Deducing Molecular Similarity Using Voronoi Binding Sites

Mary Bradley, Wendy Richardson, and Gordon M. Crippen*

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

Received May 24, 1993®

We have devised a new measure of molecular similarity with respect to given simple partitions of space into regions. The similarity is determined by numerical integration of the difference in the optimal interaction between the two molecules and the regions over a large range of interaction parameter values. Compounds differing in empirical formula are differentiated by a single infinite region; *cis/trans* or ortho/meta/para isomers are distinguishable by two adjacent regions that are half-spaces; and stereoisomers require five regions. This can be viewed as a natural classification of isomers. The concept can also be applied to drug binding studies to determine which molecules may bind alike in a given biological receptor and to elucidate a necessary starting geometry when a binding site is modeled for inhibitors whose experimental binding energies are different.

There has been a dramatic increase in interest in molecular similarity in recent years. Much of this interest is related to ligand binding and drug development, where a goal is to find compounds which have a biological activity similar to some known lead compound. There are many different methods for assessing the similarity of two molecules, ranging from molecular superpositioning[1] to calculating electrostatic potential similarities.[2] References 3 and 4 give an overview of the current state of similarity research. If we are interested in applying the concept of molecular similarity to drug binding, we must consider the *environment* of the receptor–ligand interaction, namely, the binding site. Even molecules as seemingly alike as two stereoisomers may not bind with the same affinity in a given binding site, even though they possess the same chemical formulas and physicochemical parameters.

Many methods of determining molecular similarity apply the molecular graph approach in which the atoms of the molecules serve as vertices of the graph and the bonds between atoms are the edges. These methods are useful when structural database searches are implemented for structurally similar molecules, but say little about the molecules' binding ability in a receptor. Although there are methods for assessing the similarity of molecules interacting with a given receptor, these methods either evaluate the effect of the receptor implicitly, as in pharmacophore matching,[5,6] or they require the use of the crystal structure, as in shape complimentarity analysis.[7] With Voronoi site modeling, we can determine molecular similarity with respect to a simple abstract site model that need not correspond to any physical site.

Our goal is to map different chemical structures onto real numbers. If we look at similarity in the context of drug binding, the biological receptor automatically does this in the form of the binding strength of the ligand. We have previously used this concept to model actual binding sites using Voronoi site models and experimentally determined binding energies of known inhibitor molecules.[8-11] To create a Voronoi site, we first divide space into regions which correspond to Voronoi polyhedra by selecting *generating points*, cs in eq 1. The

$$r_i = \{\mathbf{p} \mid \|c_i - \mathbf{p}\| < \|c_j - \mathbf{p}\|, \forall\, j \neq i\} \quad (1)$$

Voronoi polyhedra are then defined as the set of all points, r, which lie closer to one generating point than to any other. These regions are convex, space-filling, and separated by planes. We then define a *binding mode* for a molecule as the way in which it partitions its atoms among these various regions in space. We make no assumptions about the preferred binding modes of the molecules. The binding energies are modeled as a sum of the interactions between the ligand atoms and the regions of the site which they occupy in a particular binding mode. When we apply this method to drug binding, we define these *interaction energy parameters* in terms of physicochemical parameters (e.g. hydrophobicity or molar refractivity) and insist that the calculated energy fall within the error limits of the experimentally determined value. For the purpose of determining similarity, we insist only that the calculated energy value be unique for each molecule. Our Voronoi binding site representations are abstract, and no knowledge of the crystal structure is required. Application of ligand similarity to actual binding studies is a natural extension of this method, since we are most often concerned with modeling inhibitor molecules with similar structures.

METHODS

In order to determine what makes two molecules appear different as they approach a binding site, we made the following assumptions:

(1) Molecular similarity is only defined in a particular site geometry. This is an important point, since molecules may behave quite differently in different types of sites.

(2) For calculation of binding energies, all atomic physicochemical parameters will be assigned in terms of the alphabetic (C, H, O, ...) atom type, each atom type consisting of mutually orthogonal atomic parameter vectors:

$$V_{m,a} = [v_C, v_H, v_N, ..., v_T] \quad (2)$$

where a is an atom in molecule m, and there are T atom types, t. For a particular atom, the appropriate component in V will be 1, and the rest will be zero. This will ensure that the similarity calculations are not biased by a particular assignment of atomic physicochemical parameters. When Voronoi modeling is used for drug binding, the actual physicochemical properties are used. T then becomes the number of such parameters assigned to each atom. These parameters (v_C, v_H, v_N, ...) will henceforth be referred to as the atomic parameters, since they do not represent a true quantitative measure of the physical properties of the atoms.

