Perspectives in Drug Discovery and Design, 1 (1993) 321–328
ESCOM

PD3 015

Why are binding-site models more complicated than molecules?

G.M. Crippen*, M.P. Bradley and W.W. Richardson

College of Pharmacy, University of Michigan, Ann Arbor, MI 48109, U.S.A.

Received 16 July 1993
Accepted 2 August 1993

Key words: Quantitative structure–activity relations; Voronoi diagrams; Receptor-site mapping; Pharmacophore

SUMMARY

A commonly occurring problem in drug development is that the binding affinities for a few compounds to a particular binding site on some protein have been measured, but the crystal structure for that protein is not available. Quantitative structure–activity methods attempt to empirically correlate the binding data with various features of the chemical structures of the drug molecules, so that one can predict the binding of novel compounds and thus aid the search for improved drugs. A common feature of nearly all these methods, however, is that they rely — implicitly or explicitly — on a guess as to the positioning of each molecule when bound to the common site. If one instead assumes that each molecule is free to seek out its optimal positioning in the site, then correlating the observed activity to molecular structure becomes more difficult, and can lead to surprisingly complicated site models. Here we show with some extremely simple artificial examples how this complexity necessarily arises.

INTRODUCTION

Suppose we are given the chemical structures of several compounds, along with their experimentally measured binding affinities to some common binding site on a protein. What can one deduce, from this information alone, about the structure of the binding site and its interactions with the ligands? The traditional approach in the drug design field is to find a least-squares fit of some combination of molecular properties to the observed binding, typically in terms of a great deal fewer adjustable parameters than compounds, in accord with good statistical practice. Such correlations tend to have difficulties accounting for detailed geometric properties of the molecules, such as steric shape, conformationally restricted analogs and stereoisomerism. Furthermore, these methods tacitly assume that for a homologous series of compounds, the corresponding substituents always interact with the same part of the binding site. Not only is it unlikely that

*To whom correspondence should be addressed.

receptors are familiar with IUPAC nomenclature, but there are several well-documented cases of small changes in a ligand's chemical structure resulting in drastic changes in its orientation in the binding site [1].

Over the years we have been developing an alternative approach to explaining binding data in terms of an explicit site model that contains geometric features comparable to those of the ligand molecules [2–6]. The basic idea is that all space is divided into nonoverlapping regions, called Voronoi polyhedra, and that the contribution of any atom to the molecule's free energy of binding depends on the atom type and the region in which it lies. Since there are many different ways to orient the molecule in the site, and hence many different assignments of atoms to regions (called 'binding modes' here), we assume that the molecule chooses the mode that gives it the most favorable interaction with the site. Our sign convention is that larger positive interactions correspond to tighter binding; an interaction value of zero represents the fully solvated, unbound state of the ligand; and negative interactions are unfavorable. This free choice of binding mode in order to maximize its interaction with the site does three things: (i) the model resembles the real ligand–site interaction as the ligand randomly tries different orientations and internal conformations within the site in an equilibrium binding experiment; (ii) the computer calculations map the ligand's molecular structure into a single number, corresponding to the binding of the best mode, instead of many numbers, one for each geometrically feasible mode; and (iii) the site model sometimes becomes much more complex than one would anticipate from traditional quantitative structure–activity relations (QSARs).

While we have applied our Voronoi modeling and its associated computer program, Vorom, to a number of real datasets [7–10], the objective here is to illustrate the necessity sometimes for incredible site complexity, even for ridiculously simple examples. We have chosen these examples because they facilitate visualizing and displaying geometric features seen in real binding data, but they are otherwise completely artificial.

BINDING IN ONE DIMENSION

Instead of worrying about three-dimensional molecules tumbling in space and interacting with complicated binding cavities on proteins, consider a one-dimensional world where linear molecules slide up and down the real number line. Suppose binding depends strictly on the interaction of some chemical group, called 'A', with different portions of the site, and we are given molecule **1**, a short AA dimer, and **2**, a long AA dimer, where the intervening spacer groups play no role in binding. Also given is that the A-to-A separation in **1** is d_1, for **2** it is d_2, and $d_2 > 2d_1$. The experimentally determined binding for each is a range of values, as in standard error bars: for **1** it is $0 < L_1 < U_1$, and for **2** it is $0 < L_2 < U_2$. The exact numerical values of all these parameters are not important. From the point of view of these molecules, the world consists of an infinite line along which they are free to slide, and our task is to divide the line into regions r_1, r_2, ..., such that the optimal binding mode for each has a calculated binding within the experimental range. Whenever one A group is in region r_i, it makes an additive contribution to the binding of ε_i. How many regions are required, and hence how many adjustable ε parameters there are, depends primarily on the relative binding ranges of the short and long dimers.

