nicotine antagonism obviously is influenced by an additional effect reflected by its substantial uniqueness. The second factor represents tremorine antagonism, but not mydriatic activity, which is also based on anticholinergic effects. Placed in the third factor are the pupil test, together with lethality as main part, as well as fighting and a substantial part of the chimney test. By reason of this heterogeneity, this factor cannot be interpreted unambiguously.

Factors f_1 to f_3 (factor scores) were estimated by the regression method and are presented in Table 4. Considering their statistical characteristics [27], these estimates can be regarded as close enough to the true values to justify their identification with physicochemical molecular parameters (see Table 4) by means of multiple regression analysis. The best equations obtained (compound no. 15 not included in equations (23) and (25)) are:

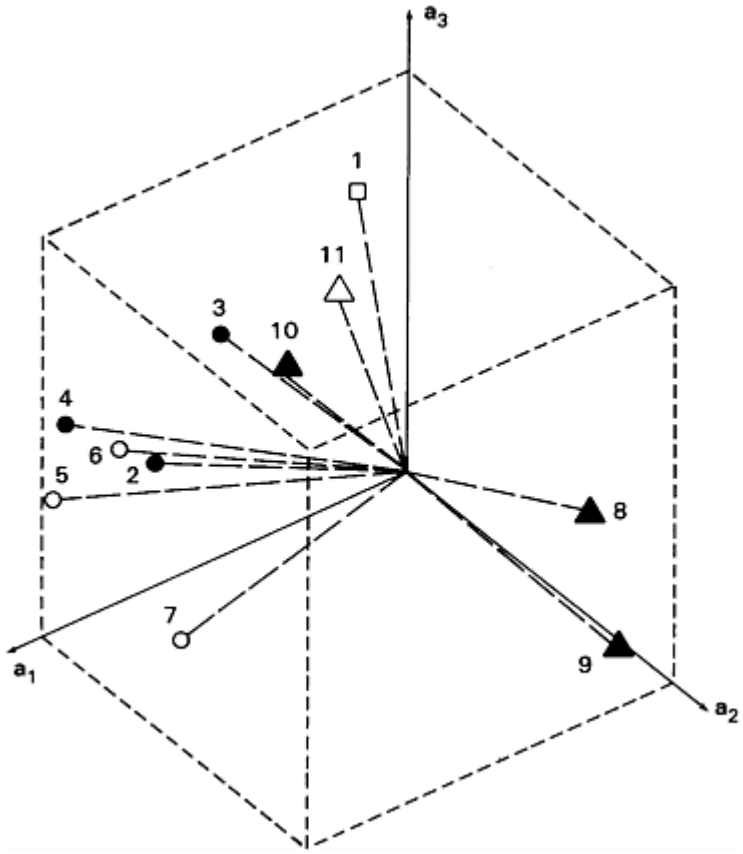


Fig. 2. Factor analysis of the Keasling/Moffett [48] data: plot of the tests in the space spanned by the three orthogonally varimax-rotated common factors. Groups of variables are: toxicity (□); simple reflexes (●) — traction (2), chimney (3), dish (4); anticonvulsant activity (○) — thiosemi carbazide (5) and nicotine antagonism (6), electroshock (7); anticholinergic actions (▲) — central (8) and peripheral tremorine antagonism (9), mydriatic activity (10); fighting (Δ) (11).

$$f_1 = -1.25(\pm 0.40)\pi_\alpha - 0.73(\pm 0.30)\pi_N + 1.63(\pm 0.40)I_a - 2.12(\pm 0.57)I_\beta + 1.51(\pm 0.42) \qquad (23)$$
$$n = 14 \quad r = 0.951 \quad s = 0.396$$

Table 3. Factor analysis of the Keasling/Moffett data. Factor loadings aik $_{of}$ the tests with respect to the three orthogonally varimax-rotated common factors, communality and uniqueness of the variables (++, high loadings; +, moderate loadings)

No.	Variable	a_{i1}	a_{i2}	a_{i3}		
1	Toxicity	0.21	0.14	0.75++	62.7	37.3
2	Traction	0.79++	0.23	0.41	84.6	15.4

No.	Variable	a_{i1}	a_{i2}	a_{i3}		
3	Chimney	0.63+	0.31	0.68+	95.3	4.7
4	Dish	0.86++	0.00	0.37	86.7	13.3
5	Thiosemicarbazide antagonism	0.87++	−0.01	0.22	80.8	19.2
6	Electroshock	0.85++	0.22	0.46	97.3	2.7
7	Nicotine antagonism	0.74++	0.31	0.08	65.1	34.9
8	Central tremorine antagonism	0.09	0.81++	0.39	82.3	17.7
9	Peripheral tremorine antagonism	0.15	0.88++	0.12	80.8	19.2
10	Mydriatic activity	0.57	0.39	0.64+	88.7	11.3
11	Fighting	0.32	0.28	0.59+	53.2	46.8

Table 4. Factor analysis of the Keasling/Moffett [48] data. Factor scores of the diphenylaminopropanols with respect to the three orthogonally varimax-rotated common factors estimated by the regression method, and descriptor variables used for the identification of the factors. Common structure: $Ph_2C(OH)R$.

No.	R=C–(C_β)–(C))–R_N	f_1	f_2	f_3							
1	CH_2CH_2Pyr	1.49	1.62	−0.17	0.	0.01	0	0	0	0	0
2	C(=CH2)$CH_2N(iPr)_2$	−0.47	−0.74	−0.42	−0.01	0.55	−1	1	0	0	0
3	CHMe$CH_2N(iPr)_2$	2.35	−0.98	−0.91	0.54	0.55	1	1	0	0	0
4	C(=C=CH_2)CH_2Pyr	−0.08	0.50	0.08	0.10	0.01	−1	1	0	0	0
5	CHEtCH_2Pyr	0.57	−0.45	−0.38	1.20	0.01	0	1	0	0	0
6	CHiPrCH_2Pyr	−0.51	0.02	1.62	1.61	0.01	0	1	0	0	0
7	CHPhCH_2Pyr	−0.99	−0.74	0.10	1.67	0.01	0	1	1	0	0
8	CHiPrCH_2NH_2	0.91	0.06	1.33	1.61	−1.77	0	1	0	0	0

No.	R=C–(C_β)–(C))–R_N	f_1	f_2	f_3							
9	CMe_2 CH_2Py r	−1.40	−0.38	−1.30	1.08	0.01	−1	1	1	0	0
10	CH_2C HMeP yr	−0.67	2.06	0.93	0.	0.01	0	0	0	1	0
11	CH_2C HMeN HMe	0.51	−0.95	0.18	0.	−1.61	0	0	0	1	1
12	CHMe Pyr	−0.05	−0.71	0.80	0.54	0.01	0	1	0	0	0
13	CH_2C HMeC H_2Pyr	−0.47	1.17	−0.72	0.	0.01	0	0	0	1	0
14	CH_2C HMeN MeCH $_2$Ph	−1.30	−0.27	−1.69	0.	0.86	0	0	0	1	0
15		0.30	0.75	−0.45	0.54	−1.26	0	1	0	0	0
16		−0.19	−0.95	0.99	–	–	–	–	–	–	–

[a] π-values calculated using the fragment approach of Leo and Hansch [19]; .
[b] π-values of the whole neutral amine substituent, calculated as π_α.
[c] 1, if R=Me and at least two C atoms between C(OH) and N; −1, if R=di-Me or if a double bond in α-position.
[d] 1, if α-branch is present.
[e] 1, if di-Me or Ph in α-position.
[f] 1, if single β-branch is present.
[g] 1, if RN is a secondary amino group.

$$f_2 = -0.67(\pm 0.44)\pi_N - 1.69(\pm 0.63)I_\alpha - 3.33(\pm 1.40)I_N + 1.29(\pm 0.52) \quad (24)$$
$$n = 15 \quad r = 0.850 \quad s = 0.570$$

$$f_3 = 0.68(\pm 0.61)\pi_\alpha - 1.27(\pm 0.90)\pi_N - 0.68(\pm 0.63)\pi_N^2 - 1.51(\pm 1.15)I_{\alpha 1} - 0.01(\pm 0.55) \quad (25)$$
$$n = 14 \quad r = 0.810 \quad s = 0.685$$

These quantitative structure-factor relationships, at least for the first two factors representing two different 'basic effects' which determine the main part of the biological activities, are well identified and interpretable. High values of the first factor representing simple reflexes and anticonvulsant action will result for compounds with low hydrophobicity of R_α and R_N, an α-Me branch, no second Me or double-bonded substituents in α-position, and no branch at the β-C atom.

For the second factor reflecting tremorine antagonism low hydrophobicity of R_N, no branch in the α-position, and the absence of a secondary amino group in R_N will result in high values. Equation (25) finally indicates that increasing values of the third factor will result with increasing hydrophobicity of substituents in R_α, an optimal hydrophobicity of R_N ($\pi_0 \approx -0.95$) and no di-Me or phenyl substitution at the α-C atom.

For the data discussed above FA seems to be more effective than PCA since it leads to a very straightforward and well-interpretable result in only three dimensions (three factors), while PCA led to an eight-dimensional space (eight components [49]). Since in many cases the properties of the data to be analysed are not known, or are not known precisely enough, it is sometimes difficult to decide whether to apply PCA or FA. A good first move then is always to start with FA. If the communalities of all variables exceed about 80% PCA can be applied and should then be used because of the advantage that the components can be exactly calculated.

A special application of FA in QSAR work is its use as a data pre-processing step in multiple regression [1, 50, 51] and other analyses [212]. The data matrix to be analysed then contains the biological potency to be considered as well as all molecular parameters to be checked as descriptors. After varimax rotation to obtain a simple structure the following can be deduced from the factor pattern:

(i) Only those factors make a contribution to the variable 'biological potency (log BR)' where that variable has a loading different from zero. The number of terms in a regression equation correlating log BR with the descriptor variables should therefore be equal to the number of these factors.
(ii) Only variables with non-zero loadings in the same factors where log BR also has a non-zero loading are important for the description of log BR.
(iii) Only variables with non-zero loadings in different factors may be combined as descriptors in one regression equation in order to avoid collinearities.

In this way, the computational work to evaluate the 'best' equation can be drastically reduced and, what is even more important, collinearities between independent variables which pose a serious statistical problem in regression analysis can be completely eliminated. As an example, Table 5 shows the pattern of the three significant factors obtained from a data matrix containing the inhibition of the NADH-oxidase system as well as the molecular descriptors π, π^2, σ, σ^2, σ^-, σ^{-2}, S P, E_S and E_S^2 for a series of 17 ring-substituted phenoxyacetic acids as variables. The variable pI_{50} (biological potency) is loaded only in the first and the third factors so that these alone are of importance for biological potency. It follows that in the space considered two parameters are necessary and sufficient to describe the inhibition of the NADH-oxidase system by phenoxyacetic acids. Starting from the factor pattern, one of the electronic substituent constants highly loaded in the first factor should then be combined with E_S in regression analysis, so that the following relations should be checked:

$$\begin{aligned} pI_{50} &= f(\sigma, E_S) \\ &= f(\sigma^-, E_S) \\ &= f(S, E_S) \end{aligned} \tag{26}$$

Equations with σ^2, σ^{-2} and can be excluded because not more than two terms may occur and there exists a close correlation of these variables with σ, σ^- and E_S, respectively, as follows from the factor pattern. The best equation which can be obtained in the space considered reads:

$$\begin{aligned} pI_{50} &= 0.75\sigma^- - 0.23\ E_S + 3.34 \\ n &= 17 \quad r = 0.93 \quad s = 0.249 \end{aligned} \tag{27}$$

This was checked by screening all conceivable combinations of descriptor variables by regression analysis (a strategy frequently followed in Hansch analysis). In comparison with this strategy the use of FA as pre-processing step saves more than 90% of computations and gives a clear picture of what one is doing.

Table 5. Example for factor analysis as pre-processing step in multiple regression analysis. Factor loadings (after varimax rotation) for the inhibition of the NADH-oxidase system by 17 ring-substituted phenoxyacetic acids. Only the high (significant non-zero) loadings are reported (from [1])

Variable	Loadings		
	First factor	Second factor	Third factor
pI_{50}	0.77		0.54
π		0.76	
π^2		0.70	
σ	0.92		
σ^2	0.82		
σ^-	0.93		
σ^{-2}	0.83		
S	0.78		
P		−0.76	
E_S			−0.86
E_S^2			−0.87

The general strategy of applying FA as pre-processing step in regression analysis is similar to that in principal component regression analysis (PCRA) (see next section and Fig. 3). As there, relationships of biological *y*-variables and 'factors' (patterns) inherent in the *x*-variables are investigated. The difference is that in PCRA all descriptors are assumed to be important while the intention of FA is to find out the relevant ones. With FA as pre-processing step before regression analysis some drawbacks of 'latent variable' models (low interpretability) and of 'pure' regression models (disturbing effects of collinearities) discussed in the next section are avoided. In addition, the

probability of getting chance correlations can be reduced. Unfortunately, the FA approach as described above has been seldom used in QSAR work. A variant more frequently applied [52] is to subject only descriptor variables to FA and to correlate biological activity with one highly loaded variable from each factor. In this way, collinearities are avoided but nothing is learned about how many terms the final regression equations should have and which of the molecular parameters are connected with biological potency.

Approaches based on principal components of independent variables

PCA is not only used as a method on its own but also as part of other mathematical techniques such as SIMCA classification (see section on parametric classification methods), principal component regression analysis (PCRA) and partial least-squares modelling with latent variables (PLS). Instead of original descriptor variables (*x*-variables), PCs extracted from a matrix of *x*-variables (descriptor matrix **X**) are used in PCRA and PLS as independent variables in a regression model. These PCs are called latent variables in this context.

The model of PCRA [53] reads

$$y = \sum_{k=1}^{p} b_k P_k(X) + e \qquad (28)$$

where y is a single biological potency (measurements from one biological test), $P_k(X)$* is the kth PC derived from the descriptor matrix **X** according to the general model presented in equation (8), b_k are regression coefficients obtained from correlating y as dependent with the $P_k(X)$ as independent variables, and e is residual.

As compared with multiple regression analysis, PCRA has the advantage that collinearities between *x*-variables are no longer a disturbing factor and that the number of *x*-variables included in the analysis may exceed the number of observations without violating statistical rules [54, 55]. On the other hand, contributions of simple *x*-variables can no longer be evaluated unless it is possible to identify the $P_k(X)$ with certain individual descriptors. This, however, is usually not done in PCRA so that this method, although of predictive capability, will not usually allow mechanistic interpretations. A danger of using latent variables in regression models as PCRA to represent whole matrices is that information contained in such matrices can be lost [55, 56]. This disadvantage is overcome by the more sophisticated PLS method developed by Wold and colleagues [55, 57–59].

---

* Since in this section we have to deal with both biological data matrices **Y** as well as descriptor matrices **X,** the respective components have to be differentiated by using the new symbols $P_k(Y)$ and $P_k(X)$.

The PLS method can be applied to the general case that several biological *y*-variables (measurements in several biological tests) forming a biological data matrix **Y** are to be related to a number of chemical *x*-variables combined into a descriptor matrix **X**. The problem of possible information loss with latent *x*-variables in regression models is circumvented in PLS by representing **X** and **Y** by vectors $P_k(X)$, and $P_k(Y)$ 'which *both* map **X** and **Y** *and* correlate with each other' [55]. To this end PCs are derived from both matrices in such a way that they simultaneously fulfil the following conditions:

(i) The components $P_k(Y)$ derived from **Y** yield an optimal description of **Y** (minimization of the residuals in the PC model of **Y**).
(ii) The components $P_k(X)$ derived from **X** yield an optimal description of **X** (minimization of the residuals in the PC model of **X**).
(iii) The vector pairs $P_k(Y)$ and $P_k(X)$ are maximally correlated according to

$$P_k(Y) = b_k\, P_k(X) + h_k \tag{29}$$

where b_k is a 'regression coefficient' and h_k the residual to be minimized.

Equation (29) represents a system of 'regression equations' in terms of latent variables. The optimized correlation between $P_k(Y)$ and $P_k(X)$ (condition iii) is obtained by computing these components as eigenvectors of $\mathbf{Y}^T\mathbf{Y}\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{X}\mathbf{Y}^T\mathbf{Y}$, respectively. These matrices contain the correlation structure between **X** and **Y** (expressed by $\mathbf{X}\mathbf{Y}^T$ and $\mathbf{Y}\mathbf{X}^T$) so that $P_k(Y)$ and $P_k(X)$ are connected. In ordinary PCA the PCs are extracted from $\mathbf{X}\mathbf{X}^T$ or $\mathbf{Y}\mathbf{Y}^T$, which makes them disconnected. The pairwise extraction of the $P_k(Y)$ and $P_k(X)$ from the matrices $\mathbf{Y}^T\mathbf{Y}\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{X}\mathbf{Y}^T\mathbf{Y}$ is performed by standard procedures in consecutive steps (compare section on PCA above) until all significant components as judged by cross-validation have been extracted. After this biological potencies (*y*-values) for new compounds can be predicted from their *x*-variables. In a first step the *x*-values are fitted to the PC model of the matrix **X** yielding *p* values of $P_k(X)$ and a residual. If this residual is in the same order of magnitude as the residuals of the training series compounds it may be concluded that the new compound is similar to these and its *y*-values may safely be predicted. If the residual is larger this is not true, and prediction is not possible. Next, values of $P_k(X)$ via equation (29) and the *y*-values are obtained from inserting the $P_k(Y)$ into the PC model of matrix **Y**.

It is certainly an advantage of PLS that the validity of each prediction can be checked mathematically. On the other hand, mechanistic interpretation in chemical terms is an essential part in predicting or designing new potentially active compounds, and this part is almost lost in PLS. In this respect we feel a bit uneasy about PLS and share the opinion of Hansch, who recently remarked that it would be better to 'take a poor mathematical description of the data, but keep this picture of how different effects are working' [60]. From our point of view, therefore, PLS should be adopted as a mathematically sound method and

included in the ever-growing 'QSAR tool-kit' but should not be considered to replace multiple regression analysis.

In addition, two further comments on the philosophy of Wold and colleagues and their view in comparing PLS with multiple regression analysis seem to be in order. First, Wold states that regression analysis requires a causal relation between *y*- and *x*-variables. This is a misleading statement since the term 'causal' does not fall into a mathematical category. For a natural scientist 'causal' means that an observed effect is, in fact, caused by the event considered causal. This, however, is not a requirement for regression analysis from a mathematical point of view, and no experienced QSAR researcher would accept the Hansch equation as a causal relationship simply because it is statistically significant. The second point concerns the number of *x*-variables to be included in an analysis. From a mathematical point of view the predictive power of PLS models increases approximately as the square of the number of relevant *x*-variables [55]. So the lesson seems to be: throw in more *x*-variables and you will get a better PLS model while, in multiple regression analysis, an increasing number of *x*-variables to be screened leads to the risk of chance correlations [65]. The point, however, is that only relevant *x*-variables are to be included in a PLS analysis and that, at the start of an analysis, it is usually not known what these are, so that each new x-variable may, in fact, just add noise instead of making the model more precise. Thus, with respect to the number of *x*-variables, the advantage of PLS over multiple regression only is that this number may become much larger. The problem of which variables are relevant and whether or not addition of certain further *x*-variables is meaningful or will obscure the picture is not solved by PLS but only shifted to a different level.

Finally the statement by Clementi *et al*. [190, 223]: PLS is claimed to be a more adequate method than multiple regression analysis (MRA) in many respects. In MRA the number of objects should be three times higher than the number of variables, the model dimensionality is fixed *a priori,* the model is based on the additivity of the effects and the grouping of objects is not detected.

PLS was frequently used to connect the variation in chemical structure of polypeptides to variation of their biological activities. Chemical structure of peptides were characterized by Z_1, Z_2 and Z_3 descriptors [45, 189, 209, 210] of their amino acid moieties. The PLS approach gave significant cross-validated predictions of the activities in most examples [191, 188, 189]. Sjöström *et al*. [193] analysed the information content of signal peptide amino acid sequences from proteins. Other applications are available in the literature [45, 61–64, 192, 222, 225–228]. The combined use of PLS and pattern recognition methods [229, 230] is reviewed in [183, 194].

In Fig. 3 we have summarized the characteristic features of several multivariate methods. Some recent applications of PCRA [3–5, 66] in the QSAR field are given in [67–69]. For case studies of PCA and PLS, see references cited above.

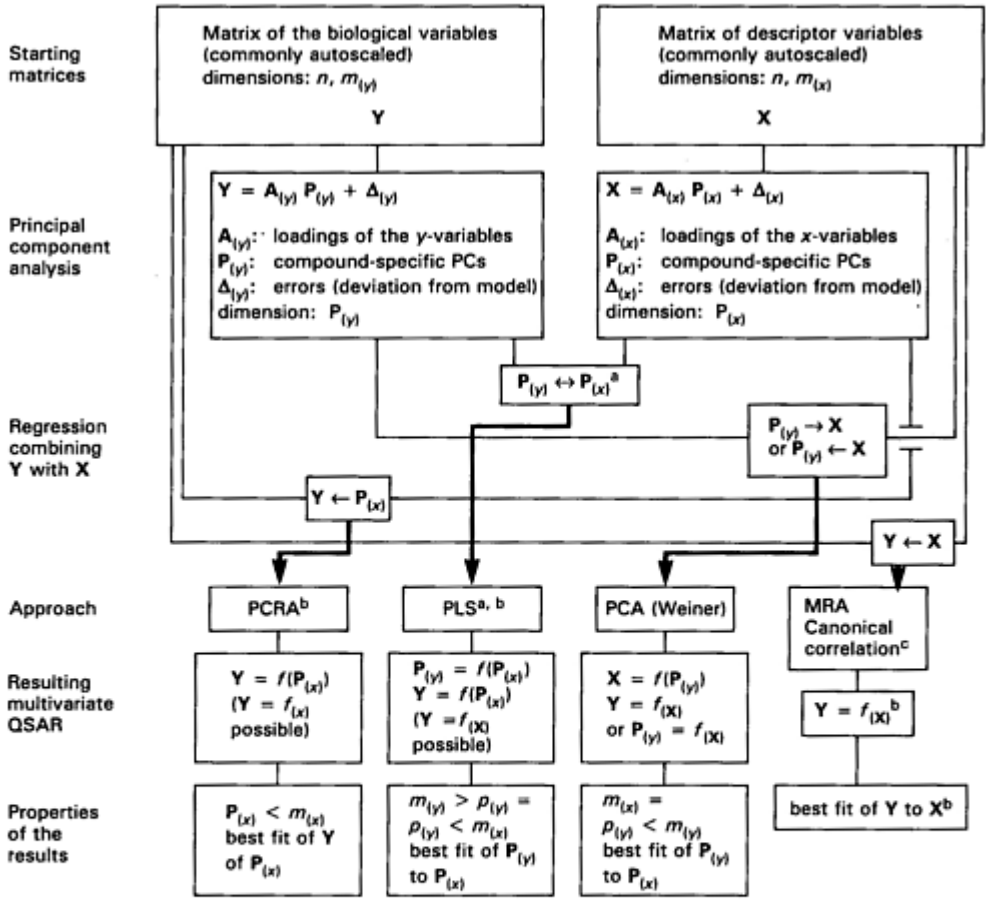
Mutlivariate regression analysis and canonical correlation [79], respectively, have been used in QSAR by Mager [3–5, 71, 72] and by Szydlo *et al*. [73]. In their analysis of pharmacokinetic parameters of different sets of compounds, Schaper and Seydel [29] applied PCA, canonical correlation as well as PLS. Cramer *et al*. [224] compared PLS with new validation techniques such as bootstrapping and cross-validation. The results indicate that PLS is particularly valuable when the physicochemical descriptors are numerous and non-orthogonal. Figure 3 is self-explanatory and need not be discussed in the text.

CLASSIFICATION METHODS

There are many cases where biological tests are not precise enough to allow the evaluation of biological potency on a continuous scale. Only semi-quantitative measurement in terms of an activity score can then be obtained, resulting in a classification of compounds with respect to their biological potency (for example, very active, active, weakly active and inactive compounds). The task then is to find an algorithm (rule of classification or classificator) which describes the distribution of compounds over the classes in terms of molecular descriptors characterizing properties of the compounds (for example, substituent constants). The algorithm is determined in a training series of compounds in such a way that the classification is reflected as consistently as possible. Once a classificator is known it becomes possible to recognize molecular or structural properties of compounds 'typical' of a particular biological effect and to predict for new compounds whether they will be (highly) active with respect to this effect. Which of these two objectives is in the foreground depends not only on the purpose of the analysis but also on the data set, the descriptors and the procedure used.

Classification methods may also be used advantageously if biological potency is on a quantitative scale in such cases where the compounds to be considered are structurally so diverse that simple linear models as, for example, regression analysis, or certain molecular descriptors based on, for example, linear free energy relationships are bound to fail or will no longer be valid in a strict sense, respectively. When applying classification methods, the rather strict requirements for the application of Linear Free Energy Relationship (LFER) — based substituent constants can be considerably relaxed since these constants are then not used to construct a quantitative linear model but only to reflect the distribution of compounds in the sense of a general pattern. It is also possible (if the right classification procedure is used) to apply substructural parameters in classification problems as descriptor variables.

Quite a variety of different methods can and have been applied in QSAR work for the evaluation of classification rules [74]. These methods may roughly be divided into two categories, namely parametric or statistical and non-parametric or heuristic techniques. While class separation in the parametric techniques is



[a]In PLS, PCA and regression steps are connected (simultaneous extraction of PC pairs p(X) and p(Y) in order to maximize their correlation; see text).
[b]Note that, in the case of a single y-variable, PCRA and PLS are based on the same general model (in PCRA, each y-variable is regarded separately), but the solutions are not numerically identical.
[c]In CCA, regression coefficients a^* (Y) and a^* (X) are calculated in order to get maximally correlated vector pairs $\mathbf{Y}^* = \mathbf{Y}a^*$ (Y) and $\mathbf{X}^* = \mathbf{X}a^*$ (X). The simple correlation coefficient between $\mathbf{Y}^*$ and $\mathbf{X}^*$ is equal to the multivariate correlation coefficient R in MRA. If the extraction of Vector pairs is continued with the residual variance not described by the first pair, an approach similar to PLS results, but the eigenvalue problems and the objectives differ from each other.

Fig. 3. Comparison of some multivariate methods used in QSAR relating a matrix of biological variables, **Y,** to a matrix of descriptor variables, **X**: PCRA (principal component regression analysis); PLS (partial least-squares method); PCA (principal component analysis according to the Weiner/Malinowski approach); MRA (multivariate regression analysis); and CCA (canonical correlation analysis).

based on statistical decision-making, the non-parametric methods start from geometrical considerations and distances in parameter space.

