Intervals and the Deduction of Drug Binding Site Models

GORDON M. CRIPPEN
College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109

Received 6 April 1994; accepted 29 August 1994

ABSTRACT

In the search for new drugs, it often occurs that the binding affinities of several compounds to a common receptor macromolecule are known experimentally, but the structure of the receptor is not known. This article describes an extraordinarily objective computer algorithm for deducing the important geometric and energetic features of the common binding site, starting only from the chemical structures of the ligands and their observed binding. The user does not have to propose a pharmacophore, guess the bioactive conformations of the ligands, or suggest ways to superimpose the active compounds. The method takes into account conformational flexibility of the ligands, stereospecific binding, diverse or unrelated chemical structures, inaccurate or qualitative binding data, and the possibility that chemically similar ligands may or may not bind to the receptor in similar orientations. The resulting model can be viewed graphically and interpreted in terms of one or more binding regions of the receptor, each preferring to be occupied by various sorts of chemical groups. The model always fits the given data completely and can predict the binding of any other ligand, regardless of chemical structure. The method is an outgrowth of distance geometry and Voronoi polyhedra site modeling but incorporates several novel features. The geometry of the ligand molecules and the site is described in terms of intervals of internal distances. Determining the site model consists of reducing the uncertainty in the interregion distance intervals, and this uncertainty is described as intervals of intervals. Similarly, the given binding affinities and their experimental uncertainties are treated as intervals in the affinity scale. The final site model specifies an entire region of interaction energy parameters that satisfy the training set rather than a single set of parameters. Predicted binding for test compounds results in an interval which, when compared to the experimental interval, may be correct, incorrect, or vague. There is a pervasive ternary logic involved in the assessment of predictions, in the search for a satisfactory model, and in judging whether a given molecule may bind in a particular orientation: true, false, or maybe. The approach is illustrated on an extremely simple artificial example and on a real data set of cocaine analogues binding to a nerve membrane receptor in vitro.

Journal of Computational Chemistry, Vol. 16, No. 4, 486–500 (1995)

CCC 0192-8651 / 95 / 040486-15

Introduction

Much of biochemistry and virtually all of modern drug therapeutics revolves about the specific binding of small molecules to biological receptor macromolecules, such as inhibitors to enzymes. We have to understand why certain ligands bind to some receptor sites and not others, and quantitively why certain ligands bind with certain strengths. There are basically two sorts of experimental approaches to these questions: measuring the binding affinity to a certain receptor for several different ligands, or directly determining the structure of the ligand/receptor complex by X-ray crystallography and/or nuclear magnetic resonance (NMR) spectroscopy in solution. The direct structural methods yield a great deal of information that can be relatively unambiguously interpreted in terms of the three-dimensional structure of the entire macromolecule and its bound ligand. Unfortunately, these studies are much more difficult to carry out than simply measuring the binding constant, assuming purity, concentration, crystallization, etc. conditions can be met at all. Second, they also tend to speak only indirectly about the energetics of the binding. Therefore, it is not surprising that the vast majority of ligand/receptor studies are simple measuremetns of binding constants. What can we extract from such data? There must be some limit to the amount of information available in a given data set. There must be some way to get this information without building in our preconceptions. Ultimately, we want to deduce objectively receptor site geometry and energetics, given only the binding constants for a series of compunds. The ideal method would be completely objective and independent of guesses by the investigator. It would also not overinterpret the data; insufficient input data should yield the lowest resolution picture of the site which is still consistent with the facts. The resulting model must be predictive in that a correct binding energy and positioning within the site can be calculated for any novel compound whatsoever, and indeed the site model should be constructed in such a way as to suggest compounds with improved binding characteristics. Competing against these goals is the need for a method which will handle problems of practical magnitude in a reasonable amount of computer time. The challenge is that this is not just a matter of adjusting a few parameters in some theoretical equation to make a least-squares fit to a plot of experimental data. Instead, we are dealing with sometimes highly inaccurate binding data for three-dimensional, conformationally flexible, organic compounds, sometimes differing greatly in covalent chemical structure. The data arise from time averages in the random approaches of these ligands to the structurally complex receptor sites, all in the presence of solvent.

Of course, there are many methods in the quantitative structure–activity relations (QSAR) field for analyzing such binding data.[1] (See ref. 2 for a recent survey and assessment of current approaches.) Most seek an empirical least-squares correlation between some measure of biological activity or binding affinity and various molecular properties, such as physiochemical features of various groups of atoms, topological features of the covalent connectivity graphs, and three-dimensional structures of energetically favorable conformers. Whether the correlation is a qualitative classification of compounds into actives versus inactives or a more quantitative correlation, the common theme is a comparison of ligand molecules with each other in some absolute sense unrelated to the particular binding site involved. Such comparisons depend on an alignment rule that tells which atoms of one molecule are supposed to correspond to which atoms of a second. Typically, the alignment is a working hypothesis supplied subjectively by the investigator, which can be problematical in examples where extremely diverse molecules have strong affinity for the same site.[3] Many ways have been suggested for automatically finding good alignments.[4–6] When an alignment can be found such that every active compound has a common arrangement of a few key functional groups in three dimensions, this commonality is called the pharmacophore, and recently algorithms have been devised to determine them.[7] However one suggests alignments or pharmacophores, the success of the main 3D QSAR methods today—the "active analogue" approach[8–10] and the recently popular CoMFA program[11]—depend on them.

The alignment problem in isolation from the binding site nevertheless remains ambiguous, because even with great similarities in chemical structure, two molecules may bind in similar spatial orientations in the experimentally determined site/ligand structure[12] or they may bind completely reversed.[13] Implicit in the different ways for choosing alignments is neglect of an important

feature of physical reality that is simultaneously a substantial constraint on the site. The real ligands randomly and repeatedly approach the site and attach in mode k (translation and rotation of the ligand relative to the site, plus internal conformational changes of the ligand and, to a lesser degree, the site), having interaction energy E_k, with probability proportional to $\exp(-(E_k - E_o)/RT)$, where $E_o \leq E_k$ is the energy for the best mode. Typically $RT < (E_k - E_o)$ for a strongly binding molecule, so that one mode predominates. Choosing an alignment corresponds to selecting the one optimal mode in advance. We instead require that there be one mode for each molecule such that the calculated binding agrees with the experimental value and that all alternative modes correspond to weaker binding. These latter constraints turn out to be crucial in determining the site model and in giving it better predictive power.[14]

Methods

INTERVAL ANALYSIS

In much of what follows, molecular geometry and binding affinity are expressed not in terms of single, scalar numbers but rather *intervals*. To facilitate the discussion, it is useful to begin by introducing some basic ideas and notation from the branch of mathematics called interval analysis.[15] Intuitively, an interval is like a set with the additional notion that all the elements of an interval are ordered from the smallest to the largest, and there are no missing elements in the middle. Consequently, many of the relations between intervals resemble the corresponding relations from set theory.

Let $\mathscr{R}$ denote the field of real numbers and $\mathscr{I}(\mathscr{R})$ the set of all real intervals, where any particular interval A is defined as

$$A = \{x | a_1 \leq x \leq a_2, a_1, a_2 \in \mathscr{R}\} \in \mathscr{I}(\mathscr{R}) \quad (1)$$

For the sake of definiteness, we will discuss only intervals of real numbers in this section, although intervals of integers are actually used in the computer programs described later. Clearly, $a_1 \leq a_2$ in general, and when $a_1 = a_2$, A should behave like a real number. We will refer to the limits of an interval by $\lambda(A) = a_1$ and $\upsilon(A) = a_2$, and we will often write $A = [a_1, a_2]$. The obvious definition of an interval's width, δ, is

$$\delta(A) = \upsilon(A) - \lambda(A) \geq 0. \quad (2)$$

Binary and unary operations on intervals are customarily defined to produce the interval that contains the results of applying that same operation to numbers within the interval(s). Subtraction is of particular interest here

$$A - B = [a_1 - b_2, a_2 - b_1] \quad (3)$$

for intervals $A = [a_1, a_2]$ and $B = [b_1, b_2]$.

