Table I. Effect of (-)-*N*-(Chloroethyl)norapomorphine (NCA) on Dopamine-Sensitive Adenylate Cyclase in Rat Striatal Homogenates[a]

preincubation concn, µM		% stimulation of cAMP productn, means ± SEM	% inhibn
NCA	DA		
0	0	100.0 ± 4.1	0
10	0	83.9 ± 6.3	16
30	0	43.9 ± 4.5	56
75	0	8.1 ± 1.3	92
30	50	102.8 ± 5.3	−3
0	50	98.0 ± 6.9	2

[a] NCA·HCl (1c·HCl) was preincubated for 10 min at 37 °C with homogenates of rat corpus striatum in a physiologic buffer, alone or with 50 µM DA added, and then washed free of the drugs. Washed tissue was then incubated with 0, 50, or 200 µM dopamine in the presence of excess ATP for 2.5 min at 37 °C, the level of cyclic AMP (cAMP) in the incubation mixture with vs. without DA was assayed by a protein-binding method,[10] and the increase in cAMP levels due to DA was estimated for all conditions ($n \geqslant 5$ replications). The typical basal level of production of cAMP without adding DA (mean ± SEM) was 1.21 ± 0.09 pmol 2.5 min^{-1} (80 µg of tissue)$^{-1}$ or (4 µg of protein)$^{-1}$.

We suggest that the process by which NCA inhibits DA-sensitive adenylate cyclase involves strong and possibly covalent bonding by receptor alkylation, analogous to the action of PBZ at the norepinephrine α receptor and other sites. Further support for the pharmacologic activity of NCA is provided by recent in vivo observations by Costall et al.[11] They found in mouse and rat that (-)-NCA when administered peripherally or intrastriatally can produce selective, potent, and long-lasting (up to 5 days) behavioral and biochemical effects indicative of DA-receptor blockade. These observations and our present results lead us to conclude that the mechanism of action of this agent is uniquely different from the dopamine receptor blockade produced by such reversible, competitive, and relatively short-acting neuroleptic agents as the phenothiazines, butyrophenones, and their congeners.

Acknowledgment. This research was supported, in part, by NIH Grants NS-15439 and NCI-NO-1-CM-53741 (Northeastern University), NIMH Grants MH-30511 and MH-31154 (McLean Hospital), and Research Career Award MH-47370 (R.J.B., Harvard Medical School). We thank Dr. Paul Vouros and Hamdy Maksoud for the mass spectral data and interpretations.

[11] B. Costall, D. H. Fortune, S. J. Law, R. J. Naylor, J. L. Neumeyer, and V. Nohria, *Nature (London)*, in press.

J. L. Neumeyer,* S.-J. Law
Department of Medicinal Chemistry
College of Pharmacy and Allied Health Professions
Northeastern University
Boston, Massachusetts 02115

R. J. Baldessarini, N. S. Kula
Laboratories for Psychiatric Research
Mailman Research Center
McLean Division of Massachusetts General Hospital, and
Department of Psychiatry
Harvard Medical School, Belmont, Massachusetts 02178

Received December 31, 1979

Articles

Structure–Activity Analyzed by Pattern Recognition: The Asymmetric Case

W. J. Dunn III*

Department of Medicinal Chemistry, College of Pharmacy, University of Illinois at the Medical Center, Chicago, Illinois 60612

and Svante Wold

Research Group for Chemometrics, Institute for Chemistry, Umea University, S 901 87 Umea, Sweden.
Received October 31, 1979

In classification studies in which pattern-recognition methods are used to distinguish active compounds from inactive ones, a type of data structure which we call "asymmetric" can be observed. This type of data structure can be quite common and its occurrence can have a profound effect on the classification analysis outcome. The origin of asymmetric data structure and a strategy for obtaining meaningful classification results when it is observed are discussed and illustrated with an example of active and inactive antimalarial quinones.

