Voronoi Binding Site Models: Calculation of Binding Modes and Influence of Drug Binding Data Accuracy

Laurent G. Boulu and Gordon M. Crippen

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

Received 22 September; accepted 2 December 1988

A new and accurate method for calculating the geometrically allowed modes of binding of a ligand molecule to a Voronoi site model is reported. It is shown that the feasibility of the binding of a group of atoms to a Voronoi site reduces to a simple set of linear and quadratic inequalities and quadratic equalities which can be solved by minimization of a simple function. Newton's numerical method of solution coupled to a line search proved to be successful. Moreover, we have developed efficient molecular and site data bases to discard quickly infeasible binding modes without time-consuming numerical calculation. The method is tested with a data set consisting of the binding constants for a series of biphenyls binding to prealbumin. After determination of the conformation space of the molecules and proposal of a Voronoi site geometry, the geometrically feasible modes are calculated and the energy interaction parameters determined to fit the observed binding energies to the site within experimental error ranges. We actually allowed these ranges to vary in order to study the influence of their broadness on the site geometry and found that as they increase, one can first model the receptor as a three-region site then as a single region site, but never as a two-region site.

INTRODUCTION

In order to understand the specific binding of small molecules to biological receptors, we have recently devised a novel approach to deduce objectively the structure and energetics of a binding site, given the observed binding energies for a series of compounds.[1,2] The method is based on modeling the site as Voronoi polyhedra,[3] is limited to nonreactive binding, and its main features have already been described.[1]

One critical step in fitting binding data with a Voronoi site model is the determination of all geometrically allowed binding modes of the molecules, and previously this was done approximately by using linear programming. Here, we report a new and mathematically more accurate way of calculating these binding modes as well as the use of molecular and site data bases to eliminate impossible binding modes without explicit calculation. The formulation of the method is given in the next section and the application to a simple example data set is considered in the third section.

FORMULATION

Calculation of the Binding Modes

The geometry of a Voronoi site model made of n_s regions is defined by the x, y, z coordinates of "generating points" $\mathbf{c}_i$, $i = 1, \ldots, n_s$ supplied by the investigator. Each one of these determines a surrounding region r_i, called a Voronoi polyhedron, defined as the set of all points $\mathbf{p}$ closer to its generating point than to the other generating points:

$$r_i = \{\mathbf{p} \mid \|\mathbf{c}_i - \mathbf{p}\| < \|\mathbf{c}_j - \mathbf{p}\|, \forall j \neq i\} \tag{1}$$

These regions happen to be convex,[4] space-filling, and separated from each other by planar surfaces. Thus every atom of a ligand molecule lies in one and only one of these regions, and one can express the orientation and internal conformation of a particular binding mode by stating in which region each atom is found. More specifically, a binding $\mathbf{b}$ for a ligand of n atoms can be described by a vector $(b_1, b_2, \ldots, b_n)$ where b_i, $i = 1, \ldots, n$, specifies the Voronoi region in

Journal of Computational Chemistry, Vol. 10, No. 5, 673–682 (1989)

CCC 0192-8651/89/050673-10$04.00

which atom i is found. An atom of the ligand can experience energetically distinct interactions with the site depending on the region it lies in. Letting $\mathbf{p}_i$ denote the position of atom a_i, the condition that a_i lies in region r_k rather than an adjacent region r_j is given by

$$\|\mathbf{p}_i - \mathbf{c}_k\|^2 < \|\mathbf{p}_i - \mathbf{c}_j\|^2 \tag{2}$$

Now, a new method to summarize the set of allowed conformations of a drug molecule or conformation space has been previously devised,[1] and greatly simplifies the solving of eq. (2). It is called the linear representation of the molecule because each atom's position is expressed as a linear combination of an overall molecular translation vector $\mathbf{w}$ and several unit vectors $\mathbf{u}_i$, $i = 1, \ldots, n_u$ chosen in such a way as to represent the orientation of the whole molecule and the relative orientation of groups of atoms linked by rotatable bonds:

$$\mathbf{p}_i = \mathbf{w} + \sum_{l=1}^{l=n_u} a_l \mathbf{u}_l \tag{3}$$

The algorithm in ref. (1) has been improved to minimize the number of unit vectors used to describe the molecule (in the case of 3,5-dichloro-4-hydroxy biphenyl illustrated herein, this number is 4). Then a single exhaustive sampling of all energetically allowed conformations can be summarized in terms of the greatest and least observed scalar products $\mathbf{u}_l \cdot \mathbf{u}_m$ between the various pairs of unit vectors. Equations (2) and (3) give

$$2\left(\mathbf{w} + \sum_{l=1}^{l=n_u} a_l \mathbf{u}_l\right) \cdot (\mathbf{c}_j - \mathbf{c}_k) < \mathbf{c}_j^2 - \mathbf{c}_k^2 \tag{4}$$

Equation (4) is seen to be linear in the Cartesian components of the unit vectors so that the geometric feasibility of a binding mode reduces to a simple set of linear and quadratic inequalities and quadratic equalities. The linear inequalities of the form of eq. (4) involve all adjacent regions for each atom, while the quadratic conditions arise from the normalization of the unit vectors and from restrictions on the dot products between pairs of unit vectors:

$$\|\mathbf{u}_l\|^2 = 1 \qquad l = 1, 2, \ldots, n_u \tag{5}$$

$$\alpha_{lm} < \mathbf{u}_l \cdot \mathbf{u}_m < \beta_{lm} \qquad l, m = 1, 2, \ldots, n_u \tag{6}$$