When comparing intervals A and B, we denote

$$\begin{aligned} A \subseteq B &\Leftrightarrow a_1 \geq b_1 \quad \text{and} \quad a_2 \leq b_2 \\ A \leq B &\Leftrightarrow a_1 < b_1 \quad \text{and} \quad b_1 \leq a_2 \leq b_2 \\ A < B &\Leftrightarrow a_2 < b_1. \end{aligned} \quad (4)$$

While the customary notation $a_1 \leq a_2$ for $a_1, a_2 \in \mathscr{R}$ includes both the cases $a_1 = a_2$ and $a_1 < a_2$, the interval comparisons defined in eq. (4) view $A < B$ and $A \leq B$ as mutually exclusive. To include both possibilities, we will write $A <\leq B$. In general, there are six possible outcomes in a comparison of intervals A and B—namely, the three listed and their respective reversals. All six are mutually exclusive except for the possibility that $A \subseteq B$ and $A \supseteq B$ iff $a_1 = b_1$ and $a_2 = b_2$, in which case we write $A = B$. Let the notation $A <\leq = B$ denote that either $A < B$ or $A \leq B$ or $A = B$. These relations allow us to generalize eq. (1) to intervals of intervals

$$\mathbb{A} = \{X | A_1 <\leq = X <\leq = A_2, A_1, A_2 \in \mathscr{I}(\mathscr{R})\} \in \mathscr{I}(\mathscr{I}(\mathscr{R})). \quad (5)$$

Just as we have overloaded the $<$, $\leq$ notation to signify similar relations when applied to real numbers and intervals, we will refer to the limits of $\mathbb{A}$ by $\lambda(\mathbb{A}) = A_1$ and $\upsilon(\mathbb{A}) = A_2$. In bracket notation, if $\mathbb{A} = [[a_1, a_2], [a_3, a_4]]$, then we can refer to $\lambda\upsilon(\mathbb{A}) = \lambda(A_2) = a_3$, $\upsilon\lambda(\mathbb{A}) = \upsilon(A_1) = a_2$, $\lambda^2(\mathbb{A}) = \lambda(A_1) = a_1$, and $\upsilon^2(\mathbb{A}) = \upsilon(A_2) = a_4$.

If A and B are sets, $A \setminus B$ denotes the set of all elements of A that are not in B. We will be concerned with the equivalent notion for intervals $A = [a_1, a_2]$ and $B = [b_1, b_2]$, where $B \subseteq A$ and $b_1 < b_2$. Define

$$A \setminus B = \begin{cases} \varnothing & \text{if } A = B \\ [a_1, b_1] & \text{if } b_1 > a_1 \text{ and } b_2 = a_2 \\ [b_2, a_2] & \text{if } b_2 < a_2 \text{ and } b_1 = a_1 \\ [a_1, b_1][b_2, a_2] & \text{if } b_1 > a_1 \text{ and } b_2 < a_2 \end{cases} \quad (6)$$

noting that the last case produces *two* disjoint intervals.

REPRESENTATION OF SHAPE

As explained in the Introduction, we are concerned with modeling the interaction of small drug molecules with a common receptor site on some macromolecule. Although the receptor is, of course, part of a large molecule, the term *molecule* in what follows always refers to the drugs, which have known structure, and *site model* or *site* refers to the receptor site, whose structure must be determined. The underlying molecular model we use is the rigid valence geometry approximation, in which atoms are points joined by covalent bonds having fixed lengths and vicinal bond angles. Thus, any molecule m has atoms $a = 1, \ldots, n_a$. A conformationally rigid molecule, such as methane or cyclopropane, has a geometrical structure that can be (redundantly) specified by the set of all interatomic distances, $\{d_{i,j}\}$. This information is insufficient unless we also include the chirality of any asymmetric centers, $\chi(i, j, k, l) \in \{-1, 0, +1\}$. More precisely, for some ordered quartet of distinguishable atoms, i, j, k, l, located at column vectors of Cartesian coordinates $\mathbf{c}_i, \mathbf{c}_j, \mathbf{c}_k, \mathbf{c}_l$, we define

$$\chi(i, j, k, l) = \operatorname{sign} \det \begin{pmatrix} 1 & 1 & 1 & 1 \\ \mathbf{c}_i & \mathbf{c}_j & \mathbf{c}_k & \mathbf{c}_l \end{pmatrix}. \quad (7)$$

Interchanging any pair of atoms in the ordered set changes the sign of χ, but it is invariant under rigid body translation and proper rotations.

Now if the molecule is conformationally flexible due to rotatable bonds, we interpret the $d_{i,j}$ to be intervals $D_{i,j}$, and for some—but not all—$\delta(D_{i,j}) > 0$, because the molecule explores all reasonably low-energy conformations. For example, for *n*-butane with carbon atoms numbered 1–4, $\delta(D_{1,4}) > 0$ as it goes from the *cis* to the *trans* conformations, but $\delta(D_{1,3}) = 0$, because the bond lengths and angles are assumed to be fixed. In the process, the χs for some arbitrarily chosen quartets of atoms may change sign, but for most molecules it is sufficient to focus on those quartets that are the substituents of asymmetric carbon atoms, and for these, χ is independent of conformation. Even when all the distance bounds are known, one cannot arbitrarily choose distances for each atom pair out of the corresponding range, because there are strong correlations between the various distances as the torsion angles are varied. Thus this summary of conformation space is a necessary restriction on the interatomic distances, but not a sufficient one.

In the applications that follow, the allowed conformation space of each molecule is explored by a systematic grid search over all combinations of torsion angle values, assuming that rings are rigid and rejecting only those conformations having van der Waals clashes. Then the distance bounds are the minimum and maximum observed values for each atom pair over all the allowed sampled conformations. Chiralities are noted for the substituents of asymmetric carbon atoms.

After the conformation space of a molecule has been searched, one may simplify the molecular structure by grouping specified sets of atoms together into united atoms, typically comprising functional groups or substituents. This is often necessary because the time required for the subsequent analysis increases rapidly with the number of atoms. If the original molecule has atoms $\{a_1, \ldots, a_{n_a}\}$, then the simplified molecule has united atoms $\{A_1, \ldots, A_{n_A}\}$, where $n_A < n_a$. A united atom A_I is just a set of atoms, and each a_i is a member of one and only one A_I. The distance bounds on the original atoms determine the distance bounds on the united atoms by

$$\begin{aligned} \lambda(D_{I,J}) &= \min \lambda(i_{i,j}) \qquad \forall a_i \in A_I,\ a_j \in A_J \\ \upsilon(D_{I,J}) &= \max \upsilon(D_{i,j}) \end{aligned} \quad (8)$$

and for any fixed conformation, the coordinates of a united atom are just the mean of the coordinates of its constituent atoms.

On the other hand, the site model is taken to consist of a set of n_r regions, $R = \{r_1, \ldots, r_{n_r}\}$, which are assumed to be nonoverlapping and convex, but otherwise of unspecified shape, and have either nonzero finite or even infinite size. Convexity means that for any two points $\mathbf{p}, \mathbf{q} \in r_i$, all points on the line segment joining them are also in that region.

$$\alpha\mathbf{p} + (1 - \alpha)\mathbf{q} \in r_i \qquad \forall\, 0 \le \alpha \le 1 \quad (9)$$

The flexibility of the real receptor site is included here in the region sizes, so the regions are taken to have fixed relative positions to one another. Specifically, let $D_{i,j}$ be the interval for distances from any point in r_i to any point in r_j. Then for $i = j$, $\lambda(D_{i,i}) = 0$, and $0 < \upsilon(D_{i,i}) \le \infty$ gives the (maximal) diameter of r_i. Otherwise, $0 \le \lambda(D_{i,j}) \le \upsilon(D_{i,j}) \le \infty$ give the minimal and maximal distance between the two points, where $\lambda(D_{i,j}) = 0$ if the two regions touch. See Figure 1, for example. Any quartet of regions may or may not have an assigned χ. Clearly, any realizable set of convex regions in $\mathscr{R}^3$ can be represented in this way, but

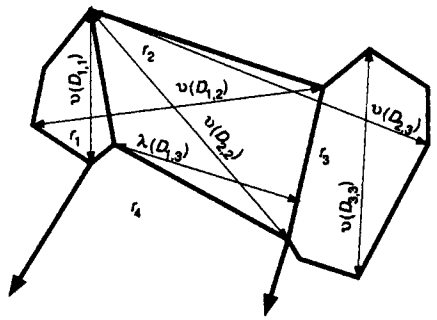


FIGURE 1. An example of four precise convex regions in the plane and the bounds on the distances between them. Not shown are $\lambda(D_{1,1}) = \lambda(D_{2,2}) = \lambda(D_{3,3}) = \lambda(D_{4,4}) = \lambda(D_{1,2}) = \lambda(D_{2,3}) = \lambda(D_{1,4}) = \lambda(D_{2,4}) = \lambda(D_{3,4}) = 0$ and $\upsilon(D_{1,4}) = \upsilon(D_{2,4}) = \upsilon(D_{3,4}) = \upsilon(D_{4,4}) = \infty$.

not every set of λs, υs, and χs corresponds to a realizable set.

