Analysis of cocaine receptor site ligand binding by three-dimensional Voronoi site modeling approach

ARTICLE *in* JOURNAL OF MEDICINAL CHEMISTRY · DECEMBER 1993

Impact Factor: 5.45 · DOI: 10.1021/jm00075a012 · Source: PubMed

2 AUTHORS, INCLUDING:

Gordon Crippen
University of Michigan
150 PUBLICATIONS **5,893** CITATIONS

SEE PROFILE

Analysis of Cocaine Receptor Site Ligand Binding by Three-Dimensional Voronoi Site Modeling Approach

Sanjay Srivastava and Gordon M. Crippen*

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

Received May 3, 1993[⊕]

The Voronoi approach has been used to obtain a three-dimensional model for the binding of the cocaine analogues at the cocaine receptor site. The method has been used to determine the geometric details and the physicochemical properties of the binding regions in the receptor site. With only eight compounds in the training set, the Voronoi site model, consisting of four regions, not only fully explains the binding affinity of the input compounds but is also successful in correctly predicting another eight compounds of the test set. The phenyl substituent at the 3-position of the tropane ring of cocaine was found to be the most significant functionality relevant for activity, while moderate contribution results from the hydrophobic interactions of the tropane ring with the binding regions. Some of the problems associated with the approach are discussed, and we report a new procedure for evaluating the validity of the model obtained from our approach.

Introduction

Natural (−)-cocaine has been known to possess several physiological properties. It is a local anesthetic and good vasoconstrictant and affects the sympathetic nervous system which leads to increased heart rate and blood pressure in some instances. Investigations[1-3] have shown that cocaine has many sites of action in the central nervous system. But these sites of action have not been well characterized. The most commonly believed site is the so-called dopamine transport inhibition where the resulting buildup of dopamine leads to the reinforcing effects of cocaine.[4] Even the dopamine transporter has been shown to possess two binding sites for (−)-cocaine, although only one of them has a very high affinity for (−)-cocaine.[5]

In order to understand the binding of cocaine at its receptor site and to design compounds subsequently which can alleviate this effect, a number of cocaine analogues have been synthesized and tested for their inhibitory potencies at these receptor sites.[6-9] It has been shown by Carroll et al.[6] that the cocaine binding is very stereospecific and the seven other stereoisomers of cocaine are less potent than the original isomer. But they synthesized several other analogues by modifying the structural features of the substituents at the 2- and 3-positions as well as the N atom in the tropane ring of cocaine.[5] Their findings showed that the modifications of the methyl group of the 2-carbomethoxy functionality exhibited only small changes in the potency. The replacement of the benzoyl functionality at the 3-position of (−)-cocaine by substituted phenyl groups led to marked enhancement in the activity over that of the natural isomer. Changes in the potency were again very marginal when the substituents at the N-position were varied. But a nitrogen group with a basic character was found to be essential for the activity.

A number of studies have correlated the inhibition of cocaine to the selectivity of transporters in the CNS system like the dopamine, norepinephrine, and serotonin transporters.[10,11] But not much insight has been provided into the structural features of ligands relevant for the inhibition of these systems. Carroll et al.[7] have recently performed both a Hansch type QSAR analysis as well as CoMFA evaluation with some of the cocaine analogues mentioned above. The classical QSAR failed to yield any meaningful results, but the CoMFA study with 12 compounds indicated high electrostatic and steric correlation. They were able to identify small regions around the substituents at the 2- and 3-positions of the tropane ring of (*R*)-cocaine as favorable and unfavorable regions for potency. This lends further support to the stereospecific nature of the receptor site.

We have been developing a novel approach called the Voronoi site modeling, for generating a 3-dimensional site model of a receptor enzyme, given the chemical structures and the binding energies of several ligands.[12,13] This approach deduces the geometry and the energetics of the site model. The method partitions space into distinct Voronoi regions which are convex in shape. The sizes, shapes, and positions of these regions are determined by the coordinates of a single *generating point* contained in each region.[12] The ligand molecules are then allowed to partition themselves in these regions such that there is an absolute fit between the calculated binding energy ($\Delta G_{m,\mathrm{calc}}$) for molecule m, and experimental binding energy, which is represented as the upper (ΔG_{m+}) and lower (ΔG_{m-}) bounds of the observed binding energy. The two bounds are introduced to take into account the experimental error involved in the measurement of these activity values. Thus,

$$\Delta G_{m-} \leq \Delta G_{m,\mathrm{calc}} \leq \Delta G_{m+} \forall m \quad (1)$$

The method does not use the bias that all the active compounds could be aligned in such a way that there is a necessary overlap of the common atoms/groups at the active site, an assumption implicit in most QSAR/modeling approaches including the CoMFA methodology. There are no outliers, and the model can be used to predict the preferred binding mode and the binding affinity of any compound not necessarily related to the training compounds. The term binding mode, in our discussion, refers to the assignment of each ligand atom to lie somewhere in a stated region. The description of the Voronoi method along with its improvements over our similar and older 3-dimensional distance geometry QSAR approach can be found in refs 12 and 14.

In this paper, we have addressed two problems which we faced with Voronoi modeling in the past. Firstly, the task of choosing the correct number of regions for a Voronoi

[⊕] Abstract published in *Advance ACS Abstracts*, October 1, 1993.

0022-2623/93/1836-3572$04.00/0

binding site model and guessing the coordinates of the corresponding generating points is rather difficult, so we introduce an automatic procedure for achieving this. Secondly, our method is capable of giving more than one solution for the site model, which is certainly appropriate, but one must have some criteria for choosing one over the others. However, due to the absolute fitting of the experimental data, we cannot use the standard tests of statistical significance, which are based on goodness of fit. Consequently, we report here a protocol which we have devised for examining the predictive power of the model itself for its validation.