Equations (4)–(6) can be rewritten as

$$\begin{aligned} \mathbf{f}_k(\mathbf{x}) &= 0 \quad k = 1, 2, \ldots, n_e \\ \mathbf{f}_k(\mathbf{x}) &< 0 \quad k = n_e + 1, \ldots, n_e + n_i \end{aligned} \tag{7}$$

where $\mathbf{x}$ is the vector of unknown unit vectors components, and n_e and n_i are the respective numbers of equalities and inequalities. Note that the n_i inequalities can be transformed into equalities by the use of the slack variables h_k:

$$g_k(\mathbf{y}) = f_k(\mathbf{x}) + h_k^2$$

so that system (7) is equivalent to $\mathbf{g}(\mathbf{y}) = 0$ and can be solved by minimization of $\Sigma_k g_k^2(\mathbf{y})$. Although this approach was successful using Newton's method, the large number of slack variables needed as the number of atoms increases makes it computationally prohibitive. For example, even fitting a molecule of benzene into a two-region site requires 12 slack variables due to eq. (4). Instead, the following vector of functions $\mathbf{e}$ was defined:

$$\begin{aligned} k > n_e: \quad & e_k(\mathbf{x}) = f_k(\mathbf{x}) \quad \text{if } f_k(\mathbf{x}) \geq 0 \\ & e_k(\mathbf{x}) = 0 \quad \text{if } f_k(\mathbf{x}) < 0 \\ k \leq n_e: \quad & e_k(\mathbf{x}) = f_k(\mathbf{x}) \end{aligned} \tag{8}$$

By eqs. (8), (7) is equivalent to

$$\mathbf{e}(\mathbf{x}) = \mathbf{0} \tag{9}$$

and can be solved by minimization of ϕ defined as

$$\phi(\mathbf{x}) = \frac{1}{2} \sum_k e_k^2(\mathbf{x}) \tag{10}$$

so that $\phi(\mathbf{x}^*) = 0$ at the solution $\mathbf{x}^*$. To do so, a modification of Newton's method[5] happened to be quite satisfactory since the second derivatives of ϕ are second order in the components of the vector $\mathbf{x}$ and are easily handled analytically. We tried other methods appropriate for the least squares problem such as the Gauss–Newton and Levenberg–Marquardt methods and found that they converge slower than the Newton method. To ensure a reliable minimization algorithm, we have used a modification of Newton's method where the pure form of Newton's method is coupled to a line search in order to get a descent method. The modified Newton algorithm has the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \gamma_k \mathbf{d}_k \tag{11}$$

where $\mathbf{d}_k$ is the Newton direction obtained as the solution of the linear system of equations

$$\nabla^2\phi(\mathbf{x}_k)\mathbf{d}_k = -\nabla\phi(\mathbf{x}_k) \tag{12}$$

and γ_k is a positive stepsize parameter. $\nabla^2\phi$ and $\nabla\phi$ are respectively the Hessian and the gradient of ϕ. The direction $\mathbf{d}_k$ can be found by Gaussian elimination of (12) and must be a descent direction, i.e., $\nabla\phi(\mathbf{x}_k) \cdot \mathbf{d}_k < 0$, in order for ϕ to decrease along the direction $\mathbf{d}_k$ (if not, one considers $-\mathbf{d}_k$). The stepsize γ_k is chosen by the Armijo rule[5]:

$$\phi(\mathbf{x}_k) - \phi(\mathbf{x}_k + \gamma_k\mathbf{d}_k) \geq -\sigma\gamma_k \nabla\phi(\mathbf{x}_k) \cdot \mathbf{d}_k \tag{13}$$

with $\sigma \in [0,1]$, $\gamma_k = (1/2)^q$ and q is the first nonnegative integer for which (13) holds. This rule ensures convergence by sufficiently decreasing ϕ along $\mathbf{d}_k$.

To test whether the binding of a ligand molecule is feasible, one does not solve the system (9) directly for all its atoms, but one tries to place the first atom, then, if this is successful, the second and so on. The algorithm used for searching the geometrically allowed binding modes is a depth-first recursive tree traversal. More explicily, starting with the first atom in the first region, one recursively places an atom in a region. The nth generation in the tree corresponds to the placement of the nth atom. As one moves down the tree from the first atom along a path corresponding to a particular binding mode, one tries to place consecutively all atoms by solving (9) at each node along the way. At each node, one has $n - 1$ "old atoms" having coordinates and unit vectors determined from the solution at the previous node, plus one "new atom." One then attempts to solve (9) for these n atoms. If the binding of a subset of n atoms is not feasible, one stops moving downward along the path and moves back upward to the parent node. In this way, a number of binding modes is discarded. On the other hand, if the binding of this subset of n atoms is successful, the solution of (9) is used as an initial guess for the one corresponding to the binding of $n + 1$ atoms. When the binding of a set of atoms is feasible, the method converges quickly to the solution (e.g., placing an atom of a molecule of benzene into a two-region site requires an average of 1.3 Newton iterations). Otherwise, the algorithm always converges to a local minimum with $\phi > 0$, in which case a new starting point is obtained by randomly generating the new unit vectors from the surface of the unit sphere. Since there is no way to know whether a global minimum has been missed, we require 10 trials as inductive proof that the proposed binding is infeasible (for the above molecule, this corresponds to about 55 iterations/atom).