In recently reported SIMCA pattern-recognition studies of the classification of 4-nitroquinoline 1-oxides,[1] polycyclic aromatic hydrocarbons,[2] and *N*-nitroso compounds[3] as carcinogens or noncarcinogens, we discovered what we term "asymmetric" data structure. This resulted from the carcinogens (active compounds) forming in descriptor space well-defined cluster(s), while the inactive compounds were more or less randomly distributed in the same data space. Such asymmetric data structures can be rather common in the application of classification methodology to the problem of predicting the type of biological response of a new or untested compound. This has an effect on the data analytical strategy used and can ruin the data analysis if not recognized. We discuss here the rationale for asymmetric structure–activity data illustrated by a recently observed example of this type of data structure. We also present a strategy and method for obtaining relevant classification results when asymmetric data structures are observed.

Origin of Asymmetric Data Structure. Asymmetric data structures are primarily encountered in classification problems and will therefore be presented in a context

[1] W. J. Dunn III and S. Wold, *J. Med. Chem.*, **21**, 1001 (1979).
[2] B. Norden, U. Edlund, and S. Wold, *Acta Chem. Scand., Ser. B*, **32**, 1 (1979).
[3] W. J. Dunn III and S. Wold, *Bioorg. Chem.*, accepted for publication.

0022-2623/80/1823-0595$01.00/0 © 1980 American Chemical Society

Table I. Data for Quinones

no.	compd type	substituents 2	3	6	7	act.	log *P*	E_{LUMO}
1	I	OH	*i*-C_5H_{11}	H	H	+	3.87	−0.384
2	I	OH	$(CH_2)_3C_6H_{11}$	H	H	+	5.31	−0.384
3	I	OH	$(CH_2)_8C_6H_{11}$	H	H	+	7.81	−0.384
4	I	Cl	*i*-C_5H_{11}	H	H	−	4.64	−0.339
5	I	CH_3	*i*-C_5H_{11}	H	H	−	4.69	−0.330
6	I	H	*i*-C_5H_{11}	H	H	−	4.20	−0.327
7	I	NH_2	*i*-C_5H_{11}	H	H	−	4.37	−0.420
8	I	OH	$C_{10}H_{21}$	OCH_3	H	−	6.47	−0.399
9	I	OH	$C_{10}H_{21}$	H	OCH_3	−	6.47	−0.397
10	I	OH	$(CH_2)_4C_6H_{11}$	H	Cl	−	6.69	−0.387
11	I	OH	$(CH_2)_3C_6H_{11}$	OH	H	−	4.64	−0.399
12	I	OH	$C_{10}H_{21}$	Br	H	−	7.17	−0.386
13	I	OH	*i*-C_5H_{11}	CH_3	H	+	4.26	−0.385
14	I	OH	*i*-C_5H_{11}	H	CH_3	+	4.26	−0.385
15	I	OH	$(CH_2)_3C_6H_{11}$	H	7,8-C_4H_4	−	6.63	−0.310

no.	compd type	substituents 2	3	5	6	act.	log *P*	E_{LUMO}
16	II	OH	$C_{11}H_{23}$	OH	H	−	5.04	−0.328
17	II	OH	$(CH_2)_3C_6H_{11}$	-$(CH_2)_4$-		−	5.93	−0.277
18	III	$(CH_2)_3C_6H_{13}$	OH			−	5.06	−0.350

no.	compd type	substituents X	Y	Z	2	3	act.	log *P*	E_{LUMO}
19	IV	N	O	O	OH	$C_{15}H_{31}$	+	7.50	−0.378
20	IV	N	O	O	$(CH_2)_8C_6H_{11}$	OH	+	6.50	−0.375
21	IV	N	O	O	OH	$S(CH_2)_{11}CH_3$	+	6.62	−0.353
22	IV	N	O	O	OH	$S(CH_2)_{13}CH_3$	+	8.62	−0.353
23	IV	N	O	O	OH	$S(CH_2)_{17}CH_3$	+	9.62	−0.353
24	IV	N	O	O	$NHC_{16}H_{33}$	H	−	7.80	−0.415
25	IV	N	NC_6H_4-*p*-Cl	O	NHC_6H_4-*p*-Cl	H	−	5.60	−0.249
26	IV	N	NC_6H_3-3,4-$(CH_3)_2$	O	NHC_6H_3-3,4-$(CH_3)_2$	H	−	5.60	−0.258
27	IV	CH	O	O	$CO(CH_2)_2C_6H_{11}$	H	−	4.10	−0.463
28	IV	N	NC_6H_4-*p*-$COOC_2H_5$	O	NC_6H_5-*p*-$COOC_2H_5$	H	−	4.50	−0.114
29	IV	N	NC_6H_4-*p*-OCH_3	O	NH_2	H	−	1.50	−0.278
30	IV	N	NH	O	NH-*n*-C_4H_9	H	−	1.50	−0.295
31	IV	N	NC_6H_4-*p*-Cl	O	NH_2	H	−	2.50	−0.260
32	IV	N	NC_6H_3-3,4-$(CH_3)_2$	O	NH_2	H	−	2.50	−0.258
33	IV	N	NC_6H_4-*p*-OCH_3	O	NHC_6H_4-*p*-OCH_3	H	−	4.50	−0.268
34	IV	N	NC_6H_5	O	NH_2	H	−	1.50	−0.257

