基于遗传算法的 PTP1B 抑制剂的二维定量构效关系研究*

潘咏梅 计明娟
(中国科学院研究生院, 北京 100039)

摘要 利用遗传算法, 并结合线性回归和交叉验证方法, 对一系列 43 个苯并呋喃/噻吩联二苯类 PTP1B 抑制剂作了二维定量构效关系的研究. 计算得到了一组效果较好的定量构效关系模型. 模型不仅具有好的回归能力, 而且还具有很好的预测能力. 同时, 通过分析在遗传优化过程中参数在精华种群中所占的比例, 还得到了可能对活性影响较大的成分. 计算结果表明, 分子的 4 个参数: lg *P*(分配系数)、Area(表面积)、MW(分子量)以及 Dip(偶极距)是影响化合物活性的最重要的参数, 这对抑制剂的设计和改造提供了指导.

关键词: 二维定量构效关系, 遗传算法, PTP1B 抑制剂
中图分类号: O641

酪氨酸蛋白磷酸酯酶(protein tyrosine phosphatases, PTPs)是动物体内的一种信号传导酶. 它通过催化蛋白质酪氨酸残基的去磷酸化, 参与调节细胞的许多生理功能. PTP1B 是最早发现的一种 PTP. 90 年代初对糖尿病产生机理的研究表明, 胰岛素通过与受体结合, 导致受体和胞内蛋白质逐级磷酸化, 从而将信号传入胞内来产生生理效应[1]. 而 PTP1B 通过催化去磷酸作用, 阻碍这个过程的进行. 这就造成了胰岛素抵抗(insulin resistance)的现象. 而胰岛素抵抗是大多数糖尿病发生的原因[2-3]. 由此可见, PTP1B 与糖尿病的发生有密切的关系, 它的抑制剂可能成为治疗糖尿病的药物. 而 1999、2000 年小鼠实验从实验上支持了这一观点[4-5]. 针对现有糖尿病药物只能缓解症状, 不能根治的缺点, 开发 PTP1B 抑制剂有可能为开发可以根治糖尿病的药物提供一条新的途径, 有很大的实用意义, 因此成为国外当前研究的一个热点.

本文涉及的苯并呋喃/噻吩联二苯类化合物(benzofuran/benzothiophene biphenyls)是 PTP1B 的抑制剂, 小鼠体内实验证明具有很高的抗高血糖活性[6]. 作者曾经对这类抑制剂做了三维构效关系 CoMFA 的研究, 从化合物周围分子场的角度考察了分子结构对活性的影响[7]. 本文则是把遗传算法应用到二维定量构效关系中, 考察了取代基参数对抑制剂活性的影响, 从取代基的角度为今后抑制剂的设计和改造提供了有用的信息.

1 研究方法

1.1 基于遗传算法的构效关系方法

在传统的二维定量构效关系研究中, 因为同系列化合物的数目和它们的物理化学参数相比往往少许多, 为了避免过拟合, 仅仅能从这些参数中选择一部分来建立回归模型. 因此怎样选择合适的参数一直是定量构效关系研究中的一个难题. 遗传算法的引入可以很好地解决这个问题[8-12]. 基于遗传算法的构效关系方法能很有效地从大量参数中选取合适的参数来构建最佳的构效关系模型. 遗传算法能够给出多个最佳构效关系模型, 而不仅仅是一个单一的结果.

在遗传算法中, 问题的每个解用种群中的一个个体来表示. 基于遗传算法的二维构效关系方法的步骤如下. 1)随机产生多个个体形成初始种群, 初始种群代表一组随机产生的构效关系模型. 在本实验中, 个体是随机挑选出的任意几个参数的组合, 每个参数由两个整数表达, 一是它的序号, 二是它的方次. 这样就得到一个待定系数的方程. 2)结合化合物活性, 用线性回归方法确定方程的参数, 得到构效关系方程. 用评价函数来评价这些个体(方程). 本文评价函数为方程的多元线性回归系数 *r*. 把评估结果好的个体保存在精华种群中. 3)根据种群中

