

一种基于三维原子场相互作用矢量的 新型氨基酸结构信息描述子

周 鹏 周 原 吴世容 李 波 田菲菲 李志良 *

(重庆大学化学化工学院, 重庆 400044; 生物医学工程教育部与重庆市重点实验室, 重庆 400044; 重庆大学生物工程学院, 重庆 400044. * 联系人, E-mail: zlli2662@163.com)

摘要 从分子的三维空间结构出发, 按照不同类型原子之间静电和立体作用得到一种新的分子结构表达方法——三维原子场相互作用矢量(3D-VAIF). 利用该矢量对 20 种天然氨基酸空间结构性质的计算, 通过主成分分析获得单个氨基酸的特征描述子——氨基酸结构信息得分(SSIA). 分别使用 58 个血管紧张素转化酶抑制剂、48 个苦味二肽以及 31 个缓激肽对 SSIA 的性能进行了测试, 所得模型的复相关系数 R_{cum}^2 和交互检验复相关系数 Q_{Loo}^2 分别为 0.789, 0.773; 0.856, 0.837 和 0.838, 0.815. 结果表明将 SSIA 用于肽类似物定量序列活性建模(定量序效建模, QSAM)效果优于传统氨基酸描述子.

关键词 三维原子场相互作用矢量 氨基酸结构信息得分 定量构效关系 定量序效建模 量子化学计算

定量构效关系(QSAR)一直是众多学科领域的研究热点, 将其应用于蛋白质、核酸等生物大分子序列与活性相关性研究亦可称为定量序列活性建模(QSAM)^[1]. 然而, 由于生物分子自身的复杂性使得对于序列活性的定量预测模型建立较为困难, 因此如何发展一种简明有效的结构表征手段成为 QSAM 研究的瓶颈. 这方面开创性的工作是由 Sneath 等^[2]于 1966 年做出的, 他们使用 20 个天然氨基酸的半定量实验参数对肽类似物序列进行了表征. 20 世纪 70 年代以后, 伴随着计算机技术和生物信息及化学计量学的发展, 对于生物大分子的构效研究逐渐活跃起来. Kidera 等^[3]首先采用因子分析将 20 种天然氨基酸的 188 种性质变换为 10 个正交因子, 并将其应用于蛋白质的高级结构的预测. Hellberg 等^[4-8]对收集到的氨基酸的 20 种物理化学性质通过正交变换得到了一组单个氨基酸的性质描述子 Z 标度, 之后又进一步加入了 9 个高效液相色谱保留指数扩充了该标度. 此后, Sandberg 等^[9]将 Z 标度推广到了包括天然氨基酸在内的 87 种氨基酸, 取得了较为成功的应用. 另外 Collantes 等^[10]使用分子性质模拟程序 HYDRAT 和 CNDO/2 法计算得到了反映氨基酸侧链立体特征和局部偶极两个基本性质参数: 全向表面积(ISA)及电荷指数(ECI)用来表达氨基酸分子结构特征, 并成功应用于一系列肽的序列活性建模当中. 近年来, Liu 等^[11]直接从氨基酸分子的二维拓扑构型出发发展了用于表征化合物结构的分子全息距离矢量(MHDV),