In the algorithm to be described, we start with full knowledge about the three-dimensional structure of the molecules involved, but we are trying to proceed from total uncertainty to barely adequate certainty concerning the site model's geometry. Thus the relative position and sizes of regions r_i and r_j are not described simply by the interval $D_{i,j} \in \mathscr{I}(\mathscr{R})$ but by the interval of intervals $\mathbb{D}_{i,j} \in \mathscr{I}(\mathscr{I}(\mathscr{R}))$. Note that the definition of an interval of intervals [eq. (5)] requires $\lambda(\mathbb{D}_{i,j}) < \leq = \upsilon(\mathbb{D}_{i,j})$, so the *intervals* describing the least and greatest interregion distances may or may not overlap, but at least $\upsilon\lambda(\mathbb{D}_{i,j}) \leq \upsilon^2(\mathbb{D}_{i,j})$ and $\lambda^2(\mathbb{D}_{i,j}) \leq \lambda\upsilon(\mathbb{D}_{i,j})$. In other words, the lower interval may not extend above the upper one, nor the upper one below the lower.

Likewise, the site motel has the possibility of chirality relations among regions whenever $n_r > 3$. If for every choice of points $\mathbf{c}_i \in r_i$, $\mathbf{c}_j \in r_j$, $\mathbf{c}_k \in r_k$, and $\mathbf{c}_l \in r_l$ we find that $\chi(\mathbf{c}_i, \mathbf{c}_j, \mathbf{c}_k, \mathbf{c}_l)$ is the same, then we can assign that value to the chirality of the corresponding four regions, $\chi(r_i, r_j, r_k, r_l)$. In the modeling procedure, however, the sizes and shapes of the regions are given only as ranges on distance bounds, which makes it impossible to deduce that some quartet of regions actually has a fixed chirality. Instead, the chirality can either be unassigned, as an expression of uncertainty, or arbitrarily specified.

Now we can clearly define what we mean by more or less certain site geometries. The geometry of site $A = (D_A, \mathrm{X}_A)$, where for the n_r regions D_A is the full set of interregion distance intervals of intervals and X_A is the set of interregion chiralities. The least certain site A has $\mathbb{D}_{A,i,j} = [[0,\infty][0,\infty]]$ for all $\mathbb{D}_{A,i,j} \in D_A$, and $\mathrm{X}_A = \varnothing$. Intuitively, A covers the whole space of site geometries, and more certain or specialized sites are subsets of it in this space. For any two site geometries A and B, we define

$$B \subseteq A \Leftrightarrow \begin{cases} \lambda(\mathbb{D}_{B,i,j}) \subseteq \lambda(\mathbb{D}_{A,i,j}) & \forall i,j \\ \upsilon(\mathbb{D}_{B,i,j}) \subseteq \upsilon(\mathbb{D}_{A,i,j}) & \forall i,j \\ \mathrm{X}_B \supseteq \mathrm{X}_A \end{cases} \quad (10)$$

and $B \subset A$ if there is strict inclusion of either the sets of chiralities, or one or more distance intervals, or both. Note that *sub*sites tend to involve *super*sets of chiralities. Suppose $B \subset A$, and we want to consider the space covered by A to be broken up into B and $A \setminus B$. Just as in the last case of eq. (6), the site geometry $A \setminus B$ may consist of more than one site. Namely, each member of $A \setminus B$ consists of D_A and X_A, except for just one of the intervals where $\lambda(\mathbb{D}_{B,i,j}) \subset \lambda(\mathbb{D}_{A,i,j})$ or $\upsilon(\mathbb{D}_{B,i,j}) \subset \upsilon(\mathbb{D}_{A,i,j})$, or for one $\chi_{B,i} \in \mathrm{X}_B \setminus \mathrm{X}_A$. By supposition there is at least one member of $A \setminus B$, and possibly many. Where an interval differs, there are one or two site geometries having $\lambda(\mathbb{D}_{A,i,j}) \setminus \lambda(\mathbb{D}_{B,i,j})$ or $\upsilon(\mathbb{D}_{A,i,j}) \setminus \upsilon(\mathbb{D}_{B,i,j})$. Where a chirality differs, $\mathrm{X}_{A\setminus B} = \mathrm{X}_A \cup \{-\chi_{B,i}\}$ for one $\chi_{B,i} \in \mathrm{X}_B \setminus \mathrm{X}_A$. In other words, we include one of the extra chiralities after inverting it.

Suppose, for example, $n_r = 4$, and site A has $\mathbb{D}_{A,i,j} = [[0,\infty],[0,\infty]]$ for all i, j and $\mathrm{X}_A = \varnothing$. If site B is the same except $\mathbb{D}_{B,1,3} = [[0,17],[0,\infty]]$ and $\mathrm{X}_B = \{\chi(1,2,3,4) = 1\}$, then $B \subset A$ on account of both 1-3 distances and one chirality relation. Consequently, $A \setminus B$ consists of two members, E and F, which are the same as A except $\mathbb{D}_{E,1,3} = [[17,\infty],[0,\infty]]$ and $\mathrm{X}_F = \{\chi(1,2,3,4) = -1\}$.

REPRESENTATION OF ENERGETICS

In this work we consider $n_p = 2$ physicochemical properties of the molecules—namely, the molar refractivity and the water/octanol partition coefficient, log P. One can empirically assign to each atom in a molecule an atomic contribution to these molecular properties, based on the atom's element and nearby covalently connected neighbors but independent of conformation.[16] Thus for each atom, a, there is a fixed property vector, $\mathbf{v}_a \in \mathscr{R}^{n_p}$. When a set of atoms is grouped into a superatom, these vectors are simply summed to give the property vector of the united atom. Similarly, each

region of the site model, r_i, has a property vector, $\mathbf{w}_i \in \mathscr{R}^{n_p}$, the components of which are treated as adjustable parameters when developing the model.

The physical picture for the interaction between a molecule and a site model is that various subsets of the atoms fall into one or another of the nonoverlapping regions, so that every atom falls into exactly one region, although some regions may contain no atoms or one region may contain all the atoms. There is no particular relationship required between the numbers of atoms and regions. To put it more formally, we define a binding mode, μ, to be a mapping from atoms in the molecule to regions in the site, i.e., $\mu : \{a \in m\} \rightarrow R$. We assume the calculated binding affinity (e.g., $-\Delta G_{\text{bind}}$) to be

$$g(m, \mu) = \sum_{a \in m} \mathbf{v}_a \cdot \mathbf{w}_{\mu(a)} \qquad (11)$$

where our sign convention is that algebraically larger g values correspond to stronger binding. The optimal binding mode, μ^*, that maximizes g for given region property vectors is then the predicted binding mode for m, and $g(m, \mu^*)$ is the predicted binding affinity. Notice that the exact orientation and conformation of the molecule in the bound state are not predicted, but only an assignment of (united) atoms to the different regions, which may or may not correspond to a detailed atomic picture in $\mathscr{R}^3$.

Besides the chemical structure of each molecule, the only other experimental inputs to the problem are the observed binding affinities, expressed as an interval, $G(m)$, from $\lambda(G(m))$ to $\nu(G(m))$. The problem then is to determine the $\mathbf{w}_i$ such that

$$g(m, \mu^*) \in G(m) \qquad (12)$$

for all the molecules in the training set. Note this is not a least-squares fit. If the error bars are large, $g(m, \mu^*)$ may turn out to be near one of the limits of the observed range and therefore be far from the midpoint.

