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).

0022-2623/80/1823-0599$01.00/0

has been that the user must guess the number of site points to be employed, their types, their relative locations, and the interaction energies. We now present algorithms to aid in these decisions and illustrate their use on a large data set, namely, 68 quinazoline inhibitors of *S. faecium* dihydrofolate reductase. The objectives of this paper are to answer three questions in turn: how many site points are required to explain the observed binding data; what must be their relative positions; and what set of interaction energies optimally reproduce the observed free energies of binding?

Methods

Determination of Number of Site Points. The number of site points required depends to a first approximation on the resolution of the representation of the ligand molecules. If whole benzene rings correspond only to a single ligand point, then many fewer site points will be required to form an attachment site than when each carbon atom and hydrogen atom in the ring is given a separate ligand point. In this work, we take the representation of the ligands as given, although we intend to explore the effects of varying the resolution in future studies. The example at hand is given in Table I, where the chemical structure of all 68 ligands is summarized. The representation of the ligands is the same as was arbitrarily chosen in ref 2: all six-membered conjugated carbon rings are given by a single point in the center of the ring, even in fused rings; NH, NH_2, CH, CH_2, CH_3, S, SO, SO_2, OH, and SH groups are all shown as a single point each with a unique type unless they differ only by a hydrogen atom; carboxyl and ester groups consist of one point each but are distinct types; and the heteroring of the quinazoline system is given two points, located at N(1) and N(3). See the left side of Table IV for a complete listing of types. In particular, the N(1) of the quinazoline is given two different types, depending on whether the ring is 2,4-diamino substituted. The intent is to simulate the shift in pK_a by assigning a different protonation state to N(1), and thus help explain the unusually favorable binding of the 2,4-diamino-substituted derivatives. The rationale behind this is discussed in ref 2. Now given this representation, we can systematically analyze the observed free energies of binding as a function of the chemical structure of the ligands. This is equivalent to assuming that each ligand binds to the site in the same orientation as all the others, with common features in contact with always the same site points. The picture is that of a site having some number of points dedicated to the common structural feature of the set of ligands and other surrounding site points to accomodate any occurence of a substituent at the various points of attachment. Of course this approach is not new, being essentially equivalent to the linear-free-energy or Free–Wilson method,[4] but we will develop a novel explicit geometric interpretation of the results of the analysis.

Making no assumptions whatever about what are customarily recognized chemical functional groups and ring systems, the decomposition algorithm locates the largest number of points common to all ligand molecules which have the same corresponding point types and have the same interpoint distances. This is referred to as the "base" group. The base group should have at least three points, so that substituents attached to it will have a well-determined position by triangulation with the distances from the substituent to the points of the base. Now the smallest set of substituent groups needed to account for all remaining points of all ligands is determined. A substituent group is characterized by the number of points involved (often in this case only one), their types, their mutual distances, and their distances to the base points. Sometimes two distinct substituent groups differ only by ligand point types and not by location relative to the base group. See the Appendix for a more complete description of the decomposition algorithm. If the observed binding free energy is assumed to be the sum of the contributions of the base group and each substituent group in the ligand, for each ligand in the data set, then a simple linear least-squares calculation gives the optimal energy contribution of each substituent.

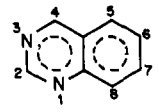
To illustrate the above procedure, consider the 68 ligands in Table I. There are two points in common to all: the N(3) of the quinazoline and the nonheteroring adjacent to it. N(1) is not common to all because for 2,4-diamino derivatives it is a different type. A third point common to all but ligands **36** and **37** is a phenyl ring (or half a naphthyl group) attached either to the 5 or 6 position. Due to flexibility of the linkages between the quinazoline and the phenyl, there is overlap in the location of the phenyl between the two attachment positions. This overlap region is excluded by the peculiar linkages in ligands **36** and **37**. Proceeding with the 66 remaining ligands, the algorithms finds 33 different substituent groups, of which only about 14 are geometrically distinct. Most substituents are only a single point and represent functional groups that a chemist would tend to pick out, such as the 2-NH_2, 5-S–, etc. However, some are rather complex, as in the case of the CONHCH group when attached in the para position of the 6-$CH_2NHC_6H_4$– group. A least-squares fit of the ΔG_{obsd} values of Table I to the sum of the contributions of each component group for each of the 66 ligands was performed with the side constraint that each contribution had to be less than or equal to zero. This is equivalent to assuming that there are no unfavorable interactions for a ligand when it fits into the site, for otherwise the unconstrained least-squares fitting might produce the physically unrealistic result of some groups contributing +100 kcal while others contribute –101 kcal. Even the weakest inhibitor of the series binds with $\Delta G = -5.8$ kcal, so it is reasonable to assume that all interaction energies are favorable. With very poor ligands in the data set, this assumption would have to be relaxed. The outcome of the fit, however, was that 66 observations could not be matched within ±1 kcal even using 34 parameters! Given the model used, this is an objective determination that the constant binding mode assumption is in error at the 1 kcal level, since any other choice of interaction energies would give an even worse fit into the data. However, it is still not clear *which* ligands violate the constant mode assumption. The nonconforming ligands were identified by removing the worst offender from the data set, repeating the fit, once again removing the worst ligand, and so on until the ΔG_{obsd} values for all the remaining inhibitors could be fit to ±1 kcal. This was achieved when ligands **25**, **48**, **26**, **35**, **64**, **63**, **68**, and **44** (in order of decreasing error with the final contribution parameters) were removed from the 66 ligands. The rms error of the fit over the 66 inhibitors was 1.17 kcal, and that was due largely to ligands **25** and **48**, for which the error was over 5 kcal. Actually, 12 of the parameters were set to zero by the least-squares algorithm, and, since they were effectively "not used", the number of degrees of freedom in the fit was 34 – 12 = 22. Nonetheless, such a failure to account for the experimental observations in some of the ligands can be taken as a clear indication that the assumption of a constant binding mode must be in error. Furthermore, ligands **25** and **48** especially are identified as probably not fitting the general pattern, and comparison of these with closely related compounds in the data set will probably suggest what the important differences might be.