* Abstract published in *Advance ACS Abstracts*, September 1, 1993.

0095-2338/93/1633-0750$04.00/0

Since we have defined the similarity between molecules in terms of their interaction with a receptor, it is necessary to have a measure of how the molecules are viewed by the receptor. This is accomplished by comparing the highest energy binding modes (eq 4) available to each of the molecules in a site with a given geometry. We then calculate the squared difference in the binding energies of a pair of molecules in that site over a large space of possible site parameters. In the Voronoi model, for each given site geometry the modes available to each molecule, m, and the atomic parameters are known; this leaves only the adjustable energy parameters which are unknown (ϵs in eq 3), where there are A atoms in molecule

$$\Delta G_{m,\text{mode}} = \sum_{a=1}^{A}\sum_{t=1}^{T} v_{a,m,t}\epsilon_{t,r(a)} \tag{3}$$

$$\Delta G_{m,\text{best}} = \max_{\forall\ \text{modes}} \Delta G_{m,\text{mode}} \tag{4}$$

m; $r(a)$ is the region assigned to atom a in a particular binding mode; and $\vec{\epsilon}_{r(a)}$ is the T-dimensional vector $\vec{\epsilon}_{r(a)} = [\epsilon_{C,r}, \epsilon_{H,r}, \epsilon_{N,r}, ..., \epsilon_{T,r}]$ assigned to that region. There may be many geometrically allowed modes for each molecule in a particular site, so we select the binding mode with the highest calculated energy (eq 4) for each molecule. If we then integrate eq 5

$$\int_{\text{finite }\vec{\epsilon}} (\Delta G_{m,\text{best}} - \Delta G_{n,\text{best}})^2\, d\vec{\epsilon} \tag{5}$$

over a very large ϵ space for a pair of molecules, m and n, and determine the difference in the energies of the best mode for each molecule at each sampled ϵ, then the difference of these energies within the finite ϵ space gives a numerical estimation of the degree of similarity between the molecules. The integral may diverge as the region size approaches infinity, but the area of integration must be large enough to allow for the entire tree of possible binding modes to be sampled, so we approximate the infinite space by allowing the range of ϵs sampled to be large. A smaller difference indicates a higher degree of similarity between two molecules. An energy overlap of 0.0 indicates that the molecules are indistinguishable in that particular site.

Exact analytical integration of eq 5 over a large ϵ space is very difficult, so we use a Monte Carlo integration scheme which allows for a reasonably accurate integration in a short time. See the Appendix for details of the numerical integration. This is a multiple integral over the vectors $\vec{\epsilon}_r$ for all regions r.

The Voronoi binding site method ensures that each molecule will achieve the binding mode which corresponds to the greatest possible value of the binding energy attainable in a given site. Therefore, if two molecules have the same number and type of atoms, they will also have the same binding energies if they are able to achieve the same binding modes. In order to distinguish between isomers, the geometry of the site must play the key role. We would like to know what types of site geometries are necessary to distinguish between the various types of geometric isomers and what geometries are required to energetically favor one isomer over the rest.

RESULTS

In this study, we were concerned mainly with geometric isomers. Geometric isomers are defined as two molecules that share the same molecular formula but differ in the bonded arrangement of their atoms. Conversion between geometric isomers requires bonds to be broken and reformed. The isomeric groups we tested are listed in Table I. Each isomeric series is made up of molecules with the same functional groups; compounds having identical atom compositions but different functional groups would be expected to behave differently in the binding site.

One Region. If we consider the set of all organic molecules, it is apparent that empirical formula is a convenient first approximation of the difference between molecules. Differences in binding energies calculated by eq 4 will most likely arise from molecules whose atomic compositions are different. If the physicochemical parameters, $V_{a,m}$, are determined solely by atom label, as in eq 2, then a single region site permits only one binding mode in eq 3 with region r_1 which maps each different empirical formula onto a unique number by choosing

$$\vec{\epsilon}_1 = [1, 10^{-2}, 10^{-4}, ...] \tag{6}$$

for a set of molecules having fewer than 100 atoms of any type. From that starting point, we can separate the molecules that share a common empirical formula into those that differ by geometric placement of atoms and those that differ merely by rotation about a single bond. The latter are usually not different for the purpose of drug binding, since most often the energy barrier for the rotation about a single bond is comparable to kT.

