PARM: A Genetic Evolved Algorithm To Predict Bioactivity

Hongming Chen, Jiaju Zhou,* and Guirong Xie

Laboratory of Computer Chemistry, Institute of Chemical Metallurgy, Chinese Academy of Sciences, P.O. Box 353, Beijing 100080, P. R. China

Received January 24, 1997[⊗]

Based on Walters' GERM (Genetic Evolved Receptor Model) algorithm, an improved algorithm PARM (Pseudo Atomic Receptor Model) was put forward. PARM uses a combination of a genetic algorithm and a cross-validation technique to produce an atomic-level pseudoreceptor model, based on a set of known structure−activity relationships. During the genetic process, an artificial interfering method, which is based on a complementary principle of ligand−receptor interaction, was used to accelerate the search speed. The evolved models show a high correlation between intermolecular energy and bioactivity and can predict the bioactivity of an unknown molecule by interpolating in the regression equation of the structure−activity relationship. This algorithm was applied to two systems and produced reasonable results.

INTRODUCTION

In recent years, numerous computer programs have appeared[1,2] that can build potential new ligands based on the three-dimensional (3D) structure of a receptor protein. However, in drug discovery, it is common to have measured activity data for a set of compounds acting upon a particular receptor protein but not to have knowledge of the 3D structure of the protein active site. In the absence of such 3D information, the Hansch method[3] is the classical quantitive structure-activity relationships (QSAR) method that correlates the bioactivity with some structure descriptors to build a structure−activity equation. As a 3D QSAR method, building a hypothetical receptor site model can provide insight about receptor characteristics. The pseudoreceptor model can be deduced from a set of compounds that has known bioactivity data.

Some methods for constructing receptor site models have been described. Comparative molecular field analysis (CoMFA) models[4] are effective receptor models that represent the 3D field properties around a set of superimposed molecules by using a probe atom to compute the interaction energies. Hahn[5] used some molecules that are known to bind the receptor to build a receptor surface model composed of many triangle meshes that attributes several kinds of surface properties to every surface point. Vedani and co-workers[6] generated a kind of pseudoreceptor by using different functional groups at the specific points in the 3D space around a series of ligands. On the other hand, Walters[7] described the GERM algorithm. In the GERM algorithm, a number of explicit receptor atoms were placed at points around a series of ligands, and the genetic algorithm was used to deduce the optimal receptor models which have high correlation between binding energy and bioactivity. In the present work, we put forward a pseudo atomic receptor model PARM algorithm that is based on Walters' GERM algorithm to predict bioactivity.

THEORY

Because our PARM algorithm is based on the GERM algorithm, we will explain the GERM algorithm briefly. The most important assumption in GERM is that the observed bioactivity is proportional to the ligand−receptor interaction energy, and that the transport and metabolic phenomena were not accounted for.

For the PARM algorithm, 15 kinds of pseudo receptor atom are defined first. Most of those atom types that are likely to be encountered in a protein have been included. Then, the molecules in the training set are superimposed on a specific pharmacophore and a set of points is generated around the common surface of the superimposed ligands. Receptor models are made by placing atoms at these points in 3D space to simulate a receptor active site and interact with the ligands. To each point, the selection of the type of atom is entirely arbitrary. So, the number of possible models can be very large. For each model, the binding energy of the atom to the ligands can be calculated. From this enormous range of possible models, those models that have good correlation between calculated binding energy and bioactivity are wanted and therefore should be selected out.

To solve this kind of problem, a genetic algorithm[9] is an effective mathematical tool. To apply a genetic algorithm, two requirements must be satisfied: the possible solution to the problem can be encoded in a linear form; and a given solution can be evaluated quantitatively. To encode the problem, each receptor model (possible solution) can be expressed in the form of a bit string, where each bit corresponds to a grid point that is placed around the superimposed ligands and one pseudoreceptor atom type should be assigned to the bit. So, each receptor site model corresponds to an individual. At the beginning of the genetic process, a large number of models are generated and each individual model is evaluated by a score function. During a genetic operation, a pair of individuals that has a high score is randomly selected to serve as parent and a pair of offspring is generated by randomly recombining the parents' genes so that each offspring is derived from part of a gene from

[⊗] Abstract published in *Advance ACS Abstracts,* December 15, 1997.

S0095-2338(97)00004-8 CCC: $15.00
Published on Web 03/23/1998

Table 1. Receptor Atom Types and Parameters[a]

atom type code	atom type	E_{min} (kcal/M)	R (Å)	partial atom charge
0	void	0.0	0.0	0.0
1	H(H on polar atom)	0.042	1.5	0.25
2	HC(H on charged N)	0.042	1.5	0.35
3	HA(aliphatic H)	0.042	1.5	0.00
4	C(carbonyl C)	0.107	1.7	0.35
5	C1(sp C)	0.107	1.7	0.00
6	C2(sp^2 C)	0.107	1.7	0.00
7	C3(sp^3 C)	0.107	1.7	0.00
8	CT(aliphatic C)	0.107	1.7	0.00
9	NP(amide N)	0.095	1.55	−0.40
10	NT(amine N)	0.095	1.55	−0.30
11	O(carbonyl O)	0.116	1.52	−0.50
12	OT(hydroxyl O)	0.116	1.52	−0.60
13	OC(carbonyl O)	0.116	1.52	−0.55
14	S	0.314	1.7	−0.20

[a] Atom charges were chosen from ref 7; the atom parameters were from TRIPOS force field[10]; the charges are values that approximate those found in the standard 20 amino acids.

each parent. Both offspring are evaluated, and the offspring that have high scores will replace the old individuals that have lower scores. By repeating this genetic operation, the average score of the population increases gradually and individuals with higher scores are generated.