As an aside for completeness, if the constant binding mode analysis *had* worked satisfactorily, proposing a distance geometry site and interaction energies is now quite easy. One site point is chosen for each point in the base group and for each point in each geometrically distinct substituent group. The relative positions of the base group points are known as are the positions of each substituent relative to the base, so calculating all the coordinates is a straightforward procedure according to the standard distance geometry algorithm.[5] The foregoing least-squares calculation has already produced the interaction energies required between site points and the appropriately fitting ligand points. Let all other interaction energies be positive to discourage alternate binding modes.

Variable Binding Mode Analysis. Having now determined that all 68 ligands apparently do not bind in the same way, we are required to suggest which ligands bind in which modes.

(3) J. B. Hynes, W. T. Ashton, D. Bryansmith, and J. H. Freisheim, *J. Med. Chem.*, 17, 1023 (1974).
(4) Y. C. Martin, "Quantitative Drug Design", G. L. Grunewald, Ed., Marcel Dekker, New York, 1978.
(5) G. M. Crippen and T. F. Havel, *Acta Crystallogr., Sect. A*, **34**, 282 (1978).

Table I. Binding of Quinazoline Derivatives to *S. faecium* Dihydrofolate Reductase



no.	group	ΔG_{obsd},[a] kcal	ΔG_{calcd},[b] kcal
1	2-H, 4-NH_2, 6-SO_2-2-$C_{10}H_7$	−5.8	−8.5
2	2,4-$(SH)_2$, 6-S-2-$C_{10}H_7$	−6.0	−6.9
3	2-SH, 4-OH, 6-S-2-$C_{10}H_7$	−6.2	−6.9
4	2,4-$(NH_2)_2$, 5-SO_2-2-$C_{10}H_7$	−6.5	−10.5
5	2-H, 4-NH_2, 6-S-2-$C_{10}H_7$	−6.5	−8.5
6	2-NH_2, 4-OH, 5-CH_3, 6-$NHCH_2C_6H_4$-4-CO_2H	−6.5	−8.0
7	2-OH, 4-SH, 6-S-2-$C_{10}H_7$	−6.8	−6.9
8	2,4-$(OH)_2$, 6-S-2-$C_{10}H_7$	−6.9	−6.9
9	2-OH, 4-NH_2, 6-S-2-$C_{10}H_7$	−6.9	−9.4
10	2,4-$(NH_2)_2$, 5-SO-2-$C_{10}H_7$	−6.9	−10.5
11	2-NH_2, 4-OH, 5-CH_3, 6-$NHCH_2C_6H_4$-4-CO_2Et	−7.1	−8.0
12	2-NH_2, 4-OH, 6-$NHCH_2C_6H_4$-4-COOH	−7.2	−8.0
13	2-H, 4-NH_2, 6-SO-2-$C_{10}H_7$	−7.2	−8.5
14	2,4-$(NH_2)_2$, 5-SOC_6H_3-3,4-Cl_2	−7.3	−10.5
15	2-NH_2, 4-OH, 5-S-2-$C_{10}H_7$	−7.4	−9.4
16	2-SH, 4-NH_2, 6-S-2-$C_{10}H_7$	−7.4	−9.4
17	2-NH_2, 4-OH, 6-$CH_2NHC_6H_4$-4-CONHCH(CO_2Et)$CH_2CH_2CO_2Et$	−7.9	−8.0
18	2-NH_2, 4-SH, 6-SO_2-2-$C_{10}H_7$	−8.0	−9.4
19	2-NH_2, 4-OH, 6-SO-2-$C_{10}H_7$	−8.2	−9.4
20	2-NH_2, 4-OH, 6-$SO_2C_6H_3$-3,4-Cl_2	−8.2	−9.4
21	2-NH_2, 4-OH, 6-$CH_2N(CH_3)C_6H_4$-4-CONHCH(CO_2Et)$CH_2CH_2CO_2Et$	−8.3	−8.0
22	2-NH_2, 4-OH, 6-$NHCH_2C_6H_4$-4-CO_2Et	−8.6	−8.0
23	2-NH_2, 4-OH, 6-$CH_2NHC_6H_4$-4-CO_2H	−8.8	−8.0
24	2-NH_2, 4-OH, 6-$CH_2NHC_6H_4$-4-CONHCH(CO_2H)$CH_2CH_2CO_2H$	−8.9	−8.0
25	2-NH_2, 4-OH, 5-SO_2-2-$C_{10}H_7$	−9.0	−9.4
