Voronoi Modeling: The Binding of Triazines and Pyrimidines to *L. casei* Dihydrofolate Reductase

Mary P. Bradley and Gordon M. Crippen*

College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109-1065

Received May 3, 1993•

Vorom is a computer-aided method of drug design which can model a biological receptor given only binding data of known ligands. Using the binding energies of known competitive, reversible ligands of a biological macromolecule, vorom can make predictions about the binding energies and conformations of other small molecules binding to that receptor as well as provide information about the geometry and physicochemical characteristics of the binding site. One such model of *L. casei* dihydrofolate reductase was made. The model was able to predict the binding energies of 31 pyrimidine and triazine inhibitors out of a total set of 47, using only eight of the molecules (four pyrimidines and four triazines) as input. The binding energy of methotrexate, which is neither a pyrimidine nor a triazine, was correctly predicted. The binding mode of methotrexate predicted by vorom is entirely consistent with known X-ray data. The predicted binding modes of the pyrimidine inhibitors and the geometry of the site model are also consistent with published NMR and crystallographic data.

We have devised a method to objectively model the binding of small molecules to a biological receptor with reasonable predictive power and meaningful geometry given the experimentally determined binding energies for a set of small inhibitor molecules. The main features of this method, based on Voronoi binding sites, have already been described.[1-4] We chose to study two types of well-known inhibitors of *L. casei* dihydrofolate reductase (DHFR), 4,6-diamino-1,2-dihydro-2,2-dimethyl-1-(substituted phenyl)-*s*-triazine and 2,4-diamino-5-(substituted phenyl)pyrimidine (Figures 1 and 2), because of the availability of binding, NMR, and X-ray data required for validating the model.

As currently implemented, the method is very flexible in the level of detail of the structure of inhibitor molecules, the required accuracy of the fit to the experimentally determined binding constants, and the detail of the structure of the resulting binding site model. As the molecular structures are simplified, and as the fitting accuracy is decreased, the site model becomes simpler in geometry, and the required computer time decreases by orders of magnitude. In this work we show that even when the inhibitors are represented by greatly simplified structures, we can obtain a fit to the observed binding within 15%, resulting in a site model that is simple but nontrivial, has strong predictive power, and agrees quantitatively in geometry and qualitatively in energetics with the crystal structure.

Methods

Molecular Simulations' Quanta was used to generate a 3-D representation of the molecules and to obtain coordinates. The molecules were converted to linearized form[5] to facilitate future calculations. Each atom is then assigned hydrophobicity and molar refractivity parameters as described in[6,7], such that the sum of each physicochemical parameter over all atoms in a molecule approximates the molecular value. A summary of the global range of conformations available (excluding those conformations prohibited by van der Waals interactions) to the flexible molecules is obtained by allowing rotation about single bonds and noting the range of interatomic distances over all allowed conformations.

In order to minimize the CPU requirements of the Vorom modeling process, which increases exponentially with the number of atoms and unit vectors in the molecule, specified atom groups are condensed into a single *pseudoatom* which retains the composite properties of the component atoms. The new pseudoatoms are subsequently treated as actual atoms for the rest of the modeling sequence. The atomic coordinates of the composite atoms in each pseudoatom are averaged so that the new atom lies at the unweighted center of mass of the old atoms. Since the assigned physicochemical parameters of hydrophobicity and molar refractivity are additive over each atom, these are summed for each atom in the new composite. The upper/lower bound on the distance between two pseudoatoms is taken to be the greatest/least upper/lower bound between one atom in the first pseudoatom and another atom in the second. Since the conformation space of the molecule remains the same for the condensed molecule, we have the effect of a molecular skeleton (the squashed molecule) with the interatomic distances defining the space-filling features of the molecule. Reduction of the number of parameters in this manner is necessary for keeping the combinatorial search for optimal binding modes to a reasonable length.

Although the atoms which comprise the pseudoatom groups must be chosen by the investigator, this is no more random or unreasonable than choosing atoms of a pharmacophore within a set of molecules. The selection of which atoms should be condensed is subjective, but the choice need not be totally arbitrary, as the atom groups chosen must be convex. In a rigid molecule, a convex set of atoms is any subset such that their convex hull contains only that subset. For a conformationally flexible molecule, different convex sets may be possible, depending on the conformation. In large molecules, the total number of subsets of atoms is much greater than the number of convex sets of atoms. Often a rigid group of atoms in a molecule, such as a methyl group or aromatic ring, is convex and may be selected for redefinition. Generally, groups which

• Abstract published in *Advance ACS Abstracts*, September 15, 1993.

0022-2623/93/1836-3171$04.00/0

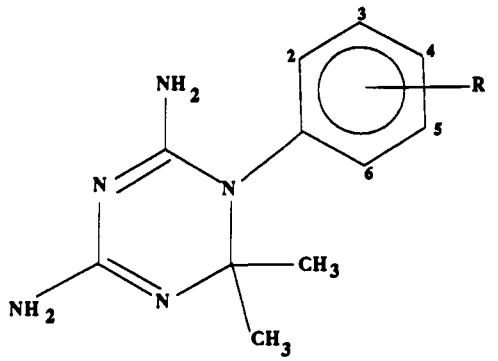

Figure 1. 4,6-Diamino-1,2-dihydro-2,2-dimethyl-1-(substituted phenyl)-*S*-triazine.

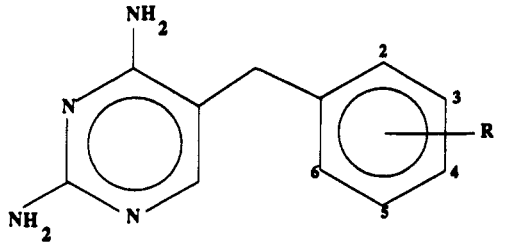

Figure 2. 2,4-Diamino-5-(substituted phenyl)pyrimidine.