Parametric classification methods

Statistical class separation can be achieved in many different ways. We have extensively used non-elementary discriminant analysis [1, 75–89] according to Ahrens and Läuter [90] and have achieved good results with this method which will, therefore, be discussed in some detail. In this type of analysis classification algorithms are determined as weighed linear combinations of molecular parameters relevant for class separation. The same parameters (substituent constants, etc.) as in Hansch analysis are used, and the resulting expressions have a similar form as Hansch equations and can be interpreted along the same lines. The difference is that the coefficients do not quantitatively reflect effects of structural prop erties on biological activity, but refer to the direction and size of influences causing, for example, active or inactive derivatives. In order to simplify matters we shall restrict the following discussion to two-class problems (for example, inactive versus active compounds: classes 1 and 2) which are the most typical in the QSAR field; it must be stressed, however, that discriminant analysis as well as most of the other classification methods are also applicable to multi-class problems [1].

Two classes are separated by a discriminant function, ω, of the following general form:

$$\omega = \sum_i a_i x_i \tag{30}$$

The a_i are the weights (coefficients) of the molecular descriptor variables x_i, and the quantity ω is called the discriminant variable. Geometrically, the discriminant variable can be regarded as the axis of a one-dimensional coordinate system where the two classes occupy certain regions (see Fig. 4). In the process of discriminant analysis, the a_i are so determined that the classes are separated as sharply as possible. Once equation (30) has been derived from the data of a training series it can be predicted into which class not yet investigated compounds will fall (they can be classified). To this end the value of the discrimimant variable is computed for each compound to be classified by inserting the respective values of x_i into equation (30), and this value then gives the position on the ω-axis. If the value of ω is smaller than a certain threshold value a compound will be classified into class 1, otherwise into class 2, so that the following simple decision rule holds:

$$\omega = \begin{cases} < \text{threshold value} \rightarrow \text{class 1} \\ > \text{threshold value} \rightarrow \text{class 2} \end{cases} \tag{31}$$

In reality the separation of classes will not always be as sharp as shown in Fig. 4, and a certain region where the two classes overlap may occur. For this reason the classification of compounds is usually based on an *F*-test. A compound is classified into class *l* (*l*=1 or 2) if

$$F_{jl} \leqslant F_{1,n_1+n_2-2;\alpha} \tag{32}$$

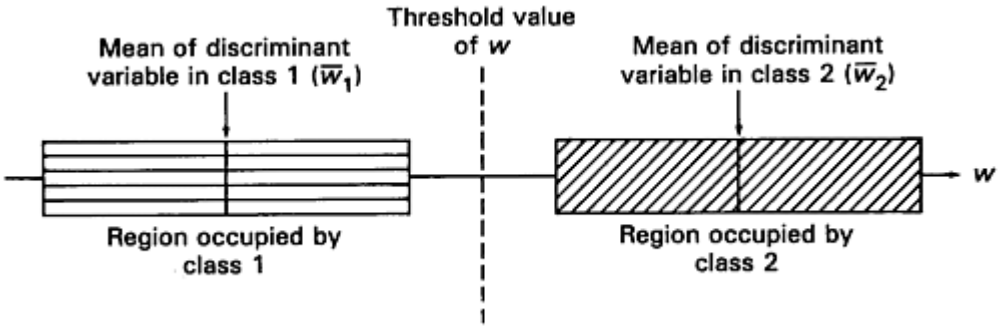


Fig. 4. Schematic representation of two classes in the one-dimensional coordinate system with the discriminant variable as axis. In many cases the separation is not as sharp as in this figure and a region may exist where the two classes overlap.

where F is taken from an F table with 1, n_1, $+n_2$–2 as degrees of freedom and α as critical probability of error. The test quantity $F_{j/l}$ is computed from

$$F_{jl} = \frac{N-2}{N-1} \frac{n_l}{n_l+1} (\omega_j - \omega_k)^2 \tag{33}$$

where N is the number of compounds in the training series, ω_j the value of the discriminant variable (calculated from the discriminant function) for compound j, and ω_l the mean value of the discriminant variable in class l (see Fig. 4).

The coefficients a_i of the discriminant function (equation (30)) are calculated from molecular parameters of the compounds in the training series by solving an eigenvalue problem according to

$$\frac{1}{N-2} \mathbf{Ha} = \lambda \mathbf{Sa} \tag{34}$$

where the matrix **H** contains the variance and covariance of variables between classes and **S** is the within-class covariance matrix, averaged over all classes:

$$\mathbf{H} = \sum_{l=1}^{2} n_l (\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\bar{\mathbf{x}}_l - \bar{\mathbf{x}}) \tag{35}$$

$$\mathbf{S} = \frac{1}{N-2} \sum_{l=1}^{2} \sum_{j=1}^{nl} (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)^T \tag{36}$$

where nl is the number of compounds in class l, is the vector of the mean values of molecular parameters in class l, is the vector of the overall mean values of molecular parameters, N is the number of compounds in the training series, and $\mathbf{x}_{ij}$ is the vector of molecular parameters of substances j in class l.

The coefficients a_i in equation (30) are the elements of the eigenvector **a** in equation (34), and the eigenvalue λ is a measure of the dicriminating power and is equal to the trace criterion Hotelling's T^2 (see below).*

For the calculation of the discriminant function only those variables (molecular parameters) are included which have been found to make a significant contribution to class separation (optimal set of variables). These are

extracted from the total set of variables considered in the analysis by means of multivariate variance analysis (MANOVA).

In MANOVA are a lot of possibilities to estimate the power of a given combination of variables to separate Q (two or, simultaneously, more than two) classes. We prefer, according to Läuter [90], the trace criterion Hotelling's T^2 with

$$T^2 = \frac{1}{N-Q}\, tr(\mathbf{H}\,\mathbf{S}^{-1}) \tag{37}$$

where **H** and **S** are the dispersion matrices defined by equations (35) and (36). The significance of class separation is tested by F-statistics; for two classes one gets (m=number of variables)

$$F = \frac{N-m}{m}\, T^2 \tag{38}$$

and the null hypothesis H_0 is rejected; that is, the class separation is significant with a critical probability of error a, if

$$F > \frac{F_{m\,(N-m-1),\,N-m-1;\,\alpha}}{N-m-2} \tag{39}$$

The optimal set among all possible sets of given variables can, in principle, be characterized by the minimal probability of error calculated from F. Commonly, stepwise procedures are used to search for optimal sets in that either the variable causing the largest increase of T^2 is added or that affecting the lowest decrease is eliminated in each step. Another F-test can then be used to calculate the contribution of each added (eliminated) variable to the class separation and, by this, to evaluate its indispensability or redundancy, respectively. With the intention to get well-interpretable results, the selected 'optimal' set should contain only indispensable variables rather than reflect the absolute minimum of the critical probability of error.

Another measure for the descriptive power of a discriminant function is the error of reclassification obtained as the percentage of compounds of the training series assigned to the wrong class when classified by means of the discriminant function. In the ideal case, all compounds are reclassified correctly (error of reclassification equal to zero), but in reality an error of reclassification up to about 10–15% (for not too small training series) is still acceptable. A high descriptive power, however, does not necessarily imply that the predictive power is also high. Predictive power can be checked by simulated predictions. If the

---

* In multi-class problems more than one discriminant function results (Q–1 for Q classes). A favourable aspect of equation (34) and, thus, of non-elementary discriminant analysis is that the separating power of the discriminant functions decreases as the eigenvalues decrease in the order 1> 2…> l…> q−1. This makes it possible to eliminate those last discriminant functions which fail to contribute significantly to class separation, thereby diminishing the dimensionality of the discriminant problem.

training series is large enough, some of the compounds can be excluded from the calculation of the discriminant function and classified as 'unknowns'. Another approach is to sue the so-called leave-*n*-out technique where groups of *n* compounds are alternatively left out of discriminant analysis and then classified with the aid of the discriminant function obtained from the remaining compounds. The discriminant function can be accepted if it remains stable throughout this procedure and if a good agreement between determined and real classification is obtained.

Using several modifications discriminant analysis and related statistical classification methods have been widely and successfully applied in the QSAR field by several workers [75–89, 91–104]. In our work we have always preferred to use discriminant analysis (or another adequate classification method) for data of low or unknown precision rather than to force such data into a regression model. A discriminant function (in multi-class problems, the first discriminant function) can often be interpreted analogously to a Hansch equation in mechanistic terms.

A simple example of applying discriminant analysis to a pharmacological problem can be provided by a QSAR investigation on the inhibition of dopamine β-hydroxylase (DBH) *in vitro* and by the antihypertensive activity of a series of picolinic acid derivatives (structure III, Fig. 5, p. 82) [105]. Inhibition of DBH blocks the synthesis of noradrenaline from dopamine, so that DBH inhibitors are potential antihypertensive agents. The picolinic acids were found to be active both as DBH inhibitors *in vitro* as well as in decreasing the blood pressure in male Wistar rats with renal hypertension. The question then arose whether the antihy pertensive activity observed *in vivo* was, in fact, related to DBH inhibition. Unfortunately, the precision of the *in vivo* data was so low that the compounds could only be divided into the classes 'inactive' (class 1) and 'active' (class 2) with respect to their effect on blood pressure. Thus, it was impossible to directly correlate *in vitro* and *in vivo* activity or to apply Hansch analysis to the latter. The problem could be solved by applying discriminant analysis to the *in vivo* data with the assumption that *in vivo* activity depends on DBH inhibition plus the transport to the site of action. This leads to the selection of pI_{50} values for DBH inhibition *in vitro* and π as measure of hydrophobicity as descriptor variables in discriminant analysis. Multivariate variance analysis shows that both variables significantly contribute to class separation, leading to the discriminant function

$$\omega = 0.383\ pI_{50} + 0.924\ \Sigma\pi \qquad (40)$$

with Hotelling's T^2=1.085 and a critical probability of error of 0.005. With equation (40) 19 out of the 21 compounds considered are correctly reclassified. Since one of the two misclassified compounds (5–OH picolinic acid) is on the borderline with respect to its *in vivo* activity and could already have been considered inactive, the separation of classes is quite satisfactory. With $\omega - 2 > \omega - 1$ it follows from equation (40) that compounds with antihypertensive activity *in vivo* must have the higher values of ω and, thus, high inhibitory potency

against DBH *in vitro* (pI_{50}), so that the connection between enzyme inhibition and antihypertensive potency *in vivo* can be regarded as fairly well established. The $\Sigma\pi$ term in equation (40) indicates that for an effective transport to the site of action a certain hydrophobicity is required.

Applying regression analysis to the *in vitro* data yielded a Hansch equation with σ, L and B_4 as independent variables. Predominant is the electronic effect, so that a statistically significant relation can already be obtained with σ alone:

$$pI_{50} = -1.32\ (\pm 0.60)\ \Sigma\sigma + 5.12\ (\pm 0.22)$$
$$n = 18 \quad r = 0.692 \quad s = 0.540 \tag{41}$$

Consequently, a significant discriminant analysis can also be obtained when the variable pI_{50} inequation (40) is replaced by σ ($\omega{-}1{=}1.537$, $\omega{-}2{=}{-}0.117$):

$$\omega = -0.139\ \Sigma\sigma + 0.990\ \Sigma\pi \tag{42}$$

Equation (42) performs almost as well as equation (40) in reclassifying the compound of the training series. The lesson to be learned from both equations is that the discriminant function (equation (42)) correctly reflects that active compounds should have electron-releasing substituents, which is to be attributed to enzyme inhibition. Without knowledge of the enzyme data equation (42) alone would have been sufficient for the selection of further analogues potentially active *in vivo*. Discriminant analysis is contrasted with different methods (for example, cluster significance analysis, multiple linear regression analysis, linear learning machine, simplex optimization, *K*-nearest-neighbours analysis, PC plots) by several authors [182, 195–197], and studies were performed to determine the influence of chance factors in discriminant analysis [202–204].

A prerequisite for applying discriminant analysis is that the classes to be separated from well-defined clusters (symmetric case). This, however, is not always the case. While in a set of compounds acting via the same mechanism physicochemical requirements for being active are usually well defined, reasons for being inactive may be extremely diverse. As a consequence, inactive compounds can be scattered irregularly around the cluster representing the class of active compounds. This situation has been referred to by Wold as the asymmetric case, where discriminant analysis (and also such hypersurface separators as, for example, the linear learning machine to be discussed below) cannot be applied. A method which can still be used here is SIMCA classification, worked out and extensively used by Wold and colleagues [9, 14, 106–113, 198] and by other authors [199, 200, 229].

The starting point in SIMCA is the same as in discriminant analysis: a training set of compounds classified with respect to their activity is characterized by a number of mostly continuous descriptor variables. With these variables (standardized before use) a PCA is performed for each class separately as outlined above. All that has already been discussed with regard to the properties of loadings, PCs and the problem of interpreting QSARs based on latent variables also applies to SIMCA.

Each compound of the training series can be represented by a point in the *m*-dimensional space of variables. If the descriptors selected are suitable to model a class of activity, the compounds of this class form a cluster of mostly less than *m,* namely *p,* relevant dimensions (in the simplest case a line). These *p* dimensions are spanned by the orthogonal coordinate system of class-specific principal axes. Class modelling means that the cluster of objects forming a class in the variable space is projected down on this system of axes. Class separation, on the other hand, is achieved if the clusters expressed by PC models (of, possibly, different dimensions) are distinct and sufficiently distant from each other.

The goodness of fit of a SIMCA model (and also the importance of single variables appearing in such a model) can be judged by two criteria: modelling and discriminatory power. Modelling power characterizes the portion of within-class structure reflected by the respective PC model, and discriminatory power is a measure of how effective the classes are separated in the variable space. Both criteria can be quantitatively expressed in terms of special mathematical quantities. In building a SIMCA model only such variables are included which make a significant contribution to its goodness. In other words, descriptor variables of low modelling and discriminatory power are eliminated so that, similar to discriminant analysis, the final result is based on a selected set of variables.

Classification of the compounds of the training series and of new compounds is based on distances between the point representing a compound in the *m*-dimensional space of variables and the *p*-dimensional hypersurfaces of the PC class models. A compound will be assigned to that class for which this distance is smallest. For this purpose, the compounds must be projected to the system of principal axes of each model, and the decision is then made by means of a special *F*-test.

This classification also works in the asymmetric case mentioned above: compounds falling in the range of the defined PC model of active compounds are assigned to this class, while compounds outside that range are assigned to the diffuse ‘all others’ class, for example, predicted to be inactive. A case in point is a SIMCA study on carcinogenic 4-nitroquinoline-*N*-oxides [108], where only the class of active carcinogens showed a substantial structure whereas the inactive compounds did not form a homogeneous class.

Non-parametric classification methods

The classification methods discussed in the previous section are all based on statistical tests which require normal data distribution. If this condition is not fulfilled the so-called ‘non-probabilistic’, ‘non-parametric’ or ‘heuristic’ classification techniques must be used. These techniques are also frequently referred to as pattern recognition methods.* They are based on geometrical and not on statistical considerations, starting from a representation of the compounds

of the training series as points in the space spanned by the descriptor variables. Typical of the application of heuristic methods in QSAR work is the use of topological descriptors derived from topological features. For this reason, a brief discussion of such descriptors is appropriate (see also section on topological pattern finders).

Topological features are two-dimensional fragments or patterns occurring in the two-dimensional structures of the compounds to be analysed, and topological descriptors describe the occurrence of these features in the compounds thus characterizing their structure. Topological descriptors can advantageously be used if, in heterogeneous and large data sets, structural variation becomes so extensive that it is no longer feasible to select substituent constants or related physicochemical quantities to describe molecular properties. Such problems may arise much more frequently than is commonly believed. Typical cases are where data from large databases, from mass screening or data collections from the literature are to be investigated.

With topological descriptors one is, of course, asking less of the data than with, for example, extrathermodynamic parameters, but they are the only possibility for large data sets of high structural diversity, and the results can still effectively be used for guiding chemical synthesis and biological testing. There are many types of such descriptors which all have in common that they can be derived directly either from two-dimensional chemical structures or from connection tables; excellent reviews are presented, for example, by Stuper *et al*. [74] and by Bawden [114]. Following certain rules, features can be computer-generated algorithmically starting from single atoms or bonds as centres, leading to atom-centred and bond-centred fragments. Such fragments are defined in terms of concentric areas of structure surrounding each atom (except hydrogens) or bond at different levels of complexity and specificity, with information about atom and bond types, non-hydrogen connections, etc. in this area. Very commonly used are so-called augmented atom fragments that describe atoms with their next neighbours.

Another possibility is to define a library of features (substructures) representing, for example, rings, functional groups, heteroatoms or other centres which may be important for drug-biosystem interaction, and distances between such centres (paths). Although linear notations as used for computer storage and retrieval of chemical structures, for example WLN [115], can serve as a source of fragment codes of this type, it seems advisable to select a more problem-oriented library on the basis of ‘common sense’ and the experience of the researcher [116, 117]. To this end a basic library of potentially active centres such as atoms or groups of atoms (functional groups) likely to be involved in

* Note that the definition of what pattern recognition methods are is not used in a unified manner. For example, methods as outlined in the previous section are also sometimes summarized under this heading. The historical definition is as used in this section.

drug-receptor interactions via van der Waals or other forces may be set up and used to create the concrete features for each problem via a set of rules in an open-ended way. Cases in point are the SSFN and the DCAM systems based on pre-defined 'descriptor centres' and distances between them [118–120]. Additional descriptors can be defined to characterize the environment of the substructures [74] (environmental descriptors).

An even better strategy is to code not only for the type of substructures but also for the molecular region where they occur, whenever this is feasible. This can be done with the help of a template generated by superimposing all structures of the training series in such a way that those features considered as potentially important (descriptor centres) are unambiguously and adequately represented. As a result an artificial reference diagram is obtained that may be considered as a hypothetical parent of all compounds. Descriptors are then derived by comparing the structure of each compound with this template. Such approaches have, for example, been used by Cammarata and Menon [30, 31], by Henry and Block [94, 95] and in our laboratory [121–125]. Philosophically, they are similar to the 'hyperstructure' of the DARC/PELCO system [126] and the MTD technique of Simon and colleagues [127].

The definition of features is the crucial step in an analysis based on substructural descriptor variables. If the features are in error the result of an analysis will be misleading. Chemical experience and intuition may play an important part here, and an interactive process in the generation, processing and selection of features is to be recommended [128]. In doing so, correlations between features must be adequately considered; otherwise, wrong conclusions may be drawn [129, 130]. Problems of redundancy may arise if features of different complexity are to be used in the same analysis [131, 132]. Sometimes substructural descriptors are used together with some physicochemical parameters such as molar refractivity or log *P*. This can be advantageous in treating complex problems but has the danger that interpretation of the results may become extremely difficult due to complicated relationships between the two descriptor sets.

Once the features are defined each compound is characterized by a set (a vector) of descriptor values (one for each feature). The simplest way to define such values is to characterize the presence or absence of features by binary indicator variables according to

$$X_{ij} = \begin{cases} 1 \text{ if the } i\text{th feature is present in the } j\text{th compound} \\ 0 \text{ if not} \end{cases} \tag{43}$$

Another possibility is to use occurrence numbers of features in the compounds:

$$X_{ij} = \begin{cases} k \text{ if feature } i \text{ occurs } k \text{ times in the } j\text{th compound} \\ 0 \text{ if feature } i \text{ is not present in the } j\text{th compound} \end{cases} \tag{44}$$

Pattern recognition methods have been successfully applied to a variety of fields (for mathematical details see, for example, [74, 133–138]), and some reviews on the application of such methods in drug design are also available [74, 117, 139, 140]. Presently, the most frequently used technique in the QSAR field seems to be the linear learning machine (LLM) according to Nilsson [133]. LLM starts from a presentation of the compounds of the training series in the space of descriptor variables. If the descriptors selected are meaningful and connected with biological potency the members of different classes will be found in different regions of feature space. Once these regions are known a new compound can be classified by observing in which region it would fall. To this end a hypersurface separating the classes is derived in LLM in the form of a simple decision rule:

$$w_0 + \Sigma_i\, w_i x_i = \begin{cases} > 0 \text{ for compounds of class 2} \\ < 0 \text{ for compounds of class 1} \end{cases} \qquad (45)$$

The term w_0 and the weights w_i assigned to the variables x_i are so determined that equation (45) is satisfied in the best possible manner, that is, for a maximum number or, ideally, for all compounds of the training series. The w_i are the elements of a vector exactly perpendicular to the separating hypersurface (normal vector) and thus describe the direction of the surface. Its position in space is fixed by the additional element w_0, which can be considered as the weight for an additional variable with a value of one for all compounds.

The evaluation of w_0 and the w_i is performed by a process which is called supervised learning and which will not be discussed here. It starts from an initial estimate which is then improved in such a way that the separation of classes becomes optimal (the computer learns how to separate the classes in a supervised manner). Prior to determining the weights connected with descriptor variables, an optimal set of variables is found in a feature selection process analogously to, for example, discriminant analysis. In this process variables not contributing to the separation of classes are eliminated. This can be done in various ways. One standard procedure is based on a leave-one-out technique. Each compound of the training series is once left out and the classificator is computed using the remaining compounds. In this way, for a series of n compounds, n sets of w_i are obtained. The more important a variable x_i is for class separation the more stable should be the value of the corresponding weight throughout this procedure. Thus, variables with a high variance of the corresponding weight will be eliminated and those with a low variance retained. Usually, stepwise (forward and/or backward) procedures are used here.

Once the classificator (w_0 and the w_i) is known new compounds can be classified by inserting their values of x_i into equation (45). Reclassification of the compounds of the training series is performed in the same way. The separating power of the classificator is judged in the usual way by the error of reclassification and the error of classification for a test set of compounds or for a simulated prediction using the leave-n-out technique as already discussed in connection with discriminant analysis. In two-class problems sign and magnitude

of w_i may permit certain conclusions with respect to the importance of individual features for classification. A high positive value for a given w_i, for example, indicates that the presence of the corresponding feature argues in favour of assignment of an object to class 2 and thus may be important for the biological activity considered. As a consequence it should, in principle, be possible to interpret LLM classifications in chemical terms and to use such information for the design of new compounds. Experience has shown, however, that this is usually not possible since LLM classificators with substructural descriptors are frequently far too complex to admit of any meaningful evaluation along such lines. Possible reasons for this will be discussed later.

An alternative to LLM which has also been applied in QSAR work is the *k*-nearest-neighbours method (KNN) [141, 142] which is based on the consideration of distances between objects in feature space and their class membership. An unknown object (compound) is classified by comparing its position in feature space with that of its *k* nearest neighbours. For this purpose the distances between the object and all objects of the training series are calculated (usually, the Euclidian distance is used here). The *k* nearest neighbours receive a vote for the class they belong to, and the unknown object is then assigned to that class which has accumulated the most votes. For vote weighing, various methods are available. For a two-class system the decision can be made very easily according to

$$\sum_{i=1}^{k} \frac{V_i}{d_i^2} = \begin{cases} < 0\text{: assignment to class 2} \\ > 0\text{: assignment to class 1} \end{cases} \tag{46}$$

where d_i denotes the Euclidean distance and V_i is defined as

$$V_i = \begin{cases} 1 \text{ for class 2} \\ -1 \text{ for class 1} \end{cases} \tag{47}$$

The KNN method has the advantage that it is applicable in cases where data are not separable by a linear decision surface and performs best on data that contain a high density of points. Since no classificator is explicitly stated, an interpretation of the results in structural terms is not possible.

LLM and KNN have quite frequently and not without success been applied in QSAR work [1, 2, 74, 117] to a variety of problems. Examples may be found in [74, 94, 116, 130, 143–160] (papers criticized by others or of doubtful validity not included). Usually, LLM performs somewhat better than KNN. In most cases a good separation of classes was achieved, and some of the classificators derived were shown to possess predictive power in simulated or real predictions. As already mentioned, however, the classificators obtained are very complex and admit of no straightforward interpretation in terms of chemical structure. For this reason, they cannot usually aid in the design of new drugs or in optimizing structures with respect to a desired activity. Their primary objective, then, is the pre-selection of already planned substances in screening programs in order to increase the incidence of biologically active compounds. Another method

suitable and successfully applied for that purpose is the heuristic index technique introduced by Hodes [129, 161–163] and further developed by Tinker [164, 165] which is based on the calculation of activity and inactivity indices from substructural features. Pajeva *et al*. [231] used a modification of the method of Hodes.

A similar computer-automated structure evaluation (CASE) program for the evaluation of biophoric and biophobic substructures was also developed by Klopman [166]. The program requires as input the molecular structures of the compounds to be evaluated, together with the experimentally measured values of the biological activity to be studied. CASE proceeds to 'fragment' each molecular structure into units of 3–10 heavy atoms together with their associated hydrogens. Biologically inactive compounds give rise to fragments that are biophobes, whereas biophores are derived from active compounds. Biologically relevant fragments expanded (differing in one position along the connectivity path) from the parent fragment are also generated. No provisions have been made within the program to account for chirality and *cis/trans* isomerism. Once all of the fragments are generated a binomial distribution is assumed, and any considerable deviation from a random distribution of a fragment among the active and inactive molecules is indicative of potential significance to the biological activity being studied. The program is capable of qualitatively distinguishing biologically active from inactive compounds based on the presence or absence of statistically relevant fragments. In order to acquire a quantitative estimation of the potency or degree of biological expression, a subset of potential descriptors must be chosen (by the 'family' method, for example) and submitted to a multivariate linear regression analysis. As an initial starting point, a fragment (either biophore or biophobe) is chosen which can classify the greatest number of active and inactive compounds. Subsequent fragments are chosen so as to account for the remaining variance. This procedure culminates in a family of largely uncorrelated fragments which serve as potential variables for the method of linear least-squares regression. Fragments are incorporated within a regression equation in a forward stepwise manner until no significant improvement is observed between calculated and actual values. The regression coefficients of the corresponding fragments are a measure of the activating/inactivating contribution made to the biological activity and are used to calculate the potency of individual compounds.

The CASE program has already been successfully applied to identify structural features responsible for the carcinogenicity [166, 215], mutagenicity [213–216] and for other different biological activities [217–221].

A new statistical method, cluster significance analysis (CSA), was presented by McFarland *et al*. [196, 233] to evaluate the statistical significance of an apparently clustered group in the graphical display of structure-activity data. CSA operates by assessing the tightness of the cluster of active compounds and determining the probability that a cluster is as tight or tighter than might have arisen purely by chance. The method is limited to analysing two-group

classifications. Applicability of CSA is shown in different examples [196, 197, 234].