26	2,4-$(NH_2)_2$, 5-$SO_2C_6H_3$-3,4-Cl_2	−9.0	−10.5
27	2-NH_2, 4-OH, 6-S-C_6H_3-3,4-Cl_2	−9.1	−9.4
28	2-NH_2, 4-OH, 5-Cl, 6-$NHCH_2C_6H_4$-4-CO_2Et	−9.3	−8.1
29	2-NH_2, 4-SH, 6-S-2-$C_{10}H_7$	−9.3	−9.4
30	2-NH_2, 4-OH, 6-SO_2-2-$C_{10}H_7$	−9.6	−9.4
31	2-NH_2, 4-OH, 6-$CH_2N(CH_3)C_6H_4$-4-CONHCH(CO_2H)$CH_2CH_2CO_2H$	−9.7	−8.0
32	2-NH_2, 4-OH, 6-$CH_2N(CHO)C_6H_4$-4-CONHCH(CO_2H)$CH_2CH_2CO_2H$	−9.9	−8.0
33	2,4-$(NH_2)_2$, 5-S-C_6H_3-3,4-Cl_2	−9.9	−10.6
34	2-NH_2, 4-OH, 6-S-2-$C_{10}H_7$	−10.2	−9.4
35	2-NH_2, 4-OH, 6-$CH_2NHC_6H_4$-4-CO_2Et	−10.6	−8.0
36	2,4-$(NH_2)_2$, 5-*trans*-CH=CH-2-$C_{10}H_7$	−10.7	−10.4
37	2,4-$(NH_2)_2$, 5-$CH_2SC_6H_4$-4-Cl	−10.9	−10.4
38	2,4-$(NH_2)_2$, 5-S-2-$C_{10}H_7$	−10.9	−10.6
39	2,4-$(NH_2)_2$, 6-$NHCH_2C_6H_4$-4-CO_2Et	−11.0	−12.5
40	2,4-$(NH_2)_2$, 5-*cis*-CH=CH-2-$C_{10}H_7$	−11.1	−10.4
41	2,4-$(NH_2)_2$, 6-$CH_2NHC_6H_4$-4-CO_2-*n*-Bu	−11.2	−12.5
42	2,4-$(NH_2)_2$, 5-CH_2S-2-$C_{10}H_7$	−11.3	−10.5
43	2,4-$(NH_2)_2$, 6-$CH_2NHC_6H_4$-4-CO_2Et	−11.4	−12.5
44	2,4-$(NH_2)_2$, 5-Cl, 6-$CH_2NHC_6H_4$-4-CO_2-*n*-Bu	−11.4	−12.5
45	2,4-$(NH_2)_2$, 6-$CH_2NHC_6H_4$-4-CONHCH(CO_2Et)$CH_2CH_2CO_2Et$	−11.5	−12.5
46	2,4-$(NH_2)_2$, 6-$CH_2NHC_6H_4$-4-CONHCH(CO_2Et)CH_2CO_2Et	−11.6	−12.5
47	2,4-$(NH_2)_2$, 5-CH_3, 6-$CH_2NHC_6H_4$-4-CO_2H	−11.8	−12.5
48	2,4-$(NH_2)_2$, 5-CH_2CH_2-2-$C_{10}H_7$	−11.9	−10.5
49	2,4-$(NH_2)_2$, 6-S-2-$C_{10}H_7$	−12.1	−13.9
50	2,4-$(NH_2)_2$, 5-CH_3, 6-$CH_2NHC_6H_4$-4-CO_2-*n*-Bu	−12.1	−12.5
51	2,4-$(NH_2)_2$, 5-Cl, 6-$NHCH_2C_6H_4$-4-CO_2Et	−12.2	−12.5
52	2,4-$(NH_2)_2$, 6-$CH_2NHC_6H_4$-4-CONHCH(CO_2H)CH_2CO_2H	−12.2	−12.5
53	2,4-$(NH_2)_2$, 6-S-C_6H_3-3,4-Cl_2	−12.2	−13.9
54	2,4-$(NH_2)_2$, 5-CH_3, 6-$CH_2NHC_6H_4$-4-CONHCH(CO_2Et)CH_2CO_2Et	−12.2	−12.6
55	2,4-$(NH_2)_2$, 5-Cl, 6-$CH_2NHC_6H_4$-4-CO_2H	−12.3	−12.5
56	2,4-$(NH_2)_2$, 5-CH_3, 6-$CH_2NHC_6H_4$-4-CO_2Et	−12.3	−12.5
57	2,4-$(NH_2)_2$, 6-SO_2-2-$C_{10}H_7$	−12.4	−13.9
58	2,4-$(NH_2)_2$, 5-Cl, 6-$CH_2NHC_6H_4$-4-CONHCH(CO_2Et)CH_2CO_2Et	−12.5	−12.6
59	2,4-$(NH_2)_2$, 5-Cl, 6-$CH_2NHC_6H_4$-4-CONHCH(CO_2H)CH_2CO_2H	−12.6	−12.6
60	2,4-$(NH_2)_2$, 5-CH_3, 6-$CH_2NHC_6H_4$-4-CONHCH(CO_2H)CH_2CO_2H	−12.7	−12.6
61	2,4-$(NH_2)_2$, 5-Cl, 6-$CH_2NHC_6H_4$-4-CO_2Et	−12.8	−12.5
62	2,4-$(NH_2)_2$, 6-SO-2-$C_{10}H_7$	−12.8	−13.9
63	2,4-$(NH_2)_2$, 6-$CH_2NHC_6H_4$-4-CO_2H	−13.0	−12.5
64	2,4-$(NH_2)_2$, 5-CH_3, 6-$NHCH_2C_6H_4$-4-CO_2Et	−13.1	−12.5
65	2,4-$(NH_2)_2$, 6-$CH_2NHC_6H_4$-4-CONHCH(CO_2H)$CH_2CH_2CO_2H$	−13.1	−12.5
66	2,4-$(NH_2)_2$, 6-$CH_2N(CHO)C_6H_4$-4-CONHCH(CO_2H)$CH_2CH_2CO_2H$	−13.3	−12.5
67	2,4-$(NH_2)_2$, 6-S-C_6H_4-3-CF_3	−13.4	−13.8
68	2,4-$(NH_2)_2$, 6-SO_2-C_6H_3-3,4-Cl_2	−13.4	−13.9

[a] See ref 3. [b] According to variable binding mode analysis.