2003-01-13 收到初稿, 2003-03-13 收到修改稿. 联系人: 计明娟(E-mail: jmj@gscas.ac.cn; Tel: 010-88258593; Fax: 010-88256092).
*国家自然科学基金(20273083、29992590)资助项目

个体的得分,结合随机方法,选择被新种群保留的个体以及被淘汰的个体. 随机选择被保留的个体进行交叉操作,产生新的个体. 进行交叉操作时,在种群中选择两个被保留的个体作为母体,然后将这两个母体随机地分为两段,而后在不同的母体中选择一部分组成新的个体. 4)在种群中随机选择个体作突变操作,得到新个体,在评估中被保留的个体和新产生的个体组成一个新的种群. 5)为了将最好的若干个个体保存下来,我们用(精华"种群来保存它们,进行完交叉和突变操作后,逐一比较新种群中的个体和精华种群中的个体,如果新种群中存在更好的个体,就把它们拷贝到精华种群中去. 6)循环 2 至 5 步的计算过程,当"精华"种群的总得分经过若干次操作后不再变化,可以认为计算结束.

通常来讲,对一个包含 200 个个体的种群,如这组数据包含 20 个左右的参数,一般需要 500 ~ 1 000 次循环;对于包含30个参数的体系,收敛则需要 1 000 ~ 1 500 次循环. 对于一般体系来讲,计算需要花费 10 min ~ 1 h 的时间(在 Pentium 150 MH 的 CPU 上测试). 整个计算结束后,从精华种群中就可以得到我们所需的信息.

在计算中,与计算有关的参数保存在一个输入文件中,其中有几个参数的正确选择与否对计算的结果会有至关重要的影响,它们分别是线性回归的结构参数数目 n、种群中所含个体的数目 N_p、交叉因子 P_c 以及突变因子 P_m. n 的选择非常重要,它直接关系到所建立模型的好坏, n 如选得过大,就可能出现过拟合;如 n 选得过小,所建立的模型就可能会丢失一些重要的信息. 一般来讲,化合物个数 N 与 n 之比应大于 2^N. 按照这个原则,本文的 n 选择了 3 和 4. 种群所包含个体的数目是一个重要的参数, N_p 一般为结构参数数目的 20 ~ 50 倍. 一般来讲, N_p 大一点有利于建立更好的模型,本文的 N_p 定义为 200. 交叉因子和突变因子则分别设为 0. 30 和 0. 005.

1.2 模型搭建以及分子特征的计算

所有的分子在 Sybyl 分子模拟软件包中搭建[13]. 搭建的分子采用 MMFF 力场[14],用分子力学方法优化,优化的收敛条件为能量的 RMS(均方根偏差)值小于 0. 42 kJ·mol^{-1}·nm^{-1}. 分配系数和摩尔折射的计算采用 Crippen 等[15]提出的原子加和法;Foct(分子从真空到正丁醇中的自由能变化值)以及 Fh2o(分子从真空到水相中的自由能变化值)的计算采用 Hopfinger[16]提出的水合壳层模型. 拓扑参数的计算采用 Dragon 软件包[17]. 所有分子参数的缩写及相应的定义如表 1 所示. 分子的结构和活性数据见表 2[6].

在计算中,根据化合物的数目,我们搭建了包含 3 个分子参数以及 4 个分子参数的构效关系模型. 种群的规模设为 200,精华种群设为 50. 当精华种群经过连续 30 次遗传操作不再变化以后,计算停止. 每 200 步交叉操作以后,进行一次部分替换操作. 为了考察模型的可靠性,我们计算了每个模型的 F、q^2 以及 PRESS 值. F 为 Fisher 检验值. q^2 为 leave-one-out 交互验证回归系数的平方,定义为 $q^2 = (SSY - PRESS)/SSY$,其中 SSY 是因变量和平均值之间的平方偏差和,PRESS 是采用 leave-one-out 交互验证得到的预测误差和的平方. 为了验证模型的实际预测能力,我们还随机挑选了 5 个化合物组成了预测集,预测集中的分子不参与模型的建立.

