中国科学 B 辑:化学 2008 年 第 38 卷 第 3 期: 259 ~ 268
www.scichina.com chem.scichina.com
《中国科学》杂志社
SCIENCE IN CHINA PRESS

将潜因多元建模技术与遗传虚拟筛选模式相结合: 改善定量构效关系

李志良[①②*], 田菲菲[①②], 吴世容[①②], 杨善彬[①②③], 杨胜喜[①②], 张巧霞[①②], 周原[①②③], 覃仁辉[①②③], 梅虎[②③④], 陈刚[②④*], 李根容[①②]

① 重庆大学化学化工学院, 重庆 400044;
② 重庆大学生物医学工程教育部与重庆市重点实验室, 重庆 400044;
③ 重庆大学生物工程学院, 重庆 400044;
④ 新加坡国立技术学院生命科学技术中心, 新加坡 139651
* 联系人, E-mail: zlli-cqu@163.com or zlli2662@163.com; gchen@sp.edu.sg

收稿日期: 2005-05-27; 接受日期: 2007-07-15
国家高技术研究发展(863)计划(批准号: 2006AA02Z312)、霍英东基金(批准号: 98-7-6)、湖南大学化学生物传感与计量学国家重点实验室基金(批准号: 2005012)、国家"春晖计划"教育部启动基金(批准号: 99-4-4+37)、重庆应用基础研究基金(批准号: 01-3-6)及重大自主创新基金攻关课题(批准号: 030506+04-9-9)资助项目

摘要 提出了改进的定量构效关系(^{mm}QSAR)研究方法, 即认为在知晓配基与受体相互作用模式前提下建立一种较可靠的定量构效关系模型, 从而克服了传统做法中仅根据样本集分子自身信息来构建预测模型的某些弊端. 将此思路应用于功能肽/蛋白质亲和活性考察, 采用遗传算法(GA)筛选虚拟受体结合靶点及相互作用模式, 结合偏最小二乘(PLS)潜因多元建模技术, 通过交互检验均方根误差(RMS_{CV})作适应度评价函数筛选变化位点和受体残基, 完成种群进化以最优个体为最终确定模型, 得到一种新颖 QSAR 方法(^{mm}QSAR): 基于功能肽/蛋白质作用模式的遗传虚拟筛选(暂称 GVSPPC). 该法成功解决了诸多 QSAR 研究难题, 即大多数情况下受体结构未知而难以了解的配基与之结合方式. 分别使用生物功能寡肽和多肽体系对 GVSPPC 加以检验, 其结果表明, GVSPPC 得到了优于传统方法 QSAR 结果(R^2_{cu} > 0.75~0.91, Q^2_{cv} > 0.71~0.86, E_{RMS} 为 0.19~0.95)且深入阐明配基与受体间作用机理, 物理意义较为明确; 同时计算选中模式作用能并与样本活性建立 PLS 模型, 以其交互检验均方根误差(E_{RMScv})作评价函数完成种群进化. 通过大规模迭代直到规定的终止条件或到达最大繁殖代数, 以该过程最优个体作为最终确定 QSAR 模型.

关键词
功能肽与蛋白质结合模式
遗传虚拟筛选
潜因多元建模
定量构效关系
偏最小二乘
遗传算法
交互检验
评价函数

在许多生命活动中功能肽等配基与蛋白质等受体形成肽/蛋白质复合体及其相互作用具有着极其重要的作用, 在化学、生物学、医药学、免疫学等众多学科备受关注 [1~3], 其中相关常见方式包括抗原与抗体、酶与底物、受体与类肽、靶标与药物等等的识别和结合. 将基于统计技术的传统定量构效关系(QSAR)用于揭示肽类生物活性与序列特征间内在本质, 至今已取得一定成效 [2~8]. 代表性工作主要包括

Hellberg等[2,3]提出的天然氨基酸Z标度(*Z*-s); Collantes等[4]以CNDO/2 法计算得到的全向表面积(ISA)及电荷指数(ECI)以及Zaliani[5]从结构出发产生的MS-WHIM指数等. 但这些传统方法往往采用以下步骤实现: a)搜集氨基酸各类理论和实验所得参数来表征生物序列并预测功能活性; b)利用主成分分析(PCA)技术压缩变量和提取信息而得到单元结构(氨基酸或碱基)描述子; c)使用该类结构描述子按相应残基出现顺序表征肽链等生物序列, 继而构建功能活性的统计模型. 综观其过程可知, 尚未考虑受体即作用对象及相关因素; 所得结构描述子对应QSAR模型, 物化意义不甚清晰而难以完成配基与受体相互作用的直接分析. 在本文中考虑到功能肽/蛋白质结合模式主要取决于侧链短程相互作用且辅以非键效应, 创建基于肽/蛋白质结合模式的遗传虚拟筛选(为方便起见暂称为GVSPPC), 以氨基酸侧基为单元定义 4 类(静电、立体、疏水和氢键)相互作用模式, 在此基础上利用偏最小二乘(PLS)作为模拟方法, 评价函数通过遗传算法(GA)对变化位点受体残基的类型及作用方式进行筛选, 以期获得物化意义较明确的QSAR模型和优于传统方法的QSAR结果($R^2_{cu}\approx0.75$~0.91, $Q^2_{cv}\approx0.71$~0.86, $E_{RMS}\approx0.19$~0.95).

1 原理与方法

1.1 功能肽与蛋白质间相互作用及复合模式

1986 年, Geysen等[8]研究得出很有意义的结论, 即: a) 绝大部分结合能由所结合的两个分子功能肽和蛋白质中少数几个关键基团的弱相互作用提供; b) 含几个关键残基的相关线性寡或多肽可模拟蛋白质上折叠表位或抗原决定簇. 而在 1988 年, Cramer等[9]则认为药物小分子和生物大分子间相互作用大多是可逆非键合效应, 如静电作用(EI)和立体作用(SI)等力场形式, 当然在本文中认为具体表现还应包括疏水作用(HI)、氢键作用(HB, 含授受方式: HA和HD)、范德瓦尔斯力(WF)等等. 基于上述观点, 可认为: 功能肽/蛋白质结合是氨基酸核心残基短程非键相互作用的综合结果, 且仅最接近的若干短程侧链基团才是维持整个复合体系稳定的关键所在. 当不同肽序列与相同蛋白质受体发生复合和识别时, 将主要体现在侧链变化上.

