Voronoi Binding Site Models

Gordon M. Crippen

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

Received 14 October 1986; accepted 15 January 1987

A frequently occurring problem in drug design and enzymology is that the binding constants for several compounds to the same site are known, but the geometry and energetic interactions of the site are not. This paper presents in detail a novel approach to the problem which accurately but compactly represents the allowed conformation space of each ligand, accurately depicts their three-dimensional structures, and realistically allows each ligand to adopt the conformation and positioning in the site which is most favorable energetically. The investigator supplies only the ligand structures and observed binding free energies, along with a proposed site geometry. With no further assumptions about how the ligands bind and what parts of the ligands are important in determining the binding, the algorithm fits the observed binding energies without leaving outliers, predicts exactly how each of the given ligands binds in the site, and predicts the strength and mode of binding of new compounds, regardless of chemical similarity to the original set of ligands. The method is illustrated by devising a simple site that accounts for the binding of five polychlorinated biphenyls to thyroxine binding prealbumin. This model then predicts the binding energies correctly for an additional six biphenyls, and fails on one compound.

INTRODUCTION

The problem at hand is: what can we deduce about the structure and energetics of a binding site given the chemical structures and observed binding energies for several compounds? In order to understand the motivation for the novel approach proposed in this paper, it is necessary to briefly review previous methods,[1,2] ranked in order of increasing physical realism. In topological methods, the ligand's three-dimensional structure is disregarded entirely, keeping only information about which pairs of atoms are bonded. Then various graph-theoretical features of the interatomic connectivity are correlated in a least-squares sense with the molecules' observed activity. Inasmuch as there is an incidental connection between these topological features and molecular size, shape, polarity, etc., one can obtain a structure-activity relationship. One is left with no picture of what the site may be, however. More physically realistic approaches start with either explicit or implicit assumptions about how the different ligands bind to the site, and then either explicitly or implicitly superimpose them. Even if the ligands are chemically very similar, this imposes a great burden on the investigator, who has little basis for assuming that, say, a ring system common to all active compounds will always be positioned the same way in the binding site. Methods of this class vary in quality of representation of the ligands all the way from stylized two-dimensional structural diagrams to full three-dimensional structures with a thorough search over all energetically favorable conformation space. The structure of the site thus deduced is implicitly the compliment of the superposition of the active compounds. The energetics are described generally as a least-squares fit of observed binding energy to a linear combination of physicochemical properties of the ligands and/or significant parts of them. What constitutes "significant" groups in these molecules is, unfortunately, another subjective choice by the investigator. Further progression toward physical realism and detailed site representation is hampered by greater computational complexity and cost. In our distance geometry approach to the problem (see reference (3) and references therein,) we have allowed the ligand molecules to explore all of their respective conformation spaces energetically available to them, and permit them to bind in the conformation and positioning, i.e. *binding mode*, which is energetically most favorable. The site is *explicitly* represented as a collection

Journal of Computational Chemistry, Vol. 8, No. 7, 943–955 (1987)

CCC 0192-8651/87/070943-13$04.00

of points or binding pockets with three-dimensional coordinates specified for each. The investigator's additional input into such a study is a working hypothesis about which pockets will be occupied by which significant parts of each ligand. Although this hypothesis may be revised by the algorithm, it certainly influences the final site model structure and energetics. Energetics of binding is expressed as a table of interaction energies between site pockets and either significant groups themselves, or proportionality constants with their physicochemical properties.[4] The possibility is kept open that structurally similar compounds may prefer to bind in very different conformations or modes, and chemically dissimilar compounds can contribute to deducing the site model and/or can have their binding energies and optimal modes predicted. This is the most physically realistic site deduction method currently available, because the calculated site geometry agrees quantitatively with the crystal structure of the receptor protein, when known, and the interaction energies agree qualitatively with interactions seen in the crystal structure.[5]

Nothing is perfect, and even if one is willing to go to the computational trouble of using the distance geometry approach, there are drawbacks which have prompted this article on Voronoi site models. This type of site model was discussed earlier,[6] but now we have implemented the method in computer programs. The drive is to be physically yet more accurate and realistic without overinterpreting the data. The ultimate goal is to have an automated algorithm which proceeds directly from the chemical structures and observed binding of the ligands, to as simple and noncommittal a model for the site as is required to account for the data. We want to eliminate all subjective decisions by the investigator. In particular, we desire improvement on the following points, and exactly how these will be implemented will be covered in the Methods section of this article.

(i) The interaction energies are derived from a least-squares fit between observed and calculated binding by the distance geometry method and many others. This is appropriate if the adjustable parameters are supposed to account for all the variation in the observations except for a physically unimportant random error factor beyond the control of the experimenter. Suppose, on the other hand, that the experimentalist has made a conservative estimate of his unavoidable errors and supplies the binding data with error ranges, saying that compound m must bind no worse than ΔG_{m-} and no better than ΔG_{m+} (in this article we quote what is actually $-\Delta G_{binding}$, so that greater positive values denote improved interaction). Then all the uncontrolled random variables of the experimental system have been expressed by these ranges, and a statistical approach is no longer appropriate. We must alter the binding model until the calculated binding values fall in their respective ranges:

$$\Delta G_{m-} \leq \Delta G_{m,calc} \leq \Delta G_{m+} \qquad (1)$$

for all m. If 99 compounds are well fitted by some model, but the 100th one is not, then it is not an "outlier" caused by a rare large value of some uninteresting random variable, but rather, the model must be revised until all 100 compounds obey eq. (1). Note that this absolute fitting approach allows us to say unambiguously whether or not the model fits the data.

(ii) If the ligand molecules are thought to be free to adopt a conformation which best fits into the site, it is computationally efficient to first globally search over all allowed conformation space and summarize the results in some compact form. The distance geometry approach does this by noting the maximum and minimum interatomic distance for every pair of atoms, taken over all energetically allowed conformations. Unfortunately, the molecule really has many fewer degrees of conformational freedom than entries in such a table, so that although each interatomic distance is correctly constrained to lie in a certain range, the values taken on by two such atom pairs are correlated in general. For example, both the 3- and 5-position substituents of a freely rotating phenyl ring will show the same range of distances to another atom, but when the 3-substituent is near, the 5-substituent is far, and vice versa. Representing the molecule by a smaller set of better chosen conformational parameters eliminates many geometrically impossible binding modes.

(iii) Distance geometry binding studies at this point "edit" the molecules, representing each one by only a subset of the original atoms, thought to be the most significant

groups in determining the binding. This is a subjective decision (common with many other methods, although usually not expressed so explicitly) that is required to keep the combinatorial search for optimal binding modes to a feasible length. It is also inherent in the model's discrete contacts between a site point and at most one atom. For example, a site point might have a favorable interaction with a methyl group, but the formalism of the model requires one of the three hydrogens or perhaps the carbon to form a contact with the site point, while the other three atoms of the group contribute no interaction. Therefore, it is preferable to remove the hydrogens and think of the carbon atom as a "significant group", namely the whole methyl, that can then form a contact with the site point.