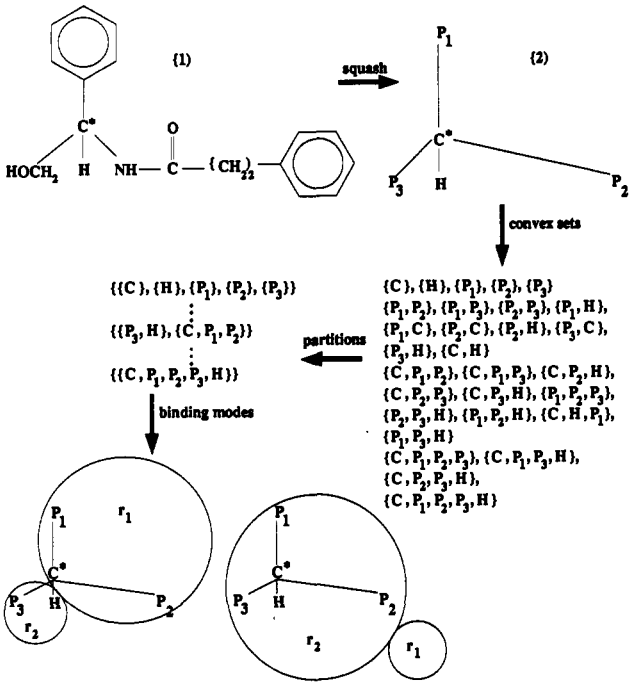

Figure 3. Illustrative flow diagram showing the treatment of one molecule: squashing, determining the convex sets (all listed), determining the partitions (only three examples listed), and finding binding modes with respect to two regions, r_1 and r_2 (only one mode shown for each of two of the listed partitions).

are common to many of the molecules in the dataset, such as the pyrimidine or triazine groups in our case, are good candidates for this process. Figure 3 shows the redefinition of a large molecule 1. Note that the condensed form of the molecule (2) still retains its chirality, and it has a distinct tetrahedral geometry which depends on the number of atoms, size, and geometry of the atom group being condensed. This helps to retain the unique shape of the molecules, even though they may be in a very simplified form. Furthermore, since the systematic search over all allowed conformations of 1 resulted in a large range of distances between the two benzene rings, the

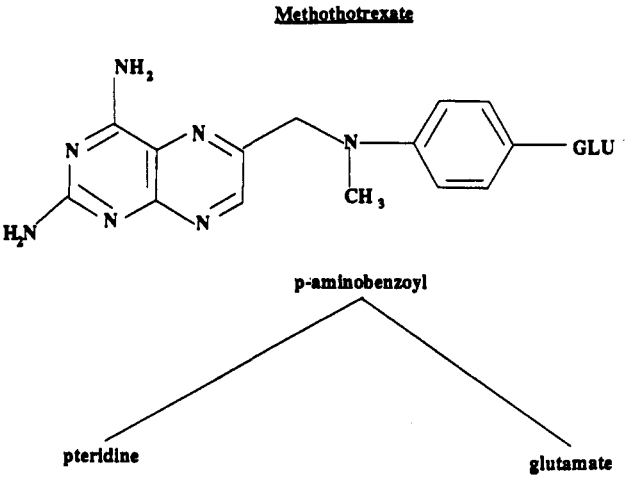

Figure 4. Composite pseudoatoms of methotrexate.

condensed form, 2, has a correspondingly large distance range between pseudoatoms P_1 and P_2. Figure 4 illustrates the molecular redefinition of the methotrexate molecule into three composite pseudoatoms: pteridine, *p*-aminobenzoyl, and glutamate.

Each molecule in the training set is broken down into convex sets of atoms. (These convex sets should not be confused with the molecular redefinition described above. In all that follows, we deal with the molecule in its abbreviated form. The term "atom" will be used to refer to any atom or pseudoatom component of the molecule.) There are $2^n - 1$ possible nonempty subsets of atoms (or pseudoatoms) for each molecule having n atoms; but not all of these are necessarily convex. Continuing along Figure 3, all the convex sets are listed for our condensed molecule, 2. Note that the maximum number of convex sets of atoms for a molecule containing 5 atoms is $2^5 - 1 = 31$, but our molecule has only 30 convex sets. The set of four atoms $\{P_1, P_2, P_3, H\}$ is not convex because the carbon lies inside the convex hull formed by the other atoms.

The convex sets are then grouped into partitions. We define a partition of a molecule to be a set of mutually exclusive and exhaustive convex sets. That is, each atom belongs to one and only one of the convex sets in the partition. For example, in 2, the convex set $\{C, P_1, P_2\}$ with three atoms may be combined with the $\{P_3, H\}$ subset to form a partition. Molecule 2 with 30 convex sets has 40 such possible partitions, of which only three are shown in the figure. The number of partitions in the molecule is generally found to be on the order of the number of convex sets.

The number of regions in the site is an unknown parameter, and it is necessary to make an initial empirical estimation. Keeping in mind our goal of determining the simplest geometry site model, we begin with simple site geometries (small number of regions). Because we are using simplified representations of the molecules, we expect our model to be a low resolution representation of the actual binding site. If all of the molecules in the training set are not fit in the first trial, the complexity of the site model is increased by adding more binding regions. The CPU required increases exponentially with the number of binding regions, so it makes sense to begin with as few regions as possible. Since a single region model will give little insight into the geometry of the site, we chose a two-region site model to begin. There was no solution in two regions, so another binding region was added.