并用于寡肽一级序列的表达中. 而 Zaliani 的研究小组^[12]则以单个氨基酸的原子空间坐标进行加权变换产生对旋转和平移的不变量——MS-WHIM 得分来描述肽的序列特征. Raychaudhury 等^[13]使用氨基酸侧链形状指数(SCI)和距离表达指数(DEI)研究了 T 细胞对结合有 28 个免疫源性多肽, 即 Th 细胞表位的主要组织相容性复合物(MHC)的响应性, 取得了相当高的正确识别率. 然而由于传统的氨基酸描述子大都以实验获得的参数为基础, 存在着数据来源随意性和不确定性等缺点, 并且在很大程度上受到了实验条件的限制, 因此难以形成一个统一的标准而限制了该类描述子的发展. 考虑到量子力学(QM)等纯粹理论计算的方法在 QSAR 研究中已有应用, 并取得了较为满意的结果, 例如 Wan 等^[14,15]就在密度泛函(DFT)水平成功地研究了卟啉氧化酶(PPO)抑制剂的生物活性与量化参数间的关系. 鉴于此, 如果使用量子化学方法对单个氨基酸的结构性质进行精确计算, 结合分子表征技术得到所谓的量化描述子, 并以该类描述子作为氨基酸三维结构特征的初始化矢量, 以此来代替传统的实验数据, 便可从一个新的角度来解决上述矛盾. 因此本文提出了一种新的分子三维结构表达方法——三维原子场相互作用矢量(3D-VAIF), 直接利用从不同量子化学计算水平获得的 20 种天然氨基酸结构和性质参数, 对氨基酸分子的空间结构加以表征, 结合主成分分析(PCA)技术对表征结果进行信息提取和变量压缩, 最终得到了一种新型氨基

酸描述子——氨基酸结构信息得分(SSIA). 使用 58 个血管紧张素转化酶抑制剂、48 个苦味肽以及 31 个缓激肽对该描述子进行测试, 取得了优于文献的结果, 同时比较了不同量子化学计算水平对模型性能的影响. 该方法也是将量子化学计算应用于QSAM研究的一次尝试.

1 原理与方法

1.1 三维原子场相互作用矢量(3D-VAIF)

分子结构表征(MSC)是当代药物分子设计与药效评价中不可缺少的重要技术, 当前分子结构描述方法大致可以分为基于二维结构的 2D 结构描述子和基于三维空间构型的 3D 结构描述子两大类. 其中 2D 描述子具有快速、简便且易于实现等特点而成为早期 MSC 研究的主流, 但是随着现代构效关系的研究对模型的物理化学意义及预测能力等方面的要求的提高, 3D 描述子迅速发展并有逐渐取代 2D 描述方法成为现代 QSAR 研究的发展趋势. 鉴于此, 在本实验室 Liu 等^[11,16~18]所发展的 2D 分子结构表达方法的基础上, 本文提出了基于分子三维空间构型的立体结构描述方法——三维原子场相互作用矢量(3D-VAIF), 该矢量从分子立体结构的不变量——原子相对距离和原子自身性质出发通过简单然而物理意义明确的变换得到, 避免了传统 3D 描述子(如 CoMFA^[19], GRID^[20]等)在计算前需要对分子空间构型进行大量叠合操作.

$$V_{mn}(E) = \sum_{i \in m, j \in n} \frac{e^2}{4\pi\epsilon_0} \cdot \frac{Z_i \cdot Z_j}{r_{ij}} \quad (1 \leq m \leq n \leq 5), \quad (1)$$

$$V_{mn}(S) = \sum_{i \in m, j \in n} \epsilon_{ij} \left[\left(\frac{R_{ij}^*}{r_{ij}} \right)^{12} - 2 \cdot \left(\frac{R_{ij}^*}{r_{ij}} \right)^6 \right] \quad (1 \leq m \leq n \leq 5), \quad (2)$$

Cramer 等^[19]认为具有药物和生物活性的分子其配基与受体作用大多是一种可逆的非键合作用, 表现为范德瓦耳斯力、静电作用等力场形式. 因此一种用于药物和生物分子活性研究的结构表达方法必须突出这些作用方式. 这里首先将有机物分子中常见原子按元素周期表的族划分为 5 类(表 1), 同一类原子具有相似的化学性质. 分子中不同类原子之间的相互作用势能可以用于表征该分子空间势场的分布信息. 对于 5 类原子间的相互作用可以得到 15 种不同的作用情况(表 1). 这样当采用 Coulomb 定律((1)式)和 Lennard-Jones 方程((2)式)分别计算一个分子内部