Two Regions. A binding site consisting of two infinite regions separated by a bounding plane is required for the dichloro-substituted benzenes (ortho, meta, and para) to achieve optimal binding modes which are different and which correspond to different binding energies. These modes are illustrated in Figure 1, where the generating points for the two regions are labeled c_1 and c_2. This site geometry is also sufficient for distinguishing the *cis* and *trans* isomers of dichloroethene (Figure 2), but the 1,1-dichloro isomer and the *cis*-1,2-dichloro isomer cannot attain optimal binding modes which differ in energy for any combination of ϵs. A more complex site is required to distinguish all three isomers.

Molecules which differ by the bonded arrangement, or connectivity, of their atoms, such as the ethers shown in Table I, were distinguishable in a two-region site. When the regions are such that

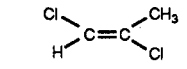
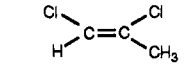
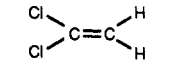
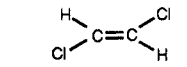
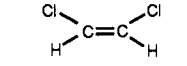
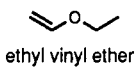
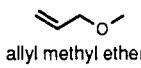
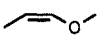
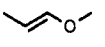
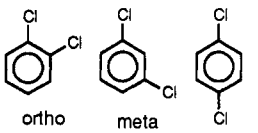
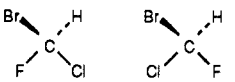
$$\begin{aligned} \epsilon_{C,j}, \epsilon_{H,j} &> \epsilon_{O,j} \\ \epsilon_{O,i} &> \epsilon_{C,i}, \epsilon_{H,i} \end{aligned} \tag{7}$$

where i and j are different regions, then the *cis* and *trans* isomers of 3-methoxy-2-propene are distinguishable by modes analogous to those shown for the ethene isomers in Figure 2. Each of the remaining ethers (tetrahydrofuran, ethyl vinyl ether, and allyl methyl ether) has a unique optimal binding mode (Figure 3) which serves to distinguish it from all of the other isomers in the region described.

Three Regions. The dichloro-substituted ethenes ($C_2H_2Cl_2$), 1,1-dichloroethene, and the *cis*- and *trans*-1,2-dichloroethene isomers, were determined to be different in a three-region site defined by two boundary planes separated by a distance of 3.0 Å. This site is pictured in Figure 4, where the generating points c_1, c_2, and c_3 are collinear; and c_1 and c_3 are 6.0 Å apart. It is easy to see that the Cl–Cl distances for the three isomers (Table II) determine what type of site differentiates the *cis-trans*-1,2- and 1,1-dichloroethenes.

We can define a *space* required in each region for a particular binding mode as the volume occupied by the molecule and nearby relevant boundary planes when the molecule is positioned in that mode. If two molecules are different in a given site, they will generally be different in a more complex site which contains the simpler site (of the same *space*

Table I. Types of Isomerism

cis/trans (E/Z) and 1,1-disubstituted alkene	 (*E*)-1,2-dichloropropene  (*Z*)-1,2-dichloropropene  1,1-dichloroethene  *trans*-1,2-dichloroethene  *cis*-1,2-dichloroethene
geometric arrangement	 ethyl vinyl ether  tetrohydrofuran  allyl methyl ether  *cis*-3-methoxy-2-propene  *trans*-3-methoxy-2-propene
ortho/meta/para	 ortho meta para
stereoisomers	 (*S*) (*R*)

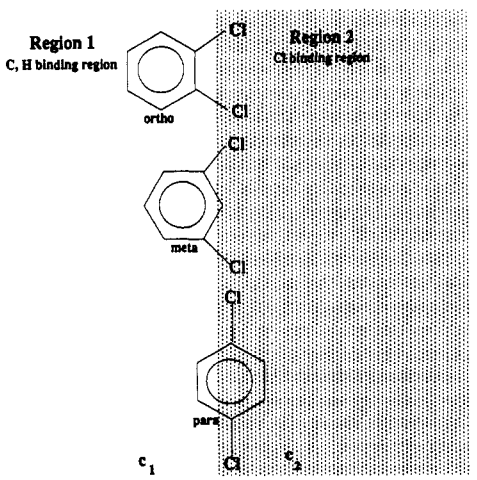


Figure 1. Two-region binding site for dichlorobenzene isomers.

