一种新的比较分子/虚拟受体相互作用分析法

周 鹏[①②] 仝建波[③] 田菲菲[①] 李志良[①②*]

(①重庆大学化学化工学院, 重庆 400044; ②生物力学与组织工程教育部重点实验室, 重庆 400044; ③山西大学化学化工学院, 太原 030006. * 联系人, E-mail: lisz2662@163.com)

摘要 将虚拟受体的概念与比较分子场法有机结合起来, 得到了一种新颖的 3D-QSAR 研究方法: 比较分子/虚拟受体相互作用分析法(CoMPIA). 该方法通过定义受体原子探针并利用遗传算法优化探针在药物分子周围的最佳分布情况来找寻模型最优解和虚拟受体模式, 从而有效提高了建模的质量和结果的可解释性. 使用 CoMPIA 对 31 个经典甾体化合物进行了系统的 QSAR 研究, 所得模型拟合复相关系数 R^2、交叉检验 Q^2 以及对测试集预测结果的均方根误差 RMSEP 分别为 0.940, 0.868 和 0.502, 该结果优于多数文献报道.

关键词 比较分子/虚拟受体相互作用分析法 定量构效关系 遗传算法 甾体

化学及其相关科学的发展至今已呈现出从描述性向推理性、从宏观状态到微观结构理论研究的发展趋势. 人们在长期实践中认识到: 分子是构成物质的基本单位, 其结构一经确定则性质也随之而定. 定量地研究和分析分子或原子中的基本结构特征与从实验中反映出的一些性质之间的关系, 即所谓定量构效关系(QSAR), 已成为现代化学基础研究与应用的重要工具. 20 世纪 60 年代, 由Hansch等[1]提出的用于表达药物活性与其理化性质关系的Hansch-Fujita分析法以及Free等[2]所创立的具有指示变量性质的Free-Wilson法成为该领域的先驱, 并被成功应用于多种药物设计. 然而由于这些方法都是建立在分子二维(2D)性质之上, 难以正确表达药物配基与受体之间的三维(3D)立体作用效果, 从而限制了模型质量的进一步提高. 和 2D方法相比, 3D方法在物理化学意义上更为明确并能直观反映药物发挥药效过程中配基与受体之间的结合方式和非键作用特征, 因此得到了迅速发展. 在当今众多 3D-QSAR方法中以Cramer等于 1988 年提出的比较分子场法(CoMFA)[3]最为著名, 该方法现已成为一种标准QSAR技术被广泛应用于各类药物分子的分析和设计. 随着CoMFA的成功, 一系列与之相似的 3D分子性质比较方法被人们提出[4~6], 并在一些具体应用领域取得了较好的效果. 但是多数方法并没有从本质上跳出CoMFA的思路, 即都是通过比较分子某些性质的差异来决定它们对生物活性贡献的大小. 很显然, 药物与受体结合过程受到诸多理化因素的影响, 这些因素一方面取决于配基自身的性质特征, 另一方面又决定于受体活性部位的立体结构. 传统 3D-QSAR 方法仅仅从药物分子性质入手却没有很好地考虑到上述两方面之间的联系, 因此在实际应用中具有一定局限. 鉴于此, 本文将虚拟受体的概念引入到 CoMFA 当中, 通过遗传算法(GA)优化分子周围空间网格上的最佳探针原子分布情况, 以此建立并比较不同药物分子与受体的作用方式, 得到了一种新的 3D-QSAR 研究方法: 比较分子/虚拟受体相互作用分析法(comparative molecule/pseudo receptor interaction analysis, CoMPIA). 与 CoMFA 相比该方法有以下几点不同之处: (1) 使用 20 种标准天然氨基酸中的 9 类原子作为格点探针, 并基于算法自动确定这些探针原子的分布情况, 这样避免了 CoMFA 中对探针类型选取的随意性和主观性等问题, 并增加了模型的可解释性; (2) 药物之间通过直接比较分子与受体相互作用方式来建立 QSAR 模型, 而不仅仅是简单比较分子间性质的差异; (3) 将 GA 用于探针类型的选择, 使得在不同位置出现不同的探针类型, 从而形成一种最佳组合, 这样提高了方法的灵活性和适应度; (4) 允许某些格点不出现探针(即出现“空”类型), 这样就将变量选择与偏最小二乘回归(PLS)建模有机地结合到了一起, 最大限度降低了模型的噪声含量.