(iv) The distance geometry approach determines the interaction energies in a way that is influenced by the investigator's proposed binding modes. If only one binding mode were geometrically possible for each molecule, then one could simply adjust the interaction energies involved so that the calculated binding energies agreed as well as possible with the observed values. Unfortunately, there are many modes that must be considered, and interaction energies which give a good fit for one mode may allow another mode to bind even more tightly. More precisely,

$$\Delta G_{m,calc} = \max_{\mathbf{b} \in B_m} \Delta G(\mathbf{b}) \qquad (2)$$

where B_m = the set of geometrically allowed binding modes for molecule m, and $\Delta G(\mathbf{b})$ is the total interaction energy for the mode **b**, maintaining the convention that algebraically greater values correspond to better binding. We finally achieve self-consistent interaction energies by either modifying the proposed modes or introducing linear inequality constraints on the least squares fit until the ΔG_{calc} values in eq. (2) agree optimally with the observed ones. Achieving self-consistency is unfortunately influenced by the investigator's original binding hypothesis, a source of subjectivity better avoided.

(v) Representing the receptor as discrete site points has certain computational advantages, but it tends to require many site points unless the binding is largely determined by only a few key groups on the ligand. As an extreme example, suppose the 12 biphenyls shown in Table III bound at a perfectly featureless receptor site, such as a large hydrophobic cavity. Distance geometry would nonetheless demand 10 site points, each positioned so that it could bind a different substituent. Only then could the model detect which substituents were available for interaction with the cavity. With fewer site points, some substituents may have no available site point with which to interact regardless of mode, so that although the group lies inside the cavity, its contribution to the total energy would be omitted. Even with all 10 site points, the predicted binding modes would be unjustifiably precise in that each molecule would be locked into a particular positioning in the site, whereas in reality, they would have a choice of many different modes, since the observed binding reflects only general hydrophobicity.

METHODS

Fitting binding data with a Voronoi site model consists of the following steps: (I) examine and summarize the conformation space of each ligand molecule; (II) propose a site geometry; (III) determine all geometrically allowed binding modes of the molecules; and (IV) determine the interaction parameters. The only step which can fail is the last, and in that event, one must return to step II and try a site of a different shape. The rest of the Methods section explains how each of these steps is carried out, and how the new approach improves on the shortcomings of the distance geometry method listed in the Introduction. The Results section will go through the steps once again for a simple example data set. All programs described here are written in the C language and run on a SUN/3 computer.[7]

Linearized Molecular Representation

Examining the conformation space (Step I) is carried out just as we and several other groups have done for some time. Molecule m is assumed to have rigid bond lengths and bond angles such that its conformation can be described in terms of a vector of dihedral angles, ϕ_m. For any choice of ϕ, there corresponds an internal energy, $E(\phi)$, which might be calculated by some molecular mechanics program. Ours consists merely of checking van der Waals contacts. Then, there is an allowed conformation space, Φ_m, consisting of

all conformations of adequately low energy:

$$\Phi_m = \{\phi_m \,|\, E(\phi_m) \leq E_m^* + \Delta E\} \quad (3)$$

where E_m^* is the global minimum of conformational energy, and ΔE is on the order of kT, but less than the observed free energy of binding. In fact, we discretize Φ by specifying a small number of allowed values for each dihedral angle and then trying all such combinations. What values to use is dictated by computer time constraints. Now instead of summarizing Φ in terms of allowed ranges for all interatomic distances as in point (ii) in the **Introduction**, we use ranges on the dot products of intramolecular unit vectors. In order to explain this, consider the linearized representation of the molecule shown in Figure 1.

The overall translation of the molecule is expressed as some unconstrained vector, **w**, chosen to point at one atom, C1 in this case. Then the overall rotation of the molecule is determined by the unit vector, $\mathbf{u}_1$, which in this example is defined as the unit vector running from C1 toward C7. For a nonlinear molecule, an additional vector is required to express the positions of atoms off the $\mathbf{u}_1$ axis, such as $\mathbf{u}_2$, defined as the unit vector running from C7 toward C8. A planar group needs two unit vectors to specify atom positions, and of course, a rigid nonplanar group would require a third. Viewing the C1-C7 and C4-O15 bonds as rotatable demands $\mathbf{u}_4$ and $\mathbf{u}_5$, respectively. Technically $\mathbf{u}_3$ could be expressed as a linear combination of $\mathbf{u}_1$ and $\mathbf{u}_2$, but it is produced due to an imperfection in the linearization algorithm. The net result is that the location of each atom in the molecule can be expressed as a linear combination of vectors, where the coefficients are given for compound **5** in Table I. Thus, for example,

$$\mathbf{Cl14} = \mathbf{w} - 4.49\mathbf{u}_1 - 3.10\mathbf{u}_4 \quad (4)$$

This is something like setting up a coordinate system to express atom positions, where the unit vectors resemble the axes, except that the axes are not orthogonal, and we need more than three axes because rotatable bonds introduce extra degrees of freedom.

A general algorithm for going from an ordinary molecule description in terms of atom coordinates and bonds to this linearized version is frankly quite complicated. Those who are interested will find it outlined in the **Appendix**. Note that the sort of distance correlation problems mentioned in item (ii) of the **Introduction** is automatically eliminated: whatever values $\mathbf{u}_1$ and $\mathbf{u}_4$ may have, the coefficients in Table I ensure that Cl13 and Cl14, for example, are on opposite sides of the ring.

Once the molecule is linearized, a straightforward recursive program searches out the available conformation space exhaustively, as outlined above, and the maximum and minimum $\mathbf{u}_i \cdot \mathbf{u}_j$ are noted for each unit vector pair. The result in the case of our example molecule is shown in Table II. This provides a compact summary of Φ in terms which are well suited to the site representation dis-

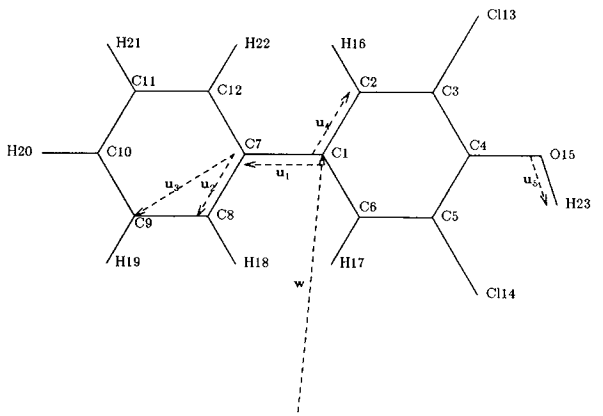


Figure 1. Representation of compound **5**, 3,5-dichloro-4-hydroxy biphenyl, as an arbitrary vector **w**, and five unit vectors, $\mathbf{u}_1, \cdots, \mathbf{u}_5$.