DEFINITION OF FIT

To develop a site model, one must look at all the different binding modes of each molecule. The combinational complexity of this task can be greatly reduced by grouping atoms and exploiting some general necessary conditions on binding modes. A somewhat arbitrary but necessary first step is the choice of united atoms discussed near eq. (8). For example, Figure 2 shows the reduction

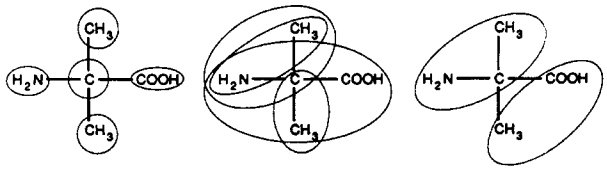


FIGURE 2. Atom groupings for α-aminoisobutyric acid. On the left is a reasonable choice of five united atoms for this molecule, where the central carbon remains a single atom. In the center are shown four of the 30 possible convex sets of these united atoms. At the right is one of the possible partitions, consisting of two convex sets.

of the 16 atoms of α-aminoisobutyric acid down to $n_a = 5$ united atoms. The second simplification follows from the assumption that every region is a convex region of space, so that the set of atoms found in one region in a particular binding mode must also be convex.[17] Here, a convex set of (united) atoms is defined to be one in which no atom outside the set lies within their convex hull, at least for one conformation. The middle of Figure 2 shows four of the 30 possible convex sets for the united atom representation of α-aminoisobutyric acid. Although there are $2^{n_a} - 1 = 31$ nonempty subsets in this example, the subset consisting of the four substituents of the central carbon atom is not convex, because the central carbon must lie within the convex hull (here a tetrahedron) formed by the substituents. With more complicated molecules, a greater fraction of the subsets tend to be eliminated this way. Our current approximation to finding convex sets uses the atomic coordinates for only one conformation, although it is possible for some sets to lose their convexity as the conformation is changed.

The third step in simplifying the combinatorics is based on the observation that a binding mode is not a completely arbitrary assignment of each of the n_a atoms to one or another of the n_r regions, a total of $n_r^{n_a}$ mappings. Regardless of the quantitative sizes and relative positions of the regions, their convexity and the demand that each atom lie in exactly one region imply that every valid mode is a *partition* of the molecule, where a partition is defined to be a set of n_r mutually exclusive and exhaustive convex sets of atoms. The right side of Figure 2 shows one partition of α-aminoisobutyric acid for $n_r \geq 2$. The advantage is that there are many fewer partitions than arbitrary mappings, typically on the order of the number of convex sets. Then in what follows, the only geometric features of the molecule that are important are the

distance intervals between convex sets and any chiral relations among them. For convex sets C_i and C_j,

$$\lambda(D_{i,j}) = \min_{a_I \in C_i,\, a_J \in C_j} \lambda(D_{I,J})$$
$$\upsilon(D_{i,j}) = \max_{a_I \in C_i,\, a_J \in C_j} \upsilon(D_{I,J}) \qquad (13)$$

and for any partition involving convex sets C_i, C_j, C_k, and C_l where there are $a_I \in C_i$, $a_J \in C_j$, $a_K \in C_k$, and $a_L \in C_l$ having $\chi(I, J, K, L) \neq 0$, we identify this with the chirality of the corresponding convex sets: $\chi(i, j, k, l) = \chi(I, J, K, L)$.

Out of all the binding modes corresponding to partitions, only a small fraction of them agree well enough in geometry between molecule and site to be permissible. Checking the agreement is done only in terms of interset versus interregion distances and chiralities, which are necessary but not always sufficient criteria. The problem is further complicated by the conformation flexibility and simplifying united atoms and derived convex sets (represented as bounds on interset distances), the flexibility and shapes of the regions (represented as bounds on the interregion distances), and the partially determined site geometry (represented as ranges on interregion distance bounds). This leads to a curious three-way logic: A binding mode may be certainly allowed ("sure"), certainly disallowed ("bad"), or indeterminate ("may"), depending on how the ranges on the interregion distance bounds will be contracted at a later time. See, for example, Figure 3.

Suppose for a given molecule and site, a particular mode assigns convex atom sets i and j to regions I and J, respectively. Then the three-way assessment of the agreement in geometry for this pair of atoms is

$$\text{bad} \Leftarrow \begin{cases} D_{i,j} < \lambda(\mathbb{D}_{I,J}) & \text{or} \\ D_{i,j} > \upsilon(\mathbb{D}_{I,J}) \end{cases}$$
$$\text{sure} \Leftarrow \begin{cases} \upsilon(D_{i,j}) > \upsilon\lambda(\mathbb{D}_{I,J}) & \text{and} \\ \lambda(D_{i,j}) < \lambda\upsilon(\mathbb{D}_{I,J}) \end{cases}$$
$$\text{may} \Leftarrow \begin{cases} D_{i,j} \leq \lambda(\mathbb{D}_{I,J}) & \text{or} \\ D_{i,j} \geq \upsilon(\mathbb{D}_{I,J}) \end{cases} \qquad (14)$$

expressed in terms of $D_{i,j} \in \mathcal{I}(\mathcal{R})$ and $\mathbb{D}_{I,J} \in \mathcal{I}(\mathcal{I}(\mathcal{R}))$. Suppose, for example, that $D_{i,j} = [4.5, 7.8]$. Then the top arrow of Figure 3 corresponds to the second "bad" case in eq. (14) when $\mathbb{D}_{I,J} = [[0, 2.1], [1.5, 4.0]]$. The middle arrow corresponds to the second "may" case when $\mathbb{D}_{I,J} = [[0, 2.1], [4.0, 7.0]]$, and the lower arrow corresponds to "sure" when $\mathbb{D}_{I,J} = [[0, 2.1], [5.0, 7.0]]$.

If, in addition, the mode assigns convex atom sets k and l to regions K and L, respectively, all four regions are distinct, and there is an assigned $\chi(i, j, k, l)$, then one must check the corresponding interregion $\chi(I, J, K, L)$ for agreement.

$$\text{bad} \Leftarrow \begin{cases} \exists\chi(I, J, K, L) & \text{and} \\ \chi(i, j, k, l) \neq \chi(I, J, K, L) \end{cases}$$
$$\text{sure} \Leftarrow \begin{cases} \exists\chi(I, J, K, L) & \text{and} \\ \chi(i, j, k, l) = \chi(I, J, K, L) \end{cases}$$
$$\text{may} \Leftarrow \{\nexists\, \chi(I, J, K, L) \qquad (15)$$

Then the entire mode is declared bad, sure, or may, depending on the outcomes for every pair of atoms [eq. (14)] and every chiral quartet of atoms [eq. (15)] in the following way. If any distance or chirality is bad, the whole mode is bad. If no distance or chirality is bad, but any one of these is may the whole mode is may. Only if all distances and chiralities are sure is the whole mode sure. In what follows, the bad modes are of little interest, but we will focus on S, the set of sure modes for a given molecule, and T, the set of may modes.

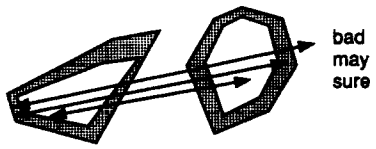


FIGURE 3. Schematic illustration of two regions for which the mutual distance limits have not yet been fully determined. The gray areas signify the ranges in the limits. Suppose an arrow represents the distance between two atoms in a molecule having a proposed binding mode that puts one atom in the left region and the other in the right. Then the top arrow is certainly bad as being too long, the middle arrow is may (depending on whether the limits will expand to the outer outlines or contract to the inner ones), and the bottom arrow is sure.

SEARCH ALGORITHM

Now that the general model for molecules and sites and their interaction has been presented, we can turn to the algorithm for finding site models. What follows has been implemented as some 4000

lines of C code collectively named "Egsite" (energy and geometry of site models).

The first simplifying observation is that there are only relatively few critical values required for interregion distances in order to discriminate between binding modes according to eq. (14). Given a set of n_m molecules interacting with n_r regions, we first determine all $\lambda(D_{i,j})$ and $\upsilon(D_{i,j})$ for all pairs of convex sets within each molecule and plot these as points along the real number line. Any two points closer than some tolerance (e.g., 0.1 Å) are counted as a single point. Then the critical distances are the midpoints between these clusters as well as some value significantly less than all $\lambda(D_{i,j})$ (the effective zero distance) and some value significantly greater than all $\upsilon(D_{i,j})$ (the effective infinite distance). Suppose, for example, the plot of intramolecular distance bounds is the ordered list [2.12, 2.16, 4.22, 4.40, 4.44, 6.20]. Clustering at a tolerance of 0.1 produces the list of four clusters, [(2.12, 2.16), 4.22, (4.40, 4.44), 6.20], which have midpoints between clusters of [3.19, 4.31, 5.32]. Adding the effective zero and infinite distances produces the critical distance list, [2.02, 3.19, 4.31, 5.32, 6.30]. Then in what follows, the ranges on the bounds on all interregion distances may take on values chosen from the discrete, finite list of critical distances, denoted by $[d_0, d_1, \ldots, \infty]$. This also eliminates technical questions about what to do in the case of equality in eq. (14), because any critical distance is always clearly greater than or less than any intramolecular distance, by construction. Chiralities among quartets of regions are already discrete, having values ± 1 or being absent.