Classification methods based on substructural descriptors and such index techniques have also been applied in the field of chemical hazards such as toxic effects [167, 168], carcinogenicity [116, 144, 149, 150, 156, 167, 168] and mutagenicity [164, 165, 168, 232] not without success. Attempts are made to provide automated predictions of such hazards for chemical compounds of interest, which is certainly helpful in setting priorities for biological testing. It is important, however, to realize that such predicted data can never be taken for granted and must not be allowed to replace experimental measurements [169].

While LLM and KNN are classical mathematical techniques applicable and applied to a variety of fields, the STRAC procedure introduced by Rozenblit and Golender was especially developed for QSAR purposes in the framework of their logico-structural approach [118, 120, 170] to drug design. STRAC also starts from a training series of compounds divided into the classes 'active' and 'inactive' and is designed for lead optimization. Substructural features are defined as substituents in their position of substitution so that STRAC is limited to more or less homologous series. A feature is considered discriminating if the probability that a compound containing this feature belongs to a certain class is greater than a certain threshold value. This probability is computed from the occurrence numbers of features in the classes, and the threshold is found by an interactive trial-and-error procedure based on the already mentioned leave-one-out technique. In many cases the discriminating features found in this way are not sufficient to classify all compounds of the training series. New discriminating features are then derived by logical operations in an interactive process until a sufficient separation of the two classes is achieved. The features so obtained are either typical of the active ('activity features') or of the inactive ('inactivity features') analogues. New compounds can then be classified according to their predominant features. A compound is rated active if

$$V_a - V_{in} > \eta \tag{48}$$

and inactive if

$$V_{in} - V_a > \eta \tag{49}$$

where V_a and V_{in} are the occurrence numbers of activity and inactivity features, respectively, in this compound and η is a certain threshold value. This technique has successfully been applied to several problems as, for example, to antimicrobial alkylaminosilanes [118] or to anticonvulsive indandione-1,3 derivatives [118, 170]; in both cases, predictions could be experimentally verified. Another part of the logico-structural approach is the program ORACLE, which was developed for the purpose of guiding the screening of compounds in a battery of biological tests [119, 120]. The basis of ORACLE is a structure and pharmacological activity file including about 9000 compounds and 55 major types of pharmacological activities. Substructural descriptors are generated using a library of substructures, rings and certain distance measures.

The compounds of the data file are divided into 55 classes according to the type of their biological activity. Each class is then separately compared with the rest of the compounds contained in the remaining 54 classes in order to find descriptors that represent the particular type of activity exhibited by the class under investigation. A descriptor is considered to represent a particular activity feature if its presence in this class is not a chance event. Sets of descriptors are found in this way which are regarded to be typical of each respective class. Once these sets are known, the type of pharmacological activity to be expected for a new compound can be predicted so that it becomes possible to decide for which of the 55 pharmacological effects considered a new compound should be screened. This method will fail if a pharmacological effect can be produced via different mechanisms of action. Nevertheless, ORACLE has successfully been applied. The system correctly recognized the presence of earlier established activities for the majority of compounds in the database, and a number of unknown activities could be predicted for many compounds.

Methods of the type mentioned in this section make the tacit assumption that the contribution of a given substructural unit to biological activity is a consistent factor. This, however, is frequently not true where large and structurally diverse data sets for relatively unspecific biological effects are considered, since in such cases the biological mechanism of action is likely to be different for different subsets of compounds. In addition, complex intramolecular interactions not accounted for by simple descriptors may strongly influence the role of certain substructures. Such interactions may have the consequence that a particular substructure is favourable for biological activity in one particular chemical environment and unfavourable in another. In such a situation a class of active compounds is not necessarily a homogeneous and well-defined entity but may consist of different clusters with quite different characteristics. None of the methods mentioned so far takes this fact into account. The very complex classificators usually obtained from an LLM analysis may well be due to such a situation; they would then present a mixture of and a compromise between several classificators needed to separate different subsets of compounds in the active class from the inactive compounds. Some of these problems can be avoided by methods which may be termed ‘topological pattern finders’ and which will be discussed in the next section.

TOPOLOGICAL PATTERN FINDERS

A topological pattern in the QSAR context is regarded as an ensemble of substructural features that is characteristic of a group of compounds possessing a desired biological property (for example, high activity) but absent from compounds devoid of this property. This ensemble, which may also be called a ‘topological pharmacophore’, is always to be considered in its entirety, so that the assumption that the contribution of a given substructural unit to biological activity is a consistent factor does not apply here. Furthermore, several

topological pharmacophores can be derived from the same data set if the compounds do not act via a uniform mechanism of action. It was for these reasons that we started to develop topological techniques of the pattern-finding type [1, 2, 121–125, 171, 174, 175, 176]. Two of them, the methods LOGANA and LOCON [171, 176] will be discussed in some detail, together with a simple example. Finally, an interactive tool for finding substructural entities—EVAL—will be described [174, 175].

The philosophy of LOGANA and LOCON is to make use of the data-handling capacity of computers while maintaining an optimal impact of the researcher's professional skill and experience, with no requirements of any special mathematical knowledge. The computer is used only to digest large amounts of data and transform the information inherent in them into a manageable form. While this part is fully formalized, the real decision making and all conclusions to be drawn for further synthesis or testing are completely left to the researcher. This implies that neither assumptions regarding probabilities or data distribution nor mathematical models are necessary.

Both methods are substructural descriptors defined according to equation (43), and they are most effective for template-derived features. Physicochemical properties such as hydrophobicity can also be included to a certain extent. This is done by selecting region(s) (or thresholds) of the corresponding molecular parameters (for example, log P or π) and defining x_{ij}=1 if the jth compound falls into the ith region, and X_{ij}=0 if not. The set of descriptors can be extended by the logical operations 'or' (disjunction, symbolized by a sign) and 'not' (negation, symbolized by a '~' sign). In a disjunction similar features to be regarded as bioisosteric are merged into a new feature. A disjunction of two features, x_i and would then read:

$$x_{ij} \text{ V } x_{kj} = \begin{cases} 1 \text{ if features } i \textit{ or } k \text{ are present in the } j\text{th compound} \\ 0 \text{ if not} \end{cases} \tag{50}$$

In a negation, the definition presented in equation (43) is reversed, meaning that the absence of the ith feature is considered important and used as a new feature:

$$\sim x_{ij} = \begin{cases} 1 \text{ if features } i \text{ is } \textit{not} \text{ present in the } j\text{th compound} \\ 0 \text{ if it is present} \end{cases} \tag{51}$$

The basic operation in LOGANA and LOCON is the stepwise and interactive construction of more complex features from the simple features by means of the logical operation 'and' (conjunction, symbolized by a sign). A conjunction of two variables x_i and x_k is defined as

$$x_{ij} \Lambda\ x_{kj} = \begin{cases} 1 \text{ if features } i \text{ and } k \text{ are present in the } j\text{th compound} \\ 0 \text{ if not} \end{cases} \tag{52}$$

Clearly, a conjunction represents a more complex chemical entity than the single variables which will usually be present in a smaller set of compounds. A set of compounds having the structural features expressed by a conjunction in common

will be called an 'object group'. As the conjunctions grow by adding a new variable in each step of the LOGANA or LOCON procedure the object groups become smaller. In other words, extending a conjunction results in an elimination of compounds. The conjunctions are constructed in such a way that the 'right' compounds are retained and the 'wrong' compounds eliminated. In LOGANA, 'right' means active and 'wrong' inactive while, in LOCON, 'right' labels the most active and 'wrong' the less active plus inactive compounds. The program LOGANA starts from a division of the training series in the classes 'active' and 'inactive', thus allowing the analysis of more or less crude data as from biological mass screening or sampled from different sources. LOCON, on the other hand, is designed to deal with continuous biological activity data in order not to lose information in those cases where the data are precise enough to allow comparison on such a scale.

LOGANA searches for combinations of features (conjunctions) that are present in as large *groups** as possible of active compounds (class 2) and absent or nearly absent from inactive compounds (class 1). In other words, those conjunctions are regarded best which yield object groups (see above) containing high numbers of class 2 compounds and the smallest possible number of class 1 compounds. As a measure of this property a simple quality criterion T is used:

$$T = \frac{N_2 + (n_2 - n_1)}{N_1 + N_2} \tag{53}$$

with T normalized to $0 \leq T \leq 1$. Here, N_1=number of compounds in class 1, N_2 = number of compounds in class 2, n_1=number of class 1 compounds in the object group belonging to the respective conjunction, and n_2=number of class 2 compounds in the object group. The evaluation of conjunctions is performed in a stepwise procedure where one more variable is added in each step using T as a selection criterion. In the first step all variables (including negations) are arranged in the order of descending T values and the best k variables (k adjustable) are selected as a starting set for the next step. Conjunctions of each of the variables of this starting set with all other variables one at a time are then formed according to equation (52), and the best k of these conjunctions are then transferred to the next step as a new starting set. Each conjunction of this set is then again combined with all variables (one at a time) by the logical operation 'and', and the best k of the resulting conjunction now comprising three variables are again transferred to the next step. The procedure is continued until a preset number of steps has been performed. Disjunctions can be introduced by a special option at any stage in the programs.

The conjunctions searched for in LOCON are those which are characteristic of the most active compounds. This means, in other words, that the corresponding object groups show a mean value of biological potency as high as possible. The quality criterion used here to characterize conjunctions (or variables) is defined as

$$D = (MS - MO)\left(1 + \frac{(NS)\,(W)}{(NO)}\right) \tag{54}$$

where *MS*=biological activity mean within the objects group of the respective conjunction, *MO*=overall mean (whole training series), *NS*=number of compounds in the object groups, *NO*=number of compounds in the training series, and *W*=adjustable weight (≥ 0). The higher the value of *D* becomes the more characteristic is the corresponding conjunction of the high-activity compounds. With the weight *W* the influence of the size of the object group on *D* can be adjusted.

In the first step of LOCON the variables (including negation) are arranged in order of descending *D* values, and the best variables are selected as a starting set for the next step. Conjunctions of these best variables with all other variables (one at a time) are then formed and the *k* best (*k* adjustable) conjunctions are displayed. From these the operator now selects those which he considers the most promising by inspection and transfers them as a starting set to the next step, where they are again combined with all variables (one at a time). The best *k* of the resulting conjunctions now comprising three variables are again displayed for each conjunction of the starting set, and a new starting set is then selected by inspection and extended by one variable in the next step. The procedure is continued until the quality of the resulting conjunctions can no longer be improved.

Together with each conjunction 'accompanying variables' are given. Accompanying variables have the property that they can be added to the respective conjunction without eliminating compounds from the corresponding object group. They must, of course, be considered when evaluating the results since they characterize structural features also present in all compounds of the object group.

The best conjunctions obtained from a LOGANA or LOCON analysis can directly be retranslated into two-dimensional chemical structures. In doing that and when interpreting the results the whole picture, including the development of the conjunctions, the object groups described by them, outliers (compounds which should have been removed from the object groups but are not), accompanying variables, etc., must be viewed. For linguistic simplicity it is tempting to label features appearing in the best conjunctions as 'favourable' (or, in the case of negations, as 'unfavourable'). Even though this can be adopted for practical purposes it must not be forgotten that it is, strictly speaking, not correct. All that can in fact be concluded is that features represented in the best conjunctions are typical of the active (LOGANA) or the most active (LOCON) compounds, but only in the context of these conjunctions and insofar as the training series is representative of chemical compound space.

---

* It is this strategy that makes LOGANA basically different from classification methods, the latter aiming at a complete separation of classes. By considering subgroups it becomes possible to account for different mechanisms of action. Another point is that simple topological descriptors by their very nature frequently afford the formation of subgroups in heterogeneous data sets [125, 171].

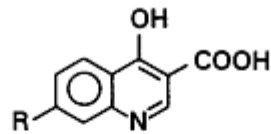
We have successfully applied LOGANA and LOCON in a number of cases [122–126, 171] and use both methods routinely. As a simple example we shall present here results [122] obtained for a series of fungicidal carboxamides as inhibitors of succinate dehydrogenase (experimental data from [172]). The series comprises 89 compounds of the general structure presented in Fig. 5. Structural variations are so extensive that QSAR methods of the Hansch or Free-Wilson type do not lead to meaningful results. Tables 6 and 7 summarize the topological descriptors used in the analysis.

Table 6. Definition of descriptors x_i used in LOGANA: x_i=1 if feature i is present[a]

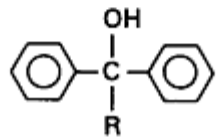
Feature	Definition	
*Z*1	–CONH–	in *Z*
*Z*2	$–CONCH_3–$	in *Z*
*Z*3	–COOH	in *Z*
*Z*4	$–COCH_2–$	in *Z*
*Z*5	–CSNH–	in *Z*
*Z*6	$–SO_2NH–$	in *Z*
X1–*X*14	*X*=structure A – O	
*X*15	– C = C – C = O	in *X*
*X*16	– O – C = C – C = O	in *X*
*X*17	– S – C – C = O	in *X*
*RX*1	$–CH_3$	in *(RX)o*
*RX*2	$–C_2H_5$	in *(RX)o*
*RX*3	–I	in *(RX)o*
*RX*4	–Br	in *(RX)o*
*RX*5	–Cl	in *(RX)o*
*RX*6	–OH	in *(RX)o*
*RX*7	–H	in *(RX)o*
*RX*8	–phenyl	in *(RX)o*
*RX*9	–F	in *(RX)o*
*RX*10	$–NH_2$	in *(RX)o*
*RX*11	$–CH_3$	in *(RX)o, m, p*
*Y*1	–phenyl	in *Y*
*Y*2	–cyclohexyl	in *Y*
*Y*3	–butyl	in *Y*
*Y*4	–H	in *Y*
*Y*5	–methyl	in *Y*
*Y*6	no ring	in *Y*
*RY*1	$3'–CH_3$	
*RY*2	$2'–CH_3$	
*RY*3	$2'–C_6H_5$	

Feature	Definition
*RY*4	3′3′–OCH_3
*RY*5	3′–Cl
*RY*6	4′–Cl
*RY*7	4′–CH_3
*RY*8	2′–C_2H_5
*RY*9	2′–Cl
*RY*10	3′–OCH_2O–4′
*RY*11	2′–OCH_3
*RY*12	4′–OCH_3
*RY*13	4′–Br
*RY*14	6′–CH_3
*RY*15	6′–C_2H_5
*RY*16	R_y with $\pi > 0$

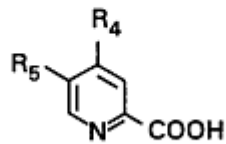
[a]Symbols as in Fig 8.



Structure I: 7-substituted 4-hydroxyquinoline-3-carboxylic acids



Structure II: diphenylaminoalcohols



Structure III: picolinic acids

Fig. 5. General structure of the carboxamides.

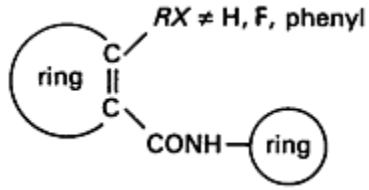


Fig. 6. Structural pattern resulting from conjunction (55).

Table 7. Additional features derived for the application of LOCON: x_i=1 if feature *i* is present

*X*18	*X10*∨*X*11∨X12∨X13∨*X*14
*X*19	*X2*∧RX11
*X*20	*X4*∧*R*X10
*X*21	*X4*∧RX11
*X*22	*X6*∧*RX*11
*X*23	*X7*∧*RX*11
*X*24	*X9*∧*RX*11
*RX*12	*RX*7∨*RX*9
*RY*13	*RX*3∨*RX*4∨*RX*5∨*RX*9
*RY*17	*RY*2∨*RY*3∨*RY*8∨*RY*11
*RY*18	*RY14*∨RY15
*RY*19	*RY6*∨RY7∨*RY*12∨*RY*13

For LOGANA, we shall consider the following two classifications:
Problem I: active versus inactive compounds

$$\text{Class 2 (I)}\ 0.60 \leqslant pI_{50} \leqslant 4.00\ N_{2(I)} = 64$$
$$\text{Class 1 (I)}\ \ pI_{50} \leqslant 0.60\ N_{1(I)} = 25$$

Problem II: very active versus inactive compounds

$$\text{Class 2 (II)}\ 3.00 \leqslant pI_{50} \leqslant 4.00\ N_{2(II)} = 14$$
$$\text{Class 1 (II)}\ \ pI_{50} \leqslant 0.60\ N_{1(II)} = 25$$

The best conjunction obtained from problem I is

$$X15 \wedge \tilde{}RX7 \wedge \tilde{}Y6 \wedge Z1 \wedge \tilde{}(X10 \vee X11) \wedge RX8 \qquad (55)$$

with $n_{2(I)}=_{61}$, $n_{1(I)}=0$ and the accompanying variables ~*X*12, ~*X*13, ~*X*14 and ~*RX*9. Translating back this conjunction into structural terms leads to the pattern presented in Fig. 6 which is present in almost all (61 out of 64) of the active compounds and does not occur in any of the inactive analogues. Obviously, this pattern reflects basic structural requirements for activity which include an intact amide group in position *Z* (variable *Z*1), a substituent different from N, F or phenyl in the *ortho* position of *X* (variables ~*RX*7, ~*RX*8, *RX*9), a C=C group adjacent to the carbon atom of the amide moiety (variable *X15*), and rings in *X* (variables *X*10, ~*X*11, ~*X*12, ~*X*13, ~*X*14) and *Y* (variable ~*Y*6).

Problem II yields the following conjunction:

$$RY16\wedge \tilde{}X6 \qquad (56)$$

with $n_{2(II)}$=10, $n_{1(II)}$=0 and the accompanying variables *Z*1, ~*X*5, ~*X*9, ~*X*10, ~*X*11, ~*X*12, ~*X*13, ~*X*14, *X*15, ~*RX*3, ~*RX*7, ~*RX*8, ~*RX*9 and *Y*1. Conjunction (56) leads to the structure presented in Fig. 7, which contains all basic elements of Fig. 6 but yields some additional information concerning the nature of the rings in *X* and *Y* as well as of substituents in these rings. The lesson to be learned from this is that compounds of medium activity may well be eliminated from a LOGANA analysis. As we have learned form a number of applications, the result will usually be sharper and by no means less representative as for the whole training series, and a lot of work can be saved if a training series is large.

After the LOGANA analysis the whole set of data was submitted to LOCON. Two relatively small conjunctions containing three variables already led to object groups with means, *MS,* well above the global mean of pI_{50} values of *MO*=1.67:

$$Z1\wedge RX1\wedge Y1 \qquad MS = 2.33 \quad NS = 47 \quad S = 0.793 \qquad (57)$$

Accompanying variables: *X*15, ~*X*11, ~*X*12, ~*X*13, ~*X*14.

$$X15\wedge RY16\wedge (RX7\vee RX9) \qquad MS = 2.46 \quad NS = 37 \quad S = 0.675 \qquad (58)$$

Accompanying variables: *Z*1, *Y*1, ~*X*10, ~*X*11, ~*X*12, ~*X*13, ~*X*14, ~*X*24.

Both conjunctions comprise large parts of the active compounds and tell much the same story as the LOGANA conjunctions characterizing basic activity features. A conjunction already more specific for the highly active compounds was obtained as

$$(X16\vee X20)\wedge RY16\wedge \tilde{}RY18\wedge \tilde{}RY19 \qquad MS = 3.04 \quad NS = 14 \quad S = 0.366 \qquad (59)$$

Accompanying variables: *Z*1, *Y*1, ~*RX*8, ~*RX*12, ~*X*3, ~*X*21, ~*X*8, ~*X*11, ~*X*12, ~*X*13, ~*X*14, ~*X*22, ~*X*23, ~*X*24.

Translating back this conjunction into structural terms leads to Fig. 8, yielding the additional information that the phenyl ring adjacent to the nitrogen should not be substituted in the *para* position, that CH_3 or C_3H_5 should not appear as *ortho* substituents, and that the structure –N=C(NH_2)–S– also is a favourable moiety to be placed in *X*. Conjunction (59) or the pattern in Fig. 8 cannot be interpreted to mean that *para* substituents, for example, are strictly incompatible with high activity or that phenyl is always an unfavourable moiety to be placed in *X*. What it in fact means is that, within the training series, a set of highly active compounds exists having the features presented by this conjunction in common so that these features may be regarded a reasonable guideline when synthesizing new compounds.

If the variable *RY*1 is added to conjunction (59) one obtains

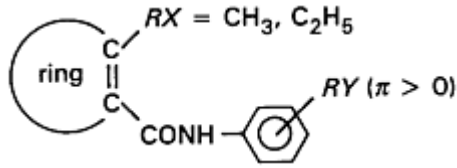


Fig. 7. Structural pattern resulting from conjunction (56).

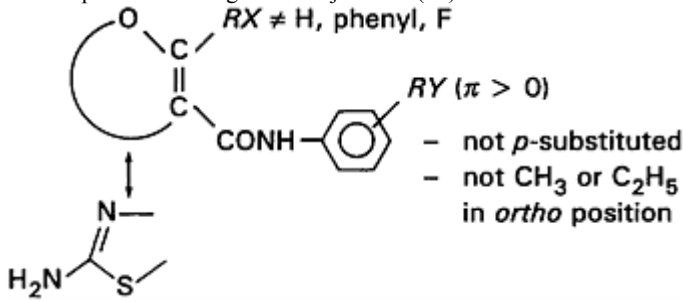


Fig. 8. Structural pattern resulting from conjunction (59).

$$(X16V\ X20)\Lambda\ RY16\Lambda\ \tilde{}RY18\Lambda\ \tilde{}RY19\Lambda\ RY1 \qquad (60)$$
$$MS = 3.42 \quad NS = 5 \quad S = 0.343$$

This conjunction yields an object group with a very high activity mean but which is less representative because of the relatively small number of compounds involved. It may, nevertheless, be taken as an indication that methyl substitution in the *meta* position of *Y* is probably another typical feature of the very active compounds.

Both LOGANA and LOCON yield consistent results allowing formulation of structural criteria for enzyme inhibitory potency. These criteria are consistent with empirical known criteria for the type of compounds concerned [173] but are much more detailed. The pattern found provides a good systematization in structural terms and a clear picture of why active compounds are active but will, in the present case, not lead to novel structures of outstanding activity. The reason is that the training series, though structurally diverse, is still too limited for elucidating structural information which could be regarded as a true surprise. For example, a compound with *X*=ring A, RX_0=CH_3, *Z*=CONH, *Y*=phenyl, *RY*=*m*–CH_3 would be predicted to be of very high activity according to the results obtained, but this compound has already been made and is the most active of the training series.

LOGANA and LOCON have the following advantages:

— No assumptions about additivity of effects or linear models are necessary.
— The contribution of a given structural feature to the biological activity need not be a consistent factor.

— A uniform mechanism of action for all compounds of the training series is not required. If different mechanisms are operating they will be reflected by different pharmacophores provided, of course, that a representative number of compounds exist for each mechanism.
— There is no dependence on data distribution or statistical formalisms.
— Full use is made of the intuition and experience of the researcher.
— The results are directly obtained in terms of a structural pattern.
— Physicochemical molecular parameters are usually not needed. Only when the variation of the substituents becomes an important property are substituent constants required to parametrize substituents via binary descriptors [123, 124]. Even then, however, approximate values are sufficient.
— The methods can be applied to extremely diverse data sets where linear free energy formalisms are bound to fail. Inactive compounds can be included in the calculation.
— For LOGANA, biological activity need not be available on a continuous scale. A simple classification is sufficient, so that it is possible to use data from mass screening or sampled over a long period of time or even from different sources.

Most specific for LOGANA and LOCON are the second to sixth points above, while the others are shared by the majority of the other topological methods mentioned in the preceding section. It should be mentioned in this context that the STRAC procedure also uses the interactive construction of conjunctions by logical operations with decision making by the operator. The objective, however, is different from LOGANA in that not a topological pattern but the evaluation of discriminating features with the final purpose of classifying compounds via probability-derived criteria is aimed at. Other differences are that STRAC cannot handle very large and diverse data sets and that it is directed at lead optimization, while LOGANA and LOCON are more of the lead generation type.

EVAL is a new interactive tool to aid the process of extracting substructural entities which are thought to be responsible for the action. The approach is based on binary (logical) descriptors. These descriptors can be understood as ‘indicator variables’ usually describing the presence or absence of certain structural features of the compounds under consideration. The features are assumed to be predefined. Normally, one will select a set of substructures and/or physicochemical properties representing the relevant structural properties. This can be done in an automated way using computer-based fragmentation techniques or simply by defining them intuitively after inspection of the training set. Physicochemical properties are also translated into binary descriptors using boundaries of corresponding molecule parameters (for example log P) as criteria. The only requirement a descriptor must meet is that its presence or absence in each of the compounds can be established. Hence, each compound is described as a set of features which are known to be ‘true’, that is, present within the given structure. Obviously, the definition of the structural descriptors is a crucial step

in the analysis. It must be done in such a way that the basic topological characteristics and the structural variation of the training series are adequately represented. For a discussion of certain requirements and possibilities see, for example, Franke *et al*. [125].

Once the features are defined and the compounds are coded according to their presence or absence a classification of the compounds regarding their activity should be performed. This is always a good approach even if continuous activity data are available. The classification into an 'active' and 'inactive' class will increase the clearness of the analysis and will speed up the procedure. If there are not obvious limits the methods of exploratory data analysis can be used to refine meaningful threshold values. The coded structures together with the classification and (if known) continuous activity data will form the input information of the EVAL procedure.

EVAL is an interactive tool that allows the logical analysis of the data. The main task is to find logical combinations of single features in such a way that the combination is 'true' for a certain set of active compounds and at the same time 'false' for all or at least most of the inactive ones. In the best case a combination of features is formed that is present in all active and in none of the inactive structures within the training series. Since the descriptors normally reflect topological information this combination can be called a 'topological pharmacophore'. The full set of logical operations known from mathematical logics can be used to construct combinations of single features to obtain complex expressions reflecting complex substructural entities. This includes negation ('not'), conjunction ('and'), adjunction and disjunction (inclusive and exclusive 'or') as well as conclusion ('if-then'), equivalence, ('if and only if'), Sheffer's function ('incompatible with', 'not together with') and Nicod's function ('neither-nor').

The EVAL analysis always starts from a sorted list of single features. Each feature is presented with its occurrence within the two classes of activity (that is, the number of compounds of the given class possessing the feature) and the mean activity of all compounds with that feature. This gives the researcher a quick review of how the single descriptors are distributed. Different sorting criteria can be applied to affect the output sequence. The goodness measures used for the sorting process are the same as implemented in LOGANA and LOCON [125] and are discussed in detail elsewhere [176].