Molecular and Site Data Bases

Molecular Topological Data Base

The very fact that most of the computation is spent on proving that a mode (or part of a mode) is infeasible led us to use a molecular topological data base in order to process rapidly those modes which can be eliminated on combinatorial knowledge of the molecule alone, thus avoiding any time-consuming numerical calculation (Figs. 1, 2). This combinatorial data base is made of convexity rules involving groups of atoms and takes advantage of the fact that the Voronoi regions are convex. When using a rule, it must involve the new atom so as to use this rule only once during the checking of the whole mode. Three kinds of rules have been considered depending on the dimensionality generated by these subsets of atoms. The basic principle underlying all of them is that if two atoms a_i and a_j lie in the same region, the segment $a_i a_j$ is also included in this region. Before stating the first rule, we define a convex hull as the surface (perimeter in two dimensions) of a convex polyhedron the vertices of which are given by a set of atoms.
The first rule can be stated as:

Let the set A defined by atoms $\{a_1, a_2 \ldots\}$ be a convex hull for all allowed conformations, and $\{b_1, b_2, \ldots\} = \mathrm{B}$ are in its interior. Then if all the atoms of A are in region r_1, all atoms of B must also be in r_1. As an example, Figure 1 shows a convex hull made of two atoms ($\{\mathrm{H}_1, \mathrm{H}_4\} = \mathrm{A}$), with two other atoms ($\{\mathrm{C}_1, \mathrm{C}_4\} = \mathrm{B}$) lying in it. Two cases may be considered depending on whether the "new" atom (i.e., the atom one tries to place) belongs to A or B.

Case 1: suppose the "new" atom is $\mathrm{H}_1 \in \mathrm{A}$. If one tries to place it in r_1 and H_4 has already been successfully placed in r_1, the above rule can be applied, i.e., the convex hull A lies entirely in r_1 and one should ask whether C_1 or C_4 has already placed, i.e., is

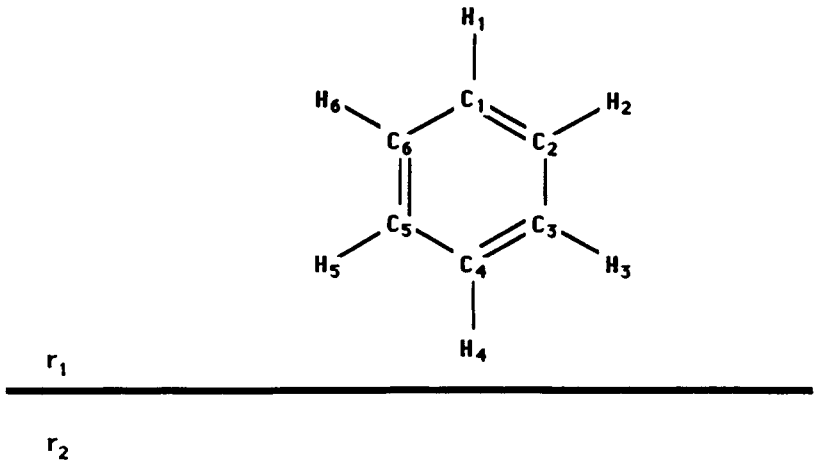


Figure 1. Illustration of rule I. If atoms H_1 and H_4 are both in region r_1, then atoms C_1 and C_4 must be also, because they lie in the interior of the convex hull defined by H_1 and H_4, and any Voronoi region is convex.

"old." If one of these, say C_1, is "old" and lies in r_2, the rule is violated and the mode is infeasible. If however C_1 lies in r_1, the rule is not violated but no useful deduction can be made as the feasiblity of the mode.

Case 2: suppose the "new" atom to be placed is $C_1 \in B$. For the rule to be used, the convex hull A should already be lying in one single region, say r_1, and it follows that C_1 must necessarily lie in r_1. In other words, if one decides to place the "new" atom C_1 in r_2, the rule is violated and the mode is infeasible. However, if one tries to place C_1 in r_1, not only is the rule not violated, but useful deduction has been made, namely the mode is feasible.

Other convex hulls made of a larger number of atoms can be considered. For example, the three atoms bound to any trigonal carbon form a hull containing the carbon. Hulls made of too large a number of atoms are less interesting since the use of the corresponding rule is less frequent.

The second rule, illustrated in Figure 2, can be stated as:

if a_1, a_2 are in region r_1, and b_1, b_2 are in region r_2, and segments a_1a_2 and b_1b_2 have a nonempty intersection for all allowed conformations, then the binding is infeasible.

The subsets of atoms to which this rule applies consist of all coplanar quartets of atoms in the molecule.

