第 13 卷 第 7 期 物 理 化 学 学 报 Vol.13, No.7
1997 年 7 月 ACTA PHYSICO-CHIMICA SINICA July, 1997

一种基于虚拟受体模型的定量构效关系研究方法 *

陈红明 周家驹 谢桂荣 任天瑞
(中国科学院化工冶金研究所，计算机化学开放实验室，北京 100080)

摘要 提出了一种可用于定量构效关系研究的 PARM(Pseudo Atomic Receptor Mode) 算法. 算法中定义了一套虚拟受体原子，并利用遗传算法构造了一系列虚拟受体模型. 这些模型具有高的受体 - 配体相互作用能和生物活性间的相关性，并能预报未知分子的生物活性. 应用此算法对钾离子通道开放剂体系进行了 QSAR 研究，获得了较好的结果.

关键词: 虚拟受体模型, 遗传算法, 定量构效关系

近年来，QSAR 研究有了很大的进展. 已从 Hansch 的传统二维 QSAR 方法，发展到三维 QSAR 方法. 其中，受体三维结构已知条件下的三维 QSAR 方法 [1,2] 发展较为成熟，出现了 LUDI, Leapfrog 等商业软件. 受体三维结构未知时的三维 SAR 研究则难度较大，这方面有代表性的方法如 CoMFA[3]. 目前人们感兴趣的体系中，大多数都是受体结构未知时的情况，因此受体结构未知时的 QSAR 研究更有意义. Walters[4] 提出了一种 GERM(Genetic Evolved Receptor Mode) 算法用于受体结构未知情况下的 QSAR 研究. 我们基于 Walters 的基本思想，提出了一种改进的 PARM 算法.

1 计算原理

大多数情况下，配体是在一个由受体原子形成的空腔中与受体发生相互作用的. 因此，配体分子生物活性的高低与这种相互作用的强弱有直接关系，如果能根据已知活性的配体分子建立一个虚拟的受体模型，计算出虚拟受体和配体之间的相互作用能，将有助于进行 QSAR 研究，GERM 和 PARM 算法都是基于这种思路.

为构造一个虚拟的受体模型，首先要选定一套已知活性的配体分子为训练集，并将它们按特定的药效团模型在三维空间中进行叠加，计算出各个分子的电荷分布. 然后在叠加好的分子周围产生一套均匀分布的网格点，每个网格点与配体分子保持一定的间隙. 在每个网格点上放一个模拟的受体原子，这样就产生了一个虚拟的受体口袋. 我们根据 20 种标准氨基酸中所包含的原子类型，共确定了 15 种模拟受体原子类型 (如表 1 所示)，其中编号为 0 的原子类型表示开放的空间，受体原子的原子参数按 TRIPOS 5.0 力场确定. 在每个网格点上放置特定类型的受体原子，就产生了一套受体模型 (如图 1 所示).

1996-12-23 收到初稿，1997-03-16 收到修改稿. 联系人：周家驹. Email: jjzhou@lcc.icm.ac.cn * 国家自然科学基金资助项目

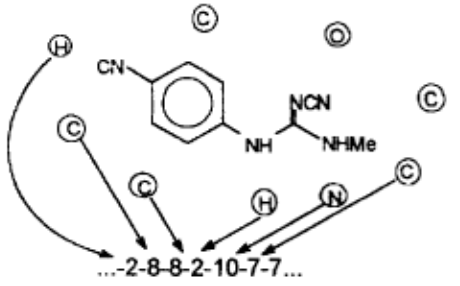

图 1 虚拟受体模型示意图

Fig.1 Illustration of pseudoreceptor model operation

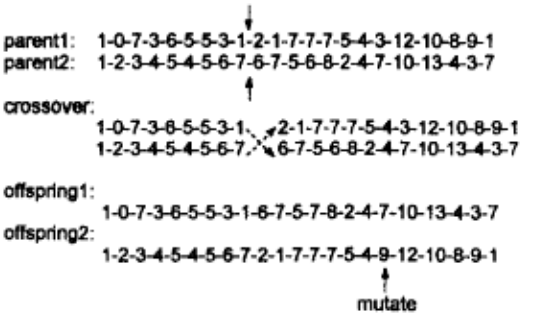

图 2 杂交遗传和变异遗传示意图

Fig.2 Illustration of crossover and mutate operation

表 1 受体原子类型和参数

Table 1 The pseudoreceptor atom types and parameters*

atom type code	atom type	E_{min} kcal·mol^{-1}	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 C)	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