The list of partitions will be used to determine the binding mode of the molecule in the binding site model. The use of partitions instead of individual atoms for binding mode determination dramatically reduces the number of combinatorial options available, since there are fewer partitions than the total number of atom subsets. The placement of the partitions among the binding regions is defined as the binding mode of the molecule. Each partition may have more than one binding mode. For example in Figure 3 if there are two regions, then the second listed partition has two possible binding modes: the one illustrated and the same with the two regions interchanged. The third listed partition also has two modes, one of which is shown. Note that a mode assigns zero or one of the subsets of a partition to each region.

The binding energy for any particular binding mode of molecule m is calculated by eq 1, in the same manner as described in ref 3

$$\Delta G_{m,\text{mode}} = \sum_{r}\sum_{a\in r}\sum_{p} V_{a,p}\epsilon_{r,p} \tag{1}$$

$$\Delta G_{m,\text{best}} = \max_{\forall\ \text{modes}} \Delta G_{m,\text{mode}} \tag{2}$$

where r is the region in the site, a is the atom assigned to one region by the mode, and p is the physicochemical property, either molar refractivity or hydrophobicity in this study. In order to satisfy molecule m, the calculated binding energy for the mode having the highest (best) binding must fall within the predetermined range of acceptable binding energies:

$$\Delta G_{m,\text{min}} \le \Delta G_{m,\text{best}} \le \Delta G_{m,\text{max}} \tag{3}$$

A new computer program, egsets, has been devised to solve for a site geometry and physicochemical parameters which satisfy the experimental binding energy ranges of all the molecules in the training set. See the next section for details of the algorithm and its treatment of a simple artificial example dataset.

In order to create a binding site with maximum predictive potential, keeping in mind that we also want to minimize the CPU time required, we have endeavored to use the smallest training set possible. The following systematic scheme for selecting molecules for the training set has been devised:

1. The number of atoms in each molecule was decreased to the smallest number possible which still retain some of the geometric characteristics of the original molecule (angle between functional groups, distance between groups, etc.). This decreases the number of combinatorial options and allows the computations to be completed within a reasonable time frame.

2. The original binding data[8,9] is in terms of the binding constant, K. Since our ΔGs need not be exactly Gibbs' free energies, but only some similar scale, we used $\Delta G_{\text{obsd}} = -\log K$. Then we took $\Delta G_{\text{min}} = 0.9\Delta G_{\text{obsd}}$ and $\Delta G_{\text{max}} = 1.1\Delta G_{\text{obsd}}$, which was always an expansion of the estimated error bars in references 8 and 9. This expansion of the error bars has the overall effect of simplifying the geometry of the site model but should still give a model with good predictive ability.

3. A molecule which is very large (in size) and has been assigned large hydrophobicity parameters and very low binding energy has been added to the training set for all of the trials. This ensures that an infinitely large and hydrophobic molecule (e.g., graphite) does not bind with infinitely large binding energy, and it allows for molecules of larger dimensions than those in the training set to fit into the site model; presumably this will provide the site model with a region which represents the infinite surrounding solvent.

4. Find the $\Delta G_{m,\text{best}}$ for all of the molecules in the data set in a single, infinite region over a large range of values for the two interaction energy parameters $\epsilon_{1,H}$ and $\epsilon_{1,\text{MR}}$ where H indicates the hydrophobicity and MR the molar refractivity parameter of the molecule which is located entirely in a single region. The difference in the calculated energies of each pair of molecules, i and j, is then integrated as in eq 4.

$$\int\int_{\text{finite }\epsilon} (\Delta G_{i,\text{best}} - \Delta G_{j,\text{best}})^2\, d\epsilon_{1,H} d\epsilon_{1,\text{MR}}\ (\forall i \neq j) \tag{4}$$

This gives a value which represents the energetic similarity of each pair of molecules in the dataset. A value of 0.0 corresponds to two molecules whose calculated energies are identical over a large range of interaction parameters.

5. The two molecules which have the greatest calculated energy difference from step 4 were used as the initial training set. These molecules are the most energetically different, and ideally both should be required for modeling.

6. Run egsets on the training set, in the simplest acceptable site geometry (i.e., fewest number of regions). If a solution is not obtained, it may be necessary to increase the complexity of the site geometry by increasing the number of regions or to expand the error bars on the experimental binding energies to allow for a simpler site geometry.

7. Attempt to predict the binding energies of the remaining molecules in the dataset with the solution obtained from egsets in step 6. If the ΔG_{best} for all of the molecules in the test set are correct (within assigned error bars, eq 3) stop; otherwise continue.

8. Use the calculated relative error (eq 5) to assess the quality of the solution and add the one molecule with the worst relative calculated error to the training set.

9. Repeat steps 6–8 until all of the molecules in the test set are predicted. For a given solution applied to test molecule i, the calculated relative error, δ_i, is

$$\delta_i = (\Delta G_{i,\text{best}} - \Delta G_{i,\text{min}})/(\Delta G_{i,\text{max}} - \Delta G_{i,\text{min}}) \tag{5}$$

and the mean relative error is

$$\delta_i = \frac{1}{N}\sum \delta_i \tag{6}$$

where N is the number of unique solutions used.

Egsets determines a complete list of all of the possible solutions which fit the given data. For some data sets, the total number of solutions determined by egsets can be quite large (>2000) but not all of the solutions are unique. Many contain the identical interaction energy parameters and interatomic distance bounds but have the regions renumbered. These solutions are identical, and only one is counted. Other solutions may have identical energy parameters or geometries. These are also considered duplicate solutions, and only the first occurrence is considered as unique. The unique solutions are evaluated for predictive power (total number of dataset molecules whose binding energies are correctly predicted, and error of the molecules which are not predicted), and the one(s) with the greatest number of compounds predicted are evaluated further for geometric feasibility. In the DHFR case, geometric feasibility was determined by the model's

ability to correctly predict the binding energy of methotrexate. Only those solutions which meet all of the above criteria are kept.