dimensions) as a subset. For example, the three-region site of Figure 4 contains the two-region site of Figure 1 as a subset (regions 1 and 2). Since the two-region site of Figure 1 is a subset of the three region site of Figure 4 in terms of binding mode space dimensions, this three-region site can differentiate all three of the dichloroethene molecules. If the generating points of the three regions were moved such that the bounding

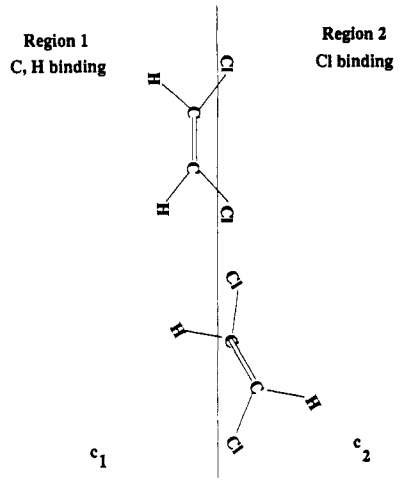


Figure 2. Different binding modes for *cis*- and *trans*-1,2-dichloroethene isomers.

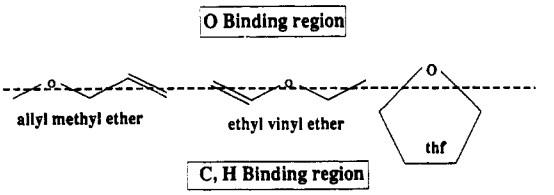


Figure 3. Different binding modes for C_4H_8O ethers in a two-region binding site.

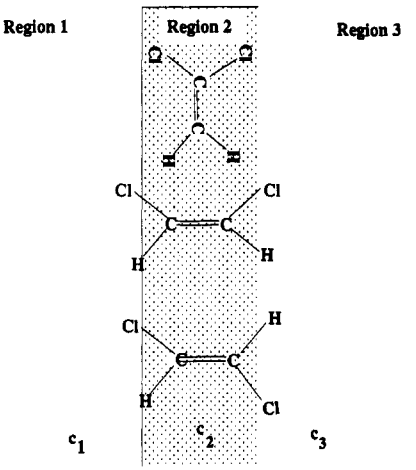


Figure 4. Binding modes for isomers of 1,2-dichloroethene.

Table II. Cl–Cl Distances in the Dichloro Substituted Ethene Isomers

	Cl–Cl distance (Å)
1,1-dichloroethene	2.91
cis-1,2-dichloroethene	3.27
trans-1,2-dichloroethene	4.23

planes were separated by 1.0 Å instead of 3.0 Å, then the *space* dimensions of the two- and three-region sites discussed above would no longer be the same for the dichloroethene isomers. In this event, the modes for the isomers shown in Figure 4 would no longer be possible, and the *cis*-1,2- and 1,1-dichloroethene isomers would not be different.

Five Regions. The chiral molecules in Table I were able to adopt the same conformation in any site which is not also chiral. A five-region chiral site specific to each particular pair of chiral molecules is required for distinguishability. This site, shown in Figure 5, consists of a finite tetrahedral center bounded by four infinite regions (R1–R4), and is made by five generating points which are close to the atomic coordinates

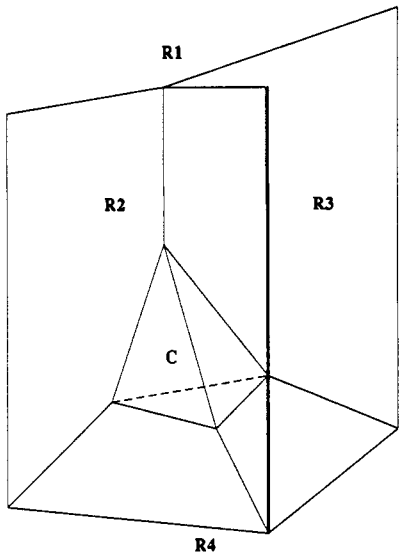


Figure 5. A five-region chiral binding site. Without the central tetrahedron, this would look like the four-region site in Figures 6 and 7.