If the $[L_1, U_1]$ and $[L_2, U_2]$ binding ranges overlap, then only one infinite region, r_1, is required, and this is the entire line. Associated with it is one parameter, ε_1, chosen so that $2\varepsilon_1$ falls in the

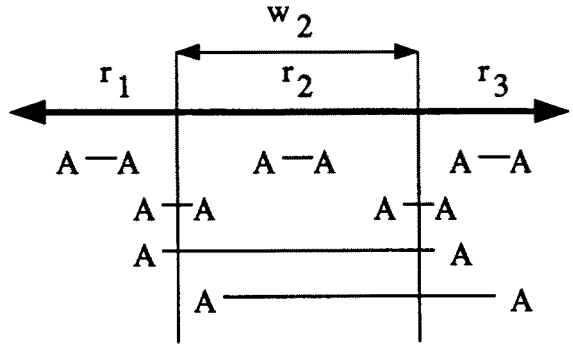

Fig. 1. A three-region, one-dimensional site where $d_1 < w_2 < d_2$. Some of the possible binding modes of the long and short AA isomers are illustrated.

overlap of the two ranges (because the interaction of either molecule with r_1 involves both A groups lying in the region).

Alternatively, suppose the short isomer binds more strongly, i.e. $U_2 < L_1$. A one-region site model fails, because both molecular structures are mapped to the same calculated binding affinity, namely $2\varepsilon_1$. A two-region model amounts to choosing an arbitrary boundary point somewhere on the line and calling the left half-infinite segment r_1 and the right one r_2. Now, depending how we slide one of the molecules up and down the line, we can position different atoms (A groups) in different regions. The binding mode where the first (left) A lies in region r_i and the second A is in r_j is denoted by $[r_i, r_j]$. The set of all geometrically allowed binding modes for **1** is

$$B_1 = \{[1,1],[1,2],[2,2]\} \tag{1}$$

but there is nothing to prevent **2** from achieving the same modes, i.e. $B_2 = B_1$. If we make the first region more favorable ($\varepsilon_1 \geq \varepsilon_2$), both molecules will prefer the first mode and slide to the left of the dividing point, producing the same calculated binding ($2\varepsilon_1$) for both. If r_2 is made more favorable, they both achieve mode [2,2], corresponding to the common calculated binding affinity of $2\varepsilon_2$. A more detailed site geometry is required to differentiate between the two compounds on a geometric basis.

In order to account for the stronger binding of the short isomer, we need to go to the three-region site shown in Fig. 1. If we choose the width w_2 of the finite middle region r_2 to be $d_1 < w_2 < d_2$, then the short molecule **1** can achieve mode [2,2], but **2** cannot, i.e. $B_2 \neq B_1$. If we choose ε_2 such that $L_1 \leq 2\varepsilon_2 \leq U_1$ and then $\varepsilon_1 = \varepsilon_3$ such that $L_2 \leq \varepsilon_1 + \varepsilon_2 \leq U_2$, the optimal binding mode for **1** will be [2,2], the two equally optimal modes of **2** will be [1,2] and [2,3], while the calculated affinities of these optimal modes will be correct. Note that we can sometimes arrange the solution so that more than one mode of a molecule is optimal, corresponding to a less detailed claim about how the molecule is supposed to bind to the site.