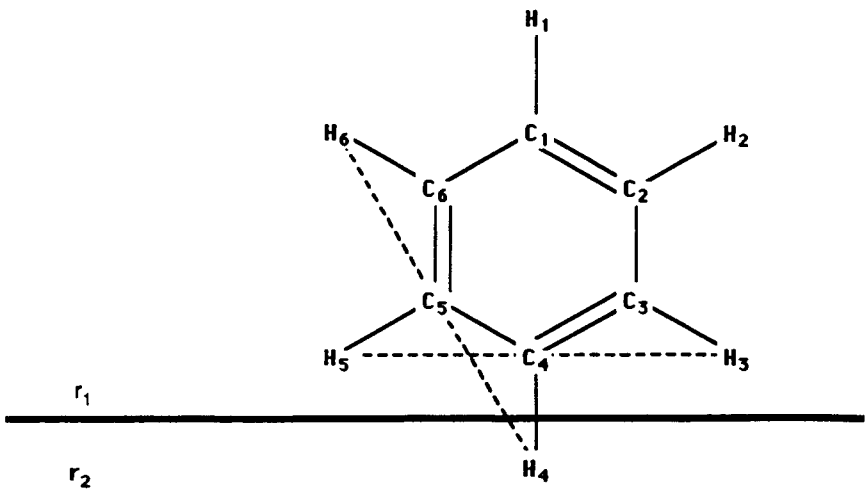


Figure 2. Illustration of rule II. If atoms H_3 and H_5 are in region r_1, and atom H_4 is in region r_2, then H_6 can not be in r_2 since the segments H_3H_5 and H_4H_6 have a nonempty intersection, and any Voronoi region is convex.

Finally, the third rule is a higher dimensional form of the second one:

if a_1, a_2, a_3 are in region r_1, and b_1, b_2 are in region r_2, and the triangle $a_1a_2a_3$ and the segment b_1b_2 have a nonempty intersection for all allowed conformations, then the binding is infeasible.

This rule applies for nonplanar molecules. Note that a one-dimensional analog of the second rule is:

if a_1 is in region r_1, and b_1 is in r_2, and $b_1-a_1-b_2$ is the ordering along a line, then b_2 cannot be in r_2.

However, it can be seen that this rule is a particular case of the first one. As an example, rules of the first and second kinds applied to the fitting of benzene into a two-region site led to the elimination of all infeasible bindings, i.e., numerical solving of (9) was attempted only when there was a solution, which reduced the computational time to a few percent of its original value when no rule was applied. When fitting benzene to a site made of three parallel regions (Fig. 3), about 3/4 of the infeasible bindings could be eliminated and the CPU time was cut by 2/3. Not all infeasible bindings can be eliminated by the use of a topological data base for a site made of a number of regions > 2 since more quantitative questions on molecular size arise and must be a priori solved numerically. In general, the most powerful rules are the ones of the first kind involving a small number of atoms like the ones shown in Figure 1.

Distances Data Base

Actually, some quantitative modes can also be discarded by using a relatively simple "distances" data base involving interatomic distances as well as the size of the regions of the investigated Voronoi site model. Let $d_{\min ij}$ and $d_{\max ij}$ be respectively the minimum and maximum distances between atoms a_i and a_j allowed in the conformation space of the molecule. Let r_{kk} denote the maximum diameter of a finite region r_k, i.e., the maximum distance between any two points of r_k. Let $r_{\min kl}$ and $r_{\max kl}$ be the minimum and maximum distance between regions r_k and r_l, i.e., respectively the minimum and maximum distance between any point of r_k and any point of r_l. Then each of the two following conditions is sufficient for the proposed binding to be infeasible:

a_i and a_j lie in two nonneighbor regions k and l and $d_{\max ij} < r_{\min kl}$ (the nonadjency is required to make the rule nontrivial).
a_i and a_j lie in two bounded regions k and l and $d_{\min ij} > r_{\max kl}$ (note that if $k = l$, $r_{\max kl}$ is r_{kk}).

The efficiency of these rules obviously depends on the relative size of the molecule and the regions considered. For example, in the case of a site made of three infinite parallel regions (Fig. 3), the only finite and nonzero quantity is $r_{\min 13}$ so that only the first rule above applies. Still, this rule becomes useless if $r_{\min 13}$ is smaller than any molecular bond. However, for benzene, if $r_{\min 13}$ is only 5/4 times the C—C bond length, it allows one to discard as much as 40% of the infeasible bindings. If used with the rules from the molecular topological data base, this percentage increases to 90%. Other distance checks involving more than two atoms can be considered, but they are more complicated and require some computation. A relatively simple check involving three atoms is given in the Appendix.

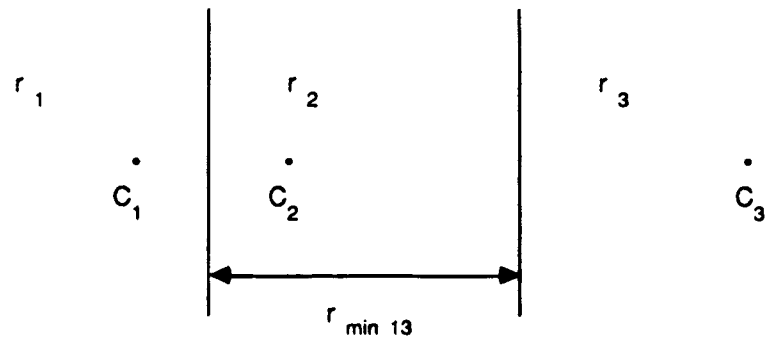


Figure 3. Three-region Voronoi site made of three infinite parallel regions r_1, r_2, and r_3 respectively generated by the points $\mathbf{c}_1$, $\mathbf{c}_2$, and $\mathbf{c}_3$. The minimum distance between r_1 and r_3 is $r_{\min 13}$ (see text).