The Voronoi approach has been used here to study and understand the binding characteristics of cocaine and its analogues at the cocaine receptor site on the dopamine transport system. For our work, we have made a key assumption that all the cocaine analogues bind to the same receptor site.

Method

The Voronoi site modeling employs the following steps: (1) Generate the 3-dimensional structure of the given ligands using some molecular modeling package. (2) Assign the physicochemical properties of the atoms in the molecules.[15] (3) Represent each molecule in linearized form[16,17] and generate all the conformers for the molecules by a systematic search in torsion angles. (4) Simplify subsequent calculations by reducing groups of atoms together into pseudoatoms. For example, a methyl group could be simply *squashed* into a single "point" which is assigned the mean coordinates of its constituent atoms' positions. The physicochemical parameters of such a pseudoatom are just a sum of its constituent atomic parameters. The implicit assumption is that we are not interested in the internal structure of the group, but rather how it is going to position itself in the site and what is its total contribution to the binding energy as a result of its interaction with some region in the site. (5) Propose a Voronoi site model to represent the actual receptor site which binds to the ligands. This involves making a correct guess for the number of binding regions in the site and choosing appropriate generating points for each region, such that certain desired atoms of the ligand fall into desired regions. This criterion of placing ligand atoms into certain regions is determined by the investigator's knowledge. (6) Determine all the geometrically allowed binding modes of all the molecules. (7) Calculate the interaction energy parameters for the regions of the proposed site, such that the calculated binding energy lies within the observed binding energy range (see eq 1). For the present work we have considered only two physicochemical energy parameters: the hydrophobicity and the molar refractivity.

The binding modes generated in step 6 are explored until for some set of binding modes not only is eq 1 satisfied but also $\Delta G_{m,\mathrm{calc}}$ is the maximum over all the geometrically allowed modes for the molecule m. If one fails in this step, then steps 5–7 are repeated, which means a different site geometry is proposed and the rest of the process is carried out again.

Our experience has shown that guessing a correct site geometry for which the above mentioned criteria are satisfied is extremely difficult. A possible alternative to this problem is to relax the full geometric reality of Vorom at least temporarily in order to deal with the combinatorics of finding the site geometry automatically. This is achieved by a program named Egsets (*e*nergetically and *g*eometrically valid *sets* of solutions). This method is briefly described below (see ref 18 for a more detailed explanation).

Given a molecule having n_a atoms that binds in n_b Voronoi binding regions, and these are necessarily convex regions, then a binding mode amounts to partitioning a molecule into n_b mutually exclusive and exhaustive convex sets of atoms, where the convexity of a set of atoms is currently assessed for one fixed conformation of the molecule. As long as n_b is small, the number of partitions is not all that great, many fewer than $n_b{}^{n_a}$. Assigning regions in various permutations to the sets amounts to generating possible binding modes. Then each mode implies restrictions on the geometry of the binding site, so the choice of optimal modes for all the molecules in the training set must make a consistent set of demands on the site geometry. If a set of proposed optimal binding modes is consistent with some site geometry, then the program attempts to find interaction energy parameters for each region such that eq 1 is satisfied and also so that all alternative modes have a worse binding energy. Without going into a lot of detail, the output is generally many sets of site geometry and interaction energy parameters, each of which is a solution to the relaxed-geometry problem described.

The geometry of a Voronoi site model, consisting of binding and nonbinding regions, is fully specified by the coordinates of the generating points. Egsets speeds up its search by instead describing the complex of Voronoi polyhedra only by the upper and lower bounds on distances within and between binding regions. Nonbinding regions are not used at all in order to keep the problem simple. However, nonbinding regions represent sterically disallowed regions and are generally required around the binding regions so as to give proper shape to the latter. But it is not always possible to find coordinates for generating points of binding and some extra nonbinding regions such that these bounds are satisfied, and such that the purportedly optimal binding modes are indeed optimal. For now, we produce 3-dimensional site models by a process resembling ensemble distance geometry, but more rigorous treatment is being sought.

Results and Discussion

The data for the QSAR Voronoi site study was taken from the work by Carroll et al. We selected all the 8 stereoisomers of cocaine[6] along with 12 other analogues which differed in the structural features at the 2-, 3-, and N-substituents of the tropane ring of cocaine.[7-9] The IC_{50}s of all these 20 compounds had been measured against the competitive binding of [^{3}H]WIN-35,428 to rat striatal membranes. In all that follows we assume these IC_{50}s are indeed due to the relative binding affinities of these compounds at the same site on the same receptor protein. Most of the other compounds which were reported had been evaluated for inhibition of different radioligand binding at the dopamine transporter. On the assumption that different parts of a ligand make approximately additive contributions to the free energy of binding, we converted the experimental IC_{50} values to a logarithmic scale,

$$\Delta G_{\mathrm{obs}} = -\log(\mathrm{IC}_{50}) \qquad (2)$$

realizing that our ΔG_{obs} is only approximately proportional to the true Gibbs' free energy of binding.