The problem now is to determine the $2n_r^2$ endpoints of the ranges on interregion distance bounds and possibly introduce some region chiralities such that eq. (12) is satisfied for all n_m molecules. Such a site will be considered a solution even if its interregion distance intervals are wide. Because every binding mode assigns a region to each atom, the regions are hypothesized to account for all space. Physically, some of them may correspond to different portions of the receptor site, but one of them must amount to the solvent outside the receptor. Accordingly, let r_1 represent the solvent, and we may fix $\upsilon(\mathbb{D}_{1,I}) = [\infty, \infty]$ for all $I = 1, \ldots, n_r$. This guarantees that every molecule has available at least the one sure mode, where all atoms fall into r_1. Of course, we have also fixed $\lambda(\mathbb{D}_{I,I}) = [d_0, d_0]$ for all $I = 1, \ldots, n_r$. Otherwise, at the outset the site is taken to have no chiralities and all other $\lambda(\mathbb{D}_{I,J}) = \upsilon(\mathbb{D}_{I,J}) = [d_0, \infty]$.

Finding a solution is viewed as exploring a decision tree, where the initial, maximally vague site is the root at the top. The first child B of any node A is a particular strict subsite of A, and the one or more other children are the various subsites in $A \setminus B$, as discussed above at eq. (10). At every node in the tree, the current site model is tested to see whether it is a solution, according to methods described below. If it is a solution (the node is "sure"), one may either halt at the first solution or backtrack up the tree to find all solutions eventually. The second possible outcome at a node ("bad") is that it is not a solution and, furthermore, there is no possibility of a solution deeper down from here in the tree. In this case, one backtracks up a level in the tree and continues the search. The third and last possible outcome is that the node is "may," meaning that although the tests have not demonstrated a solution here, there is still a possibility that a solution may be found by searching further down the tree from here. In this case, a particularly promising subsite is chosen by a method described below to be the first child, and this consequently determines the geometries of the other children. Because the search is carried out in depth first order, taking the first child first, the first solution (if any) tends to be located quickly. A step down in the decision tree always corresponds to making a more restrictive subsite geometry. Because of the definitions of subsites and sure/bad/may modes, any sure mode in the parent node is still sure in the child, and any bad mode remains bad. However, zero or more of the may modes are changed into either sure of bad modes. Lowering some $\upsilon^2(\mathbb{D}_{i,j})$ or raising some $\lambda^2(\mathbb{D}_{i,j})$ may change the classification of no modes or move some from may to bad. Similarly, raising some $\lambda\upsilon(\mathbb{D}_{i,j})$ or lowering some $\upsilon\lambda(\mathbb{D}_{i,j})$ can at most move some modes from may to sure. Adding a chirality can have either effect. Thus descending the decision tree not only makes the site geometry more definite, but it more clearly delineates how the molecules must interact with the site.

Evaluation of a node in the search tree consists of a sequence of increasingly more time-consuming tests that are halted as soon as a sure/bad/may decision can be made.

1. Because r_1 is infinite, every molecule has at least one sure mode. If $\cup_m T_m = \varnothing$, then make test 2, else test 5.
2. The site is so narrowly specified that no molecule has any may modes. Let $C_S =$

$\cup_m\{g(m, \mu) \leq v(G(m))| \mu \in S_m\}$ be the set of linear constraints on the ws requiring the affinity of all sure modes to be less than the greatest observed affinity. If C_S has no solution (i.e., the constraints are mutually inconsistent), the node is bad. Otherwise try test 3.

3. If there is one $\mu_m^{(\lambda)} \in S_m$ for each molecule such that each of the n_m sets of inequalities $C_S \cup \{\lambda(G(m)) \leq g(m, \mu_m^{(\lambda)})\}$ is consistent, then try test 4. Otherwise, at least one of these sets is inconsistent and the node is bad.
4. If there is some combination of one $\mu_m^{(\lambda)} \in S_m$ for each molecule such that $C_S \cup (\cup_m\{\lambda(G(m)) \leq g(m, \mu_m^{(\lambda)})\})$ is consistent, then the node is sure. If there is no such combination, then the node is bad.
5. From test 1, it has been established that every molecule has at least one sure mode, and at least one molecule has at least one may mode. As in test 2, if C_S has no solution, the node is bad. Otherwise try test 6.
6. As in test 3, if C_S can be augmented by one sure lower bound, considering each molecule independently, go on to test 8. Otherwise try test 7.
7. Consider the following test applied to each of the n_m molecules independently. Is there one $\mu_m^{(\lambda)} \in T_m$ for molecule m such that when the site geometry is temporarily minimally modified to make it sure (thus changing S to S' for all molecules), the inequalities $C_{S'} \cup \{\lambda(G(m)) \leq g(m, \mu_m^{(\lambda)})\}$ are consistent? If the test is true for all molecules, then the node is may. Otherwise it is bad.
8. As in test 4, if there is some combination of one sure lower bound for each molecule that can be simultaneously added to C_S and still be consistent, then the node is sure. Otherwise it is may.

Clearly the node evaluation procedure is complicated, but the brief list of tests given here outlines the logic without getting lost in the details. Some additional comments may make the individual steps seem less cryptic. Remember that searching for a solution amounts to descending a tree of these nodes, proceeding to greater geometric restrictions, which can also be viewed as moving may modes to the sure and bad categories. Test 1 asks whether this procedure has terminated by finally eliminating all may modes for all molecules. If so, tests 2, 3, and 4 in the left branch of Figure 4 are the relatively clear-cut questions to ask, as

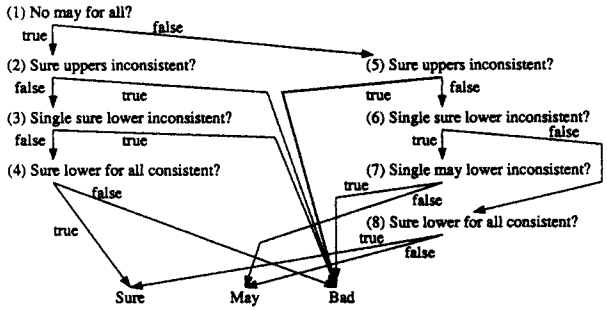


FIGURE 4. Decision tree for evaluating a node in the geometric search tree. See text for details.

opposed to the alternative versions in tests 5, 6, 7, and 8 along the right branch. Now for a solution (i.e., a sure search node), all sure modes of each molecule must bind no more strongly than the experimental upper limit, and there must be at least one sure mode for each molecule that binds more strongly than the experimental lower limit. Consequently, test 2 asks whether all the current sure modes are compatible with the requirement that no molecule can bind more strongly than its experimental upper limit, regardless of which mode it tries. Finding ws that satisfy the set of inequalities C_S is equivalent to proving compatibility. Failing test 2 certainly determines the node is bad, regardless of what mode might be chosen as optimal for each molecule, because those choices will only add to the set of inequalities to be solved. Test 3 makes a first attempt at adding these extra inequalities by treating each molecule independently. If for some molecule m there is no sure mode $\mu_m^{(\lambda)}$ whose calculated binding can be raised above the experimental lower limit without violating some upper bound in C_S due to any of the molecules, then there is no hope of a solution. Only in test 4 do we try the much more difficult problem of designating one sure mode as $\mu_m^{(\lambda)}$ for each molecule simultaneously and adding these constraints to C_S. Because each molecule may have hundreds of sure modes, there are many combinations to try, so that proving a node bad can be tedious, whereas the first successful combination of $\mu_m^{(\lambda)}$s is sufficient to prove the node is sure.