为模型简化起见, 在GVSPPC中将侧链残基抽象为实际van der Waals体积之球体(表 1)[10]. 在一组

表 1 在 GVSPPC 中使用到的 20 个标准天然氨基酸的相关参数

	氨基酸	V_{vdW} a)[10]	R b)	$q_{NCI}\times10^3$ c)	A_{SASA} d)[11]	a e)[12]	C d, f)	C a,g)	ECI	ZZ_1[13]	ZZ_2	ZZ_3	ZZ_4	ZZ_5
A	Ala	57.02	2.39	7.187	118.1	0.42	−3.27	0.00	0.05	0.24	−2.32	0.60	−0.14	1.30
C	Cys	149.07	3.29	−36.613	146.1	1.34	−3.26	1.04	0.15	0.84	−1.67	3.71	0.18	−2.65
D	Asp	210.51	3.69	−23.824	158.7	−1.05	−5.82	2.80	1.25	3.98	0.93	1.93	−2.46	0.75
E	Glu	328.37	4.28	6.802	186.2	−0.87	−8.00	2.80	1.31	3.11	0.26	−0.11	−3.04	−0.25
F	Phe	552.98	5.09	37.552	222.8	2.44	−7.63	0.00	0.14	−4.22	1.94	1.06	0.54	−0.62
G	Gly	2.79	0.87	179.052	88.1	0.00	−1.09	0.00	0.02	2.05	−4.06	0.36	−0.82	−0.38
H	His	376.94	4.48	−10.692	202.5	0.18	−6.27	3.32	0.56	2.47	1.95	0.26	3.90	0.09
I	Ile	376.71	4.48	21.631	181.0	2.46	−9.81	0.00	0.09	−3.89	−1.73	−1.71	−0.84	0.26
L	Leu	378.76	4.49	51.672	193.1	2.32	−9.81	0.00	0.10	−4.28	−1.30	−1.49	−0.72	0.84
K	Lys	453.63	4.77	17.708	225.8	−1.35	−12.54	4.24	0.53	2.29	0.89	−2.49	1.49	0.31
M	Met	388.72	4.53	2.683	203.4	1.68	−7.63	1.04	0.34	−2.85	−0.22	0.47	1.94	−0.98
N	Asn	229.72	3.80	5.392	165.5	−0.82	−6.00	3.52	1.31	3.05	1.62	1.04	−1.15	1.61
P	Pro	227.90	3.79	239.531	146.8	0.98	−6.54	0.00	0.16	−1.66	0.27	1.84	0.70	2.00
Q	Gln	348.61	4.37	49.211	193.2	−0.30	−8.18	3.52	1.36	1.75	0.50	−1.44	−1.34	0.66
R	Arg	589.46	5.20	43.587	256.0	−1.37	−12.27	5.90	1.69	3.52	2.50	−3.50	1.99	−0.17
S	Ser	92.04	2.80	4.627	129.8	−0.05	−4.46	1.40	0.56	2.39	−1.07	1.15	−1.39	0.67
T	Thr	193.41	3.59	3.352	152.5	0.35	−6.64	1.40	0.65	0.75	−2.18	−1.12	−1.46	−0.40
V	Val	256.74	3.94	57.004	164.5	1.66	−7.63	0.00	0.07	−2.59	−2.64	−1.54	−0.85	−0.02
W	Trp	793.51	5.74	37.977	266.3	3.07	−7.36	1.66	1.08	−4.36	3.94	0.59	3.44	−1.59
Y	Tyr	611.50	5.27	23.599	236.8	1.31	−7.95	1.40	0.72	−2.54	2.44	0.43	0.04	−1.47

a) 侧链 van der Waals 体积; b) 侧链半径, 由 V_{vdW} 计算得到; c) 侧链净电荷指数; d) 残基溶剂可及表面积; e) 侧链疏水参数, 被表示为相对于甘氨酸的溶解自由能变化; f) 氢键给体贡献因子; g) 氢键受体贡献因子

肽类样本集中以结合最强(或活性最高)者为模板并认为该肽与受体侧链能恰好匹配(图 1 实线), 而其他肽将可能重叠或脱离(图 1 虚线).

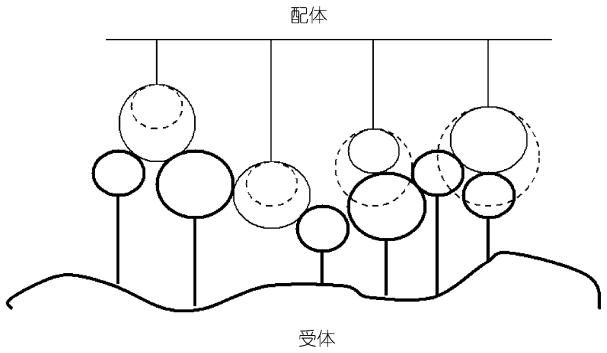

图 1 GVSPPC 中配体(如肽类似物)与受体(如蛋白质)匹配方式示意图

将残基视为整体, 则相互作用情况包括静电、立体、疏水和氢键等效应. 静电作用(EI)是一类重要非键效应, 用经典 Coulomb 定律来计算侧链间静电相互作用能:

$$E_{ij}(\mathrm{EI}) = \frac{1}{4\pi\varepsilon_0}\frac{q_{i\mathrm{CNI}}q_{j\mathrm{NCI}}}{d_{ij}} = \frac{1}{f_0}\frac{q_{i\mathrm{CNI}}q_{j\mathrm{NCI}}}{d_{ij}}, \quad (1)$$