Table I. Coefficients for representing the atomic coordinates in Fig. 1 as linear combinations of an arbitrary vector, **w**, and five unit vectors, $\mathbf{u}_1, \cdots, \mathbf{u}_5$.

atom	**w**	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$
C1	1.0	0.0	0.0	0.0	0.0	0.0
C2	1.0	0.0	0.0	0.0	1.39	0.0
C3	1.0	−1.39	0.0	0.0	1.39	0.0
C4	1.0	−2.78	0.0	0.0	0.0	0.0
C5	1.0	−2.78	0.0	0.0	−1.39	0.0
C6	1.0	−1.39	0.0	0.0	−1.39	0.0
C7	1.0	1.48	0.0	0.0	0.0	0.0
C8	1.0	1.48	1.39	0.0	0.0	0.0
C9	1.0	1.48	0.0	2.41	0.0	0.0
C10	1.0	1.48	−2.78	4.82	0.0	0.0
C11	1.0	1.48	−4.17	4.82	0.0	0.0
C12	1.0	1.48	−2.78	2.41	0.0	0.0
Cl13	1.0	−1.39	0.0	0.0	3.10	0.0
Cl14	1.0	−4.49	0.0	0.0	−3.10	0.0
O15	1.0	−4.13	0.0	0.0	0.0	0.0
H16	1.0	1.05	0.0	0.0	2.44	0.0
H17	1.0	−1.39	0.0	0.0	−2.44	0.0
H18	1.0	1.48	3.49	−1.82	0.0	0.0
H19	1.0	1.48	1.05	2.41	0.0	0.0
H20	1.0	1.48	−3.83	6.63	0.0	0.0
H21	1.0	1.48	−6.27	6.63	0.0	0.0
H22	1.0	1.48	−3.83	2.41	0.0	0.0
H23	1.0	−4.13	0.0	0.0	0.0	1.00

Table II. Dot product ranges for the vectors $\mathbf{u}_1, \cdots, \mathbf{u}_5$, as shown in Fig. 1. Upper bounds are shown in the upper triangle and lower bounds in the lower.

	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$
$\mathbf{u}_1$	1.0	0.50	0.87	−0.50	−0.31
$\mathbf{u}_2$	0.50	1.0	0.87	0.50	0.67
$\mathbf{u}_3$	0.87	0.87	1.0	0.00	0.21
$\mathbf{u}_4$	−0.50	−1.00	−0.87	1.0	0.98
$\mathbf{u}_5$	−0.31	−0.98	−0.74	−0.67	1.0

cussed in the next section. It is also strongly reminiscent of orientational correlation between the ends of a polymer chain,[8] although I was not conscious of the connection at the time I devised this method.

Voronoi Site Representation

In order to reduce unjustifiable detail in the site model's shape and excessive precision in the predicted binding modes (item (v) in the **Introduction**), we choose to represent the site not as points, but as non-overlapping regions covering all space. Thus, in the extreme example of the featureless hydrophobic cavity, one could use a single infinite region. Each atom would always lie in one and only one region, and a binding mode would consist of a listing of the region in which each atom is located. In the featureless cavity case, any molecule would have only a single binding mode, namely all atoms lying in the one region, and the orientation of the molecule remains appropriately vague. A convenient way to define such a subdivision of all space is in terms of Voronoi polyhedra,[9,10] sometimes known as Dirichlet tesselations. Suppose the investigator has decided to try a site model with n_s regions. Then he must supply the coordinates of "generating points" $\mathbf{c}_i$, $i = 1, \ldots, n_s$, chosen such that corresponding to each is a Voronoi polyhedron, or "site region" r_i, consisting of the locus of all points that are closer to it than to any other generating point:

$$r_i = \{\mathbf{x} \,|\, \|\mathbf{c}_i - \mathbf{x}\| < \|\mathbf{c}_j - \mathbf{x}\|, \forall j \neq i\} \quad (5)$$

For example, Figure 2 shows a two-dimensional site model consisting of five regions arising from the chosen generating points $\mathbf{c}_1, \ldots, \mathbf{c}_5$. In this paper, we will deal only directly with eq. (5), but Voronoi polyhedra have other potentially useful properties explained in the **Appendix**.

Given the binding site structure in terms of Voronoi polyhedra, and a ligand molecule m consisting of n_m atoms, each binding mode $\mathbf{b}$ can be encoded as a vector, where the ith component, $b_i = k$ when atom i lies in region r_k. Because the site regions are non-overlapping and altogether space filling, there is always a k for each of the n_m atoms, and the choice is unambiguous. Letting $\mathbf{p}_i$ denote the position of atom i, and supposing r_j is adjacent to r_k, then for $b_i = k$ to be true, by eq. (5), we must have

$$\|\mathbf{p}_i - \mathbf{c}_k\|^2 < \|\mathbf{p}_i - \mathbf{c}_j\|^2 \quad (6)$$

Since $\mathbf{p}_i$ is really a linear expression of the form of eq. (4) in the Cartesian components of $\mathbf{w}, \mathbf{u}_1, \ldots$, it is easy to show that eq. (6) is actually linear in these components. In other words, a geometrically feasible binding mode has a solution to a set of linear inequalities of the form of eq. (6) involving all adjacent regions for each atom. In addition, there are nonlinear equations and inequalities to be satisfied: the unit vectors must each have

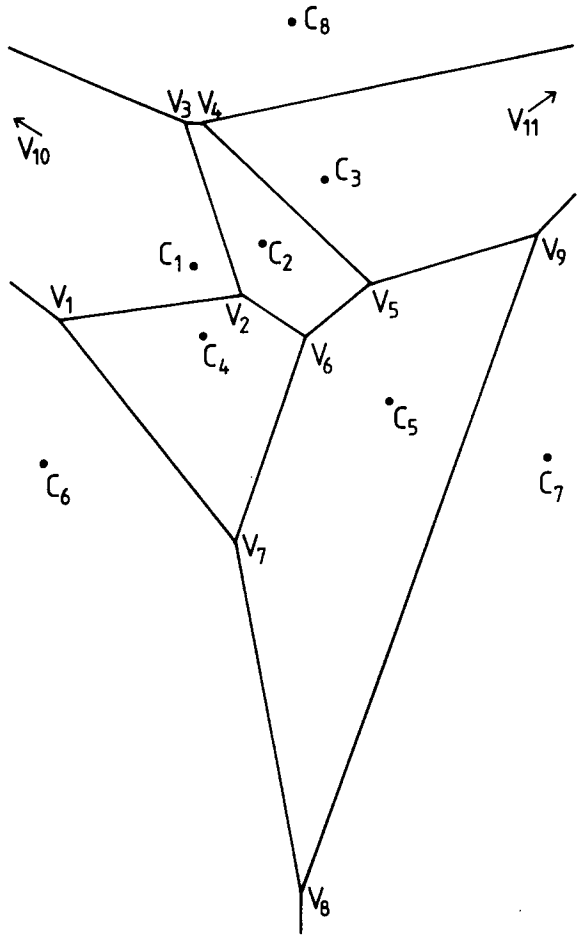


Figure 2. An example of a two-dimensional Voronoi site model consisting of five finite regions specified by generator points $\mathbf{c}_1, \ldots, \mathbf{c}_5$ and outriggers (see **Appendix**) $\mathbf{c}_6$, $\mathbf{c}_7$, $\mathbf{c}_8$. Vertices $\mathbf{V}_{10}$ and $\mathbf{V}_{11}$ lie at a large but finite distance out of view. Solid lines are boundary edges between regions.

