Genetically Evolved Receptor Models: A Computational Approach to Construction of Receptor Models

D. Eric Walters* and R. Michael Hinds

Department of Biological Chemistry, Finch University of Health Sciences/The Chicago Medical School, 3333 Green Bay Road, North Chicago, Illinois 60064-3095

Received April 7, 1994®

Given the three-dimensional structure of a receptor site, there are several methods available for designing ligands to occupy the site; frequently, the three-dimensional structure of interesting receptors is not known, however. The GERM program uses a genetic algorithm to produce atomic-level *models* of receptor sites, based on a small set of known structure–activity relationships. The evolved models show a high correlation between calculated intermolecular energies and bioactivities; they also give reasonable predictions of bioactivity for compounds which were not included in model generation. Such models may serve as starting points for computational or human ligand design efforts.

Introduction

In recent years there have appeared numerous computer programs which can identify potential new ligands, based on the three-dimensional structure of a receptor site determined by X-ray crystallography, NMR spectroscopy, or homology modeling.[1,2] Some of these programs search libraries of structures or fragments to find molecules complementary to the binding site,[3,4] and others construct molecules de novo[5–10] to maximize favorable interactions with the binding site. In every case, however, the structure of the receptor site or related sites must be known. The goal of the present research is to produce atomic-level *models* of receptor sites, based on a small set of known structure–activity relationships. Such models can then serve as starting points for computational or human ligand design efforts.

Others have used three-dimensional quantitative structure–activity relationships (QSAR) to map out steric and electrostatic interactions on a grid surrounding a series of ligands. In particular, Cramer et al.[11] have used statistical methods to correlate such interactions with binding for a series of steroids binding to carrier proteins. More recently, Davis et al.[12] have used Goodford's GRID force field[13] to map possible receptor binding surfaces in a 3D-QSAR study of calcium channel agonists. Snyder et al. have recently reviewed efforts to construct atomic models of receptors, which they classify as "pseudoreceptors" (connected sets of atoms or functional groups) or "minireceptors" (unconnected sets of atoms or functional groups).[14] This group constructed a pseudoreceptor model for the NMDA receptor by linking together four different functional groups at specific points in three-dimensional space around a series of agonists.[15] In the present work, we place a number of explicit model atoms (e.g., 40–60 atoms) at points in space around a series of ligands and calculate intermolecular interactions between ligand and receptor model atoms. By changing the types of atoms at the various positions, we produce models which have a high correlation between calculated binding energy and bioactivity.

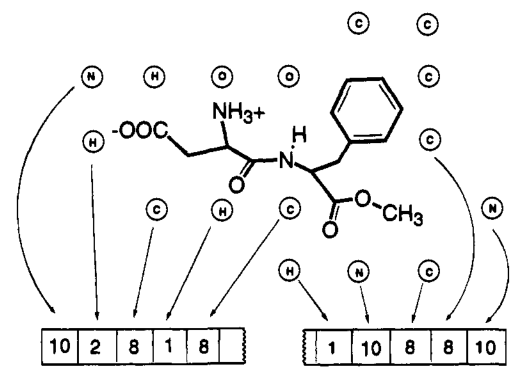


Figure 1. Each model is coded as a linear string of atom types. Each position in the string corresponds to a particular point in coordinate space around the chosen ligands.

Theory

There are several implicit assumptions in our approach. First, it is assumed that the observed bioactivity is proportional to the ligand–receptor interaction energy; there is no attempt to account for transport or metabolic phenomena. Second, it is assumed that the compounds selected for study act at a common receptor site. Third, an active conformation and alignment of ligands in the receptor site is assumed (for the time being); work on examining alternate conformations and alignments is currently in progress. Finally, ligands and receptor models are treated as rigid entities; the current implementation of the program does not take into account any conformational changes.

Receptor models are made by placing atoms at points in three-dimensional space in which they can simulate a receptor surface and interact with the ligands. Figure 1 illustrates this schematically. Since we have no prior knowledge of the receptor structure, the selection of number of atoms, types of atoms, and their positions is entirely arbitrary, and the number of possible models is essentially infinite (a model consisting of 60 atoms chosen from eight possible atom types could exist in $>10^{54}$ forms). From this nearly infinite range of possible models, we wish to identify models for which calculated ligand binding energy correlates with bioactivity. Such

* Address correspondence to: D. Eric Walters, Department of Biological Chemistry, Finch University of Health Sciences/The Chicago Medical School, 3333 Green Bay Road, North Chicago, IL 60064-3095. Telephone: 708-578-3000, extension 498. Fax: 708-578-3240. e-mail: walterse@mis.fuhscms.edu

® Abstract published in *Advance ACS Abstracts*, July 1, 1994.

Table 1. The "Genetic Code" and Parameters Used in This Work[a]

atom type code	CHARMm type	E_{min} (kcal/M)	R_{min} (Å)	partial atomic charge
0	–	–	–	–
1	H (H on polar atom)	−0.0498	0.800	0.25
2	HC (H on charged N)	−0.0498	0.600	0.35
3	HA (aliphatic H)	−0.0450	1.468	0.00
4	C (carbonyl C)	−0.1410	1.870	0.35
5	CH1E (CH group)	−0.0486	2.365	0.00
6	CH2E (CH_2 group)	−0.1142	2.235	0.00
7	CH3E (CH_3 group)	−0.1811	2.165	0.00
8	CT (aliphatic C)	−0.0903	1.800	0.00
9	NP (amide N)	−0.0900	1.830	−0.40
10	NT (amine N)	−0.0900	1.830	−0.30
11	O (carbonyl O)	−0.2000	1.560	−0.50
12	OT (hydroxyl O)	−0.2000	1.540	−0.60
13	OC (carboxylate O)	−0.1591	1.560	−0.55
14	S	−0.0430	1.890	−0.20

[a] Atom types were chosen from the CHARMm force field, and charges are values which approximate those found in the standard 20 amino acids in the commercially distributed version of the CHARMm force field.[17] Type 0 corresponds to having no atom at all in a given position.

a highly multidimensional task is clearly beyond the capabilities of a systematic approach, so we have chosen to attack this problem using a genetic algorithm.

Genetic algorithms have been used successfully in rapidly finding good solutions to very high-dimensional problems for which systematic solution is not practical.[16] The requirements for applying a genetic algorithm are (1) that a possible solution to the problem can be encoded in a linear form, and (2) that a given solution can be evaluated quantitatively. A population of possible solutions is generated (often at random), each linearly-coded solution is treated as a "gene" or an individual member of the population, the "fitness" of each individual is calculated, pairs of individuals are selected to serve as "parents", and pairs of "offspring" solutions are generated by randomly recombining the parent genes so that each offspring derives part of its gene from each parent. Each offspring is evaluated; if the fitness of an offspring is sufficiently high, it replaces a less fit member of the population and can serve as a parent in sucessive generations. By continuing the parent solution–recombination process for a number of generations, the overall fitness of the population increases—natural selection and survival of the fittest takes place.

The genetic algorithm is implemented as follows for our problem, as illustrated in Figure 1. First, a shell of atoms (typically 45 to 60) is created around the series of superimposed ligands. In the current implementation of the program, the ligands are fully surrounded by receptor atoms; in a subsequent version, we expect to allow for an "open face" for the receptor site model. A "gene" consists of a list of atom types (aliphatic H, hydrogen bonding H, aliphatic C, carbonyl C, hydroxyl O, etc.), with each location in the gene corresponding to a specific location in space.