所有原子间的静电和立体相互作用势能时就可以得到 30 个不同类型的作用项(其中前 15 项代表分子的静电作用能分布情况, 后 15 项则表示立体作用能分布情况)作为该分子的 3D-VAIF 描述子. 本文基于量子化学计算方法分别在 AM1, PM3, Hartree-Fock 和 DFT(使用 B3LYP 杂化泛函)4 个水平上使用 Berny 能量梯度法和广义梯度近似对 20 种天然氨基酸的立体几何构型进行了优化, 并采用 Mulliken 布居分析法以单点方式计算出分子中每个原子所带净电荷数量. 其中采用 Hartree-Fock 和 DFT 进行几何结构优化时选用的分子轨道基函数为双重分裂基组并在每个重原子上加入了 d 轨道极化函数(6-31G*), 而单点计算所采用的轨道基组是在几何优化基组的基础上再在氢原子上加入 p 轨道极化函数得到(6-31G**). 对于 AM1, PM3 的计算使用半经验量子化学软件 MOPAC 6.0 完成, Hartree-Fock 和 DFT 的计算使用 Gaussian 98W 实现, 氨基酸分子的初始立体结构输入文件由 ChemOffice 8.0 自动生成, 将量化计算结果转化为 3D-VAIF 描述子是由 C 语言编写的应用程序 OBTVAIF 完成.

表 1 按元素周期表划分的 5 类常见原子及 15 种相互作用

原子类型	每类原子	1	2	3	4	5 ^{a)}
1	H	1-1	1-2	1-3	1-4	1-5
2	C		2-2	2-3	2-4	2-5
3	N, P			3-3	3-4	3-5
4	O, S				4-4	4-5
5 ^{a)}	F, Cl, Br, I					5-5

a) 计算氨基酸分子 3D-VAIF 描述子时无第 5 类卤素原子作用项

1.2 氨基酸结构信息得分(SSIA)

由于天然氨基酸中不含第 5 类卤素原子, 因而 30 种原子相互作用情况中有 10 项计算结果为零. 对于余下的 20 个非零项将其按静电和立体作用分成两部分(每一部各有 10 项), 这样 20 个天然氨基酸就可以得到两个由非零 3D-VAIF 描述子构成的 20×10 变量矩阵. 对这两个矩阵首先采用自定标法进行标准化处理, 即按列将所有数据变换为服从均值为 0, 方差为 1 的标准正态分布, 然后使用主成分分析得到相应的主成分得分以及每一个主成分的特征值和特征向量. 这里采纳 Kaiser 建议, 即认为特征值大于 1 的主成分为显著主成分而予以保留(SYSTAT 11.0 默认标准), 并作为氨基酸结构信息得分 SSIA. 对不同量子化学计算水平得到的氨基酸 SSIA 描述子的显著主

成分得分和相应特征值列于附表1~4中¹⁾。

对于一组肽类似物,每个位点上的氨基酸结构特征可由该氨基酸的SSIA描述子所表征,例如使用SSIA-AM1来表征一个二肽分子将得到 $2 \times (3+4)=14$ 个描述子,而SSIA-HF则为 $2 \times (3+5)=16$ 个。并将表征所得的肽分子整体变量与其活性建立偏最小二乘(PLS)^[21]模型。由于氨基酸序列的不同残基以及用于描述同一残基的不同变量对活性贡献是不相同的,因此通常在建立QSAM模型之前要进行变量筛选,以剔除对模型有干扰效果的变量。常见的变量筛选方法如正交化变换^[22]、逐步回归(SMR)、遗传算法(GA)^[23]、模拟退火(SA)^[24]等,其中当变量数目较少时逐步回归以其计算简便、迅速而成为一种有效的变量选择方法。本文采用SMR按不同变量的F检验显著性大小依次引入,并结合PLS进行建模采用留一法交互检验(LOOCV)的复相关系数 Q_{LOO}^2 作为最佳变量数目的确定标准。需要说明的是,对于SMR每一步生成的原始变量矩阵X,其PLS潜隐变量个数的确定是按照软件Simca-P 10.0默认的标准得到的,即通过非线性迭代偏最小二乘(NIPALS)算法进行迭代并依据解释原始变量方差大小依次计算各个主成分,同时检测每次得到的主成分对模型交互检验复相关系数 Q_{LOO}^2 的贡献,当某一个被引入的主成分对模型的 Q_{LOO}^2 相关性小于0.097时即认为该主成分已不显著,并停止计算更多的主成分。主成分分析、逐步回归和偏最小二乘分别使用SYSTAT 11.0, SPSS for Windows 13.0以及Simca-P 10.0软件实现,所有计算工作是在联想P4 2.0G微机上完成。