其中f_0为常数项; i, j分别代表配体和受体侧链; d为配体/受体侧链球心距离; q_{NCI}为本文中所定义的用于表征侧链整体电性相关指数称为侧链净电荷指数(NCI, q_{NCI}), 即残基原子净电荷算术加和值, 源自但不同于Collantes等[4]提出的电荷指数(ECI)(参见附表 2). ECI 指数用于描述侧链电荷分布, 定义为侧链原子所带净电荷绝对值加和, 显然只反映侧基电荷分布不均匀性和极化特征; NCI则为整体电性量度, 计算如下: 先用软件Chemoffice 8.0 自带数据库生成 20 个氨基酸分子初始立体结构; 继以分子模拟HyperChem 7.5 作分子力学(MM+力场)优化构象; 再用量化软件Gaussian 98W在密度泛函(DFT)[14,15]水平基于广义梯度近似法最终优化得三维分子结构, 以Mulliken布居分析法[16]借单点(s-p)形式计算原子净电荷数量(因基函数非正交, DFT对Mulliken电荷作修正); 最后加和得相应NCI指数q_{NCI}(表 1). 其中DFT采用B3LYP[17]泛函类型, 几何优化时选分子轨道基函为双重分裂基组并在重原子上加入d轨道极化函数(6-31G*), 而单点计算所用轨道基组则再向氢原子加入p轨道极化函数(6-31G**). 立体作用(SI)是原子间偶极诱导或偶极-非偶作用. 传统上表示为Lennard-Jones 6~12 势能函数, 而在近距时计算作用能量曲线变化太陡而致结果涨落. 在GVSPPC中用MM3 力场[18]的Buckingham势能函数来计算立体相互作用能:

$$E_{ij}(\mathrm{SI}) = \varepsilon_0\left[\frac{6}{\alpha-6}\exp\frac{-(-d_{ij}-R_{ij}^*)}{R_{ij}^*} - \frac{\alpha}{\alpha-6}\left(\frac{R_{ij}^*}{d_{ij}}\right)^6\right], \quad (2)$$

式中E_{ij}(SI)为立体作用能; ε_0为势阱深校正系数; α为调整因子, 取 14; R_{ij}^*为两侧链半径和(由van der Waals体积按球积公式计算, 参见表 1). 显然上述缺点因排斥项使用准Gaussian形指数函数表示而被有效克服. 疏水作用(HI)是药物与生物结合重要影响因素, 往往表现为体系熵变而难用统一公式描述. 鉴于Kellogg等[19~22]提出的Hint方法最为有效并被成功采用[23~28], 故在GVSPPC中用以定义疏水作用, 但因前面立体势能已单独考虑而删去van der Waals作用项(r_{ij}):

$$E_{ij}(\mathrm{HI}) = [A_{i\mathrm{SASA}}\cdot a_i\cdot A_{j\mathrm{SASA}}\cdot a_j\cdot\exp(-d_{ij})\cdot T_{ij}], \quad (3)$$

其中E_{ij}(HI)为疏水相互作用能; A_{SASA}为残基X在三肽G-X-G中溶剂可及面积(SASA, $A_{i\mathrm{SASA}}$, 表 1, Rose G D等[11])即以水分子(R_{vdW} = 1.4 Å)为探针在该残基表面滚动其球心所形成表面积; a为侧链疏水常数, 其量度取相对溶解自由能变化(表 1, $\Delta G = 1.36\times(\lg P(\mathrm{X}) - \lg P(\mathrm{G}))$为Eisenberg D和McLachlan A D定义[12]); T_{ij}为符号判别函数(取+1 或–1)以表明不同类型残基疏水作用熵变(表 2).

表 2 按Hint[19~22]定义不同疏水基团相互作用符号判别函数T_{ij}取值

类型	疏水性	亲水性
疏水性	+1	–1
亲水性	–1	–1

氢键(HB)是分子间特种氢与孤对电子间弱静电吸引, 相关作用强弱至今尚无有效计算公式, 往往将其视为静电和立体的特殊形式而不予考虑[9,29~31]或用指示变量来代表[32]. 在本文中借AMBER场[33]定义类Lennard-Jones10-12 函数来确定氢键强弱, 而势能阱深项系数以Raevsky等氢键热力学(HYBOT)[34,35]含 13500 多氢键体系数据库经$\Delta G = \beta C^{\mathrm{d}}\times C^{\mathrm{a}} + \beta_0$回归计算最优匹配状态下的自由能变化, 其$C^{\mathrm{d}}$和$C^{\mathrm{a}}$分别为氢键给受体自由能贡献因子. 故在GVSPPC中用上述

两法结合函数作侧链间氢键定量计算:

$$E_{ij}(\mathrm{HB})=\{g_0[(C^{a}_{i}\cdot C^{d}_{j}+C^{d}_{i}\cdot C^{a}_{j})\cdot[(R^{*}_{ij}/d_{ij})^{12}-2(R^{*}_{ij}/d_{ij})^{10}]\}. \quad (4)$$

式中E_{ij}(HB)为氢键相互作用能; 常数g_0为有效氢键结合比例; C^{d}和C^{a}分别为残基氢键贡献因子且具线性累加性[35], 即可简单表示为该体系所有非氢原子(受体)或氢原子(给体)算术加和, $(C_i^{a}\times C_j^{d}+C_i^{d}\times C_j^{a})$为$i$, j两侧链分别作氢键给受体时相互作用自由能贡献. 由此计算出20种氨基酸侧链氢键贡献因子(参见表1和附表 1, 取自文献[35]), 需附加说明的是, 因各常数项对模型结果最后没有影响, 故在计算中未予考虑; 由此使用上面4种非键势能作用函数计算得到的结果将被直接用于QSAR拟模中.

1.2 遗传虚拟受体筛选

实际常遇到已知配基肽类物与未知受体蛋白质结合强度(或活性)问题, 由此可分析结合强度与肽序特征间的内在联系, 找到受体靶点残基及相互作用, 继而发展更为可靠且容易解释的QSAR模型. 统计方法是有效途径, 但逐一检验由于氨基酸多样性及作用复杂性难以实现. Holland[36]发展遗传算法(Genetic Algorithm, GA), 模仿"适者生存, 优胜劣汰"进化法则, 用于解决大规模组合优化的非数值并行算法, 当今已比较成熟; 结合Wold等[37]提出的偏最小二乘(PLS)可有效解决变量间的多重共线性, 并适合在样本容量小于变量个数情况下建模和评价. 故GVSPPC中虚拟蛋白质受体及作用模式的筛选用GA完成, 基于Matlab 7.0中统计和GA工具箱经自编衔接函数GVSPPC-tool.M实现; 非键作用势能函数用C语言编写应用程序NBOND.EXE计算完成. 具体实现过程如下: 对n个位点变化的肽类样本集, 每位点都能与蛋白质受体的一或几个近距侧链发生某些非键作用, 可为静电、立体、疏水和氢键作用中一或几种. 则全面考察所有可能模式需将算法搜索空间域预定为每位点均与20种氨基酸单元发生各种非键作用, 即多肽与受体结合含$n\times(20\times4)=80n$种模式. 将所有模式以二进码成一条基因数目为$80n$的染色体串: 1为选中模式, 0则没选中; 同时计算样本选中模式作用能并与活性建立PLS模型, 以交互检验均方根误差($\mathrm{RMS_{CV}}$)作为适应度评价函数完成种群进化; 经大规模迭代至规定终止条件或达最大繁殖代数, 以其最优个体作最终模型, 完成QSAR拟模过程.