The "genetic code" is shown in Table 1 and is based on the CHARMm force field.[17] Most of the atom types likely to be encountered in a protein receptor site have been included. There are types for aliphatic and polar hydrogens; carbonyl and neutral carbon atoms; amide and amine nitrogens; carbonyl, hydroxyl, and carboxylate oxygens; sulfur. Extended atom types (CH, CH_2,

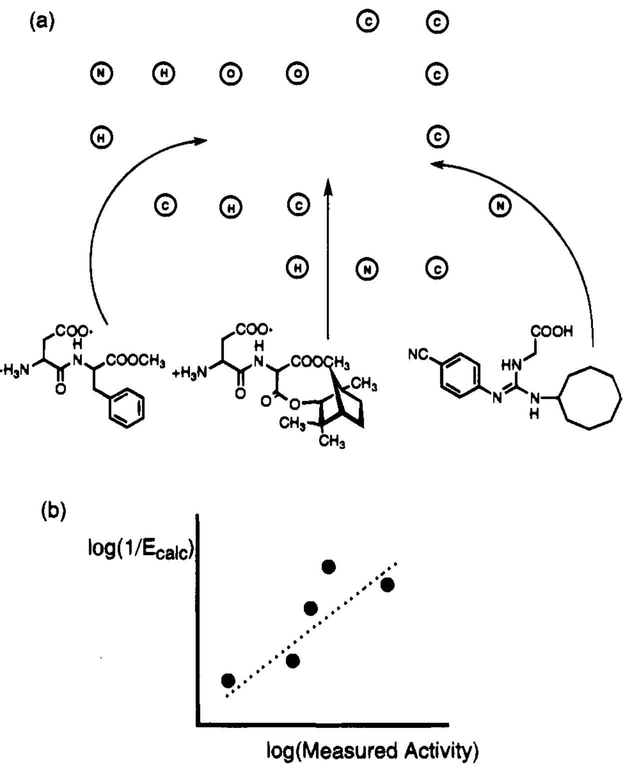


Figure 2. Fitness score for a model is computed by first calculating an interaction energy with each ligand, then calculating the regression coefficient for 1/exp(energy) versus log(bioactivity).

```
parent 1
  10736553121770755531724423303323266632
  01667743721004055373321756237224562O2
parent 2
```

crossover:

```
1073655312177   4055373321756237224562O2
0166774372100   0755531724423303323266632
```

```
offspring 1
  10736553121774055373321756237224562O2
  01667743721000755531724423303323206632
offspring 2
                                   mutate
```

Figure 3. Illustration of the recombination and mutation of a pair of models to form two new models. (a) A crossover point is selected at random. (b) Crossover is applied to form two new models, each deriving a portion of its code from each parent. (c) Random mutation may be applied to one or both of the new models.

CH_3 groups) have been included to permit the inclusion of more steric bulk at a given position; such steric bulk may be important in distinguishing affinities of different ligands for a receptor site. There is also include a "null" atom type, i.e., no atom at all at a given position, to allow for the possibility of open space on the receptor surface. The effect of using different atom types or parameter sets has not been extensively investigated. However, we have carried out a limited number of experiments using a severely reduced genetic code (aliphatic C, polar H, carboxylate O, null type) and

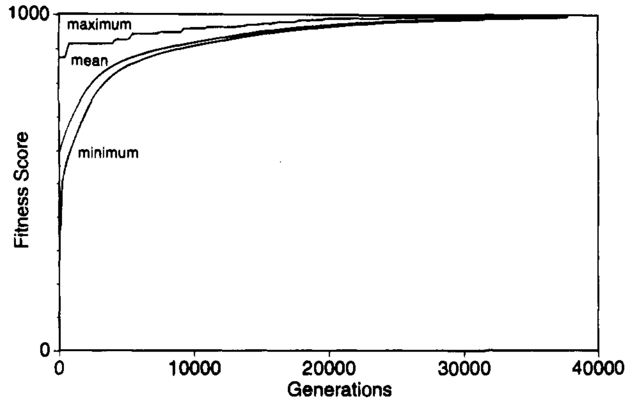


Figure 4. Typical evolution of a population of 2000 receptor models, illustrating the minimum, mean, and maximum fitness scores as a function of generation number.

found the resulting models to be substantially less satisfactory (data not shown).

A population is generated by randomly assigning an atom type code to every position of every individual "gene". The "fitness" score for a given gene is produced by first calculating intermolecular van der Waals and electrostatic energies between a gene (model) and each individual ligand; the correlation coefficient for 1/exp(energy) versus log(bioactivity) is our criterion for measuring fitness (Figure 2). In this way, if a model gives a better correlation between calculated binding energy and bioactivity, it is assigned a higher fitness score. While the fitness function bears some resemblance to the equation relating equilibrium binding constants to Gibbs free energy ($K = \exp(-\Delta G/RT)$), it is not imagined that calculations at this level give real free energies. At best, they are approximations of intermolecular enthalpy.

After the initial population is generated and evaluated, "parents" are selected in a fitness-weighted random manner. Thus, any member of the population may be selected, but members with higher fitness are more likely to be chosen. The generation of two new "offspring" models from two parents is illustrated in Figure 3. First, a point on the gene is chosen at random, and the two parent genes are broken at that point. Then, the tail end of parent 2 is connected to the head end of parent 1, and vice versa, so that each offspring derives part of its "genetic material" from each parent. Following the generation of offspring, random mutation may be carried out at a user-selected rate. A Poisson distribution is applied to the mutation rate, so that if the overall mutation rate is 1.0 per generation, the probabilities of having 0, 1, 2, 3, 4, or 5 mutations in a given gene are 0.368, 0.368, 0.184, 0.062, 0.016, and 0.004, respectively. Mutation then takes the form of assigning random atom types to the appropriate number of randomly selected positions on the gene. Fitness scores for the two offspring are then calculated. If an offspring has a fitness score higher than the least fit member of the population, it takes the place of that member; otherwise, it is discarded. The only exception is the case where the offspring is identical to an existing member of the population; duplicates are not allowed, in order to maintain genetic diversity. Without this restriction, a reasonably fit member of the population sometimes takes over the entire population before significant evolution can occur. The result is that good *partial solutions* to the problem (e.g., atoms which provide good discrimination at one region of the receptor) may be combined with other good partial solutions to produce even better models. This version of natural selection insures that, in the long run, better solutions survive and reproduce, while worse ones are eliminated.

Methods

Hardware and Software. All GERM calculations were programmed in ANSI C and carried out on Unix workstations or PCs. For intermolecular energy calculations, we used eqs 13 and 14 and the van der Waals parameters from ref 12. Partial atomic charges for receptor model atoms are as listed in Table 1; these are representative values from the amino acid parameter set in CHARMm, which are template-based and smoothed. Partial atomic charges for ligand atoms were calculated using the CNDO method as implemented in the Quanta/CHARMm software package, version 3.2.[18] In this way, ligand and receptor atoms having the same atom type have comparable partial atomic charges (e.g., carboxylate oxygens on ligands typically have partial charge of -0.55 ± 0.05). All conformational analysis was carried out using the Quanta/CHARMm program. Local minima corresponding to those previously identified[19] were superimposed manually.