*Atom charges were chosen from ref.4, the atom parameters were from TRIPOS force field[5]. The charges are values which approximate those found in the standard 20 amino acids.

现在的问题是每个网格上应放什么类型的原子. 由于受体的三维结构未知, 每个网格点上应放置的受体原子类型在一定程度上是任意的, 当网格点数达到一定数目时, 这 15 种受体原子在网格点上进行排列组合所形成的可能的受体模型数目将十分庞大. 因此需要从这数量巨大的模型集合中选出最优的受体模型. 对于这种多维空间最优化问题, 遗传算法是一种强有力的数学工具. 要选取最优的受体模型, 首先要确定一个模型评判标准. 考虑到配体分子的生物活性与配体 - 受体间的相互作用有密切的关系, 对于每个受体模型, 先计算出训练分子与受体模型的相互作用能. 这里主要考虑立体作用能和静电作用能, 力场数采用 TRIPOS 5.0[5] 力场的参数. 其计算公式如

下所示：

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

$$E_{\mathrm{elec}} = 332.17 * \sum_{i=1}^{n}\sum_{j=1}^{m} Q_i Q_j / D_{ij} r_{ij} \tag{2}$$

$$E_{\mathrm{inter}} = E_{\mathrm{vdw}} + E_{\mathrm{elec}} \tag{3}$$

其中： E_{elec} 表示静电作用能， E_{vdw} 表示立体作用能， E_{inter} 表示相互作用能； n= 配体分子的原子个数， m= 受体原子数目 (网格点数目); $E_{ij} = \sqrt{E_i}\sqrt{E_j}$, E_i, $E_j(\mathrm{kcal\cdot mol^{-1}})$ 分别是表示配体原子和受体原子的范德华常数； $a_{ij} = r_{ij}/(R_i + R_j)$, $r_{ij}(\text{Å})$ 表示 i, j 原子间距离， R_i, $R_j(\text{Å})$ 分别是表示 i, j 原子的范德华半径； $D_{ij} = i, j$ 原子间的介电函数； Q_i 原子净电荷.

然后用相互作用能对训练集分子的生物活性值进行回归，与 GERM 算法不同的是， PARM 算法采用 Log(bioactivity) 与 E_{inter} 进行回归，而不是与 $1/\exp(E_{\mathrm{inter}})$ 进行回归，这样可以防止计算时的浮点溢出. 我们在 PARM 算法中用回归方程的带交叉验证的复相关系数 R^2_{cross} 作为评价受体模型的标准，而不是象 GERM 算法那样用普通的相关系数 R 作为标准. 这样将减少回归时的过拟合现象. 模型的打分函数为 f=$\exp(k * R^2)$, k 为一经验值，取为 5.0. 这样每个模型的好坏就用分值 f 来衡量.

遗传算法 [6] 作为一种模拟生物进化过程的非数值并行计算方法，近年来日益受到人们的重视. 能有效地解决多维空间的最优化问题. 遗传算法中每个受体模型按表 1 中的原子类型编码表示成一个字符串的形式，每个字符串被称为一个基因. 基因中的每一位对应于受体模型的每一个网格点，每一位上填入一个 0 到 14 间的整数，即表 1 中的 15 种受体原子的原子类型编码. 这样不同的受体模型就用不同的字符串来表示. 进行遗传计算时，首先用随机的方法产生大量的个体，即受体模型. 这一步是通过对基因的每一位随机地给定一个原子类型编码来实现的. 通常的计算中，开始要产生 1000 到 3000 个个体，初始个体产生后，计算出每个模型的分值，以对每个模型进行质量评价.