2 结果与讨论

2.1 二肽血管紧张素转化酶抑制剂 QSAR 分析

肾素与血管紧张素系统(rennin-angiotensin system, RAS)在调节人体血压过程中起着关键作用. 作为降血压药物作用靶标的血管紧张素转化酶(angiotensin converting enzyme, ACE)将肝脏产生的血管紧张素原经肾素催化断裂为非活性血管紧张素I的十肽后再催化裂解为具极强血管收缩功能的血管紧张素II的八肽. 在尚未探明ACE三维结构的研究[38,39]中, 借助同属性锌蛋白酶的羧肽酶A来分析复合方式得到ACE与底物的结合情况(参见图2), 可发现底物二肽(58个血管紧张素转化酶抑制剂ACEI, 该组样本见表3, $\mathrm{pIC_{50}}$

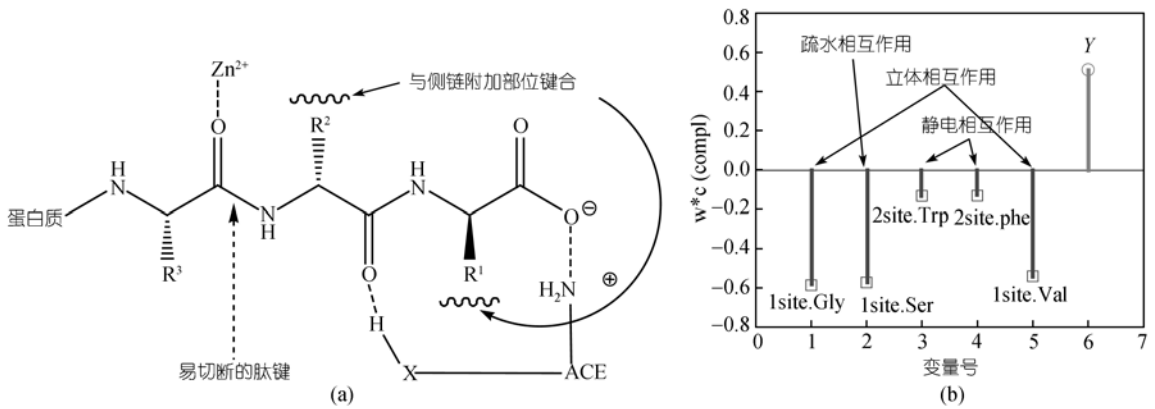

图 2 羧肽酶 A 复合物的结合情况

(a) ACE 与底物血管紧张素 I 结合示意; (b) GVSPPC 分析得到的二肽与 ACE 之间 5 种作用模式对 PLS 第一个主成分载荷贡献图(site: 位点; Y: 因变量)

表 3 按照HYBOT[34]定义的各类原子氢键贡献因子

H-bond donor	C_d			H-bond acceptor	C_a		
	Min	Max	Mean		Min	Max	Mean
—C—H	–2.06	–0.12	–1.09	—O—	0.78	2.01	1.40
—S—H	–1.92	–0.24	–1.08	═O	2.40	4.02	3.21
—N—H	–3.32	–0.50	–1.91	—S═O	1.16	3.25	2.21
PhO—H	–4.24	–0.76	–2.50	—S—	0.57	1.51	1.04
AlkO—H	–3.50	–1.06	–2.28	—N	0.58	3.65	2.12
—C(O)O—H	–4.78	–2.50	–3.64	═N	0.55	2.77	1.66

表示抑制活性[40])与ACE受体之间结合 4 类主要作用: a) 带负电荷羧基的底物与带正电荷氨基的ACE受体以离子键结合; b) 二肽第 2 位点羰基与ACE受体形成氢键; c) Zn^{2+}在外端将其切断; d) 侧链R^1和R^2辅助总体有利于亲和性结合. 因所有二肽主链相同, 样本对抑制活性影响的差异主要体现在侧链的影响因素(d)上. 对两残基(n=2)二肽样本集, GVSPPC搜索空间区域为 160(为 80n=80×2)种模式(GA设置见表 4). 重复实验 10 次得最优个体含 5 种作用模式(具体含义见表 5). 经分析这 5 种作用模式对PLS显著主成分载荷贡献(图 3)可知, 第一位点(二肽从右向左编号)是活性主要影响因素, 表现为与ACE受体上具较小侧链两残基G和V的立体相互作用, 说明该位点二肽配体含较大侧链, 有利空间匹配: 具较高抑制强度者该位点往往含较大基团如W, Y, F, P等(表 3 数据证实). 同时该位点可与ACE受体上Ser残基有一定疏水作用, 在GVSPPC中定义疏水参数为–0.05 的Ser侧链具一定亲水性, 可推测二肽抑制剂在此位点须具强疏水性残基, 这样与之结合才不致自由能损失过多, 这与

表 4 遗传算法的参数设置

参数	设置
软件	Matlab 7.0
数据预处理	自定标
适应度函数	PLS-交互检验均方根误差 $RMSE_{CV}$
PLS 算法	非线性迭代偏最小二乘(NIPALS)
主成分数选取标准	Q^2_{cv}
初始种群大小	500
最大繁殖代数	200
交叉互换数	2
交叉概率	50%
变异概率	0.5%
重复试验次数	10