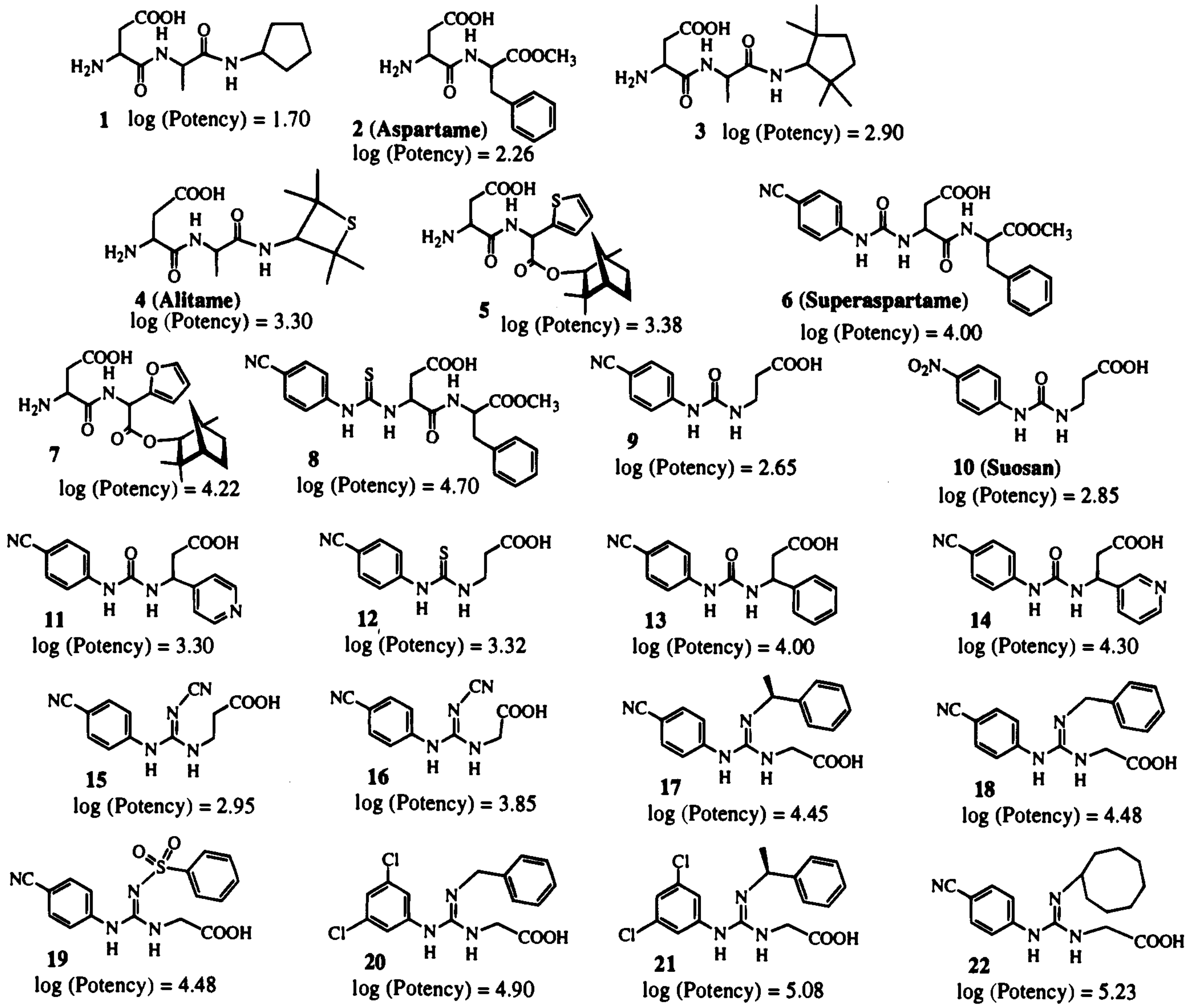
Generation of Receptor Models. The shell of atoms constituting the model is constructed as follows. First, the superimposed ligands are centered with respect to Cartesian coordinate space. Next, points are distributed evenly over a sphere surrounding the superimposed ligands. A model aliphatic carbon atom is placed at each point on the surface of the sphere, and its position is adjusted by optimizing its radius in a spherical coordinate system so as to get maximal van der Waals attraction between the model carbon atom and the ligand molecules. Finally, the radial distance is optionally adjusted by addition of a fixed distance (the "cushion" parameter) to the optimized radius. Thus, the receptor model atoms form a surface reasonably close to the largest ligands. This method of placing atom positions produces a set of points which are well spaced out, although it is not guaranteed to locate the ideal position for a particular receptor atom, and it would not be optimal for surfaces which have very large concave faces. The purpose of adding a "cushion" of 0.1–1.0 Å is to compensate for the lack of flexibility of receptors and ligands in our current implementation.

The population of starting models is generated by filling an $(m \times n)$ array with random numbers ranging from 0 to 14, where m corresponds to the number of members of the population (typically 500–2000) and n corresponds to the number of atoms constituting a model (typically 50–60). The numbers 0–14 correspond to the atom types listed in Table I.

Evolution of Receptor Models. After the initial models are generated, their fitness scores are calculated in the following way. First, intermolecular interaction energies are calculated between a given model and each ligand in the input set. The correlation between 1/exp(energy) and log(bioactivity) is then calculated for the model, and the correlation coefficient becomes the fitness score. Thus, a model is considered to be a good one if it provides a good correlation between bioactivity and calculated binding energy.

A generation consists of the following sequence of steps: (1) selection of two parents, using a fitness-weighted random scheme; (2) selection of a random point in the gene for crossover; (3) generation of two new genes by switching the parental genes from the crossover point onward; (4) carrying out random mutation(s) on the offspring; (5) evaluation of the fitness scores for the offspring; and (6) replacement of less fit members of the existing populations with offspring having higher fitness scores. The final step is omitted if the offspring is identical to an existing member of the population, in order to maintain genetic diversity in the population. The generation process is repeated for a specified number of cycles or until the mean fitness of the population does not change significantly over some number of generations. Typically, if the mean fitness does not increase by 0.001 over 250 generations, we consider that convergence has occurred.

Chart 1



Prediction of Bioactivity from Models. After a model (or population of models) has been evolved, it can be used to calculate bioactivities of other ligands when docked onto the model. Intermolecular van der Waals and electrostatic energies are calculated between the model(s) and ligand, and the predicted bioactivity is interpolated from the energy vs bioactivity correlation which constituted the fitness score of the model. We typically look at predictions from 50 to 100 models in a population and report the mean value and standard deviation. Inclusion of more models does not substantially change the mean or standard deviation.

Results

The user has control of a number of parameters in each calculation, including the size of the population, the number of atoms constituting a model, the mutation rate, and the number of generations to be run. A small population will rapidly converge—the population quickly becomes very similar, and additional generations produce little or no improvement in fitness scores. If the population is too small (100 or less), there is not enough "genetic diversity" to evolve very good solutions. Larger populations (5000 or more) have broader genetic diversity and may evolve to much higher levels of fitness, but they also evolve much more slowly. We have typically worked with population sizes of 500–2000.

We find that mutation rate has much less effect on fitness scores than does the recombination of genes. When crossover is prevented and the only changes are due to mutation, evolution is extremely slow. Conversely, if recombination is applied with a mutation rate of 0, evolution progresses well. Our experience to date suggests that a mutation rate of 1 per generation is somewhat better than no mutation at all, but higher mutation rates can degrade performance significantly. Especially after a population has reached advanced stages of evolution, random change is more likely to degrade than to improve a model.

Typically, we use sets of 4–10 compounds to generate models. We are consistently able to generate models for which the correlation between calculated energy and bioactivity is in the range r = 0.90–0.99. Figure 4 shows results of a typical model evolution run. Fitness scores are initially fairly low and cover a broad range; as the population evolves, there is a rapid increase in mean fitness, then a leveling out and convergence.