2 结果与讨论

2.1 构效关系模型

表 1 构效关系计算中采用的分子参数及缩写

Table 1 Abbreviations used in the QSAR analysis of the data set

Abbreviation	Definition	Abbreviation	Definition
Dip	Dipole vector	lg P	Partition coefficient
RadOfGyration	Radius of gyration	Fh2o	Aqueous desolvation free energy
Area	Surface area	Foct	1-octanol desovation free energy
MW	Molecular weight	MR	Molecular refractivity
Density	Molecular density	JX	Balaban index
Rotbonds	Number of rotatable bonds	Zegreb	Zagrebe index
Hbond acceptor	Number of H bond acceptors	lg Z	Hosaya index
Hbond donor	Number of H bond donors	Hf	Final heat of formation

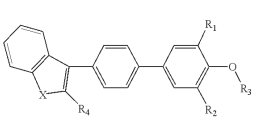
表 2 化合物活性的实验值及预测值

Table 2 Experimental and calculated biological activities of benzofuran and benzothiophene biphenyls

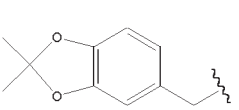
No.	R_1	R_2	R_3	R_4	X	$lg(1/IC_{50})$ Obs. [6]	$lg(1/IC_{50})$ Cal. [b]	Residue
1	H	H	H	benzyl	O	0.04	−0.12	−0.16
2	H	H	$CH(CH_2Ph)CO_2H(R)$	benzyl	O	0.46	0.74	0.28
3	H	H	$CH(Ph)CO_2H(R)$	benzyl	O	0.40	0.56	0.16
4	H	H	CH_2Ph-4-CO_2H	benzyl	O	0.44	0.57	0.13
5	NO_2	H	$CH(CH_2Ph)CO_2H(R)$	benzyl	O	0.64	0.44	−0.20
6[a]	H	H	$CH(CH_3)CO_2H(R)$	benzyl	O	−0.12	0.18	0.30
7	CH_3	CH_3	$CH(CH_2Ph)CO_2H(R)$	benzyl	O	1.13	1.16	0.03
8	cyclopentyl	H	CH_2COOH	benzyl	O	0.77	0.65	−0.12
9	$NHCH_2CO_2H$	H	CH_2CH_2Ph	benzyl	O	1.09	0.96	−0.13
10	$NHCH_2CH_2CO_2H$	H	CH_2CH_2Ph	benzyl	O	0.85	1.08	0.23
11	$NHCOCH_2CH_2CO_2H$	H	H	benzyl	O	0.04	0.10	0.06
12	$NHCOCH=CHCO_2H$	H	H	benzyl	O	0.34	0.09	−0.25
13[a]	$NHCO-C_6H_4$-2-CO_2H	H	H	benzyl	O	0.80	0.40	−0.40
14	H	H	$CH(CH_2Ph)CO_2H(R)$	benzyl	S	1.02	0.72	−0.30
15	Br	H	H	benzyl	S	−0.03	0.22	0.25
16	Br	Br	H	benzyl	S	0.35	0.48	0.13
17[a]	I	I	H	benzyl	S	0.28	0.62	0.34
18	Br	H	$CH(CH_2Ph)CO_2H(R)$	benzyl	S	1.24	1.03	−0.21
19	Br	Br	$CH(CH_2Ph)CO_2H(R)$	benzyl	S	1.60	1.42	−0.18
20	4-Cl-Ph	H	$CH(CH_2Ph)CO_2H(R)$	benzyl	S	1.28	1.61	0.33
21	Br	H	CH_2COOH	benzyl	S	0.44	0.35	0.09
22	Br	Br	CH_2COOH	benzyl	S	1.00	0.61	−0.39
23	4-OCH_3-Ph	H	CH_2COOH	benzyl	S	1.10	0.81	−0.29
24	2, 3-di-OCH_3-Ph	H	CH_2COOH	benzyl	S	1.15	0.98	−0.17
25[a]	4-OCH_3-Ph	Br	CH_2COOH	benzyl	S	1.54	1.12	−0.42
26	3, 4, 5-tri-OCH_3-Ph	H	CH_2COOH	benzyl	S	1.00	1.13	0.13
27	2, 4-di-OCH_3-Ph	Br	CH_2COOH	benzyl	S	1.33	1.33	0
28	4-OCH_3-Ph	4-OCH_3-Ph	CH_2COOH	benzyl	S	1.60	1.54	−0.06
29	H	H	H	butyl	O	0.13	−0.05	−0.18
30	H	H	CH_2COOH	butyl	O	−0.34	0.01	0.35
31	H	H	H	butyl	S	0.15	0.04	−0.11
32	H	H	$CH(CH_2Ph)CO_2H(R)$	butyl	S	0.77	0.97	0.20
33	H	H	$CH(Ph)CO_2H(R)$	butyl	S	0.96	0.76	−0.20
34[a]	H	H	$CH(CH_2Ph)CO_2H(R)$	benzoyl	O	0.17	0.57	0.40
35	H	H	$CH(CH_2Ph)CO_2H(R)$	CH(OH)phenyl	O	0.96	0.78	−0.18
36	H	H	H	4-OH-benzyl	S	−0.03	0.04	0.07
37	H	H	H	2, 4-di-OH-benzyl	S	0.24	0.13	−0.11
38	H	H	$CH(CH_2Ph)CO_2H(R)$	4-OCH_3-benzyl	S	1.11	0.94	−0.17
39	H	H	$CH(CH_2Ph)CO_2H(R)$	2, 4-di-OCH_3-benzyl	S	1.07	1.16	0.09
40	H	H	$CH(CH_2Ph)CO_2H(R)$	2, 4-di-OH-benzyl	S	0.92	0.98	0.06
41	H	H	$CH(CH_2Ph)CO_2H(R)$	c	S	1.11	0.99	−0.12
42	H	H	$CH(CH_2Ph)CO_2H(R)$	d	S	−0.06	0.48	0.54
43	H	H	$CH(CH_2Ph)CO_2H(R)$	e	S	−0.19	0.57	0.76