unit length ($\|\mathbf{u}\|^2 = 1$) and dot products between pairs of unit vectors must obey the bounds found by exploring conformation space, for example as in Table II. How all this is solved, even approximately, is given in the Appendix. Suffice it to say, a computer program uses the generation point coordinates supplied by the investigator together with the linearized description of the molecule to produce a list of all geometrically allowed binding modes, using a moderately efficient algorithm that solves sets of linear inequalities of the form of eq. (6). Because all atomic positions are determined by a few vectors, no atoms need be excluded, as in item (iii) of the Introduction. Necessary conditions for geometric realism are checked, but in the present state of development, they are not always sufficient. Therefore, at the end of a binding study, when the energetically optimal mode has been located for each molecule, it is necessary to check by computer graphics whether those modes can in fact be achieved.

Energy Fitting

If each molecule had only one binding mode, determining the interaction energies would be a simple matter of solving a set of linear inequalities. As an illustration of the real difficulties, however, consider the extremely simple case shown in Figure 3 of two molecules residing in a small one-dimensional site consisting of two regions, r_1 and r_2. The short isomer, AA, consists of two "A" atoms and can fit into either r_1, r_2, or straddle the boundary, for a total of three modes: (1,1), (1,2), and (2,2). The long isomer, A—A, has the same two atoms fixed at such a long separation that only the mode (1,2) is possible. There are only two interaction energies to determine: e_1 for r_1 and A, and e_2 for r_2 and A. Determining the energies amounts to picking an appropriate point in the e_1e_2-plane. If the given binding of A—A is the range 4 to 5 kcal, then the one mode (1,2) translates into two inequalities: $e_1 + e_2 > 4$ and $e_1 + e_2 < 5$. In Figure 3 this corresponds to the diagonal band. In general, for any mode $\mathbf{b}$, there is a corresponding energy space vector β, whose integer components indicate how many times each interaction energy is invoked in binding. In this example, $\mathbf{b} = (1,2)$ and $\beta = (1,1)$. Now for the short AA isomer, if the binding range is 3.0 to 3.5 kcal, the modes, taken from left to right in the figure, correspond to $3 < 2e_1 < 3.5$, $3 < e_1 + e_2 < 3.5$, and $3 < 2e_2 < 3.5$. Since the molecule always prefers the highest (most favorable) energy mode, the first is taken when $e_1 > e_2$, the third when $e_2 > e_1$, and all three are equally favored when $e_1 = e_2$. This corresponds to the bent region in Figure 3. Note that the upper bounds for one molecule must always be satisfied, but it is only necessary that at least one lower bound be satisfied. The desired interaction energies clearly correspond to the *two* shaded regions. Linear programming could deal with the total set of upper bounds, which must all be satisfied and define the convex region outlined in heavy lines. However, it cannot cope with satisfying at least one lower bound for each molecule, which causes the feasible region to be possibly discontiguous. In the Appendix we explain our method for determining the interaction energies, based on subgradient optimization.

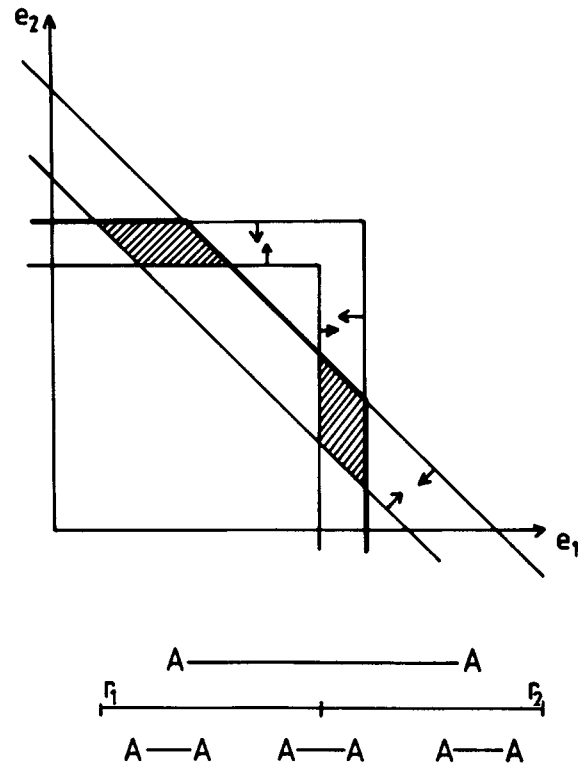


Figure 3. Below is depicted a one-dimensional site model consisting of two small regions, r_1 and r_2. The long A—A molecule can have only the one binding mode shown, but a short AA isomer could have three different modes. Above is the corresponding interaction energy space diagram, where the diagonal band corresponds to the A—A allowed energies, the L-shaped band is those energies permitted by AA, and the shaded parallelograms are the solution sets. The region determined by all upper bounds is marked with heavier lines.

The important concepts to note at this point are that determining the interaction energy parameters is not a least squares fitting procedure, but rather, the absolute fitting advocated in item (i) of the Introduction; the problem is relatively difficult because the molecules are allowed the realistic freedom to seek the binding mode of optimal energy; and the investigator *does not* have to supply suggestions about interaction energy values or optimal binding modes (item (iv) of the Introduction).

EXPERIMENTAL DATA

We selected the study of Rickenbacher, et al.[11] as a small but nontrivial test of the new method. They measured the specific binding of twelve biphenyl derivatives to prealbumin, a serum protein important in thyroxine transport.

They give the experimentally determined binding of these compounds in terms of I_{50}, the molar concentration at 50% total binding, but without any explicit statement of estimated error. The experiments may indeed have been done very well, yet absolute fitting requires given ranges on the observed binding. The only hint their paper gives is that their measured K_a for tetraiodothyronine binding to prealbumin gives "good agreement" with the values found by others. Altogether the quoted K_as range from 8.6×10^7 M^{-1} to 1.3×10^7 M^1, a multiplicative factor of 6.6. In this paper we add $\pm\ln 6.6 = \pm 1.9$ to their $-\ln I_{50}$ values to get the ΔG_- and ΔG_+ values given in Table III. This is a conservative interpretation of their data, but in the absence of further experiments, it is not unreasonable.