The principle of PARM method is similar to that of GERM. The difference between the two methods is mainly in the genetic process. The PARM algorithm is implemented as follows. A set of pseudoreceptor atom types is defined (shown in Table 1). The atom parameters are based on TRIPOS 5.0 force field and partial atom charge is the same as that of GERM. In the Table 1, the atom type 0 means an open space and no atom is put there. When a ligand molecule interacts with the target protein (receptor), some parts of the molecule may not be in contact with the receptor. Considering this situation, arranging an open space in there is necessary. Then, a set of grid points is set around the common surface of the superimposed ligands. In the genetic operation, a large number of individuals (model) are generated first. Considering the complement between the electrostatic field of the ligand and that of the receptor, unlike GERM, every initial individual is generated not entirely on the basis of random mechanism but in a charge-dependent, random manner. We assumed that the charge on the receptor surface point is complementary to the partial atomic charge of the nearest ligand atom that can interact with the receptor surface point. When we generated an individual, one atom type must be assigned to each bit (grid point position). In our algorithm, this selection of atom type isn't entirely random but dependent on the charge of atom that is closest to the grid point. So, we define a formal charge on every grid point. If the receptor model is constructed over a single molecule, each grid point is given a formal charge that is equal to but opposite in sign to the charge of a ligand atom that is closest to the grid point. If the model is constructed over a set of ligands, each grid point is given a formal charge that is equal but opposite in sign to the average partial atomic charge of the closest ligand atoms in the whole molecule set.

After the formal charge of every grid point is calculated, we can assign atom type to each grid point (bit). We divided the 15 pseudoreceptor atom types into three group. The atom types which have positive partial atom charge are included into the positive group, the atom types which have negative partial atom charge are included into the negative group and the neutral atom types are included into neutral group. To each grid point, the selected possibility of atom types in different group is different and that of atom types in the same group is same. The selected possibility of atom types to different grid point is expressed in following formula:

$$Q_f \geq 0.15: \quad P_{\text{positive}} = 0.75, P_{\text{neutral}} = 0.25, P_{\text{negative}} = 0 \tag{1}$$

$$Q_f \leq -0.15: \quad P_{\text{negative}} = 0.75, P_{\text{neutral}} = 0.25, P_{\text{positive}} = 0 \tag{2}$$

$$-0.15 \leq Q_f \leq 0.15: \quad P_{\text{neutral}} = 0.5, P_{\text{negative}} = 0.25, P_{\text{positive}} = 0.25 \tag{3}$$

where Q_f is the formal charge of the grid point and P is the selection possibility of atom type in a certain group.

This selecting mechanism means that if the closest atom in the ligand to the grid point position is an atom that has negative partial atomic charge, then according to the electrostatic complement principle, a pseudoreceptor atom type that has a positive charge is more likely to be assigned to this point and vice versa. From the view point of interaction between ligand and receptor, this mechanism is more reasonable and can speed up the genetic process. When we assigned a pseudoreceptor atom at one grid point, we assumed that the process of assigning atoms at each grid point is independent; that is that placing an atom at one grid point will not affect the choice of an atom at an adjacent grid point.

After the initial population is generated, all the individuals (receptor model) should be evaluated. To get the fitness score of a given model, the interaction energy between each ligand and the receptor model(individual) should be computed first. This energy term, which is comprised of van der Waals energy and electrostatic energy, is determined with eqs 5−7. Then, the linear regression method is used to correlate the bioactivity with the interaction energy data to get a QSAR equation that is in the linear form of $bioactivity = A + B \cdot E_{\text{inter}}$. The cross-validated R^2 of the QSAR equation is used as the criterion to evalute the quality of each individual model. In GERM, the conventional correlation coefficient for 1/exp(energy) versus activity data is the criterion for measureing fitness. But, in the initial stage of genetic process, we found the value of exp(energy) may be so large that it often overflows in the computer and, at the same time, using the conventional correlation coefficient may cause overfitting. For QSAR research, the overfitting is a serious problem that means that in the training set you can get a good correlation coefficient for the regress equation but obtain poor results for the predicting set. To avoid this problem, we used the cross-validation method to measure the ability of the genetic-evolved model to predict. In a cross-validation experiment involving n molecules for a given model, a regression equation is built from all but the first molecule, and this equation is used to predict the activity of the first molecule. Then, all but the second molecule are used to create a regression equation that predicts the second molecule, and so on. In this way, each molecule is predicted