Table II. Role of Proposed Site Points

point	binds
1	N(1) unprotonated
2	N(1) protonated (when 2,4-diamino substituted)
3	second quinazoline ring
4	2-position substituent
5	4-position substituent ordinarily
6	4-position substituent when 2,4-diamino substituted
7	5-position linkage ordinarily
8	5-position linkage when 2,4-diamino substituted
9	proximal half of naphthalene (carbon-1 to -4, next to linkage)
10	distal half of naphthalene (carbon-5 to -8)
11	6-position linkage

Table III. Final Proposed Dihydrofolate Reductase Binding Site Showing Site Point Numbering, Types, and Coordinates

point no.	point type	coordinates		
		x	y	z
1	t_{s1}	−3.25	0.24	0.28
2	t_{s2}	−3.82	−0.27	−0.99
3	t_{s3}	−1.16	−1.53	−0.69
4	t_{s4}	−5.54	0.16	0.83
5	t_{s4}	−0.57	−1.56	2.72
6	t_{s4}	−0.73	1.92	−1.41
7	t_{s5}	1.12	−2.53	0.60
8	t_{s6}	1.01	0.96	−1.41
9	t_{s3}	3.83	−0.42	0.22
10	t_{s3}	6.57	0.19	0.98
11	t_{s5}	1.84	−1.96	−2.23

Following the subjective reasoning suggested in ref 2, we continue to propose that the quinazoline ring is free to rock in the binding site about an axis running from the 2 to the 6 position. Then there would be two different sites for the N(1), depending on whether it was protonated or not (i.e., whether or not the ring is 2,4-diamino substituted), and corresponding to these would be two different site points each for 4- and 5-position substituents. The proposed binding scheme can be summarized in Table II. This scheme can be thought of as a preliminary sketch which must now be filled in as a detailed list of the *desired* binding modes for each of the 68 inhibitors. Once that is done, a great deal has implicitly been said about the geometry of the site, and it is a straightforward matter to devise an algorithm to deduce the upper and lower bounds on the distance among the site points. The basic logic is that the upper and lower bound distances between any two ligand molecule points are known for any given ligand, and if the desired binding mode has two of these in contact with some two site points then the lower bound distance between the two site points must be at least as great as the lower bound for the corresponding ligand points, etc. For more detail, see the Appendix. It is possible that due to errors in proposing the binding modes, there is no arrangement of the site points in space which will meet the deduced geometric constraints. Fortunately, one can usually trace back to the source of such inconsistencies with our computer programs. In the case of the DHFR inhibitors, more than one configuration of site points will satisfy the constraints, so the standard distance geometry algorithm[5] produces a variety of possibilities from which to choose. In particular, the angle swept out in the rocking motion is not determined, so that the distance between site points 7 and 8 can lie in the range 1.0 to 6.8 Å, corresponding to a slight tilt or a full 180° flip. In an effort to keep the 11 site points well separated in space, we opted for the low end of the tilt range. It is difficult to assess the number of degrees of geometric freedom in our proposed binding site, but an upper bound can be derived as follows: There are 11 points with 3 coordinates each, but only their relative positions are of interest, so we must subtract 6 degrees of freedom for rigid translation and rotation, leaving 27. Now 45 of the interpoint distances were constrained to some degree, including 9 cases where the upper and lower bounds were nearly equal. Further, at the solution configuration we chose, 3 more inequalities were nearly being violated, so subtracting these 12 strong constraints leaves 15 degrees of freedom at most. Table III shows the coordinates of the proposed site points.

Having now established the geometric part of the proposed site, we next turn to the energetic parameters. In the original representation of the ligands step, 19 different ligand point types were arbitrarily chosen, corresponding to the groups shown along the left side of Table IV. Given the specifications for site points' roles in Table II, we chose only six different site point types. For example, site points 3, 9, and 10 are all intended to bind aromatic, six-membered carbon rings, so it is reasonable to attribute the same sorts of interaction energies to each, and hence all three are given type t_{s3}. Some of the entries in Table IV can be specified immediately. For example, site point 8, having type t_{s6}, is supposed to be repulsive to >SO and $>SO_2$ groups in order to account for the anomolously low binding of ligands **4** and **10**. Hence, entries 14,6 and 13,6 were set to +10.0 kcal (clearly the exact value is unimportant). Many other entries correspond to interactions not employed by the desired binding modes, so they are given the faintly repulsive standard value of +0.1 kcal. That way they can be used in an unanticipated binding mode without over-

Table IV. Final Proposed Dihydrofolate Reductase Interaction Energy Table[a] for the 19 Ligand Point Types and the 6 Site Point Types

		site point types					
	ligand point types	t_{s1}	t_{s2}	t_{s3}	t_{s4}	t_{s5}	t_{s6}
t_{l1}	N(3) quinazoline	0.1	0.1	0.1	0.1	0.1	0.1
t_{l2}	N(1) quinazoline	−1.409	10.0	0.1	0.1	0.1	0.1
t_{l3}	phenyl, etc.	0.1	0.1	−0.067	0.1	10.1	0.1
t_{l4}	$-CF_3$	0.1	0.1	−0.001	0.1	0.1	0.1
t_{l5}	$-CH_2CH_2$	0.1	0.1	0.1	0.1	0.1	−0.017
t_{l6}	$-CH_3$	0.1	0.1	0.1	0.1	−0.067	−0.083
t_{l7}	-COOH	0.1	0.1	0.1	0.1	0.1	0.1
t_{l8}	-COO-	0.1	0.1	0.1	0.1	0.1	0.1
t_{l9}	$-NH_2$	0.1	0.1	0.1	−3.443	−2.036	0.1
t_{l10}	-CONH-	0.1	0.1	0.1	0.1	0.1	0.1
t_{l11}	-Cl	0.1	0.1	−0.075	0.1	−0.083	−0.083
t_{l12}	-S- or -SH	0.1	0.1	0.1	−0.950	−3.409	−0.084
t_{l13}	>SO	0.1	0.1	0.1	0.1	−3.405	10.0
t_{l14}	$>SO_2$	0.1	0.1	0.1	0.1	−3.405	10.0
t_{l15}	-OH	0.1	0.1	0.1	−0.950	0.1	0.1
t_{l16}	>C=C<	0.1	0.1	0.1	0.1	0.1	−0.067
t_{l17}	$N(1)H^+$ quinazoline	−0.016	−3.409	0.1	0.1	0.1	0.1
t_{l18}	(unused)	0.1	0.1	0.1	0.1	0.1	0.1
t_{l19}	-CHO	0.1	0.1	0.1	0.1	0.1	0.1