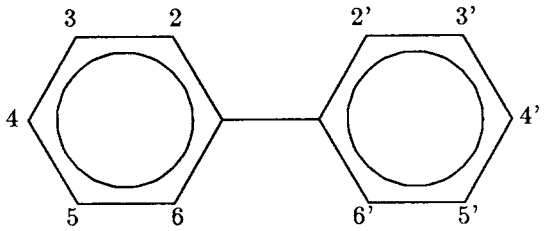
RESULTS AND DISCUSSION

Step I proceeded quite routinely because there are few rotatable bonds in these molecules, so conformation space could be explored with 60° increments in dihedral angles, or smaller. The only energy calculations done were to check van der Waals contacts. Linearization created in each case a small number of unit vectors, but not the minimum, although that is not an important shortcoming. Even so, as in Figure 1, it is a much more compact description than considering all atoms independently, due to the rigid phenyl rings. The outcome for each molecule is much as that shown in Tables I and II. Clearly this step is only a preliminary manipulation of the twelve molecules, and is not directly related to constructing the site model.

Step II was a little more challenging. The simplest possible site consists of a single large region, but it is impossible to fit the data for even only compounds **1**, **2**, and **3** this way. The next most complicated site has two large regions separated by a plane, and this also failed to explain the data for the first three compounds. Going to three regions at last requires a choice as to the shape of the three regions. Noting that **2** binds much worse than **3**, in spite of their identical atomic composition, naturally leads to the idea that the ability to adopt a planar conformation permits better binding. Of course, the *o*-Cl substitutions in **2** force it to adopt conformations where the two phenyl rings do not lie in the same plane, and this was properly encoded in the allowed dot product ranges. The simplest sort of site which could detect planarity vs. nonplanarity would consist of a thin slab separated on either side from large regions by two parallel planes. The generating point coordinates corresponding to a 1 Å thick slab are simply (−1,0,0), (0,0,0), and (+1,0,0), or any rigid translation or rotation of them. This final site model geometry follows from a consideration of only compounds **1**, **2**, and **3**.

The most time consuming part of the whole process was step III, determining all geometrically allowed binding modes of the molecules. This data set happens to be not very chemically diverse, in that each molecule has exactly the same carbon skeleton, and the only other atoms are Cl, O, and H. We therefore decided that the interaction of C with all three regions would be fixed in advance at zero, in an effort to reduce the number of adjustable energy parameters. This simplifies the binding modes considerably and consequently makes the search for all of them faster per mode and shorter altogether. In addition, it is not necessary for the subsequent energy fitting to include extra modes which are *energetically* identical. For example, for **1** if all atoms lie in region 1 except for a single H in region 2, there are 10 different ways this can be done because there are 10 hydrogen atoms in the molecule, and the large size of region 1 permits all the different orientations required to push each H across the planar boundary into region 2. The calcu-

Table III. Observed and Calculated Binding of Biphenyls to Prealbumin.



Compound	$\Delta G_{-}^{a,b}$	$\Delta G_{+}^{a,b}$	ΔG_{calc}^{a}	Optimal Mode[c]
1 biphenyl	0.0	14.2	14.2	all in r_3
2 2,2′,6,6′-Cl_4	0.0	14.2	14.2	2-6, 2′ in r_2 rest in r_1
3 3,3′,5,5′-Cl_4	17.9	21.7	18.7	2 in r_3 rest in r_2
4 3,3′,4,4′,5,5′-Cl_6	17.7	21.5	21.4	2 in r_3 rest in r_2
5 3,5-Cl_2-4-OH	18.6	22.4	18.7	OH, 2′, 3′ in r_3 rest in r_2
6 3,5-Cl_2-2-OH	15.3	19.1	19.1	3, 4, 5, in r_2 rest in r_3
7 2,4,6-Cl_3-4′-OH	15.7	19.5	17.1	4, 2′ in r_1 2, 3, 5, 6, 3′ in r_2 rest in r_3
8 3,5,4′-Cl_3-4-OH	18.1	21.9	20.0	OH, 2′, 3′ in r_3 rest in r_2
9 2,3,4,5-Cl_4-4′-OH	17.6	21.4	20.7	2′ in r_1 OH, 2′, 3′ in r_3 rest in r_2
10 2,3,5,6,-Cl_4-4,4′-$(OH)_2$	17.9	21.7	21.7	2′, 3′, 4′ in r_3 rest in r_2
11 3,3′,5,5′-Cl_4-4,4′-$(OH)_2$	18.4	22.2	22.3	4-OH, 4′-H in r_3 rest in r_2
12 3,3′,4,4′,5,5′,-Cl_6-2-OH	17.3	21.1	23.9	OH in r_3 rest in r_2

[a]In terms of $-\ln I_{50}$, where I_{50} is the molar concentration at 50% total binding.
[b]Derived from reference (11).
[c]Each atom must lie in one and only one region, but we are neglecting the carbon atoms in this series. Thus 2 through 6 and 2′ through 6′ refer to the substituents at those positions on the two rings.

lated energy of all these modes is the same, regardless of the interaction energy values, so even if one of these modes is the optimal one, it suffices to include only one representative of the ten. Even so, the largest molecules with the most diversity of atom types produced as many as 900 binding modes. Just as step I, this step involves processing all twelve molecules for subsequent purposes, but it does not directly affect the site model.

Step IV, determining the interaction parameters, was fairly straightforward once it became clear that three regions were required. Compounds **1**, **2**, **4**, **5**, and **6** were enough to guide the random search to the parameters shown in Table IV. In other words, these five compounds constituted the training set, and the other seven were used for testing predictions. Some fifty other binding "strategies" were explored but discarded as they appeared not to converge on a solution.

Table IV. Interaction Parameters ($\ln I_{50}$ units) for Thin Slab Site Model Binding Biphenyls.

	Site Regions		
Atom	r_1	r_2	r_3
Cl	0.468	2.677	−3.746
H	0.591	1.320	1.422
O	0.088	1.030	2.460

Curiously enough, the optimal binding modes corresponding to these energies did not reflect the conscious strategy that led to proposing the thin slab model. Namely, **2** indeed has a nonplanar structure that forces it to put some of its atoms in r_1 and the others in r_2, but

the planar isomer **3**, which was to lie completely in r_2, prefers to push one H into r_3. The ΔG_{calc} column in Table III was produced from these parameters. Of the remaining seven compounds, only **12** was substantially mispredicted. Apparently additional site geometric detail is needed to account for the unexpectedly low contribution to binding made by the chloro substitutions. At this level of site model structure, all we can say is that it explains the binding of five compounds and correctly predicts the binding of six more, yet any simpler site cannot even do that. It is unreasonable to expect the model's structure to correspond to that of the real prealbumin. Instead, the model should be viewed as a device to discriminate among the twelve compounds and map their perceived structure into a calculated binding energy, just as the real binding site does, but not necessarily by the same mechanism. Eventually, a more detailed site model should have attractive regions where the real site has attractive binding pockets, and the calculated optimal binding modes should reflect how the molecules actually prefer to bind to the real site. Figure 4 illustrates how the compounds typically bind by showing **11** in its optimal mode. Notice that **11** could equally well be rotated and translated about within the slab, and that the C-O bonds have considerable freedom of rotation while still preserving the predicted mode. Even more extreme is the predicted binding mode of **1**: all atoms lie in the infinite r_3, in any translation, any orientation, and any conformation. That such a simple representation of the site accounts for so much binding data, tells us that more data on more varied compounds are needed to construct anything like a reasonable picture of the site.