Two questions might come to mind about this three-region solution. First, why did we need four adjustable parameters (ε_1, ε_2, ε_3 and w_2) to account for the binding of two molecules? The answer is that we had to fit not only two observed binding ranges, but we also had to exclude or make suboptimal several other binding modes, because the molecules are permitted to seek their

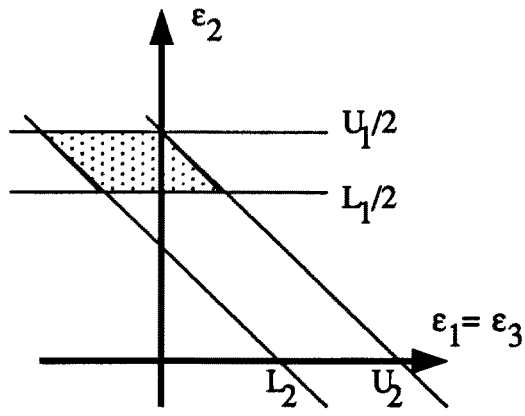

Fig. 2. Interaction-energy parameter solution set (shaded), corresponding to the site geometry of Fig. 1, where the short AA dimer **1** binds more tightly.

most favorable binding modes. For example, if we make the middle region so wide that $w_2 > d_2$, then both molecules will have available the same five modes (illustrated for the short dimer in Fig. 1), $B_2 = B_1$, and no amount of adjusting the values of ε will produce a solution. Alternatively, if we stay with the correct geometry but let, say, ε_1 become large, then both molecules will prefer the [1,1] mode, i.e. both will slide over to the far left infinite region and give the same calculated binding affinity. The second question is: Why is the answer expressed as a set of inequalities, instead of some single value for each variable? The reason is that the solutions are in general one or more patches in the parameter space (the feasible region consists of one or more possibly disjoint polytopes, in the language of linear programming), and these patches may be small, or large, or even open out to infinity in some directions. For example, if $L_2 = 1$ kcal, $U_2 = 2$ kcal, $L_1 = 3$ kcal, $U_1 = 4$ kcal, $d_1 = 2$ Å and $d_2 = 5$ Å, then w_2 can range from 2.1 to 4.9 Å, taking 0.1 Å as the margin for strict inequality. Independent of the w_2 choice, we can pick ε_2 anywhere between 1.5 and 2 kcal. At the lower end of the ε_2 range, ε_1 and ε_3 can be between −0.5 and +0.5 kcal, or at the upper end, between −1 and 0 kcal (see Fig. 2). Clearly, this type of binding-site model gives answers that are qualitatively different from those of linear regression models.

The situation becomes much worse when the long isomer, **2**, binds distinctly better than **1**, i.e. $U_1 < L_2$. Of course, one- and two-region site models fail as before, but now even the three-region model of Fig. 1 fails because the short isomer can always slip into whatever region has the most favorable interactions and always binds at least as well as the long isomer. It is tedious, but possible, to prove that the simplest solutions are two four-region models, one of which is shown in Fig. 3. Its geometry is characterized by region widths $w_2 < d_1$, $w_3 > d_1$ and $d_1 < w_2 + w_3 < d_2$ (the alternative solution has $w_3 < d_1$, $d_1 < w_2 < d_2$ and $w_2 + w_3 > d_2$; mirror images of sites are not counted). The only way this site can achieve a solution is for **2** to have the optimal mode [2,4]. Intuitively, the strategy of this site is to strongly grasp the left end of a molecule in r_2, which positions it rather accurately, due to the small width, and then measure whether the right end can reach across r_3 into r_4. In order to prevent other modes of **2** being better than [2,4], and incidentally to make [4,4] the optimal mode for **1**, we have to make r_1 and r_3 very unfavorable by choosing $\varepsilon_1, \varepsilon_3 \ll 0$, set ε_4 by $L_1 \le 2\varepsilon_4 \le U_1$, and then adjust ε_2 so that $L_2 \le \varepsilon_2 + \varepsilon_4 \le U_2$.