1 原理及方法

(1) 探针. 由于药物作用对象往往是蛋白质、酶、活性肽等氨基酸组成物, 因此CoMPIA中定义的探针为 20 种标准天然氨基酸中出现的 8 类不同杂化状态的原子. 另外考虑到配基周围的受体原子分布是不均匀的, 即网格上某些位置可能处于结合腔的背面

或药物分子内部，而这些地方受体往往难以抵达，故又定义了一个“空”类型原子探针. 为了反映这些探针的性质，我们分别给其赋予了平均电荷指数、van der Waals指数和平均疏水指数的概念(表 1). 平均电荷指数(MCI): 探针原子电性特征取其出现于氨基酸中的平均电荷数. 具体计算如下: 首先使用分子模拟软件HyperChem 7.5 自带数据库生成 20 个天然氨基酸的初始分子立体结构，同时利用该软件进行分子力学构象优化；其结果使用量子化学软件Gaussian 03W进行Hartree-Fock水平的进一步优化；最终以单点形式在密度泛函水平计算出所有原子的净电荷数量. van der Waals指数(VWI): 通常文献给出的是孤立原子van der Waals半径，但实际原子由于其所处化学微环境和自身杂化状态不同该半径有所变化，因此本文使用经校正后的原子van der Waals半径作为探针原子的van der Waals指数，VWI= $C_h \times R_{VDW}^*$，其校正因子C_h参见文献[7]. 另外，各类原子的标准van der Waals半径取自Bondi 等的报告[8]. 平均疏水指数(MHI): 类似于MCI，MHI取每类探针的疏水性在天然氨基酸中出现的平均值(单位为kJ/Å^2). 这里以Pei等定义的原子溶解参数(ASP)[9]作为疏水性量度.

表 1 9 类探针的 MCI, VWI 和 MHI 取值

编号	探针	MCI	VWI	MHI
0	Φ(空)	0.000	0.000	0.000
1	H(s)	0.227	1.200	-6.384×10^{-3}
2	C(sp3)	−0.220	1.700	-8.527×10^{-1}
3	C(sp2)	0.296	1.615	3.035×10^{-2}
4	N(sp3)	−0.714	1.550	4.006×10^{-1}
5	N(sp2)	−0.592	1.473	-8.931×10^{-2}
6	O(sp3)	−0.578	1.520	1.347×10^{-3}
7	O(sp2)	−0.470	1.444	-5.602×10^{-2}
8	S(sp3)	0.011	1.800	7.618×10^{-2}

(2) 叠合. CoMPIA 基于公共骨架进行分子叠合. 其主要操作步骤是: 首先选定一个药效活性最高的分子作为模板，并确定分子公共骨架上的几个关键原子，采用最小二乘法调整样本集中每个药物分子的空间位置使得它们与模板分子在这几个关键原子上的距离均方根误差(RMS)达到最小.

(3) 网格. 网格空间区域大小至少要包含待研究化合物中最大的分子，并且给予一定的空余. 本文网格覆盖区域对叠合后的分子体系给出 0.4 nm 的多余空间，同时将 X, Y, Z 3 个维度都平均分为 4 等份，这样可得到 5×5×5=125 个格点，并作为安置探针原子的位置.

(4) 势能. 静电作用是一类重要的非键效应，经典的点电荷作用方式服从Coulomb定理. 对于处在某一格点上的探针p与药物分子的总静电作用能计算参见(1)式，其中n是药物分子包含的原子数；r_{pi}是探针到分子中某原子i的Euclid距离；e为单位电荷电量；ε_0为真空中的介电常数；Z为原子净电荷数. 立体作用是空间原子间存在的非偶极-偶极或偶极诱导作用，这里采用Lennard-Jones方程来描述这种作用方式((2)式)，式中$\varepsilon_{pi}=(\varepsilon_{pp}\cdot\varepsilon_{ii})^{1/2}$ 为探针-原子对势能阱深[10]；$R_{pi}^*=(\mathrm{VWI}_p+C_h\cdot R_i^*)/2$，为经过校正后的探针-受体原子间van der Waals碰撞半径. 立体作用势能取 126 kJ/mol作为能量截断值. 疏水作用是影响药物分子与生物体结合的重要因素，由于其往往表现于体系熵的改变，因此很难用一个统一的公式来描述. 为了深入到药物分子内部原子与探针的相互作用，CoMPIA采用Hint方法[11]定义的原子之间疏水相互作用表达式((3)式)，其中S为原子溶剂可及面积；a为原子疏水常数；T是作用形式的二值判别函数，以表明不同类型原子疏水作用的熵效应变化方向.

$$E_p(E)=\sum_{i=1}^{n}\frac{e^2}{4\pi\varepsilon_0}\cdot\frac{\mathrm{MCI}_p\cdot Z_i}{r_{pi}}, \tag{1}$$

$$E_p(S)=\sum_{i=1}^{n}\varepsilon_{pi}\cdot\left[\left(\frac{R_{pi}^*}{r_{pi}}\right)^{12}-2\cdot\left(\frac{R_{pi}^*}{r_{pi}}\right)^{6}\right], \tag{2}$$

$$E_p(H)=\sum_{i=1}^{n}S_p\cdot\mathrm{MHI}_p\cdot S_i\cdot a_i\cdot e^{-r_{pi}}\cdot T_{pi}. \tag{3}$$

(5) 算法. 预备工作: 首先对样本集中所有分子进行逐一几何优化，进而将优化得到的分子立体结构参照模板化合物加以叠合，并划分出 5×5×5 的空间网格. 遗传筛选建模工作: 在每一个格点位置随机放置一种探针类型，同时计算该探针与样本集中每一个药物分子的 3 种作用势能大小，这样 125 个格点可计算出 375 项作用能，并作为遗传算法中一个个体的染色体串，这个个体就代表了格点上探针原子放置的一种组合方式. 使用该个体对每个分子的 375 项作用能作为自变量 X 并与其生物活性(因变量Y)建立 PLS 模型，并采用模型的交叉检验均方根误差($\mathrm{RMS_{CV}}$)作为个体适应度评价函数. 对包含多个个体的遗传种群通过复制、交叉、变异操作反复叠代直到收敛，并以其中最优个体作为最终确定模型，而该个体的探针原子排布模式即为优化得到的虚拟受体.