APPLICATION TO A SIMPLE DATA SET

As a test of the new method, the data set used in reference 1 was considered. It consists of twelve biphenyl derivatives binding to prealbumin.[6] In fitting these data to a Voronoi site, steps other than the calculation of the binding modes were implemented as in reference 1. The molecules were first linearized, which allowed us to summarize their conformation space (step I), and a site geometry was proposed (step II). After calculation of the binding modes by the method described in the previous section (step III), the interactions parameters were determined so that the calculated binding energy $\Delta G_{m,\text{calc}}$ for each ligand molecule m falls within its respective experimental range (step IV). In other words, we require an absolute fit to the given ranges:

$$\Delta G_{m-} \leq \Delta G_{m,\text{calc}} \leq \Delta G_{m+} \quad \text{for all } m$$

where ΔG_{m-} and ΔG_{m+} are the bounds of the experimental range for molecule m. Note that for all the feasible modes, the molecule m is said to have a calculated binding energy $\Delta G_{m,\text{calc}}$ corresponding to that of the energetically most favorable mode, i.e.,

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

where B_m is the set of geometrically feasible binding modes for molecule m, and $\Delta G(\mathbf{b})$ is the total interaction energy for the mode **b** (in this article we take the convention that algebraically greater values denote better interaction). The free energy of binding $\Delta \mathbf{G}(\mathbf{b})$ is formally broken down into a sum of contributions from all atoms of the molecule:

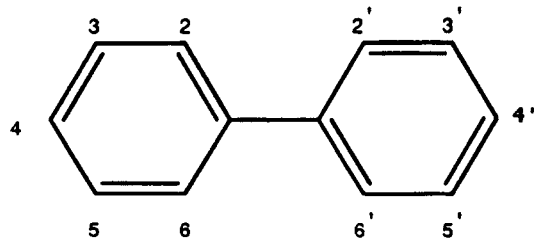
$$\Delta G(\mathbf{b}) = \sum_{\text{atoms } i} X_{\text{type i, region}(\mathbf{b}, i)}$$

where the X's are the interaction energies to be determined. The method for determining the interaction energies is based on subgradient optimization of an error function of the X's.[1]

In reference 1, the experimental ranges were set to a particular and relatively arbitrary value since no estimated error was given with the experimental binding energies.[6] Here, we allow them to vary in order to study the influence of their broadness on the site geometry. In other words, starting from a particular range, we find a geometry and interaction parameters fitting the data, then we increase the range until a simpler geometry is found and so on until the data can be fitted to a single large region.

We first started from the range used in ref. (1), i.e., $\Delta G_{m-} - \Delta G_{m+} = 3.8$ (here, the ΔG are given as $-\ln I_{50}$ where I_{50} is the molar concentration at 50% binding). Again, it was found that fitting the data requires at least three regions and the three-region site of Figure 3 was chosen with $r_{\min 13} = 1$ Å. This simple site model discriminates between planar and nonplanar molecules and can explain why compound **2** (which can't adopt planar conformations because of its o – Cl substituents and thus can't fit in the thin central region) binds much worse than its isomer **3** (Table I). Since all compounds have the same carbon skeleton, we have set (as in ref. 1) the interaction of the carbon atom with all three regions to zero in order to reduce the number of adjustable energy parameters. Table I gives the optimal modes found in this way for the twelve compounds with their corresponding binding energy $\Delta G_{m,\text{calc}}$ calculated from the interactions parameters found in step IV (Table II). As in ref. (1), these parameters were determined from the consideration of compounds **1, 2, 4, 5,** and **6.** We found that for five compounds (**1, 5, 6, 8,** and **12**), the optimal modes are identical to the ones given in reference 1, and they differ only slightly for four other compounds. Also, in both cases, only the $\Delta G_{m,\text{calc}}$ of compound **12** is out of bounds (but significantly). These facts are not surprising since in the random search for the parameters shown in Table II, we started from the solution found in ref. (1). However, by starting far enough from it, another and better solution was found which becomes exact if one increases very slightly (by 0.3 units) the ranges for three compounds (Tables III and IV). This is probably due to the fact that the approximate method of ref. (1) for calculating the binding modes missed some modes. Comparison of Tables I and III show that the biphenyls tend to bind all three regions more often in the second solution. In both solutions, however, we note that the oxygen atom (or at least one oxygen) always lies in the region of highest energy parameter value, thus hinting at a good interaction of the oxygen with the receptor.