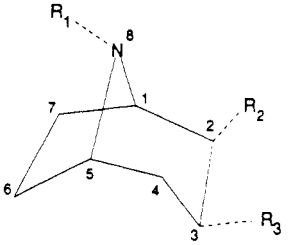
Since our approach does not simply seek a least-squares fit to the ΔG_{obs} values, but rather an absolute fit to a given range of binding for each compound, we needed some way to estimate the accuracy of the binding assay results. While it is true that a particular worker in a given laboratory may achieve a high degree of reproducibility following a particular experimental protocol, we note some considerable variation between research groups attempting to carry out the same experiment. For example, (*R*)-allopseudococaine had activity reported as 28.5 and 5.0 μM,[6,19] while (*S*)-cocaine was reported as 15.8 and 30.0 μM.[6,20] These yield deviations of 82% and 90%, respectively, on the measured IC_{50} scale. Although these values had been obtained by different groups, their analysis was carried out against the binding of the same [^{3}H]WIN-35,428. Therefore, we assumed an 85% error bar, even though we used data coming only from one laboratory. Thus,

$$\Delta G_{+} = -\log_{10}(1.85 \times \mathrm{IC}_{50})$$

$$\Delta G_{-} = -\log_{10}(0.15 \times \mathrm{IC}_{50}) \qquad (3)$$

The crystal structure of (−)-cocaine was used for (*R*)-cocaine, while all the remaining compounds were modeled

Table I. Potencies of Cocaine Analogues for Inhibition of the Binding of [^{3}H]WIN-35,428 at the Dopamine Transporter

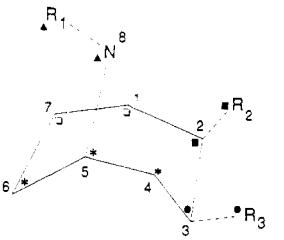


compd	config	common name	R_1	R_2	R_3	IC_{50}[a] (μM)
1	*R*	cocaine (C)	Me	β-CO_2Me	β-O(CO)Ph	0.102
2	*S*	cocaine	Me	β-CO_2Me	β-O(CO)Ph	15.8
3	*R*	pseudo-C	Me	α-CO_2Me	β-O(CO)Ph	15.8
4	*S*	pseudo-C	Me	α-CO_2Me	β-O(CO)Ph	22.5
5	*R*	allo-C	Me	β-CO_2Me	α-O(CO)Ph	6.16
6	*S*	allo-C	Me	β-CO_2Me	α-O(CO)Ph	9.82
7	*R*	allopseudo-C	Me	α-CO_2Me	α-O(CO)Ph	28.5
8	*S*	allopseudo-C	Me	α-CO_2Me	α-O(CO)Ph	67.7
9	*R*		Me	β-CO_2Ph	β-O(CO)Ph	0.112
10	*R*		Me	β-$CO_2CH_2CH_2Ph$	β-O(CO)Ph	0.248
11	*R*	tropa-C	Me	H	β-O(CO)Ph	5.18
12	*R*	benzoylecgonine	Me	β-CO_2H	β-O(CO)Ph	195.0
13	*R*		Me	β-$CO_2(Me)_2$-(*p*-NH_2Ph)	β-O(CO)Ph	0.072
14	*R*		Me	β-CH_2OH	β-O(CO)Ph	0.561
15	*R*		CH_2Ph	β-CO_2Me	β-O(CO)Ph	0.668
16	R		H	β-CO_2Me	β-O(CO)Ph	0.303
17	*R*		Me	β-CO_2Me	β-Ph	0.023
18	R		Me	β-CO_2Me	β-Ph-*p*-F	0.016
19	R		Me	β-CO_2Me	β-Ph-*p*-NH_2	0.025
20	*R*		Me	β-CO_2Me	β-Ph-*p*-OMe	0.008

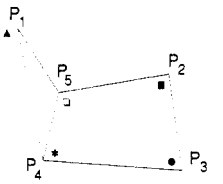
[a] Data from refs 6–9.

through the Quanta program by appropriate modifications of the (*R*)-cocaine structure. It was first necessary to determine the conformation of the tropane ring. A search of the Cambridge Structural Database was conducted in which a total of 79 structures were found having a tropane ring. Of these, 73 were found to exist in the chair conformation. The six remaining structures had either a planar or a boat conformation of the tropane ring, but only one out of six structures was found to resemble the cocaine structure. The bulky substituent in this compound was probably responsible for forcing this nonchair conformation. Moreover, semiempirical and molecular dynamics studies by other workers have also shown evidence for the predominance of the chair form of the tropane ring for cocaine and its diastereomers.[21] The same calculations also established the equatorial position of the methyl group on the nitrogen atom. Hence, we decided to consider only the chair form and equatorial methyl for our studies. Otherwise, the molecules were permitted free rotation of their substituents in a systematic search over all substituent single bonds. The results of this search for each molecule were summarized in terms of the greatest and least observed value for every interatomic distance, taken over all sterically allowed conformations.

The next step involved the simplification of the data set. We squashed each molecule into five pseudoatoms, combining atoms and substituents at positions 1 and 7; 4, 5, and 6; 2; 3; and 8 (see Figure 1). In doing so, we maintained the stereochemistry of the tropane ring with respect to the bridge and assigned individual atom identity to the three main functional groups in the molecule (R_1, R_2, and R_3 in Figure 1). The rest of the tropane ring was broken up into two fragments. Because the coordinates of a united atom are simply the centroid of the constituent atoms, α substitution is still distinguishable from β, and so on. The squashing procedure[18] furthermore preserves the conformational flexibility of the molecules by setting the lower and upper distance bounds between two



(a) Structure of Cocaine analogue



(b) Squashed structure of Cocaine

Figure 1. General structure and squashed structure of the cocaine analogues. The atom positions marked by a common symbol in a have been squashed together into a common pseudoatom in b. The coordinates of the pseudoatom are the mean of the coordinates of its constituent atoms.

pseudoatoms as the least lower bound and greatest upper bound between an atom in one pseudoatom and an atom in the other pseudoatom, respectively. Even with 5 pseudoatoms per molecule, it was not possible to use all 20 molecules in the training set, as the estimated CPU time was several weeks to search through all the possible combinations of binding modes. Instead, we began with a very small training set of four compounds and sought solutions, i.e. site geometry and interaction energy parameters, having high predictive power. Egsets can generate numerous solutions from a very small number of input molecules, but while each one certainly satisfies the training set compounds, most of the resulting site models are not able to predict the binding of test compounds. Thus, it is clear that the solution which yields the maximum

prediction would most likely be the closest to the real model of the receptor site. The following heuristic approach was adopted for locating significantly predictive solutions.