[a]These compounds were used as a test set and not included in the derivation of equations.

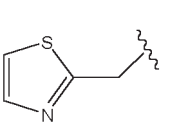
[b]The values of $lg(1/IC_{50})$ were calculated using Eq. 6.



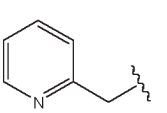
compound structure



c



d



e

计算结束以后, 从精华种群中, 就可以得到 50 个包含 3 个分子参数的构效关系模型以及 50 个包含 4 个分子参数的构效关系模型. 最佳的 5 个三参数的构效关系模型如方程 1 ~ 5 所示; 10 个最佳的四参数的构效关系模型如方程 6 ~ 15 所示.

1) $\lg(1/IC_{50}) = -5.40 + 0.0055\mathrm{Area} + 1.18JX'' + 0.154728\lg P$

$r^2 = 0.75$, $q^2 = 0.70$, $F = 38.74$, PRESS = 3.62

2) $\lg(1/IC_{50}) = -3.22 + 0.0021\mathrm{Area} - 0.020Foct'' + 0.23\lg P$

$r^2 = 0.73$, $q^2 = 0.67$, $F = 34.60$, PRESS = 3.89

3) $\lg(1/IC_{50}) = -7.20 + 0.0070\mathrm{Area} + 2.23JX'' + 0.0096MR$

$r^2 = 0.72$, $q^2 = 0.67$, $F = 32.61$, PRESS = 3.96

4) $\lg(1/IC_{50}) = -2.63 + 0.0040\mathrm{Area} - 0.033Dip + 0.13\lg P$

$r^2 = 0.71$, $q^2 = 0.66$, $F = 31.60$, PRESS = 4.13

5) $\lg(1/IC_{50}) = -5.57 + 1.34JX + 0.0081\mathrm{Area} - 0.066\text{Hbond acceptor}$