Table I. Observed and calculated binding of biphenyls to prealbumin for the three-region site of Figure 3[a] and for an energy range of 3.8.[b]



Compound	ΔG_{m-}	ΔG_{m+}	$\Delta G_{m,\text{calc}}$	Optimal mode[c]
1 biphenyl	0.0	14.2	13.9	all in r_3
2 2,2′,6,6′-Cl_4	0.0	14.2	14.2	3′–6′ in r_1 2,6,2′ in r_2 rest in r_3
3 3,3′,5,5′-Cl_4	17.9	21.7	18.7	2,4′ in r_3 rest in r_2
4 3,3′,4,4′,5,5′-Cl_6	17.7	21.5	21.2	all in r_2
5 3,5-Cl_2-4-OH	18.6	22.4	18.6	OH,2′,3′ in r_3 rest in r_2
6 3,5-Cl_2-2-OH	15.3	19.1	18.8	3–5 in r_2 rest in r_3
7 2,4,6-Cl_3-4′-OH	15.7	19.5	18.3	2′,3′ in r_1 5′,6′,OH in r_3 rest in r_2
8 3,5,4′-Cl_3-4-OH	18.1	21.9	19.9	OH,2′,3′ in r_3 rest in r_2
9 2,3,4,5-Cl_4-4′-OH	17.6	21.4	21.3	2–5,2′ in r_2 rest in r_3
10 2,3,5,6-Cl_4-4,4′-$(OH)_2$	17.9	21.7	21.4	2′ in r_1 4′–6′,4-H in r_3 rest in r_2
11 3,3′,5,5′-Cl_4-4,4′-$(OH)_2$	18.4	22.2	22.2	4-OH,4′-H,2′ in r_3 rest in r_2
12 3,3′,4,4′,5,5′-Cl_6-2-OH	17.3	21.1	23.7	OH in r_3 rest in r_2

[a] $r_{\min 13} = 1$ Å.
[b] $\Delta G_{m+} - \Delta G_{m-} = 3.8$. The ΔG are given as $-\ln I_{50}$ where I_{50} is the molar concentration at 50% binding.
[c] Since we are neglecting the carbon atoms, 2–6 and 2′–6′ refer to the substituents at those positions, i.e., Cl, O, and H.

Next, we increased the energy ranges in order to fit the data to a two-region site (r_1, r_2). The main difficulty came from fitting both isomers **2** and **3** since their observed binding energies are quite different. Good differentiation of these compounds by the site of course requires that they have different optimal binding modes, which can only be obtained if they span both regions in their respective optimal modes. It is easy to see that in this case, the regions of highest energy parameter value must be different for the H and Cl atoms. In other words, the molecules tend to place all their H atoms in one region and all their Cl atoms in the other in order to get the highest possible binding energy. Examination of the feasible binding modes of **2** and **3** obtained with a two-region site showed that the best these molecules can do is *for both* to place two Cl atoms in one region and the remaining atoms in the other. Thus their optimal modes are still energetically identical and the only way for the data to fit the site is that their ranges become large enough to overlap. This is obtained if both ranges are increased by 1.9. As for the

Table II. Interaction Parameters ($\ln I_{50}$ units) for the data of Table I.

Atom	Site regions r_1	r_2	r_3
H	0.549	1.336	1.390
Cl	0.454	2.642	−3.746
O	0.088	1.030	2.460

Table III. Observed and calculated binding of biphenyls to prealbumin for the three-region site of Figure 3[a] and for an energy range of 3.8[b] (second solution).

Compound	ΔG_{m-}	ΔG_{m+}	$\Delta G_{m,\text{calc}}$	Optimal mode[c]
1 biphenyl	0.0	14.2	14.2	all in r_2
2 2,2′,6,6′-Cl_4	0.0	14.2	14.2	2 in r_3 4–6 in r_1, rest in r_2
3 3,3′,5,5′-Cl_4	17.9	21.7	17.6	3,3′ in r_3 rest in r_2
4 3,3′,4,4′,5,5′-Cl_6	17.7	21.5	19.2	3–5 in r_3 rest in r_2
5 3,5-Cl_2-4-OH	18.6	22.4	18.3	O in r_1 rest in r_2
6 3,5-Cl_2-2-OH	15.3	19.1	18.5	O in r_1, 5 in r_3 rest in r_2
7 2,4,6-Cl_3-4′-OH	15.7	19.5	16.3	6 in r_3 2–4 in r_1 rest in r_2
8 3,5,4′-Cl_3-4-OH	18.1	21.9	19.2	O in r_1, 4′ in r_3 rest in r_2
9 2,3,4,5-Cl_4-4′-OH	17.6	21.4	19.2	O, 6 in r_1, 2–4 in r_3 rest in r_2
10 2,3,5,6-Cl_4-4,4′-$(OH)_2$	17.9	21.7	21.2	4′-O, 2, 3 in r_1 5, 6 in r_3, rest in r_2
11 3,3′,5,5′-Cl_4-4,4′-$(OH)_2$	18.4	22.2	22.0	4-O in r_1, 5′ in r_3 rest in r_2
12 3,3′4,4′5,5′-Cl_6-2-OH	17.3	21.1	21.3	O in r_1, 5 in r_3, rest in r_2 and 3′–5′ in r_3, rest in r_2

[a] $r_{\min 13} = 1$Å.
[b] $\Delta G_{m+} - \Delta G_{m-} = 3.8$. The ΔG are given as $-\ln I_{50}$ where I_{50} is the molar concentration at 50% binding.
[c] Since we are neglecting the carbon atoms, 2–6 and 2′–6′ refer to the substituents at those positions, i.e., Cl, O, and H.