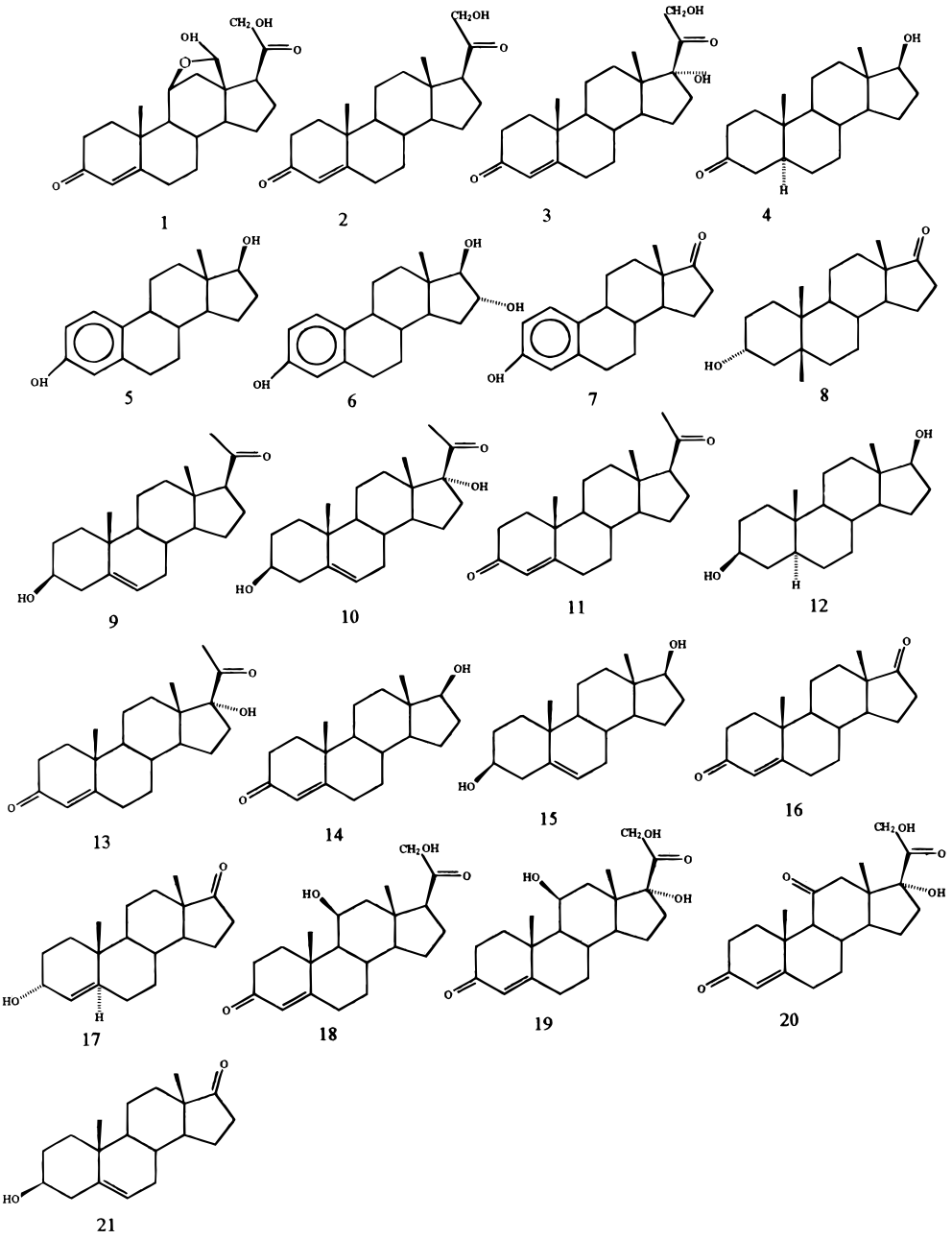

Figure 1. The structures of steroids in the training set.

based on this model and the quality of the model is measured by the cross-validated R^2. Cross-validated R^2 is defined as follows:

$$R^2 = 1 - \frac{\sum(a_i - p_i)^2}{\sum(a_i - \bar{a})^2} \tag{4}$$

where a_i are the assayed activities of the molecules, $\bar{a}$ is the mean of the a_i, and p_i are the predicted molecular activities. The numerator is the squared errors of the predictions, and the denominator is a measure of how much variation there is in the actual activities.

Our fitness score function is $score = exp(\alpha \cdot R^2)$, where α is an adjustable factor. Because the cross-validated R^2 can express the predictive ability of the model, it can reduce the overfitting possibility. In this way, if a model gives a better correlation between binding energy and bioactivity, it can obtain a higher fitness score.

$$E_{vdw} = \sum_{i=1}^{n}\sum_{j=1}^{m} E_{ij}(1.0/a_{ij}^{12} - 2.0/a_{ij}^{6}) \tag{5}$$

$$E_{elec} = 332.17 \cdot \sum_{i=1}^{n}\sum_{j=1}^{m} Q_i Q_j / D_{ij} r_{ij} \tag{6}$$

$$E_{inter} = E_{vdw} + E_{elec} \tag{7}$$

where E_{elec} is the electrostatic energy, E_{vdw} is the steric energy; E_{inter} is the total binding energy; n is the atom number of ligand; m is the grid point number; $E_{ij} = \sqrt{E_i}\sqrt{E_i}$, E_i, E_j (kcal/mol) are the epsilon of ligand atom and pseudoreceptor atom; $a_i = r_{ij}/(R_i + R_j)r_{ij}$ (Å) is the distance between the i atom and the j atom. R_i, R_j are the van der Waals radius of