$r^2 = 0.71$, $q^2 = 0.65$, $F = 31.48$, PRESS = 4.14

6) $\lg(1/IC_{50}) = -3.21 + 0.21\lg P + 0.0077\mathrm{Area} - 0.015Zagreb - 0.021Foct$

$r^2 = 0.80$, $q^2 = 0.74$, $F = 33.81$, PRESS = 2.62

7) $\lg(1/IC_{50}) = -3.57 + 0.21\text{Hbond donor} - 0.025Zagreb - 0.17\text{Rotbonds} + 0.019\mathrm{Area}$

$r^2 = 0.80$, $q^2 = 0.74$, $F = 33.29$, PRESS = 2.63

8) $\lg(1/IC_{50}) = -4.52 + 0.0053\mathrm{Area} + 0.13\lg P - 0.042Dip + 0.89JX$

$r^2 = 0.80$, $q^2 = 0.74$, $F = 33.03$, PRESS = 2.65

9) $\lg(1/IC_{50}) = -2.28 - 0.0088Zagreb + 0.0085 \cdot \mathrm{Area} - 0.23\text{RadOfGyration} + 0.14\lg P$

$r^2 = 0.80$, $q^2 = 0.73$, $F = 32.62$, PRESS = 2.66

10) $\lg(1/IC_{50}) = -1.98 + 0.0036\mathrm{Area} + 0.22\lg P - 0.32\text{RadOfGyration} + 0.067\ \text{Rotbonds}$

$r^2 = 0.79$, $q^2 = 0.73$, $F = 31.92$, PRESS = 2.70

11) $\lg(1/IC_{50}) = -4.66 - 0.017Fh2o + 0.35\lg P + 0.81JX + 0.096\text{Rotbonds}$

$r^2 = 0.79$, $q^2 = 0.73$, $F = 31.86$, PRESS = 2.71

12) $\lg(1/IC_{50}) = -2.25 - 0.083Dip + 0.0026MW - 0.0042MR + 0.0037\mathrm{Area}$

$r^2 = 0.79$, $q^2 = 0.73$, $F = 31.49$, PRESS = 2.73

13) $\lg(1/IC_{50}) = -3.65 + 0.0052\mathrm{Area} - 0.13 \cdot \mathrm{RadOfGyration} + 0.54JX + 0.16\lg P$

$r^2 = 0.79$, $q^2 = 0.73$, $F = 31.36$, PRESS = 2.75

14) $\lg(1/IC_{50}) = -3.25 + 0.24\lg P - 0.0056MR + 0.14\text{Hbond donor} + 0.0039\mathrm{Area}$

$r^2 = 0.79$, $q^2 = 0.73$, $F = 31.12$, PRESS = 2.75

15) $\lg(1/IC_{50}) = -2.31 + 0.0054MW + 0.15\lg P - 2.03\text{Density} + 0.89JX$

$r^2 = 0.79$, $q^2 = 0.72$, $F = 30.85$, PRESS = 2.78

一般来讲, 当采用多重线性回归来建立构效关系模型时, 要尽量避免在一个模型中包含互相相关的变量, 因为变量之间的相关往往会使计算结果偏离正确的数学模型. 我们对以上模型所有分子参数进行了相关分析, 发现其中所有的参数均是独立的.

为了检验模型的实际预测能力, 我们采用最佳的四参数模型预测了预测集中 5 个分子的活性预测值. 表 2 显示了这 5 个化合物的预测活性, 图 1 显示了预测集中化合物的预测活性和实际活性之间的线性相关图. 从预测值上看, 最佳的四参数模型对于预测集中的 5 个化合物的活性都可以比较准确地预测. 图 2 则显示了训练集中化合物的预测活性和实际活性之间的线性相关图.