no.	compd type	substituents X	3	4	6	act.	log *P*	E_{LUMO}
35	V	N	H	$N(CH_2)_5CH_3$	CH_3	−	1.50	−0.409
36	V	N	H	NH_2	H	−	1.20	−0.407
37	V	N	Cl	NHC_6H_4-*p*-CH_3	H	−	1.50	−0.403
38	V	N	Cl	NHC_6H_5	H	−	1.30	−0.399
39	V	CH	H	NHC_6H_4-*p*-OCH_3	H	+	1.90	−0.412
40	V	CH	H	$N(C_6H_5)(CH_2)_3N(CH_3)_2$	H	+	3.40	−0.409
41	V	CH	CH_3	OH	H	−	0.20	−0.534
42	V	CH	H	OC_2H_5	H	−	1.20	−0.388
43	V	CH	H	$OCH_2CH(CH_3)C_3H_7$	H	−	3.38	−0.388
44	V	CH	H	$O(CH_2)_4CH_3$	H	−	3.38	−0.388
45	V	CH	H	$NH(CH_2)_4$-c-C_6H_{11}	H	+	6.11	−0.422
46	V	CH	H	$NH(CH_2)_4N(C_2H_5)$	H	+	3.79	−0.422
47	V	CH	H	$N(C_2H_5)_2$	H	+	3.28	−0.422
48	V	CH	H	$NH(CH_2)_3N(C_5H_{11})_2$	H	+	5.47	−0.422

applicable to such problems. In the pattern-recognition approach to structure–activity studies, compounds in different classes are described by structural variables (descriptors) which are assumed to determine the type and level of pharmacological activity of the compounds in the classes. Such data (compounds and their descriptors) are represented in matrix form in Figure 1. The asymmetric case is most often encountered when two classes, the active and inactive compounds, are specified.

If the descriptors for the compounds of the members of the active class are represented graphically in the data space as in Figure 2a (here shown in three dimensions), ideally the class will be represented by a well-defined cluster. The inactive compounds are also graphed in the same space. If the number of inactive compounds is small (~2–3), a well-defined cluster will not be obvious. It is also possible, if the number of inactives is large (>6), that they will be randomly scattered around the active class. The latter case is shown in Figure 2b. In both cases the result is the same; i.e., no mathematical description of the inactive class is possible. We call this result asymmetric structure,[4] which can result from a number of factors but is primarily due to the testing strategy. In tests designed

[4] A more descriptive term is "embedded structure", Dr. H. Wold, Upsala University, personal communication.

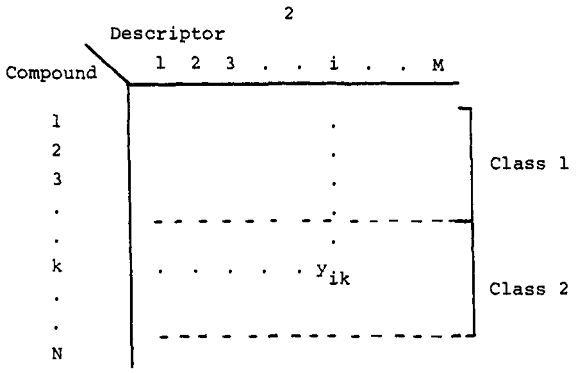


Figure 1. Data matrix for a pattern-recognition classification problem.

