MEDICINAL CHEMISTRY

A SERIES OF MONOGRAPHS

VOLUME 19

Quantitative Structure–Activity Relationships of Drugs

Edited by

John G. Topliss
Warner-Lambert/Parke-Davis
Pharmaceutical Research Division
Ann Arbor, Michigan

1983

ACADEMIC PRESS

A Subsidiary of Harcourt Brace Jovanovich, Publishers

New York London
Paris San Diego San Francisco São Paulo Sydney Tokyo Toronto

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1 7DX

Library of Congress Cataloging in Publication Data

Main entry under title:

Quantitative structure-activity relationships of drugs.

(Medicinal chemistry series)
Includes index.
Contents: Introduction / C. John Blankley -- Synthetic anti-infective agents / M.S. Tute -- Semisynthetic antibiotics / Yvonne C. Martin and Elizabeth W. Fischer -- Antitumor agents / W.J. Dunn III -- [etc.]
1. Structure-activity relationship (Pharmacology)
I. Topliss, John G. II. Series.
QP906.S75Q36 1983 615'.7 82-25280
ISBN 0-12-695150-0

PRINTED IN THE UNITED STATES OF AMERICA

83 84 85 86 9 8 7 6 5 4 3 2 1

Quantitative
Structure–Activity
Relationships of Drugs

Contributors

Numbers in parentheses indicate the pages on which the authors' contributions begin.

V. Austel (437), Department of Chemistry, Dr. Karl Thomae GmbH, Biberach an der Riss, West Germany

Joel G. Berger (329), Schering-Plough, Research Division, Bloomfield, New Jersey 07003

C. John Blankley (1), Warner-Lambert/Parke-Davis, Pharmaceutical Research Division, Ann Arbor, Michigan 48105

Richard D. Cramer III (253), Research and Development Division, Smith, Kline and French Laboratories, Philadelphia, Pennsylvania 19101

W. J. Dunn III (137), College of Pharmacy, Department of Medicinal Chemistry and Pharmacognosy, University of Illinois, Chicago, Illinois 60612

Elizabeth W. Fischer (77), Department of Chemistry, Barat College, Lake Forest, Illinois 60045

James Y. Fukunaga (329), Schering-Plough, Research Division, Bloomfield, New Jersey 07003

Peter Gund (285), Merck Sharp & Dohme, Research Laboratories, Rahway, New Jersey 07065

Norman P. Jensen (285), Merck Sharp & Dohme, Research Laboratories, Rahway, New Jersey 07065

E. Kutter (437), Department of Chemistry, Dr. Karl Thomae GmbH, Biberach an der Riss, West Germany

Philip S. Magee (393), Chevron Chemical Company, Agricultural Chemicals Division, Richmond, California 94804

Yvonne C. Martin (77), Abbott Laboratories, Abbott Park, North Chicago, Illinois 60064

John G. Topliss (497), Warner-Lambert/Parke-Davis, Pharmaceutical Research Division, Ann Arbor, Michigan 48105

M. S. Tute (23), Central Research, Pfizer Ltd., Sandwich, Kent, England

Stefan H. Unger (177), Syntex Research, Palo Alto, California 94304

*Manfred E. Wolff** (351), Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, California 94143

*Present address: Allergan, Irvine, California 92713.

Preface

Some eighteen years have elapsed since the landmark publications of Hansch and Fujita and of Free and Wilson ushered in the era of quantitative methodology analyzing the structure–activity relationships of drugs, or QSAR as it is commonly known. Over this period there has been an increase in the sophistication, depth, and number of methods available. A range of statistical methodologies has now been tapped. Cluster analysis, discriminant analysis, and principal component and factor analysis have been employed beyond the original use of multiple regression analysis. In addition to the linear free-energy based method, others such as pattern recognition, topological methods, and molecular modeling have come into use, and there has been continued development of quantum mechnical methods.

It might be asked, What has all this accomplished? Previous books in the QSAR field have concentrated on methodology; discussions of the principles and theories underlying the different methods and descriptions of how to apply these methods to actual problems have been provided. The objective of this book is to critically review applications of various QSAR methodologies in different drug therapeutic areas and examine the results in terms of their contribution to medicinal chemistry. There is now a sufficient body of accumulated information on this subject so that an undertaking of this kind appears timely and should shed some light on the question of how useful QSAR is in the broad context of medicinal chemistry.