Selecting a feature which is present in at least many of the active compounds to be the 'stem variable' all combinations with it can be formed now in a simple interactive way. The results are again presented in a sorted list and can be used as a stem for further combinations. As these combinations grow the pattern of substructures reflected by them becomes more complex and is therefore present in fewer compounds. As a consequence compounds are eliminated at each step. In order to end up with a feature combination typical of the active class the strategy of combining features always is to eliminate inactive and retain active compounds. Since the programme is fully interactive the professional skill of the

medicinal chemist has full impact at any step of the analysis so that, for example, the generation of meaningless complex features can be avoided. The researcher has complete freedom to check any combination, to return to previous steps of the analysis in order to try another hypothesis or to define new variables at any time. This makes EVAL a very flexible tool to extract meaningful substructural information from large and diverse data sets.

Although the programme does not follow the fixed lines of a built-in algorithm in most cases a simple strategy will help in obtaining a quick review about pharmacophoric pattern. Using an option that we call 'garbage collection', one can reduce the number of variables and inactive compounds to be inspected at the beginning of the analysis.

This means that all features occurring only in inactive compounds and never present in active ones are combined into an adjunction (inclusive 'or') and will then be negated to form a new variable. Note that this does not mean that one or all of them are responsible for inactivity. But using this generated variable as a stem for further analysis will reduce the inactive class and will improve the clarity of the procedure. The desired feature combinations for the active class can now be constructed in a much simpler way. The resulting pharmacophoric pattern should then be compared with the previously excluded inactive objects. A joint occurrence of such a pattern together with one or more of the variables of the 'garbage' would be an indication of a structural misfit that leads to inactivity.

When applying topological methods one must, of course, keep in mind several restrictions and possible pitfalls, already discussed in part. As in any QSAR method the properties of the training series are crucial for the validity of the results. Structural variations not adequately represented in the training series can also not adequately be reflected by the results. For LOGANA and LOCON, for example, this has the consequence that the pattern evaluated cannot be regarded as a general or final solution. All that can be said is that they are typical of the highly active analogues in that part of the chemical compound space that is covered by the training series and that there is a high probability of finding other highly active compounds if these patterns are used as a design criterion.

A very serious limitation of all topological approaches results from the fact that the actual events of drug-receptor interactions are three-dimensional and dynamic, while these methods are static and operate in only two dimensions. Topological approaches will, therefore, yield valid results only to the extent that two-dimensional chemical structures reflect the much more complicated processes operating when drugs interact with biological targets. This implies that in a number of cases such methods are bound to fail because the topological description of structures is inadequate; even if geometrical descriptors are added which can, in principle, be done in all methods, this would still be true because of the static kind of approach. Molecular modelling techniques are then the most promising way to go but even in connection with such methods topological analyses can be very valuable since their results may provide hypotheses which can be used as an input when applying the more sophisticated techniques.

REFERENCES

1 [] Franke, R. (1984) *Theoretical Drug Design Methods.* Elsevier. Amsterdam, and Akademie-Verlag, Berlin.

2 [] Franke, R. (1980) *Optimierungsmethoden in der Wirkstofforschung.* Akademie-Verlag, Berlin.

3 [] Mager, P.P. (1980) in *Drug Design,* Vol. IX (ed. Ariens, E.J.), Academic Press, New York, p. 188.

4 [] Mager, P.P. (1980) in *Drug Design,* Vol. X (ed. Ariens, E.J.), Academic Press, New York, p. 343.

5 [] Mager, P.P. (1983) *Multidimensional Pharmacochemistry: Design for Safer Drugs.* Academic Press, New York.

6 [] Lewi, P.J. (1976) in *Drug Design,* Vol. VII (ed. Ariens, E.J.), Academic Press, New York, p. 209.

7 [] Lewi, P.J. (1980) in *Drug Design,* Vol. X (ed. Ariens, E.J.), Academic Press, New York. p. 308.

8 [] Lewi, P.J. (1982) *Multivariate Data Analysis in Industrial Practice,* Chemometrics Series (ed. Bawden, D.), Wiley, Chichester.

9 [] Wold, S. and Sjöström, M. (1977) in *Chemometrics: Theory and Practice* (ed. Kowalski, B.R.), *ACS Symp. Ser.,* **52,** 243.

10 [] Wold, S., Albano, C., Dunn, W.J. III, Edlund, U., Esbensen, D., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W. and Sjöström, M. (1984) in *Chemometrics: Mathematics and Statistics in Chemistry* (ed. Kowalski, B.R.), Reidel, Dordrecht, p. 17.

11 [] Malinowski, E.R. and Howery, D.G. (1980) *Factor Analysis in Chemistry,* Wiley, New York.

12 [] Rummel, R.J. (1970) *Applied Factor Analysis,* Northwestern University Press, Evanston.

13 [] Harman, H.H. (1967) *Modern Factor Analysis,* 3rd edn, University of Chicago.

14 [] Wold, S. (1978) *Technometrics,* **20,** 387.

15 [] Malinowski, E.R. (1977) *Anal. Chem.,* **49,** 606, 612.

16 [] Dove, S., Coats, E., Scharfenberg, P. and Franke R. (1985) *J. Med. Chem.,* **28,** 447.

17 [] Shah, K.J. and Coats, E.A. (1977) *J. Med. Chem.,* **20,** 1001.

18 [] Coats, E.A., Shah, K.J., Milstein, S.R., Genther, C.S., Nene, D.M., Roesener, J., Schmidt, J., Pleiss, M., Wagner, E. and Wagner, J.K. (1982) *J. Med. chem.,* **25,** 57.

19 [] Hansch, C. and Leo, A. (1979) *Substituent Constants for Correlation Analysis in Chemistry and Biology,* Wiley, New York.

20 [] Lukovits, I. and Lopata, A. (1980) *J. Med. Chem.,* **23,** 449.

21 [] Darvas, F., Meszaros, Z., Kovacs, L., Hermecz, I., Balogh, M. and Kordos, J. (1979) *Arzneimittel-Forsch,* **29,** 1334.

22 [] Lukovits, I. and Lopata, A. (1980) *J. Med. Chem.,* **23,** 449.

23 [] Davies, R.H. and Morris, T.R. (1983) *Int. J. Quant. Chem.,* **23,** 1385.

24 [] Coats, E.A., Cordes, H.-P., Kulkorni, V.M., Richter, M., Schaper, K. -J., Weise, M. and Seydel, J.K. (1985) *Quant. Struct.—Act. Relat.,* **4,** 99.

25 [] Dove, S. (1978) Doctoral thesis. M. Luther University, Halle-Wittenberg.

26 [] Szigeti, Z., Cserhati, T. and Bordos, B. (1985) *Gen. Physiol. Biophys.,* **4,** 321.

27 [] Dove, S., Kühne, R. and Franke, R. (1985) in *QSAR in Toxicology and Xenobiochemistry,* Vol. 8 (ed. Tichy, M.), Elsevier, Amsterdam, p. 277.
28 [] Dove, S. and Franke, R. (1977) *Wiss. Beitr. M.-Luther-Univ. Halle-Wittenberg,* **42(R34),** 216.
29 [] Schaper K.-J. and Seydel, J.K. (1985) in *QSAR and Strategies in the Design of Bioactive Compounds* (ed. Seydel, J.K.), VCH, Weinheim, p. 173.
30 [] Cammarata, A. and Menon, G.K. (1976) *J. Med. Chem.,* **19,** 739.
31 [] Menon, G.K. and Cammarata, A. (1977) *J. Pharm. Sci.,* **66,** 304.
32 [] Grossy, G., Tenlade, J.-C., Chapat, J.P., Simeon de Buochberg, M. and Attisso, M. (1982) *Eur. J. Med. Chem.,* **17,** 109.
33 [] Grossy, G., Rival, Y., Tenlade, J.-C., Chapet, J.-P., Simeon de Buochberg, M. and Attisso, M. (1985) *Eur. J. Med. Chem.,* **20,** 199.
34 [] Mawden, D., Gymer, G.E., Marriott, M.S. and Tute, M.S. (1983) *Eur. J. Med. Chem.,* **18,** 91.
35 [] Chen, B.-K., Horvath, C. and Bertino, J.R. (1979) *J. Med. Chem,* **22,** 483.
36 [] Dove, S., Streich, W.J. and Franke, R. (1980) *J. Med. Chem.,* **23,** 1456.
37 [] Cramer, R.D. (1980) *J. Am. Chem. Soc.,* **102,** 1837.
38 [] Cramer, R.D. (1983) *Quant. Struct.-Act. Relat.,* **2,** 13.
39 [] Franke, R., Dove, S. and Kühne, R. (1979) *Eur. J. Med. Chem.,* **14,** 363.
40 [] Franke, R., Kühne, R. and Dove, S. (1983) in *Quantitative Approaches to Drug Design,* Vol. 6 (ed. Dearden, J.C.), Elsevier, Amsterdam, p. 15.
41 [] Dove, S., Kühne, R. and Franke, R. (1980) in *Chemical Structure—Biological Activity Relationships* (ed. Darvas, F.), Akademiai Kiado, Budapest, p. 247.
42 [] Dunn, W.J.III and Wold, S. (1978) *Acta Chem. Scand.,* **B32,** 536.
43 [] Dunn, W.J.III and Johansson, E. (1983) *Quant. Struct.—Act. Relat.,* **2,** 156.
44 [] Sjöström, M. and Wold, S. (1985) *J. Mol. Evol.,* **22,** 272.
45 [] Hellberg, S., Sjöström, M. and Wold, S. (1986) *Acta Chem. Scand.,* **B40,** 135.
46 [] Hill, M.O. (1974) *Appl. Statist.,* **23,** 340.
47 [] Dore, J.-C., Gilbert, J., Ojasoo, T. and Raynaud, J.-P. (1986) *J. Med. Chem.,* **29,** 54.
48 [] Keasling, H.H. and Moffett, R.B. (1971) *J. Med. Chem.,* **14,** 1106.
49 [] Weiner, M.L. and Weiner, P.H. (1973) *J. Med Chem.,* **16,** 655.
50 [] Franke, R. (1978) *Il Farmaco,* 34, 545.
51 [] Franke, R. (1982) in *Strategy in Drug Research,* Vol. 4 (ed. Keverling Buisman, J.A.), Elsevier, Amsterdam, p. 355.
52 [] Miyashita, Y., Takahashi, Y., Daiba, S., Abe, H. and Sasaki, S. (1982) *Anal. Chim. Acta,* **143,** 35.
53 [] Draper, N.R. and Smith, H. (1981) *Applied Regression Analysis,* 2nd edn, Wiley, New York.
54 [] Wold, S. and Dunn, W.J.III (1983) *J. Chem. Inf. Comput. Sci.,* **23,** 6.
55 [] Dunn, W.J.III, Wold, S., Edlund, U., Hellberg, S. and Gasteiger, J. (1984) *Quant. Struct.—Act. Relat.,* **3,** 131.
56 [] Joliffe, I.T. (1982) *Appl. Statist.,* **31,** 300.
57 [] Wold, S., Albano, C., Dunn, W.J. III, Esbensen, K., Hellberg, S., Johansson, E. and Sjöström, M. (1983) in *Food Research and Data Analysis* (eds Martens, H. and Russwurm, H.), Elsevier, Amsterdam, p. 147.
58 [] Norden, B., Edlund, U., Johnels, D. and Wold, S. (1983) *Quant. Struct.—Act. Relat.,* **2,** 73.

59 [] Wold, H. (1982) in *Systems under indirect Observation,* Part 2 (eds Joreskog, K.G. and Wold, H.), North-Holland, Amsterdam, p. 1.
60 [] Hansch, C. (1984) contribution to discussion in *Drug Design: Fact or Fantasy?* (eds Jolles, G. and Wooldridge, K.R.H.), Academic Press, London, p. 93.
61 [] Lindberg, W., Persson, J.-A. and Wold, S. (1983) *Anal. Chem.,* **55,** 643.
62 [] Hellberg, S., Wold, S., Dunn, W.J.III, Gasteiger, J. and Hutchings, M. G. (1985) *Quant. Struct.—Act. Relat.,* **4,** 1.
63 [] Berntsson, P. and Wold, S. (1986) *Quant. Struct.—Act. Relat.,* **5,** 45.
64 [] Clementi, S. (1984) in *Drug Design: Fact or Fantasy?* (eds Jolles, G. and Wooldridge, K.R.H.), Academic Press, London, p. 73.
65 [] Topliss, J.G. and Edwards, R.P. (1979) *J. Med. Chem.,* **22,** 1238.
66 [] Mager, P.P. (1984) in *QSAR in Design of Bioactive Compounds. Proceedings of the First International Telesymposium on Medicinal Chemistry,* (ed. Kuchar, M.), J.R.Prous, Barcelona, p. 379.
67 [] Mager, P.P. (1982) *Toxicol. Lett.,* **11,** 67.
68 [] Lukovits, I. (1983) *J. Med. Chem.,* **26,** 1104.
69 [] Kubota, T., Hanamura, J., Kano, K. and Uno, B. (1985) *Chem. Pharm. Bull.,* 33, 1488.
70 [] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis,* Academic Press, New York.
71 [] Mager, P.P. and Seese, A. (1981) *Pharm. Unserer Zeit,* **10,** 97.
72 [] Mager, P.P. (1982) *Med. Res. Rev.,* **2,** 93.
73 [] Szydlo, R.M., Ford, M.G., Greenwood, R. and Salt, D.W. (1985) in *QSAR and Strategies in the Design of Bioactive Compounds* (ed. Seydel, J.K.), VCH, Weinheim, pp. 219, 229.
74 [] Stuper, A.J., Bruegger, W.E. and Jurs, P.C. (1979) *Computer Assisted Studies of Chemical Structure and Biological Function,* Wiley, New York.
75 [] Franke, R. and Meiske, W. (1976) *Acta Biol. Med. Ger.,* **35,** 73.
76 [] Franke, R., Vassiliev, G.N. and Dove, S. (1979) *Compt. Rend. Acad. Bulg. Sci.,* **32,** 369.
77 [] Dove, S., Franke, R., Mndshojan, O.L., Shkuljev, W.A. and Chashakian, W.L. (1978) *J. Med. Chem.,* **22,** 90.
78 [] Schönfelder, D. and Franke, R. (1978) in *Quantitative Structure—Activity Analysis* (eds Franke, R. and Oehme, P.), Akademie-Verlag, Berlin, p. 303.
79 [] Tonew, M., Laass, W., Tonew, E., Franke, R., Goldner, H. and Zschiesche, W. (1978) *Acta virol.,* **22,** 287.
80 [] Dove, S. and Franke, R. (1978) in *Quantitative Structure-Activity Analysis* (eds Franke, R. and Oehme, P.), Akademie-Verlag, Berlin, p. 257.
81 [] Krause, G., Franke, R. and Vassiliev, G.N. (1979) *Biochem. Physiol. Pfl.,* **174,** 128.
82 [] Hagemann, R. and Franke, R. (1976) *Pharmazie,* **31,** 130.
83 [] Laass, W., Bercher, H., Schult, E., Franke, R. and Grisk, A. (1979) *Pharmazie,* **34,** 334.
84 [] Bercher, H., Laass, W., Schult, E., Grisk, A. and Franke, R. (1979) *Pharmazie,* **34,** 336.
85 [] Krause, G., Klepel, M., Jumar, A. and Franke, R. (1984) *Tag.—Ber. Akad. Landwirtsch.—Wiss. DDR, Berlin,* **222,** 229.

86 [] Krause, G., Klepel, M. and Franke, R. (1983) in *Quantitative Approaches to Drug Design* (ed. Dearden, J.C.), Elsevier, Amsterdam, p. 223.
87 [] Krause, G., Franke, R. and Klepel, M. (1985) in *QSAR and Strategies in the Design of Bioactive Compounds* (ed. Seydel, J.K.), Verlag Chemie GmbH, Weinheim, P. 416.
88 [] Krause, G. and Franke, R. (1985) in *QSAR in Toxicology and Xenobio-chemistry* (ed. Tichy, M.), Elsevier, Amsterdam, p. 309.
89 [] Krause, G. and Zanke, D. (1986) *Biochem. Physiol. Pfl.,* **181,** 497.
90 [] Ahrens, H. and Läuter, J. (1974) *Mehrdimensionale Varianzanalyse,* Akademie-Verlag, Berlin.
91 [] Martin, Y.C., Holland, J.B., Jarboe, C.H. and Plotnikoff, N. (1974) *J. Med. Chem.,* **17,** 409.
92 [] Prakash, G. and Hodnett, E.H. (1978) *J. Med. Chem.,* **21,** 369.
93 [] Dunn, W.J.III and Greenberg, M.J. (1977) *J. Pharm. Sci.,* **66,** 1416.
94 [] Henry, D.R. and Block, J.H. (1979) *J. Med. Chem.,* **22,** 465.
95 [] Henry, D.R. and Block, J.H. (1980) *Eur. J. Med. Chem.,* **15,** 133.
96 [] Moriguchi, I., Komatsu, K. and Matsushi, Y. *J. Med. Chem.,* **23,** 20.
97 [] Smith, C.C., Genther, C.S. and Coats, E.A. (1979) *Eur. J. Med. Chem.,* **14,** 271.
98 [] Ordukhanyan, A.A., Sarkisyan, A.S., Landau, M.A., Mndzhoyan, S.L. and Ter-Zakharyan, Yu. Z. (1983) *Khim. Farm. Zh.,* **17,** 63.
99 [] Ordukhanyan, A.A., Sarkisyan, A.S., Landau, M.A., Magakyan, A.R., Hudzhoyan, S.L. and Ter-Zakharyan, Yu. Z. (1983) *Khim. Farm. Zh.,* **17,** 66.
100 [] Sekiya, T., Hata, S. and Yamada, S.-I. (1983) *Chem. Pharm. Bull.,* **31,** 2432.
101 [] Sekiya, T., Imada, S., Hata, S. and Yamada, S.-I. (1983) *Chem. Pharm. Bull.,* **31,** 2779.
102 [] Takahashi, Y., Miyashita, Y. and Tanaka, Y. (1984) *J. Pharm. Sci.,* **73,** 737.
103 [] Nigmatullin, R.S., Osipov, A.L., Puzatkin, A.P. and Koplyug, V.A. (1985) *Khim. Farm. Zh.,* **19,** 179.
104 [] Unger, S.H. (1983) in *Quantitative Approaches to Drug Design* (ed. Dearden, J.C.), Elsevier, Amsterdam, p. 43.
105 [] Dove, S., Kühne, R. and Franke, R. (1984) in *QSAR in Design of Bioactive Compounds. Proceedings of the First International Telesymposium on Medicinal Chemistry* (ed. Kucharc, M.), J.R. Prous, Barcelona.
106 [] Wold, S. (1976) *Patt. Recog.,* **8,** 127.
107 [] Dunn, W.J.III, Wold, S. and Martin, Y. (1978) *J. Med. Chem.,* **21,** 922.
108 [] Dunn, W.J.III and Wold, S. (1978) *J. Med Chem.,* **21,** 1001.
109 [] Norden, B., Edlund, U. and Wold, S. (1978) *Acta Chem. Scand.,* **B32,** 602.
110 [] Dunn, W.J.III and Wold, S. (1980) *Bioorganic Chem.,* **9,** 505.
111 [] Dunn, W.J.III and Wold, S. (1980) *J. Med Chem.,* **23,** 595.
112 [] Wold, S., Dunn, W.J.III and Hellberg, S. (1984) in *Drug Design: Fact or Fantasy?* (eds Jolles, G. and Wooldridge, K.R.H.), Academic Press, London, p. 95.
113 [] Serrano, J.L., Marcos, M., Melendez, E., Albano, C., Wold, S. and Elguero, J. (1985) *Acta Chem. Scand.,* **B39,** 329.
114 [] Bawden, D. (1983) *J. Chem. Inf. Comput. Sci.,* **23,** 14.
115 [] Smith, E.G. and Baker, P.A. (1975) *The Wiswesser Line-Formula Chemical Notation,* 3rd edn, Chemical Information Management, Cherry Hill, USA.
116 [] Jurs, P.C., Chou, J.T. and Yuan, M. (1979) *J Med. Chem.,* **22,** 476.

117 [] Kirschner, G.L. and Kowalski, B.R. (1979) in *Drug Design,* Vol. VIII (ed. Ariens, E.J.), Academic Press, New York, p. 73.
118 [] Golender, V.E. and Rozenblit, A.B. (1980) in *Drug Design,* Vol IX (ed. Ariens, E.J.), Academic Press, New York.
119 [] Avidon, V.V., Pomerantsev, I.A., Golender, V.E. and Rozenblit, A.B. (1982) *J. Chem. Inf. Comput. Sci.,* **22,** 207.
120 [] Rozenblit, A.B. (1982) in *Strategy in Drug Design* (ed. Keverling Buisman, J.A.), Elsevier, Amsterdam, p. 287.
121 [] Hübel, S., Rösner, T. and Franke, R. (1980) *Pharmazie,* **35,** 424.
122 [] Franke, R., Streich, W.J. and Hübel, S. (1983) *Studia Biophys.,* **97,** 11.
123 [] Franke, R. and Streich, W.J. (1985) *Quant. Struct.—Act. Relat.,* **4,** 51.
124 [] Franke, R. and Streich, W.J. (1985) *Quant. Struct.—Act. Relat.,* **4,** 63.
125 [] Franke, R., Hübel, S. and Streich, W.J. (1985) *Environ. Health Perspect.,* **61,** 239.
126 [] Mercer, C. and Dubois, J.E. (1979) *Eur. J. Med. Chem.,* **14,** 415.
127 [] Simon, Z., Chiriac, A., Holban, S., Cinbotaru, D. and Minalas, G. (1984) *The MTD Method for QSAR Studies,* Research Studies Press, Letchworth.
128 [] Mathews, R.J. (1975) *J. Am. Chem. Soc.,* **97,** 935.
129 [] Hodes, L., Hazard, G.F., Geran, R.F. and Richman, I. (1977) *J. Med. Chem.,* **20,** 469.
130 [] Perrin, C.L. (1974) *Science,* **183,** 551.
131 [] Feldman, A. and Holdes, L. (1975) *J. Chem. Inf. Comput. Sci.,* **15,** 147.
132 [] Hodes, L. (1976) *J. Chem. Inf. Comput. Sci.,* **16,** 88.
133 [] Nilsson, N.J. (1965) *Learning Machines,* McGraw-Hill, New York
134 [] Meisel, W.S. (1972) *Computer-Oriented Approaches to Pattern Recognition,* Academic Press, New York.
135 [] Batchelor, B.G. (1974) *Practical Approach to Pattern Classification,* Plenum Press, New York.
136 [] Fu, K.S. (ed.) (1976) *Digital Pattern Recognition,* Springer-Verlag, New York.
137 [] Jurs, P.C. and Isenhour, R.L. (1975) *Chemical Applications of Pattern Recognition,* Interscience, New York.
138 [] Andrews, H.C. (1972) *Introduction to Mathematical Techniques in Pattern Recognition,* Wiley, New York.
139 [] Jurs, P.C., Chou, J.T. and Yuan, M. (1979) in *Computer-Assisted Drug Design* (eds Olson, E.C. and Christoffersen, R.E.), *ACS Symp. Series,* **112,** 103.
140 [] Jurs, P.C., Stouch, T.R., Czerwinski, M. and Narvaez, J.N. (1985) *J. Chem. Inf. Comput. Sci.,* **25,** 296.
141 [] Cover, T.M. and Hart, P.E. (1967) *IEEE Trans. Inf. Theory,* IT–13, 21.
142 [] Hart, P.E. (1968) *IEEE Trans. Inf. Theory,* IT–14, 515.
143 [] Chu, K.C., Feldman, R.J., Shapiro, M.B., Hazard, G.F. and Geran, R.I. (1975) *J. Med. Chem.,* **18,** 539.
144 [] Chou, J.T. and Jurs, P.C. (1979) *J. Med. Chem.,* **22,** 792.
145 [] Clerc, J.T., Naegeli, P. and Seibl, J. (1973) *Chimia,* 27, 639.
146 [] Chu, K. (1974) *Anal. Chem.,* **46,** 1181.
147 [] Stuper, A.J. and Jurs, P.C. (1975) *J. Am. Chem. Soc.,* **97,** 182.
148 [] Stoltzberg, C.J. and Wilkins, C.L. (1977) *J. Am. Chem. Soc.,* **99,** 439.
149 [] Yuan, M. and Jurs, P.C. (1980) *Toxicol, Appl. Pharmacol.,* **52,** 294.
150 [] Yuta, K. and Jurs, P.C. (1981) *J. Med. Chem.,* **24,** 241.