We consider that the ultimate test of the evolved models will be to use them in ligand design; such work is currently in progress. For initial evaluation of the method, however, we have chosen to see how well the

Table 2. The Compounds Used in This Study, Their Structural Types, and Their Reported Potencies

compound	structural type[a]	log(potency[b])	ref
1	A	1.699	20
2 (aspartame)	A	2.255	21
3	A	2.903	20
4 (alitame)	A	3.301	20
5	A	3.380	22
6 (superaspartame)	A, U	4.000	23
7	A	4.216	22
8	A, U	4.699	23
9	U	2.653	23
10 (suosan)	U	2.845	24
11	U	3.301	25
12	U	3.322	23
13	U	4.000	25
14	U	4.301	25
15	G	2.954	23
16	G	3.845	23
17	G	4.447	23
18	G	4.477	23
19	G	4.477	23
20	G	4.903	23
21	G	5.079	23
22	G	5.230	23

[a] A = aspartic derivative, U = arylurea or arylthiourea derivative, G = guanidine–aliphatic acid derivative. [b] Potencies are stated on a weight basis relative to 2% sucrose.

Table 3. Design of Cross-Validation Studies

(a) L-Aspartic Acid Derivatives

compd	Asp set 1	Asp set 2	Asp set 3	Asp set 4
1		√	√	√
2	√		√	√
3	√	√		√
4	√	√	√	
5		√	√	√
6	√		√	√
7	√	√		√
8	√	√	√	

(b) Arylurea and Arylthiourea Derivatives

compd	urea set 1	urea set 2	urea set 3	urea set 4
9		√	√	√
10	√		√	√
11	√	√		√
12	√	√	√	
6		√	√	√
13	√		√	√
14	√	√		√
8	√	√	√	

(c) Guanidine–Aliphatic Acid Derivatives

compd	Guan set 1	Guan set 2	Guan set 3	Guan set 4
15		√	√	√
16	√		√	√
17	√	√		√
18	√	√	√	
19		√	√	√
20	√		√	√
21	√	√		√
22	√	√	√	

evolved models can "predict" bioactivity for a series of compounds. From a structure–activity series, we select a subset around which to evolve models and then see how well the models calculate bioactivities for the entire series. It is expected that the models should work well for the compounds around which they were constructed. The real test is whether they can also predict the omitted compounds. By running successive subsets, we can generate predictions for all compounds in the series.

Table 4. Results of Calculations on (a) L-Aspartic Acid Derivatives, (b) Arylurea and Arylthiourea Derivatives, and (c) Guanidine–Aliphatic Acid Derivatives[a]

(a) L-Aspartic Acid Derivatives

	log(potency)				
		calculated form			
compd	actual	set 1 $r = 0.984$	set 2 $r = 0.979$	set 3 $r = 0.990$	set 4 $r = 0.996$
1	1.70	**2.04 ± 0.29**	1.81 ± 0.07	1.82 ± 0.07	1.71 ± 0.04
2	2.26	2.35 ± 0.07	**2.10 ± 0.15**	2.18 ± 0.08	2.23 ± 0.05
3	2.90	3.03 ± 0.07	3.00 ± 0.11	**3.17 ± 0.48**	2.89 ± 0.06
4	3.30	3.05 ± 0.05	2.95 ± 0.06	3.17 ± 0.11	**2.43 ± 0.17**
5	3.38	**3.75 ± 0.43**	3.55 ± 0.10	3.44 ± 0.12	3.46 ± 0.05
6	4.00	4.00 ± 0.11	**3.82 ± 0.58**	4.02 ± 0.11	4.01 ± 0.07
7	4.22	4.24 ± 0.09	4.20 ± 0.09	**3.75 ± 0.32**	4.15 ± 0.06
8	4.70	4.70 ± 0.09	4.70 ± 0.11	4.71 ± 0.09	**3.85 ± 0.87**

(b) Arylurea and Arylthiourea Derivatives

	log(potency)				
		calculated from			
compd	actual	set 1 $r = 0.952$	set 2 $r = 0.947$	set 3 $r = 0.963$	set 4 $r = 0.981$
9	2.65	**2.98 ± 0.09**	2.94 ± 0.05	2.87 ± 0.02	2.66 ± 0.05
10	2.85	3.11 ± 0.04	**3.00 ± 0.07**	3.00 ± 0.02	2.80 ± 0.04
11	3.30	3.43 ± 0.05	3.46 ± 0.07	**3.76 ± 0.07**	3.41 ± 0.09
12	3.32	3.01 ± 0.02	2.99 ± 0.04	2.96 ± 0.02	**2.38 ± 0.15**
6	4.00	**3.83 ± 0.73**	4.01 ± 0.12	4.00 ± 0.06	4.00 ± 0.11
13	4.00	4.05 ± 0.05	**4.16 ± 0.23**	3.99 ± 0.07	4.08 ± 0.09
14	4.30	4.15 ± 0.06	4.14 ± 0.09	**3.77 ± 0.07**	4.15 ± 0.08
8	4.70	4.71 ± 0.05	4.75 ± 0.07	4.71 ± 0.05	**5.25 ± 0.96**

(c) Guanidine–Aliphatic Acid Derivatives

	log(potency)				
		calculated from			
compd	actual	set 1 $r = 0.998$	set 2 $r = 0.997$	set 3 $r = 0.952$	set 4 $r = 0.943$
15	2.95	**3.45 ± 0.18**	2.97 ± 0.02	3.31 ± 0.02	3.34 ± 0.02
16	3.85	3.84 ± 0.01	**2.81 ± 0.12**	3.45 ± 0.02	3.44 ± 0.01
17	4.45	4.45 ± 0.03	4.44 ± 0.05	**4.01 ± 0.41**	4.45 ± 0.05
18	4.48	4.50 ± 0.02	4.45 ± 0.05	4.50 ± 0.06	**4.46 ± 0.07**
19	4.48	**4.68 ± 0.08**	4.47 ± 0.05	4.51 ± 0.05	4.51 ± 0.05
20	4.90	4.89 ± 0.02	**4.77 ± 0.23**	4.87 ± 0.06	4.86 ± 0.04
21	5.08	5.08 ± 0.02	5.08 ± 0.05	**4.78 ± 0.40**	5.11 ± 0.04
22	5.23	5.23 ± 0.02	5.25 ± 0.04	5.25 ± 0.05	**5.00 ± 0.41**

[a] For each set, r is the correlation coefficient ("fitness score") as described in the text. Boldface numbers are "predictions" for compounds which were not inlcuded in the set of compounds around which models were built. log(potency) was calculated for the first 100 models in each population. Reported values are mean ± standard deviation.