A broad definition of QSAR has been used here so that applications of all methods that employ some type of quantitative measure have been included. Also, the term "drug" will be interpreted in its broadest sense as meaning a biologically active substance. An attempt has been made to standardize the way equations are presented. Many different formats and levels of statistical information have appeared over the years so that a completely uniform presentation is not possible. The statistical significance of equations and of individual terms in them should be considered adequate unless stated otherwise by the chapter author.

Not every published paper on QSAR applications is mentioned. The intent is

to cover the most significant work and to be analytical and critical as opposed to encyclopedic. Where possible, information and conclusions from different papers have been integrated in an attempt to provide new insight and generalizations.

It is hoped that after reading this book, medicinal chemists will have a better understanding of just what QSAR has contributed to the field of medicinal chemistry and what it might reasonably be expected to contribute in the future.

Finally, I would like to express my gratitude to all of the chapter authors whose expertise and hard work have made this book possible and to other fellow scientists who have provided valuable comments and advice.

Contents

1

Introduction: A Review of QSAR Methodology

C. JOHN BLANKLEY

I. INTRODUCTION

Before embarking upon a critical examination of the applications of quantitative structure–activity relationships (QSAR) in biology, it is nec-

QUANTITATIVE STRUCTURE–ACTIVITY
RELATIONSHIPS OF DRUGS

ISBN 0-12-695150-0

essary to review what methods have been used to obtain QSAR. The purpose of this chapter is to gather together in one place and describe briefly the various procedures, the results of which will be discussed in the subsequent chapters of the book. Some reviews that cover the same ground in other contexts may also be consulted (*119,120,132*).

II. FREE ENERGY MODELS

A. The Hansch Equation

First and foremost among the QSAR methods is the model proposed by Hansch and co-workers (*67,68,72*). It was the seminal contribution of this group to propose that the early observations of the importance of relative lipophilicity to biological potency (*51,125,133*) be incorporated into the useful formalism of linear-free energy relationships (LFER) (*104*) to provide a general model for QSAR in biological contexts. As a suitable measure of lipophilicity, the partition coefficient, log P, between 1-octanol and water was proposed, and it was further demonstrated that this was roughly an additive and constitutive property and hence calculable in principle from molecular structure (*56,84*). Using a probabilistic model for transport across biological membranes, Hansch derived Eq. (1a) (Eq. 1b is an alternate form), which is now known by his name (*68*).

$$\log(1/C) = -k\pi^2 + k'\pi + \rho\sigma + k'' \tag{1a}$$

$$\log(1/C) = -k(\log P)^2 + k'(\log P) + \rho\sigma + k'' \tag{1b}$$

C is the molar concentration (or dose) that elicits a constant biological response (e.g., ED_{50}, MED, IC_{50}), π is the substituent lipophilicity, log P is the partition coefficient, σ is the substituent electronic effect of Hammett (*49*), and k, k', ρ, and k'' are the regression coefficients derived from the statistical curve fitting. The reciprocal of the concentration reflects the fact that higher potency is associated with lower dose, and the negative sign for the π^2 or $(\log P)^2$ term reflects the expectation of an optimum lipophilicity, designated π_0 or $\log P_0$.

The statistical method used to determine the coefficients in Eq. (1) is multiple linear regression (*33,40,176*). A number of statistics are derived in conjunction with such a calculation, which allow the statistical significance of the resulting correlation to be assessed. The most important of these are s, the standard error of the estimate (in many papers called simply the standard deviation), r^2, the coefficient of determination or percentage of data variance accounted for by the model (r, the correlation coefficient is also commonly cited), F, a statistic for assessing the overall

significance of the derived equation (statistics table list critical values for the appropriate number of degrees of freedom and confidence level), and *t* values (also compared with statistics tables) and confidence intervals (usually 95%) for the individual regression coefficients in the equation. Also very important in multiparameter equations are the cross-correlation coefficients between the independent variables in the equation. These must be low to assure true "independence" or orthogonality of the variables, a necessary condition for meaningful results.