Tests 5, 6, 7, and 8 run similarly to 2, 3, and 4, but now the possibility exists that a child of this node (a more restricted subsite) may be sure, even if this node is may. Because any child node will have at least this current set of sure nodes and maybe more, test 5 checks the mutual consistency of C_S as a first necessary condition. Test 6 continues examining the adequacy of the current sure

modes by seeking one $\mu_m^{(\lambda)}$ to add a lower bound energy constraint to C_S for each molecule independently. Passing test 6 suggests that the current set of sure modes may prove adequate for a solution, but failing it leaves open the possibility that there exists a child search node in which some may modes have been moved to the sure category such that the child will pass test 6. Test 7's job is to examine this latter possibility by an inexpensive, one-molecule-at-a-time calculation. For each molecule m in turn, it chooses a may mode to be the $\mu_m^{(\lambda)}$ corresponding to which a lower energy bound will be added to the set of all upper energy bounds from all molecules. The reason it is a may mode is that one of the two may cases of eq. (14) was true for at least one pair of convex sets or the may case of eq. (15) was true for at least one chiral quartet of convex sets. To make this mode sure temporarily, one would have to tighten the corresponding interregion distance intervals just enough to pass the sure conditions of eq. (14) and add any required chiral relations among the regions. The effect of these minimal site modifications is to produce a subsite in which $\mu_m^{(\lambda)}$ is sure for molecule m while possibly moving other may modes for all the molecules into the sure and bad categories. Therefore, the corresponding new set of energy upper bound inequalities, $C_{S'}$, is combined with the one energy lower bound inequality and checked for consistency. Passing test 7 still does not prove there is some child node that is sure. Therefore, the current node is marked may, so its children will be examined later in the tree search. Finally, test 8 follows up on the success of test 6, checking that the current sure modes may be sufficient to prove there is a solution. Like in test 4, the many combinations of $\mu_m^{(\lambda)}$s make this a much more expensive test than 6. Failing 8 still leaves open the possibility of a successful child node, so the current node is only marked as may.

Any time a search tree node is declared may by test 7, there is at least one molecule m that failed test 6 but passed test 7 for a particular mode $\mu_m^{(\lambda)} \in T_m$. Making this mode sure involves creating a tailored subsite, which then becomes the designated site geometry of that node's first child. In the much rarer case (ca. 200 times rarer in our experience) that the node is declared may by test 8, we choose the subsite of the first may mode of the first molecule that has one. Thus the initial descent of the decision tree from parent to first child always corresponds to moving more and more may modes to the sure category rather than moving them to bad, as long as no chiralities are

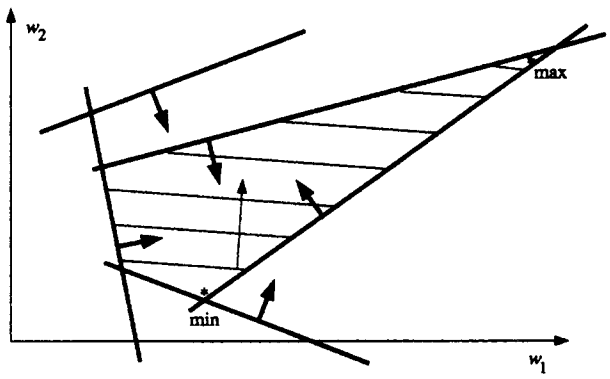


FIGURE 5. An artificial example of the feasible region for $\mathbf{w} \in \mathscr{R}^2$ delineated by heavy lines and arrows corresponding to nonredundant inequalities. Inside the feasible region, light lines represent the level lines for $g(m, \mu)$ for one particular binding mode μ, and the light arrow shows the gradient of $g(m, \mu)$. One redundant inequality is also indicated.

introduced. Eventually backtracking and trying subsequent children has more the effect of converting may modes to bad.

Although it is theoretically possible to fail tests 2 or 5, this almost never occurs in practice. Clearly the most time-consuming steps are decisions 4 and 8, but these are performed seldom, typically just to verify the existence of a solution. Test 7 is the decisive one generally, in which there may be many hundreds of attempts to create a set of several hundred inequalities in a dozen variables and seek a solution for them. Standard linear programming by the simplex algorithm either locates a set of $\mathbf{w}_i$ for $i = 1, \ldots, n_r$ that satisfy all the inequalities, or it determines unambiguously that there is no such solution. (Numerical instabilities can cause problems here with naive linear programming code. We currently use an interior point method as implemented in the LOQO program.) When the inequalities are consistent, there is not only a solution point in $\mathscr{R}^{n_r n_p}$, but a whole finite or infinite polyhedron, called the feasible set, bounded by hyperplanes corresponding to some of the inequalities (see Fig. 5). These are referred to as the nonredundant inequalities because altering any one of them would change the feasible set. We take the energetic solution to be not just the one set of $\mathbf{w}_i$ values, but the whole feasible set as described by the set of nonredundant inequalities, C^*. Denote the feasible set by $W^* = \{\mathbf{w} \in \mathscr{R}^{n_r n_p} | \mathbf{c}_i \mathbf{w} < b_i \ \forall \mathbf{c}_i, b_i \in C^*\}$. In the terminology of linear programming, we identify the nonredundant inequalities by a series of "phase 2" mini-

mizations of the slack variable of each inequality for which the redundancy status is yet unknown. If the minimal value of the slack is zero, that inequality is nonredundant, as are any that are in the basis at the solution. The approach is simple and reliable, but not the emost efficient. As there is typically a lot of degeneracy in the inequalities, so-called weakly redundant inequalities may be declared nonredundant by this procedure, but the feasible set so described is still correct.

PREDICTION

The successful outcome of the search algorithm is the geometry of the regions in terms of chiralities and ranges on interregion distance bounds and the energetics in terms of W^*. Because the feasible region delineated by C^* is larger than a single point, and because the geometry of the site may permit several different binding modes for each test molecule, the predicted binding is not a single value, but a range.

To put it more precisely, consider the prediction of binding for molecule m, where the geometry of the site model permits a nonempty set of sure modes, $S_m \neq \emptyset$. Then for any $\mu \in S_m$, $g(m, \mu)$ is a linear function of $\mathbf{w}$. Now maximizing or minimizing $g(m, \mu)$ subject to C^* is just the linear programming problem illustrated in Figure 5, resulting in the greatest and the least calculated binding affinity, respectively, for that mode. In the figure, these values of $\mathbf{w}$ are the vertices marked "max" and "min." Of course, for different μ the linear function to be optimized will be different in general, resulting in locating possibly different vertices of W^* corresponding to different extremal values of the calculated affinity. Let

$$\lambda(G^*(m)) = \max_{\mu \in S_m} \min_{\mathbf{w} \in W^*} g(m, \mu)$$

$$\upsilon(G^*(m)) = \max_{\mu \in S_m} \max_{\mathbf{w} \in W^*} g(m, \mu) \qquad (16)$$

be the calculated lower and upper bounds on the affinity, which in general occur for different binding modes.

Equation (12) gives the criterion for agreement between experiment and calculation when the predicted binding is a single number. Instead, we now have to compare the experimental interval, $G(m)$, with the calculated interval, $G^*(m)$. If the calculated range lies entirely within the experimental one, then we judge the site model to agree with experiment completely. Otherwise, there are two kinds of disagreement: benign excess calculated range and outright error. In the excess range case (denoted by "xs" in Table I in the Results section), the calculated range overlaps the experimental one to some degree but extends beyond it, so the model is not necessarily in disagreement, but it cannot exclude erroneous binding affinities. As an extreme example, $[-\infty, +\infty]$ is a prediction that covers the experimentally observed range but is so vague as to be trivial. In summary, we define the prediction error, $E(m)$, by

$$E(m) = \begin{cases} 0 & \text{if } G^*(m) \subseteq G(m) \\ -\lambda(G^*(m) - G(m)) & \text{if } G^*(m) < G(m) \\ \lambda(G^*(m) - G(m)) & \text{if } G^*(m) > G(m) \\ \max\begin{pmatrix} \lambda(G(m)) - \lambda(G^*(m)) \\ \upsilon(G^*(m)) - \upsilon(G(m)) \end{pmatrix} & \text{otherwise} \end{cases} \qquad (17)$$

where the first case covers correct predictions, the second and third outright errors, and the fourth excess errors.