表 5 GVSPPC 分析 ACE 受体与二肽抑制剂作用

位点 a)	作用对象 b)	作用类型
1	Gly	SI
1	Val	SI
1	Ser	HI
2	Trp	EI
2	Phe	EI

a) 二肽位点: ACE 抑制剂从右向左编号; b) 作用对象: 二肽位点对应作用的 ACE 氨基酸残基

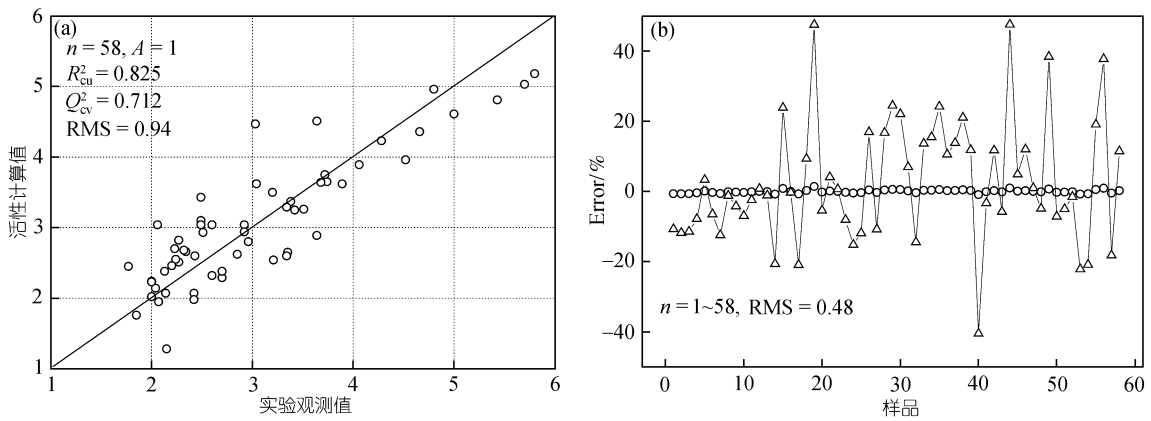

图 3 GVSPPC 的模拟计算结果

(a) GVSPPC 模型计算值与实验观测值的相关情况; (b) GVSPPC 模型计算值的残差分布散点图

该点出现W或Y并不降低抑制活性事实相符(表3, W和Y疏水参数分别为3.07和1.31). 而二肽第二位点与ACE受体Trp和Phe残基主要发生静电作用, 在GVSPPC中这两种氨基酸侧链电性指数很高且相近: q_{NCI}分别为0.037977和0.037552, 从而可互相取代. 为便于比较, 将GVSPPC所建5参数PLS模型相关统计量及部分文献报道结果共同列出(表6). 本文中模型含一个主成分(潜变量), 解释Y变量76.6%方差, 交互检验解释Y变量71.2%方差. 可看到该模型质量(R^2_{cu} = 0.765, Q^2_{cv} = 0.712)优于多数文献(R^2_{cu}为0.700~0.744, Q^2_{cv}=0.637或未知), 仅比Z scale拟合结果(R^2_{cu} = 0.770, 遗憾的是未给出Q^2_{cv}因而难以说明其外部预测能力)略低.

2.2 苦味二肽生物活性QSAR模拟

苦味活性二肽(bitter tasting thresholds, BTT)位于人体舌表面, 由味觉细胞组成的味蕾是重要的味觉受体器官, 具有刺激性的物质溶于唾液随之进入味蕾尖端小孔并被吸附于受体膜表面而产生味觉. 由于受体膜主要是由蛋白质等生命大分子物质组成, 由此可见能够刺激味觉产生的分子同样也是一类与蛋白质受体有关的化合物. 取48个苦味活性二肽(BTT)[7~9]验证GVSPPC描述子有效性, 同时建立QSAR模型, 其活性用浓度负对数pT表示. 同样每个苦味二肽可用24个GVSPPC描述子变量表征, 采用GPLS挑选变量, 参数设置为: 初始群体大小200, 最大遗传代数200, 收敛标准80%, 交叉频率50%, 变异概率0.5%. 从10个被训练模型中, 选择一个相对最优3个显著主成分模型, 分别解释68.0%、15.6%和7.61%方差: RMS = 0.187, R^2_{cu} = 0.916, Q^2_{cv} = 0.862. 表明所建模型对活性估计和预测效果均较好.

2.3 Camel抗体双位点残基变异特性

1998年, Transue等[42]研究表明, camel抗体蛋白cAb-Lys3中第99~112号残基序列能形成一凸出环状结构并插到其抗原蛋清溶菌酶(lysozyme)凹陷部位中以抑制该酶活性. 1996年, Desmyter等[43]已利用X射线衍射法测得抗体和抗原复合物晶体结构, 它在蛋白质结构数据库(PDB)[44]中编号为1 mel. 2002年, Genst等[45]通过多元实验设计系统改变101和105号两位点残基, 并测出变异后抗体对抗原亲和性K_d, 被表示为离解常数k_d与复合常数k_a之比. 这里列举自然品种及实验测定此两位点17次变异残基类型及其相应亲和性用($\ln K_d$表示). 抗体蛋白cAb-Lys3第99~112号残基序列为DS□IYA□YYECGHG, 其中"□"代表101和105位置变异位点. 因考察两位点, 故GVSPPC搜索空间亦为160种模式, 遗传设置同表3. 对10次重复计算所得最优个体包含11种模式: 分别为抗体101位点与受体残基Asn氢键和静电作用, Met静电和立体作用, Ala立体和疏水作用及Val静电作用; 而105位点则与受体Trp立体和疏水作用, Val疏水作用及Ala静电作用. 可看到cAb-Lys3与lysozyme结合过程非常复杂, 涉及残基间多种作用. 与此相呼应, 在Freyhult等[46]报告中也有类似结论即结合过程受众多物化因素影响: 他们用Sandberg等[13]氨基酸扩展描述子ZZ-scale(参见表1)及分子力学(Molecular Mechanics, MM)优化所得残基α碳原子3D空间坐标作初始参数, 继经遗传变量挑选最终建立3个主成分PLS模型(A=3, R^2_{cu}=0.800, Q^2_{cv} = 0.710, RMS未给)[47], 相比较本文用较少主成分获更优效果(A = 2, R^2_{cu} = 0.824, Q^2_{cv}=0.781, RMS = 0.95): 总体GVSPPC优于文献, 合理解释了camel抗体蛋白cAb-Lys3与抗原酶lysozyme复合情况.