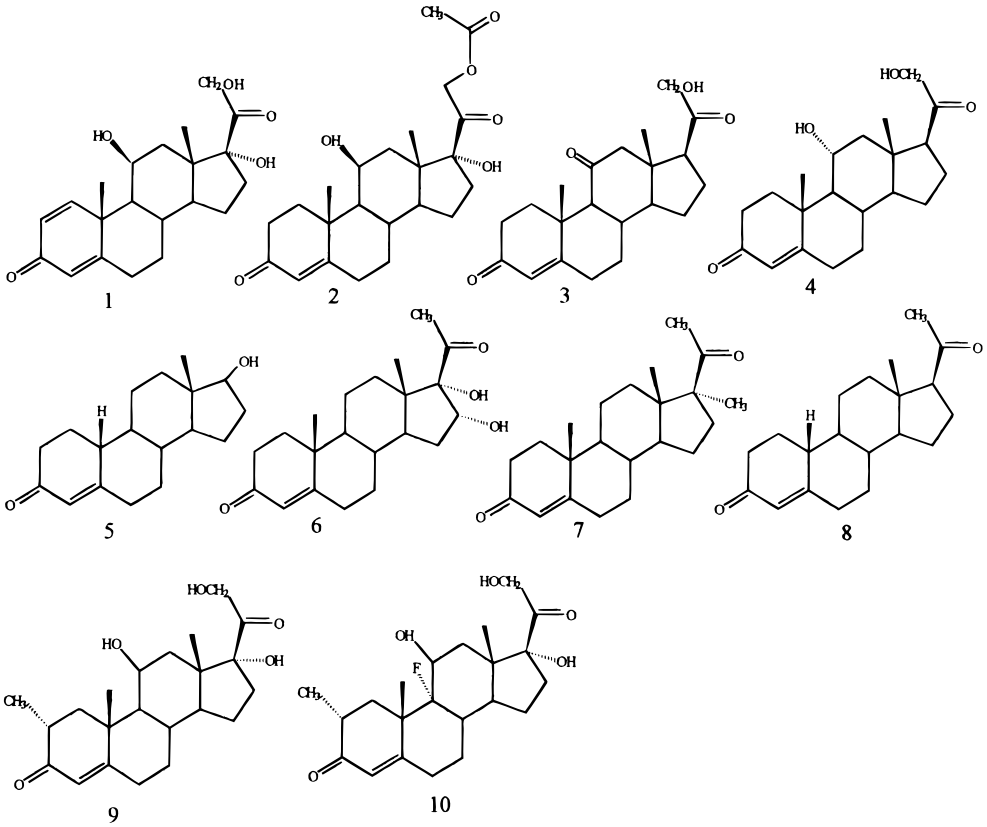

Figure 2. The structures of steroid in the prediction set.

Table 2. Computation Results of Steroids[a]

molecule	CBG	PARM predicted	PARM residual	E_{inter}(kcal/mol)	CoMFA predicted	CoMFA residual
1	6.279	5.803	0.476	−8.277	—	—
2	7.653	7.418	0.235	−24.158	—	—
3	7.881	8.210	−0.329	−31.937	—	—
4	5.919	6.095	−0.176	−11.147	—	—
5	5.000	4.953	0.047	0.0812	—	—
6	5.000	4.464	0.536	4.882	—	—
7	5.000	4.808	0.192	1.503	—	—
8	5.255	5.590	−0.335	−6.187	—	—
9	5.255	5.849	−0.594	−8.733	—	—
10	5.000	6.030	−1.030	−10.508	—	—
11	7.380	6.708	0.672	−17.176	—	—
12	5.000	5.339	−0.339	−3.714	—	—
13	7.740	6.714	1.026	−17.237	—	—
14	6.724	6.163	0.561	−11.818	—	—
15	5.000	5.303	−0.303	−3.364	—	—
16	5.763	6.100	−0.337	−11.196	—	—
17	5.613	5.388	0.225	−4.199	—	—
18	7.881	7.856	0.025	−28.459	—	—
19	7.881	8.027	−0.146	−30.137	—	—
20	6.892	7.004	−0.112	−20.085	—	—
21	5.000	5.292	−0.292	−3.253	—	—
$CBG = 4.961 - 0.102 \cdot E_{inter}$	$r = 0.913$	SD = 0.491	$R^2_{cross} = 0.806$			
1[b]	7.512	7.449	0.063	−24.459	6.544	−0.968
2[b]	7.553	8.037	−0.484	−30.241	7.540	−0.013
3[b]	6.779	6.601	0.178	−16.118	6.526	−0.253
4[b]	7.200	6.015	1.185	−10.360	7.546	0.346
5[b]	6.114	6.246	−0.132	−12.630	5.955	−0.159
6[b]	6.247	5.742	0.505	−7.680	7.057	0.810
7[b]	7.120	6.925	0.195	−19.309	5.384	−1.736
8[b]	6.817	6.100	0.717	−11.204	7.009	0.192
9[b]	7.688	6.108	1.580	−11.272	7.227	−0.461
10[b]	5.797	5.991	−0.194	−10.128	6.937	1.140
SD* = 0.504					SD* = 0.637	

[a] All the data of CoMFA computation are from ref 4; —, CoMFA predicted and residuals are not available. [b]Predicting set.

the *i* and *j* atoms, respectively; D_{ij} is dielectric function between the *i* and *j* atoms; and Q_i is the is atomic charge of the *i* atom.