2 结果及分析

2.1 血管紧张素转化酶抑制剂的QSAM研究

人体内肝脏产生的血管紧张素原经肾素催化断裂为非活性的十肽血管紧张素(Ang⁻),再由血管紧张素转化酶(angiotensin converting enzyme, ACE)催化裂解为具有极强血管收缩功能的八肽血管紧张素(Ang⁻),因此ACE就成为降血压药物研究的作用靶点。用于降低ACE与底物结合活性的58个二肽血管紧张素转化酶抑制剂是一个经典的抗高血压肽类物质样本集,其活性用 pIC_{50} 来表示。Hell-berg

等^[7]使用Z scale, Cocchi等^[25]采用GRID分析法, Collantes等^[10]使用ECI-ISA指数, Zaliani等^[12]使用MS-WHIM scores, Li等^[26]使用MEEV描述子对其进行了构效研究。本文分别采用SSIA-AM1, SSIA-PM3, SSIA-HF和SSIA-DFT对该58个ACE抑制剂的序列结构进行表征,并使用SMR结合LOO-CV技术取得具有最佳变量数目的PLS模型。表2列出了文献以及我们使用SSIA描述子的建模结果,其中除了两个MEEV模型是使用多元线性回归(MLR)得到的之外,其他模型都是采用PLS技术建立。可看出SSIA-PM3和SSIA-HF的预测能力 Q_{LOO}^2 显著优于已有的文献报道,其余两个SSIA模型的交互检验复相关系数 Q_{LOO}^2 仅略低于MEEV(M2)的结果。另外,SSIA-HF模型对58个样本的拟和度($R_{\text{cum}}^2 = 0.797$)为最高,同时它的 Q_{LOO}^2 也达到了较高的值。而综合考虑,使用SSIA-PM3建立了最优的PLS模型,该模型中使用了1个显著的主成分,其解释了Y变量78.9%的方差,交互检验所解释的Y变量方差为77.3%,相差仅为1.6%,可见模型的稳定性极佳。为了便于比较,这里将SSIA-HF以及文献使用GRID和MS-WHIM的建模结果绘制于图1中。

表2 ACE抑制剂的QSAM模型比较

编号	方法	R_{cum}^2 ^{a)}	Q_{LOO}^2 ^{b)}	RMSEE ^{c)}
1	Z scale ^[7]	0.770	nd ^{d)}	nd ^{d)}
2	t scores ^[25]	0.744	nd ^{d)}	0.50
3	ISA-ECI ^[10]	0.700	nd ^{d)}	nd ^{d)}
4	MS-WHIM(rotameric) ^[12]	0.657	0.541	nd ^{d)}
5	MS-WHIM(extended) ^[12]	0.708	0.637	0.54
6	MEEV(M1) ^{[26] e)}	0.792	0.677	0.50
7	MEEV(M2) ^{[26] e)}	0.741	0.711	0.50
8	SSIA-AM1	0.769	0.699	0.49
9	SSIA-PM3	0.789	0.773	0.47
10	SSIA-HF	0.797	0.760	0.46
11	SSIA-DFT	0.734	0.678	0.52

a) R_{cum}^2 : 复相关系数; b) Q_{LOO}^2 : 交互检验复相关系数; c) RMSEE: 训练集均方根误差; d) nd: 文献未报道; e) MEEV模型由MLR法建立