CONCLUSIONS

Voronoi site models are a qualitatively different approach to accounting for given binding data. The accompanying computer algorithms so far enable one to handle relatively simple, but nontrivial, data sets. Solutions can be found which agree *completely* with experiment, leaving no outliers. The solution in this test case has considerable predictive power, although it is geometrically so sketchy that we can conclude very little about the real site's structure from these data alone.

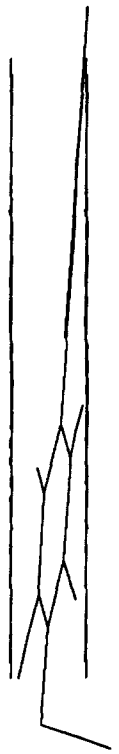


Figure 4a. Compound **11** in its predicted optimal binding mode. **(a)** In this view, the boundary planes between the regions are viewed almost exactly edge on. Unfortunately, this means that both phenyl rings are also viewed from the edge. To the left is region 1; the center slab is region 2; and to the right is region 3. The 4′-OH group extends at the top into r_3 while leaving the 3′,5′-Cl_2 atoms in their preferred r_2. At the bottom, the 4-OH remains in r_2, which is better than being in r_1, but moving to the preferred r_3 would force Cl atoms into the energetically repulsive r_3 also. Still the H of the 4-OH can reach the preferred r_3.

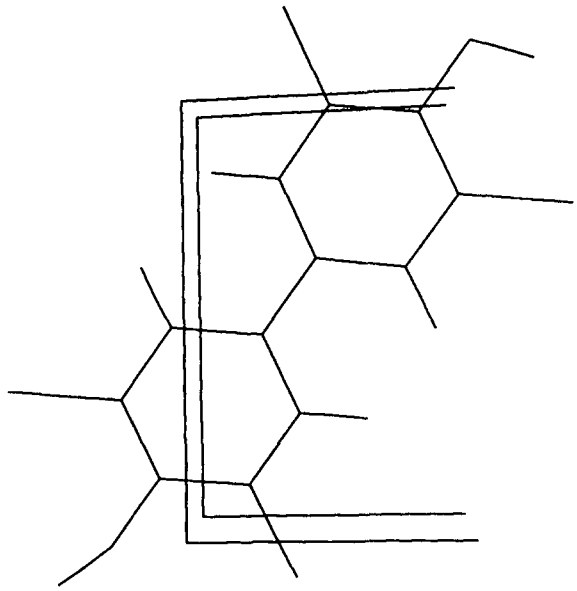


Figure 4b. A perspective view of the site, where the two boundary planes are indicated by rectangles. Here the molecule lies nearly in the plane of the paper, which corresponds to the plane of the thin slab between the boundary planes.

The importance of this study is not really in what we have learned about prealbumin, but rather in the concepts illustrated.

The first idea is that if the data can be fit with such incredibly vague models as two or three broad regions, then the limited binding studies really are not telling us much about the structure of the site. The bias of the method toward simplicity is a welcome counterbalance to our natural tendency of deducing too much detail. In fact, it is the only honest conclusion we can make from these data.

The second idea is that requiring $\Delta G_- \leq \Delta G_{calc} \leq \Delta G_+$ for all molecules amounts to an unambiguous test of whether the model fits the data. We do not have to resort to statistical criteria, such as standard deviation, correlation coefficient, explained variance, t-test, etc. It is not really a statistical question, anyway.

The third concept that emerges is that of conformation space of a molecule with respect to a model. Usually the most time-consuming step in any QSAR method that takes into account the conformational freedom of drug molecules is exploring all the conformational possibilities. When several rotatable bonds are involved, there is always the concern that the exploration was not thorough enough. In the trivial case of one region, conformation can be completely disregarded. For our three-region example, only fairly gross structural features were of interest. In general, the conformation space of a molecule with respect to a given site can be completely expressed as a list of geometrically allowed binding modes. All other detail is superfluous.

Fourthly, we can parameterize molecular comparisons in a similar fashion. Many QSAR studies hinge upon superimposing sometimes structurally different molecules upon one another, but there are always nagging questions: *which atoms of the one molecule are to be brought into coincidence with which atoms of the other? How close is "coincidence"? What about the left-over pieces of the larger molecule?* If instead one looks at comparisons with respect to a model, the ambiguities largely vanish. Two molecules, m and m', match with respect to a given model if there are $\beta \in B_m$ and $\beta' \in B_{m'}$ such that $\beta = \beta'$. Since the βs are always vectors in the same interaction energy space, they can be directly compared, even for very dissimilar molecules. Note that the site geometry determines the dimensionality of the energy space and the contents of the mode sets, but finding a solution set of interaction energies is not required.

Clearly the implementation of this approach can be improved on a number of fronts, so that larger, more flexible molecules can be treated in more elaborate site geometries. In its present stage of development, however, it is already remarkably objective, requiring only a suggested site structure. No longer do we have to decide in advance which portions of the ligand molecules are important and how these important parts superimpose or bind to the site.

This work was supported by grants from the National Institutes of Health (GM30561) and the National Science Foundation (PCM-8314998). Part of this work was carried out while I was the recipient of a Fulbright Senior Scholar Fellowship administered by the Australian-American Educational Foundation. I am very grateful for the support of my host institutions, the School of Pharmaceutical Chemistry, Victorian College of Pharmacy, Ltd. and the La Trobe University Chemistry Department, both in Melbourne, Australia.

APPENDIX

Linear Representation of Molecules

Suppose a molecule is described by the atomic Cartesian coordinates for a particular conformation, a list of which pairs of atoms are bonded, and a list of which bonds are rotatable. Suppose further, that one bond has been deleted from each (rigid) ring, so that the molecule is viewed as formally acyclic but fully connected in a tree graph. All rings are assumed to be rigid groups, i.e., pseudorotations are not treated here. Then the following algorithm produces a linearized representation of the molecule, as in Figure 1 and Table 1.

Choose the first atom involved in the greatest number of rotatable bonds.
Rearrange the tree representation of the molecule so this atom is the root.
Express root atom coordinates = $\mathbf{p}_0 \equiv \mathbf{w}$
And conversely define $\mathbf{w}$ by $\mathbf{w} \equiv \mathbf{p}_0$.
Recursively for each son i of the root of the current subtree o
  This atom $\mathbf{p}_i$, has as its "critical bond" the last rotatable bond in the chain from the

tree root to it, unless it is an endpoint of the bond.
If only **w** has so far been determined,
Define a new unit vector

$$\mathbf{u}_1 = \frac{\mathbf{p}_i - \mathbf{p}_o}{\| \mathbf{p}_i - \mathbf{p}_o \|}$$