A fitness-weighted random manner is used in the selection of parents. This procedure means that any member of the population may be selected, but the higher the fitness score

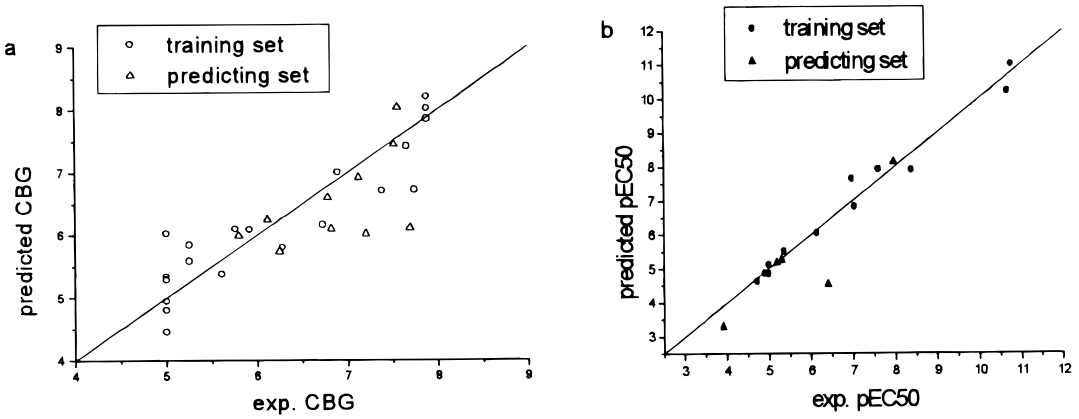

Figure 3. The plots of predicted versus experimental bioactivity data: (a) CBG of steroids; (b) pEC_{50} of KCOs.

Table 3. Top 10 Receptor Site Models of Steroid

1	CBG = 5.246 - 0.065*E_{inter}, r = 0.924, SD = 0.462, R^2_{cross} = 0.828 -5-11-10-4-3-3-7-0-7-1-10-5-11-13-2-1-9-0-4-4-4-11-0-9-11-0-2-8-2-4-13-13-1-8-6-7-3-0-5-3-12-2-10-0-4-12-3-1-12
2	CBG = 4.862 - 0.089*E_{inter}, r = 0.922, SD = 0.465, R^2_{cross} = 0.827 -3-11-11-3-9-7-10-8-0-6-12-2-2-14-4-4-8-6-4-4-0-2-9-14-8-1-8-14-7-4-6-10-3-1-12-4-3-5-0-4-9-2-9-13-0-11-9-10-7
3	CBG = 4.971 - 0.103*E_{inter} r = 0.919, SD = 0.475, R^2_{cross} = 0.820 -4-9-0-10-5-2-3-0-5-10-12-2-2-14-5-1-8-7-8-0-2-5-6-0-7-13-4-8-4-1-0-5-7-6-10-3-0-6-4-8-12-2-11-1-11-0-2-7-10
4	CBG = 5.534 - 0.104*E_{inter} , r = 0.921 SD=0.469, R^2_{cross} = 0.818 -9-14-10-4-0-0-8-10-5-13-14-3-11-13-2-1-9-4-4-4-4-1-8-3-12-12-8-12-4-1-0-5-7-6-3-10-3-1-1-6-9-1-6-11-2-0-0-
5	CBG = 5.194 - 0.087*E_{inter}, r = 0.920,SD = 0.471, R^2_{cross} = 0.817 -8-5-9-5-1-10-6-8-1-12-12-5-10-1-5-8-14-2-3-5-6-5-0-6-9-7-2-10-1-4-0-11-0-2-6-3-7-8-7-8-12-8-11-13-6-11-9-5-3
6	CBG = 5.019 - 0.116*E_{inter} , r = 0.918, SD = 0.476, R^2_{cross} = 0.814 -8-5-9-5-1-3-3-0-5-11-13-8-8-5-1-7-5-6-6-3-7-0-4-5-7-14-4-0-4-1-6-0-6-1-12-3-5-4-0-4-9-2-12-13-0-11-7-5-3
7	CBG = 5.266 - 0.073*E_{inter} ,r = 0.919,SD = 0.473, R^2_{cross} = 0.813 -5-11-10-4-3-3-7-0-7-1-10-8-5-7-1-2-10-3-1-13-4-1-0-9-13-1-8-7-1-2-14-13-6-1-12-3-3-4-0-4-9-2-12-0-4-12-3-1-12
8	CBG = 5.685 - 0.072*E_{inter} , r=0.914,SD=0.489, R^2_{cross} = 0.811 -2-14-6-5-13-5-5-7-4-10-12-5-1-1-5-8-14-2-3-5-6-5-0-6-9-11-2-5-4-2-0-10-12-7-13-3-12-6-0-0-11-7-7-12-4-12-2-7-9
9	CBG = 5.914 - 0.105*E_{inter}, r = 0.918, SD = 0.478, R^2_{cross} = 0.809 -4-5-8-3-3-13-2-0-7-0-11-4-11-6-12-3-8-11-6-13-1-2-9-14-13-1-5-10-1-4-11-1-9-4-10-3-11-1-2-0-9-1-4-1-13-6-6-0-13
10	CBG = 4.961 - 0.102*E_{inter}, r = 0.913, SD = 0.491, R^2_{cross} = 0.806 -14-9-8-3-2-7-3-0-14-5-13-8-8-5-3-7-5-6-6-7-7-0-4-5-7-14-4-0-4-1-6-0-6-1-12-3-3-4-0-4-9-2-12-13-0-11-7-7-14

an individual has, the more likely it will be chosen. Similar to GERM, we select two parents and use a crossover and mutation operation to generate new offspring models. In crossover operation, a point in one individual is selected randomly and two parent individuals are broken at that point. Two parents cross over and exchange their different sections to recombine to form two children. In mutation operation, a bit of each child is selected randomly and its value changed