Heuristic Validation Procedure

A model is expected to perform at least somewhat better with an input of additional information. Therefore, we made the requirement that the predictive power of the model should either increase or remain unchanged with addition of extra molecules to the training set. In particular, this should hold when the worst predicted molecule from the test set is added to the training set. For each training set several solutions were obtained for which the predictive power was tested using the remaining test molecules. The solutions from different training sets could be arranged in what we call a *solution tree*. Each level of this tree corresponds to a particular training set, and all the *nodes* on that level are its solutions. The branches or *children* underneath each node correspond to solutions of training sets consisting of an additional molecule. These children have the same optimal binding mode as their parent node for all the molecules common to them.

It was clear that not all the solutions would converge to a good model. The success of any single solution in exhibiting a good prediction could easily be attributed to the usual chance correlation, so common in any QSAR study. Thus, the only way to have confidence in a particular solution, or more specifically in a set of proposed optimal binding modes, is to trace its predictive power as more molecules are added to refine the model. The model could be considered sound only if the predictive power never drastically drops. Naturally, the site geometry will alter as well as the interaction energy parameters, but then again, we do expect them to be refined as we add molecules to our training set. Eventually, after we have considered enough diverse molecules in the training set, we should end up with a final site geometry and energy values which should not change significantly with addition of extra molecules. Thus our protocol was to begin with a small training set and find a large number of solutions distinguished by which binding mode was the optimal for each molecule. For each of those solutions, we add the worst predicted compound to the training set and find once again many solutions. One of these new solutions is deleted if the predictive power has been consistently decreasing with increasing training set size, or if its predictive power is markedly worse than the alternative solutions at that stage. More specifically, we used the following two empirical rules, applied independently of each other:

I. Test the condition: $(1/n)\sum_{i=1}^{n} P_{x_i} - P_a \leq \delta$, where P_a and P_{x_i} refer to the predictions of any particular parent node and its ith child x_i, respectively, in the solution tree (thus, node a contains n branches here). δ is the error allowed in the computation (in our case it is 0.1). If the above condition is satisfied, then those solutions of the second training set (consisting of x_i's) are retained for which the predictive power (P_{x_i}) is either better or at least the same as that of its parent solution (P_a).

II. Test the condition: $P_{x_i} - P_a \leq \delta$, and collect all the x_i that meet this criterion for each parent node a. Retain only that half of these solutions which have now higher predictive power than the median of the set.

The predictive power (P) for a particular solution (a or x_i) is calculated by the following relations:

$$E_m = \frac{|\Delta G_{m,\mathrm{obs}} - \Delta G_{m,\mathrm{calc}}|}{\Delta G_{m,\mathrm{obs}}} \tag{4}$$

$$\chi_m = \begin{cases} 0.4E_m & \text{for } E_m \leq 5 \\ 0.2E_m + 1 & \text{for } 5 < E_m \leq 20 \\ 5E_m & \text{for } E_m > 20 \end{cases} \tag{5}$$

$$P = 100\left(\frac{\sum_{m=1}^{N} \chi_m^{\,2}}{5^2 N}\right)^{1/2} \tag{6}$$

Here E_m is defined as the relative error in prediction between the observed binding energy (ΔG_{obs}) and the calculated binding energy (ΔG_{calc}) of a test molecule m. χ_m is the scaled prediction error of the mth test molecule contained in a test set which has a total of N molecules. The exact choice of definition of χ_m is somewhat arbitrary, but the intent is to be sensitive to small relative errors, not overly penalize moderate errors, but finally emphasize large ones. P is the prediction power for a particular solution. Note that a lower value of P denotes a better prediction because this number represents the fraction of incorrect predictions in the test set.

Further, if a sufficiently large number of training sets can be created, then rule I is tested among three generations simultaneously. In such a case the solutions are discarded only if the condition in rule I is not met in two successive generation comparisons. This ensures that a solution may not be thrown away just because it fares a little worse in any one particular level.

It was important that we test the usefulness of this procedure by doing a full search of all the possible binding modes that a set of molecules could achieve. This was not possible with the cocaine dataset due to problems described above. Consequently, we decided to construct an artificial dataset with substituted ethanes and ethylenes. We considered a total of eight compounds: unsubstituted ethylene, monochloro- and *cis*- and *trans*-1,2-dichloroethylenes, ethanol, ethylamine, 2-aminoethanol, and chloroethane. These compounds were assigned the physicochemical values as usual. They were then squashed into four atoms each, with the two central carbons along with all but one hydrogen being labeled as two pseudoatoms and the substituents on the carbons (either a hydrogen or a heavy atom) being the other two pseudoatoms. A model having two regions was constructed by the Vorom program[14] using a training set of only two molecules, which were assigned some arbitrary activity values. This model was then used to calculate the binding energy of the remaining compounds. The purpose behind this was to determine some activity values for this artificial test which were related to some arbitrary receptor site. It does not matter whether such a site really exists or not. Egsets was used to generate solutions for training sets containing one to four molecules. These solutions were compared and evaluated in the manner described above. The test set used for determining the predictive power of each of these solutions consisted of the remaining molecules not used in any of the training sets. Table II lists a brief summary of the results obtained from this exercise. A total of 7, 54, 155, and 28 solutions were obtained for the four training sets, respectively, for a complete search of all the possible binding modes which are not only geometrically feasible