Table 5. Average Errors for Bioactivities Calculated from Evolved Models

model sets	average error for compounds included in model evolution	average error for compounds excluded from model evolution
all Asp sets	0.08	0.44
all urea sets	0.06	0.41
all Guan sets	0.04	0.36
all sets	0.06	0.40

We describe here the results of such calculations on 22 sweet-tasting structures from three structural classes. We have chosen these structures for several reasons. First, the measured bioactivity (potency relative to sucrose standards) should be free from complication by uptake, transport, and metabolism factors, since the receptors are located on the surface of the tongue. Second, the bioactivities of this set span a range of 3.5

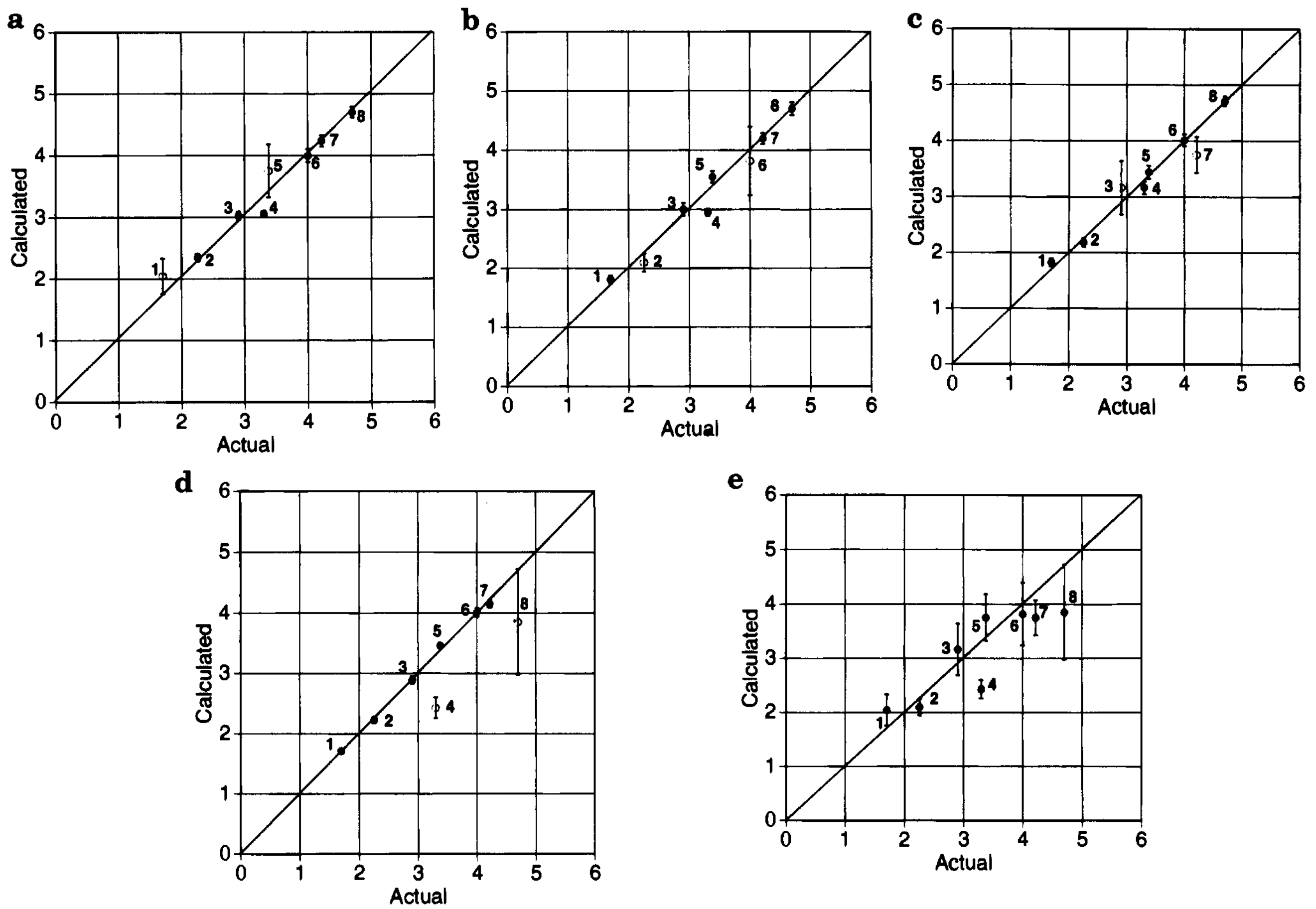


Figure 5. Calculated bioactivities for aspartic derivatives. Data points are calculated as averages over the first 100 models in the population, and error bars indicate standard deviation. (a) Calculated values from set 1. (b) Calculated values from set 2. (c) Calculated values from set 3. (d) Calculated values from set 4. (e) Composite calculated values for all eight aspartic derivatives, each taken from the set in which it was not part of the model-generating process.

orders of magnitude. Third, based on previous modeling studies,[19] we believe that these three classes of compounds act at a common receptor because all have some common features (carboxylate, polar NH groups, hydrophobic groups) which can be superimposed in low-energy conformations. Fourth, we have previously addressed the conformation and alignment issues for these compounds. Finally, these compounds have sufficient structural diversity that they are very difficult to evaluate with classical QSAR methodology; there is no single "core" structure from which standard substituent constants can be applied. Among the ureas, for example, there are compounds with different sized chains connecting urea to carboxylate, compounds with no substitution on the connecting chain, as well as compounds with aryl substitution and peptide substitution, ureas, and thioureas.

Chart 1 shows the structures included in the present study and their reported log(potencies). Three structural classes are represented: aspartame and other L-aspartic acid derivatives (**1**–**8**); arylurea– and arylthiourea–acetic acids (**6**, **8**-**14**); and guanidine-aliphatic acids (**15**–**22**). Table 2 lists the compounds and their reported potencies.[20–25] The potencies listed are based on the concentration of sweetener which matches the sweetness recognition threshold of sucrose (approximately 2%), generally measured as described by DuBois et al.[26] In Table 3, we show the design of the initial studies. Using the eight aspartic acid derivatives as an example, we generate four sets of models. Each set uses six compounds as templates and then calculates bioactivity for all eight compounds, so that each of the eight compounds is predicted from models which were not specifically built around that particular compound. In all of these calculations, we used the following parameters: 60 atoms per model, population size = 2000, mutation rate = 1.0/generation, cushion = 0.5 Å, no. generations = 10,000. These calculations typically used 1–2 h of cpu time on a Silicon Graphics 4D-120 workstation with 16-MHz processors. Table 4 summarizes bioactivities calculated from the populations of models.

First, it is apparent from Table 4 that the models are able to discriminate analogs on the basis of calculated binding energy. In each set of six compounds, the correlation coefficient for 1/exp(energy) versus log(bioactivity) is between 0.943 and 0.998 ($r^2 \geq 0.89$ in every case). Table 5 shows the average errors in calculated bioactivities. "Predicted bioactivity" for a ligand which was a part of the model generation does not vary greatly from one model to another within a population, but predictions for ligands which were not part of the initial model construction have higher variability. For compounds which were part of the model-building dataset, average error is 0.04–0.08 log unit, and for compounds which were not included in model building, average errors for aspartic derivatives, urea derivatives, and guanidine derivatives are 0.44, 0.41, and 0.36 log units, respectively.