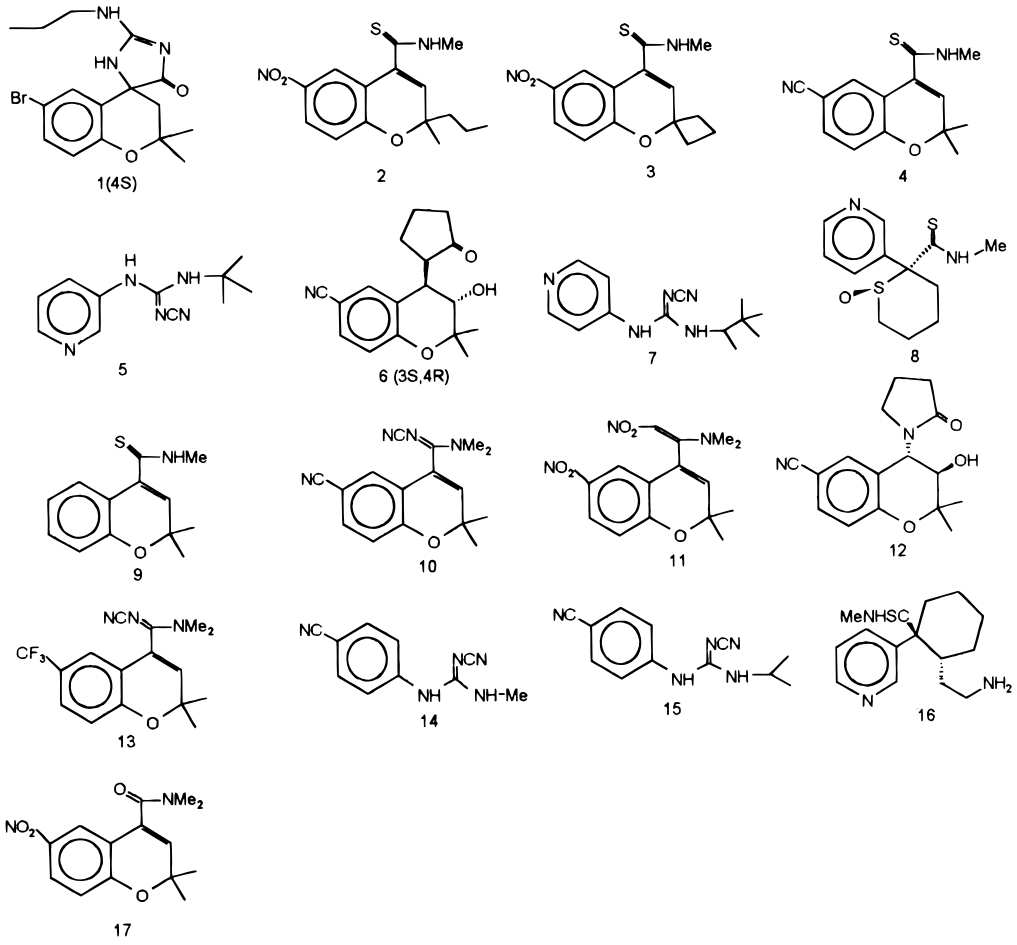

Figure 4. The structures of the KCO molecules.

randomly to get a new individual. After two children are generated and evaluated, they replace two old individuals that have lower fitness scores. To maintain the diversity of the population, if the offspring is identical to any existing member of the population, this offspring will be discarded. This generation process continues over and over again until the maximum generation is reached. In our experiment, often after 2000−3000 generations, some good models can be obtained. After the reasonable models are found, we use these good models to predict the bioactivity of the molecules in the predicting set. First, the binding energy between the predicting molecule and receptor model is computed and, by interpolating this energy in the QSAR equation of that model, we can compute the predicted bioactivity.

We know that there are some groups that sometimes can be regarded as a hydrogen bond donor and some times can be regarded as a hydrogen bond acceptor, such as the hydroxyl group. In a given model, the role of the hydroxyl group depends on which atom is used to represent it in the grid point. If a H atom is put in the grid point, it means that the hydroxyl group is regarded as a hydrogen bond donor and vice versa.

Table 4. The Computation Results of KCO Compounds[a]

molecular	actual pEC_{50}	predicted pEC_{50}	residual	E_{inter} (kcal/mol)
1	8.400	7.906	0.494	−17.537
2	10.770	11.008	−0.238	−37.021
3	10.680	10.223	0.457	−32.089
4	7.610	7.926	−0.316	−17.665
5	7.040	6.821	0.219	−10.725
6	6.970	7.648	−0.678	−15.914
7	6.140	6.058	0.082	−5.932
9	5.370	5.520	−0.150	−2.555
12	5.000	4.857	0.143	1.609
13	4.900	4.875	0.025	1.499
15	4.720	4.625	0.095	3.069
16	5.000	5.132	−0.132	−0.118
$EC_{50} = 5.114 - 0.159*E_{inter}$		$R = 0.988$	SD = 0.345	
$R^2_{cross} = 0.965$				
8*	6.400	4.536	1.864	3.628
10*	5.300	5.271	0.029	−0.992
11*	5.190	5.175	0.015	−0.390
14*	3.910	3.296	0.614	11.416
17*	7.970	8.143	−0.173	−19.025
SD* = 0.4698				

[a] Note: the molecules with asterisks are predicting set molecules.