Egsets Algorithm

1. Assume rigid molecules. For each molecule note the following.
2. Find all convex sets of atoms.
3. Group convex sets into all possible partitions.
4. For every convex subset pair s_i and s_j, intersubset distance bounds are

$$u_{ij} = \max_{a_i \epsilon s_i,\ a_j \epsilon s_j} d(a_i,a_j)$$

and

$$l_{ij} = \min_{a_i \epsilon s_i,\ a_j \epsilon s_j} d(a_i,a_j)$$

where $d(a_i,a_j)$ is the distance between atom a_i and a_j.

5. For every ordered quartet of convex sets occurring in a partition, note its chirality, which is defined as the chirality of the four centers of mass, $\check{c}_1$, $\check{c}_2$, $\check{c}_3$, $\check{c}_4$.

$$\chi([1,2,3,4]) = det\begin{pmatrix} \check{c}_1 & \check{c}_2 & \check{c}_3 & \check{c}_4 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Record χs only if $\neq 0$.

6. For all u_i and l_i over all molecules, order them on the real line and find the set of midpoints between clusters of values, plus a grand lower and upper bound. These are the *critical distances* that need to be used in constructing significantly different site parameters.
7. For number of regions = 1, 2, ..., given upper bound, find solutions.
8. Consider the tree of possible choices of optimal binding modes, where the nth layer of nodes are the choices for the nth molecule. Carry out a depth-first search of the tree, checking each node that a solution can be found. If there is a solution, try the children of the node; otherwise backtrack up the tree one layer.
9. Given choices for the optimal partitions of the first n molecules and the mode for each partition.
10. Choose the tightest geometry site that permits the proposed optimal modes. Site geometry is described only in terms of upper, u'_{ij}, and lower, l'_{ij}, bounds on distances between regions i and j. Also keep a list of some chiral relations among quartets of regions. Then put convex set s_i in region i' and set s_j in region j' only if ranges $[l_{ij},u_{ij}]$ and $[l'_{ij},u'_{ij}]$ overlap. Also if chirality $\chi(s_h,s_i,s_j,s_k) > 0$ and these go in regions h',i',j',k', respectively, then we must have $\chi(h',i',j',k') > 0$ also. *Tightest geometry* is least u's and greatest l's and correct χs such that all of the chosen optimal modes are geometrically allowed.
11. Then for each molecule m under consideration, $m = (1,...,n)$, and each nonoptimal but geometrically allowed mode, note inequality $\Delta G_{m,\text{allowed}} < \Delta G_{m,\text{optimal}}$ in addition to $\Delta G_{m,\min} < \Delta G_{m,\text{optimal}} < \Delta G_{m,\max}$. Solve all these linear inequalities. If there is a solution, record it.
12. Go to the next tree node in 8.

Egsets Example

As an artificial illustration of how egsets works, suppose that carbon monoxide binds to some receptor site with $\Delta G_{\text{obs}} = 100 \pm 0.1$ kcal/mol and carbon dioxide binds to it with $\Delta G_{\text{obs}} = 10 \pm 0.1$ kcal/mol. Let the unique atom labels of the oxygens of CO_2 be O2 and O3. Imagine there are two "physicochemical parameters" such that any C atom has $V_{C,1} = 1$ and $V_{C,2} = 0$, whereas any O atom has the reverse: $V_{O,1} = 0$ and $V_{O,2} = 1$. For simplicity of notation, denote the corresponding first adjustable parameter of region r by $\epsilon_{r,C}$ and the second by $\epsilon_{r,O}$. Then referring to the outline of steps above, egsets carries out the following calculations.

(Step 2) CO has three convex sets, {C}, {O}, and {C,O}, while CO_2 has six: {C}, {O2}, {O3}, {C, O2}, {C, O3}, and {C, O2, O3}. Note that {O2, O3} is not a convex set for CO_2 because it is a linear molecule with C in the middle. (Step 3) The partitions of CO are {{C}, {O}} and {{C, O}}; for CO_2 they are {{C}, {O2}, {O3}}, {{C, O2}, {O3}}, {{C, O3}, {O2}}, and {{C, O2, O3}}. (Step 4) If the C–O bond length is 1.3 Å, the distance between subsets {O2} and {C, O3} lies in the range 1.3–2.6 Å, because the closest pair of atoms between the two subsets is O2 and C, and the most distant pair is O2 and O3. For conformationally flexible molecules, one would use the pairs with the closest lower bound and most distant upper bound. By the same reasoning, {C, O3} has a diameter of 1.3 Å. (Step 5) These linear molecules have no chirality to worry about. (Step 6) The critical distances are "small" = 0.6, 1.8, and "large" = 3.2 Å because every intersubset distance falls between these values. In this example, the distances will not turn out to be important.

(Step 7) Try for a solution having only one region. This implies only those partitions consisting of one subset are applicable. Trivially CO has only the binding mode where {C, O} is in region r_1, and CO_2 has only one binding mode, putting {C, O2, O3} into r_1. According to eq 1, r_1 has adjustable parameters $\epsilon_{1,C}$ and $\epsilon_{1,O}$, corresponding to its interaction with C and O atoms, respectively. We must solve

$$99.9 \le \epsilon_{1,C} + \epsilon_{1,O} \le 100.1 \text{ for CO}$$

$$9.9 \le \epsilon_{1,C} + 2\epsilon_{1,O} \le 10.1 \text{ for } CO_2$$

but these inequalities are inconsistent.

(Step 8) We must try for a two-region model. Both partitions of CO and the one- and two-subset partitions of CO_2 can be used. For a one-subset partition, there are two modes, depending on whether the subset is put into r_1 or r_2, and for a two-subset partition there are also two modes, depending on whether the first subset is put into r_1 and the second into r_2 or vice versa. There are now four ϵs to be determined.