151 [] Stuper, A.J., Brugger, W.E. and Jurs, P.C. (1977) in *Chemometrics: Theory and Application* (ed. Kowalksi, B.R.), *ACS Symp. Ser.,* **52,** 165.
152 [] Schönfelder, D. and Franke, R. (1978) in *Quantitative Structure Activity Analysis* (eds Franke, R. and Oehme, P.), Akademie-Verlag, Berlin, p. 303.
153 [] Abe, H., Kumazawa, S., Tajn, T. and Sasaki, S. (1976) *Biomed, Mass Spectrom.,* **3,** 151.
154 [] Henry, D.R., Jurs, P.C. and Denny, W.A. (1982) *J. Med. Chem.,* **25,** 899.
155 [] Rose, S.L. and Jurs, P.C. (1982) *J. Med. Chem.,* **25,** 769.
156 [] Jurs, P.C., Hasan, M.N., Henry, D.R., Stouch, T.S. and Whalen-Pedersen, E.K. (1983) *Fund. Appl. Toxicol.,* **3,** 343.
157 [] Dierdorf, D.S. and Kowalski, B.R., cited in [117].
158 [] Bodor, N., Harget, A.J. and Phillips, E.W. (1983) *J. Med. Chem.,* **26,** 318.
159 [] Byers, W.A., Frelser, B.S. and Perone, S.P. (1983) *Anal. Chem.,* **55,** 620.
160 [] Shulman, V.S. (1984) *Khim. Farm. Zh.,* **18,** 1087.
161 [] Hodes, L. (1979) in *Computer-Assisted Drug Design* (eds Olsen, E.C. and Christoffersen, R.E.), *ACS Symp. Ser.,* **112,** 583.
162 [] Hodes, L. (1981) *J. Chem. Inf. Comput. Sci.,* **21,** 128.
163 [] Hodes, L. (1981) *J. Chem. Inf. Comput. Sci.,* **21,** 132.
164 [] Tinker, J.F. (1981) *J. Chem. Inf. Comput. Sci.,* **21,** 3.
165 [] Tinker, J.F. (1981) *J. Comput. Chem.,* **2,** 231.
166 [] Klopman, G. (1984) *J. Am. Chem. Soc.,* **106,** 7315.
167 [] Enslein, K. and Craig, P.N. (1978) *J. Environ. Pathol. Toxicol.,* **2,** 115.
168 [] Enslein, K. and Craig, P.N. (1979) *Status Report on Development of Toxicological Endpoints,* Genese Corp., Rochester.
169 [] Rekker, R.F. *TIPS,* 383.
170 [] Golender, V.E. and Rozenblit, A.B. (1978) *Computer Assisted Methods of Drug Design,* Zinatne, Riga (in Russian).
171 [] Streich, W.J. and Franke, R. (1985) *Quant. Struct.—Act. Relat.,* **4,** 13 (1985).
172 [] White, G.A. and Thorn, G.D. (1975) *Pestic. Biochem. Physiol.,* **5,** 380.
173 [] ten Haken, P. and Dunn, C.C. *Proc. 6th Bri. Insectic. Fungi. Conf.,* **2,** 453.
174 [] Hübel, S. and Franke, R. (1991) in *Software Development in Chemistry.,* Vol. 5 (ed. Gmehling, J.), Springer-Verlag Heidelberg (in press).
175 [] Hübel, S. and Franke, R. (1990) *Pharmacochemistry Library.,* Elsevier, Amsterdam (in press).
176 [] Streich, W.J., Hübel, S. and Franke, R. (1988) *Math. Comput. Modelling.,* **11,** 674.
177 [] Hellberg, S., Sjöström, M., Skagerberg, B., Wikstrom, C. and Wold, S. (1987) *Acta Pharmacol. Jugosl.,* **37,** 53.
178 [] Hellberg, S., Sjöström, M., Skagerberg, B., Wikstrom, C. and Wold, S. (1987) in *QSAR in Drug Design and Toxicology* (eds Hadzi, D. and Blazic, B.J.), Elsevier, Amsterdam, p.·64.
179 [] Henry, D.R., Lavine, B.K. and Jurs, P.C. (1987) *Mutat. Res.,* **179,** 115.
180 [] Niemi, G.J., Regal, R.R. and Veith, G.D. (1985) *ACS Symp. Ser.,* **100.**
181 [] Wiese, M., Seydel, J.K., Pieper, H., Krüger, G., Noll, K.R. and Keck, J. (1987) *Quant. Struct.—Act. Relat.,* 6, 164.
182 [] Stouch, T.R. and Jurs, P.C. (1986) *J. Med. Chem.,* **29,** 2125.
183 [] Wold, S., Dunn, W.J.III and Hellberg, S. (1985) *Environ. Health Perspect.,* **61,** 257.

184 [] Esaki, T. (1987) *Chem. Pharm. Bull.,* **35,** 3105
185 [] Dunn, W.J.III, Koehler, M.G. and Grigoras, S. (1987) *J. Med. Chem.,* 30, 1121.
186 [] Chastrette, M., Zakarya, D. and Elmouaffek, A. (1986) *Eur. J. Med. Chem.,* **21,** 505.
187 [] Devillers, J., Zakarya, D., Chambon, P. and Chastrette, M. (1987) *C.R. Acad. Sci. Paris,* S. III, **304,** 195.
188 [] Wold, S., Eriksson, L., Hellberg, S., Jonsson, J., Sjöström, M., Skagerberg, B. and Wilstrom, C. (1987) *Can. J. Chem.,* **65,** 1814.
189 [] Skagerberg, B., Sjöström, M. and Wold. S. (1987) *Quant. Struct.—Act. Relat.,* **6,** 158.
190 [] Clementi, S., Coata, G., Ebert, C., Lassiani, L., Linda, P., Hellberg, S., Sjöström, M. and Wold. S. (1987) in *QSAR in Drug Design and Toxicology* (eds Hadzi, D. and Blazic, B.J.), Elsevier, Amsterdam p. 19.
191 [] Hellberg, S., Sjöström, M., Skagerberg, B. and Wold, S. (1987) *J. Med. Chem.,* **30,** 1126.
192 [] Alunni, S., Clementi, S., Ebert, C., Linda, P., Musumarra, G., Sjöström, M. and Wold. S. (1985) *J. Chem. Soc. Perkin Trans. II,* 485.
193 [] Sjöström, M., Wold. S., Wieslander, A. and Rilfors, L. (1987) *EMBO J.,* **6,** 823.
194 [] Dunn, W.J.III and Wold, S. (1986) *Trends Anal. Chem.,* **5,** 53.
195 [] Leavitt, S.A. and Mass, M.J. (1985) *Cancer Res.,* **45,** 4741.
196 [] McFarland, J.W. and Gans, D.J. (1986) *J. Med. Chem.,* **29,** 505.
197 [] McFarland, J.W. and Gans, D.J. (1987) *J. Med. Chem.,* **30,** 46.
198 [] Dunn, W.J.III, Wold, S. and Stalling, D.L. (1985) *ACS Symp. Ser.,* **292,** 243.
199 [] Miyashita, Y., Takahashi, Y., Takayama, C., Sumi, K., Nakatsuka, K., Ohkubo, T., Abe, H. and Sasaki, S. (1986) *J. Med. Chem.,* **29,** 906.
200 [] Henry, D.R. and Craig, A.M. (1989) *ACS Symp. Ser.,* **413,** 70.
201 [] Dunn, W.J.III and Wold, S. (1990) in *Pattern Recognition Techniques in Drug Design* (eds Hansch, C. *et al*.), Pergamon Press, Oxford, 691.
202 [] Stouch, T.R. and Jurs, P.C. (1986) *Quant. Struct.—Act. Relat.,* **5,** 57.
203 [] Stouch, T.R. and Jurs, P.C. (1985) *J. Chem. Inf. Comput. Sci.,* **25,** 45.
204 [] Stouch, T.R. and Jurs, P.C. *J. Chem. Inf. Comput. Sci.,* **25,** 92.
205 [] Gombar, V.K., Jaeger, E.P. and Jurs, P.C. (1988) *Quant. Struct.—Act. Relat.,* **7,** 225.
206 [] Koehler, M.G., Grigoras, S. and Dunn, W.J.III (1988) *Quant. Struct.—Act. Relat.,* **7,** 150.
207 [] Nendza, M. and Seydel, J.K. (1988) *Quant. Struct.—Act. Relat.,* 7, 165.
208 [] Seydel, J.K., Cordes, H.-P., Wiese, M., Chi, H., Croes, N., Hanpft, R., Lullmann, H., Mohr, K., Patten, M., Padberg, Y., Lullmann-Rauch, R., Vellguth, S., Meindl, W.R. and Schönenberger, H. (1989) *Quant. Struct.—Act. Relat.,* **8,** 266.
209 [] Eriksson, L., Jonsson, J., Sjöström, M. and Wold, S. (1988) *Quant. Struct.—Act. Relat.,* **7,** 144.
210 [] Jonsson, J., Eriksson, L., Hellberg, S., Sjöström, M. and Wold, S. (1989) *Quant. Struct.—Act. Relat.,* **8,** 204.
211 [] Skagerberg, B., Bonelli, D., Clementi, S., Cruciani, G. and Ebert, C. (1989) *Quant. Struct.—Act. Relat.,* **8,** 32.
212 [] Ford, M.G. and Livingstone, D.J. (1990) *Quant. Struct.—Act. Relat.,* **2,** 107.
213 [] Klopman, G. and Rosenkranz, H.S. (1984) *Mutat. Res.,* **126,** 227.

214 [] Rosenkranz, H.S., Klopman, G., Chankong, V., Pet-Edward, J. and Haimes, Y.Y. (1984) *Environ. Mutagen.,* **6,** 231.
215 [] Klopman, G. (1985) *Environ. Health Perspect.,* **61,** 269.
216 [] Klopman, G., Frierson, M.R. and Rosenkranz, H.S. (1985) *Environ. Mutagen.,* 7, 625.
217 [] Klopman, G. and Venegas, R.E. (1986) *Acta Pharm. Jugosl.,* **36,** 189.
218 [] Klopman, G. and Kalos, A.N. (1986) *J. Theor. Biol.,* **118,** 199.
219 [] Klopman, G. and Macina, O.T. (1987) *Mol. Pharmacol.,* **31,** 457.
220 [] Klopman, G., Macina, O.T., Simon, E.J. and Heller, J.M. (1985) *Theor, Biol.,* **113,** 637.
221 [] Klopman, G. and Macina, O.T. (1986) *J. Mol. Struct. (Theochem.),* **134,** 299.
222 [] Norinder, U. and Högberg, T. (1991) *Quant. Struct.—Act. Relat.,* **10,** 1.
223 [] Cruciani, G., Baroni, M., Bonelli, D., Clementi, S., Ebert, C. and Skagerberg, B. (1990) *Quant. Struct.—Act. Relat.,* **2,** 101.
224 [] Cramer, R.D.III, Bunce, J.D. and Patterson, D.E. (1988) *Quant. Struct.—Act. Relat.,* **7,** 18.
225 [] Berntsson, P. and Wold, S. (1986) *Quant. Struct.—Act. Relat.,* **5,** 45.
226 [] Johnels, D., Gillner, M., Norden, B., Toftgard, R. and Gustafsson, J.-A. (1989) *Quant. Struct.—Act. Relat.,* **8,** 83.
227 [] Cocchi, M., Menziani, M.C., Rastelli, G. and DeBenedetti, P.G. (1990) *Quant. Struct.—Act. Relat.,* **9,** 340.
228 [] Lindgren, F., Eriksson, L., Hellberg, S., Jonsson, J., Sjöström, M. and Wold. S. (1991) *Quant. Struct.—Act. Relat.,* **10,** 36.
229 [] Nilsson, L.M., Carter, R.E., Sterner, O. and Liljefors, T. (1988) *Quant. Struct.—Act. Relat.,* **7,** 84.
230 [] Consentino, U., Moro, G., Pitea, D., Todeschini, R., Brossa, S., Gualandi, F., Scolastico, C. and Giannessi, F. (1990) *Quant. Struct.—Act. Relat.,* **9,** 195.
231 [] Pajeva, I. and Golovinsky, E. (1990) *Quant. Struct.—Act. Relat.,* **9,** 216.
232 [] Walsh, D.B. and Claxton, L.D. (1987) *Mutat. Res.,* **182,** 55.
233 [] McFarland, J.W. and Gans, D.J. (1990) in *Cluster Significance Analysis* (eds Hansch C. *et al*.), Pergamon Press, Oxford, p. 675.
234 [] McFarland, J.W. and Gans, D.J. (1987) in *QSAR in Drug Design and Toxicology* (eds Hadzi, D. *et al*.), Elsevier, Amsterdam, p. 25

5
Biotransformation of drugs in man: characterization and prediction

D.D.Breimer

Division of Pharmacology, Center for Bio-Pharmaceutical Sciences, University of Leiden, PO Box 9503, 2300 RA Leiden, The Netherlands

SUMMARY

Individual patients show a wide variation in response to the same dose of a drug. Much of this interindividual variability in drug response can be explained by pharmacokinetic factors, for example absorption, distribution, metabolism and excretion. The rate at which these processes occur and which consequently determines the concentration time-course of a drug at the receptor sites is influenced by many factors pertaining to the drug and its dosage form, to pathophysiological circumstances, to genetic variables especially those determining the rate of drug metabolism and to the effects caused by other drugs taken concurrently.

The major cause for the large interindividual differences in the kinetics of drugs in man is the variation in drug-metabolizing enzyme activity. The basal rate of drug metabolism of a certain individual is determined by genetic factors and in recent years several deficiences have been detected in the capability of certain individuals to metabolize certain drugs. But it also varies with age, sex and environmental factors like diet, smoking, occupational exposure to chemicals, disease states and concurrent use of other drugs (inducers, inhibitors). Methodology that makes it possible to characterize and possibly also predict an individual's drug-metabolizing capacity by measuring the kinetics of certain model substrates *in vivo* has been developed in recent years. It has, however, turned out to be quite impossible to assess an individual subject, for example, for oxidative drug-metabolizing enzyme activity by using one probe drug only, because of the existence of multiple isozymes of the cytochrome P–450 family. Different substances may be oxidized by different isozymes (intersubstrate selectivity) and/or from a single substance different metabolites are produced by different isozymes (intrasubstrate selectivity).

The latter seems to be the case for the widely applied probe drug antipyrine; however, this compound does not provide information on any of the hitherto identified genetically determined deficiences in drug oxidation.

In order to characterize an individual subject's oxidative metabolizing capacity, several monofunctional probe drugs or a limited number of

polyfunctional probes are required, whose rate of metabolite formation *in vivo* is assessed. We have recently developed an experimental strategy, which involves the simultaneous administration of a number of probe drugs (cocktail approach) in order to investigate correlations of metabolic pathways of different drugs. This may allow predictions to be made from one substance to another. Furthermore, this approach has proved that it is possible to assess the incidence of the poor metabolizer pheno-types of sparteine and of mephenytoin in one experimental session, as well as the differential effects of enzyme induction and inhibition on *in vivo* oxidative enzyme activity.

INTRODUCTION

In drug metabolism the chemical reactions of enzymatic biotransformation are characterized as either phase I or phase II reactions. Phase I reactions cause a structural change in the parent compound by oxidation, reduction or hydrolysis, which in general gives rise to the formation of more polar metabolites. This metabolite may be biologically inactive (drug metabolism → bio-inactivation), less active or occasionally more active than the parent compound (drug metabolism → bio-activation). Occasionally a phase I reaction leads to the formation of chemically reactive intermediates, which may covalently bind to macromolecules (toxification). If the metabolite of a drug is the only pharmacologically active substance, the parent drug is a so-called prodrug. Phase II reactions, which are also called conjugation reactions, involve coupling of the drug (or often a metabolite originating from a phase I reaction) with an endogenous substrate, such as glucuronate, sulphate, acetate, or amino acids (glutathione). These conjugates are most often biologically inactive, but there are exceptions to this rule when chemically reactive conjugates are formed (toxification). Hence from a therapeutic and from a toxicological point of view drug metabolism is an extremely important process, with major implications in terms of bio-inactivation or bio-activation. It is therefore of great importance to obtain in-depth knowledge of the mechanisms of drug metabolism and also of the factors influencing it. In this presentation only one aspect will be dealt with in some greater detail, that is, the existence of large interindividual differences in drug oxidation and the experimental strategies to characterize and possibly predict the rate of drug oxidation in man by using model compounds. This variability explains—at least partially—why people (individual patients) differ widely in their sensitivity towards the desirable and also undesirable (toxic) effects of drugs and other chemical substances.

INTERINDIVIDUAL VARIABILITY IN DRUG OXIDATION

Enzymatic oxidation in the liver and to a lesser extent other organs is one of the most important processes in the disposition of many drugs and other xenobiotics

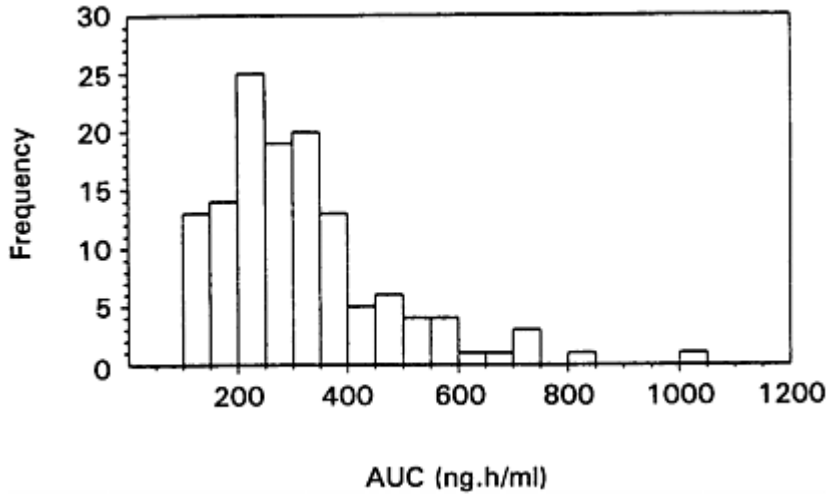


Fig. 1. Frequency distribution of the area under the plasma concentration versus time curve (AUC) of nifedipine in plasma after oral administration of 20 mg nifedipine to 130 healthy subjects (from [5]). There is 10-fold interindividual variability in AUC in this population.

in animals and humans. This process is often the rate-limiting step in the elimination of a compound from the body and/or in the formation of active or reactive metabolites. Large differences exist between different individuals within one species, including humans, in their capacity to oxidize drugs and other xenobiotics. This interspecies and interindividual variability is caused by both genetic and environmental factors [1–4]. The basal rate of drug metabolism in a particular individual is determined primarily by genetic constitution, but also varies with age, gender and environmental factors such as diet, disease states and concurrent use of other drugs (inducers, inhibitors). The latter factor also may cause intraindividual variability. A certain individual may, for instance, start or stop smoking, contract a disease, or start or stop taking other drugs, all of which may change drug-metabolizing activities to a greater or lesser extent and for shorter or longer periods. Under basal conditions the degree of intraindividual variability is far less than the degree of interindividual variability. However, perturbation of the basal conditions by some environmental or disease factors may have a profound influence on the rate of drug metabolism in that individual. In Fig. 1, the frequency distributions of the area under the plasma concentration-time curve (AUC) is shown for the calcium channel blocker nifedipine (dihydropyridine-type) after administration of 20 mg nifedipine to 130 young healthy subjects [5]. There is a very large intersubject variability in nifedipine kinetics on oral administration, which is most likely caused by large differences in drug-metabolizing enzyme activity. No significant differences were observed in nifedipine kinetics between males and females, or between smokers and non-

smokers or females using contraceptive steroids. On the other hand, it has been established that nifedipine metabolism is quite sensitive to environmental factors, like induction by barbiturates and inhibition by cimetidine [6]. In addition, grapefruit juice was recently shown to inhibit nifedipine metabolism, as well as the metabolism of other dihydropyridine calcium channel blockers [7, 8]. This is caused by relatively high concentrations of specific flavonoids which seem to be present in grapefruit juice [9]. Obviously such a factor also causes intraindividual variability in the kinetics of dihydropyridines, because the intake of food or drink constituents like grapefruit is not constant with time.

There are important interactions between genetic and environmental factors determining drug-metabolizing enzyme activity. To achieve expression, genetic factors often require participation of environmental factors; conversely environmental factors sometimes require a genetic apparatus to exert their effects on enzyme activity. There is in general a clear need for quantitative information on drug-metabolizing enzyme activity *in vivo*.

ASSESSMENT OF DRUG OXIDATION *IN VIVO* AND THE MULTIPLICITY OF THE P–450 SYSTEM

Several different approaches have been undertaken to characterize and assess drug-metabolizing enzyme activity in humans and to predict the rate of metabolism of a drug *in vivo* in individual subjects. The use of marker drugs (substrate probes) has found relatively wide application, in particular in determining changes in enzyme activity as caused by induction or inhibition as assessed in longitudinal studies in the same subjects [10]. One of the major shortcomings, however, of any marker drug currently used to measure oxidative enzyme activity *in vivo* is that the results obtained with one drug only rarely possess predictive value towards other drugs. For example, several attempts have been made to correlate plasma elimination rates (half-lives) or plasma clearance values of antipyrine with those of another oxidized drug in the same panel of subjects. In most studies the correlations were low or non-existent [11]. A number of possible causes may explain these findings, of which the multiplicity of the cytochrome P–450 system is the most important one. Multiplicity implies that different compounds may be oxidized by different isozymes (intersubstrate selectivity) and/or that from a single compound different metabolites may be produced by different isozymes (intrasubstrate selectivity). The latter, is, for instance, likely to be the case for antipyrine. The formation of its major primary metabolites, 3-hydroxymethylantipyrine (HMA), norantipyrine (NORA) and 4-hydroxyantipyrine (OHA), is selectively inducible in rats and to a lesser extent also in man [12]. In pigs OHA is the only metabolite that is found in appreciable quantities, whereas HMA and NORA are almost totally absent [13]. Furthermore, the occurrence of genetic deficiences in oxidative drug-metabolizing enzyme activity supports the fact that certain isozymes exhibit considerable selectivity in substrate binding and metabolite formation [14–17].

Fig. 2 represents some of the isozymes of the cytochrome P–450 system which are independently regulated and which have different, but in many cases overlapping, substrate and product selectivities. The arrows represent single primary metabolic pathways with metabolite formation of some frequently used probe drugs *in vivo*. While some substrates (for example, debrisoquine, sparteine) are predominantly metabolized by the same isozyme [18], phenazone oxidation occurs independently of both debrisoquine or sparteine metabolism [19] and of nifedipine [20]. The scheme presented in Fig. 2 is not complete but it is of conceptual interest, since it suggests that several different probe drugs (or a limited number of polyfunctional substrates like phenazone), whose rate of metabolite formation is assessed, are required to obtain an overall picture of oxidative drug-metabolizing enzyme activity *in vivo* [21].

'COCKTAIL' STRATEGY

The 'cocktail' strategy is characterized by giving combinations of substrate probes simultaneously (in relatively low doses) and measuring either plasma kinetics of unchanged drugs or urinary excretion of metabolites or both. The kinetic parameters that most closely reflect enzyme activity *in vivo* are the intrinsic clearance of unchanged drug or the clearance for production of separate metabolites [11]. Intrinsic clearance can be determined by measuring the area under the plasma concentration versus time curve (AUC) after oral administration of the drug: CL(int)=dose/AUC and clearance of formation of metabolite: CL(→ met)= *f(m)*. CL(int); *f(m)* is the fraction of dose excreted into urine as that metabolite. So-called metabolite ratios are often used in phenotyping panels of subjects for the existence of genetically determined deficiences in drug-oxidizing enzyme activity in order to differentiate between poor and rapid metabolizers. They are, however, not parameters which reflect enzyme activity quantitatively and are therefore less suitable in metabolic correlation studies, which are meant to elucidate common regulation (co-segregation) of metabolic pathways of different substrates [3].

There are in principle at least two major advantages associated with the cocktail approach: a practical one and a conceptual one. It is obviously of great practical advantage when, for example, phenotyping for different types of polymorphism can be achieved in one session. Conceptually this approach is of advantage in metabolic correlation studies, because it does exclude the influence of intraindividual variability in metabolizing enzyme activity with time. In the past in such studies two or more drugs were given to the same panel of subjects, but on different occasions. Although oxidative drug-metabolizing enzyme activity seems to be primarily controlled by genetic factors, there may still be a considerable influence of environmental factors which potentially may cause intraindividual differences in activity from one point in time to another. A complication of the cocktail approach is that several drugs and metabolites have to be analysed in the same biological samples, which requires a high degree of

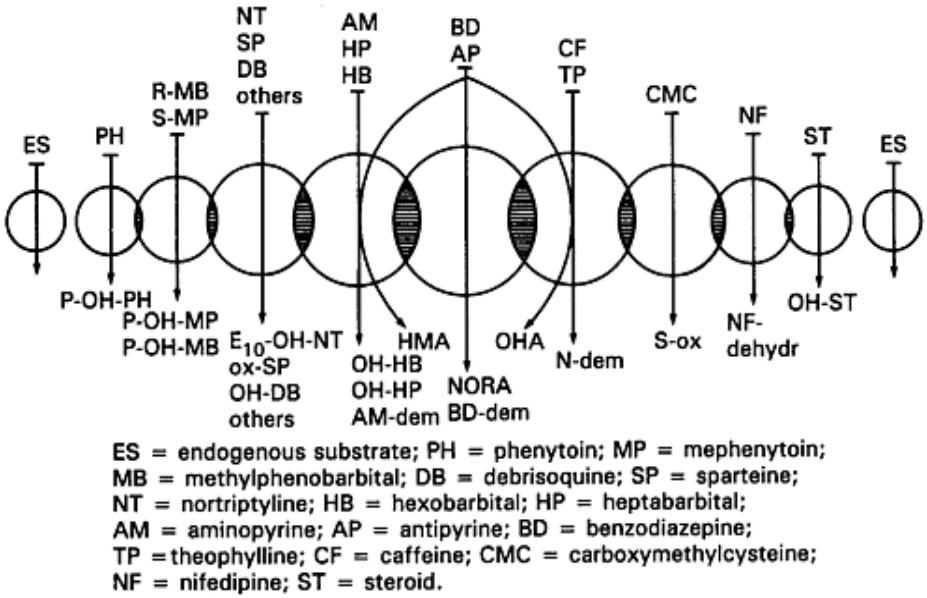


Fig. 2. The multiplicity of the cytochrome P–450 system. The circles represent the different isozymes and the arrows single metabolic pathways of the various drugs (from [21]).

selectivity and sensitivity of the analytical methodology required for their determination. With the currently available techniques, however, that is quite feasible. In recent years we have applied the cocktail approach for the following purposes:

(1) To assess in a single session the incidence of the poor-metabolizer phenotypes for sparteine and mephenytoin, and the variability in nifedipine metabolism, in a Dutch population of 172 subjects [20]. A 7.4% incidence of poor metabolizers of sparteine was detected, which is quite similar to that found in other Caucasian populations. For mephenytoin 2.3% of this population was found to poorly metabolize it to *para*-hydroxymephenytoin. In a similar study in 130 healthy subjects a cocktail of phenytoin, sparteine and nifedipine was administered [22]. The results of this study for nifedipine have been presented in Fig. 1, whereas a similar extent of variability in the plasma kinetic (AUC) for phenytoin was observed. Correlations between relevant kinetic and metabolic parameters of the three probe drugs were all low and non-significant. None of the data of nifedipine and phenytoin were different between extensive and poor metabolizers of sparteine. Thus the major oxidative metabolic pathways of nifedipine, sparteine and phenytoin are not related to each other. The three compounds can in principle be used

in combination to characterize the activity of different cytochrome P–450 enzymes *in vivo* in man in one experimental session.

(2) To elucidate metabolic pathways of different drugs which are mediated by the same enzyme(s) through correlation studies. A very strong correlation of kinetic parameters reflecting enzyme activity of different drugs is highly indicative of a common enzyme. Such investigations have been shown to be feasible in rats, in healthy subjects and in patients with liver disease, in correlating the rate of single metabolic reactions of different probe drugs with each other [23–26]. In this respect it is important to take, for example, the different primary metabolic pathways of antipyrine into consideration, which are probably mediated by different P–450 enzymes (Fig. 2). In Fig. 3 these pathways are shown, with their chemical structures. In correlation studies including antipyrine, hexobarbital and theophylline strong indications have been obtained that theophylline metabolism and 4-OH-antipyrine formation are mediated by the same enzyme(s), whereas hexobarbital metabolism is closely related to norantipyrine and 3-OH-methylantipyrine formation. This suggests that the rate of antipyrine metabolite formation has predictive values towards the rate of metabolism of those other drugs.