2.2 苦味二肽的QSAM研究

48个苦味活性二肽(BTT)取自Collantes报告^[10],其活性用浓度的负对数 pT 表示,这组样本集同样常被用于测试氨基酸描述子的有效性。这里我们利用

1) 请向作者索取

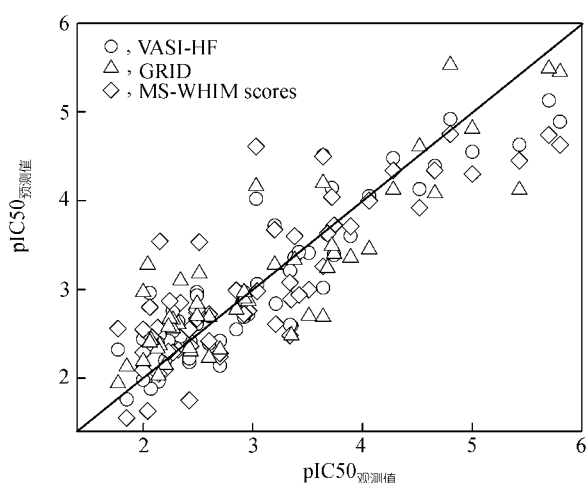


图 1 ACE 抑制剂的活性观测值与 SSIA-HF, GRID 及 MS-WHIM 预测值相关

SSIA 矢量对其进行表征, 结合 SMR 技术最终得到了具有最佳变量数目的 PLS 模型. 表 3 给出了 SSIA 模型的统计结果以及已有文献的报道值, 可以看到由 SSIA 得到的 4 个 PLS 模型都不同程度优于其他建模方法的结果, 拥有很高的建模拟合度和对外部样本预测能力. 其中, 具有最高预测稳定性的 PLS 模型是 SSIA-AM1, 其留一法交互检验 Q^2_{Loo} 对 Y 变量解释了 83.3% 的方差, 而其他 3 个 SSIA 模型的 Q^2_{Loo} 也不同程度地优于已有的文献报道. 另外, 比较不同的建模方法可以看到, 只有 SSIA 描述子以及 Z 标度和 ISA-ECI 建立的模型对于苦味肽活性值的拟和结果

R^2_{cum} 达到了 0.8 以上, 遗憾的是 Z 标度和 ISA-ECI 没有给出交互检验结果, 因而难以说明它们对外部样本的预测能力. 这里将 SSIA-PM3, Z 标度和 MS-WHIM 对该 48 个苦味二肽活性计算值与实验值相关性绘于图 2.