(Step 9) Try making optimal the CO mode where {C} goes into r_1 and {O} goes into r_2. (Step 10) The tightest compatible geometry is that the distance between the regions is 1.3 Å, and the diameter of each is "small". (Step 11) Energetically this choice of optimal binding mode implies that $99.9 \le \epsilon_{1,C} + \epsilon_{2,O} \le 100.0$ in order to satisfy the observed binding, $\epsilon_{1,O} + \epsilon_{2,C} \le \epsilon_{1,C} + \epsilon_{2,O}$ to make the opposite mode of the same partition suboptimal, and $\epsilon_{1,C} + \epsilon_{1,O} \le \epsilon_{1,C} + \epsilon_{2,O}$ and $\epsilon_{2,C} + \epsilon_{2,O} \le \epsilon_{1,C} + \epsilon_{2,O}$ to make the two modes of the {{C},{O}} partition suboptimal. At this point, the last two inequalities are not required because putting all of CO into either one of the regions is precluded by their small diameter. The rest of the inequalities can be solved, so we proceed to the next molecule.

(Step 9) Try making optimal the CO_2 mode where {O3} goes into r_1 and {C, O2} goes into r_2. (Step 10) Clearly the diameter of r_2 must be expanded to greater than 1.3 Å, but

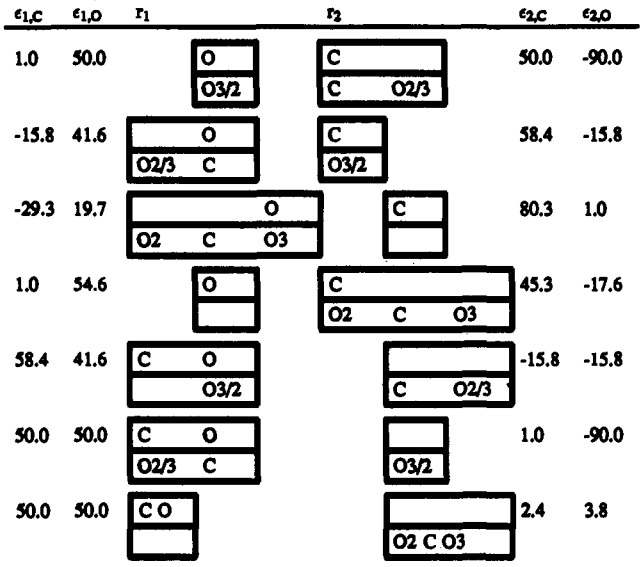

Figure 5. All egsets solutions to the artificial example of CO and CO_2 binding data. The optimal binding modes are shown in each solution for both molecules in boxes sized to represent approximately the diameters and separations of the two regions.

this simultaneously permits CO to put both atoms into r_2. (Step 11) Combine the three relevant energy inequalities for CO with the two arising in this situation for CO_2 and solve. There is a solution, namely $\epsilon_{1,O} = \epsilon_{2,C} = 50$, $\epsilon_{2,O} = -90$, and $\epsilon_{1,C} = 1$. It is easy to verify that the designated optimal modes of the two molecules have calculated binding energies in agreement with the observed values, and all other modes are either excluded on geometric grounds or have no better binding energies.

(Step 12) Since there are no more molecules, we have reached a solution, and one can go back and try other choices of optimal modes for the two molecules. In all, there are seven distinct solutions, shown in Figure 5. These differ in geometric terms by exploiting differences in distances between atoms in the two molecules. Due to the symmetry of CO_2, some of the solutions can be realized in two ways by exchanging O2 and O3.

Results

A solution with three binding regions was obtained for the binding of 47 pyrimidines and triazines to *L. casei* DHFR. The model required only eight molecules (four pyrimidines and four triazines) to correctly predict the binding energies of the remaining 23/39 molecules in the test set (see Tables I and II). The binding energies of the training set molecules were correct, as required by the modeling method. The adjusted average relative error, calculated from eq 11 for the molecules whose binding energy was not correctly predicted, was 0.83

$$\bar{\delta} = \frac{1}{\tilde{n}} \sum_i \begin{Bmatrix} |\delta_i|, \text{ if } \delta_i \le 0 \\ \delta_i - 1, \text{ if } \delta_i > 1 \end{Bmatrix} \qquad (11)$$

where $\tilde{n}$ is the number of nonpredicted molecules and only the errors of nonpredicted molecules are included.

In general, the relative calculated error among the nonpredicted molecules decreased as the number of molecules in the training set increased. This is a positive sign, since if all of the molecules were included in the training set, the error would be zero by design. The addition of the "infinitely" large hydrophobic molecule also increased the predictive power of the model.

Table I. Calculated Binding of Triazine Inhibitors

			predicted binding mode[c,d]		
R[a]	$\Delta G_{max} - \Delta G_{min}$[b]	ΔG_{calc}[c]	region 1	region 2	region 3
H	4.23–5.17	4.23	T	Ar	
3-I	4.66–5.70	5.47	R	Ar	T
3-OBzCl$_2$	5.01–6.13	6.07	R	Ar, T	
4-OMe	3.69–4.51	4.51	T	Ar, R	
3-SO_2NH_2	2.64–3.22	4.79	T	Ar, R	
3-$COCH_3$	3.82–4.68	4.68*	T	Ar, R	
3-OH	3.465–4.274	4.65	T		Ar
3-CF_3	4.29–5.25	4.41*	T	Ar, R	
3-F	4.39–5.37	4.30	R	Ar	T
3-CN	4.78–5.84	4.72	T		Ar
3-CH_3	4.46–5.46	4.76*	R	Ar	T
3-Et	4.68–5.49	4.55	T	Ar, R	
3-OMe	4.07–4.97	4.51*	T	Ar, R	
3-OEt	4.67–5.71	4.90*	T	Ar	R
3-OPr	5.02–6.14	4.79	T	Ar, R	
3-OHx	5.12–6.26	5.42*	R	T	Ar
3-OBz	5.11–6.25	5.27*	T	Ar, R	
3-CH_2OPh	5.91–7.23	5.27	T	Ar, R	
4-OH	4.42–5.40	4.65*	T		Ar
4-NH_2	3.55–4.33	5.53	T		Ar
4-I	3.99–4.87	4.60*	T	Ar, R	
3-CH_3	3.75–4.59	4.41*	T	Ar, R	
4-F	4.18–5.12	4.24*	R	Ar	T