This four-region site certainly accounts for the given binding data, but it has an interesting

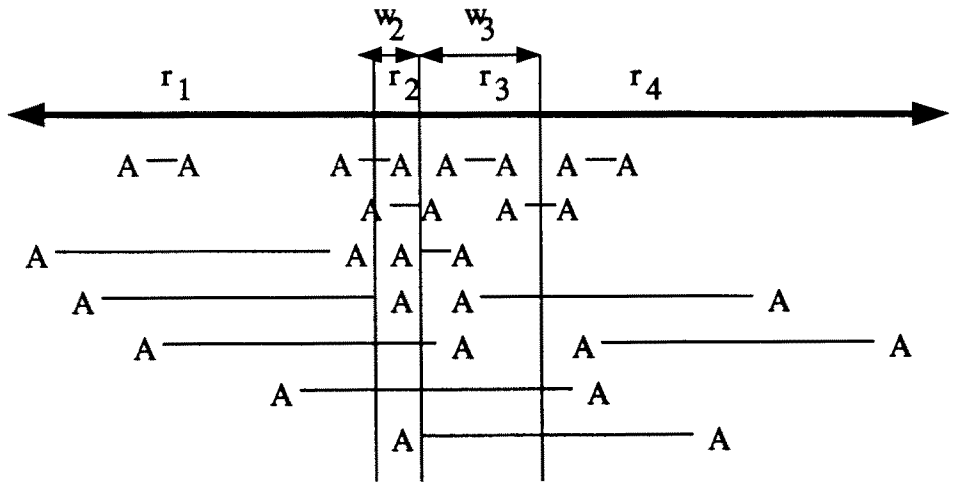

Fig. 3. A four-region, one-dimensional site permitting stronger binding by the long AA dimer than by the short one. All seven possible binding modes for each molecule are illustrated.

failing when it is used to predict the binding of other molecules. Because **1** is supposed to bind to the site ($0 < L_1$), the interaction with r_4 must be at least mildly favorable ($\varepsilon_4 > 0$). Since the site must account for all space, an end region, such as r_4, is necessarily infinite and can therefore accommodate arbitrarily large molecules. Therefore, a long-chain polymer, such as polyA, can always achieve the binding mode [4,4,4,...] with an associated binding energy that becomes arbitrarily large (favorable) as the degree of polymerization increases. One way to fix this unacceptable behavior is to add a fifth region on the right of Fig. 3, positioned so that $w_4 < d_2$ and $w_3 + w_4 > d_2$. Then one can make [2,4] the optimal mode of **2** (as before) by satisfying $\varepsilon_1 = \varepsilon_5 = 0$, $L_2 \le \varepsilon_2 + \varepsilon_4 \le U_2$, $L_1 \le 2\varepsilon_4 \le U_1$, $\varepsilon_2 \le U_1$ and $L_1 \le 2\varepsilon_3 \le U_1$. Now that both infinite regions have zero interaction parameters, thus simulating the solvent, very large molecules will have predicted binding based only on the numbers of atoms or groups they can fit into the finite middle regions. Note that this leaves us with fitting the binding of two molecules, each having only two atoms, in terms of eight parameters (five ε's and three widths)!

DICHLOROETHENE

The advantage of the one-dimensional example is that, even for five regions, there are so few modes for two-atom molecules that one can enumerate all the possibilities, draw pictures of all modes, and prove that certain kinds of site models have the stated properties. Of course, we are really interested in more realistic representations of molecules moving in three dimensions and interacting with somewhat more realistic models of binding sites. This entails more difficult movements to visualize, many more possible binding modes, and much more elaborate calculations to determine whether proposed binding modes are actually geometrically feasible. In our research group, much effort has been devoted to constructing the suite of programs collectively referred to as Vorom, which automatically carry out many of these calculations for real drug-binding problems. It is not surprising that all the general features seen in the one-dimensional example occur again in this much more realistic setting.

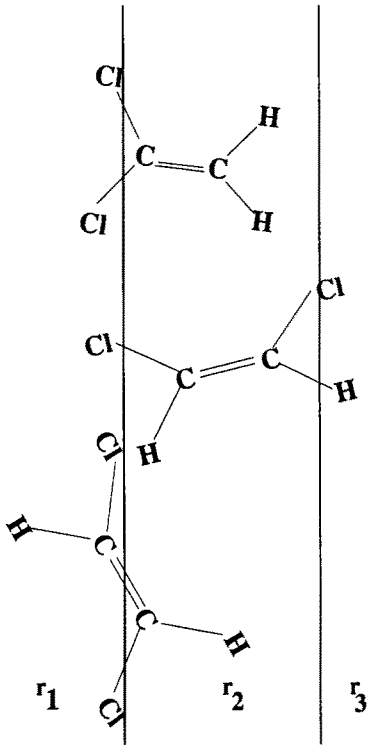

Fig. 4. Optimal binding modes in a three-dimensional three-region 'sandwich' site chosen so that 1,1-dichloroethene binds better than *cis*-1,2-dichloroethene, which in turn binds better than *trans*-1,2-dichloroethene. The two parallel boundary planes are seen edge-on.