COMPUTATION

The PARM algorithm was programmed in ANSI C and run on a SGI Personal Iris workstation with a friendly Motif user interface. Using the TRIPOS 5.0 force field to calculate binding energy between pseudoreceptor and ligand, the dielectric function that was used in the energy calculation was $D_{ij} = r$, and the α parameter in the score function was 5.0. Partial atomic charges and atomic parameters of pseudoreceptor atoms are listed in Table 1. Two data sets were investigated: one is the binding affinity to corticosteroid binding globulin (CBG) of 21 steroids that were studied by CoMFA; and the other is the pEC_{50} of 17 K^+ channel opener (KCO),[11] which we have studied before. Partial atomic

Table 5. Top 10 Receptor Models of KCO

1	$pEC_{50} = 5.114 - 0.159*E_{inter}$	r = 0.988 SD = 0.346 $R^2_{cross} = 0.966$
	6-10-3-3-11-6-1-10-2-6-1-8-3-10-2-2-2-5-6-0-10-0-3-4-2-14-8-4-5-11-6-10-3-2-6-5-8-5-11-2-4-8-2-2-4-13-4-0-7-9-1-	
2	$pEC_{50} = 4.193 - 0.149*E_{inter}$	r = 0.987 SD = 0.393 $R^2_{cross} = 0.960$
	6-10-3-3-11-5-7-8-3-3-1-8-3-0-9-2-4-13-4-6-4-3-3-8-4-14-8-4-5-11-6-10-3-2-0-8-8-1-11-14-4-8-2-2-4-13-4-0-7-9-1-	
3	$pEC_{50} = 4.936 - 0.130*E_{inter}$	r = 0.985 SD=0.392 $R^2_{cross} = 0.959$
	8-10-2-7-12-6-7-7-3-5-4-0-3-10-7-2-1-8-4-6-4-5-8-0-4-8-1-4-5-11-6-10-3-2-0-10-3-1-11-2-3-2-5-2-4-13-4-3-4-12-4-	
4	$pEC_{50} = 5.017 - 0.118*E_{inter}$	r = 0.986 SD = 0.376 $R^2_{cross} = 0.959$
	12-12-2-7-12-4-1-10-2-6-1-8-3-0-9-2-4-0-3-1-7-9-4-4-4-14-3-7-5-14-6-10-3-2-6-5-8-5-11-2-4-8-2-2-4-13-4-3-0-13-2-	
5	$pEC_{50} = 4.252 - 0.150*E_{inter}$	r = 0.985 SD = 0.384 $R^2_{cross} = 0.958$
	6-10-2-7-12-4-1-10-2-5-4-0-3-7-9-2-4-5-4-6-4-3-3-4-2-7-3-14-5-11-6-6-0-6-6-5-8-5-11-2-4-8-2-2-4-13-4-3-13-9-1-	
6	$pEC_{50} = 4.167 - 0.162*E_{inter}$	r = 0.984 SD = 0.405 $R^2_{cross} = 0.958$
	6-12-2-7-12-6-1-10-2-10-1-8-3-8-9-2-4-12-2-0-8-5-13-14-4-7-7-4-4-10-6-10-3-1-7-8-8-1-11-2-3-8-4-2-4-7-2-8-8-14-1-	
7	$pEC_{50} = 5.046 - 0.125*E_{inter}$	r = 0.985 SD = 0.396 $R^2_{cross} = 0.956$
	1-1-10-5-12-7-6-1-3-13-4-0-3-10-7-2-1-8-4-6-4-5-8-0-4-8-1-4-7-10-6-10-0-2-0-10-14-3-0-2-7-6-2-2-4-9-4-3-5-12-4-	
8	$pEC_{50} = 5.012 - 0.163*E_{inter}$	r = 0.982 SD = 0.421 $R^2_{cross} = 0.953$
	6-10-3-3-11-6-1-10-2-6-1-8-3-10-2-2-2-5-6-3-6-6-1-4-2-14-8-4-5-11-8-10-3-2-6-5-8-5-11-2-4-8-2-2-4-13-4-0-7-9-1-	
9	$pEC_{50} = 5.731 - 0.124*E_{inter}$	r = 0.983 SD = 0.414 $R^2_{cross} = 0.953$
	1-1-10-5-13-11-7-6-7-1-0-9-8-13-6-2-6-14-2-7-4-5-8-0-4-14-8-4-5-11-6-10-3-2-0-10-14-1-11-2-3-1-2-2-4-13-4-3-0-13-2-	
10	$pEC_{50} = 5.338 - 0.107*E_{inter}$	r = 0.982 SD = 0.425 $R^2_{cross} = 0.951$
	1-10-3-0-11-5-7-10-2-6-1-8-3-0-6-2-4-5-4-6-8-5-8-0-4-5-12-1-1-0-3-7-1-2-0-10-8-1-11-2-3-8-2-2-4-13-4-3-4-11-3-	

charges of ligand atoms were calculated by PM3 semiemprical method as implemented in SYBYL 6.0.[8] The conformation optimization and pharmacophore superimposing work were all carried out in SYBYL 6.0 software.