other biphenyls, we found that increasing their range by 1.2 led to two solutions depending on which set of compounds was chosen for the random search of the interaction parameters. Consideration of the set $C = \{\mathbf{2, 3, 4, 5, 10}\}$ gave optimal modes using both regions while the set $D = \{\mathbf{1, 2, 3, 5, 10}\}$ led to modes where all biphenyls lie in r_2 only (Tables V and VI). Note that there is no need to increase again the ranges in order to fit the data to a single region site since the solution already exists, and consists of the interaction parameters for r_2 in the set D solution. In other words, it has not been possible to find ranges such that one obtains a solution for the two-region site and no solution for the single-region site. This means that depending on the estimated experimental errors on the observed binding energies, one can model the receptor as a three-region site (or more if the ranges get smaller) or a single-region site, but never as a two region site.

Table IV. Interaction Parameters ($\ln I_{50}$ units) for the data of Table III.

Atom	Site regions r_1	r_2	r_3
H	0.169	1.42	−2.94
Cl	1.57	2.12	2.40
O	2.65	2.10	1.74

The programs to calculate the binding modes by the method described in the second section as well as the ones dealing with the molecular and site data bases are written in the C language. The most time consuming part of the whole algorithm is the calculation of the modes. For example, compound **8** consists of 11 atoms (since carbon atoms are neglected) and needs five unit vectors for its linear representation. Fitting it with a two region site and using the data bases to eliminate infeasible modes takes respectively 154 and 21 seconds of CPU time on an IRIS 2400 and on a SUN/4 computer. When going to the three-region site of Figure (3), these numbers increase to 206 and 26 minutes, respectively.

A very rough estimate of the computation order of our method with respect to the num-

Table V. Observed and calculated binding of biphenyls to prealbumin for a two-region site (r_1 and r_2).

Compound	ΔG^{a}_{m-}	ΔG^{a}_{m+}	$\Delta G^{a,b}_{m,calc}$	Optimal mode[c]
1 biphenyl	0.0	15.4	13.7(14.1)	all in r_1
2 2,2′,6,6′-Cl_4	0.0	16.1	16.0(16.1)	2,2′ in r_2 rest in r_1
3 3,3′,5,5′-Cl_4	16.0	21.7	16.0(16.1)	3,3′ in r_2 rest in r_1
4 3,3′,4,4′,5,5′-Cl_6	17.1	22.1	17.2(17.1)	3–5 in r_2 rest in r_1
5 3,5-Cl_2-4-OH	18.0	23.0	18.0(18.2)	3 in r_2 rest in r_1
6 3,5-Cl_2-2-OH	14.7	19.7	18.0(18.2)	3 in r_2 rest in r_1
7 2,4,6,-Cl_3-4′-OH	15.1	20.1	18.3(18.7)	2 in r_2 rest in r_1
8 3,5,4′-Cl_3-4-OH	17.5	22.5	18.3(18.7)	3 in r_2 rest in r_1
9 2,3,4,5-Cl_4-4′-OH	17.0	22.0	19.6(19.2)	3–5 in r_2 rest in r_1
10 2,3,5,6-Cl_4-4,4′-$(OH)_2$	17.3	22.3	22.3(22.3)	5,6 in r_2 rest in r_1
11 3,3′,5,5′-Cl_4-4,4′-$(OH)_2$	17.8	22.8	22.3(22.3)	3,3′ in r_2 rest in r_1
12 3,3′4,4′5,5′-Cl_6-2-OH	16.7	21.7	20.3(20.2)	3′–5′ in r_2 rest in r_1

[a]The ΔG are given as $-\ln I_{50}$ where I_{50} is the molar concentration at 50% binding.
[b]Values obtained when using the set *C*. The values in parenthesis are the ones obtained when using set *D* (see text).
[c]Modes obtained when using set *C*. Set *D* gave modes for which all atoms are in r_2. Since we are neglecting the carbon atoms, 2–6 and 2′–6′ refer to the substituents at those positions, i.e., Cl, O, and H.

ber of atoms and regions can be given. Another important parameter is the number of unit vectors n_u used to represent the molecule since the order of the Hessian in the system (12) to be solved by Gaussian elimination is $3(n_u + 1)$. A Newton iteration is in our case rather limited by solving this system (the update of the Hessian and gradient is pretty fast), which gives roughly $O(n_u + 1)^3$ flops per iteration. It is not yet clear what is the influence of n_u on the number of Newton iterations needed to find a solution. The influence of the number of regions on the computational cost is extremely complex to evaluate. The total number of nodes (i.e., systems to be soved by Newton's method) in the recursive tree algorithm for the placement of n atoms in r regions is $r^2(1 - r^{n-1})/(1 - r)$, but only a few percent of them are solved depending on the structure of both the molecule and the site model.

Table VI. Interaction Parameters ($\ln I_{50}$ units) for the data of Table V.

Atom	Site regions r_1[a]	r_2[a]
H	1.37(0.22)	−4.25(1.41)
Cl	1.70(−2.94)	2.19(1.91)
O	3.15(2.05)	1.14(3.10)

[a]Values obtained when using the set *C*. The values in parentheses are the ones obtained when using set *D* (see text).