[a] Training set molecules in bold. [b] $-\log(K) \pm 10\%$, ref 9. [c] Calculated by egsets, * indicates correct prediction. [d] See Table IV for description of regions.

Table II. Calculated Binding of Pyrimidine Inhibitors

			predicted binding mode[c,d]		
R[a]	$\Delta G_{max} - \Delta G_{min}$[b]	ΔG_{calc}[c]	region 1	region 2	region 3
H	4.68–5.72	5.72	Ar		P
3-OBu	5.52–6.74	6.34	Ar	R	P
4-I	6.00–7.34	6.00	Ar	R	P
3,4,5-(OMe)$_3$	6.19–7.57	6.30	Ar	R1,R2,R3	P
3-F	4.84–5.92	5.64*	Ar	R	P
3-CH_2OH	5.10–6.24	6.00*	Ar	R	P
4-NH_2	4.92–6.02	5.86*	Ar	R	P
3,5-$(CH_2OH)_2$	5.16–6.30	6.28*	Ar	R1,R2	P
4-F	5.10–6.24	5.64*	Ar	R	P
3,4-$(OH)_2$	5.26–6.42	5.72*	Ar	R1,R2	P
3-OH	5.24–6.40	5.72*	Ar	R	P
4-CH_3	5.25–6.41	4.33	P	Ar, R	
3-CH_2OBu	4.94–6.04	6.58	Ar	R	P
3-CH_3	5.20–6.36	5.88*	Ar	R	P
4-OMe	5.62–6.88	4.43	P	Ar, R	
4-OBu	5.73–7.00	4.96	P	R	Ar
4-$NHCOCH_3$	5.44–6.66	4.81	P	R	Ar
3-OMe	5.34–6.52	5.91*	Ar	R	P
3-OBz	5.54–6.76	6.68	Ar	R	P
3-CF_3	5.54–6.78	5.88*	Ar	R	P
3-CF_3,4-OMe	6.57–8.03	6.07	Ar	R1,R2	P
3,4-$(OMe)_2$	6.22–7.61	6.11	Ar	R1,R2	P
3,5-$(OMe)_2$	5.78–7.06	6.11*	Ar	R1,R2	P
3,5-$(OH)_2$	3.04–3.72	5.72	Ar	R1,R2	P

[a] Training set molecules in bold face. [b] $-\log(K) \pm 10\%$, ref 8. [c] Calculated by egsets, * indicates correct prediction. [d] See Table IV for description of regions.

The model of the binding site from egsets superimposed on the *L. casei* DHFR crystal structure was created using the distance geometry program, DGEOM.[10] The site point, c_r, represents the center of the spherical binding region, r. The distance bounds from Table III were used as interatomic distance constraints between molecules when superimposing the optimal binding modes of the training set molecules onto the site points. The upper triangle of the matrix represents the maximum distance between two atoms which lie in different regions, and the lower triangle

Table III. Distance Bounds Matrix for Site Model from Egsets[a]

region	1	2	3
1	1.7	∞	7.0
2	0.0	∞	8.0 (5.2)
3	4.3	4.1 (3.8)	3.7 (3.0–4.0)

[a] Actual distances from Cheung et al.[16] in parentheses.

Table IV. Binding Site Characteristics of Model

region	hydrophobicity[a]	binds MTX moiety	binds pyrimidine moiety[b]
1	0.069	*N*-methyl-*p*-aminobenzoyl	benzene ring
2	−0.066	glutamate	−X
3	−1.150	pteridine	pyrimidine

[a] From egsets solution. [b] In all correctly predicted molecules.

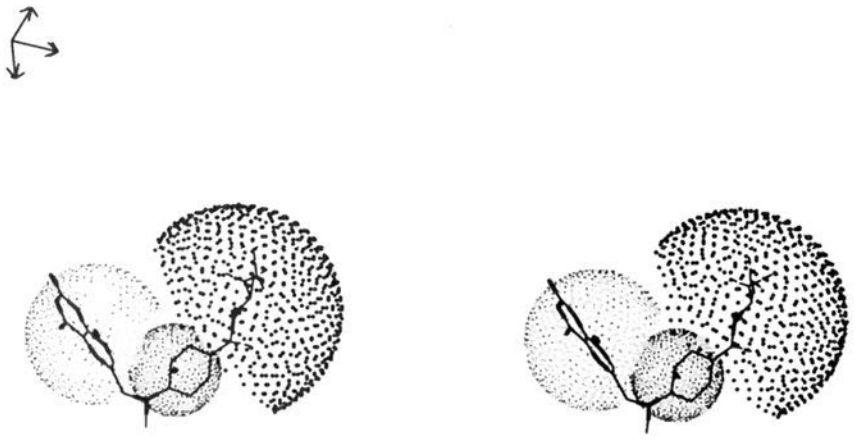

Figure 6. Site model of methotrexate bound to *L. casei* dihydrofolate reductase.