Results

SIMPLE EXAMPLE

Before considering the performance of the algorithm on real experimental data, it is helpful to examine a tiny, artificial test case. Suppose we have two molecules consisting of two important functional groups "A" apiece separated by unimportant spacer groups. In the molecule denoted by "AA," the spacer is short, so that the distance between A's is 1 length unit; the other molecule, "A-A," is 3 units long. Let the observed binding affinities be $G(AA) = [1,2] < G(A - A) = [6,7]$ in some arbitrary units. Let $n_p = 1$ and $\mathbf{v}_A = 1$, so that $\mathbf{w}_i$ is just the contribution to binding when an A group lies in region r_i. The objective is to calculate the simplest and vaguest site model that gives $E(AA) = E(A - A) = 0$. Simple models involve

few regions, and vague ones have few chiralities, wide interregion distance intervals, and broad W^*.

Regardless of the complexity of the site model, the only critical distances can be assigned formal values of [0, 2, 4], where 4 is effectively infinite. The simplest site has one region, r_1, where the method demands a fixed $\mathbb{D}_{1,1} = [[0, 0], [4, 4]]$. We will write binding modes as ordered tuples (i, j), where the first A lies in r_i and the second in r_j. The sets of sure modes are $S_{AA} = \{(1, 1)\}$ and $S_{A-A} = \{(1, 1)\}$, whereas the sets of may modes $T_{AA} = T_{A-A} = \varnothing$. Because each molecule has only a single sure mode, the unique set of inequalities to be solved in step 4 of the search algorithm is just $\{2w_1 < 7,\ 2w_1 < 2,\ 2w_1 > 6,\ 2w_1 > 1\}$. Clearly these are inconsistentm, so that node in the search tree is bad; and because it is the only node, there is no site model possible with only one region.

The next step up in complexity is two regions, starting the search tree with the maximally vague geometry compatible with an infinite first region —namely, $\mathbb{D}_{1,1} = [[0, 0],\ [4, 4]]$, $\mathbb{D}_{1,2} = [[0, 4], [4, 4]]$, and $\mathbb{D}_{2,2} = [[0, 0], [0, 4]]$. With only two regions in three spatial dimensions, there can be no chiralities. Then $S_{AA} = \{(1, 1)\}$, $S_{A-A} = \{(1, 1)\}$, $T_{AA} = \{(1, 2), (2, 1), (2, 2)\}$, and $T_{A-A} = \{(1, 2), (2, 1), (2, 2)\}$. This passes step 5 because $\{2w_1 < 7,\ 2w_1 < 2\}$ is consistent, but in test 6 the introduction of the lower bound for A − A creates the inconsistent set of inequalities $\{2w_1 < 7, 2w_1 < 2, 2w_1 > 6\}$, forcing test 7. Test 7 is passed by altering temporarily $\lambda(\mathbb{D}_{1,2})$ to [0, 2] so that still $S_{AA} = \{(1, 1)\}$, but now $S_{A-A} = \{(1, 1), (1, 2), (2, 1)\}$. This suggested allteration is passed back to the search tree, which consequently branches to either $\lambda(\mathbb{D}_{1,2}) = [0, 2]$ or [4, 4]. Taking the first branch, test 8 declares the node to be sure for the choice $\mu^{(\lambda)}_{AA} = (1, 1)$ and $\mu^{(\lambda)}_{A-A} = (1, 2)$, corresponding to the consistent set of inequalities $\{2w_1 < 7, w_1 + w_2 < 7, 2w_1 < 2, w_1 + w_2 > 6, 2w_1 > 1\}$. Then eliminating the redundant inequality, we have $C^* = \{w_1 + w_2 < 7, 2w_1 < 2, w_1 + w_2 > 6, 2w_1 > 1\}$. Checking the solution by "predicting" the training set, we find $G^*(AA) = G(AA)$ and $G^*(A - A) = G(A - A)$. Calculating sure modes amounts to assuming the effective $\mathbb{D}_{i,j} = [[\upsilon\lambda(\mathbb{D}_{i,j}),\ \upsilon\lambda(\mathbb{D}_{i,j})],\ [\lambda\upsilon(\mathbb{D}_{i,j}),\ \lambda\upsilon(\mathbb{D}_{i,j})]]$, so the picture of the site model consists of a large r_1 having a weak $w_1 \in [0.5, 1.0]$, separated in space from r_2 by a gap too long for AA to span but short enough for $A - A$. Then r_2 is so small that it will accept only a single A group, but it has a sufficiently strong w_2 to make $g(A - A, (1, 2))$ in the range [6, 7].

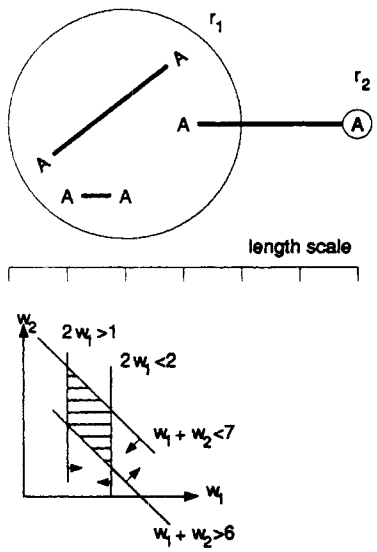


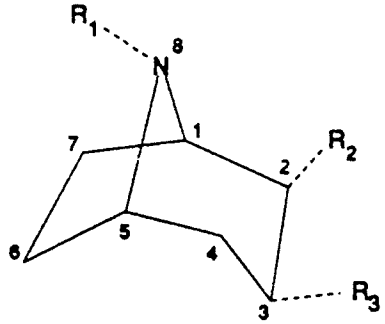
FIGURE 6. The final site model corresponding to the short *AA* dimer binding weakly and the long *A* − *A* isomer binding strongly. At the top, the regions are depicted as circles with the distance scale shown for reference. Sure binding modes are displayed. Below is the energy parameter space showing the nonredundant inequalities defining the hatched feasible region.

COCAINE ANALOGUES

For the sake of comparison, we considered the same set of 20 stereoisomers and analogues of cocaine binding to a receptor site in brain cell membranes, as we had studied earlier.[18] All molecules were simplified as before by grouping all atoms of each molecule into five superatoms: the atoms and substituents at positions 1 and 7; at 4, 5, and 6; at 2; at 3; and at 8 (see structure in Table I). The experimental binding affinities[19–22] were given as best estimated IC_{50s}, so the binding ranges G in Table I are $-\log(IC_{50}) \pm$ the previously estimated errors. We took the same final training set of eight compounds (**1**, **2**, **3**, **4**, **7**, **12**, **15**, and **17**), leaving 12 test compounds. The only difference is that the earlier Egsets program has been replaced by the current Egsite program.

Once the training set was chosen and the grouping of atoms into superatoms was selected, Egsite requires no manual intervention. Convex sets are calculated once and for all for each molecule. Then solutions are sought for first one region, then two, three, and finally four. For a given number of regions, the partitions of each molecule are calculated, which determines the full set of binding modes. Then the solution tree was searched until

TABLE I.
Potencies of Cocaine Analogues for Inhibition of the Binding of [^{3}H]WIN-35428 at the Dopamine Transporter.



Molecule[a] configuration		Common name	R_1	R_2	R_3	Observed G[b]	Calculated G*	Error E
1	R	Cocaine (C)	Me	β-CO_2Me	β-O(CO)Ph	[6.72, 7.82]	[6.72, 7.82]	0
2	S	Cocaine	Me	β-CO_2Me	β-O(CO)Ph	[4.53, 5.47]	[4.53, 5.47]	0
3	R	Pseudo-C	Me	α-CO_2Me	β-O(CO)Ph	[4.53, 5.63]	[4.53, 5.47]	0
4	S	Pseudo-C	Me	α-CO_2Me	β-O(CO)Ph	[4.38, 5.47]	[4.53, 5.47]	0
5	R	Allo-C	Me	β-CO_2Me	α-O(CO)Ph	[4.94, 6.03]	[7.24, 11.08]	+1.20
6	S	Allo-C	Me	β-CO_2Me	α-O(CO)Ph	[4.74, 5.83]	[7.24, 11.08]	+1.40
7	R	Allo Pseudo-C	Me	α-CO_2Me	α-O(CO)Ph	[4.28, 5.37]	[4.53, 5.37]	0
8	S	Allo Pseudo-C	Me	α-CO_2Me	α-O(CO)Ph	[3.90, 4.99]	[4.53, 5.37]	0.38 xs
9	R		Me	β-CO_2Ph	β-O(CO)Ph	[6.68, 7.78]	[7.43, 13.12]	5.34 xs
10	R		Me	β-CO_2 CH_2CH_2Ph	β-O(CO)Ph	[6.34, 7.43]	[6.55, 16.31]	8.88 xs
11	R	Tropa-C	Me	-H	β-O(CO)Ph	[5.02, 6.11]	[4.77, 10.08]	4.22 xs
12	R	Benzoyl-ecgonine	Me	β-CO_2H	β-O(CO)Ph	[3.44, 4.53]	[3.56, 4.53]	0
13	R		Me	β-$CO_2(Me)_2$-(p-NH_2-Ph)	β-O(CO)Ph	[6.88, 7.97]	[6.76, 21.53]	13.68 xs
14	R		Me	β-CH_2OH	β-O(CO)Ph	[5.98, 7.07]	[3.10, 4.55]	−1.44
15	R		CH_2 Ph	β-CO_2Me	β-O(CO)Ph	[5.91, 7.00]	[5.91, 7.00]	0
16	R		H	β-CO_2Me	β-O(CO)Ph	[6.25, 7.34]	[4.52, 8.98]	3.38 xs
17	R		Me	β-CO_2Me	β-Ph	[7.37, 8.46]	[7.37, 8.46]	0
18	R		Me	β-CO_2Me	β-Ph-p-F	[7.53, 8.62]	[7.32, 9.27]	0.86 xs
19	R		Me	β-CO_2Me	β-Ph-p-NH_2	[7.34, 8.43]	[4.63, 7.96]	2.71 xs
20	R		Me	β-CO_2Me	β-Ph-p-OMe	[7.83, 8.92]	[6.72, 7.88]	1.11 xs