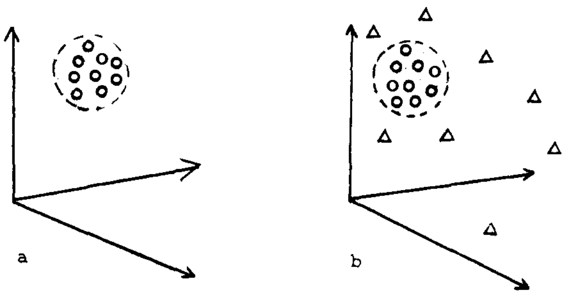


Figure 2. Three-dimensional data spaces: (a) a class of active compounds with well-defined structure; (b) the same class with inactive compounds randomly distributed around the actives. In practice, the number of dimensions is larger than three, but such spaces are difficult to visualize.

to determine if compounds are active at a particular receptor, only the response of that receptor is monitored and only those substances with rather rigid structural requirements will cause an observable response.

From the work of Hansch[5] we know that from a class of structurally similar substances the level of biological activity of the members of the class can be a regular and well-behaved function of structurally related parameters. In this view, the structure–activity system (active principal: biological system) can be considered in terms of control theory.[6] As long as structural variation within the series is not drastic, the system will respond predictably, but any drastic change in structure may result in a discontinuity of the structure–activity relationship and inactivity will result. This corresponds in Figure 2b to moving far away from the active class in any one of the *M* dimensions in descriptor space.

Methods of Classification

To obtain a successful classification result in cases with asymmetric data structures, the method of analysis must be able to distinguish the well-defined class from the one with no structure. A number of methods are commonly used in chemical classification studies: the linear learning machine (LLM),[7] linear discriminant analysis (LDA),[8] the *k* nearest neighbor (KNN),[9] and the recently developed SIMCA[10] method. The two former methods are so-called hyperplane methods, which obtain the equation for a plane or hyperplane in data space which is inserted between two classes in a separation problem. This discriminant function can be used to classify an unknown or untested object by determining on which side of the hyperplane this object lies. From Figure 2b it is obvious that such methods cannot be used in classification problems in which embedded structures are found. No meaningful hyperplane can be inverted between the classes. Also, since they are regression methods, they require that in the *initial stages* of the analysis,[11] the number of compounds be four to five times the number of descriptors. This is not possible in cases where the number of compounds in a class is only four or five. The KNN method classifies an object on the basis of the identity of its two nearest neighbors. Therefore, this method can be used to obtain classification results in cases with asymmetric data structures, but it suffers from the inability to obtain results beyond classification.[12]

SIMCA encloses the well-structured class in a region of space, and is therefore the most obvious method to use in problems with

Table II. Statistical Data for the Quinones

no.	RSD[a]	u_1
1	0.71	-0.37
2	0.24	0.14
3	0.58	0.92
4	1.70	1.14
5	2.00	1.41
6	2.20	1.34
7	0.46	-1.21
8	0.56	0.06
9	0.50	0.12
10	0.29	0.47
11	0.04	-0.54
12	0.42	0.66
13	0.56	-0.27
14	0.56	-0.27
15	1.90	2.61
16	1.90	1.58
17	3.00	3.30
18	1.30	0.97
19	0.31	0.99
20	0.10	0.75
21	0.68	1.40
22	0.02	2.06
23	0.30	2.39
24	1.40	0.05
25	3.90	3.98
26	3.70	3.73
27	1.60	-2.50
28	8.10	7.40
29	4.50	1.82
30	4.00	1.35
31	4.60	2.65
32	4.70	2.71
33	3.70	3.06
34	5.00	2.41
35	0.79	-1.85
36	0.95	-1.89
37	0.96	-1.68
38	1.10	-1.63
39	0.58	-1.80
40	0.17	-1.22
41	2.30	-5.77
42	1.50	-1.36
43	0.76	-0.64
44	0.76	-0.64
45	1.10	-0.70
46	0.32	--1.46
47	0.16	-1.63
48	0.87	-0.91

[a] RSD (residual standard deviation) for class = 0.57.