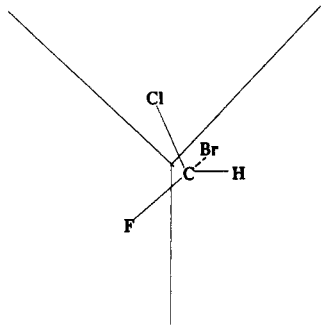


Figure 6. *R*-Fluorochlorobromomethane in a four-region site. The site has four bounding planes, pairs of which meet to form four edges. These four edges meet at one point in the center of the illustration, but in this view one of the edges comes straight out of the page.

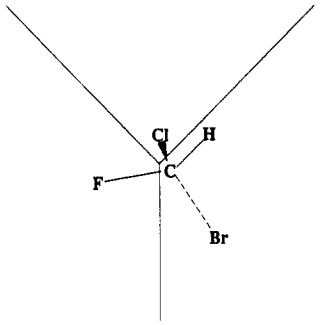


Figure 7. *S*-Fluorochlorobromomethane in a four-region site. Same view as in Figure 6.

of the chiral center and its bound substituents. Intuitively, it may seem that the stereoisomers should be distinguishable in a four-region site in which each substituent on the chiral center is placed in a separate region. However, both enantiomers are able to achieve this mode (Figures 6 and 7), even when the generating points for the four-region site are chosen to be the substituents bound to the chiral center of one of the enantiomers. Note that the three-dimensional site and molecules pictured have been projected into two dimensions, so that the bond lengths and angles may appear somewhat different. Creating a chiral center in the site by adding a finite region at the center allows one isomer to achieve a unique binding mode.

Ordering the Binding Energies. Returning to the dichlorobenzene isomers in the two-region site described above, if

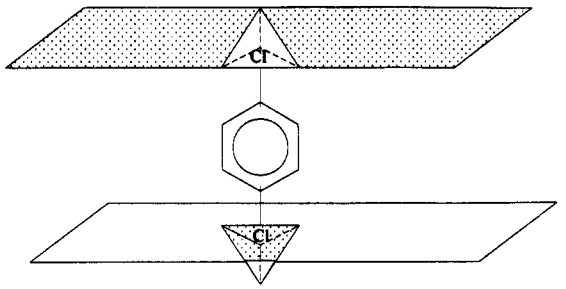


Figure 8. Three region binding site for $\Delta G_{\text{para,best}} > \Delta G_{\text{meta,best}} \geq \Delta G_{\text{ortho,best}}$.

we define an energy penalty such that there is a preferred atom type in each region with $\epsilon_{t,r(a)} > 0$ and all other atom types in that region interact with $\epsilon_{t,r(a)} \leq 0$, then there is no combination of ϵs such that

$$\Delta G_{\text{para,best}} > \Delta G_{\text{meta,best}} \geq \Delta G_{\text{ortho,best}} \tag{8}$$

or

$$\Delta G_{\text{meta,best}} > \Delta G_{\text{ortho,best}} \geq \Delta G_{\text{para,best}} \tag{9}$$

In other words, the ortho isomer will always be able to attain the highest energy binding mode, or, at the very worst, all of the calculated energies of the three isomers will be equal. In order to achieve the greatest binding energy, the molecules will bind such that the greatest number of atoms will be placed in a favorable region. In the dichlorobenzene case, there does not exist a set of parameters such that the ortho isomer will not have the greatest binding energy *in this two-region site* (Figure 1). In order to change the ordinal binding energies of these three isomers, it is necessary to change the binding modes which are available to each of the molecules, so that an energetic solution can be found which satisfies our criterion from eq 4.

The geometry of the site is again more important than the energetics. Not only the number of regions which comprise the site but also their relative size and proximity will change the way in which the molecules can orient themselves in the site. The placement of the atoms in turn affects the calculated binding energy of the molecules, which determines their similarity. We can construct a site with geometry, as pictured in Figure 8 and allow the two finite regions to be separated by a distance large enough so that only the *para* isomer is able to achieve a binding mode such that both Cl atoms are in these regions. We further define the two finite regions to have $\epsilon_{\text{Cl,finite}} > 0$ and $\epsilon_{\text{C,finite}} < 0$, $\epsilon_{\text{H,finite}} < 0$, and the larger center region is defined as having $\epsilon_{\text{Cl,center}} < 0$, then the molecules can achieve binding modes so as to satisfy eq 4.