A similar correlation approach (but not as a 'cocktail') was used to elucidate similar rate-limiting steps in the metabolism of dihydropyridine calcium channel blockers. Racemic felodipine, racemic nitrendipine and nifedipine were investigated in a randomized cross-over study in healthy subjects, using stereoselective enantiomers. The high correlations between the AUCs of all compounds strongly suggest the involvement of the same or very similar enzyme(s) in their primary oxidative metabolism, which in all cases involves aromatization of the dihydropyridine ring structure [22].

(3) To assess selectivity in the inducing and inhibiting effects of environmental factors (for example, drug treatment) on the activity of the different P–450 enzymes. In one of such studies, nifedipine, sparteine, mephenytoin and antipyrine were administered simultaneously to 15 healthy subjects, including four poor metabolizers of sparteine and four poor metabolizers of mephenytoin [6]. They received the 'cocktail' on three different occasions: without pretreatment, after pentobarbital pretreatment and together with cimetidine. Concentrations of nifedipine, its pyridine metabolite, and of sparteine and its dehydro metabolite were measured in the plasma; nifedipine metabolites, sparteine, dehydrosparteine, 4-hydroxymephenytoin, antipyrine and its three major metabolites were all measured in urine. The kinetic parameters obtained clearly indicated that nifedipine metabolism is very sensitive to pentobarbital induction (Fig. 4), whereas antipyrine metabolism is sensitive only to a moderate degree and sparteine metabolism is unaffected. Cimetidine inhibited the metabolic clearance of all three compounds to a similar extent and also inhibited renal clearance of sparteine, most likely at the level of tubular secretion. Urinary excretion of 4-

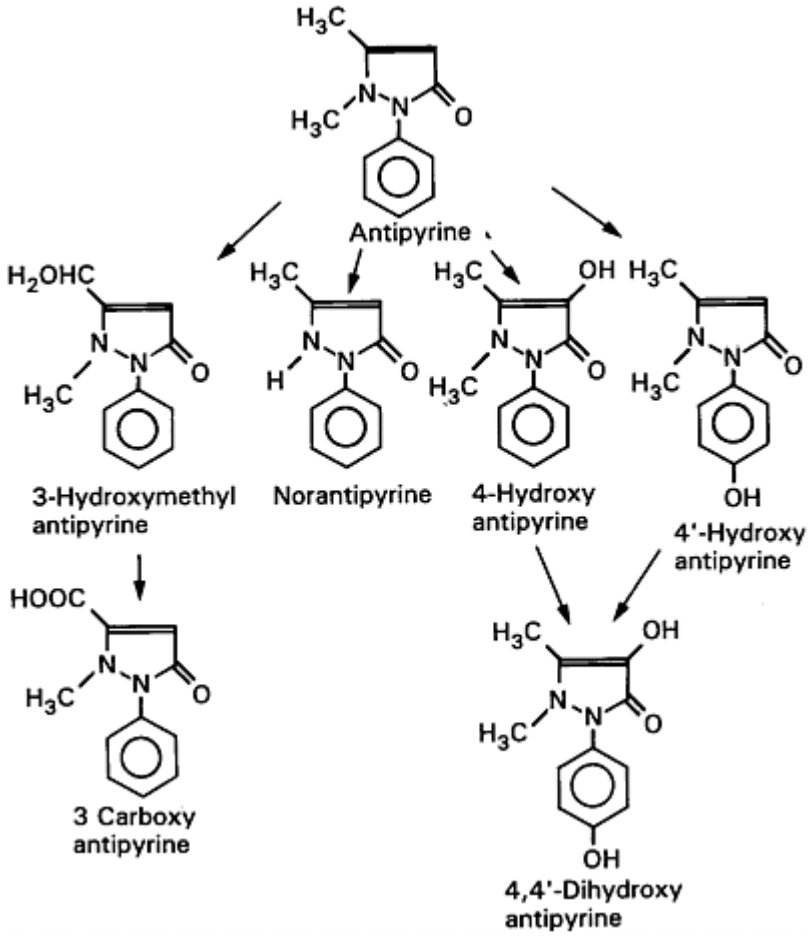


Fig. 3. The different oxidative metabolic pathways of antipyrine.

hydroxymephenytoin was not affected by either treatment, nor was the metabolic ratio of sparteine/dehydrosparteine in 8 hour urine [6, 27]. The results of this study have clearly illustrated that the 'cocktail' allows the assessment of the differential effects of drug treatment on oxidative enzyme activity. This approach could also be useful in new drug development in order to assess whether or not a new compound will give rise to potential risks of drug-drug interactions through induction or inhibition of drug metabolism. Rather than performing several different studies with different (probe) drugs, the 'cocktail' approach should allow sufficient pertinent information to be obtained in the context of one experimental protocol.

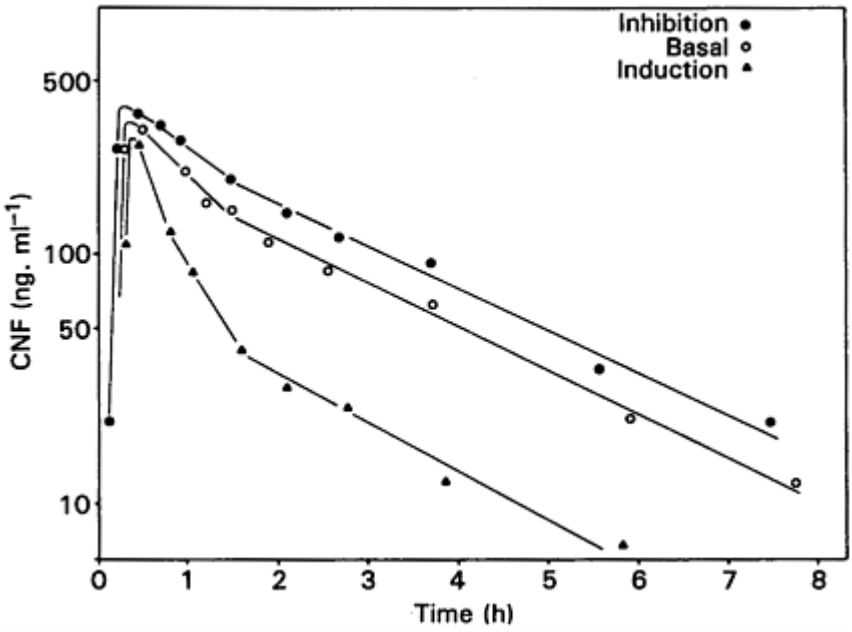


Fig. 4. Plasma concentration versus time curves of nifedipine in an individual subject after oral intake of 20 mg nifedipine on three different occasions: without pretreatment (○), after enzyme induction (▲) with pentobarbital and inhibition (●) by cimetidine (from [6]).

CONCLUSIONS

Methodology is becoming available that makes it feasible to characterize and possibly also predict oxidative drug-metabolizing enzyme activity *in vivo* in man. The multiplicity of the P–450 enzyme system dictates that different probe drugs are used for this purpose. By combining suitable probe drugs ('cocktail' strategy) it is possible to obtain relevant information in one experimental session on the activity of different enzymes. This approach can be extended to, or combined with, probe drugs representing other enzyme systems (for example, those involved in phase II reactions).

REFERENCES

1 [] Vesell, E.S. (1977) *Clin. Pharmacol. Ther.,* **22,** 659.
2 [] Breimer, D.D. (1983) *Int. J. Clin. Pharmacol. Res.,* **III,** 399.
3 [] Breimer, D.D. (1983) *Clin. Pharmacokinet.,* **8,** 371.
4 [] Clark, D.W.J. (1986) *Drugs,* **29,** 342.
5 [] Schellens, J.H.M., Sons, P.A. and Breimer, D.D. (1988) *Biochem. Pharmacol.,* **37,** 2507.

6 [] Schellens, J.H.M., van der Wart, J.H.F., Brugman, M. and Breimer, D.D. (1988) *J. Pharmacol. Exp. Ther.,* **249,** 638.
7 [] Bailey, D.G., Spencer, J.D., Munoz, C.D. and Arnold, J.M.D. (1991) *Lancet,* **337,** 269.
8 [] Soons, P.A., Vogels, B.A. P.M., Roosemalen, M.C.M., Schoemaker, H.C., Uchida, E., Edgar, B., Lundahl, J., Cohen, A.F. and Breimer, D.D. (1991) *Clin. Pharmacol. Ther.,* in press.
9 [] Guengerich, F.P. and Kim, D.-H. (1990) *Carcinogenesis,* **11,** 2275.
10 [] Vesell, E.S. (1979) *Clin. Pharmacol. Ther.,* **26,** 27.
11 [] Breimer, D.D. *et al*. (1984) in *Pharmacokinetics: A Modern View* (eds Benet, L.Z., Levy, G. and Ferraiolo, B.L.), Plenum Press, New York, p. 191.
12 [] Teunissen, M.W.E. (1984) Factors and conditions affecting antipyrine clearance and metabolite formation, PhD thesis, University of Leiden, The Netherlands.
13 [] Van den Broek, J.M., Teunissen, M.W.E. and Breimer, D.D. (1981) *Drug Metab. Dispos.,* **9,** 541.
14 [] Guengerich, F.P. *et al.* (1986) *Xenobiotica,* **16,** 367.
15 [] Guengerich, F.P. (1989) *Annu. Rev. Pharmacol. Toxicol.,* **29,** 241.
16 [] Gonzalez, F.J. (1989) *Pharmacol. Rev.* **40,** 243.
17 [] Wilkinson, G.R. (1989) in *Topics in Pharmaceutical Sciences 1989* (eds Breimer, D.D., Crommelin, D.J.A. and Midha, K.K.), International Pharmaceutical Federation (FIP), The Hague, p. 541.
18 [] Wilkinson, G.R., Guengerich, F.P. and Branch, R.A. (1989) *Pharmacol. Ther.,* **43,** 53.
19 [] Danhof, M. *et al.* (1981) *Pharmacology,* **22,** 349.
20 [] Breimer, D.D., Schellens, J.H.M. and Soons, P.A. (1988) in *Microsomes and Drug Oxidation* (eds Miners, J.O., Birkett, D.J., Drew, R., May, B.K. and McManus, M.E.), Taylor & Francis, London, p. 232.
21 [] Breimer, D.D. and Schellens, J.H.M. (1990) *Trends Pharmacol. Sci.,* **11,** 223.
22 [] Soons, P.A. Variability and stereoselectivity in pharmacokinetics, metabolism and effects of dihydropyridine calcium entry blockers in man, PhD thesis, University of Leiden, The Netherlands.
23 [] Van der Graaff, M., Vermeulen, N.P.E., Hofman, P.H. and Breimer, D.D. (1983) *J. Pharmacol. Exp. Ther.,* **227,** 459.
24 [] Teunissen, M.W.E., de Leede, L.G.J., Beoijinga, J.K. and Breimer, D.D. (1985) *J. Pharmacol. Exp. Ther.,* **223,** 770.
25 [] Schellens, J.H.M., van der Wart, J.H.F., Danhof, M., van der Velde, E.A. and Breimer, D.D. (1988) *Br. J. Clin. Pharmacol..,* **26,** 372.
26 [] Schelles, J.H.M., Janssen, A.R., van der Wart, J.H.F., van der Velde, E.A. and Breimer, D.D. (1989) *Eur. J. Clin. Invest.,* **19,** 472.
27 [] Schelles, J.H.M. Characterization of human drug oxidation *in vivo:* development of a 'cocktail' study design, PhD thesis, University of Leiden, The Netherlands.

6
The role of computational chemistry in molecular modelling

J.P.Tollenaere

Department of Theoretical Medicinal Chemistry, Janssen Research Foundation, 2340 Beerse, Belgium

SUMMARY

Molecular modelling and computational chemistry, linked together through the advent of high-performance computers and the advances in software, are currently employed successfully to simulate the structure and properties of drug molecules.

A review is presented of the techniques of conformational searching such as the systematic grid search, minimization schemes and molecular dynamics. Also discussed are the relative merits and shortcomings of the various semi-empirical quantum mechanical approximations such as CNDO, MNDO, AM1, PM3 and PCILO. The problem of whether the current force fields are satisfactory will be addressed and some recent progress in the area of second-generation force fields are discussed. Finally, some examples will be presented illustrating the progress that has been made in our ability to calculate free energy differences.

All this is put in the perspective of the goal of molecular modelling, which is to provide insight into complex molecular phenomena and which stimulates the creativity of all those involved in innovative drug research.

INTRODUCTION

During the last decade, molecular modelling has become a well-established technique owing its use in the chemical laboratory to the spectacular developments in computer hardware and the advances in software. Whereas in the early 1980s molecular modelling was largely a technique to visualize and manipulate chemical structures, the last few years saw progressing integration of molecular modelling and computational chemistry. In particular, the advent of high-performance RISC (reduced instruction set computer) workstations replacing to a large extent the department minicomputer has given a powerful boost to the computational power of the theoretical chemist. In fact, the present-day super-workstations offer the raw computer power of the mainframes of the 1980s at a price level affordable to virtually any interested researcher.

Because computational chemistry and the underlying methodologies are now at the heart of molecular modelling it is worthwhile to scrutinize the various computational schemes and approximations that are employed to simulate the properties of a molecule. Because the accuracy with which molecular properties can be computed very much depends on the size of the molecular system, a number of problems are easily identified.

The first fundamental problem and probably the most difficult one for macromolecular structures is that of finding the global minimum. Structures such as oligopeptides and proteins having many hundreds or thousands of atoms possess a multitude of possible low-energy conformations.

The second problem, closely associated with the study of proteins and nucleic acids is the use of empirical force fields (FFs) or the molecular mechanics (MM) method. Can the MM method, treating a molecular system as a collection of balls and springs, be employed to reliably describe the structure of small molecules and the interaction energetics of macromolecular systems?

Thirdly, the quantity of real relevance is not the enthalpy but actually the free energy of the molecular system which is needed for the correlation with experimental energetics of drug-receptor complexes.

In this contribution, a brief review will be presented of the techniques of conformational searching such as the systematic grid search, minimization schemes and molecular dynamics. Also discussed will be the relative merits and shortcomings of the various semi-empirical quantum mechanical approximations such as CNDO, MNDO, AM1, PM3 and PCILO.

The problems involved with the current FFs will be addressed and the direction taken in the area of second-generation FFs will be discussed. Molecular dynamics (MD) calculations as a means to locate minimum energy conformations and to study time-dependent phenomena will be described, as well as free energy perturbation (FEP) calculations.

All this can be put in the perspective of the goal of molecular modelling, which is to provide insight into complex molecular phenomena and which stimulates the creativity of all those involved in innovative drug research.

MOLECULAR ENERGY

Ever since the formulation of quantum mechanics in the 1920s by Schrödinger, Dirac and many others, numerical solutions to the Schrödinger equation have been developed to describe accurately the energy of a molecular system. Extremely accurate predictions are nowadays possible for small systems. The larger the molecule the less rigorous the formulation tends to become. The present-day armoury at the disposal of the computational chemist ranges from highly sophisticated large basis set *ab initio* techniques over semi-empirical allvalence electron molecular orbital (MO) to MM or FF methods. The choice as to which method is to be used is largely dictated by the size and the nature of the problem in hand. For all practical purposes the size of the problem is closely linked

to the available computational resources. Similarly, the nature of the problem will influence the choice of computational sophistication used for the description of the molecular system.

In this section a brief discussion will be presented of the computational schemes which are of practical interest to the computational medicinal chemist. This restriction rules out to a large extent *ab initio* MO methods which are only amenable to systems whose size is hardly of interest to the practising medicinal chemist.

All-valence electron molecular orbital methods

Ever since the introduction of the semi-empirical Extent Huckel Theory (EHT) by Hoffman [1] in the early 1960s and followed by the Complete Neglect of Differential Overlap (CNDO) method [2], there has been a constant evolution in the field of semi-empirical MO methods. This research has evolved into a situation where at present the computational chemist using the recent Modified Neglect of Differential Overlap (MNDO) [3] version and its more elaborate successors Austin Model (AM1) [4] and Parametrized Model (PM) [5] has a number of impressive tools at his or her disposal to describe the molecular structure and properties of molecules containing several hundreds of atoms.

Like their related methods such as the Intermediate Neglect of Differential Overlap (INDO) [2] and Modified Intermediate Neglect of Differential Overlap (MINDO) [6] the methods already mentioned have a number of features in common. They are all self-consistent field (SCF) methods that take into account electrostatic repulsion and exchange stabilization and where all integrals are evaluated by approximate means. They are further characterized by the fact that they all use a restricted basis set, that is, one s orbital and three p orbitals per atom, except for the hydrogen atom, which is represented by an s orbital.

Because of their widespread use, not least because most of them are implemented in the MOPAC [5] single program package, what follows is a brief enumeration of the strong points and drawbacks of MNDO, AM1 and PM3.

MNDO, being the oldest of the three, is on all counts the least accurate. From the medicinal chemist's point of view one of its most serious limitations is the fact that hydrogen-bonded systems cannot be treated. A further disturbing characteristic is the systematic failure of MNDO to treat oxygenated substituents on aromatic rings. The most telling example is nitrobenzene, where the plane of the NO_2 group is predicted to be perpendicular to the phenyl plane as the most stable rotamer with a rotational barrier of the order of 2 kcal/mol and a calculated heat of formation ΔH_f=20 kcal/mol—too high compared to the experimental ΔH_f [5]. In general, sterically crowded molecules are calculated to be too unstable and the COC angle in ethers is too large by approximately 9°.

AM1, introduced in 1985, is a significant improvement over MNDO in several respects. In particular, the hydrogen bond problem has been overcome in the sense that, for example, the hydrogen bond interaction energy of the water dimer

is calculated to be 5.5 kcal/mol in line with experimental data [5]. Compounds such as nitrobenzene are now predicted to be planar, whereas orthogonal and planar benzaldehyde are equi-energetic. On the other hand, the calculated ΔH_f of nitrobenzene is still 10 kcal/mol too high, whereas the calculated ΔH_f=−8.9 kcal/mol of benzaldehyde is virtually identical to the experimental value of −8.8 kcal/mol.

PM3, which is the most recent addition in the arsenal of semi-empirical MO methods [5] appears to be an improvement over AM1 as far as the accuracy of ΔH_f is concerned. Nitrobenzene, predicted to be planar, is an example where the calculated ΔH_f=14.5 kcal/mol is very near the experimental value of 15.4 kcal/mol. Much more experience must be gained with PM3 in order to judge its overall accuracy relative to AM1.

This brief account of the various semi-empirical MO methods would not be complete without mentioning the Perturbative Configuration Interaction using Localized Orbitals (PCILO) [7]. PCILO differs from the methods discussed above in the fact that it is not an SCF method but one which is based on Raleigh-Schrödinger perturbation theory, in which third-order terms of the series are included. PCILO has a long-standing record of success in the correct prediction of the minimum energy conformation of a variety of small molecules [8, 9]. Due to its non-SCF nature, PCILO is considerably faster than the other semiempirical methods, despite the fact that owing to its third-order-based perturbation it includes some electron correlation effects.

In summary, all semi-empirical MO methods discussed above have their strong and weak points, which unfortunately are not always known prior to the actual calculation. As with all evaluations of theoretical methodologies, making contact and comparing with experimental data is always a wise thing to do. It should be borne in mind that quantum chemical calculations are most often done on the isolated molecule, that is, in absolute vacuum. More often than not, the results of these calculations are compared to the results obtained from X-ray (crystalline state) or nuclear magnetic resonance (NMR) data (liquid state), from which it follows that discrepancies between theoretical and experimental results are not necessarily a reflection on the validity and accuracy of the former.

Molecular mechanics

As molecular systems become larger the less rigorous are the methods used to describe their chemical and physical properties. As molecular structures containing many hundreds or thousands of atoms are far beyond the reach of even the fastest super-computers, drastic simplifications are needed for a reasonable description of chemicals such as oligopeptides and proteins.

The basic assumption in MM methods is that a molecule can be considered as a collection of classical Newtonian masses held together by hypothetical springs holding the atoms at an equilibrium position in configuration space [10]. The generic form of an MM force field is given by

$$V = \Sigma V_{\text{bonds}} + \Sigma V_{\text{angles}} + \Sigma V_{\text{tors}} + \Sigma V_{\text{vdW}} + \Sigma V_{\text{coul}} \quad (1)$$

where V_{bonds} and V_{angles} are the contributions to the total potential energy V arising from bond stretching and valence angle bending, respectively. The V_{tors} term reflects the torsional angle strain in the molecule and V_{vdW} represents the contribution from the non-bonded dispersion and repulsion interactions and V_{coul} the coulomb electrostatic interactions.

A typical FF potential is given by

$$V = \Sigma K_r (r - r_0)^2 + \Sigma K_\theta (\theta - \theta_{eq})^2 + \Sigma \frac{1}{2} K_\varphi [1 + \cos(n\varphi - \gamma)] + \sum_{i<j} \frac{b_{ij} - a_{ij}}{r_{ij}^{12} r_{ij}^{6}} + \sum_{i<j} \frac{1}{\varepsilon} \frac{q_i q_j}{r_{ij}} \quad (2)$$

where K_r, K_θ and K_φ pertain to the empirical force constants for bond stretching, bond angle deformation and torsion angle twisting, respectively. The constants a_{ij} and b_{ij} are associated with the dispersion and repulsion contribution, respectively, between atoms i and j at a distance r_{ij}. The last term is the coulombic interaction between charges q_i and q_j in a dielectric medium ε.

More elaborate FFs include out-of-plane deformation terms and cross-terms reflecting, for example, coupled bond stretching and bond angle deformations.

As it is beyond the scope of this chapter to present an in-depth analysis and evaluation of the current FFs in use, it suffices to bear in mind that interesting developments [11–17] are taking place in the generation of the so-called second-generation FFs. In fact, a number of basic questions can be put forward by looking at equation (2). These include, among others, the question of the assumed isotropy of the dispersion and repulsion constants a_{ij} and b_{ij} [18]. Experimental evidence suggests that atoms in molecules are ellipsoidal rather than spherical in shape [12, 20]. Another disturbing aspect of equation (2) is the absence of an adequate description of atomic polarizability [18] and the as yet unsolved problem of an appropriate treatment of the dielectric constant ε of the intervening medium [21].

Because the development of older FFs suffered from the lack of experimental data to accurately determine the various empirical constants, new FFs are developed on the basis of fitting analytical FF expressions such as equation (2) to accurate *ab initio* results. It turns out that these second-generation FFs not only contain harmonic but also cubic and quartic terms combined with a (6–9) *v*d*W* term instead of the familiar Lennard-Jones (6–12) potential [16].

CONFORMATIONAL SEARCH METHODS

One of the fundamental problems in the computer-aided simulation of a molecular system is that of finding the global energy minimum [22–25]. In the case of small molecules with a few torsional degrees of freedom exhaustive methods may readily lead to the global minimum energy conformation. With

larger molecules such as oligopeptides or proteins containing hundreds to thousands of atoms having a correspondingly large number of conformational degrees of freedom, various strategies for locating the global energy minimum must be employed.

Systematic grid search

For molecules having n torsional degrees of freedom the systematic variation of the n torsional angles in steps of $\Delta\alpha$ leads to the generation of N_{conf} conformations:

$$N_{\text{conf}} = \left(\frac{360}{\Delta\alpha}\right)^n$$

with energies E_i:

$$E_i = f(\tau_1, \tau_2, \ldots, \tau_n)$$

It is readily seen that the systematic grid search rapidly leads to a combinatorial block. For example, n=4 and $\Delta\alpha$=10° results in the generation of $N=36^4 \approx 1.6\times10^6$ conformations. If this brute-force approach may still be feasible under the rigid body rotation (non-adiabatic) approximation, it rapidly becomes beyond the capacity of present-day computers when full geometry relaxation (adiabatic) is employed for each conformation generated. In this particular application comfort may come from coarse-grained parallelized computer systems where many hundreds or even thousands of processors may share the computing burden.

Energy minimization techniques

Finding the point in configuration space where all the forces on the atoms are balanced is not a trivial task for systems with a large number of conformational degrees of freedom. However, quite a few strategies for finding the minimum energy conformation (MEC) are available. Before giving a brief discussion of each of them it is extremely important to realize that all standard energy minimization methods always proceed downhill. This means that these methods always lead to the minimum nearest to the starting conformation. Thus the energy minimum and corresponding structure is strongly dependent on the starting conformation and may or may not be the global energy minimum.

As it is beyond the scope of this chapter to elaborate on the mathematical details of each of the minimization algorithms, the following discussion merely serves to indicate the pros and cons of each of them.

Steepest descents

The steepest descents (SD) algorithm uses the gradient or derivative of the energy function as the direction towards the minimum. Using simple gradients leads to oscillations on the way to the minimum, and its convergence slows near the minimum. However, SD is a robust minimizer even for systems far from

harmonic behaviour. SD is therefore often used for the initial minimization of structures far from the minimum or for highly strained systems.

Conjugate gradients

The conjugate gradients (CG) algorithm improves on the SD algorithm by minimizing along directions which are mutually conjugate so that a next-direction vector does not undo earlier progress made by a previous step. In doing so each successive step continuously refines the direction toward the minimum. CG is very well suited for large systems and converges more efficiently than SD.

Newton-Raphson

The Newton-Raphson (NR) algorithm not only applies the gradient (first derivative) but also the second derivative to predict the curvature of the function or, in other words, to predict where the function will pass through a minimum. Because NR performs best at harmonic surfaces, it is far from ideal to use NR for systems far from the minimum. In addition, the calculation of the second derivatives (Hessian matrix) is costly in computer time and requires additional memory storage and is therefore only suited for molecular systems of limited size.

Constrained and restrained minimization

By including additional energy terms in the target function such as equation (2), one can direct the minimization process toward a specific goal. Constraints are degrees of freedom which are not allowed to vary during the minimization. A restraint is an additional term which biases the minimization toward a certain value angle to adopt a predefined value. By systematically changing the predefined torsion angle one may explore the conformational space in an adiabatic fashion whereby the whole structure is fully minimized for each of the values of the predefined torsion angle. In a similar way, distances may be restrained to certain values obtained, for instance, from nuclear Overhauser effects (NOE) experiments. Other restraints include template forcing and tethering atoms to points in space. Template forcing is particularly useful to establish the energy price paid by one molecule when it is forced to adopt the conformation of another molecule which is, for example, known or supposed to contain the pharmacophore of a given class of compounds exhibiting a particular pharmacological activity.