The grid points were generated evenly around a maximum van der Waals common surface of the superimposed ligands. The distance between grid point and its closest atom in ligands was a user adjustable parameter (called cushion). In our computation, the range of cushion distance was set to 0.5−1.0 Å. The initial population was normally 1000−3000, and a maximum generation was set. When the generation reached the maximum generation, the genetic process is over.

RESULTS AND DISCUSSION

In the genetic process, the population is important to get a good solution. If the population number is too small, there is not enough genetic diversity to evolve a good solution. Larger populations can broaden the genetic diversity, which may evolve into a much higher fitness score, but this will need more time. The cushion parameter and grid point number are two adjustable factors. Normally, the total number of grid point was set between 40 and 60, and the cushion was 0.5−1 Å.

We applied PARM to the steroid binding affinity prediction problem. This standard data set (shown in Figures 1 and 2) has been studied by Cramer's CoMFA method.[4] So, we can compare the computation results of PARM algorithm with those of CoMFA and see the validity of PARM. The first 21 molecules were used as the training set and the remaining 10 molecules were included in the predicting set. In our computation, the population was 1000, maximum generation was 2000, the grid point was set to 49, and the cushion was 0.5 Å. In PARM computation, at last, we can get a series of receptor site models that have high conventional correlation coefficients and cross-validated R^2. Usually, the top 20 models are used to predict bioactivity of predicting set. The computation results of the 10th model are listed in Table 2 (this model has the lowest standard

prediction deviation) and the calculated data versus the actual data is plotted in Figure 3. It can be seen that the experimental data are fairly well reproduced over the entire activity range. The optimal model also can predict the activity of molecules in the predicting set. From Table 2 it seems that the computation results are comparable to those of the CoMFA computation. The computation results of top 10 models are listed in Table 3.

Another system we studied is the K^+ channel opener (KCO). In total, 17 molecules were included in this data set (shown in Figure 4). The conformation was aligned according to our previously studied pharmacophore model,[11] where 12 molecules were used as the training set and the remaining five molecules were used as a predicting set. We set the initial population to 1000, the maximum generation was 3000, the cushion parameter was 0.6 Å, and the grid point number was 51.

The computation results of the model that has the highest score are listed in Table 4 and the predicted pEC_{50} and the observed pEC_{50} are plotted in Figure 3. The conventional correlation coefficient for this model is 0.9882 and the cross-validated R^2 is 0.965, indicating that the results are quite good with this method. In the prediction set, the mean standard deviation is 0.4698. The prediction for compound **8** is poor and lowers the total predicting precision. However, compared with the mean standard deviation of the training set, this mean deviation is acceptable. So, this model has some predictive ability and can be used to screen potential ligands. The top 10 models are listed in Table 5, and they all have high correlation coefficients and cross-validated R^2 values.

From our computation results, it seems that the PARM algorithm can generate some receptor models that can predict unknown bioactivities based on relating molecular structures to known biological activities. With regard to the regression equation, we can see that the larger the binding energy, the higher the activity. This relationship is reasonable from the viewpoint of ligand−receptor interaction energy.

It should be understood that the derived receptor model cannot be regarded as the real receptor site model and the computed interaction energy has significance only to the bioactivity data. But, this method still can be used as a tool to predict bioactivity of unknown molecules.

CONCLUSION

We have described a new genetic evolved algorithm, PARM, which is based on Walters' GERM algorithm. Using this algorithm, we generated a set of pseudoreceptor site models that have high correlation between bioactivity and ligand−receptor interaction energy and high cross-validation R^2. Two data sets were investigated by this method and the computation results show that the generated receptor models have some prediction ability. This algorithm can be used as a tool to predict bioactivity of unknown molecules.

REFERENCES AND NOTES

(1) Bohm, H. J. The Computer Program LUDI: A New Method for the De Novo Design of Enzyme Inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61.
(2) 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.
(3) Hansch, C.; Fujita, T. ρ-σ-π Analysis. A Method for the Correlation of Biology Activity and Chemical Stricture. *J. Am. Chem. Soc.* **1964**, *86*,1616.
(4) Cramer, R. D.; 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.
(5) Hahn, M., Receptor Surface Models . 1, Definition and Construction. *J. Med. Chem.* **1995**, *38*, 2080.
(6) Snyder, J. P.; Rao, S. N.; Koehler, K. F.; Vedani, A. In *Minireceptors and Pseudoreceptors in 3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; Eson: Leiden, 1993; p 336.
(7) Walters, D. E.; Hinds, R. M. Genetically Evolved Receptor Models: A Computational Approach to Construction of Receptor Models. *J. Med. Chem.* **1994**, *37*, 2527.
(8) *SYBYL 6.0 Molecule Modeling System*; Tripos Associates: St. Louis, MO, 1992.
(9) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithm as a Strategy for Feature Selection. *J. Chemom.* **1992**, *6*, 267−281.
(10) *SYBYL Tutorial Manual*; Tripos Associates: St. Louis, MO, 1992; p 236.
(11) Chen, H. M.; Zhou, J. J.; Xie , G. R.; Pang, S. H., The Studies on Pharmacophore Model of K^+ Channel Openner. *Acta Phys.-Chim. Sinica* **1997**, *13*, 101−105.

CI970004W