Table II. Number of Solutions Generated or Left for Each Training Set of Substituted Ethanes, under Different Validation Rules[a]

	number of molecules in the training set			
	1	2	3	4
total solutions	7	54	155	28
solutions left after:				
rule I		42 (77%)	84 (54%)	7 (25%)
rule II		19 (35%)	19 (12%)	1 (4%)

[a] See text for explanation of rules.

but also yield binding energies which are in agreement with the assigned values for the molecules. In the four molecule training set case, only one solution was found to be valid under rule II and seven solutions under rule I, which also included the surviving solution under the former rule. A close inspection of the predictability of all the 28 solutions showed that there is only one other nonsurviving solution which is slightly better than the solution left under the rule II. This shows that our procedure for determining the best solution can be successful. Both the rules should be tried depending on the dataset being considered. Rule II performs a more drastic reduction on the total solutions and hence could be used if there are only a few molecules available. But if several training sets can be created of different sizes, then rule I would do a better and more thorough job of locating solutions with high and persisting increase in prediction power because it is more conservative in discarding solutions. For example, this rule retained seven solutions while rule II ended up with a solitary solution.

Finally, our procedure is certainly not very rigid in that it attempts to locate only near-optimal solutions and not the optimal solution itself. As long as the near-optimal solutions are not very different from the optimal solution (which was found to be true in the datasets we have examined) and the investigator is satisfied with an answer close enough to the optimal solution, our exercise has justification. Thus, we are quite aware that we might lose the optimal solution hidden somewhere while pruning the solution tree by our method, but we reach our objective when we obtain some other solutions which are close enough to this optimal solution. This was certainly found to be true in the ethane dataset that we investigated here. The single solution surviving under rule II had a predictive power of 52.5% while the maximum prediction (best solution) from the exhaustive search was around 50%. The number representing the predictive power stands for the proportion of molecules incorrectly predicted. Thus, a predictive value x% means that our of 100 test molecules about x were incorrectly predicted by the model.

Analysis of Cocaine Analogues

We started with four of the stereoisomers of cocaine (**1**, **2**, **4**, and **17**) in the training set. While the choice is somewhat arbitrary, and not claimed to be optimal, a good policy is to choose molecules where small changes in chemical structure correspond to large changes in activity. Because the binding of **2** and **4** are much worse than that of **1**, due to changes in stereochemistry, the practically inactive **2** and **4** tell us a great deal about the geometry of the site. There must be specific features preventing them from binding as well as **1** does. On the other hand, deleting the carboxyl in **1** to produce the more active **17** tells us something important about the energetic preferences of regions involved in binding R_3.

Table III. Validation Results of the Solutions Obtained from Egsets Program for Training Sets of Cocaine Analogues, under Different Validation Rules[a]

	training set (compounds)			
	A (1, 4, 17, 2)	B (A + 3)	C (B + 7)	D (C + 12, 15)
total solutions	120	154	108	214
solutions left after				
rule I		82 (53%)	71 (65%)	36 (17%)
rule II		49 (32%)	10 (9%)	1 (0.5%)

[a] Training set consists of compounds **1**, **4**, **17**, **2**; **3**, **7**, **12** and **15**. Test set consists of **5**, **6**, **8**, **9**, **10**, **11**, **13**, **14**, **16**, **18**, **19**, **20**.

Solutions were generated with the Egsets program. Then we proceeded to add the worst predicted molecule to the training set, as explained above. This was continued until we ended up with eight molecules in our training set. Because the four added compounds (**3**, **7**, **12**, and **15**) all have low observed potency, we conclude that the most important refinement of successive site models was always some way to prevent relatively inactive compounds from binding too well. This set ran for several weeks on a Silicon Graphics Indigo workstation but could produce only four solutions while searching through less than 25% of the whole tree. When tested against our validation procedure, only one of these four solutions could survive. But this solution was found to correctly predict only 6 of the 12 compounds in the test set. The other six had rather high errors. Realizing that this could at best be described a modest prediction, we decided to shuffle the order of our compounds in the training set in order to explore different parts of the tree by starting at a new point. Most of such trials resulted in either no solution or in very few solutions. It should be noted that we are not attempting to explore the entire tree now. One such trial (starting point), however, yielded 214 solutions. We now compared all the solutions generated from the training sets of four to eight molecules. The results are summarized in Table III.

Both the rules effect major reduction in the solutions of the final training set (D). Interestingly enough, the single solution left after rule II for training set D was not present in the list of those which survived rule I for the same set D. We still decided to choose one of the solutions under rule I because some of the solutions were fairly high in predictive power compared to the single solution in rule II.