The applicability of Eq. (1) to a broad range of biological SAR has been convincingly demonstrated by the Hansch group and many others in the years since 1964 (*15,69,70,73*). The success of this model led early to its generalization to include additional parameters. In attempts to minimize the residual variance in such correlations, a wide variety of physicochemical parameters and properties, structural and topological features, molecular orbital indices, and, for constant but theoretically unaccountable features, indicator or "dummy" (1 or 0) variables have been employed (*119,140,173*). In fact, the widespread use of Eq. (1) has provided an important stimulus for the review and extension of established scales of substituent effects (*18*), and even for the development of new ones (*2,48,76,109,174*). It should be cautioned, however, that the general validity or, indeed, the need for these latter scales has not been established.

Lipophilicity in particular, as reflected in partition coefficients between aqueous and nonaqueous media, most commonly water (or aqueous buffer) and 1-octanol after the initial suggestion of Fujita *et al.* (*56*), has received much attention (*74,105,142–144*). Log *P* for the octanol–water system has been shown to be approximately additive and constitutive and hence schemes for its *a priori* calculation from molecular structure have been devised using either substituent π values or substructural fragment constants *f* (*74,143,144*). A computer adaptation of the method of Leo *et al.* (*74*) has recently been reported (*20*). The approximate nature of any partition coefficient calculation has been frequently emphasized (*17,74,143*), and indeed, some of the structural features that cause unreliability have been identified and accommodated (*142*). Other complications such as steric effects (*36*), conformational effects (*135,136*), and substitution at the active positions of heteroaromatic rings (*162,164,181*) have been observed but cannot as yet be accounted for systematically. Theoretical (*79,145*), statistical (*43,53*), and topological (*128*) methods to approach some of these problems have been reported. The observations, originally by Collander (*23*), of linear relationships among partition coefficients between water and various organic solvents have been extended and qualified (*105,143*). New methodology for the more convenient measurement of log *P* or relative lipophilicity by thin-layer chromatography

(TLC) (*87,163*) or high-pressure liquid chromatography (HPLC) (*87, 120,129,130,171,172*) procedures has been reported. Parameters other than partition coefficients have been proposed as measures of relative lipophilicity, but apart from the chromatographically derived R_m values, these have not as yet been widely used. In several of the cited reviews these are discussed (*69,73,119,120,132,173*).

It is not the present purpose to review all of the other parameters that have been employed in Hansch correlations because these have been adequately discussed elsewhere (*15,74,119,120,132,173*). Several compilations of the most commonly used substituent constants in QSAR work have been published (*18,50,74,113,119,131,171*). The fact that the values listed in these tables do not always agree simply underlines the need for caution in accepting any "critical" collection of constants as definitive.

Another consequence of the empirical and statistical nature of the Hansch model, especially with the proliferation of variables that have been used to seek correlation of biological data, has been the heightened awareness of statistical requirements and constraints. Problems with multicolinearity or cross correlation of independent variables have been noted and discussed (*13,24,114,115*). The potential for chance correlation when too many variables are surveyed to correlate too few data has been pointed out (*166,169*). Misleading results due to "cluster correlation" (*111*) or inappropriate scaling of parameters (*112*) have recently been discussed. The effect of error in the independent variables on the reliability of the regression results has been scarcely mentioned (*52,61,119*) and not at all studied, even though it is an important assumption of regression analysis that the independent variables have minimal error (*33*). "Overfitting" the data, that is obtaining standard deviations lower that the experimental error of the biological measurements, should arouse suspicion (*61*). It is very important, but also difficult to evaluate properly, that any statistically derived equation make good chemical and/or biological sense. This has been urged before (*170,171*), and one study of the steric effects of alkyl groups (*38*) is particularly instructive in this connection.