From all this we can make conclusions about the feasibility of carrying out an analysis of drug binding using Voronoi polyhedra, and about the kind of results one can hope to achieve. First, the limiting step in such a study is searching out all the binding modes allowed, given the ligand and the geometry of the site. The correct numerical solution of the binding feasibility equations presented here is a qualitative advance over the earlier approach, which missed some allowed modes and admitted some incorrect modes (curiously, this had little effect on the final site model[7]). All this has been done in such a way as to search out efficiently all allowed binding modes, keeping in mind this is a substantial combinatorial problem. Second, the new nonnumerical rules for eliminating incorrect modes have produced a significant speed up. We anticipate they will permit us to attempt binding studies involving much larger molecules and more geo-

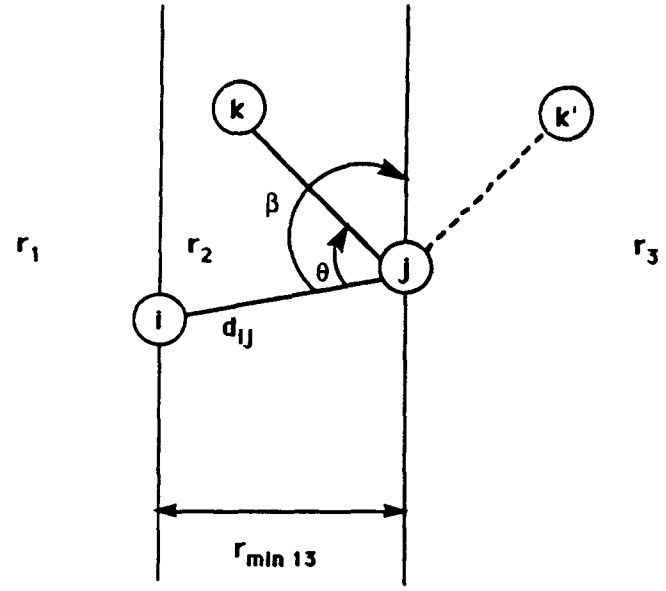


Figure 4. Geometrical check for the binding of atoms a_i, a_j, and a_k to the Voronoi region of Figure 3. If $\theta > \beta$ (position k' of atom a_k), the binding is infeasible.

metrically detailed site models. The third point is that aside from improving the methodology, we have shown the accuracy of the given experimental binding data strongly affects the level of structural detail required of the site model. To put it another way, as the accuracy is improved for the same set of ligands (much less adding new ligands to the data set) one is justified in building in more geometric detail in the deduced site. Low accuracy data can and should be fit by extremely simple pictures of what the real site looks like. This ability to produce both low and high resolution pictures is in our opinion the major strong point of the whole Voronoi site model formalism. We hope it will help counter the natural tendency to overinterpret the data.

APPENDIX

Let a_i and a_j be two atoms belonging to a common rigid group and lie in respectively two nonadjacent regions of a Voronoi site. For simplicity, we consider the regions r_1 and r_3 of the site of Figure 3. Let β be the maximum angle between the segment $a_i a_j$ of length d_{ij} and the boundary separating r_3 from r_2, defined as in Figure 4. This angle is obtained in placing a_i and a_j on the boundaries so that $a_i a_j$ spans exactly the central region r_2. From Figure (4), we have $\sin \beta = r_{\min 13}/d_{ij}$. Now, consider a third atom a_k from the same rigid group, and let θ be the angle a_i–a_j–a_k. Then it is clear that if $\theta > \beta$ (position k' on Figure 4) and a_i is in r_1 and a_j is in r_3, it is impossible to place a_k in r_2 and this rule allows one to eliminate infeasible binding modes. If $\theta < \beta$, no deduction can be made since we have a priori considered both the best positions for a_i and a_j (which only span r_2) and the minimum interregion distance $r_{\min 13}$. Note that by symmetry, if θ' is the angle a_j–a_i–a_k, a similar check should be made, i.e., $\theta' > \beta$ implies the mode is infeasible.

References

1. G.M. Crippen, *J. Comp. Chem.* **8,** 943 (1987).
2. L.G. Boulu and G.M. Crippen, Voronoi Receptor site models in *Proceedings of the 1988 OHOLO Conference on Computer-Assisted Modeling of Receptor-Ligand Interactions: Theoretical Aspects and Drug Design,* R. Rein and A. Golumbek, eds., Alan R. Liss Inc., New York, 1989, pp. 267–277.
3. G.M. Crippen, *Ann. New York Acad. Sci.*, **439,** 1 (1984).
4. A region r is defined to be convex if for any two points p and q of r, the segment pq also lies in r.
5. D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods,* Academic Press, New York, 1982, pp. 40–44.
6. U. Rickenbacker, J.D. McKinney, S.J. Oatley, and C.C.F. Blake, *J. Med. Chem.*, **29,** 641 (1986).
7. That this is the case is probably because the fitting procedure to the binding data depends on the calculated energy of the optimal modes which represent only a small fraction of the total number of modes. Likely, the errors on the experimental data were too large to generate a site detailed enough to discriminate the missing modes.