Additional support for our overall method, in the form of validation, could be obtained if it can be shown that the average predictive power of the model increases with more input. In other words, if we have several solutions for each training set, then the question to be asked is: Do the models for each training set predict correctly a higher number of test molecules as we increase the number of input molecules? Table IV gives a listing of the predictive power of the solutions from training sets consisting of four, five, six, and eight molecules (labeled as sets A, B, C, and D, respectively). The entries in the table show the fraction of the solutions found for a given training set that correctly predict the given test compound. We observe that there is a general increase in the predictability of each test compound when we move from A to D training sets. The last row of the table represents the fraction of correct predictions averaged over all test molecules. The higher this number the greater is the predictive strength of our method. In fact, the average predictions do increase gradually from 24% (four molecules) to 34% (eight molecules). Moreover, all the molecules are predicted at

Table IV. Fraction (%) of the Solutions Found for a Given Training Set of Cocaine Analogues that Correctly Predict the Given Test Compounds

test compound	percent of solutions correctly predicting test compound			
	A	B	C	D
5	33	40	55	73
6	21	23	28	33
8	4	10	22	38
9	7	23	27	31
10	3	8	12	24
11	14	31	36	44
13	11	13	12	9
14	17	5	8	2
16	24	32	28	33
18	86	82	88	84
19	27	10	0	0
20	43	33	38	35
average	24	26	30	34

Table V. Interaction Energy Parameters of the Proposed Four-Region Voronoi Site for the Binding of the Cocaine Analogues

regions	hydrophobicity	refractivity
1	0.131	0.081
2	1.591	−0.025
3	1.528	0.081
4	0.763	0.011

Table VI. Interregion Distance Bounds (Å) for the Four Regions of the Optimal Solution of Training Set D, Obtained from the Egsets Program[a]

regions	r_1	r_2	r_3	r_4
r_1	1.35	7.47	6.19	10.00
r_2	3.82	1.35	8.50	8.50
r_3	0.00	2.91	2.91	6.75
r_4	4.57	2.91	2.91	2.75

[a] Upper triangle and diagonal are upper bounds; lower triangle lists lower bounds.

least once by every training set, except for molecule **19** in sets C and D. This is probably due to insufficient searching of all the possible binding modes in the solution tree, especially since sets A and B do predict this compound. On analyzing the solution tree we found that the program had not reached that part of the tree for set D which was successful in predicting this molecule.

One of the solutions of set D, left surviving under validation rule I (see Table III) was chosen as our best solution. This solution could correctly predict eight out of the 12 test molecules (see Table VIII). Of the remaining, one was within 4% error, two were totally incorrectly predicted, while one could not fitted into the model. The last mentioned molecule (**13**) has a rather big R_1 functionality, which probably explains why it could not find any optimal binding mode for the geometry obtained from rather simpler molecules in our training sets. The solution consists of four regions, the interaction parameters for which are listed in Table V while the interregion geometric details can be found in Table VI. From Table V we can see that the region r_2 and r_3 are the most hydrophobic and region r_1 is the least hydrophobic. Region r_2 seems to disallow sterically bulky groups while regions r_1 and r_3 seem to partially accommodate bulky ligand groups due to their slightly higher molar refractivity parameter values. In Table VI the diagonal elements represent the diameter of the region in the corresponding row and column, while the entries below and above the diagonal are the minimum and the maximum distances, respectively, for the regions represented by the corresponding rows and columns. A

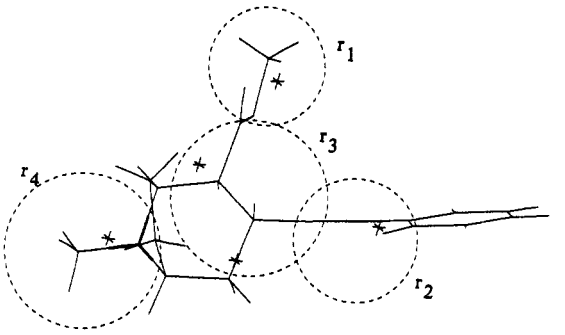


Figure 2. Schematic representation of the four-region Voronoi binding site, with (*R*)-pseudococaine. Regions are represented by dotted spheres. The positions of the pseudoatoms are marked by an asterisk (*) (refer to text for details). The 2-carboxymethyl substituent (R_2) lies in region r_1.

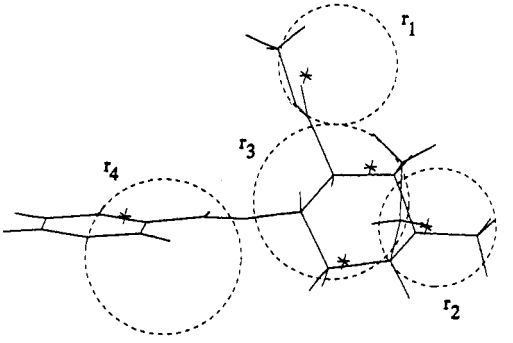


Figure 3. Schematic representation of the four-region Voronoi binding site, with (*S*)-pseudococaine. Regions are represented by dotted spheres. The positions of the pseudoatoms are marked by an asterisk (*) (refer to text for details). The 2-carboxymethyl substituent (R_2) lies in region r_1.

Table VII. Optimal Binding Modes of the Eight Molecules (Training Set) Used in Generating the Picture of the Voronoi Site[a]

compd	ΔG_{obs}[b]	ΔG_{calc}[b]	binding mode				
			P_1	P_2	P_3	P_4	P_5
1	6.99	7.44	2	1	3	4	4
2	4.80	5.47	2	1	4	3	3
3	4.80	5.62	4	1	2	3	3
4	4.65	5.47	2	1	4	3	3
7	4.54	4.78	1	4	2	3	3
12	3.71	3.98	3	2	4	1	3
15	6.17	6.13	1	2	4	3	3
17	7.64	7.37	2	1	3	4	4

[a] For the identity of the pseudoatoms (P_i) refer to Figure 1. The numbers 1, ..., 4 under the columns headed P_1, ..., P_5 indicate which of the four regions the corresponding pseudoatom lies in when the molecule binds in its optimal mode. [b] ΔG is expressed as $-\log(IC_{50})$.

schematic illustration of the four binding regions of the proposed Voronoi site can be seen in Figures 2 and 3, which show the binding of (*R*)-pseudococaine and (*S*)-pseudococaine, respectively. The regions are (clockwise from the top) r_1, r_2, and r_4, respectively, while the one in the center is r_3. The isolated starred coordinates represent the location of the pseudoatoms falling in the corresponding regions. In fact, these regions have been constructed using the pseudoatom representation, by first superimposing the squashed ligand molecules such that there is common overlapping of those pseudoatoms which occupy same regions, and maintaining the distance constraints shown in Table VI. Therefore, it is not surprising that parts of the actual ligand molecule does not fit entirely into the respective regions that they are meant to occupy.