[a] In kcal.

whelming penalty, if the optimal binding algorithm so desires. For the remaining 20 energy parameters, we require the values giving optimal (rms) fit of the calculated binding free energies to the observed values, subject to the constraints that the individual energy interactions must be at least not unfavorable (i.e., values must be ≤ 0.0) and that there must not be any other geometrically allowed binding mode having a lower calculated energy than that of the desired mode for each of the ligands. The optimization part is simply a linear least-squares fitting problem, since the calculated energy is a sum of the interaction energies in the mode. The constraints are linear inequalities for the same reason. The total problem is exactly a "quadratic programming" operation, which can be reliably solved by standard methods such as the Wolfe algorithm.[6] As implemented, the quadratic programming step does not use a starting guess for the energy parameters, so the values found in the constant binding mode analysis are not needed here. Since there are very many geometrically allowed alternate binding modes, and only a few of these contribute significant inequalities to the problem, the approach was to first optimize without constraints, note the first ligand that found an energetically optimal binding mode other than the desired one given the energy parameters, add the corresponding inequality to the problem to eliminate the unwanted binding mode, reoptimize with current constraints, and so on. The iteration ends when at last the energy parameters that give an optimal fit to the observed binding free energies for the desired binding modes also give a higher (worse) calculated binding energy for any other mode for all ligands. The resultant energy parameters are shown in the body of Table IV.

With this method of refining the energies, we are able to give a good accounting of the number of degrees of freedom. In all, 20 parameters were adjustable, but at the final optimum one of these were set to zero by the quadratic programming algorithm; that is, it was "not used". By the time the final iteration was reached, 27 inequalities had been generated due to alternate better binding modes found along the way by some 20 of the 68 ligands. Of these inequalities, only 11 were still "active" at the solution, which means that they were restraining the optimal solution from moving into their disallowed zones. Hence, there were $20 - 1 - 11 = 8$ adjustable energy parameters.

Discussion

In summary, the systematic approach to rationalizing binding data as outlined in the previous section consists of five steps. (i) The series of ligands must be represented as points corresponding to atoms or groups of atoms. So far we have not developed any guidelines for these choices, and certainly the outcome of a study can depend upon them. (ii) The data set of ligands is then probed by seeing if the assumption of a constant binding mode for similar ligands is valid. This exploratory calculation is performed without further guidance by the investigator according to the decomposition algorithm, which is explained in more detail in the Appendix. If the constant binding mode assumption yields a satisfactory fit to the observed binding free energies, then the site and energy parameters can be easily produced. However, if the assumption is significantly in error, the analysis points to specific ligands as not fitting the general pattern. The geometries of the deduced base group and substituent groups are quite helpful in deciding how many different site points are required, especially by noting how many different substituents there are and which of them are structurally similar to others. (iii) Comparison of the indicated deviant ligands with similar but "regular" ones should inspire the investigator to propose certain alternate modes of binding for some of the ligands. This is the most subjective part of the whole procedure and should be considered as the formation of a hypothesis aided by the analysis of the data done in the previous step. The subsequent steps test that hypothesis objectively and thoroughly. Our model of ligand points coinciding with site points requires one to express the vague roles of site points (as shown in Table II) in a much more concrete and *testable* form. (iv) The set of proposed binding modes imply constraints on the arrangement of the site points in space. The geometry determination algorithm in the Appendix objectively deduces what these constraints are, and the distance geometry algorithms of ref 5 then produce a sampling of allowed site point coordinate sets. If the constraints are very confining, all the generated site configurations will be quite similar, and any one will do. Alternatively, if there is a wide range of permissible configurations, the investigator is faced with an arbitrary choice that probably will not make much difference for the present data set but could be important when the proposed site is tested in the future by including new ligands. At least one is made aware in this stage of the calculations of the *range* of geometric possibilities, as we have already mentioned under Methods. Note that deducing the geometry of the site is entirely separated in this scheme from deducing the interaction energies. (v) Only in the last stage do we determine the energy parameters. Some of them have already been tacitly fixed at large positive values by requiring that otherwise geometrically allowed binding modes for certain ligands be avoided (e.g., a pocket that will accomodate a methyl group, but not an ethyl). The others are determined objectively by the quadratic programming optimization of fit to the observed binding energies subject to maintaining the desired binding modes.

One might well ask if there is enough information in a set of binding data such as these 68 quinazolines to determine all the geometric and energetic parameters describing the site. After all, other QSAR methods produce only a few parameters to define their fit (Hansch et al.[7] using the same data set employ only six empirical parameters). Mathematical information theory[8] provides at least a rough answer by the following "order of magnitude" argument. The site, its geometry and interaction energies, can be thought of as a source of a lengthy message. Each ligand, along with its observed free energy of binding and its (flexible) structure, corresponds to a signal sent to us over a noisy channel, where the noise is the experimental error in the binding energies and the conformational variability of the ligand. The question then becomes: how many signals must be sent at least in order to deduce the original message, namely, the site? The information theory measure of uncertainty is

$$H = \sum_i p_i \log_2 p_i$$

where the p_i are the probabilities of receiving the ith type of signal, assuming the different possibilities are mutually independent. If some events are dependent, H will be less. H attains its maximum value, $\log_2 n$, when each of the n possibilities are equally probable. We first crudely estimate the maximum possible uncertainty in the proposed site in this work, and then from an equally rough estimate of the information conveyed by each ligand we can show how many ligands must be included in the data set to reduce the site uncertainty to zero. In the geometry of the site, there are $11 \times 3 - 6 = 27$ possibly mutually independent coordinates to be determined to an accuracy of ± 0.5 Å, the site flexibility parameter in this study. We see from Table III that the coordinates run over at most

(6) J. L. Kuester and J. H. Mize, "Optimization Techniques with Fortran", McGraw-Hill, New York, 1973, pp 106-119.