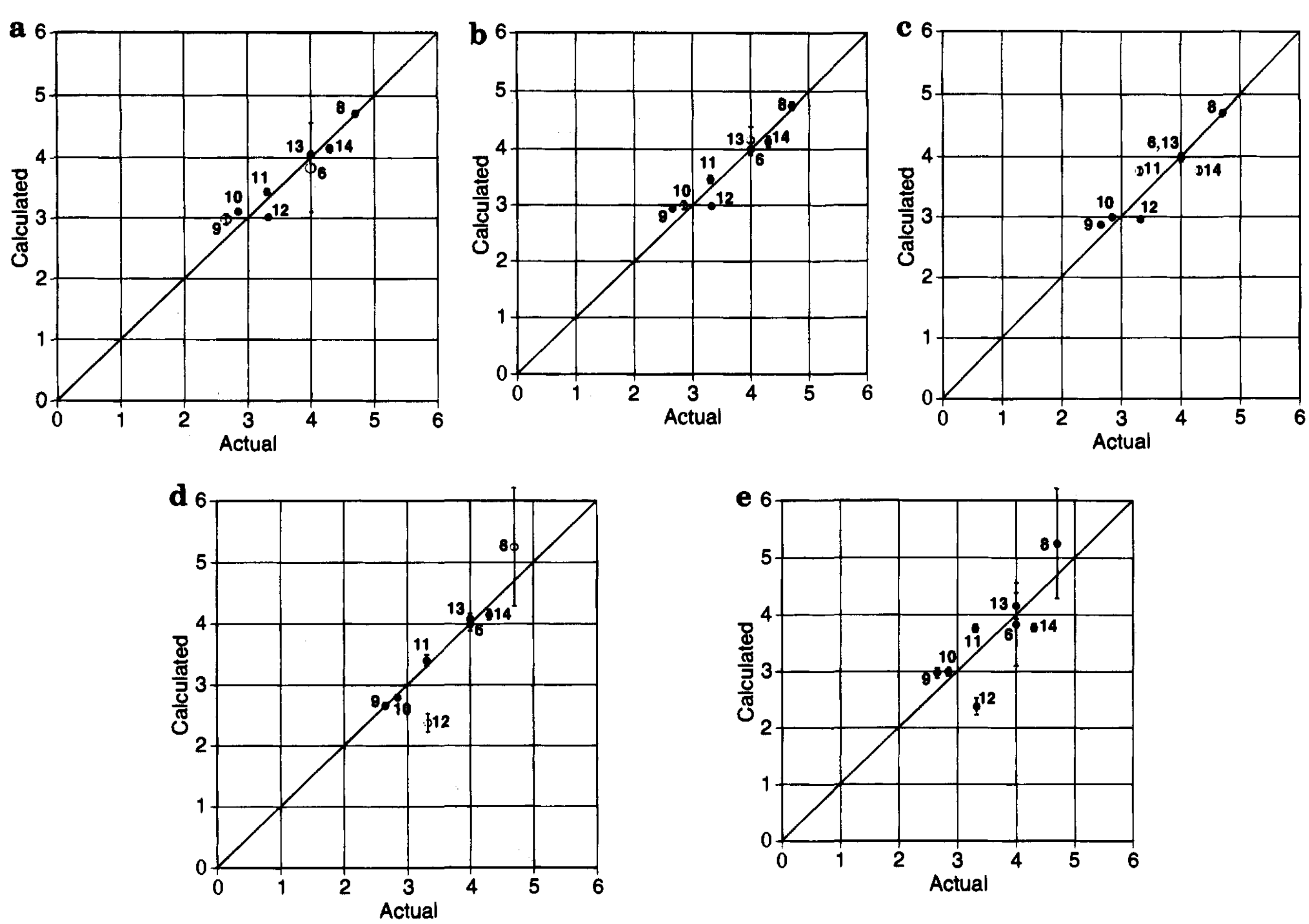


Figure 6. Calculated bioactivities for urea/thiourea derivatives. Data points are calculated as averages over the first 100 models in the population, and error bars indicate standard deviation. (a) Calculated values from set 1. (b) Calculated values from set 2. (c) Calculated values from set 3. (d) Calculated values from set 4. (e) Composite calculated values for all eight urea derivatives, each taken from the set in which it was not part of the model-generating process.

Figures 5–7 show calculated bioactivities for the included and excluded compounds in the aspartic, urea, and guanidine series, respectively. Not surprisingly, we found that the biggest errors in calculated potencies correspond to the compounds for which alignment with other compounds is poorest.

Next, we wished to see how well the method can handle structural diversity. QSAR and related methods are notoriously less effective on structurally diverse series than on homologous series. We selected 11 of the 22 compounds (see Table 6) so as to include the full range of potencies and structural types and evolved a population of models around them. For this series, we used models containing 46 atoms each, a population size of 5000 and ran the calculations for 50 000 generations. The final mean fitness score was r = 0.944. Not surprisingly, these calculations used substantially more cpu time (12–20 h) because of the large number of structures and larger population sizes. Results are shown in Table 6 and Figure 8. Average error for compounds around which models were built was 0.20 log units; for compounds which were not included in model construction, average error was 0.44. In all cases, the residual error in predicted bioactivity was less than 0.75. Such a set of models would be more than adequate for purposes of screening potential synthetic target molecules and identifying those most likely to have desired bioactivity.

Finally, it is important to show that the evolved models are not simply artifacts from the large number of variables.[27] To address this question, we generated several series of models around the aspartic derivatives (**1**–**8**). Using a population size of 2000, mutation rate of 1.0, running for 10 generations, and all eight compounds, we carried out 10 sets of calculations. The resulting r^2 values had a mean of 0.955 ± 0.003. We then carried out 10 more sets of calculations, scrambling the bioactivity data each time, as shown in Table 7. The resulting r^2 values had a mean of 0.344 ± 0.292, indicating that the method is not able to generate models which can correlate arbitrary data.

It is important to recognize that the genetic algorithm method is not designed to find a "global best" solution, but to *rapidly* find many "very good" solutions. Since we prevent duplicate members in our populations, we can evolve thousands of models with very high correlation coefficients. With such a highly combinatorial problem, it is not surprising that there may be thousands of different models with very high fitness scores. Visual inspection of the final population from a model calculation shows that some atom positions have only one or two atom types, while others have a range of possible types. This may indicate which sites are most important for ligand recognition. Sequence analysis of evolved populations will be the subject of a subsequent study.

Figure 9 shows a representative receptor model from the population of 5000 constructed around 11 of the compounds in the dataset. Compound **22**, the most