从模型的回归能力上看, 四参数的模型并没有明显的差别. 我们用这些模型预测了预测集中的化合物, 发现对于大部分分子, 这些模型给出的预测结果基本相似, 但对于某些分子给出的预测结果也存在较大的差别. 比如模型 7 对分子 1 给出的预测结果为 0.98, 而模型 11 对分子 1 给出的预测结果却

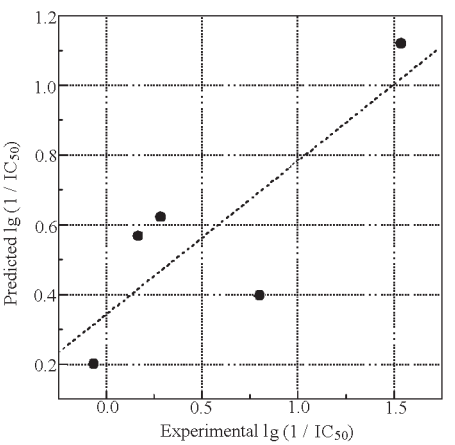


图 1 预测集中的化合物预测活性和实际活性之间的线性相关图

Fig. 1 Plot of the actual and predicted activity for five tested compounds

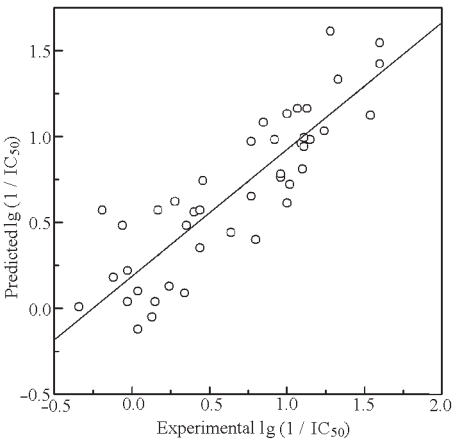


图 2　训练集中的化合物预测活性和实际活性之间的线性相关图

Fig. 2　Comparison of experimental lg(1/IC_{50}) with predicted lg(1/IC_{50}) obtained from Eq. 6

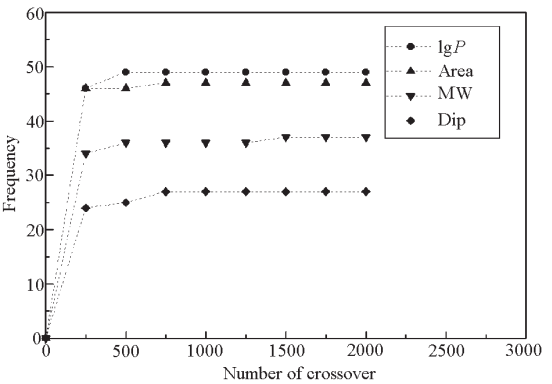


图 3　精华群中几种重要的分子参数的数目随循环次数的变化

Fig. 3　Change in the descriptor used in the evolution procedure of the elite population with four descriptors

为 0.39. 这说明某些模型还是存在了过拟合的情况, 原因在于随机相关以及训练集中的分子不够充分,而且一个单一的构效关系模型可能只对某些化合物具有好的预测能力 . 这其实是构效关系研究中一个广泛存在的问题 .

2.2　重要的分子参数

图 3 显示了在遗传优化过程中, 精华种群中几种重要的分子参数的数目随循环次数的变化 . 从这个图中可以看出, 不同的分子参数在精华种群中出现的频率是有很大差别的 . 有四种分子参数, 包括: lg *P*(分配系数), Area(表面积), MW(分子量)以及 Dip(偶极距), 它们出现的频率要明显高于其它的参数 . 当遗传算法优化达到平衡以后, 精华种群中所有模型中都包含 lg *P* 这个参数;94% 的模型中包含 Area 这个参数;70% 的模型中包含 MW 这个参数;50% 的模型中包含 Dip 这个参数 . 因此, 这几个参数可能是影响化合物活性最重要的参数 . 除了这几个参数, RadOfGyration(分子回旋半径)、Hbond acceptor(氢键受体数目)、Rotbond(分子中可旋转键个数)以及 Density(分子密度)在精华种群的模型中也有较高的出现频率 . 而其它的参数在模型中出现的频率就很低了, 因此它们对化合物活性的贡献可能比较小 .