2.4 抗菌18肽QSAR探索

抗菌肽(antimicrobial peptides, AMP)一直是工农产业、医药卫生等行业的研究重点, 微生物抗药性提

表6 关于ACE抑制剂QSAR模型比较

记号	方法	文献	模型	A[a]	R^2_{cu}[b]	Q^2_{cv}[c]	RMS[d]
1	ISA-ECI	[4]	PLS	2	0.700	nd[e]	nd[e]
2	MS-WHIM	[5]	PLS	2	0.708	0.637	0.54
3	Z scale	[40]	PLS	2	0.770	nd[e]	nd[e]
4	t scores	[41]	PLS	1	0.744	nd[e]	0.50
5	GVSPPC	本文	PLS	1	0.765	0.712	0.48

a) 潜隐变量数目; b) 复相关系数; c) 交叉检验复相关系数; d) 均方根误差; e) 文献未报道

高使之日益受到高度重视[48~51]. 在生物体内抗菌肽在防御外来微生物病原体自然免疫过程中发挥着极重要的作用, 过去几十年来在哺乳和两栖动物及昆虫和植物中相继发现众多抗菌活性多肽, 部分抗菌肽已作为候选药物进入临床, 有望取代传统抗生素成为一类新型抗菌药物. 本文选用的 12 个十八肽序列抗菌活性数据来自文献[51](标度为 2 h内杀死金黄色葡萄球菌(staphylococcus aureus ATCC 6583)数目取对数). 先用GVSPPC描述子表征肽序列. 用GPLS 挑选变量和回归建模得2 个显著成分, $R^2_{cu}=0.851$, $Q^2_{cv}=0.787$, $RMS_e=0.927$, $RMS_p=1.726$; 分别解释 Y变量57.2%和 22.2%方差. 所有样本PLS 得分都位于椭圆置信区间内, 表明无特异点且模型预测能力较强; 从各描述子PLS载荷中, 可看出与Y之间相关性强的变量分别是第 1 残基疏水性和氢键, 第 6 残基氢键, 第 11 残基立体, 第 13 残基电性, 第 13 和 14 残基疏水性(注意第 1 残基氢键和第 13 残基电性对抗菌活性呈较大负贡献). 根据这些位点相应氨基酸不同特性, 可改变残基获得具有更高活性的抗菌肽.

3 结语

对于肽类物与蛋白质的亲和能力及生物活性的研究是一件十分重要但是非常困难的任务, 这是由于该过程受到生物体内众多复杂因素影响而产生难以预知的效应. 传统的做法仅从药物配体自身结构特征出发, 通过统计学方法与技术建立相应的 QSAR 模型. 显然由于此类方法忽略了作为配体作用对象即受体(通常是蛋白质)的重要性从而使得模型难以具有较高可靠性能和较强泛化能力. 因此发展所谓“合理”模制 QSAR 具有良好理论和应用前景, 其中必须克服的问题是如何找到配基对受体的结合靶点及其相互作用情况. 本文提出了一种新颖的功能肽和蛋白质结合模式遗传虚拟筛选(GVSPPC)技术, 该法将基于非数值优化技术的遗传算法作为筛选工具, 在此基础上通过优选得到虚拟作用模式建立样本集的模制QSAR模型. 此法在一定程度上解决了QSAR建模所遇到的矛盾, 其思想有望进一步完善和推广, 特别是在分子模拟与药物设计中.

参考文献

1 Cooper W J, Waters M L, Molecular recognition with designed peptides and proteins, Curr Opinion in Chem Biol, 2005, 9(6): 627—631
2 Hellberg S, Sjostrom M, Wold S. The prediction of bradykinin potentiating potency of pentapeptides. An example of a peptide quantitative structure-activity relationship. Acta Chem Scand, 1986, 40: 135—140
3 Hellberg S, Sjostrom M, Skagerberg B, Wold S. Peptide quantitative structure-activity relationships, a multivariate approach. J Med Chem, 1987, 30: 1126—1135[DOI]
4 Collantes E R, Dunn W J. Amino acid side chain descriptors for quantitative structure activity relationship studies of peptide analogues. J Med Chem, 1995, 38: 2705—2713[DOI]
5 Zaliani A, Gancia E. MS-WHIM scores for amino acids: A new 3D-description for peptide QSAR and QSPR studies. J Chem Inf Comput Sci, 1999, 39: 525—533[DOI]
6 Kidera A, Konishi Y, Oka M, Ooi T, Scheraga H A. A statistical analysis of the physical properties of the 20 naturally occuring amino acids. J Protein Chem, 1985, 4: 23—55[DOI]
7 Raychaudhury C, Banerjee A, Bag P, Roy S. Topological shape and size of peptides: identification of potential allele specific helper T cell antigenic sites. J Chem Inf Comput Sci, 1999, 39: 248—254
8 Geysen H M, Rodda S J, Mason T. A priori delineation of a peptide which mimics a discontinuous antigenic determinant. J Mol Immunol, 1986, 23: 709—716[DOI]
9 Cramer R D, Patterson D E, Bunce J D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc, 1988, 110: 5959—5967[DOI]
10 Matta C F, Bader R W. Atoms-in-molecules study of the genetically encoded amino acids. Ⅲ. Bond and atomic properties and their correlations with experiment including mutation-induced changes in protein stability and genetic coding. Proteins Struct Funct Genet, 2003, 52: 360—399[DOI]
11 Rose G D, Geselowitz A R, Lesser G J, Lee R H, Zehfus M H. Hydrophobicity of amino acid residues in globular proteins. Science, 1985, 229: 834—838[DOI]
12 Eisenberg D, McLachlan A D. Solvation energy in protein folding and binding. Nature, 1986, 319: 199—203[DOI]
13 Sandberg M, Eriksson L, Jonsson J, Sjöström W S. New chemical descriptors for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem, 1998, 41: 2481—2491[DOI]
14 Hohenberg P, Kohn W. Inhomogeneous electron gas. Phys Rev, 1964, 136: B864—B871[DOI]
15 Kohn W, Sham L J. Self-consistent equations including exchange and correlation effects. Phys Rev, 1965, 140: A1133—A1138[DOI]
16 Mulliken R S. Electronic population analysis on LCAO-MO molecular wave functions I. J Chem Phys, 1955, 23: 1833—1840[DOI]
17 Becke A D. Density-functional thermochemistry. Ⅲ. The role of exact exchange. J Chem Phys, 1993, 98: 5648—5652[DOI]