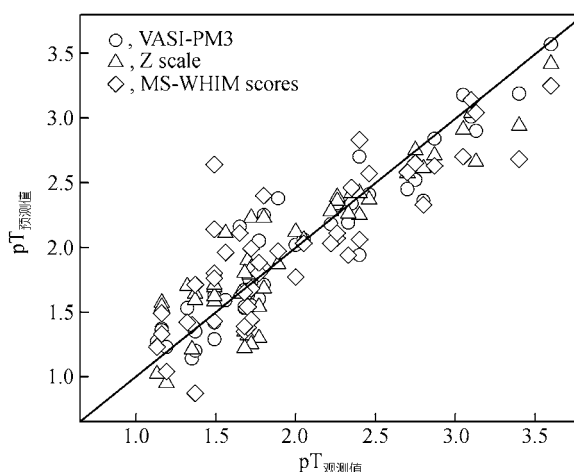


图 2 BTT 的活性观测值与 SSIA-PM3, Z 标度及 MS-WHIM 预测值相关

2.3 缓激五肽的 QSAM 研究

通过激肽原转化得到的胰激肽在酶催化裂解作用下生成具有血管扩张功能的舒缓激肽, 该类肽激素可以增加血液渗透性、引起局部疼痛以及刺激前列腺素合成. 31 个血管舒缓激肽 (BPP) 始见报道于 Ufkes 等^[27, 28] 的论文中, 后被 Niwa^[29], Hellberg^[4, 5], Collantes^[10] 和 Sung^[30] 等作为一组标准的 QSAM 样本集多次使用, 最近又被 Kim 等^[31] 用于测试他们所发展的溶解自由能密度模型 (SFED) 对单个氨基酸残基疏水指标的计算值. BPP 的活性分两步于 1978 年和 1982 年被测出, 其值 $\log \text{RAI}$ 被表示为相对于第一条序列 VESSK 活性大小的对数值. 在 Sung 的研究中, 他们分别采用了 ISA-ECI 指数、Z 标度以及软件 MOLCON 生成的 460 个拓扑指数, 在此基础上利用程序 Focus-2D 基于 GA-PLS 技术对该组样本中的 28 个序列建立了 QSAM 模型, 并以另外 2 个具有最高活性的序列 VEWAK, VKWAP 作为他们设计虚拟组合肽库的探针化合物. 表 4 列出 Sung 的 GA-PLS 建模结果以及我们分别使用 4 种 SSIA 描述子的 SMR-PLS 建模情况. 由于利用 SSIA 表征五肽化合物将得到 $5 \times 7 = 35$ 或 $5 \times 8 = 40$ 个整体描述子, 但是随 SMR 引入变量的显著性依次递减, 因此这里只对前 20 步进行

表 3 BTT 的 QSAM 模型比较

编号	方法	R^2_{cum} ^{a)}	Q^2_{Loo} ^{b)}	RMSEE ^{c)}
1	Z scale ^[7]	0.824	nd ^{d)}	0.26
2	t scores ^[25]	nd ^{d)}	0.780	nd ^{d)}
3	ISA-ECI ^[10]	0.847	nd ^{d)}	nd ^{d)}
4	MS-WHIM(rotameric) ^[12]	0.704	0.633	nd ^{d)}
5	MS-WHIM(extended) ^[12]	0.754	0.710	0.32
6	MEEV(M1) ^{[26] e)}	0.711	0.475	0.34
7	MEEV(M2) ^{[26] e)}	0.649	0.570	0.37
8	MEEV(M3) ^{[26] e, f)}	0.773	0.588	0.33
9	MEEV(M4) ^{[26] e, f)}	0.735	0.677	0.32
10	SSIA-AM1	0.850	0.837	0.25
11	SSIA-PM3	0.888	0.829	0.22
12	SSIA-HF	0.844	0.798	0.25
13	SSIA-DFT	0.856	0.741	0.24

a) R^2_{cum} : 复相关系数; b) Q^2_{Loo} : 交互检验复相关系数; c) RMSEE: 训练集的均方根误差; d) nd: 文献未报道; e) MEEV 模型由 MLR 法建立; f) 参与建模样本数为 47 个

表4 BPP的QSAM模型比较

编号	方法	R_{cum}^2 ^{a)}	Q_{Loo}^2 ^{b)}	RMSEE ^{c)}
1	Z scale ^[30]	0.831	0.633	0.30
2	ISA-ECI ^[30]	0.840	0.725	0.29
3	Topological Indices (M1) ^[30]	0.470	0.367	0.53
4	Topological Indices (M2) ^[30]	0.630	0.533	0.46
5	Topological Indices (M3) ^[30]	0.899	0.845	0.23
6	SFED(M1) ^{[31] d)}	0.635	0.450	nd ^{e)}
7	SFED(M2) ^{[31] d)}	0.750	0.617	nd ^{e)}
8	SSIA-AM1	0.874	0.786	0.27
9	SSIA-PM3	0.816	0.760	0.32
10	SSIA-HF	0.838	0.815	0.31
11	SSIA-DFT	0.817	0.778	0.32

a) R_{cum}^2 : 复相关系数; b) Q_{Loo}^2 : 交互检验复相关系数; c) RMSEE: 训练集均方根误差; d) SFED模型由MLR法建立; e) nd: 文献未报道