(5) C. Hansch, *Acc. Chem. Res.*, **2**, 232 (1967).

(6) O. I. Elgerd, "Control Systems Theory", McGraw-Hill, New York, 1967.

(7) P. C. Jurs and T. L. Isenhour, "Chemical Applications of Pattern Recognition", Wiley-Interscience, New York, 1967.

(8) Y. C. Martin, "Quantitative Drug Design", Marcel Dekker, New York, 1978.

(9) B. R. Kowalski, in "Computers in Chemical and Biochemical Research", Vol. 2, C. E. Klopfenstein and C. L. Wilkens, Eds., Academic Press, New York, 1974.

(10) S. Wold, *Pattern Recognition*, **8**, 127 (1976).

(11) J. G. Topliss and R. P. Edwards, *J. Med. Chem.*, **22**, 1238 (1979).

(12) C. Albano, W. J. Dunn III, U. Edlund, E. Johansson, B. Norden, M. Sjostrom, and S. Wold, *Anal. Chim. Acta*, **103**, 429 (1978).

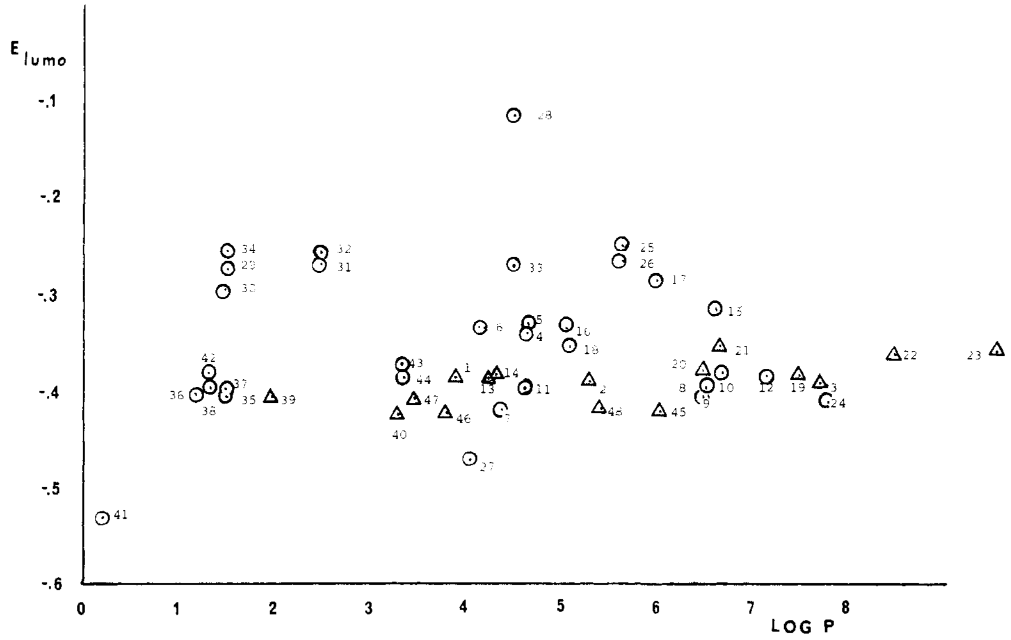
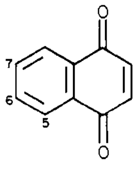


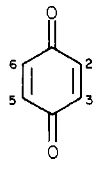
Figure 3. Graphic display of quinone data before regularization: (△) active compounds; (○) inactive compounds.

asymmetric data structures. With SIMCA, an area or volume element in descriptor space is defined in which the probability of finding the active compounds is the highest. Compounds outside of this region are expected to be nonmembers of this class. An advantage of this approach is that the size of the inactive class can be very small, e.g., two or three compounds, and meaningful classification results can still be obtained.