18 Allinger N L, Yun Y H, Li J H. Molecular mechanics: the MM3 force field for hydro-carbons. J Am Chem Soc, 1989, 111: 8551—8565 [DOI]
19 Kellogg G E, Semus S F, Abraham D J. HINT—a new method of empirical hydrophobic field calculation for CoMFA. J Comput -Aided Mol Des, 1991, 5: 545—552[DOI]
20 Wireko F C, Kellogg G E, Abraham D J. Allosteric modifiers of hemoglobin. 2. Crystallographically determined binding sites and hydrophobic binding/interaction analysis of novel hemoglobin oxygen effectors. J Med Chem, 1991, 34: 758—767[DOI]
21 Kellogg G E, Joshi G S, Abraham D J. New tools for modeling and understanding hydrophobicity and hydrophobic interactions. Med Chem Res, 1992, 1: 444—453
22 Kellogg G E, Abraham D J. Key, lock and Locksmith. Complementary hydrophobicity map predictions of drug structure from a known receptor/receptor structure from known drugs. J Mol Graph, 1992, 10: 212—217[DOI]
23 Nayak V R, Kellogg G E. Cyclodextrin-barbiturate inclusion complexes: a CoMFA/HINT 3-D QSAR study. Med Chem Res, 1994, 3: 491—502
24 Welch W, Ahmad S, Airey J A, Gerzon K, Humerickhouse R A, Besch Jr H R, Ruest L, Deslongchamps P, Sutko J L. Structural determinants of high-affininty binding of ryanoids to the vertebrate skeletal muscle ryanodine receptor: a comparative molecular field analysis. Biochem, 1994, 33: 6074—6085[DOI]
25 Oprea T I, Waller C L, Marshall G R. 3D-QSAR of human immunodeficiency virus I protease inhibitors. Ⅲ. Interpretation of CoMFA results. Drug Design Discovery, 1994, 12: 29—51
26 Waller C L. A three dimensional technique for the calculation of octanol/water partition coefficients. Quan Struct Act Relat, 1994, 13: 172—176
27 Wright H T, Scarsdale J N. Structural basis for serpin inhibitor activity. Proteins, 1995, 22: 210—225[DOI]
28 Welch W, Sutko J L, Mitchell K E, Airey J, Ruest L. The pyrrole locus is the major orienting factor in ryanodine binding. Biochemistry, 1996, 35: 7165—7173[DOI]
29 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—276[DOI]
30 Walters D E, Hinds R M. Genetically evolved receptor models: A computational approach to construction of receptor models. J Med Chem, 1994, 37(16): 2527—2536[DOI]
31 Polański J. The receptor-like neural network for modeling corticosteroid and testosterone binding globulins. J Chem Inf Comput Sci, 1997, 37: 553—561[DOI]
32 Hahn M. Receptor surface models. 1. Definition and construction. J Med Chem, 1995, 38(12): 2080—2090[DOI]
33 Weiner S J, Kollman P A, Case D A, Singh U C, Ghio C, Alagona G, Profeta J S, Weiner P. A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc, 1984, 106: 765—784[DOI]
34 Raevsky O A, Sapegin A, Zefirov N. The QSAR discriminantregression model. Quant Struct Act Relat, 1994, 13: 412—418
35 Raevsky O A. Quantification of noncovalent interactions on the basis of the thermodynamic hydrogen bond parameters. J Phys Org Chem, 1997, 10: 405—413[DOI]
36 Holland J H. Adaptation in Natural and Artificial System. Ann Arbor: University of Michigan Press, 1975
37 Wold S, Ruhe A, Wold H, Dunn W J. The collinearity problem in linear regression. The Partial Least Squares (PLS) approach to generalized inverses. Siam J Sci Statist Comput, 1984, 5: 735—743[DOI]
38 Hassell C H. The Design and synthesis of new triazolo-, pyrazolo-, and pyridazo-pyridazine derivatives as inhibitors of angiotensin converting enzyme. J Chem Soc Perkin Trans I, 1984, 23: 155—162[DOI]
39 DePriest S A, Mayer D, Naylor C B, Marshall G R. 3D-QSAR of angiotensin converting enzyme and thermolysin inhibitors: A comparison of CoMFA models based on deduced and experimentally determined active site geometries. J Am Chem Soc, 1993, 115: 5372—5384 [DOI]
40 Hellberg S, Eriksson L, Jonsson J, Lindgren F, Sjöström M, skagerberg B, Wold S, Andrews P. Minimum Analogue Peptide Sets (MAPS) for quantitative structure-activity relationships. Int J Pept Protein Res, 1991, 37: 414—424
41 Cocchi M, Johansson E. Amino acids characterization by GRID and multivariate data analysis. Quant Struct Act Relat, 1993, 12: 1—8 [DOI]
42 Transue T R, Genst E D, Ghahroudi M A, Wyns L, Muyldermans S. Camel single-domain antibody inhibits enzyme by mimicking carbohydrate substrate. Proteins, 1998, 32: 515—522[DOI]
43 Desmyter A, Transue T R, Ghahroudi M A, Dao-Thi M H, Poortmans F, Hamers R, Muyldermans S, Wyns L. Crystal structure of a camel single-domain Vh antibody fragment in complex with lysozyme. Nat Struct Biol, 1996, 3: 803—811[DOI]
44 Berman H M, Westbrook J, Feng Z. The protein data bank. Nucl Acid Res, 2000, 28: 235—242[DOI]
45 Genst D, Areskoug E D, Decanniere K, Muyldermans S, Andersson K. Kinetic and affinity predictions of a protein-protein interaction using multivariate experimental design. J Biol Chem, 2002, 277: 29897—29907[DOI]
46 Freyhult E K, Andersson K, Gustafssonz M G. Structural modeling extends QSAR analysis of antibody-lysozyme interactions to 3D-QSAR. Biophys J, 2003, 84: 2264—2272
47 Zhou P, Tian F F, Wu S R, Li Z L. Genetic algorithm-based virtual screening of combinative mode for peptide/protein, Acta Chimi Sini, 2006, 64(7): 691—697
48 Mei H, Zhou Y, Liang G, Li Z. Support vector machine applied in QSAR modelling. Chin Sci Bull, 2005, 50(20): 2291—2296[DOI]
49 Hancock R E W. Cationic peptides: Effectors in innate immunity and novel antimicrobials. Lancet Infect Dis, 2001, 1: 156—164 [DOI]
50 Boman H G. Antibacterial peptides: Key components needed in immunity. Cell, 1991, 65: 205—207[DOI]
51 Lehrer R I, Lichtenstein A K, Ganz T. Defensins: Antimicrobial and cytotoxic peptides of mammalian cells. Annu Rev Immunol, 1993, 11: 105—128[DOI]