因为 lg *P*、Area、MW 和 Dip 四个参数在模型中出现频率较高, 他们可能对化合物的活性有着非常重要的影响 . lg *P* 是出现频率最高的参数 . 从模型方程的系数为正可以看出, 疏水作用越大, 抑制剂的活性越高 . 这可能是因为 PTP1B 活性区域是一个疏水口袋, 而抑制剂就处在这个活性口袋中 . 如果它含有疏水性强的基团, 就会增加配体分子和受体分子之间的疏水相互作用, 从而提高抑制剂的活性. 如含有较多苯环的分子 28, 是活性最高的分子;而含有苯环较少的分子, 如分子 30, 则活性很低 . Area 和 MW 的模型方程的系数也为正, 说明分子的尺寸越大, 抑制剂的活性越高 . 当分子的尺寸增加时, 配体分子可以和受体分子产生更多的表面接触, 进而产生更好的几何匹配 . 当然, 尺寸效应会存在一定的范围, 当分子增加到一定的大小时, 尺寸效应的影响就不明显了, 甚至可能损害配体和受体之间的几何匹配 . 而 Dip 的模型方程的系数为负, 说明分子的极性越大, 越不利于它的活性的提高 . 其原因可能是配体分子的偶极和受体分子的偶极的方向是相同的,存在一定的排斥, 因此配体分子偶极的增加可能不利于配体分子和受体分子之间的能量匹配.

实际上这四个参数对于活性的影响是一致的 . 疏水性强的基团, 如苯基、大的烃基等, 他们的体积、分子量往往比较大, 而他们的极性较小;疏水性小的基团如羟基、氨基等, 他们的体积、分子量往往较小, 而极性较大 . 反映在方程的系数上, 就是 lg *P*、Area 和 MW 对分子活性的影响是一致的, 都为正相关;而 Dip 对活性的影响正相反, 为负相关 . 因此总结以上参数对抑制剂活性的影响, 可以主要归结为疏水作用、空间效应和分子的极性 . 而实际上这几个因素之间是有联系的 .

3 结 论

将遗传算法引入二维定量构效关系中, 结合线性回归和交叉验证方法, 对一系列 43 个苯并呋喃/噻吩联二苯类 PTP1B 抑制剂作了二维定量构效关系的研究, 得到了一组效果较好的定量构效关系模型, 这为以后化合物的改造提供了定量的预测工具. 计算分析了不同参数对抑制剂活性影响的差别. 研究表明, 4 个参数 lg*P*、Area、MW 和 Dip 是影响化合物活性的最重要的参数. 即疏水作用和立体效应是受体和抑制剂之间作用的主要因素, 为今后抑制剂的设计和改造提供了指导.

致谢 感谢北京大学化学与分子工程学院的侯廷军博士和徐筱杰教授提供了本文做遗传算法的二维构效关系计算的程序并给予了大力指导和支持.