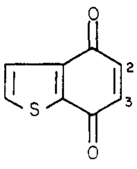
Example. Recently, a study was reported in which an obvious relationship between physical properties and drug class was recognized for naphthoquinones as active or inactive against *Plasmodium berghei*, one of the parasites responsible for malaria.[13] However, LDA or quadratic discriminant analysis were not able to separate the classes.[14] The structures of the compounds are given below and a total of 48 compounds are included.



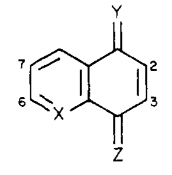
I



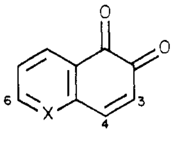
II



III



IV



V

The antimalarial response is thought to be due to their ability to compete with coenzyme Q and disrupt mitochondrial electron transport.[15] It has been proposed that in order to be active, quinones must be passively transported to the site of action where reduction occurs.[15] Hence, proper transport properties and reduction potential are required for activity. These properties were modeled by the 1-octanol/water partition coefficient and the energy of the lowest unoccupied molecular orbital, E_{LUMO}, respectively.[13] The compounds and their activities and data (log P and E_{LUMO}) are given in Table I.

This is a two variable problem and, therefore, can be presented in graphic form (Figure 3). In this figure it can be seen that the active compounds fall on or near a straight line, while the inactives are scattered randomly in the space around this line. This is a very clear exhibition of embedded structure, and we use it to illustrate problems inherent in classification studies where such data are observed.

The Basis of Classification with the SIMCA Method. The methodology on which SIMCA is based assumes that the regularity within the data for a class, such as that shown in Figure 3, can be approximated by a principal components model, eq 1. Here

$$Y_{ik} = m_i + \sum_{a=1}^{A} b_{ia}u_{ak} + e_{ik} \qquad (1)$$

Y_{ik} is the observed value of variable i for the kth object, m_i is the mean value of variable i in the class, and A is the number of component or product terms in the model. Associated with each product term is a variable specific vector, b_{ia}, and a compound specific vector, u_{ak}. The difference between Y_{ik} and the calculated value is the residual, e_{ik}.

Since the problem is a two variable one, it can be used to visualize how SIMCA actually classifies. These steps are summarized as follows: (1) First the data are adjusted to zero mean and unit variance. This gives each variable equal weight in the analysis. (2) The data for the active class are fitted to a similarity model. In this case, the model has one component and is graphically represented by a straight line. (3) The residuals for the objects in the training set are then calculated. From these a standard deviation for the class can be obtained. Classification of the training-set compounds is done by comparing their standard deviations with that of the class. If inside the standard deviation for the class, an object is classified as being a member of that class; if outside, the compound is classified as being a nonmember of the class. (4) The unknown or test set compounds are classified. Since a standard deviation above and below the model defines an area of descriptor space, an unknown compound can then be classified as being a member of the class if it falls in this area. Those compounds outside the area are nonmembers of the class. These operations were carried out on the data in Table I, and the statistical data for the analysis are given in Table II.

Results

From the standard deviations for the active compounds it can be seen that 9/16 fall within 1 standard deviation for the class (0.57), and all lie within 2 standard deviations. Compound **45** is the one with poorest fit to the similarity model with a standard deviation of 1.10. This puts it on

[13] Y. C. Martin, T. M. Bustard, and K. R. Lynn, *J. Med. Chem.*, **16**, 1089 (1973); Y. C. Martin, "Quantitative Drug Design", Marcel Dekker, New York, 1978, p 106.

[14] Y. C. Martin and G. Chao, private communication.

[15] J. D. Turnbull, G. L. Biagi, A. J. Merala, and K. D. Cornwell, *Biochem. Pharmacol.*, **20**, 1383 (1971).

the "edge" of the region where those compounds expected to have antimalarial activity are to be found.

Of the inactive compounds, some are classified to be active. These are **7-12**, which are false positives.

Discussion

Asymmetric data structures discovered in recent structure-activity studies[1-3] seem to be rather common. They can be expected in classification studies in which a class of active compounds is analyzed together with a nonactive class. Such structures can result mainly from the existence in the analysis of one of two factors: (1) an inactive class which contains too few members to obtain a statistically significant mathematical description of the class or (2) an inactive class which contains no systematic structure. In the former case, the number of compounds in a class must be at least five to justify a similarity model with one component ($A = 1$). The second case is illustrated well by the quinone data in this report. The number of inactive compounds is 32, and there is no apparent structure in their data while there is obvious structure in the data for the active class.