PLS交互检验. 表4中SSIA对于前28个作为训练集的样本具有优于ISA-ECI指数、Z标度、SFED指标以及前两类拓扑描述子的模型稳定性, 仅略低于第三类拓扑模型, 这可能是因为文献使用了遗传算法从460个庞大的描述子集群中优选出了23个变量参与建模, 从而使该模型的质量显著提高. 另外对SSIA-AM1计算结果的残差分布情况分析可知其中26个缓激肽的模型计算残差值都分布在正负2倍总体均方根误差以内, 对于12, 13号样本其残差绝对值也小于3倍总体均方根误差(0.74和0.58<0.81), 因而不能认为它们是异常点. 包括文献在内的所有方法都对两个探针化合物预测结果偏低, 这可能有以下原因: (1) 实验结果偏高; (2) 模型不具有对外插值样本的泛化能力; (3) 样本本身的特殊性所导致. 具体情况有待进一步探讨. 另外, 我们对该文献没有给出的第31号样本(GGGG)进行了预测, 4个SSIA模型的结果都较好地表明了该肽不具有生物活性特征.

3 结语

本文从一个新的角度提出了分子立体结构表达方法: 三维原子场相互作用矢量(3D-VAIF), 基于此得到了一种新的天然氨基酸描述子: 氨基酸结构信息得分(SSIA). 分别使用三组标准肽类似物样本集进行测试, 取得了满意结果. 由于SSIA完全是由理论计算获得, 具有操作简单、意义明确等特点. 另外, 本文得到的SSIA描述子只包了含静电和立体作用两部分, 可以推测如果加入其他有关生物活性的理化性质, 如疏水、氢键等必将进一步提高模型质量, 相关工作有待进一步深入. 通过比较发现各种量化计算

所得结果较为一致, 说明该法具有良好的稳定性. 需要注意的是, 从序列重复结构单元表征具有不同链段数目的生物分子时, 必将面临描述子个数的差异, SSIA同样难以摆脱此类困境, 此时可以考虑使用如交差协方差(ACC)^[32]等方法来对变量数目进行归一化处理.

致谢 谨以此文献给恩师黄里河和张秀英老师七十寿辰. 李真、李晟和、李声时、黄莺、杨善彬、梅虎、梁桂兆等提供协助. 本工作为霍英东基金(批准号: 98-7-6)、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)及重庆大学创新基金(批准号: 03-5-6+04-10-10)资助项目.

参 考 文 献

- Jonsson J, Norberg T, Carlsson L, et al. Quantitative sequence-activity models (QSAM)—tools for sequence design. *Nucleic Acids Res*, 1993, 20: 733~739
- Sneath P H. Relations between chemical structure and biological activity in peptides. *J Theor Biol*, 1966, 12: 157~195
- Kidera A, Konishi Y, Oka M, et al. A statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J Protein Chem*, 1985, 4: 23~55[DOI]
- 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
- Hellberg S, Sjostrom M, Skagerberg B, et al. Peptide quantitative structure-activity relationships, a multivariate approach. *J Med Chem*, 1987, 30: 1126~1135[DOI]
- Wold S, Eriksson L, Hellberg S, et al. Principal property values for six non-natural amino acids and their application to a structure-activity relationship for oxytocin peptide analogues. *Can J Chem*, 1987, 65: 1814~1820
- Hellberg S, Eriksson L, Jonsson J, et al. Minimum analogue peptide sets (MAPS) for quantitative structure-activity relationships. *Int J Pept Protein Res*, 1991, 37: 414~424
- Jonsson J, Eriksson L, Hellberg S, et al. Multivariate parameterization of 55 coded and non-coded amino acids. *Quant Struct-Act Relat*, 1989, 8: 204~209
- Sandberg M, Eriksson L, Jonsson J, et al. 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]
- 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
- Liu S S, Yin C S, Li Z L, et al. A novel MHDV descriptor for dipeptide QSAR studies. *J Chin Chem Soc*, 2001, 48: 253~260
- 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]