If we consider the interaction energy parameters for the one-region site given in eq 6, then the molecule which has the greatest (most favorable) binding is always largest (greatest number of C's). To make an intermediate sized molecule the most favorable one, we have to adjust the geometry such that there is an attractive region of finite size, and all other regions are $\epsilon_{t,r} \leq 0$. Otherwise a very large molecule having the correct sort of atom type ratios in its empirical formula will have better calculated binding. This condition must be taken into account when the binding site of an actual biological receptor is modeled. For example, say we have a binding site in which benzo[*a*]pyrene is known to bind strongly. Benzo[*a*]pyrene is quite hydrophobic, so we want to maximize the tendency for a hydrophobic molecule to bind more strongly, while at the same time ensuring that a very large, hydrophobic molecule, such as graphite, does not bind much more strongly. This effect can be achieved by altering the geometry of the site to

Table III. Similarity of Geometric Isomers

type of isomer	minimum no. of regions required for distinguishing
empirical formula	1
cis-/trans-alkene	2
cis-/trans-1,2- and 1,1-disubstituted alkene	3
ortho/meta/para substituted benzene	2
geometric arrangement	2
chiral	5[a]

[a] The site must be constructed such that each atom is attached to the chiral C, and the chiral Cs are close to the generating point for that region.

exclude the binding of molecules of a particular size or shape.

Table III gives the minimum number of regions required for distinguishing between the isomers listed in Table I. Note that all of the molecules within each set of isomers were indistinguishable in a one-region site. This is as expected since the atom types for each set of geometric isomers are identical, thereby giving identical binding energies for the only binding mode available in a one-region site. In general, the more subtle the difference between the isomers, the more complex the site required for distinguishing them.

In a situation where all of the optimal binding modes found are identical for each molecule, there will be no difference between the molecules, as in the case of the two enantiomers in a one-, two-, three-, or four-region site, for example. It is the determination of the most energetically favorable binding mode for each Monte Carlo point that determines whether the molecules will be different in a particular site. The molecules become indistinguishable when there is no difference in their binding energies at every point in the sampled ϵ space. It may also happen that, for a set of three molecules, there are at least two which share the same energy at every sampled point. Then the summed energy differences over the entire sample space may be nonzero for each pair of molecules, yet there will be no combination of ϵs such that all of the three molecules can obtain simultaneously different binding energies.

CONCLUSIONS

We can differentiate geometric isomers by mapping their structures onto unique real numbers. This information may be applied to facilitate a structural database search for molecules structurally similar to some lead compound. If actual physicochemical parameters are assigned to the atoms, then we can use this concept of quantifying similarity to determine a degree of physicochemical similarity or bioisosterism between molecules. Since we also use Voronoi binding sites for modeling receptor sites, similarity measurements of the data set molecules are useful for choosing a necessary starting geometry and for splitting the data set into a training set and a prediction validating set.

ACKNOWLEDGMENT

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

APPENDIX. NUMERICAL INTEGRATION

Evaluating the integral in eq 5 over the space of all ϵs may not yield a finite result, so we integrate each variable over the range $[-E, +E]$ for some positive value of E large enough to show the full range of variations in best binding modes. The

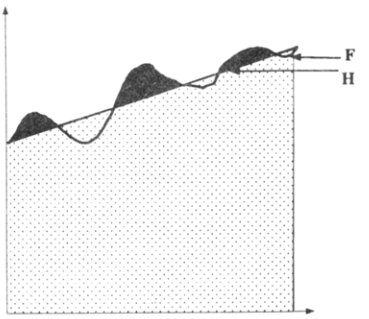


Figure 9. Monte Carlo approximation of the true area under a curve using reduction of variance.

function F (eq 10) cannot be integrated analytically. In

$$F(\vec{\epsilon}) = (\Delta G_{m,\text{best}} - \Delta G_{n,\text{best}})^2 \tag{10}$$

addition, $F(\vec{\epsilon})$ varies so greatly with $\vec{\epsilon}$ that a direct Monte Carlo estimation will converge poorly. To overcome this difficulty, a reduction of variance technique was implemented. We can create a function, H (eq 11), which is a simplification of the function F.

$$H = (h_m - h_n)^2 \tag{11}$$

where

$$h_m = \sum_{a=1}^{A}\sum_{t=1}^{T} v_{a,m,t}\epsilon_{t,\max} \tag{12}$$

$$\epsilon_{t,\max} = \max_{\forall \text{ regions}} \epsilon_{t,r}$$

and a is an atom in molecule m. $\int H$ can be calculated analytically, and $F - H$ has a smaller variance, so the Monte Carlo estimation of $\int(F-H)$ is more accurate.[12] Now we are solving

$$\int_{\text{finite }\vec{\epsilon}} F\,d\vec{\epsilon} = \int_{\text{finite }\vec{\epsilon}} (F-H)\,d\vec{\epsilon} + \int_{\text{finite }\vec{\epsilon}} H\,d\vec{\epsilon} \tag{13}$$

A pictorial representation of this can be seen in Figure 9. The area under the line (H) is added to the darkly shaded area ($F - H$).