(7) C. Hansch, J. Y. Fukunaga, P. Y. C. Jow, and J. B. Hynes, *J. Med. Chem.*, **20**, 96 (1977).

(8) R. Ash, "Information Theory", Wiley Interscience, New York, 1965.

a 10-Å range each, so with the given accuracy, there are 10 distinguishable possibilities for each coordinate. Therefore, there are at most $H_{S:coord} = 27 \log_2 10 = 89.69$ bits of uncertainty in the site coordinates. Similarly, we suppose the 20 possibly independent energy parameters need to be determined within ±0.1 kcal out of a range of 3.4 kcal (see Table IV). Hence, there are 17 distinguishable energy levels for each parameter and $H_{S:energy} = 20 \log_2 17 = 81.75$ bits. The total maximum estimated uncertainty in the site $H_S = H_{S:coord} + H_{S:energy} = 171.4$ bits. The optimistic view is that the uncertanties in the data set, which measure the maximum number of bits of information obtainable from it, can be fully used to determine the large uncertainty in the site. The data set as a whole (considering now only the 66 ligands treated in the constant binding mode analysis decomposition algorithm explained under Methods) contains structural variability in the form of the different substituents. The base group does not count because it is found in all the 66 ligands and is therefore completely predictable. Of the 33 substituent groups found by the decomposition algorithm, 9 were *geometrically* identical with some of the others and therefore do not contribute to the variability. If the regions of space for each substituent (as determined by their upper and lower distance bounds to the base group) are all equal in volume and do not overlap (i.e., the most favorable case), then the total data set geometric uncertainty is $24 \log_2 24 = 110.04$ bits. If all this variability can be fully applied to deducing the site, then there remains 171.4 − 110.0 = 61.4 bits to be determined. Now from Table I we can see that the range of observed binding free energies is only about 8 kcal, and since the estimated error may be ±1 kcal, there are only four distinguishable energy levels. The distribution of ΔG_{obsd} values is fairly even over the data set and yields $H_{L:energy} = 1.971$ bits as the average uncertainty in energy per ligand. It is interesting to note how small this value is, which is due to the small range of values and large experimental error. In addition, each ligand also has an uncertainty due to its composition in terms of substituents. The geometry of each substituent has already been accounted for. Assuming the distribution of substituents through the set of ligands is mutually independent, the known frequencies of occurrences of each substituent in the data set yields $H_{L:comp} = 4.4$ bits as the average uncertainty per ligand in structure. Note that more information is potentially conveyed in the chemical structure of ligands than in their binding energies. The most information that could be extracted per ligand then is $H_{L:energy} + H_{L:comp} = 6.4$ bits, assuming that composition and binding energy *appear* to be independently distributed. Our very rough estimate for the minimum number of ligands we need to observe is 61.4/6.4 = 10 ligands. Thus, we can say that deducing a site of the complexity shown in Tables III and IV is at least conceivable in principle, although the above argument says nothing of how that may be accomplished. Clearly, so many approximations have been made that our conclusions must remain only qualitative.

Tables III and IV summarize the deduced site. The choices of number of site points and their types are somewhat arbitrary and are the same as in our previous paper.[2] Although it would have been entirely reasonable to include more than 11 site points in an effort to accomodate some of the very lengthy groups attached to position 6, such as in ligand **58**, we chose to see how well we could fit the 68 ligands with the same site points we had used previously, when only a subset of 20 of the simpler inhibitors had been considered. Note that there are only six distinct types of site points allowed, even though there is no good a priori reason for supposing that site points 4 and 5, for example, interact with the same types of ligand points in the same way. Instead, our objective has been to show how one can deal with a substantial data set and how the fit can be refined in spite of such arbitrary limitations in the proposed site. With the methodology presented here, we can devote more effort to improving the agreement with experiment in future studies.

Table IV shows the outcome of the quadratic programming refinement of the energy parameters. As we have already explained, the +10 kcal entries are an arbitrarily large repulsive interaction, and the exact value is undetermined. Similarly, because no contacts with an aldehyde group, t_{l19}, were encountered, its interaction energy with all site-point types remains at the default 0.1 kcal. All entries of +0.1 occur for this same reason. Of the remaining entries, many have similar values; for example, −0.084 ± 0.001 can be found four times. This is an artifact of the optimization procedure, where a typical constraint amounts to making one interaction energy slightly more favorable than another so that the desired binding mode will be at least marginally preferred over some other undesirable mode. The constrained optimum of fit tends to lie on the boundary of the region defined by these inequalities, and hence many near equalities of interaction energies are to be seen. This computational artifact is also the reason why the interaction energies of Table IV are quoted to such high accuracy. If they are rounded off, undesirable binding modes are calculated when several modes have the same (best) energy. Otherwise, the entries seem to make some chemical sense, although they were certainly not constrained to do so. Site point type 3 was intended as binding sites for hydrophobic groups, and indeed there are weakly favorable interactions with phenyl, CF_3, and chloro ligand points. Site points 4–6, being of type 4, consistently interact favorably with electronegative groups. Since these are empirically determined parameters that depend on the desired binding modes and the set of ligands under consideration, it is unreasonable to expect them to correspond closely to the ΔG_{bind} of the corresponding ligand point type to some cluster of atoms in a real protein, as calculated, say, by quantum mechanical methods.