- 13 Raychaudhury C, Banerjee A, Bag P, et al. 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[DOI]
- 14 Wan J, Zhang L, Yang G F, et al. Quantitative structure-activity relationship for cyclic imide derivatives of protoporphyrinogen oxidase inhibitors: A study of quantum chemical descriptors from density functional theory. *J Chem Inf Comput Sci*, 2004, 44: 2099~2105[DOI]
- 15 Wan J, Zhang L, Yang G F. Quantitative structure-activity relationships for phenyl triazolinones of protoporphyrinogen oxidase inhibitors: A density functional theory study. *J Comput Chem*, 2004, 25: 1827~1832[DOI]
- 16 Liu S S, Cai C X, Li Z L. Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance edge (MDE) vector λ . *J Chem Inf Comput Sci*, 1998, 38: 387~394[DOI]
- 17 Liu S S, Cai S X, Li Z, et al. Molecular electronegative-distance vector (MEDV) relating to 15 properties of alkanes. *J Chem Inf Comput Sci*, 2000, 40: 1337~1348[DOI]
- 18 Liu S S, Yin C S, Li Z, et al. QSAR study of steroid benchmark and dipeptides based on MEDV-13. *J Chem Inf Comput Sci*, 2001, 41: 321~329[DOI]
- 19 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]
- 20 Goodford P J. A computational procedure for determining energetically favorable binding sites on biologically important molecules. *J Med Chem*, 1985, 28: 849~857[DOI]
- 21 Wold S, Ruhe A, Wold H, et al. 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]
- 22 Šoškić M. Link between orthogonal and standard multiple linear regression models. *J Chem Inf Comput Sci*, 1996, 36: 829~832[DOI]
- 23 Rogers D, Hopfinger A J. Application of genetic function approximation to quantitative-structure activity relationships and quantitative structure-property relationships. *J Chem Inf Comput Sci*, 1994, 34: 854~866[DOI]
- 24 Sutter J M, Dixon S L, Jurs P C. Automated descriptor selection for quantitative structure-activity relationships using generalized simulated annealing. *J Chem Inf Comput Sci*, 1995, 35: 77~84[DOI]
- 25 Cocchi M, Johansson E. Amino acids characterization by GRID and multivariate data analysis. *Quant Struct-Act Relat*, 1993, 12: 1~8
- 26 Li S Z, Fu B H, Wang Y Q, et al. On structural parameterization and molecular modeling of peptide analogues by molecular electronegativity-edge vector (VMEE): Estimation and prediction for biological activity of pentapeptides. *J Chin Chem Soc*, 2001, 48: 937~944
- 27 Ufkens J G, Visser R J, Heuvel G, et al. Structure-activity relationships of bradykinin potentiating peptides. *Eur J Pharmacol*, 1978, 50: 119~122
- 28 Ufkens J G, Visser R J, Heuvel G, et al. Further studies on the structure-activity relationships of bradykinin potentiating peptides. *Eur J Pharmacol*, 1982, 79: 155~158[DOI]
- 29 Niwa T S, Ogino A. Evaluation of the hydrophobic parameters of the amino acid side chains of peptides and their application in QSAR and conformational studies. *Theochem*, 1997, 392: 43~54
- 30 Sung J, Wei F, Tropsha A. Rational combinatorial library design. 2. Rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches. *J Chem Inf Comput Sci*, 1998, 38: 259~268[DOI]
- 31 Kim J, Nam K Y, Cho K H, et al. Theoretical study on hydrophobicity of amino acids by the solvation free energy density model. *Bull Korean Chem Soc*, 2003, 24(12): 1742~1750
- 32 Andersson P M, Sjöström M, Lundstedt T. Preprocessing peptide sequences for multivariate sequence-property analysis. *Chemom Intell Lab Syst*, 1998, 42: 41~50 [DOI]
- 33 Miyashita Y, Li Z, Sasaki S. Chemical pattern recognition and multivariate analysis for QSAR studies. *Trend Anal Chem (TrAC)*, 1993, 12(2): 50~60[DOI]

(2005-04-09 收稿, 2005-07-02 接受)