The Monte Carlo evaluation of the first term in eq 13 produces

$$\int(F-H) = \langle F-H\rangle v \pm \text{error}$$

$$\text{error} = \left[\frac{\langle (F-H)^2\rangle - (\langle F-H\rangle^2)}{N}\right]^{1/2} \tag{14}$$

Here, N is the number of random points, and v is the volume of the space we are integrating over. $\langle\cdots\rangle$ denotes the mean value. The term under the square root is one standard deviation. This term gives some measure of the error expected from the approximation. Since the error does not necessarily have a Gaussian distribution, the standard deviation may not be the exact error, but it is a sufficiently close approximation.

In the second term of eq 13, the function H can be evaluated exactly, but the *max* function makes it difficult. Notice that the summation of the atomic parameters (remember: we are using mutually orthogonal vectors to represent each atom type in the molecule as a separate physicochemical parameter, and T is the number of total atom types) over all atoms in molecule m is just the molecular parameter.

$$\bar{V}_m = \sum_a V_{a,m} \tag{15}$$

Table IV. Terms and Coefficients To Use When Three Atomic Parameters Are Assigned to Each Molecule

	$V_{m,1}$	$V_{m,2}$	$V_{m,3}$	$V_{n,1}$	$V_{n,2}$	$V_{n,3}$
$V_{m,1}$	a	b	b	$-a$	$-b$	$-b$
$V_{m,2}$	b	a	b	$-b$	$-a$	$-b$
$V_{m,3}$	b	b	a	$-b$	$-b$	$-a$
$V_{n,1}$	$-a$	$-b$	$-b$	a	b	b
$V_{n,2}$	$-b$	$-a$	$-b$	b	a	b
$V_{n,3}$	$-b$	$-b$	$-a$	b	b	a

Table V. Coefficients Needed To Compute the Exact Integral of the Artificial Energy Function

no. of regions	a	b
1	$2^{N_p}/3$	0
2	$2^{N_p}/3$	$2^{N_p}/9$
3	$2^{(3N_p+1)}/(5 \times 3^{N_p})$	$2^{(3N_p-2)}/3^{N_p}$
4	$(7 \times 2^{2N_p})/15$	$(9 \times 2^{2N_p})/25$
5	$(11 \times 2^{5N_p})/(21 \times 5^{N_p})$	$2^{(5N_p+2)}/(9 \times 5^{N_p})$
6	$2^{(5N_p+2)}/(7 \times 3^{N_p})$	$(25 \times 2^{5N_p})/(49 \times 3^{N_p})$

Now eq 12 becomes

$$h_m = \bar{V}_m \cdot \vec{\epsilon}_{\max} \tag{16}$$

The second integral becomes:

$$\int_{-E}^{+E} \cdots \int_{-E}^{+E} H \, d\epsilon_{1,1} \ldots d\epsilon_{r,T} \tag{17}$$

where all of the different ϵs are shown explicitly. There are

$$k = rT \tag{18}$$

variables in the problem, where r is the number of regions in the site and T is the number of atom types. For each atom type we have r associated variables, and only one of these can be the maximum at any given time. The integral can be divided such that there is one case when the first ϵ is the maximal, another for the second, etc. In addition, the ϵs are interchangeable—the solution will remain unchanged if they are renumbered. The problem can be simplified by only evaluating the case where the first ϵ associated with each property is the largest and by multiplying by a factor of r^T to account for every case.

For convenience, the maximal ϵ values are numbered as the first r items. Since the nonmaximal variables must be less than the maximal one, the integral becomes:

$$r^T \int_{-E}^{+E} \cdots \int_{-E}^{+E} \int_{-E}^{\epsilon_1} \cdots \int_{-E}^{\epsilon_1} \cdots \int_{-E}^{\epsilon_T} \cdots \int_{-E}^{\epsilon_T} H \, d\epsilon_{1,1} \ldots d\epsilon_{r,T} \tag{19}$$

The integral was entered into the Maple program and evaluated for a variety of different values of T and r, and the results generalized. In general, the solutions involved one term for every possible pair of molecular parameters. For two molecules with three physicochemical parameters each, there are $6 \times 6 = 36$ different terms in the solution, each term having a coefficient associated with it. For three physicochemical parameters, the final result is

$$\int H \, d\vec{\epsilon} = r^T E^{(rT+2)} (a V_{m,1}{}^2 + b V_{m,1} V_{m,2} + b V_{m,1} V_{m,3} - \ldots)$$

where the a and b coefficients depend on the size of the problem (see Tables IV and V).