在进行了初始个体的产生和评价后，就模拟自然界的生物进化现象，挑选两个个体作为父辈进行繁殖以产生两个后代. 挑选个体作为父辈时，采用一种以模型分值作为权重的随机挑选机制，即分值越高的模型被选中的概率越大. 繁殖过程采用了单点杂交和变异两种遗传操作，其原理 [6] 如图 2 所示. 确定了两个后代后，就要计算出各自的分值，对它们进行评价. 最后，把这两个新个体加入个体集合中，同时随机取代两个分值低于它们的个体，这就完成了一代繁殖过程. 通过反复重复这种繁殖过程，将产生越来越多的分值高的个体，整个群体的平均分值都会提高. 当繁殖代数达到指定的最大繁殖代数，整个遗传过程停止. 这样我们就得到了一系列近似最优的受体模型. 通常我们取分值排列前 10 位的模型.

利用这些模型可以预报未知分子的活性值. 首先将预报集分子与训练集分子相叠加，选择通过遗传得到的合理的受体模型，计算出预报分子与虚拟的受体模型间的相互作用能，用此相互作用能对模型的回归方程进行内插而得到预报的活性值.

2 计算

根据上述的算法，用 ANSIC 编制了程序 PARM. 我们选取钾离子通道开放剂 (KCO) 体系进行计算，分子按我们前文研究的药效团模型 [7,8] 进行叠加，分子构象优化和 MOPAC 电荷的计算均在 SYBYL 6.0 分子模型化软件包中进行. 立体能和静电能的计算参数均采用 TRIPOS 5.0 力场参数，静电能计算中的介电函数取 $D_{ij} = r_{ij}$. 遗传计算中的初始个体数为 1000, 最大遗传代数为 3000, 网格点与配体分子的间隙设定为 0.6Å, 以考虑配体分子的柔性因素. 网格点数目定为 51, 程序在 SGI 工作站上运行.