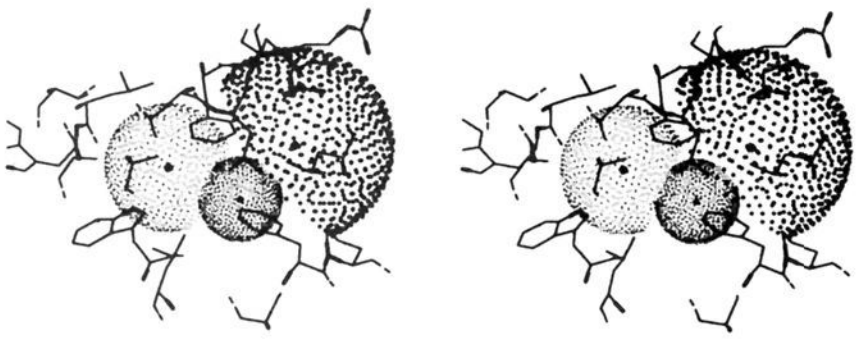

Figure 7. Site model of *L. casei* dihydrofolate reductase superimposed on binding site from crystal structure.[11]

represents the minimum distance between two points in different regions. The maximum intraregion distances lie on the diagonal of the matrix, and the minimum intraregion distances are 0.0 for all regions. The upper bounds on the distances from the site points, c_r, to the atoms in region r is set to one-half the maximum intraregion distance, since these are to be the center points of the regions. The interregion distance bounds between the site points is kept as per egsets during the DGEOM calculations. To evaluate how well the model approximates the actual site, the resulting site points from DGEOM were superimposed as a rigid body on the structure of bound methotrexate from the crystal structure[11] using Quanta Molecular Similarity. The site point representing the center of the pteridine binding region was placed close to the center of the pteridine ring, and so on. Since we know from egsets which region binds which MTX group, and the site points were treated as a rigid body, the overall result of this exercise is merely to translate and rotate the coordinates of the binding site region centers to the same coordinate frame as the crystal structure.

The binding energy ($-\log(K_i) > 9.0$)[12] of methotrexate (MTX) was correctly predicted. The binding mode of MTX calculated by egsets (see Table IV) corresponds to the one shown in the crystal structure.[11] Dihydrofolate was correctly predicted to bind 3 orders of magnitude worse than methotrexate, although in a different binding mode than is indicated by X-ray data.[13] Since dihydrofolate binds its pteridine ring rotated 180° compared to MTX, and the pteridine ring has been compressed and cannot therefore reflect this rotation, we would expect the difference in the binding energies of the two compounds to be reflected by different binding modes. The prediction of the subtle difference in the binding modes of the two compounds may have been achieved by egsets if the pteridine rings had been left intact, or the number of binding regions increased.

The site records from the DHFR crystal structure define four distinct binding regions. We combined two of these regions (*N*-methyl and *p*-aminobenzoyl) to give a larger single region with more meaningful geometric and physicochemical characteristics. Region 1 interacts with the *N*-methyl and *p*-aminobenzoyl groups of the MTX, region 2 interacts with the glutamate, and region 3 binds the pteridine. For purposes of prediction, MTX was squashed in the same manner as the other data set molecules into three pseudoatoms: pteridine, *N*-methyl-*p*-aminobenzoyl, and glutamate groups. The predicted regional placement of these groups corresponds to the regional placement in the crystal structure, see Figure 6. In addition, all of the pyrimidine inhibitors whose binding energies were correctly predicted had the same binding mode: pyrimidine in region 3 (pteridine region), aromatic ring in region 1 (paba region), and the aromatic constituent in region 2. This mode of binding is entirely consistent with the binding mode of trimethoprim, a pyrimidine inhibitor, to *L. casei* DHFR proposed by Cayley *et al.*[14] and confirmed by Roberts.[15] Comparison of the inter- and intraregion distances derived from egsets with those determined by NMR,[16] listed in Table III, show that our model is geometrically compatible with the actual binding of pyrimidine inhibitors. See Figure 7. The majority of the triazines whose binding energies were correctly predicted tended to have optimal binding modes which placed the triazine group in the paba region and the aromatic ring in the glutamate region (Table I). None of the binding modes placed the aromatic ring in the paba region analogous to the pyrimidines or MTX. This inconsistency in the

Table V. Binding Site of *L. casei* Dihydrofolate Reductase[11]

region	residues	binds MTX moiety	no. of hydrophilic residues	calcd hydrophobicity[a]
1	LEU19, LEU27, PHE30, SER48, PHE49, PRO50, LEU54	*N*-methyl-*p*-aminobenzoyl	1	−1.03
2	LEU27, HIS28, PHE30, ARG31, LEU54, ARG57	glutamate	3	−6.24
3	LEU4, TRP5, ALA6, LEU19, ASP26, LEU27, PHE30, ALA97, THR116,K HOH201, HIOH217, HOH253	pteridine	5	−9.35

[a] Hydrophobicity parameters calculated from[7] using free amino acis and then summed over all residues in the region.

triazine binding may be due to the fact that the triazines tend to be better inhibitors of vertebrate DHFR, and we are using the data for inhibition of bacterial DHFR.

Table V shows the relative hydrophobicities of the combined residues in the active site. The hydrophobicities were obtained using the same method as that used for calculating the physicochemical parameters of the ligands. Values for the uncharged free amino acids were used for calculating the hydrophobicity parameter of the regions. (While this hydrophobicity value does not reflect the *actual* hydrophobicity of the binding region, it gives a good indication of the *trend* of relative hydrophobicity of the side chains in the binding region, since the hydrophobic contribution of the backbone is the same for all the amino acids. This trend is also reflected in the number of hydrophilic residues versus calculated hydrophobicity.) The resulting values for each residue were then summed for all of the residues in each region. The hydrophobicities derived from the egsets solution are not identical, but note that the trend of the pteridine binding in the most hydrophilic region, the *N*-methyl-*p*-aminobenzoyl group binding in the most hydrophobic region, and so on is reproduced by egsets. The differences in the calculated hydrophobicities of the regions may be due in part to the use of the free amino acids in the calculation.