[a]Numbering as in Table 1 of ref. 17.
[b]Data from Carrol et al.[18-21] as $-\log(IC_{50})$ (μM).

TABLE II.
First Site Model Found for the Training Set of Eight Cocaine Derivatives.

Regions	r_1	r_2	r_3	r_4
r_1	$[0, \infty]$	$[4.04, \infty]$	$[4.32, \infty]$	$[4.60, \infty]$
r_2		$[0, 0]$	$[4.32, 8.50]$	$[2.86, 2.86]$
r_3			$[0, 0]$	$[5.33, 6.19]$
r_4				$[0, 2.86]$
w_{HP}	5.34	−13.65	1.55	−4.16
w_{MR}	−0.36	0.76	−0.04	0.57
$\chi(r_1, r_2, r_3, r_4) = +1$				

The intervals are the interregion distance intervals in Å. Also shown is a representative set of values for the hydrophobicity (HP) and molar refractivity (MR) interaction energy parameters associated with each region. At the bottom, the one required interregion chiral relation is indicated.

either a solution was found or all branches terminated in "bad" nodes. Egsite discarded the one, two, and three region sites quickly and then located the first four-region solution in 330 seconds of central processing unit (CPU) time on a Sun Sparcstation 2 (compared to ca. 2 months' CPU for Egsets). The solution's most conservative interregion distance intervals and representative interaction parameters are given in Table II. The full description of the eight interaction parameters in terms of 26 linear inequalities has been omitted. Its predictions are given in Table I in terms of calculated G^* and errors E, where of course $E = 0$ for all members of the training set. The best site model we had been able to find with Egsets mispredicted compounds **14**, **19**, and **20**, and it was unable to fit **13** into the site structure at all. Now Egsite's first solution mispredicts compounds **5**, **6**, and **14**, all compounds certainly are accommodated due to the large first region, and the remaining nine predictions all agree with the observed binding intervals but extend beyond them by varying amounts. As an extreme example, Egsite predicts that **13** should bind at least rather well, and possibly extraordinarily tightly; but from this explanation of this training set, it is impossible to be more precise.

In our previous work, Egsets produced literally thousands of relatively narrowly specified site models, and searching through these for ones of high predictive power was a great challenge. Now Egsite produces only 52 solutions for this training set, of which 27 are unique. The other solutions give predictions of quality similar to that presented here. The difference is that Egsite solutions are much more broadly specified and correspond to whole classes of Egsets solutions.

According to the abstract mathematical rules of the game, Table II is indeed a solution, but at this stage of development of the method, a detailed physical interpretation of the result is a dubious undertaking. Although r_2 and r_3 have very small diameters, they can still be occupied by a single superatom apiece. In at least one direction r_4 has a moderate size, whereas r_1 is very large without approaching any of the other regions closely. Technically these distance intervals are not exactly embeddable in three-dimensional space, and introducing such a constraint would constitute a future improvement of the method. On the energetic side, r_1 does not have zero interaction with all types of atoms, leaving the possibility that a large molecule (which could certainly always fit in r_1) would have a predicted extremely favorable or unfavorable binding to the site. Because the experimental results we are trying to fit are generally the measured binding at the site relative to being free in solution, it would be reasonable to introduce a constraint that all ws for r_1 be fixed to zero. Pending these and other improvements, we can at least conclude that Egsite is capable of handling some realistic data sets with modest computational requirements. This abstract approach with interval analysis and ternary logic is not only conceptually intriguing, but also useful.

Acknowledgments

This project was supported by a grant from the National Institute on Drug Abuse (DA06746).

Thanks are due to V. N. Maiorov for many helpful discussions and to Robert Vanderbei for kindly letting us use his LOQO program.

References

1. C. Hansch, P. G. Sammes, J. B. Taylor, and C. A. Ramsden, Eds., *Comprehensive Medicinal CHemistry*, Vol. 4, Pergamon Press, New York, 1990.
2. H. Kubinyi, Ed., *3D QSAR in Drug Design Theory, Methods and Applications*, ESCOM Science Publishers, Leiden, The Netherlands, 1993.
3. W. E. Haefely, *Int. Anesthesiol. Clin.*, **26**, 262 (1988).
4. H. R. Karfunkel, *MATCH*, **19**, 67 (1986).
5. V. E. Kuz'min and S. Krutius, *Khim.-Farm. Zh.*, **20**, 791 (1986).
6. I. Motoc, *MATCH*, **5**, 275 (1979).
7. Y. C. Martin, M. G. Bures, E. A. Danaher, J. DeLazzer, I. Lico, and P. A. Pavlik, *J. Comp.-Aided Molec. Design*, **7**, 83 (1993).
8. C. Humblet and G. R. Marshall, *Drug Devel. Res.*, **1**, 409 (1981).
9. I. Motoc, *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.*, **3**, 43 (1984).
10. I. Motoc, G. R. Marshall, and J. Labanowski, *Z. Naturforsch., A: Phys., Phys. Chem., Kosmophys.*, **40A**, 1121 (1985).
11. R. D. Cramer and J. D. Bunce, *Pharmacochem. Libr.*, **10**, 3 (1987).
12. H. T. A. Cheung, M. S. Searle, J. Feeney, B. Birdsall, and G. C. K. Roberts, *Biochem.*, **25**, 1925 (1986).
13. G. C. K. Roberts, J. Feeney, A. S. V. Burgen, and S. Daluge, *FEBS Lett.* **131**, 85 (1981).
14. A. Ghose and G. Crippen, In J. Dearden, Ed., *Proceedings of the 4th European Symposium on Chemical Structure—Biological Activity: Quantitative Approaches*, Elsevier, Amsterdam, 1983, p. 99.
15. G. Alefeld and J. Herzberger, In *Computer Science and Applied Mathematics*, W. Rheinboldt, Ed., Academic Press, New York, 1983.
16. A. K. Ghose, A. Pritchett, and G. M. Crippen, *J. Comp. Chem.*, **9**, 80 (1988).
17. M. P. Bradley and G. M. Crippen, *J. Med. Chem.*, **36**, 3171 (1993).
18. S. Srivastava and G. M. Crippen, *J. Med. Chem.*, **36**, 3572 (1993).
19. F. Carroll, A. Lewin, P. Abraham, K. Parham, J. Boja, and M. Kuhar, *J. Med. Chem.*, **34**, 883 (1991).
20. F. Carroll, Y. Gao, M. Rahman, P. Abraham, K. Parham, A. Lewin, and J. Boja, *J. Med. Chem.*, **34**, 2719 (1991).
21. A. Lewin, Y. Gao, P. Abraham, J. Boja, M. Kuhar, and F. Carroll, *J. Med. Chem.*, **35**, 135 (1992).
22. P. Abraham, J. Pitner, A. Lewin, J. Boja, M. Kuhar, and F. Carroll, *J. Med. Chem.*, **35**, 141 (1992).