Molecular dynamics

While all methods of conformational analysis discussed thus far assume molecules to be static in nature, MD offers the opportunity to study the time-

dependent behaviour of molecules [26–29]. In MD the well-known Newton equation of motion $F_i=m_ia_i$, where F_i is the force acting on the mass m_i and a_i the acceleration of atom i. From the derivative with respect to the coordinates of the potential such as that of equation (2), the force F_i can be computed:

$$F_i = -\frac{\partial V}{\partial r_i} = m_i \frac{\partial^2 r_i}{\partial t^2} \qquad (3)$$

As equation (3) cannot be solved analytically for the n-body problem (n>2) numerical methods must be utilized to solve n-body systems. In practice, MD calculations are carried out in time steps of the order of 10^{-15} s and are repeated several thousands or even millions of times.

Analysis of a well-equilibrated molecular system for several picoseconds, the socalled MD trajectory, may yield an interesting picture of the conformational flexibility of a molecular system. Periodically minimizing conformations visited during the MD trajectory may yield lower energy conformers which would have been missed with the static minimization techniques.

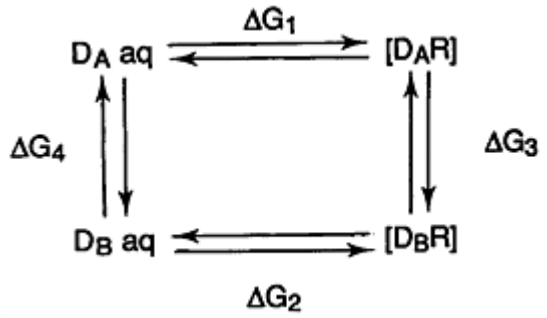
Because of computer storage limitations and the large amount of required computer times, MD calculations sampling many thousands of conformational states are limited in practice to trajectories covering picoseconds to nanoseconds. Many interesting motional phenomena in peptides and proteins, however, take place on the milliseconds to minutes time scale. Unfortunately, current mass storage devices fall short of storing and handling many thousands of megabytes of information. Equally so, number-crunching devices should be available having processing speeds at least four orders of magnitude larger than the current fastest super-computers. Notwithstanding this severe limitation, MD calculations prove to be of invaluable interest to study time-dependent properties of large molecular systems.

Free energy perturbation calculations

Although free energy perturbation (FEP) calculations are not used to evaluate minimum energy conformations, this technique is discussed here because FEP calculations involve many of the methodologies discussed thus far. Furthermore, FEP calculations open avenues to the computation of free energies which may be directly compared with experimental data such as solvation energies and binding constants of drug-receptor binding experiments. For an overview of the many applications to chemical and biomolecular systems of FEP calculations the interested reader is referred to the literature [28, 30, 31].

In FEP calculations MD is used, not to generate trajectories, but as a method to generate ensembles of conformational states permitting statistical mechanical quantities to be evaluated. The free energy difference ΔG between two states A and B of a system can be expressed as

$$\Delta G = G_B - G_A = RT \ln \langle e^{-\Delta V/RT} \rangle_A \qquad (4)$$



where $\Delta V=V_B-V_A$ is the difference of the potential energy of the states B and A, and $<>_A$ stands for the ensemble average to be taken over the reference state A. In the computational implementation of the FEP calculations a coupling parameter λ is introduced such that $V(\lambda=0)=V_A$ and $V(\lambda=1)=V_B$. Thus as λ is gradually changed from λ=0 to λ=1 the system is mutated from the A to the B state. The free energy is then given by

$$\Delta G = \sum_{\lambda=0}^{\lambda=1} \Delta G_\lambda \qquad (5)$$

Using the thermodynamic cyclethe free energy of interaction between drugs D_A and D_B with their receptor R is then given by

$$\Delta G_2 - \Delta G_1 = - RT\ln \frac{K_2}{K_1}$$

Calculating ΔG_1 or ΔG_2 is virtually impossible since it would involve the calculation of the desolvation of D_A or D_B. Since the free energy is a state function and can be calculated by any pathway, even a non-physical one, it follows that

$$\Delta G_2 - \Delta G_1 = \Delta G_3 - \Delta G_4$$

where ΔG_3 and ΔG_4 are the calculated free energies involved in the mutation of D_A into D_B complexed with R and in their solvated state, respectively. In the actual computations the drug D_A is gradually mutated into D_B by changing the coupling parameter λ=0 *(D_A)* to λ=1 *(D_B)*.

In comparing $(\Delta G_3-\Delta G_4)$ with $(\Delta G_2-\Delta G_1)$ which can bemeasured experimentally, FEP calculations offer the opportunity to make direct contact with experimental data.

MINIMIZATION OF A LARGE STRUCTURE: SOME CAVEATS

The problems involved and the precautions to be taken in a minimization study of a macromolecular structure can best be illustrated by means of an actual example.

As the G-protein coupled receptors are of fundamental importance in modern pharmacology [32] tentative physical models of the 7-transmembrane helix complexes of these type of receptors were constructed. The particular structure that was modelled on the computer consisted of seven helices with their longitudinal axes orientated in a more or less parallel fashion. The model to be energy minimized consisted of a total of 156 amino acids containing 2485 atoms. The *de novo reproducible* construction of a model of such complexity where seven helices have to be mutually orientated is not a trivial task. However, preliminary trial-and-error manipulation eventually led to a starting structure which could be minimized. As this computational experiment was also a feasibility study testing the computational capacity of the laboratory it was decided to include the more CPU-costly cross terms in the potential energy function of equation (2). The criterion for convergence of the minimization was set at $dE/dr=0.01$ kcal/Å. The calculations were performed on the Silicon Graphics 4D–220 GTX in single-processor mode using the Biosym Discover FF version 2.6.

From Fig. 1 it can be seen how the energy of the structure steadily converges after some 265 h (11 days) of CPU time. More important in Fig. 1 is the behaviour of the root-mean-square (RMS) of the energy derivative as a function of the number of iterations. Whereas the energy monotonically decreases, the RMS may remain constant or may go up during many thousands of iterations. This brings up the question of the choice of the convergence criterion. If incidentally the criterion had been set at 0.08 kcal/Å the structure would have converged at an energy of about 2780 kcal/mol after approximately 7000 iterations. Another additional 10000 iterations however, let the energy drop by some 80 kcal/mol!

Thus, if quantitative estimates and comparisons of energies of macromolecular structures or interaction energetics are to be calculated, a judicious choice of the convergence criterion is of critical importance if a physical meaning is to be attributed to the numbers which come out of lengthy and/or costly simulation experiments. Unfortunately no rule can be given that may solve the judicious choice of a convergence criterion. This, combined with the fact that all standard minimization techniques merely locate the nearest minimum, add to the notion that all alleged so-called minimum energy conformations of molecular structures in general and macromolecular structures in particular should be looked at with the utmost caution, if not suspicion.

CONCLUDING REMARKS

An overview has been presented of the various computational schemes aiming at the calculation of molecular properties such as the minimum energy conformation and free energy differences of molecules. It was seen how scientific rigour goes inversely with the size of the molecular system. Each computational formalism has its own niche where it supposedly performs best. Of the semi-empirical MO methods, the AM1 formalism and its possible successors

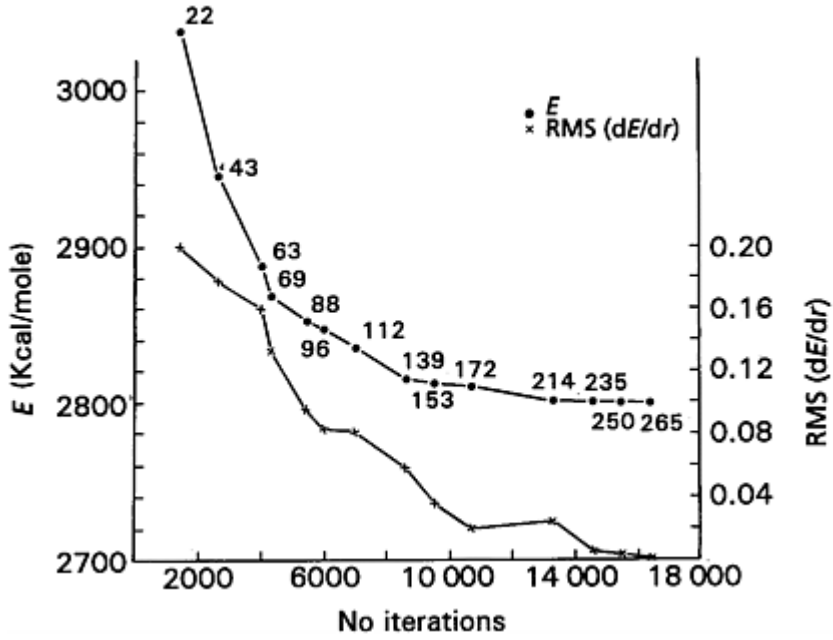


Fig. 1. The minimization energy of a 7-transmembrane helical model (156 amino acids or 2485 atoms) and the RMS of the energy derivative as a function of the number of iterations.

seem to be the most promising for further development in the direction of higher accuracy. Similarly, the development of second- and higher-generation FFs capable of describing small molecules equally well in addition to condensed phases and macromolecular systems is of utmost importance for a more accurate description of the physical properties and chemical behaviour of macromolecular systems. MD calculations covering time periods in the milliseconds or seconds range will yield a better understanding of time-dependent phenomena. FEP calculations may usher in the era allowing the computation of the long-sought fundamental quantities such as free energy and entropy and derived quantities such as sublimation energies, heat of vaporization and heat capacities.

All this critically depends on two important factors. The first is a sustained effort in the development of the theoretical framework necessary for ever-increasing accuracy and realism of the models describing the physical world of molecular systems and, secondly, the development of computer hardware and the clever implementation of modelling software. If the past is anything of an indication of the future then one may safely state that by the end of this century computational chemists will have at their disposal computing environments allowing them to describe either larger systems of real biological interest at today's level of sophistication or the molecular systems of today at a higher level of realism and accuracy.

The real impact of computational chemistry in molecular modelling comes from the researcher using these tools [33–37]. No matter the speed of computers or the sophistication of software, the relevant questions are to be asked by the researcher. Having more than a working knowledge of the theory behind it and being critically aware of the issues which still need to be resolved, the researcher may provide answers and insight into complex molecular phenomena [38]. Perhaps the even greater importance of computational chemistry in molecular modelling stems from the fact that it allows the researcher to ask pertinent questions and that it stimulates the creativity of all those involved in chemical research in general and in innovative drug research in particular.

REFERENCES

1 [] Hoffman, R.J. (1963) *J. Chem. Phys.,* **39,** 1397.
2 [] Pople, J.A. and Beveridge, D.L. (1970) *Approximate Molecular Orbital Theory,* McGraw-Hill, New York.
3 [] Dewar, M.J.S. and Thiel, W. (1977) *J. Am Chem. Soc.,* **99,** 4899.
4 [] Dewar, M.J. S., Zoebisch, E.G., Healy, E.F. and Stewart, J.J.P. (1985) *J. Am. Chem. Soc.,* **107,** 3902.
5 [] Stewart, J.J. P. (1990) *J. Comput. Aided Mol. Des.,* **4,** 1.
6 [] Baird, N.C. and Dewar, M.J. S. (1969) *J. Chem. Phys.,* **50,** 1269.
7 [] Diner, S., Malrieu, J.P., Jordan, F, and Gilbert, M. (1969) *Theor, Chim. Acta,* **15,** 100, and references cited therein.
8 [] Tollenaere, J.P., Moereels, H. and Raymaekers, L.A. (1980) in *Drug Design,* Vol. X (ed. Ariëns, E.J.), Academic Press, New York, p. 71, and references cited therein.
9 [] Tollenaere, J.P., Moereels, H. and Van Loon, M. (1986) *Drug Dev. Res.,* **8,** 141.
10 [] Burkert, U. and Allinger, N.L. (1982) *Molecular Mechanics,* American Chemical Society Monograph, 177, Washington DC.
11 [] Allinger, N.L., Yuh, Y.H. and Lii, J.-H. (1989) *J. Am. Chem. Soc.,* **111,** 8551.
12 [] Lii, J.-H. and Allinger, N.L. (1989) *J. Am. Chem. Soc.,* **111,** 8566.
13 [] Allinger, N.L., Li, F. and Yan, L.J. (1990) *Comput. Chem.,* **11,** 848.
14 [] Allinger, N.L., Li, F., Yan, L. and Tai, J.C. (1990) *J. Comput. Chem.,* 11, 868.
15 [] Maple, J.R., Dinur, U. and Hagler, A.T. (1988), *Proc. Natl Acad. Sci. USA,* **85,** 5350.
16 [] Consortium for the Research and Development of Potential Energy Functions (1991) Biosym. Technologies Inc.
17 [] Halgren, T.A. (1990) *J. Am. Chem. Soc.,* **112,** 4710.
18 [] Stone, A.J. and Price, S.L. (1988) *J. Phys. Chem.,* **92,** 3325.
19 [] Hyburg, S.C. and Fearman, C.H. (1985) *Acta Cryst.,* **B41,** 274.
20 [] Ishikawa, M., Ikuta, S., Katada, M. and Sano, H. (1990), *Acta Cryst.,* **B46,** 592.
21 [] Gilson, M.K., Rashin, A., Fine, R. and Honig, B. (1985) *J. Mole. Biol.,* **183,** 503.
22 [] Howard, A.E. and Kollman, P.A. (1988) *J. Med. Chem.,* **31,** 1669.
23 [] Lipton, M. and Still, W.C. (1988) *J. Comput. Chem.,* **9,** 343.
24 [] Chang, G., Guida, W.C. and Still, W.C., (1989) *J. Am. Chem. Soc.,* **111,** 4379.
25 [] Murall, N.W. and Davies, E.K. (1990) *J. Chem. Inf. Comput. Sci.,* **30,** 312.

26 [] McCammon, J.A. and Harvey, S.C. (1987) *Dynamics of Proteins and Nucleic Acids,* Cambridge University Press, Cambridge.
27 [] Van Gunsteren, W.F. (1989) *Mol. Simul.,* **3,** 187.
28 [] Computer Simulation of Biomolecular Systems (1989) *Proceedings of two colloquia organized by Alliant Computer Systems Corporation, December 1987 Princeton and April 1988 Amsterdam* (eds Van Gunsteren, W.F. and Weiner, P.K.) ESCOM, Leiden.
29 [] Karplus, M. and Petsko, G.A. (1990) *Nature,* **347,** 631.
30 [] Jorgensen, W.L. (1989) *Acc. Chem. Res.,* **22,** 184.
31 [] Beveridge, D.L. and DiCapua, F.M. (1989) *Annu. Rev. Biohys. Chem.,* **18,** 431.
32 [] Findlay, J. and Eliopoulos, E. (1990) *Trends Pharmacol. Sci.,* **11,** 492.
33 [] Tollenaere, J.P., Moereels, H. and Van Loon, M. (1986) *Progr. Drug Res.,* **30,** 92.
34 [] Vinter, J.G. (1986) in *Molecular Graphics and Drug Design* (eds Burgen, A.S.V., Roberts, G.C.K. and Tute, M.S.), Elsevier, Amsterdam, p. 15.
35 [] Tollenaere, J.P. (1987) in *Topics in Pharmaceutical Sciences* (eds Breimer, D.D. and Speiser, P.) Elsevier, Amsterdam, p. 355.
36 [] Cohen, N.C., Blaney, J.M., Humblet, C., Gund, P. and Barry, D.C. (1990) *J. Med. Chem.,* **33,** 883.
37 [] Richards, W.G. (1989) *Computer Aided Molecular Design,* IBC Technical Services, London.
38 [] Kollman, P.A. and Merz, K.M., Jr (1990) *Acc. Chem. Res.,* **23,** 246.

7
The importance of labelling of bioactive compounds in the development of new drugs

A.Benakis

Laboratory of Drug Metabolism, Department of Pharmacology, University of Geneva, Geneva, Switzerland.

SUMMARY

The registration of drugs requires, in practically all countries, very precise metabolic and pharmacokinetic data, both from animal and human studies. These studies currently involve the use of radioactive labelled compounds. New guidelines, recently proposed by the European Community Committee dealing with drug regulation, require whole-body autoradiography data, which can only be obtained with the use of radioactive labelled compounds. These compounds are also essential to establish other parameters such as plasma protein binding.

The synthesis of the radioactive labelled compounds needed for these studies is a very specialized discipline and it is necessary to observe a number of rules in order to obtain compounds with the desired specifications. The main points to be considered are: choice of radioactive isotope, labelling position, specific activity, radiochemical purity, number of steps used, radiochemical yield and, finally, the cost of the procedure. This latter point must be considered because the cost of obtaining radioactive labelled compounds is high, mainly because of the great amount of time required rather than because of the cost of the labelled starting materials. Nevertheless, this cost is justified by the quality of the results obtained.

Among the numerous compounds synthesized in our laboratories, the results obtained with a few representative ones will be presented in order to illustrate certain problems of labelling related to the chemical structure of the drug.

Although the procedures used for labelling by total chemical synthesis are those of organic chemistry, the labelling of natural compounds extracted from plants, the use of which has become increasingly widespread, involved a photosynthetic procedure. This consists of the cultivation of all or part of the plant under controlled conditions of light, humidity and temperature, and submitting it to $^{14}CO_2$ with the highest possible specific activity and a large amount of radioactivity in view of the enormous isotropic dilution occurring in the plant. This technique has recently been used in our laboratories and the data will be presented.

Another category of labelled compounds comprises those used for the radio-immunoassay technique applied for blood level quantification of drugs for

pharmacokinetic studies. This technique involves the use of tritiated compounds with particular specifications which will be described during the presentation.

Another application of ^{14}C-labelled compounds is their use as substrates for the evaluation of liver function. In this case, the labelled position is very important for the determination of the enzymatic mechanism involved. Results obtained from the aminopyrine breath test, an example of this type of application, will be presented.

INTRODUCTION

Although traditional analytical techniques are essential for the quantification of drugs and all other compounds related to human life and the environment, the techniques involving radioactive isotopes have now become determinant. Not only are they required by the health authorities in all countries, but there are many data that could not be obtained without them, as for example, whole-body auto-radiographic results and protein binding values.

Furthermore, traditional analytical techniques like gas-liquid chromatography (GLC), GLC/mass spectrometry, high-performance liquid chromatography (HPLC) and HPLC/mass spectrometry can only be used on compounds extracted from biological materials after their administration to animals or humans. The use of radioactive labelled techniques, on the other hand, permits the quantification of both extractable and bound compounds. Nevertheless, in order to do this, it is necessary to make additional separations so that the radioactivity can be attributed to a specific compound.

Although the use of labelled compounds in biomedical science dates back to 1945 when a number of these compounds became available [1], it was not until the 1960s that they became widely used, especially because of certain health authority requirements.

The most challenging problems related to the use of labelled compounds are the synthesis and the specifications of the compounds. These aspects have been extensively discussed in books, review papers and at international symposia [1–22].

There are two main procedures used for labelling: total chemical synthesis, which is the more important; and photosynthesis, a procedure that has recently gained importance due to the increasing use of drugs extracted from plants.

For total chemical synthesis, the type of isotope to be used for labelling is the first point that must be decided. This choice depends essentially on the chemical structure of the product to be labelled and on the type of study in which it will be applied. Among the radioactive isotopes ^{14}C, ^{3}H, ^{35}S, ^{32}P and ^{36}Cl, two are partic ularly important: ^{14}C and ^{3}H. Both of these isotopes have specific advantages and disadvantages. The other isotopes are limited to utilization in products that contain sulphur, phosphorus or chlorine in their chemical structure. Furthermore, the relatively short half-life of each of these isotopes can be a problem.

^{14}C has particularly interesting properties as a tracer isotope. It emits only β particles, which are sufficiently energetic so that radioactivity measurements can be made fairly easily using available techniques, particularly liquid scintillation counting. The half-life of ^{14}C (see Table 1) makes it unnecessary to correct for decay.

In addition, current production capacity by reactors has been developed in many countries, making it easier to obtain isotopes in adequate quantities and with high specific activities at competitive prices.

For the synthesis of ^{14}C-labelled compounds, the following points must be considered: the labelled position of the molecule, which must be chosen with regard to the type of experiment to be performed; the difficulty of the synthesis, that is, the number of steps necessary for labelling at the desired position; and finally the cost. The specificity of the labelled position must also be appropriate for the type of study and the results that are being sought. This problem is closely related to that of the biodegradation, etc., of the labelled products.

In fact, for ^{14}C-labelled compounds, if the labelled position (for example, methyl, ethyl, alkyl or carboxylic groups) is too 'exposed' to enzymatic systems, they can be rapidly and extensively metabolized and thus generate $^{14}CO_2$, which would be eliminated in the expired air. This property is used for the evaluation of liver function. Compounds such as aminopyrine, erythromycine and xylocaine have been proposed and used in the so-called $^{14}CO_2$ breath test [23].

The specific activity of the labelled product must be as high as possible and chosen as a function of the sensitivity of the radioactivity measurement techniques and the quantity of radioactivity it will be possible to introduce into the biological system being studied in the course of the experiment. A high specific activity is needed to provide mass spectra for confirming the labelled position of the labelled molecule.

The quantity of labelled compound (in mCi) to be synthesized depends on the type of study, whether it will be carried out on plants, in animals or in soil. In any case, as it is usual to underestimate the quantity needed for the entire study, it is recommended that the greatest possible quantity be synthesized.

Table 1. Radioactive isotope specifications

Radioisotopes	Half-life	β energies (max.)
^{14}C	5730 years	0.159 MeV
^{3}H	12.3 years	0.018 MeV
^{32}P	14.3 days	1.71 MeV
^{35}S	87.2 days	0.167 MeV

Other aspects of the synthesis that must be considered are the radioactive starting material, the number of steps and the radiochemical yield for each step, the chemical and radiochemical purity of the product, the stability, especially for molecules with a high specific activity, and finally the storage conditions.

The starting compound for ^{14}C-labelled molecules must be as simple as possible, not only to reduce costs, but also for easy availability and the ease with which the specific activity can be chosen. For all of these reasons, barium carbonate or calcium carbonate is the best starting material.

The procedure for the synthesis of ^{14}C-labelled compounds consists of the following four steps:

(1) Preliminary study.
(2) Preliminary non-radioactive synthesis.
(3) Preliminary labelled synthesis.
(4) Final synthesis with high specific activity.

PRELIMINARY STUDY

This consists of obtaining the necessary information regarding the specific application of the labelled compound. Careful consideration must be given to the different synthesis procedures (laboratory, pilot, industrial) that can be applied to obtain the final product as well as to the availability of the intermediate non-labelled compounds, their specifications and purity. Also, analytical data, such as chromatography results for intermediate and final compounds, must be collected. The labelling position and starting material must then be chosen. This preliminary step also establishes the amount of starting material required for the final synthesis.

PRELIMINARY NON-RADIOACTIVE SYNTHESIS

Once all the information is in hand, a preliminary non-radioactive synthesis should be performed. In this step, the techniques, material and intermediate compound in the quantities planned for the final synthesis must be used to evaluate the feasibility of the proposed synthesis. This synthesis gives information regarding the yield and specifications of the synthesis of the product, for example in terms of purity. The product thus synthesized can also be used for preliminary purification trials and determination of physicochemical data, such as ultraviolet (UV), infrared (IR), nuclear magnetic resonance (NMR) and mass spectra.

PRELIMINARY LABELLED SYNTHESIS

After completion of the previous two steps, it is recommended that a preliminary labelled synthesis be undertaken, under identical conditions, to produce around 1 mCi or less of the final labelled product with low specific activity (around 1–2 mCi/mM).

This synthesis allows a calculation to be made of the yield in terms of radioactivity. It also permits a determination to be made of the radiochemical

purity of the final product and the purification procedure required. Finally, this sample can be used for information concerning radioanalysis and storage conditions required, as well as physicochemical data such as UV, IR and mass spectra, and gives a good indication of the amount of time necessary for the synthesis and the price of the final product.

FINAL LABELLED SYNTHESIS WITH HIGH SPECIFIC ACTIVITY

As mentioned before, there are several reasons why it is important to increase the specific activity of the synthesized product, although this should not interfere with the stability of the product.

On the basis of the yield of the preliminary labelled synthesis, the amount of radioactivity of starting material necessary can be calculated. Usually, the final specific activity should be around 20–40 mCi/mM.

After the synthesis, purification is the next important step. Numerous analytical techniques must be used for determining the chemical and radiochemical purity of the product. It is further recommended that for greater sensitivity different techniques, for example NMR, IR, UV and radioactivity measurement by liquid scintillation, be used for determining the specific activity of the molecule. Finally, the determination of the physicochemical data, that is the labelled position, must be performed by mass spectrometry, provided the specific activity is high enough, of course.

Another problem is that of the evaluation of the radiochemical stability of the compound with different specific activities. Very often it is necessary to make further purifications. It is advisable that any quantity of the product not required for immediate utilization be stored under optimum conditions in order to avoid radiolysis.

COST

The question of the final cost of the synthesis is an important one and it must be carefully evaluated. This evaluation must factor in not only the cost of the starting materials but, even more important, the cost of the labour which naturally will depend on how long the synthesis and purification(s) will take to perform. The cost of these syntheses tends to be high because we are dealing with compounds with complicated chemical structures [24].

EXAMPLES OF A ^{14}C-LABELLED SYNTHESIS

Phenobarbital, a well-known hypnotic and antiepileptic agent, is a good example of a labelled synthesis [25]. The number of steps necessary for the labelling depends on the position of labelling; that is, labelling in the 2 position of the

pyrimidinic ring only requires one step, whereas labelling in the 5 position requires seven steps.

— In order to label in the 2 position, [^{14}C]urea is condensed with ethyl phenyl malonic acid diethyl ester, according to the plan in Fig. 1.
— In order to label in the 5 position, the steps shown in Fig. 1 are necessary.

The starting material is [^{14}C]benzoic acid and the following products are synthesized:

[7-^{14}C]benzyl alcohol
[7-^{14}C]benzyl chloride
[2-^{14}C]phenyl acetic acid ethyl ester
[2-^{14}C]phenyl malonic acid diethyl ester
[2-^{14}C]phenyl ethyl malonic acid diethyl ester and, finally, the condensation of this ester with urea.