Express $\mathbf{p}_i = \mathbf{p}_o + \| \mathbf{p}_i - \mathbf{p}_o \| \mathbf{u}_1$
Else if currently there are **w** and $\mathbf{u}_1$
If $\mathbf{p}_i$ is on the line $\mathbf{u}_1$ from **w**
Express $\mathbf{p}_i$ as above
Else define $\mathbf{u}_2$ as the unit vector from $\mathbf{p}_o$ to $\mathbf{p}_i$ and express $\mathbf{p}_i$ in terms of $\mathbf{p}_o$ and $\mathbf{u}_2$.
Else if we have **w**, $\mathbf{u}_1$, and $\mathbf{u}_2$.
Let $\mathbf{v} = \mathbf{p}_i - \mathbf{p}_o$

$$\text{Let } \beta = \frac{\mathbf{v} \cdot \mathbf{u}_2 - (\mathbf{u}_1 \cdot \mathbf{u}_2)(\mathbf{v} \cdot \mathbf{u}_1)}{1 - (\mathbf{u}_1 \cdot \mathbf{u}_2)^2}$$

Let $\alpha = \mathbf{v} \cdot \mathbf{u}_1 - \beta \mathbf{u}_1 \cdot \mathbf{u}_2$
Let $\mathbf{u}_3 = \mathbf{v} - \alpha \mathbf{u}_1 - \beta \mathbf{u}_2$
If u_3^2 is small, $\mathbf{p}_i$ is coplanar
Express $\mathbf{p}_i = \mathbf{p}_o + \alpha \mathbf{u}_1 + \beta \mathbf{u}_2$
Else define $\mathbf{u}_3$ as the unit vector from $\mathbf{p}_o$ to $\mathbf{p}_i$ and express $\mathbf{p}_i$ in terms of $\mathbf{p}_o$ and $\mathbf{u}_3$.
Else if we have **w**, $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$
Let $\mathbf{v} = \mathbf{p}_i - \mathbf{p}_o$

$$\text{Let } r = \frac{\mathbf{v} \cdot \mathbf{u}_2 - (\mathbf{u}_2 \cdot \mathbf{u}_3)(\mathbf{v} \cdot \mathbf{u}_3)}{1 - (\mathbf{u}_2 \cdot \mathbf{u}_3)^2}$$

$$\text{Let } s = \frac{(\mathbf{u}_2 \cdot \mathbf{u}_3)(\mathbf{u}_1 \cdot \mathbf{u}_3) - \mathbf{u}_1 \cdot \mathbf{u}_2}{1 - (\mathbf{u}_2 \cdot \mathbf{u}_3)^2}$$

Let

$$\alpha = \frac{\mathbf{v} \cdot (\mathbf{u}_1 - \mathbf{u}_3) + r(\mathbf{u}_2 \cdot \mathbf{u}_3 \cdot \mathbf{u}_1 \cdot \mathbf{u}_2)}{1 - \mathbf{u}_1 \cdot \mathbf{u}_3 + s(\mathbf{u}_1 \cdot \mathbf{u}_2 - \mathbf{u}_2 \cdot \mathbf{u}_3)}$$

Let $\beta = r + \alpha s$
Let $\gamma = \mathbf{v} \cdot \mathbf{u}_3 - \alpha \mathbf{u}_1 \cdot \mathbf{u}_3 - \beta \mathbf{u}_2 \cdot \mathbf{u}_3$
Express $\mathbf{p}_i = \mathbf{p}_o + \alpha \mathbf{u}_1 + \beta \mathbf{u}_2 + \gamma \mathbf{u}_3$
Having treated atom i, do the same for any sons it may have in the tree, where if a rotatable bond had just been crossed to reach i,
then the sons will have only **w** in their vector sets.

Voronoi Polyhedra Properties

The boundaries between regions are the perpendicular bisecting lines (planes in three dimensions) of the lines joining the generating points. For instance in Figure 2, $\mathbf{V}_6$—$\mathbf{V}_7$ bisects $\mathbf{c}_4$—$\mathbf{c}_5$. It is convenient to always deal with finite regions, so we add a large equilateral triangle (tetrahedron in three dimensions) of "outrigger" points, $\mathbf{c}_6$, $\mathbf{c}_7$, $\mathbf{c}_8$. As long as the chosen generator points lie within this triangle, they will produce finite regions. There will always be one region for every generator, but the size, shape, and number of edges (faces) depends on their relative positions. A region r_i can conveniently be described in terms of its set of vertices $\{\mathbf{V}\}_i$, where the three edges (four faces) intersect. The vertex sets for adjacent regions of course have some members in common. For instance, $\{\mathbf{V}\}_2 = \{\mathbf{V}_2, \mathbf{V}_3, \mathbf{V}_4, \mathbf{V}_5, \mathbf{V}_6\}$ and $\{\mathbf{V}\}_4 = \{\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_6, \mathbf{V}_7\}$. Any point **p** in a region r_i can be expressed as a convex combination of its vertices:

$$\mathbf{p} = \sum_{\mathbf{V}_j \in \{\mathbf{V}\}_i} \alpha_j \mathbf{V}_j$$

and

$$\sum_j \alpha_j = 1$$

A vertex is equidistant to three (four) generating points while not being closer to any other generator. Due to the interest in Voronoi polyhedra for applications in geography, there are a number of rapid algorithms for locating all vertices of a large set of regions, but these are restricted to two dimensions.[12,13,14] For our purposes, dealing with relatively few regions, the following algorithm suffices to determine the vertex set of each region.

For some large number $B > 0$, place outrigger points about the origin at

$$\left(\frac{-B}{2}, \frac{-B}{2\sqrt{3}}, \frac{-B}{2\sqrt{6}}\right), \quad \left(\frac{B}{2}, \frac{-B}{2\sqrt{3}}, \frac{-B}{2\sqrt{6}}\right),$$

$$\left(0, \frac{B}{\sqrt{3}}, \frac{-B}{2\sqrt{6}}\right), \text{ and } \left(0, 0, \frac{B\sqrt{3}}{2\sqrt{2}}\right).$$

For all quartets of generators $\mathbf{c}_i$, $\mathbf{c}_j$, $\mathbf{c}_k$, and $\mathbf{c}_l$ (including outriggers),
Solve the linear system

$$\begin{pmatrix} 2(\mathbf{c}_j - \mathbf{c}_i) \\ 2(\mathbf{c}_k - \mathbf{c}_i) \\ 2(\mathbf{c}_l - \mathbf{c}_i) \end{pmatrix} \mathbf{V} = \begin{pmatrix} \mathbf{c}_j^2 - \mathbf{c}_i^2 \\ \mathbf{c}_k^2 - \mathbf{c}_i^2 \\ \mathbf{c}_l^2 - \mathbf{c}_i^2 \end{pmatrix}$$

where the rows of the matrix have been indicated.
If the equation cannot be solved
The four generators are coplanar,
so simply proceed to the next quartet.
If **V** is no closer to any generator than it is to $\mathbf{c}_i$,

It is a vertex and belongs in $\{\mathbf{V}\}_i$, $\{\mathbf{V}\}_j$, $\{\mathbf{V}\}_k$, and $\{\mathbf{V}\}_l$.