In order to assess the statistical validity of the model, the usual statistical methods employed in QSAR modeling, such as standard deviation, were not employed here, since our method is a global combinatorial search and not a statistical fitting method. To determine that our model was based on the experimental binding energies of the data set molecules, the binding energies were scrambled (reassigned among the training set molecules), and egsets was rerun on the training set with the incorrect energies. When the resulting model was used to predict the energies of the remaining molecules, only five were correctly predicted. It would be preferable to have had many more of these scrambling validations; however the length of time required for numerous egsets runs is prohibitory. We have employed the scrambled energy validation on other test sets with identical results. Along the same lines, we have found that even changing the ordinal arrangement of the binding energies can lead to a substantially different solution.[17]

We have used experimental binding data to fit very simplified molecules to a model with three binding regions. This should not be viewed as an oversimplification of the problem but rather an avoidance of overinterpretation of the problem. If we had used the complete molecules (of approximately 35 atoms), we would not expect the experimental binding data to support a model with 35 binding regions! A model which predicts all of the dataset molecules within the given error bars certainly seems attainable, but time constraints prohibited further exploration at this time, as there were indications that the complexity of the site geometry may have to be increased.

Acknowledgment. This work was supported by grants from the National Institutes of Health (GM37123) and the National Institute of Drug Abuse (DA06746).

References

(1) Crippen, G. M. Deduction of binding site structure from ligand binding data. *Annals N.Y. Acad. Sci.* **1984**, *439*, 1–11.
(2) Boulu, L. G.; Crippen, G. M. Voronoi binding site models: Calculation of binding modes and influence of drug binding data accuracy. *J. Comput. Chem.* **1989**, *10*(5), 673–682.
(3) Boulu, L. G.; Crippen, G. M.; Barton, H. A.; Kwon, H.; Marletta, M. A. Voronoi binding site model of a polycyclic aromatic hydrocarbon binding protein. *J. Med. Chem.* **1990**, *33*, 771–775.
(4) Crippen, G. M. Voronoi binding site models. *J. Comput. Chem.* **1987**, *8*(7), 943–955.
(5) Crippen, G.; Smellie, A.; Richardson, W. Conformation sampling by a general linearized embedding algorithm. *J. Comput. Chem.* **1992**, *13*(10), 1262–1274.
(6) Ghose, A. K.; Pritchett, A.; Crippen, G. M. Atomic physicochemical parameters for three dimensional structure directed quantitative structure activity relationships iii: Modeling hydrophobic interactions. *J. Comput. Chem.* **1988**, *9*(1), 80–90.
(7) Viswanadan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for the three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*(163), 163–172.
(8) Hansch, C.; Hathaway, B. A.; Guo, Z. r.; Dias Selassie, C.; Dietrich, S. W.; Blaney, J. M.; Langridge, R.; Volz, K. W.; Kaufman, B. T. Crystallography, quantitative structure-activity relationships, and molecular graphics in a comparative analysis of inhibition of dihydrofolate reductase from chicken liver and lactobacillus casei by 4,6-diamino-1,2-dihydro-2,2-dimethyl-1-(substituted-phenyl)-s-triazines. *J. Med. Chem.* **1984**, *27*, 129–143.
(9) Hansch, C.; Li, R.-l.; Blaney, J. M.; Langridge, R. Comparison of inhibition of Escherichia coli and Lactobacillus casei dihydrofolate reductase by 2,4-diamino-5-(substituted-benzyl)pyrimidines: Quantitative structure-activity relationships, x-ray crystallography, and computer graphics in structure-activity analysis. *J. Med. Chem.* **1982**, *25*, 777–784.
(10) Blaney, J. M.; Crippen, G. M.; Dearing, A.; Dixon, S. *DGEOM, 590*. Quantum Chemistry Program Exchange, Indiana University, Bloomington, IN, 1990.
(11) Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 angstroms resolution. i. General features and binding of methotrexate. *J. Biol. Chem.* **1982**, *257*, 13650–62.
(12) *Principles of Medicinal Chemistry*, 3rd ed.; Foye, W. O., Ed.; Lea and Feibiger: 1989.
(13) Freisheim, J. H.; Matthews, D. A. Folate antagonists as therapeutic agents. In *The Comparative Biochemistry of Dihydrofolate Reductases*; Sirotnak, F. M., Birchall, J. J., Ensminger, W. B., Montgomery, J. A., Eds.; Academic Press, Inc.: 1984; Vol. 1, pp 94–117.
(14) Cayley, P. J.; Albrand, J. P.; Roberts, G. C. K.; Piper, E. A.; Burgen, A. S. V. Nuclear magnetic resonance studies of the binding of trimethoprim to dihydrofolate reductase. *Biochemistry* **1979**, *18*(18), 3886–3894.
(15) Roberts, G. C. K.; Feeney, J.; Burgen, A. S. V.; Daluge, S. The charge state of trimethoprim bound to Lactobacillus casei dihydrofolate reductase. *FEBS Lett.* **1981**, *131*(1), 85–88.
(16) Cheung, H. T. A.; Searle, M. S.; Feeney, J.; Birdsall, B.; Roberts, G. C. K.; Kompis, I.; Hammond, S. J. Trimethoprim binding to Lactobacillus casei dihydrofolate reductase: A ^{13}C NMR study using selectively ^{13}C enriched trimethoprim. *Biochemistry* **1986**, *25*, 1925–31.
(17) Bradley, M.; Richardson, W.; Crippen, G. M. Deducing molecular similarity using voronoi binding site models. *J. Chem. Inf. Comput. Sci.*, in press.