It is also obvious from Figure 3 that classification methods which rely on the insertion of a plane or hyperplane between the classes in order to separate the classes will fail if asymmetric or embedded structures are in the data. For this reason, LDA could not separate the active antimalarials from the inactive compounds; likewise, the LLM would also fail to separate the two classes. The KNN method, which classified an object on the basis of its nearest neighbors, can be expected to give good classification results.

While SIMCA can be expected to give good results in such classifications, the interpretation of the results must be carefully made. The result of classifying a new compound in the active class means that the compound is a member of that class and its residual standard deviation is a measure of the probability of this assignment. Compounds within 1 standard deviation have the highest probability of being a class member. Compounds with larger standard deviations, but within 2 such deviations, will have lower probabilities of being class members. Larger standard deviations than this suggest that the new compound is a "nonmember" of the class.

Quantitative Structure-Activity Relationships by Distance Geometry: Systematic Analysis of Dihydrofolate Reductase Inhibitors[1]

Gordon M. Crippen

Department of Chemistry, Texas A&M University, College Station, Texas 77843. *Received November 5, 1979*

Extensions are presented for the distance geometry approach to rationalizing ligand binding data. These are algorithms to (i) detect when homologues are not binding with the same orientation in the binding site although they are chemically similar; (ii) deduce what the binding site's size and shape must be; and (iii) calculate the optimal set of interaction energies between parts of the site and parts of the ligand molecules. This improved methodology is tested on a set of 68 quinazoline inhibitors of *S. faecium* dihydrofolate reductase. Results are discussed and compared with the Hansch method of QSAR, and an improved inhibitor is predicted.

This is the second paper in a series on a novel method for deducing quantitative structure-activity relationships (QSAR) for drugs. We treat the following idealized problem: (i) binding is observed to occur on a single site of a pure receptor protein (or other macromolecule); (ii) each ligand has a well-determined chemical structure and stereochemistry but may be flexible due to rotation about single bonds; (iii) no chemical modification of the ligands occurs during the binding experiment, although the conformation of the ligand may change upon binding to accomodate the binding site; (iv) the free energy of such a conformational change is small compared to the free energy of binding; (v) the experimentally determined free energy of binding is given and is approximately the sum of the "interaction energies" for all "contacts" between parts of the ligand molecule and parts of the receptor site; (vi) the site itself may be slightly flexible, although no major conformational changes are permitted, and the energetic cost of any deformation is negligible. The previous paper in this series[2] explained how the series of ligands may each be represented as a collection of points corresponding to atoms or small groups of atoms, and the conformational flexibility can be treated as upper and lower bounds on the distances between all pairs of points making up the ligand. Similarly, the binding site was represented as points positioned rigidly in space with respect to each other. The site points are best thought of as corresponding in the real site to the locations of pockets of various types or in some cases as the positions of steric blocking groups. The interaction energies between ligand points of the various types and the site points of their types are given as entries in an energy table. It is through this table that a certain type of site point may be characterized as being a hydrogen bond donor, or a small pocket accomodating ethyl groups or less, etc. The first paper went on to outline computer algorithms for finding the energetically most favorable but still geometrically allowed binding *mode* for each ligand in the data set. The binding mode consists of specifying which ligand points are to coincide with which site points. The calculated binding free energy for a given mode is taken to be the sum of the interaction energies for each coincidence (or "contact").

In the present work, all of the above has remained the same, and the interested reader is urged to read ref 2 for more detail. The major shortcoming of the method so far

[1] This work was supported by grants from the National Resource for Computation in Chemistry under a grant from the National Science Foundation and the U.S. Department of Energy (Contract W-7405-ENG-48) and by the National Science Foundation directly under Grant PCM78-05468. We are also grateful for the use of the UCSF Computer Graphics Laboratory (NIH RR 1081).

[2] G. M. Crippen, *J. Med. Chem.*, **22**, 988 (1979).