附表 1 ACE 抑制剂的氨基酸序列及其抑制活性观察值与计算值

编号	肽	观测值 a)	计算值 b)	绝对误差	相对误差/%
1	VW	5.80	5.18	–0.62	–10.690
2	IW	5.70	5.03	–0.67	–11.754
3	IY	5.43	4.81	–0.62	–11.419
4	AW	5.00	4.61	–0.39	–7.801
5	RW	4.80	4.96	0.16	3.333
6	VY	4.66	4.36	–0.3	–6.439
7	GW	4.52	3.96	–0.56	–12.39
8	VF	4.28	4.23	–0.05	–1.168
9	AY	4.06	3.89	–0.17	–4.187
10	IP	3.89	3.62	–0.27	–6.941
11	RP	3.74	3.65	–0.09	–2.406
12	AF	3.72	3.75	0.03	0.806
13	GY	3.68	3.64	–0.04	–1.087
14	AP	3.64	2.89	–0.75	–20.604
15	RF	3.64	4.51	0.87	23.901
16	VP	3.38	3.37	–0.01	–0.297
17	GP	3.35	2.65	–0.7	–20.896
18	GF	3.20	3.50	0.3	9.375
19	IF	3.03	4.47	1.44	47.526
20	VG	2.96	2.80	–0.16	–5.405
21	IG	2.92	3.04	0.12	4.110
22	GI	2.92	2.94	0.02	0.686
23	GM	2.85	2.62	–0.23	–8.071
24	GA	2.70	2.29	–0.41	–15.185
25	YG	2.70	2.38	–0.32	–11.853
26	GL	2.60	3.04	0.44	16.924
27	AG	2.60	2.32	–0.28	–10.771
28	GH	2.51	2.93	0.42	16.733
29	GR	2.49	3.10	0.61	24.499
30	KG	2.49	3.04	0.55	22.088
31	FG	2.43	2.60	0.17	6.996
32	GS	2.42	2.07	–0.35	–14.463
33	GV	2.34	2.66	0.32	13.675
34	MG	2.32	2.68	0.36	15.518
35	GK	2.27	2.82	0.55	24.231
36	GE	2.27	2.51	0.24	10.573
37	GT	2.24	2.55	0.31	13.840
38	WG	2.23	2.70	0.47	21.076
39	HG	2.20	2.46	0.26	11.819
40	GQ	2.15	1.28	–0.87	–40.466
41	GG	2.14	2.07	–0.07	–3.272
42	QG	2.13	2.38	0.25	11.738
43	SG	2.07	1.95	–0.12	–5.798
44	LG	2.06	3.04	0.98	47.573
45	GD	2.04	2.14	0.1	4.903
46	TG	2.00	2.24	0.24	12.001
47	EG	2.00	2.02	0.02	1.001
48	DG	1.85	1.76	–0.09	–4.866
49	PG	1.77	2.45	0.68	38.418
50	LA	3.51	3.26	–0.25	–7.124
51	KA	3.42	3.25	–0.17	–4.972
52	RA	3.34	3.29	–0.05	–1.498
53	YA	3.34	2.60	–0.74	–22.156
54	AA	3.21	2.54	–0.67	–20.873
55	FR	3.04	3.62	0.58	19.079
56	HL	2.49	3.43	0.94	37.752
57	DA	2.42	1.98	–0.44	–18.182
58	EA	2.00	2.23	0.23	11.501

a) 实验观测活性值, pIC_{50}; b) GVSPPC 模型计算活性值

附表 2　Collantes等[4]提出的侧链电荷指数ECI以及Sandberg等[46]扩展的氨基酸描述子ZZ-scale

序号	氨基酸	ECI	ZZ_1	ZZ_2	ZZ_3	ZZ_4	ZZ_5
1	Ala, A	0.05	0.24	–2.32	0.60	–0.14	1.30
2	Arg, R	1.69	3.52	2.50	–3.50	1.99	–0.17
3	Asn, N	1.31	3.05	1.62	1.04	–1.15	1.61
4	Asp, D	1.25	3.98	0.93	1.93	–2.46	0.75
5	Cys, C	0.15	0.84	–1.67	3.71	0.18	–2.65
6	Gln, Q	1.36	1.75	0.50	–1.44	–1.34	0.66
7	Glu, E	1.31	3.11	0.26	–0.11	–3.04	–0.25
8	Gly, G	0.02	2.05	–4.06	0.36	–0.82	–0.38
9	His, H	0.56	2.47	1.95	0.26	3.90	0.09
10	Ile, I	0.09	–3.89	–1.73	–1.71	–0.84	0.26
11	Leu, L	0.10	–4.28	–1.30	–1.49	–0.72	0.84
12	Lys, K	0.53	2.29	0.89	–2.49	1.49	0.31
13	Met, M	0.34	–2.85	–0.22	0.47	1.94	–0.98
14	Phe, F	0.14	–4.22	1.94	1.06	0.54	–0.62
15	Pro, P	0.16	–1.66	0.27	1.84	0.70	2.00
16	Ser, S	0.56	2.39	–1.07	1.15	–1.39	0.67
17	Thr, T	0.65	0.75	–2.18	–1.12	–1.46	–0.40
18	Trp, W	1.08	–4.36	3.94	0.59	3.44	–1.59
19	Tyr, Y	0.72	–2.54	2.44	0.43	0.04	–1.47
20	Val, V	0.07	–2.59	–2.64	–1.54	–0.85	–0.02