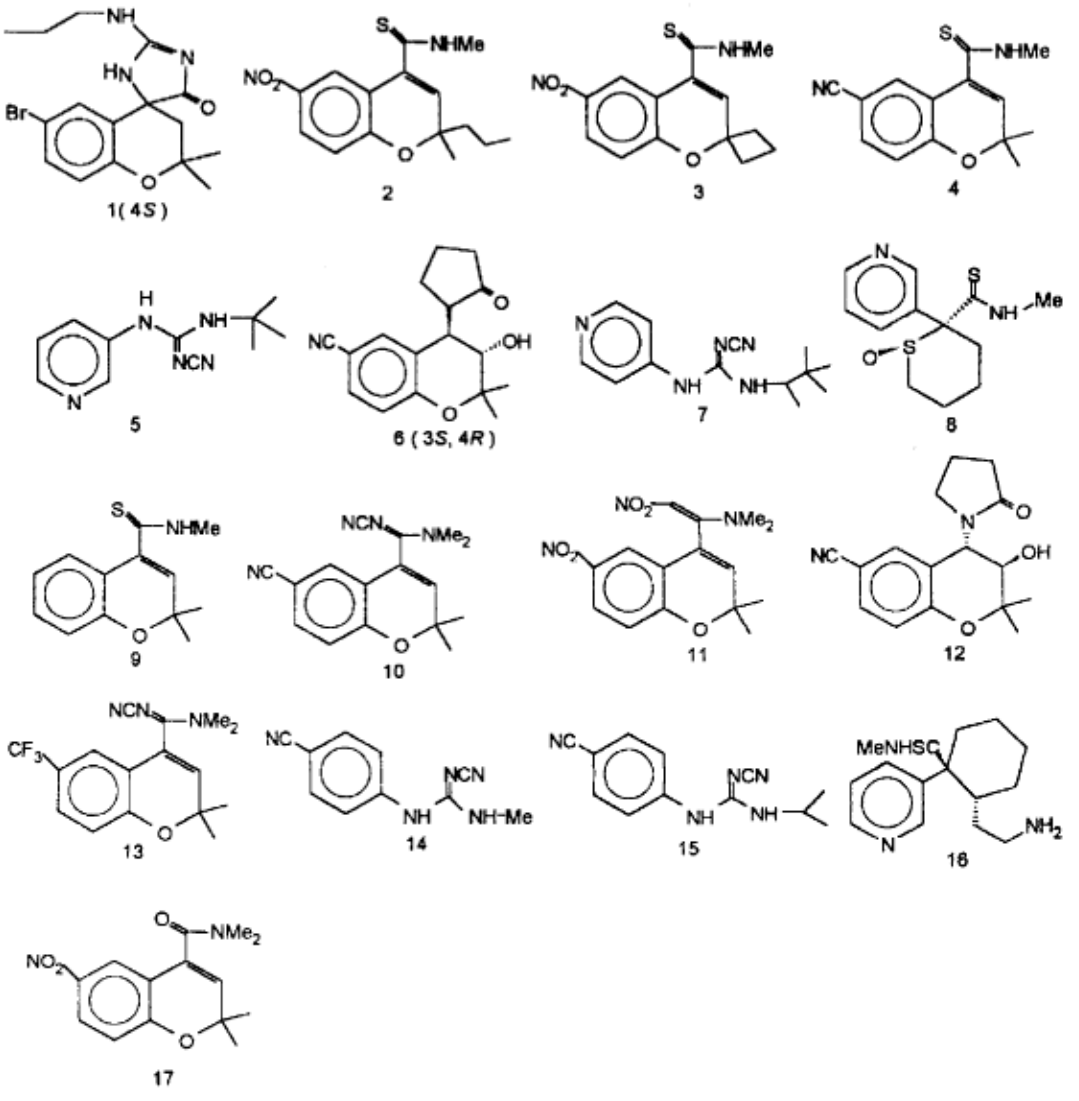

图 3 KCO 分子结构

Fig.3 The KCO molecules' structure

3 结果与讨论

共选取了 17 个分子进行计算，其中 12 个分子为训练集，余下 5 个分子为预报集. 分子结构如图 3 所示. 计算结果表明，通过遗传计算能得到一系列具有高相关系数的受体模型. 模型分值排第一位的模型的计算结果如表 2 所示. 表中 E_{inter} 表示配体分子与受体模型之间的相互作用能， R 表示回归方程的相关系数， R^2_{cross} 表示带交叉验证的复相关系数， SD 为训练集标准偏差， SD^* 表示预报集的标准偏差. 对于 KCO 体系，遗传算法的计算结果是令人满意的，对于预报集的分子，也能给出较准确的预报值. 每个模型都存在一个形如 bioactivity=$a + b^* E_{inter}$ 的回归方程. 它表示配体与受体模型之间的相互作用能越大，配体分子的生物活性就越高. 这种相互作用能表示的是配体和受体相互作用形成复合物前后，体系能量的差值. 相互作用能是与分子的三维结构信息有关的因素，与传统的 Hansch 二维 QSAR 方法相比，这种相互作用能有关的回归方程具有更明确的物理意义，而且避开了传统方法中选择变量这一难点.