SYNTHESIS OF [^{14}C]DDT

Another example is the synthesis of ^{14}C-labelled DDT and once again the number of steps involved depends on the starting material. The first route starts with [^{14}C] *p*-chlorobenzoic acid in order to synthesize [^{14}C]*p*-chlorbenzoyl chloride, which yields [^{14}C]trichloromethyl *p*-chlorophenyl ketone. Next the [1-^{14}C]1-(*p*-chlorophenyl)-2,2,2-trichloroethanol reacts with anhydrous chlorobenzene and sulphuric acid (Fig. 2) [28, 29].

Another route of synthesis begins with [1-^{14}C]ethanol toobtain [1-^{14}C]chloral [30, 31]. By condensing [1-^{14}C]chloral with chlorobenzene in the presence of sulphuric acid [^{14}C]DDT can be obtained. Other labelled positions are possible, notably in the benzene rings.

The second procedure used for labelling is one that is based on photosynthesis. This procedure is used for compounds extracted from plants.

LABELLING BY PHOTOSYNTHESIS

Drugs extracted from plants have attracted increasing attention in recent years. Issuing from traditional medicine, these are an important source of new chemical structures that have allowed the development of new drugs. In spite of considerable efforts to determine the chemical structure of these compounds, total synthesis remains very difficult, if not impossible. Nevertheless, it has been possible to synthesize derivatives of the natural compounds in order to improve some of their pharmacological properties.

Despite the natural origin of these compounds extracted from plants, the registration procedure requires practically the same expertise reports as needed for synthetic compounds.

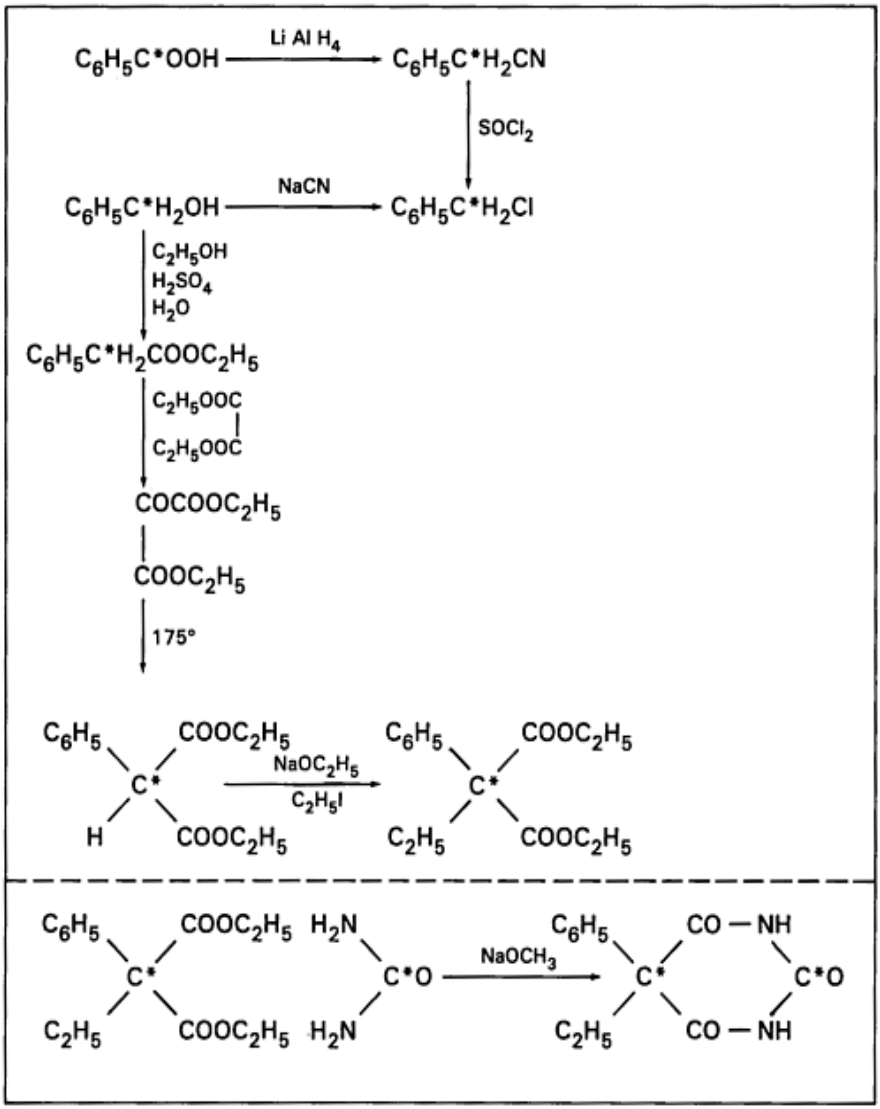


Fig. 1. Synthesis of [2-^{14}C]- and [5-^{14}C]phenobarbital.

In order to conduct a metabolic study of a new vasodilator drug, the drug was synthesized using $^{14}CO_2$ in a classical photosynthetic procedure. The experimental conditions have been described in a previous publication [33]. The same technique has recently been used for the ^{14}C-labelling of Artemisin, an antimalarial drug of Chinese origin.

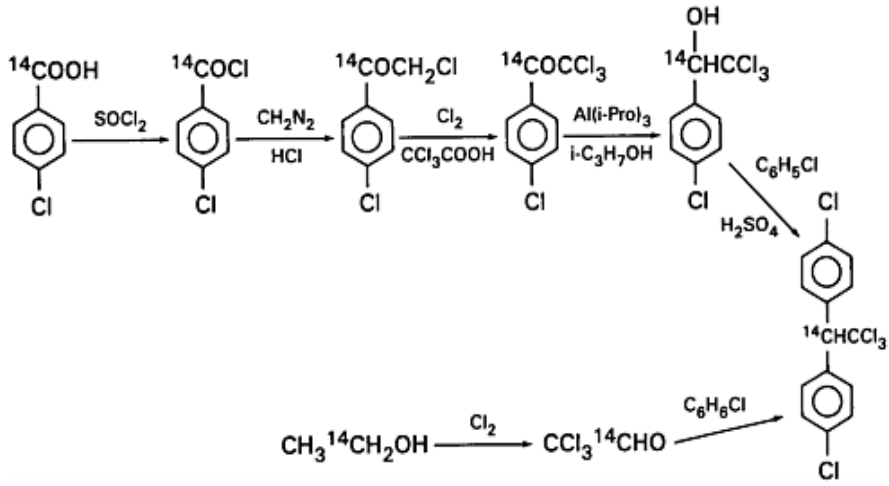


Fig. 2. Synthesis of [^{14}C]DDT.

Although the chemical structure of Artemisin has been established, total synthesis for labelling with ^{14}C in the appropriate position requires at least 17 steps, some of which are unreliable and with very poor yield. By using $Ba^{14}CO_3$ with a specific activity of 56 mCi/mM, we obtained by photosynthesis ^{14}C-labelled Artemisin with a specific activity of 1.0 μCi/mg, which was sufficient to conduct the metabolic study.

Although the technique of $^{14}CO_2$ labelling by photosynthesis can be used for any compound of plant origin, the procedure must be adapted according to the plant and the conditions of its cultivation.

Furthermore it must be noticed that:

— The yield of extractable labelled compound is often related to the geographical origin of the plant.
— The yield also varies according to the part of the plant used (proximal, distal).
— The yield as well as the specific activity of the extracted product can vary according to the cultivation condition, for example field/greenhouse/phytotron (total plants or part of plants in hydroponic cultivation), and can also be related to other parameters such as light, temperature, humidity and cultivation medium.
— Appropriate analytical techniques for quantification and particularly for separation and purification must be carefully chosen in order to assure maximum yield and purity.
— It is recommended that $Ba^{14}CO_3$ or $Ca^{14}CO_3$ (56 mCi/mM) be used as a start material, not only because it is less expensively available, but also because it has a higher specific activity.

The disadvantage of using, for a metabolic study, a compound labelled in all the carbons of the molecule is largely compensated for by the reliability of the procedure, which allows the labelled compound to be obtained in a relatively short period of time. Nevertheless, it is clear that total synthesis allowing for ^{14}C labelling at a specific position of the molecule is of great interest.

In view of the problems related to the use of a compound labelled at all of the positions of the molecule, we are now conducting studies on the simultaneous labelling with $^{14}CO_2$ and $^{13}CO_2$ in order to obtain NMR spectra that could provide useful information regarding the ^{14}C-labelled position.

TRITIUM-LABELLED COMPOUNDS

Tritium-labelled compounds are also widely used, particularly in biochemistry and in pharmacology for receptor studies. The essential problem with tritiated compounds is the integrity of the carbon-hydrogen bond and the specificity of the labelling. Yet, there are also certain advantages, for example the low price and the extremely high specific activity that can be obtained.

Although in the past the β emission created difficulties in working with these products, the advent of the use of liquid scintillation counting has resolved these.

The introduction or substitution of tritium into organic molecules is much easier than ^{14}C substitution and tritium-labelled compounds are consequently more easily prepared. Conversely, the label is also more readily lost. This aspect is particularly important in biological experiments, as the loss of tritium is not due to the biological system, even though there is some uncertainty regarding the ability of biological systems to replace the tritium with hydrogen without modification of the chemical structure.

The primary forms are elementary tritium or tritiated water. All methods of labelling with tritium involve ultimately either exchange, substitution or addition reactions. A number of these reactions are shown in Fig. 3. The primary labelling processes are as follows.

Exchange with heterogeneous catalysts

This is generally the easiest and quickest method of tritium labelling. The unlabelled organic substance is heated with, for example, tritium-labelled water or acetic acid and a catalyst such as palladium or platinum. After removal of the excess solvent, labile tritium is removed by repeated equilibration with water or other appropriate solvents and the product purified by suitable methods. This results in labelling which is general but not, as a rule, uniform, and the precise determination of the labelling is usually so laborious that it is not often attempted. The method is not applicable to compounds which are unstable under the conditions used, and it is not often possible to obtain a theoretical equilibrium concentration without excessive breakdown of the starting material. Purification

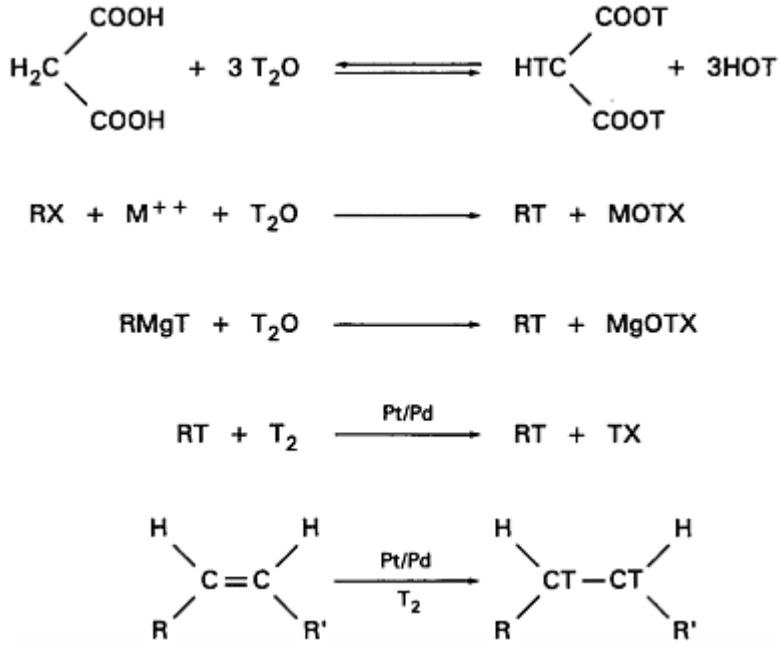


Fig. 3. Synthesis route for tritium labelling.

of the products is not usually too difficult and the method, although rather unpredictable, is generally useful.

Exchange with homogeneous catalysts

Hydrogen atoms in labile positions, for example the hydrogen of carboxylic acid, undergo rapid exchange and equilibration with tritiated water. These products are not very useful for biotransformation studies because the tritium label is very susceptible to exchange in aqueous solutions.

Wilzbach method [21, 22]

This method consists of radiation-induced labelling and often appeals to the beginner in tracer studies because of its simplicity. In its original form, the organic compound in finely pulverized form is simply exposed to elementary tritium. Hydrogen exchange occurs to a certain extent with compounds, but is complicated by secondary reactions, such as to unsaturated centres and above all by extensive radiation decomposition. This decomposition limits the specific activities attainable. A very complex mixture of labelled compounds results and special care is necessary in purification procedures. With really rigorous purification and analysis it will often be found that the desired product is present

only in very small concentration compared with labelled impurities. The success of labelling by the Wilzbach method or its variants is quite unpredictable, although it can occasionally be very successful.

Exchange by reduction

The reduction of a halogen compound by a metal catalyst such as zinc in the presence of tritiated water or tritium-labelled acetic acid will obviously produce a compound with a tritium atom in place of the original halogen. Reduction of this kind, although sometimes used for labelling, requires an excess of tritium in the water, acetic acid or other solvent.

This excess can be avoided by converting the halide into an organo-metallic compound, such as a Grignard reagent, and then having this react with an equivalent quantity of tritiated water.

Another variant is reduction, using elementary tritium and a catalyst such as platinum or palladium. This is often a very efficient method, but it should not be assumed that the labelling by this means is completely specific, as hydrogen migration can occur in the presence of catalysts. As in labelling by addition of tritium, care is needed in the selection of solvents and other conditions such as pH.

Hydrogenation to unsaturated centres

This method, particularly on carbon-carbon double and triple bonds catalysed by platinum or palladium, is a generally useful and efficient method of tritium labelling. There is always some risk that exchange or hydrogen migration will occur at the same time and it cannot be assumed that tritium will be located only at the places indicated by simple addition to the double or triple bond.

^{35}S-LABELLED COMPOUNDS

Use of ^{35}S for labelling is, of course, limited to compounds that contain sulphur in their chemical structure; however, this is often the case with many pharmaceutical products and pesticides.

^{35}S emits a soft radiation similar to that emitted by ^{14}C and the counting techniques employed are practically the same. The essential difference between these two is the relatively short half-life of ^{35}S, as shown in Table 1. This means that the specific activity and the period of use must be precisely calculated before the synthesis.

The starting materials for synthesis can be [^{35}S] sulphur dioxide ($^{35}SO_2$), [^{35}S] sulphuric acid, sodium [^{35}S]sulphate, [^{35}S]sulphuryl chloride, [^{35}S]thionyl chloride, etc.

^{32}P-LABELLED COMPOUNDS

The same comments made for ^{35}S are valid for ^{32}P labelling, in particular the importance of the programming of the synthesis and that of its use.

Starting materials can be [^{32}P]phosphorus trichloride ($^{32}PCl_3$), [^{32}P]phosphoric acid ($H_3{}^{32}PO_4$), [^{32}P]phosphorus oxychloride ($^{32}POCl_3$), etc.

This radioactive isotope can be very useful for labelling many organophosphorus pesticides.

COMPOUNDS LABELLED WITH STABLE ISOTOPES

Several stable isotopes are currently available, such as ^{15}N, ^{13}C and ^{18}O. Compounds labelled with stable isotopes are useful because they involve no radiation hazard in human experiments. The disadvantage, however, is the problem of quantitative assay. Mass spectrometry may be used, but problems of sensitivity and cost of instrumentation are limiting factors.

An emmission spectrometry technique of sufficient sensitivity has recently been introduced for quantitative determination of ^{15}N and promises to encourage the wider use of the isotope in the future.

Stable isotopes can be especially useful when associated with radioactive isotopes in double-labelled compounds. The use of stable isotopes with radioactive isotopes is an easy solution to many problems. The radioactive isotope can be applied to the separation and purification of metabolites, while the stable isotope can be applied to determine the chemical structures of the metabolites.

CONCLUSION

We have tried to emphasize not only the importance of the problem of the availability, but also that of the precise specifications of ^{14}C, ^{35}S, ^{32}P (etc.) labelled compounds, namely the specificity of the labelling position and the radiochemical purity. These two parameters are of vital importance for the future application of the labelled molecule in medicine, biology and the environmental sciences. Furthermore, two factors must be considered for the reliability of metabolic studies of drugs and pesticides: correct choice of labelling position, since a position that is too exposed can result in loss of the label and consequently incomplete results; the other factor is that of stable isotopes and the so-called isotopic effect.

Because of the high cost of synthesis, it is always best to start with the simplest possible starting material, although it must be kept in mind that the most expensive factor in synthesizing is the time spent for the synthesis, data collection and preliminary trials. Nevertheless, labelled molecules with correct and sufficient specifications not only facilitate the work of the researchers, but

also allow for experimental approaches to various problems that would otherwise be impossible.

BIBLIOGRAPHY

In addition to the publications cited here, there are numerous papers giving the principles of the synthesis of labelled compounds, appearing along with results in scientific journals of many fields. Only a relatively limited number of descriptions of synthesis of labelled compounds are to be found in the specialized publication, *Journal of Labelled Compounds and Radiopharmaceuticals,* published by John Wiley & Sons.

REFERENCES

1 [] Calvin, M. (1949) in *Isotopic Carbon: Techniques in its Measurement and Chemical Manipulation,* Wiley, New York.

2 [] Burner, M. and Schimidt, L. (1966) in *Die Synthese Kohlenstoff-14-markierter organischer Verbindungen,* VEB George Thieme, Leipzig.

3 [] Calvin, M., Heidelberger, C., Reid, J.C., Tolbert, B.M. and Yankwich, P.F. (1949) in *Isotopic carbon: Techniques in its Measurement and Chemical Manipulation,* Wiley, New York.

4 [] Catch, J.R. (1961) in *Carbon-14 Compounds,* Butterworths, London.

5 [] Catch, J.R. (1971) in *Radionuclides in Pharmacology,* Pergamon Press, Oxford, p. 97.

6 [] Evans, E.A. (1966) *Tritium and its Compounds,* Butterworths, London.

7 [] European Atomic Energy Community (1963) *Proceedings of the Conference on Methods of Preparing and Storing Marked Molecules,* Brussels, 13–16 November. European Atomic Energy Community Report EUR. 1625e.

8 [] European Atomic Energy Community (1964) *Preparation and Biomedical Application of Labelled Molecules,* EAEC, Venice, p. 23.

9 [] Feinendegen, L.E. (1967) in *Tritium-labelled Molecules in Biology and Medicine,* Academic Press, New York.

10 [] Gutlbauer, F., Getoff, N. and Wess, K. (1967) in *Osterreichische Studiengesellschaft für Atomenergie,* report SGAE-CH-48.

11 [] International Atomic Energy Agency (1962) *International Directory of Radioisotopes,* 2nd edn, IAEA, Vienna.

12 [] International Atomic Energy Agency (1960) *Tritium dosage, preparation de molecules marquées et applications biologiques,* IAEA, Vienna, p. 56.

13 [] Kaufman, W.F. *et al.* (1962) *Proceedings of the Symposium on the Detection and Use of Tritium in the Physical and Biological Sciences,* Vol. 1, IAEA, Vienna, p. 249.

14 [] Nevenzel, J.C. *et al.* (1954) Bibliography of syntheses with carbon isotopes. *California University Report,* UCLA, p. 136.

15 [] Nevenzel, J.C. *et al.* (1957) A bibliography of syntheses with carbon isotopes. *California University Report,* UCLA, p. 395.

16 [] Murray, A. and Williams, D.L. (1958) *Organic Syntheses with Isotopes. Part I. Compounds of Isotopic Carbon,* Interscience, New York.

17 [] Murray, A. and Williams, D.L. (1958) *Organic Syntheses with Isotopes. Part II. Organic Compounds Labeled with Isotopes of the Halogens, Hydrogen, Nitrogen, Oxygen, Phosphorus and Sulfur,* Interscience, New York.
18 [] Sommerville, J.L. (1962) *The Isotope Index 1962,* Scientific Equipment Co., Indianapolis.
19 [] Schütte, H.R. (1966) *Radioaktive Isotope in der organischen Chemie und Biochemie,* VEB Deutscher Verlag der Wissenschaften, Berlin.
20 [] Wenzel, M. and Schulze, P.E. (1962) *Tritium-markierung, Darstellung, Messung und Anwendung nach Wilzbach ^{3}H-markierter Verbindungen,* W. de Gruyter, Berlin.
21 [] Wilzbach, K.E. (1957) *J. Am. Chem. Soc.,* **79,** 1013.
22 [] Wilzbach, K.E. (1962) *Proceeding of the Symposium on the Detection and Use of Tritium in the Physical and Biological Sciences,* Vol. II, IAEA, p. 3.
23 [] Irving, C.S., Schoeller, D.A., Nakamura, K.I., Baker, A.L. and Klein, P. D. (1982) *J. Lab. Clin. Med.,* **100,** 356.
24 [] Pichat, L. (1971), in *Symposium des Progrès des techniques nucléaires en pharmacodynamie,* Masson, Paris.
25 [] Benakis, A. (1963) in *Proceedings of the Conference on Methods of Preparing and Storing Marked Molecules.* European Atomic Energy Community EUR 162e, p. 1283.
26 [] Blitz, Seydel (1911) *Berichte der Deutschen Chemischen Gesellschaft,* **44,** 441.
27 [] Henze, H.R. and Long, L.M. (1941) *J. Am. Chem. Soc.,* **63,** 1936.
28 [] Henze, H.R. and Nunn, L.C., Jr (1947) *J. Org. Chem.,* **12,** 543.
29 [] Buhler, D.R., Wang, C.H. and Christensen, B.E. (1953) *J. Am. Chem. Soc.,* **75,** 4336.
30 [] Murray, A. and Williams, D.L. (1958) in *Organic Syntheses with Isotopes. Part I,* Interscience, New York, p. 887.
31 [] Pearce, G.W. and Jensen, J.A. (1953) *Science,* **118.**
32 [] Murray, A. and Williams, D.L. (1958) in *Organic Syntheses with Isotopes. Part I,* Interscience, New York, p. 889.
33 [] Benakis, A., Sugnaux, F.R., Collet, G.F., Kradolfer, J.P., Berney, C., Sion, R., Necciari, J. and Cautreels, W. (1986) Carbon 14 photosynthesis labeling of natural compounds and drugs from plants. *Synthesis and Applications of Isotopically Labeled compounds 1985, Proceedings of the Second International Symposium,* p. 219.

8
Epilogue—closing remarks: trends in and future of medicinal chemistry

P.N.Kourounakis and E.A.Rekka

Department of Pharmaceutical Chemistry, School of Pharmacy, Aristotelian University of Thessaloniki, Thessaloniki 540 06, Greece

The preceding chapters—examples of modern research work in pharmacochemistry, treating the subject of drug discovery from different viewpoints—induce thoughts in the reader about the current situation, the trends in and the future of medicinal chemistry and, in particular, on the discovery of new pharmacomolecules.

Pharmaceutical chemistry is among the most rapidly advancing disciplines, with constant change of the methods and techniques used towards its principal target, the drug. Thus, we have reached the present state of the art of medicinal chemistry, with promising application of rational drug design in the discovery of drugs. This has been realized by coordinated efforts through various approaches. The art of synthetic chemistry has always been a significant preoccupation of the pharmacochemist, thus the progress of medicinal chemistry can be predicted to be sustained in discoveries in this field. However, synthetic chemistry in medicinal chemistry becomes progressively structure based, using physical organic, crystallographic techniques and computational chemistry [1, 2]. The issues in the future of medicinal chemistry are not technical, since synthetic methodologies are rapid, sophisticated and stereospecific. Structure determination permits, or will soon permit, elucidation and interpretation of drug-receptor interactions, and therefore the development of more selective drug molecules [3].

The cooperation of chemistry with the techniques and philosophy of molecular biology is becoming very productive. Chemistry is increasingly mimicking the biological world—drug-receptor interaction/host-guest chemistry, design of self-replicating molecules, whereby the formation of a molecular template catalyses its own generation from starting components (model receptors for nucleic acid components) [4–6]. The impact of the combination of chemistry with biology goes further than the synthetic, preparative methods for the production of bioactive macromolecules. It extends to the conferring of drug selectivity and to drug delivery processes [7, 8]. The molecular biological approach of medicinal chemistry can provide the fundamental definition of drug sensitivity, define the location of drug binding sites and may offer indications for the aetiology of a disease [9].

Medicinal chemistry of the near future is going to contain a significant component with a strong molecular biological character. Shortly, data from the complete human genome project will be available, and this will define the targets for the pharmacomolecules of the approaching century. Hesitantly, some positive results of this approach have appeared, for example definition of genetic deficits and gene therapy [10]. The solution to drug resistance by molecular biological approaches of medicinal chemistry has already been presented. The physicochemical properties of the receptor site of macromolecular drug targets could be determined, offer the solution of the sequence of human genes and the protein-folding problem and will make available three-dimensional drug targets. Therefore, the cloning of new receptor subtypes and of mutant receptors and enzymes obtained by site-directed mutagenesis has brought us closer to understanding the structure and function of these macromolecules.

It is only hoped that the immense capabilities that are widely opened by molecular biological approaches of medicinal chemistry will not be outweighed by various problems, limitations and restrictions—ethical, social, medical, scientific, financial—which will not be discussed here [11].

The impact of molecular graphics studies in lead generation and optimization is overwhelming. Advances in mathematical and computational sciences, coupled to the fast technological evolution of computers, will continue to offer to pharmacochemistry tools of ever-increasing sophistication and efficiency [12].

REFERENCES

1 [] Kuntz, I.D. (1992) *Science,* **257,** 1078.
2 [] Hendrickson, J.B. (1990) *Angew. Chem. Int. Ed.,* **29,** 1286.
3 [] Whitesides, G. (1990) *Angew. Chem. Int. Ed.,* **29,** 1209.
4 [] Suckling, C.J. (1991) *Experientia,* **47,** 1093.
5 [] Breckenridge, B.J. (1991) *Experientia,* **47,** 1148.
6 [] Rebek, J. (1991) *Experientia,* **47,** 1096.
7 [] Culver, K.W., Ram, Z., Wallbridge, S., Ishii, H., Oldfield, H. and Blaese, M.R. (1992) *Science,* **256,** 1550.
8 [] Barr, E. and Leiden, J.M. (1991) *Science,* **254,** 1507.
9 [] Van Tol H.H.M., Wu, C.M., Guan, H.-C., Ohara, K., Bunzow, J.R., Civelli, O., Kennedy, J., Seeman, P., Niznik, H.B. and Jovanovic, V. (1992) *Nature,* **358,** 149.
10 [] Pyevitz, R.E. (1992) *FASEB J.* **6,** 2761.
11 [] Kevles, D.J. and Hood, L. (eds) (1992) in *The Code of Codes,* Harvard University Press, Cambridge, MA.
12 [] Testa, B. (1992) *Chimia,* **46,** 297.

INDEX