Suppose we are given experimental binding data on three compounds: 1,1-dichloroethene (**3**), *cis*-1,2-dichloroethene (**4**), and *trans*-1,2-dichloroethene (**5**). We will view these molecules as being built out of six atoms each, where we can distinguish only three atom types: C, H, and Cl. Any region r_i now has three interaction parameters, denoted by $\varepsilon_{i,C}$, $\varepsilon_{i,H}$ and $\varepsilon_{i,Cl}$. Suppose the three given binding ranges have some common overlap range. Then a one-region site model (all three-dimensional space) is sufficient to map their common empirical formula, $C_2H_2Cl_2$, into that overlap range by the expression $2\varepsilon_{1,C} + 2\varepsilon_{1,H} + 2\varepsilon_{1,Cl}$.

Alternatively, suppose **3** and **4** have overlapping binding ranges, but **5** binds distinctly worse. Now the model requires two regions, namely two half-spaces separated by a dividing plane, which is just the three-dimensional analog of dividing the line by a boundary point. If $\varepsilon_{1,Cl} > \varepsilon_{2,Cl}$ and $\varepsilon_{1,C} < \varepsilon_{2,C}$, then the 1,1 and the cis isomers can put both Cl atoms in r_1 while keeping the rest of the molecule in r_2. The trans isomer, however, can put at most one Cl in r_1 without putting C atoms on that side, too.

Now suppose all three compounds have distinct binding ranges, and the relative binding order is **4** > **3** > **5**. Distinguishing **4** and **5** can be done on the basis of two regions, which may form a part of the required site, but to distinguish **3** from **4** requires a measurement of the Cl–Cl distance. One solution is to construct a three-region site consisting of two half-spaces, r_1 and r_3, sandwiching a thin slice, r_2, between two parallel dividing planes (see Fig. 4). The nine ε's are adjusted so

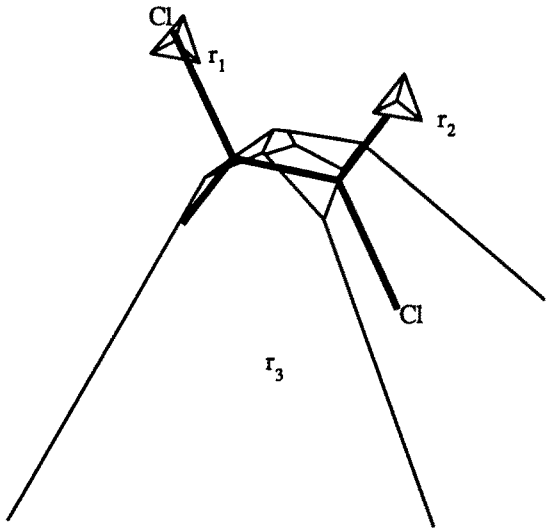

Fig. 5. A 14-region Voronoi binding-site model for dichloroethenes when *trans*-1,2-dichloroethene must bind the most strongly. Only edges forming the binding regions are shown (light lines), along with the trans isomer in its optimal mode (heavy lines).

that r_1 is strongly favorable to Cl, neutral to H and mildly repulsive to C; r_2 is repulsive to Cl and H but favorable to C; and r_3 is favorable to H, neutral to C and repulsive to Cl. The width of r_2 must be chosen so that the two Cl atoms of the cis isomer **4** can span across r_2, but the closer Cl atoms in the 1,1 isomer **3** cannot. Then the optimal binding modes are those in the figure, and the ordering of binding strengths is as required.