Table VII shows the calculated binding energies and the optimal binding modes of the molecules in the training

Table VIII. Optimal Binding Modes of the 12 Molecules in the Test Set[a]

compd	ΔG_{obs}[b]	ΔG_{calc}[b]	binding mode P_1	P_2	P_3	P_4	P_5	error[c] (%)
5	5.21	5.47	2	1	4	3	3	0
6	5.01	5.63	2	4	1	3	3	0
8	4.17	3.91	1	2	4	3	3	0
9	6.95	7.44	4	1	2	3	3	0
10	6.60	6.40	3	2	4	1	3	0
11	5.28	5.11	4	3	1	2	3	0
13[d]	7.14							–
14	6.25	3.86	3	2	4	1	3	36
16	6.52	7.16	4	1	3	4	4	0
18	7.79	7.60	2	1	3	4	4	0
19	7.60	6.55	2	1	3	4	4	10
20	8.10	7.51	2	1	3	4	4	4

[a] For the identity of the pseudoatoms (P_i) refer to Figure 1. The entries in the table correspond to the four regions of the site. [b] ΔG is expressed as $-\log(IC_{50})$. [c] Relative error (E_m) for the compound, from eq 4. [d] No binding modes could be found for compound 13.

set D, determined from the "best" solution described above. The following interesting observations could be made about their hypothesized binding at the cocaine receptor site. Compounds 1 and 17 possess the maximum observed activity and they both seem to bind in the same fashion at the site. Their pseudoatoms P_4 and P_5, which is most of the tropane ring, occupy the same region r_4. The interaction energy parameters for that region suggests that there is moderate contribution toward the binding affinity due to the hydrophobic interactions with the tropane ring. In fact, in all the molecules the tropane ring (P_4 and P_5) seem to always occupy the same region r_3, except for **12**. Since region r_3 is more hydrophobic than region r_4, we infer that the less active molecules actually bind in a way that allows more hydrophobic interaction for the tropane ring. But the chief contribution to the activity comes from the interaction of the 3-benzoyl and 3-phenyl (P_3) groups in region r_3 for molecules 1 and 17, respectively. Table V shows that this region is very hydrophobic and has a moderately large molar refractivity. The other main functional group, 2-carbomethoxy, occupies region r_1, which has the least hydrophobicity and moderate refractivity. This suggests that even the contribution due to the carbomethoxy group is very marginal, and most of the binding affinity is due to the interaction of the substituent at the 3-position in both these highly active molecules. Among the two substituents for these two molecules at the 3-position, phenyl offers more favorable interaction because it is more hydrophobic. The importance of each part of these compounds could be summarized as follows.

Effect of Stereochemistry. Stereochemistry plays an important role in modulating the binding of the cocaine analogues as evidenced by the results of the Voronoi model. The binding mode for every pair of enantiomers was found to be different. And that difference always seems to involve only two groups in the ligand which switch the regions they were occupying for the corresponding enantiomer. For example, the pseudoatoms P_3 and P_1 of (*R*)-pseudococaine occupied regions r_2 and r_4, whereas in (*S*)-pseudococaine the same atoms were found to occupy regions r_4 and r_2, respectively. This can be seen from Figures 2 and 3. All the other atoms in these two enantiomers had the same region occupancy. The notable observation was that one of the two atoms involved in such a switch (for all such pair of enantiomers) was always P_3. This further highlights the importance of the functionality at the 3-position.

Effect of 3-Substituents. As stated above, the functionality at the 3-position seems to be the most important for the ligand binding. For those cases where the only structural difference lay in the substitution at the 3-position with respect to (*R*)-cocaine, the binding mode remained the same as that for (*R*)-cocaine. The difference in the activity between these compounds then is a result of the difference in the physicochemical properties of the substituents. The substituted phenyl seems to be more favorable than the benzoyl group. For most compounds, the pseudoatom containing this functionality (P_3) seems to fall into either region r_3 or r_4. Both these regions have high hydrophobicity interactions. Thus, the phenyl group contributes to the activity mostly through the hydrophobic interactions. However, the role of electron-donating para substituents is not very clear.

Effect of 2-Substituents. In most active compounds the 2-substituent does not impart significant activity. Since this is a hydrophilic group in most cases, it makes a compound less active whenever the final geometry forces it to occupy the hydrophobic region r_2 as in **12**. Otherwise we observe this to lie mostly in the relatively less hydrophobic regions r_1 and r_4.

Effect of N-Substituents. The only compound in our training set which has a different substituent at the N atom is compound **15**. The bulky substituent in this compound cannot fit into the smallest region r_2, as in (*R*)-cocaine. Hence it goes into r_1 and forces P_2 into region r_2. The net result is that the tropane ring and benzoyl group are also forced into different regions. Thus, a bulky substituent at the N atom results in a totally different binding mode of the ligand.

Conclusion

We have used the Voronoi approach to obtain a 3-dimensional binding site model for a single cocaine receptor site. Our results show the importance of the substituent at the 3-position of the tropane ring, in agreement with results of other workers. Two of the four regions which are very hydrophobic seem to accommodate this functionality. The model also establishes the importance of the tropane ring which presumably acts as an anchor to the binding of the molecule by engaging in favorable hydrophobic interactions. However, these results were obtained by making the assumption that the binding of all the cocaine analogues occur at a single site on the dopamine transport system. Hence they should be viewed with caution until the assumption is verified by other experiments.