Search For Geometrically Allowed Binding Modes

Given the coordinates of the site generating points, $\mathbf{c}_i, i = 1, \ldots, n_s$, and a ligand molecule expressed in linearized form, we need to find all geometrically allowed binding modes. In order to avoid solving systems of quadratic equations, we make the approximation that each unit vector can be oriented toward any of the six faces of a cube. (Choosing polyhedra with more faces gives a more accurate approximation at the expense of more combinations to examine). If two unit vectors point to the same face, their dot products are in the range 0.65 to 1.0; for adjacent faces we take the range to be −0.9 to 0.9; and for opposite faces it is −1.0 to −0.65. Then the following recursive algorithm examines all possibilities of placing the atoms in different regions and pointing the unit vectors toward different faces, eliminating geometrically impossible choices as early in the search as possible.

Sort the atoms and unit vectors so that the position of the first atom depends only on **w**, the next atoms need only **w** and $\mathbf{u}_1$, etc.
Starting with the first atom in the first region, recursively place an atom in a region
  For all the atoms so far placed and the unit vectors so far involved
    Solve set of equations (6), where **w** is variable, but the unit vector components are given their corresponding face values with some additional freedom to point anywhere on the face.
    This is a standard linear programming problem.
    If a solution is not found,
      Try placing the atom in the next region.
    Else atom placement was successful
      If that was the last atom
        A new allowed mode has been found.
      Else if the next atom needs no new unit vectors,
        Try placing it, initially in the first region
      Else recursively set the next required unit vector
        For each face
          The formal dot product range with previously placed vectors must overlap the conformationally observed range
          If so,
            If the next atom needs no extra unit vectors
              Place that atom
            Else set the next unit vector.

Determining Interaction Energies

Although we generally think of the interaction energies as a table, such as Table IV, with a row for each atom type and a column for each region, we could write all these entries as a linear vector, **e**. Then for any **e**, we define an error function

$$F(\mathbf{e}) = \max\begin{cases} 0 \\ \max\limits_{m} \min\limits_{\beta \in B_m} (\Delta G_{m-} - \beta \cdot \mathbf{e}) \\ \max\limits_{m} \max\limits_{\beta \in B_m} (\beta \cdot \mathbf{e} - \Delta G_{m+}) \end{cases}$$

In other words, $F(\mathbf{e})$ is either zero if **e** is a solution, or the largest upper bound violation, or the least lower bound violation for the molecule which is worst off in that respect. Unfortunately it is possible to have $F > 0$ for all **e** (i.e., no solution is possible), or to have local minima where $F > 0$ while having $F = 0$ for some other **e**. Hence, the algorithm for finding at least one solution is:

For all molecules m and their sets of modes B_m
  Calculate the corresponding energy space mode vectors, β, and inequalities, $\beta \cdot \mathbf{e} < \Delta G_{m+}$ and $\beta \cdot \mathbf{e} > \Delta G_{m-}$.
  Remove duplicate β', where $\beta, \beta', \in B_m$ and $\beta = \beta'$.
For all pairs of molecules, m and m', and for respective pairs of binding modes, $\beta \in B_m$ and $\beta' \in B_{m'}$
  If $\beta = \beta'$
    If $\Delta G_{m'+} > \Delta G_{m+}$ and $\Delta G_{m'-} > \Delta G_{m-}$
      Remove β' because it can never be an optimal mode.
    If $\Delta G_{m'+} > \Delta G_{m'+}$ but $\Delta G_{m'-} \leq \Delta G_{m-}$
      $\beta' \cdot \mathbf{e} < \Delta G_{m'+}$ is a redundant upper bound, but β' might be the optimal mode for m'.
Repeatedly try
  Taking a random starting $\mathbf{e}^{(0)}$ with each component in the range $[-R, +R]$, where $R = \max_m \Delta G_{m+}$.

Locate a solution, $\mathbf{e}^*$, by subgradient optimization.[15]
Iterate for $k = 0, 1, 2, \ldots$

$$\mathbf{e}^{(k+1)} = \mathbf{e}^{(k)} + t\,\nabla F(\mathbf{e}^{(k)})$$

$$t = \lambda \frac{F(\mathbf{e}(k)) - F(\mathbf{e}^*)}{\| \nabla F(\mathbf{e}^{(k)}) \|^2}$$

Where $F(\mathbf{e}^*) = 0$, and we let $\lambda = 0.5$.
Until $k = 2000$ or $F(\mathbf{e}^{(k)}) < 10^{-5}$.
Until an attempt converged or too many attempts have been made.

If this algorithm succeeds in finding a solution, then one can go back and see which modes are indeed optimal. It is possible that more than one will have the same maximal value for some molecule, particularly (but not necessarily) if it corresponds to a duplicated mode. In our simple example of Figure 3, if A — A could rotate in the plane, there would be modes $\mathbf{b} = (1,2)$ and $\mathbf{b}' = (2,1)$, but $\beta = \beta' = (1,1)$. There may also be other very different solutions that could be found from different random starts. On the other hand, failure to find a solution after many random starts and lengthy iterations indicate there is no solution (and a different site geometry must be tried), but it is not proof. With this algorithm, the only case of proven failure is when all modes for one of the molecules have been eliminated, according to the eighth line of the above algorithm. Unfortunately other causes of failure manifest themselves only in lack of convergence of the subgradient optimization process.

References

1. Y. C. Martin, *Quantitative Drug Design*, Medicinal Research Series, 8, Marcel Dekker, Inc. New York, 1978.
2. R. Franke, *Theoretical Drug Design Methods*, Akademie Verlag, Berlin, DDR, 1984.
3. A. K. Ghose and G. M. Crippen, *J. Comp. Chem.*, **6,** 350–359 (1985).
4. A. K. Ghose and G. M. Crippen, *J. Comp. Chem.*, **7,** 565–577 (1986).
5. A. K. Ghose and G. M. Crippen, *J. Med. Chem.*, **28,** 333–346 (1985).
6. G. M. Crippen, *Ann. New York Acad. Sci.*, **439,** 1–11 (1984).
7. Those interested in the programs at this early stage of development should contact the author.
8. P. J. Flory, *Statistical Mechanics of Chain Molecules*, Wiley Interscience, New York, 1969.
9. G. F. Voronoi, *Reine Angew. Math.*, **134,** 198 (1908).
10. P. F. Ash and E. D. Bolker, *Geometriae Dedicata*, **19,** 175–206 (1985).
11. U. Rickenbacher, J. D. McKinney, S. J. Oatley, and C. C. F. Blake, *J. Med. Chem.*, **29,** 641–648 (1986).
12. P. J. Green and R. Sibson, *Comput. J.*, **21,** 168–173 (1978).
13. M. Iri, K. Murota, and T. Ohya, *A Fast Voronoi-Diagram Algorithm with Applications to Geographic Optimization*, Lecture Notes in Control and Information Sciences, 59, Springer Verlag, 1984.
14. T. Ohya, M. Iri, and K. Murota, *Inf. Proc. Lett.*, **18,** 227–231 (1984).
15. C. Sandi, in *Combinatorial Optimization*, ed. C. Sandi, pp. 73–91, Wiley, New York, 1979.