B. Other Free Energy Models

The success of Hansch in demonstrating that free energy correlations can be successfully applied to biological processes has prompted many workers to reexamine the derivation of the Hansch equation. Using the principles of theoretical pharmacology (*1,60*) or pharmacokinetics (*147,148,153,154*), they have sought to provide improved theoretical models to accommodate more complex relationships between biological activity and chemical structure or properties, or to broaden the scope of

Eq. (1) to include, for example, ionizable compounds. Excellent discussions of these models have been provided by Martin (*119,120*) and Kubinyi (*101*). With one exception, these models have not as yet been tested to a degree that would permit a reasonable evaluation of them to be made. The semiempirical "bilinear" model of Kubinyi (*94,97,99–101*) is a more flexible version of Eq. (1) that still allows for an optimum log P but provides linear ascending and descending portions of the curve with separately determinable slopes. The bilinear model is given by Eq. (2),

$$\log(1/C) = a \log P - b \log(\beta P + 1) + c \tag{2}$$

where C and log P have the same meaning as in Eq. (1), and a, b, β, and c are the coefficients derived by nonlinear regression analysis (*99*).

III. FREE–WILSON MATHEMATICAL MODEL

The idea that substituents ought to contribute constant increments or decrements to biological activity in an analog series has probably been a long-held intuition of medicinal chemists trained in organic chemistry. However, very few solid demonstrations of this can be found in the literature prior to 1964 (*9,11,182*). The existence of linear Hansch correlations is one verification of this idea. However, at about the same time that the Hansch model was proposed, Free and Wilson demonstrated a general mathematical method both for assessing the occurrence of additive substituent effects and for quantitatively estimating their magnitude (*55*). According to their method, the molecules of a drug series are structurally decomposed into a common moiety or core that is variously substituted in multiple positions. A series of linear equations of the form

$$BA_i = \sum_j a_j X_{ij} + \mu \tag{3}$$

are constructed where BA is the biological activity, X_j is the jth substituent with a value of 1 if present and 0 if not, a_j is the contribution of the jth substituent to BA, and μ is the overall average activity. All activity contributions at each position of substitution must sum to zero. The series of linear equations thus generated is solved by the method of least squares for the a_j and μ. There must be several more equations than unknowns and each substituent should appear more than once at a position in different combinations with substituents at other positions. Craig (*25*) and Purcell *et al.* (*82,140*) have discussed in detail the requirements and constraints of the Free–Wilson model as originally formulated. The attractiveness of this model, also referred to as the de novo method, is fourfold:

(1) any set of quantitative biological data may be employed as the dependent variable, (2) no independently measured substituent constants are required, (3) the molecules of a series may be structurally dissected in any convenient manner, and (4) multiple sites of variable substitution are easily accommodated. There are also several limitations: a substantial number of compounds with varying substituent combinations is required for a meaningful analysis; the derived substituent contributions give no reasonable basis for extrapolating predictions outside of the substituent matrix analyzed; and the model will break down if nonlinear dependence on substituent properties is important or if there are interactions between the substituents.

Fujita and Ban (*57*) suggested two modifications of the original formulation. First, the biological activity should be expressed as $\log(1/C)$ or an equivalent measure proportional to a free energy change so that the derived substituent constants might be compared with other free energy related parameters, and second, that μ, the overall average, become analogous to an intercept, that is the calculated activity of the unsubstituted or reference compound of the series. This obviates the need for the cumbersome symmetry or restriction equations of the original method. The Fujita–Ban modification is the form of the Free–Wilson method in common use today. Simplified methods for calculating (*98*) or estimating (*146*) solutions to this model have been reported.

The mathematical implications of the Free–Wilson model have been discussed on several occasions (*14,58,150,151,156*), and the relationship of it to the Hansch model has been noted. Kubinyi (*93,95,96*) has provided the definitive discussion of the interrelationship between the two models. These may be pictured as opposite extremes of the same multiple regression model, the Hansch equation using continuous independent variables and the Free–Wilson model using only discrete (1 or 0) variables. The use of indicator variables in a Hansch equation or of $(\log P)^2$ or π^2 terms to accommodate nonlinearity in a Free–Wilson model, as suggested by Kubinyi, illustrates a mixed model (*95*).

IV. OTHER STATISTICAL MODELS

A. Discriminant Analysis

In many cases of interest the biological measurements available are only semiquantitative or qualitative in nature, and activity assessments such as highly active (+++), moderately active (++), slightly active (+), or inactive (0), or simply active/inactive, must be evaluated. Such data