Table I gives the experimental ΔG values of binding and those calculated using the variable mode analysis. The rms error is 1.33 kcal over the 68 ligands, compared to a 0.99 kcal error in our previous paper[2] using the same number of site points and types but considering only a subset of 22 of the ligands. The 1.33 rms error is rather unevenly distributed, there being ligands with an error between 3 and 4 kcal (no. **4**, **10**, and **14**), five in error between 2 and 3 kcal (no. **1**, **35**, **9**, **15**, and **16**), and 28 having errors between 2 and 4 kcal. Apparently, the fitting procedure has been biased by the preponderance of strongly binding 2,4-diamino derivatives, since ligands **36–68** all are fit relatively well. The rocking alternate binding modes for the N(1) (which were intended to account for the good binding of ligands **36–68** while explaining the poor binding of ligands 4, **10**, and **14**) were subverted by the energy parameter optimization to yield a good rms fit overall, but left ligand 4 with the worst single fit in spite of maintaining the desired binding modes in all cases. We believe this can be remedied by introducing extra site points to interact with the lengthier substituents and by relaxing the requirements that certain site points have the same type (see Table III). Our justification for this belief is that ligand **4** was fit perfectly well in our previous paper with the more

Table V. Comparison of Results of QSAR on 68 Quinazoline Inhibitors of Dihydrofolate Reductase by Three Methods

	method		
	Hansch et al.[a]	fixed mode[b]	variable mode
rms error[c]	1.05	1.17	1.33
max error[c]	4.82	5.46	4.0
no. of parameters	6	22	≤23

[a] See ref 7. [b] Parameters derived for 60 ligands; errors calculated for 66. See text. [c] In kcal.

limited data set, which included only 5- or 6-naphthyl derivatives, but otherwise employed the same kinds of desired binding modes with the same numbers and types of site points. Apparently, we are seeing difficulties more with empirical parameter fitting than with the type of model of ligand binding.

Table V gives a comparison of our overall results on the 68 quinazoline inhibitors of DHFR using fixed and variable mode analysis with those of Hansch et al.[7] Clearly, the Hansch approach is far superior in terms of number of variable parameters used, and their root mean square error in calculating the binding energies is somewhat better. Given that the experimental error in binding studies may be roughly ±1 kcal, it is difficult to say whether the differences in the rms fitting error are significant. Our variable binding mode calculations show some improvement in fitting the worst case. Note that the inhibitor with the largest error is quite different, depending on the method. That of Hansch et al. was no. **25** (for which our variable mode analysis has an error of 0.4 kcal), whereas our worst was no. 4 (which they fit quite well). Hence, there is no good reason to be suspicious of the experimental data at this point. The real differences in the methods lie in their predictions and general predictive power.

The Hansch method does not give much geometric detail (see Figure 1 of ref 7), although that is a goal of their work in ref 7. In summary, they find a "polar space" near the 6 position of quinazoline, a "hydrophobic space" near the 5 position, and a "sterically sensitive" region between the 4 and 5 positions.

If the constant binding mode assumption is actually correct, then the predicted best inhibitor would be simply the nonoverlapping combination of substituent groups in the data set that gives the lowest combined binding energy. There is considerable geometric detail in the locations of hypothesized binding sites for 14 geometrically distinct substituent groups.

On the other hand, if the variable binding mode hypothesis is correct and if the proposed rocking mode is realistic, then the derivative proposed in the previous paragraph happens to have a much worse predicted ΔG_{bind}. Because the variable binding mode analysis produces coordinates of site points, much more unusual sorts of high binding affinity ligands can be predicted. It is particularly interesting to see how one might proceed. Referring to Figure 1, we begin with the strongest ligand, no. **68**, except that now the best known group to bind at site point 8 is attached at the 5 position, and a short chain from the benzene ring should be able to reach site point 7 for a very favorable interaction. It should be clear that no other QSAR method would be able to make this kind of proposal, because it depends on a precise deduction of the three-dimensional structure of the binding site. A promising line of future experiments on this enzyme would be to probe the binding site by adding small aliphatic groups at various positions in hopes of detecting the steric

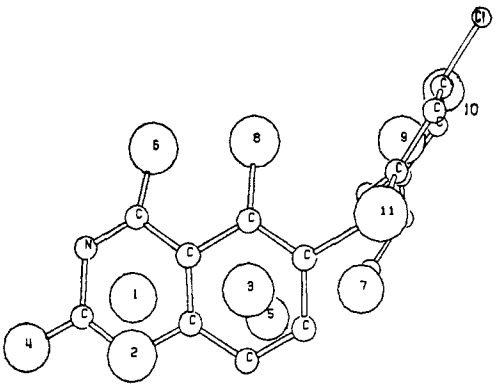


Figure 1. Proposed dihydrofolate reductase binding site geometry with a predicted type of high-affinity inhibitor in place. Large spheres are the locations of site points, numbered as in Tables II and III, and the small spheres are the nonhydrogen atoms of the ligand connected by bonds. Site point 1 lies in the foreground in front of the plane of the quinazoline ring system, while site points 2, 3, 4, 6, 8, and 11 are in the plane, and 5, 7, 9, and 10 are behind it. Where a ligand atom or group occupies a site point, only the large site point is shown. Thus, the linkage between the quinazoline and benzene rings occupies site point 11, etc. The benzene ring is seen edge on. The slight distortions from planarity and desired bond angles are the inconsequential result of minor numerical errors.

blocking groups that must surround the ligand in some directions. One could experimentally determine the allowed angle of rocking (if any) by linking the groups of the ligand in Figure 1 occupying sites 7 and 8 by varying lengths of aliphatic chain. We believe that the strength of this approach to QSAR lies in such stimulations to future research.