References

1 Zhang, Z. Y. *Curr. Opin. Chem. Biol.,* **2001, 5**: 416
2 Kenney, B. P. ; Ramachandran, C. *Biochem. Pharmacol.,* **2000, 60**: 877
3 Goldstein, B. J. *Receptor,* **1993, 3**: 1
4 Elchebly, M. ; Payette, P. ; Michaliszyn, E. ; Cromlish, W. ; Collins, S. ; Loy, A. L. ; Normandin, D. ; Cheng, A. ; Hagen, J. H. ; Chan, C. C. ; Ramachandran, C. ; Gresser, M. J. ; Tremblay, M. L. ; Kennedy, B. P. *Science,* **1999, 283**: 1544
5 Klaman, L. D. ; Boss, O. ; Peroni, O. D. ; Kim, J. K. ; Martino, J. L. ; Zabolotny, J. M. ; Moghal, N. ; Lubkin, M. ; Kim, Y. B. ; Sharpe, A. H. ; Krongrad, A. S. ; Shulman, G. I. ; Neel, B. G. ; Kahn, B. B. *Mol. Cell. Biol.,* **2000, 20**: 5479
6 Malamas, M. S. ; Sredy, J. ; Moxham, C. ; Katz, A. ; Xu, W. ; McDevitt, R. ; Adebayo, F. O. ; Sawicki, D. R. ; Seestaller, L. ; Sullivan, D. ; Taylor, J. R. *J. Med. Chem.,* **2000, 43**: 1293
7 Pan, Y. M. ; Ji, M. J. ; Ye, X. Q. ; Kuang, P. X. *Chin. J. Org. Chem.,* **2003, 23**(2): 167 [潘咏梅, 计明娟, 叶学其, 邝平先. 有机化学(*Youji Huaxue*), **2003, 23**(2): 167]
8 Hou T. J. ; Wang, J. M. ; Li, Y. Y. ; Xu, X. J. *Chin. Chem. Lett.,* **1998, 9**: 651
9 Hou, T. J. ; Wang, J. M. ; Xu, X. J. *Chemometr. Intell. Lab.,* **1999, 45**: 303
10 Hou, T. J. ; Wang, J. M. ; Liao, N. ; Xu, X. J. *J. Chem. Inf. Comp. Sci.,* **1999, 39**: 775
11 Hou, T. J. ; Xu, X. J. *J. Mol. Graph. Model.,* **2001, 19**: 455
12 Hou, T. J. ; Xu, X. J. *J. Mol. Model.,* **2002, 8**: 337
13 Tripos 6. 3 Users Guide. St. Louis, USA: Tripos Inc., 1996
14 Dauber-Osguthorpe, P. ; Roberts, V. A. ; Osguthorpe, D. J. ; Wolff, J. ; Genest, M. ; Hagler, A. T. *Proteins: Struct., Funct., Genet.,* **1988, 4**: 31
15 Ghose, A. K. ; Crippen, G. M. *J. Comput. Chem.,* **1986, 7**: 565
16 Hopfinger, A. J. Conformational properties of macromolecules. New York: Academic Press, 1977
17 Dragon v2. 1 user guide. Milano Chemometrics and QSAR Research Group, 2002. http: // www. disat. unimib. it/chm/

Applications of Genetic Algorithms on 2D-QSAR Analysis of Benzofuran and Benzothiophene Biphenyls as PTP1B Inhibitors *

Pan Yong-Mei Ji Ming-Juan
(*Graduate School of Chinese Academy of Sciences, Beijing 100039*)

Abstract Quantitative structure-activity relationships (QSARs) for 43 benzofuran and benzothiophene biphenyls were studied. By using a genetic algorithm (GA), a group of multiple regression models with high fitness scores (r^2 was up to 0. 70) were generated. From the statistical analyses of the descriptors used in the evolution procedure, four of them, including the partition coefficient(lg*P*), the molecular surface area(Area), the molecular weight(MW), and the dipole vector(Dip) were found to be the principal features affecting the biological activity. For example, the molecular surface area appeared in 94% of the models in the elite populations. That is to say, the hydrophobic interactions between the inhibitors and the receptors are very important to the biological activity, which supplies a guide for the design and reconstruction of new PTP1B inhibitors.

Keywords: 2D-QSAR, Genetic algorithm, PTP1B inhibitors

Received: January 13, 2003; Revised: March 13, 2003. Correspondent: Ji Ming-Juan (E-mail: jmj@gscas. ac. cn; Tel: 010-88258593; Fax: 010-88256092). * The Project Supported by NSFC (20273083, 29992590)