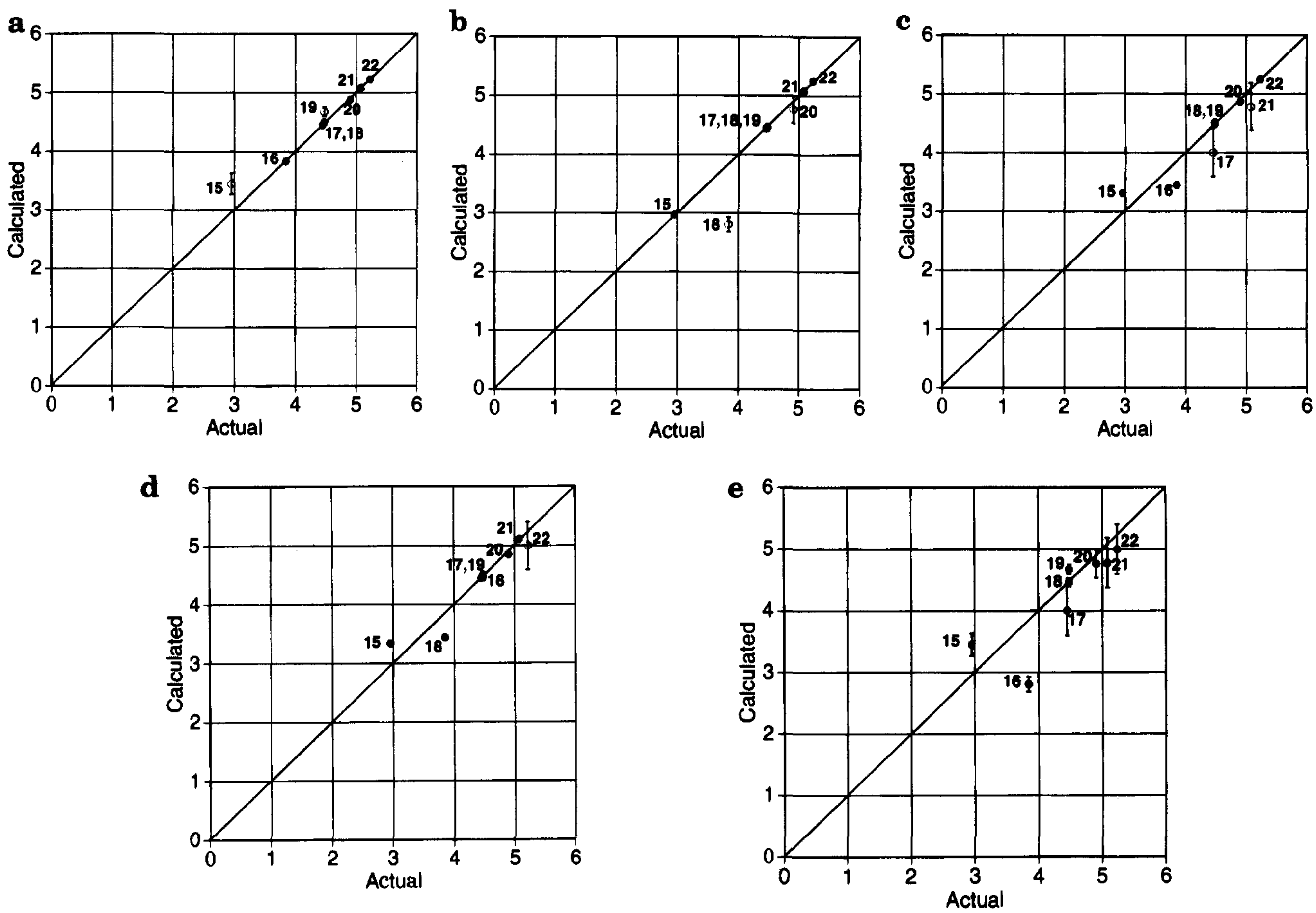


Figure 7. Calculated bioactivities for guanidine derivatives. Data points are calculated as averages over the first 100 models in the population, and error bars indicate standard deviation. (a) Calculated values from set 1. (b) Calculated values from set 2. (c) Calculated values from set 3. (d) Calculated values from set 4. (e) Composite calculated values for all eight guanidine derivatives, each taken from the set in which it was not part of the model-generating process.

Table 6. Actual and Calculated Potencies of All 22 Compounds from Models Built around 11 of the Compounds[a]

compound	log(potency) actual	log(potency) calculated	residual error
1	**1.70**	**2.09 ± 0.07**	**0.39**
2 (aspartame)	2.26	2.64 ± 0.26	0.38
3	2.90	3.39 ± 0.83	0.48
4 (alitame)	**3.30**	**3.12 ± 0.09**	**0.18**
5	**3.38**	**3.43 ± 0.13**	**0.05**
6 (superaspartame)	**4.00**	**3.90 ± 0.15**	**0.10**
7	4.22	3.65 ± 0.66	0.57
8	**4.70**	**4.69 ± 0.15**	**0.01**
9	2.65	2.98 ± 0.17	0.33
10 (suosan)	**2.85**	**3.25 ± 0.09**	**0.41**
11	3.30	4.05 ± 0.26	0.75
12	3.32	2.82 ± 0.12	0.50
13	**4.00**	**4.01 ± 0.15**	**0.01**
14	4.30	4.51 ± 0.36	0.21
15	2.95	3.41 ± 0.11	0.45
16	**3.85**	**3.10 ± 0.07**	**0.74**
17	**4.45**	**4.67 ± 0.10**	**0.22**
18	4.48	4.04 ± 0.16	0.44
19	4.48	4.73 ± 0.14	0.26
20	**4.90**	**4.86 ± 0.11**	**0.04**
21	5.08	5.59 ± 0.40	0.52
22	**5.23**	**5.23 ± 0.11**	**0.00**

[a] Boldface indicates compounds included in evolution of the population of models. Calculated potencies are based on the first 100 models in the population, ± standard deviation, with residual error in the final column.

potent in the series, is shown in the model to show orientation and points of favorable interaction.

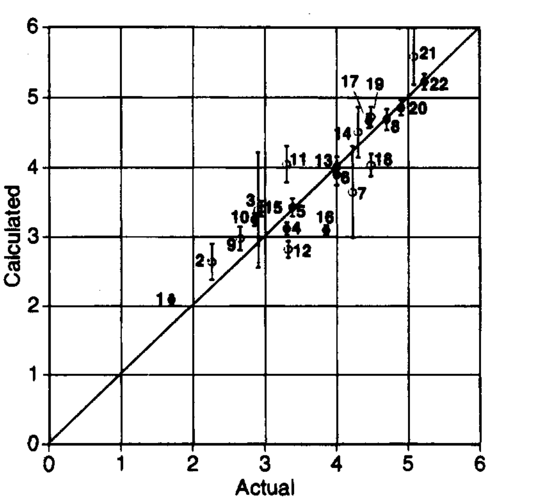


Figure 8. Calculated bioactivities for all 22 compounds. Filled circles indicate the 11 compounds used for evolution of the models, and open circles are calculated values for the 11 compounds which were not included in model generation. All values are averages calculated from the first 100 models in the population.

As is the case for QSAR and other methods based on interpolation, the calculated numbers are quite good for compounds which were part of the model building, and the least accurate results are seen in cases which lie outside the range of structures originally considered. Since energy calculations are based in part on a van der Waals potential, any compound which extends

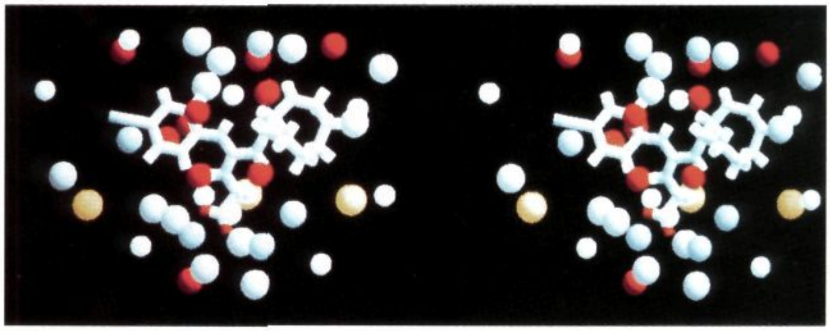


Figure 9. Stereo view of compound **22** enclosed in one of the evolved receptor models. This model has a fitness score (correlation coefficient r for the 11 compounds used in its evolution) of 0.945.

Table 7. Results of Calculations for Compounds **1**–**8** with Bioactivity Data Scrambled in Order to Test for Over-Fitting

run number	bioactivity data order	final r^2
1	5 1 7 2 4 6 3 8	0.797
2	3 5 8 2 4 1 6 7	0.459
3	5 6 2 4 3 1 8 7	0.300
4	6 5 8 2 1 3 7 4	0.170
5	5 3 8 1 4 6 7 2	0.552
6	7 8 2 5 4 1 6 3	0.033
7	7 4 2 8 3 5 1 6	0.050
8	2 1 3 7 5 4 6 8	0.794
9	6 7 1 2 3 8 4 5	0.208
10	7 3 5 8 1 2 6 4	0.079
		mean r^2 = 0.344
		standard deviation = 0.292

substantially beyond space occupied by the starting set of compounds will have a very unfavorable interaction energy and, therefore, a low predicted potency. A solution to this problem which we will address in a subsequent version of the program will be to allow for an open face on the model receptor site. In any case, we believe that the primary value of these models will be as tools to design novel lead structures. A model which is consistent with known structure–activity relationships should be useful in designing and evaluating new molecular frameworks on which to construct active compounds. It is useful to keep in mind that these models are not real receptors in any sense—they are composed of isolated atoms which are not connected to form any kind of protein structure. No intramolecular interactions are taken into account. The models are simply a collection of atom types and positions which distinguish relative potencies of bioactive compounds, on the basis of calculated binding energies. Our goal is to derive a working model which aids in drug design, not to discern the way in which a receptor protein sequence is folded.