The integrals are evaluated at randomly generated points until either a predetermined maximum number of points have been evaluated or the estimated error is within a specified tolerance. If two molecules have identical energy surfaces over the defined space, then the calculated energy difference will be exactly zero. In this event, the estimated error will also be exactly zero.

Since the volume of integration is usually large, the function must be evaluated at many points to get an accurate value, a CPU intensive process. If the minute errors due to numerical rounding and truncation inherent in floating point calculations are multiplied by the very large volume of integration, the resulting error due to floating point notation ceases to be insignificant. This effect is most noticeable when the mean function value is close to zero. To avoid this problem, we do not multiply by the volume, but instead compute the energy difference per unit volume. This may be more meaningful, since it allows for a direct comparison between values obtained using different limits of integration.

In order to assure an accurate representation of difference between molecules, it should be noted that *all* of the binding modes available to each molecule be used in the energy overlap calculation. Say that two molecules—in a given site—have binding modes which are identical except for a single mode. If that single different mode never becomes optimum during the course of the overlap calculation, the two molecules will show a difference of 0.0. A different selection of Monte Carlo points may yield a very different answer if that mode becomes optimum. This problem may be alleviated by using additional Monte Carlo points for integration or by using a more exhaustive type of integration, such as a grid method. The isomers should also all be energy minimized for reasonable bond lengths and dihedral angles by the same method to ensure that small differences in minimization techniques do not create differences in bond lengths and dihedral angles in the groups of isomers where none exist, thereby causing the molecules to appear more distinguishable than they really are.

REFERENCES AND NOTES

(1) Kato, Y.; Inoue, A.; Yamada, M.; Itai, A. Automatic superposition of drug molecules based on their common receptor site. *J. Comput.-Aided Mol. Des.* **1992**, *6* (5), 475–486.

(2) Good, A. C. The calculation of molecular similarity: alternative formulas, data manipulation and graphical display. *J. Mol. Graph.* **1992**, *10*, 144–151.

(3) Johnson, M. A., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.

(4) Livingstone, D. J. Quantitative structure-activity relationships. In *Similarity Models in Organic Chemistry, Biochemistry and Related Fields*; Zalewski, R. I., Krygowski, T. M., Shorter, J., Eds.; Elsevier: Amsterdam, 1991.

(5) Humblet, C.; Marshall, G. R. Pharmacophore identification and receptor mapping. *Annu. Rep. Med. Chem.* **1980**, *15*, 267–276.

(6) Randić, M. Design of molecules with desired properties. In *Concepts and applications of molecular similarity*; Maggiora, G. M., Johnson, M. A., Eds.; Wiley: New York, 1990; pp 77–146.

(7) Desjarlais, R. L.; Sheridan, R. P.; Siebel, G. L.; Dixon, J. S.; Kuntz, I.; Venkataraghavan, R. Using shape complimentarity as an initial screen in designing ligands for a receptor-binding site of known three-dimensional structure. *J. Med. Chem.* **1989**, *31*, 722–729.

(8) Crippen, G. M. Deduction of binding site structure from ligand binding data. *Ann. N. Y. Acad. Sci.* **1984**, *439*, 1–11.

(9) Boulu, L. G.; Crippen, G. M. Voronoi binding site models: Calculation of binding modes and influence of drug binding data accuracy. *J. Comput. Chem.* **1989**, *10* (5), 673–682.

(10) Boulu, L. G.; Crippen, G. M.; Barton, H. A.; Kwon, H.; Marletta, M. A. Voronoi binding site model of a polycyclic aromatic hydrocarbon binding protein. *J. Med. Chem.* **1990**, *33*, 771–775.

(11) Crippen, G. M. Voronoi binding site models. *J. Comput. Chem.* **1987**, *8* (7), 943–955.

(12) Kalos, M. H.; Whitlock, P. *Monte Carlo Methods*; Wiley: New York, 1986; Vol. 1.