Amusingly enough, life becomes much more complicated if the binding order is altered, so that **3** < **4** < **5**. Specifically, suppose the 1,1 isomer has $[L_3,U_3] = [1.0, 2.0]$, cis has $[L_4,U_4] = [3.0, 4.0]$, and trans has $[L_5,U_5] = [5.0, 6.0]$. Instead of simply assigning to each region an interaction parameter for C, H and Cl, we applied our more general physicochemical-parameter technique that is used in receptor modeling studies. Each of the three molecules has assigned to each atom a an atomic contribution [11,12] to that molecule's hydrophobicity $v_{a,hp}$ and molar refractivity $v_{a,mr}$. Then to each region r_i there correspond an adjustable $\varepsilon_{i,hp}$ and $\varepsilon_{i,mr}$ so that placing atom a in r_i contributes $\varepsilon_{i,hp}v_{a,hp} + \varepsilon_{i,mr}v_{a,mr}$. It so happens that the Cl atoms in **3** have different v values than those in **5**, as do the H atoms. The simplest solutions we can find involve three binding regions precisely separated in space, much as the one-dimensional site in Fig. 3 employs r_2 and r_4 with carefully adjusted widths and separation. This one-dimensional example used r_1 and r_3 as energetically repulsive spacer regions, designed to shape r_2 and its separation from r_4. The equivalent operation for the dichloroethenes in three dimensions is much more complicated, resulting in attractive regions r_1, r_2 and r_3, as shown in Fig. 5, shaped by another 11 repulsive regions, which are not shown in the illustration. Figure 5 shows the optimal binding mode for **5**, and those for **3** and **4** are similar. Region r_1 always contains a Cl substituent, r_2 contains an H for **3** and **5** but a Cl for **4**, and the rest of the molecule falls in r_3. Even though **3** and **5** place the same atoms (C, H or Cl) in the same regions, the differences in the v values for the Cl and H atoms in the two molecules are sufficient to allow the site to differentiate between them. The distances, sizes and

shapes of the three regions have to be adjusted to close tolerances, so that only the long C–Cl bond can reach from r_1 to r_3, but both a short C–H and a long C–Cl bond can stretch from r_2 to r_3. While r_3 is open-ended toward the bottom of the illustration, its top is closely adjusted so that the molecules have little freedom to move without hitting the repulsive walls. There are a wide range of ε's in the solution polytope, so that adjusting them is not difficult.

CONCLUSION

Allowing molecules in the computer to explore the different binding modes available to them, given the site-model's geometry, is a touch of realism that forces a conceptual advance in QSAR. The penalty is that much more elaborate calculations are required than in traditional methods, but the reward is that the calculations in effect wring much more information out of each molecule in the training set. The result is sometimes a necessarily remarkably detailed model for the binding site.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Institutes of Health (GM-37123) and the National Institute of Drug Abuse (DA-06746). W.W.R. was an E. Mead Johnson Memorial AFPE Fellow.

REFERENCES

1 Roberts, G.C.K., Feeney, J., Burgen, A.S.V. and Daluge, S., FEBS Lett., 131 (1981) 85.
2 Crippen, G.M., J. Med. Chem., 22 (1979) 988.
3 Ghose, A.K. and Crippen, G.M., J. Med. Chem., 28 (1985) 333.
4 Crippen, G.M., J. Comput. Chem., 8 (1987) 943.
5 Crippen, G.M. and Havel, T.F., Distance Geometry and Molecular Conformation, Research Studies Press Ltd. (Wiley), 1988.
6 Ghose, A.K. and Crippen, G.M., In Ramsden, C. (Ed.) Comprehensive Medicinal Chemistry: The Rational Design, Mechanistic Study, and Therapeutic Application of Chemical Compounds, Vol. 4, Pergamon Press, Oxford, 1990, pp. 715–733.
7 Boulu, L.G. and Crippen, G.M., J. Comput. Chem., 10 (1989) 673.
8 Boulu, L.G., Crippen, G.M., Barton, H.A., Kwon, H. and Marletta, M.A., J. Med. Chem., 33 (1990) 771.
9 Bradley, M.P. and Crippen, G.M., J. Med. Chem., 36 (1993) 3171.
10 Srivastava, S. and Crippen, G.M., J. Med. Chem., 1993, in press.
11 Ghose, A.K., Pritchett, A. and Crippen, G.M., J. Comput. Chem., 9 (1988) 80.
12 Viswanadhan, V.N., Ghose, A.K., Revankar, G.R. and Robins, R.K., J. Chem. Inf. Comput. Sci., 29 (1989) 163.