Conclusion

Using a genetic algorithm, we have developed a program which can empirically generate *models* of receptors, based on a limited structure–activity series.[28] These models give very high correlation between calculated binding energy and bioactivity. The models also give a good indication of bioactivity for compounds which were not a part of the model-building process, so they should be useful in screening candidates for new analog synthesis. Finally, these receptor models are atom-based and should be adaptable for use with programs which design novel ligands based on three-dimensional receptor structure. It is expected that such models should fill the large existing gap for cases where the three-dimensional structure of a receptor is not known from crystallographic studies.

Acknowledgment. We acknowledge support from the Illinois Division, American Cancer Society, Grant No. 93-07, for some preliminary portions of this research. D.E.W. thanks The NutraSweet Company/NSC Technologies, Mount Prospect, IL, for a gift of computer equipment, and Molecular Simulations, Inc., Burlington, MA, for providing the Quanta/CHARMm and Cerius^2 software. We also thank Prof. Tony Dean, Mr. Eric Kamprath, and Dr. Ki H. Kim for helpful discussions.

References

(1) Martin, Y. C. 3D Database Searching in Drug Design. *J. Med. Chem.* **1992**, *35*, 2145–2154.
(2) Humblet, C.; Dunbar, J. B., Jr. 3D Database Searching and Docking Strategies. *Annu. Rep. Med. Chem.* **1993**, *28*, 275–284.
(3) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.
(4) Bartlett, P. A.; Shea, G. T.; Telfer, S. J.; Watermann, S. CAVEAT: A Program to Facilitate the Structure-Derived Design of Biologically Active Molecules. In *Molecular Recognition in Chemical and Biological Problems*; Roberts, S. M., Ed.; Royal Society of Chemistry: London, 1989; Vol. 78, pp 182–196.
(5) Moon, J. B.; Howe, W. J. Computer Design of Bioactive Molecules: A Method for Receptor-Based *de Novo* Ligand Design. *Proteins: Struct. Funct. Genet.* **1991**, *11*, 314–328.
(6) Nishibata, Y.; Itai, A. Automatic Creation of Drug Candidate Structures Based on Receptor Structure. Starting Point for Artificial Lead Generation. *Tetrahedron* **1991**, *47*, 8985–8990.
(7) Böhm, H.-J. The Computer Program LUDI: A New Method for the De Novo Design of Enzyme Inhibitors. *J. Comput.-Aided Mol. Design* **1992**, *6*, 61–78.
(8) Lawrence, M. C.; Davis, P. C. CLIX: A Search Algorithm for Finding Novel Ligands Capable of Binding Proteins of Known Three-Dimensional Structure. *Proteins: Struct. Funct. Genet.* **1992**, *12*, 31–41.
(9) Verlinde, C. L. M. J.; Rudenko, G.; Hol, W. G. J. In Search of New Lead Compounds for Trypanosomiasis Drug Design: A Protein Structure-Based Linked-Fragment Approach. *J. Comput.-Aided Mol. Design* **1992**, *6*, 131–147.
(10) Rotstein, S. H.; Murcko, M. A. GroupBuild: A Fragment-Based Method for *De Novo* Drug Design. *J. Med. Chem.* **1993**, *36*, 1700–1710.
(11) Cramer, R. D. III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
(12) Davis, A. M.; Gensmantel, N. P.; Johansson, E.; Marriott, D. P. The Use of the GRID Program in the 3-D QSAR Analysis of a Series of Calcium-Channel Agonists. *J. Med. Chem.* **1994**, *37*, 963–972.
(13) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
(14) Snyder, J. P.; Rao, S. N.; Koehler, K. F.; Vedani, A. Minireceptors and Pseudoreceptors. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; Escom; Leiden, 1993; pp 336–354.
(15) Snyder, J. P.; Rao, S. N.; Koehler, K. F.; Pellicciari, R. Drug Modeling at Cell Membrane Receptors: The Concept of Pseudoreceptors. In *Trends in Receptor Research*; Angeli, P., Giulini, U., Quaglia, W., Eds.; Elsevier; Amsterdam, 1992; pp 367–403.

(16) Michalewicz, Z. *Genetic Algorithms + Data Structures = Evolution Programs*; Springer-Verlag: Berlin, 1992.
(17) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
(18) CHARMm, version 22. Molecular Simulations Inc., Burlington, MA 01803.
(19) Culberson, J. C.; Walters, D. E. Three-Dimensional Model for the Sweet Taste Receptor: Development and Use. In *Sweeteners: Discovery, Molecular Design, and Chemoreception*; Walters, D. E., Orthoefer, F. T., DuBois, G. E., Eds.; American Chemical Society: Washington, DC, 1991; Symposium Series Vol. 450; pp 214–223.
(20) Glowaky, R. C.; Hendrick, M. E.; Smiles, R. E.; Torres, A. Development and uses of Alitame, a Novel Dipeptide Sweetener. In *Sweeteners: Discovery, Molecular Design, and Chemoreception*; Walters, D. E., Orthoefer, F. T., DuBois, G. E., Eds.; American Chemical Society: Washington, DC, 1991; Symposium Series Vol. 450; pp 57–67.
(21) Mazur, R. H.; Schlatter, J. M.; Goldkamp, A. H. Structure-Taste Relationships of Some Dipeptides. *J. Am. Chem. Soc.* **1969**, *91*, 2684–2691.
(22) Janusz, J. M.; Young, P. A.; Blum, R. B.; Riley, C. M. High-Potency Dipeptide Sweeteners. 2. L-Aspartylfuryl-, Thienyl-, and Imidazolylglycine Esters. *J. Med. Chem.* **1990**, *33*, 1676–1682.
(23) Tinti, J.-M.; Nofre, C. Design of Sweeteners: A Rational Approach. In *Sweeteners: Discovery, Molecular Design, and Chemoreception*; Walters, D. E., Orthoefer, F. T., DuBois, G. E., Eds.; American Chemical Society: Washington, DC, 1991; Symposium Series Vol. 450, pp 88–99.
(24) Peterson, S.; Müller, E. Über eine neue Gruppe von Süssstoffen. (On a New Group of Sweeteners.) *Chem. Ber.* **1948**, *81*, 31–38.
(25) Muller, G. W.; Madigan, D. L.; Culberson, J. C.; Walters, D. E.; Carter, J. S.; Klade, C. A.; DuBois, G. E.; Kellogg, M. S. High-Potency Sweeteners Derived from β-Amino Acids. In *Sweeteners: Discovery, Molecular Design, and Chemoreception*; Walters, D. E., Orthoefer, F. T., DuBois, G. E., Eds.; American Chemical Society: Washington, DC, 1991; Symposium Series Vol. 450, pp 113–125.
(26) DuBois, G. E.; Walters, D. E.; Schiffman, S. S.; Warwick, Z. S.; Booth, B. J.; Pecore, S. D.; Gibes, K.; Carr, B. T.; Brands, L. M. Concentration-Response Relationships of Sweeteners. A Systematic Study. In *Sweeteners: Discovery, Molecular Design, and Chemoreception*; Walters, D. E., Orthoefer, F. T., DuBois, G. E., Eds.; American Chemical Society: Washington, DC, 1991; Symposium Series Vol. 450, pp 261–276.
(27) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.
(28) For information about availability of the GERM program, contact D.E.W.