Appendix

Decomposition Algorithm. Given the chemical structures of all the ligands, coded as ligand points of various types with given upper and lower bound distances among the points, we wish to find the common structural feature, if any, and a small set of substituent groups that describe the differences. The substituent groups will be characterized by the number of ligand points involved, their types, their relative positions as specified by upper and lower distance bounds, and by their locations relative to the base group as specified by upper and lower distance bounds. Therefore, it is necessary that the base consist of at least three points (and preferably four) which are not all colinear (and preferably four which are not all coplanar) in order to correctly fix the locations of the substituents. Of course, it is possible that there may be no such base group common to *all* the ligands, in which case it would be desirable to subdivide the original set of ligands into more than one subset, each of which would have a common base. Once the base group has been determined, there are still a variety of ways to choose the substituents. For example, if at a given position in a series of analogues there is always either a $-H$ or a $-CH_2CH_3$, it is reasonable to deduce two substituents, $-H$ and $-CH_2CH_3$, rather than the three groups, $-H$, $-CH_2-$, and $-CH_3$. However, if an isopropyl group also occurs in that position, the latter set of substituents would be preferable, since the isopropyl group could be described as a methyne (taken to be the same type as a methylene) and two methyls. The policy we have taken is to choose the smallest number of substituents necessary to describe the entire set of ligands, even if that means having to represent any one ligand as a greater number of parts. From these considerations, we propose the following algorithm.

Step 1: Determine the Base Group. Of all the ways to pair some of the points in the first ligand with an equal number of points in the second, there is an optimal, not necessarily unique, matching that involves the greatest number of points such that the corresponding points in the two ligands have the same type and such that the range of allowed distances between two points in the first ligand overlaps the corresponding range in the second ligand, for all pairs of points used in the matching in the first ligand. We call the "intersection" of two ligands the collection of points in the optimal matching, their types, and the set of upper and lower distance bounds among them, always choosing for a given pair of points in the intersection the greater lower bound between the two ligands and the lesser upper bound. The matching is carried out by exhaustive enumeration of all the viable possibilities in a tree search that excludes classes of disallowed matches whenever possible. Clearly, the ordering of the points within a single ligand has no effect on the matching process, and the intersection always has no more points than the smaller of the two ligands. To find the base group of a set of ligands, first take the intersection of the first and second ligands, then take the intersection of the result with the third ligand, and so on until all ligands have been considered. If at any time in the process the number of points in the intersection should drop below 3, the latest ligand is declared to have no recognizable common structural feature with the others and is excluded from further consideration. The base group is the resultant intersection. It is certainly possible that the results depend on the sequence of the ligands when some of the ligands must be excluded, and the algorithm does not attempt to overcome this shortcoming.

Step 2: Determine the Substituent Groups. Having found a subset of ligands with a common base in the previous stage of the calculation, we now remove the base points from each ligand. The definition of intersection required in this step is the same as before except that, in addition, the ranges of distances from ligand point to the constant base must also overlap, and the overlap distance ranges to the base are included in the description of the intersection. For the points remaining in the first ligand and those left in the second ligand, find the intersection according to the above definition. Intersect the result with the remaining points of the third ligand and so on until no points remain in the intersection. Then the outcome of the previous intersection is a substituent group, and every occurrence of it in the set of ligands is removed from further consideration. Beginning the process again with the first ligand that still has some points left, determine the next substituent group, iterating until every point of every ligand has been accounted for.

Step 3: Termination. The initial set of ligands has now been completely broken down into a single common base group and a collection of substituents occurring one or more times each in some but not all the ligands. However, there may be other ligands that were excluded in step 1 on the grounds of having no common base with the others. We simply repeat steps 1 and 2 with these alone, independently of the results obtained so far. If any ligands still are excluded, iterate until all have been taken care of. Clearly the worst case is that every ligand stands alone in its separate class, having no common structure with any of the others.

Site Geometry Determination. Given the structure of the ligands and their desired modes of binding, we wish to calculate the bounds on the distances among the site points. It is important to remember that the calculation of the optimal binding mode given a ligand and a full site description allows the site the same small flexibility, δ, in each interpoint distance. Thus, a contact is allowed between ligand point La and site point Sa, when for every other contact between ligand point Lb and site point Sb both $d_{Sa,Sb} + \delta \geq l_{La,Lb}$ and $d_{Sa,Sb} - \delta \leq u_{La,Lb}$, where the subscripted d, l, and u are the distance, lower bound, and upper bound, respectively. Hence, the algorithm for deducing upper and lower bounds on the distances among the site points proceeds as follows.

Initially let the upper bounds be some large value and the lower bounds be zero. For each ligand in turn, consider every pair of contacts La–Sa and Lb–Sb. If the site point upper bound, $u_{S:Sa,Sb}$, is greater than the corresponding ligand point upper bound, $u_{L:La,Lb} + \delta$, then set the site point upper bound to the latter value. Similarly, whenever $l_{S:Sa,Sb} < l_{L:La,Lb} - \delta$, the site point lower bound is raised to the latter value. If the desired binding modes are unfortunately chosen, at some point in the process, a site point lower bound will exceed the corresponding upper bound. In such a case, it is clear that the desired binding mode of the ligand under present consideration is geometrically incompatible with that of some preceding ligand. A more detailed analysis enables one to pinpoint at least two conflicting ligand binding modes and the relevant desired contacts. Even is there is no conflict, inconsistencies may appear later if the δ used in this calculation is not slightly smaller than the δ for the subsequent optimal binding mode calculations. The reason is that the distance geometry algorithm for producing coordinates of site points from the bounds deduced here tends to result in some distances that are equal to or very slightly beyond the specified bounds. In such cases, an equal δ value in the optimal binding calculation may yield very slight geometric violations when attempting the desired binding mode, whereas a somewhat larger δ permits the desired mode.