表 2 受体模型计算结果

Table 2 The computation results of the top receptor model

molecular	actual pEC_{50}	calculated pEC_{50}	residual	E_{inter}/kcal·mol^{-1}
1	8.4000	7.9059	0.4941	−17.5371
2	10.7700	11.0080	−0.2380	−37.0210
3	10.6800	10.2232	0.4568	−32.0890
4	7.6100	7.9260	−0.3160	−17.6650
5	7.0400	6.8210	0.2190	−10.7250
6	6.9700	7.6475	−0.6775	−15.9144
7	6.1400	6.0580	0.0820	−5.9320
9	5.3700	5.5203	−0.1503	−2.5552
12	5.0000	4.8573	0.1427	1.6089
13	4.9000	4.8747	0.0253	1.4990
15	4.7200	4.6248	0.0952	3.0688
16	5.0000	5.1323	−0.1323	−0.1184
	pEC_{50}=5.1135−0.1592*E_{inter}	R=0.9882	SD=0.3450	R^2_{cross}=0.9650
8*	6.4000	4.5360	1.8640	3.6277
10*	5.3000	5.2710	0.0290	−0.9916
11*	5.1900	5.1750	0.0150	−0.3895
14*	3.9100	3.2960	0.6140	11.4162
17*	7.9700	8.1430	−0.1730	−19.0250
SD^*=0.4698				

Note: The molecules with asterisk are predicting set molecules.

同时，我们还研究了群体的平均分值与遗传代数的变化关系，如图 4 所示. 在遗传的初期，群体的平均分值上升很快，但随着遗传代数的延长，分值的变化逐渐趋缓. 这说明遗传代数大到一定的程度时，模型改善的余地会逐渐变小. 因此，最大遗传代数定得过大，会延长计算时间，而模型质量的改善有限，从效率上讲并不合适.

从计算得出的分值排前十位的模型来看，一些网格上某些类型原子的出现机率较大，这应与受体和配体的相互作用有关，这一点还有待于进一步的研究. 应当指出，得到的受体模型只是虚拟的模型，并不代表真正的模型. 因此，计算得到的相互作用能只具有相对意义，并不能代表实际的配体和受体间的相互作用能.

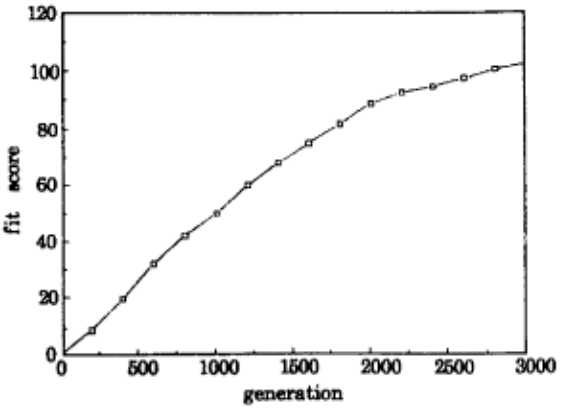

图 4 遗传个体的平均分值与遗传代数的关系

Fig.4 The relationship between the average fit score of population and generation

参 考 文 献

1 Moon J B, Howe W J. *Proteins: Struct. Func. Genet.*, **1991, 11**:314

2 Böhm H J. *J. Comput. -Aided Mol. Design*, **1992, 6**:61

3 Cramer R D III, Paterson D E, Bunce J D. *J. Am. Chem. Soc.*, **1988, 110**:5959

4 Walters D E, Hinds R M. *J. Med. Chem.*, **1994, 37**:2527

5 SYBYL Molecular Modeling System, Tripos Associate, St Louis

6 刘 勇，康立山等. 非数值并行算法 (第二册) 遗传算法，北京：科学出版社，1995

7 陈红明，庞素华，周家驹. 计算机与应用化学，**1997, 14**:31

8 陈红明，周家驹，谢桂荣，庞素华. 物理化学学报，**1997, 13**(2): 101

A QSAR Research Method Based on Pseudoreceptor Model

Chen Hongming　Zhou Jiaju　Xie Guirong　Ren Tianrui

(Laboratory of Computer Chemistry, Chinese Academy of Science, Beijing 100080)

Abstract In this paper, PARM algorithm which can be used in QSAR research was put forward. In this algorithm, a set of pseudo atom was defined and a series of pseudo receptor model was generated by using genetic algorithm. These models which have high correlation between receptor-liagnd interaction and bioactivity can predict bioactivity of unknown molecules. This algorithm was used to investigate the K^+ channel opener system. The reasonable results were obtained.

Keywords: Pseudoreceptor model, Genetic algorithm, Quantitative structure-activity-relationship