Acknowledgment. This project is supported by grants from National Institute of Health (GM37123) and National Institute on Drug Abuse (DA06746). We would also like to thank Joel Dinverno and Ian Nordan for their helpful contributions at the early stages of this project.

References

(1) Coyle, J. T.; Snyder, S. H. Antiparkinsonian Drugs: Inhibition of Dopamine Uptake in the Corpus Striatum as a possible Mechanism of Action. *Science* **1969**, *166*, 889–901.
(2) Horn, A. S.; Cuello, C.; Miller, R. J. Dopamine in the Mesolimbic System of the Rat Brain: Endogenous Levels and the Effect of Drugs on the Uptake Mechanism and Stimulation of Adenylate Cyclase Activity. *J. Neurochem.* **1974**, *22*, 265–270.
(3) Javitch, J. A.; Blaustein, R. O.; Snyder, S. H. [^{3}H]mazindol Binding Associated with Neuronal Dopamine and Norepinephrine Uptake Sites. *Mol. Pharmacol.* **1984**, *26*, 35–44.

(4) Kuhar, M. J.; Ritz, M. C.; Boja, J. W. The Dopamine Hypothesis of the Reinforcing properties of Cocaine. *Trends Neurosci.* **1991**, *14*, 299–302.
(5) Carroll, F. I.; Lewin, A. H.; Boja, J. W.; Kuhar, M. J. Cocaine Receptor: Biochemical Characterization and Structure-Activity Relationships of Cocaine Analogues at the Dopamine Transporter. *J. Med. Chem.* **1992**, *35*, 969–981.
(6) Carroll, F. I.; Lewin, A. H.; Abraham, P.; Parham, K.; Boja, J. W.; Kuhar, M. J. Synthesis and Ligand Binding of Cocaine Isomers at the Cocaine Receptor. *J. Med. Chem.* **1991**, *34*, 883–886.
(7) Carroll, F. I.; Gao, Y.; Rahman, M. A.; Abraham, P.; Parham, K.; Lewin, A. H.; Boja, J. W. Synthesis, Ligand Binding, QSAR and CoMFA study of 3β-(p-substituted phenyl)tropane-2β-carboxylic Acid Methyl Esters. *J. Med. Chem.* **1991**, *34*, 2719–2725.
(8) Lewin, A. H.; Gao, Y.; Abraham, P.; Boja, J. W.; Kuhar, M. J.; Carroll, F. I. 2β-Substituted Analogues of Cocaine. Synthesis and Inhibition of Binding to the Cocaine Receptor. *J. Med. Chem.* **1992**, *35*, 135–140.
(9) Abraham, P.; Pitner, J. B.; Lewin, A. H.; Boja, J. W.; Kuhar, M. J.; Carroll, F. I. N-Modified Analogues of Cocaine: Synthesis and Inhibition of Binding to the Cocaine Receptor. *J. Med. Chem.* **1992**, *35*, 141–144.
(10) Ritz, M. C.; Cone, E. J.; Kuhar, M. J. Cocaine Inhibition of Ligand Binding at DA, NE and Serotonin Transporters: A Structure Activity Relationship Study. *Life Sci.* **1990**, *46*, 635–645.
(11) Carroll, F. I.; Abraham, P.; Lewin, A. H.; Parham, K. A. Studies of Uptake of Serotonin(5-HT), Norepinephrine(NE) and Dopamine(DA). *J. Med. Chem.* **1992**, *35*, 2497–2500.
(12) Crippen, G. M. Voronoi Binding Site Models. *J. Comput. Chem.* **1987**, *8* (7), 943–955.
(13) 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.
(14) Srivastava, S.; Richardson, W. W.; Bradley, M. P.; Crippen, G. M. 3-Dimensional Receptor Modeling Using Distance Geometry and Voronoi Polyhedra. In *3D-QSAR: Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers: Leiden, Netherlands, 1993 (in press).
(15) Viswanadan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for the Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic Interactions and their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (163), 163–172.
(16) Crippen, G. M. Linearized Embedding: A New Metric Matrix Algorithm for Calculating Molecular Conformations Subject to Geometric Constraints. *J. Comput. Chem.* **1989**, *10*, 896–902.
(17) Crippen, G. M.; Smellie, A. S.; Richardson, W. W. Conformational Sampling by a General Linearized Embedding Algorithm. *J. Comput. Chem.* **1992**, *13*, 1262–1274.
(18) Bradley, M. P.; Crippen, G. M. Voronoi Modeling: The Binding of Triazines and Pyramidines to L. Casei Dihydrofolate Reductase. *J. Med. Chem.*, in press.
(19) Reith, M. E. A.; Meisler, B. E.; Sershen, H.; Lajtha, A. Structural Requirements for Cocaine Congeners to Interact with Dopamine and Serotonin Uptake Sites in Mouse Brain and to Induce Stereotyped Behavior. *Biochem. Pharmacol.* **1986**, *35*, 1123–1129.
(20) Madras, B. K.; Spealman, R. D.; Fahey, M. A.; Neumeyer, J. L.; Saha, J. K.; Milius, R. A. Cocaine Receptors labeled by [^{3}H]2β-Carbomethoxy-3β-(4-fluorophenyl)tropane. *Mol. Pharmacol.* **1989**, *36*, 518–524.
(21) Villar, H. O.; Loew, G. H. A Conformational Study of Cocaine and its Diastereomers. *J. Comput. Chem.* **1990**, *11* (9), 1111–18.