2 数据集

31 个甾体化合物是一组经典的药物分子样本集, 始见于Dunn等[12,13]的报告中. 自从 1988 年Cramer等[3]首次将其用于CoMFA分析以来, 已有大量关于其QSAR研究的报道[4~6,15~24]. 该组化合物的活性被表示为与一种转移蛋白即键合肾上腺皮质层球蛋白(CBG)亲和能的负对数值pK_a(表 2).

3 结果与讨论

首先在Chemoffice 8.0中搭建31个甾体分子的初始结构, 并使用半经验量子化学软件MOPAC 6.0 在AM1 水平上对这些分子进行几何优化和原子净电荷数计算. 这里以具有最大亲和活性之一的 11 号分子deoxycortisol为模板化合物进行分子叠合. 对于样本集划分我们采用Cramer等[3]的方案, 即将后 10 个样本作为测试集对前 21 个样本所建模型进行检验. 首先使用训练集 21 个样本来建立CoMPIA模型, 其遗传算法相关参数为: 初始群体大小 100; 最大遗传代数 200; 变异概率 0.5%; 交叉互换 2 点; 交互检验: 留一法; 数据预处理: 标准化; 其他均为默认设置. 筛选结果(最优个体): 主成分数(A) 4; 适应度(RMS_{CV}) 0.395. 通过比较遗传算法优化得到的排在前 10 位的CoMPIA模型可以发现它们的复相关系数R^2 及交叉检验Q^2均在 0.9 和 0.8 以上, 表明模型具有良好的自身拟合能力及稳定性. 值得注意的是, 这些个体的非空探针数目大约为 80 个左右, 也就是说算法确定的空探针数约占到总格点数(125 个)的 36.0%, 而实际上空探针平均出现概率为 11.1%, 这说明在CoMPIA的计算结果中体现出了受体没有抵达网格的某些部位(如分子内部和结合腔的背面等), 因此赋予空值, 这样就避免了噪声和干扰因素的引入. 进一步对最优个体进行分析表明, 该模型包含的 4 个主成分解释了Y变量 94.0%的方差, 交叉检验所解释的Y变量方差为 86.8%. 另外在Defaye等[14]的报告中提到甾体化合物与受体蛋白结合时, 其B, C, D环的侧链分别与半胱氨酸、蛋氨酸和组氨酸等残基发生作用. 为了分析最优模型是否体现出了这些作用方式, 我们首先以距离远近程度将 125 个格点分配给甾核上A, B, C, D 4 个环, 它们分别得到的格点数目为 38, 28, 24, 35, 进而计算上述氨基酸中的特殊原子硫(sp^3)和氮(sp^2)在对应环所占据格点上的出现频率, 发现硫原子(sp^3)在B环和C环出现次数分别占格点数目的 21.4%和16.7%, 而氮原子(sp^2)在D环周围出现频率为 25.7%, 均大于平均值 11.1%, 该结果表明CoMPIA所得到的虚拟受体分布模式确实能够在一定程度上反映出受体的真实状态. 这里我们在 11 号甾体分子的立体结构图上将不同环所对应的硫(sp^3)和氮(sp^2)原子用线连接起来绘于图 1 中, 可以看到虚拟受体中特殊原子较好地分布在环侧链周围. 另外为了便于比较, 表 3 列出了已有文献的报道以及本文的建模结果, 其中CoMPIA模型的复相关系数R^2 仅比Tuppurainen等[16]使用电子特征值(EEVA)指数的拟合值略低, 而交叉检验的Q^2 也优于大多数文献报道. 继而使用该模型对测试集 10 个样本活性进行预测, 亦取得了较优效果(表 4). 图 2 是CoMPIA模型对该组样本计算值与实验观测值的相关情况, 可以看到大部分样本都较好地落在过原点的 45°斜线附近,

表 2 31 个甾体化合物

编号	化合物	pK_a	编号	化合物	pK_a	编号	化合物	pK_a
1	aldosterone	−6.279	12	dihydrotestosterone	−5.919	23	cortisolacetate	−7.553
2	androstanediol	−5.000	13	estradiol	−5.000	24	4-pregnene-3,11,20-trione	−6.779
3	androstenediol	−5.000	14	estriol	−5.000	25	epicorticosterone	−7.200
4	androstenedion	−5.763	15	estrone	−5.000	26	19-nortestosterone	−6.144
5	androsterone	−5.613	16	etiocholanolone	−5.225	27	16a,17a-dihydroxyprogesterone	−6.247
6	corticosterone	−7.881	17	pregnenolone	−5.225	28	17a-methylprogesterone	−7.120
7	cortisol	−7.881	18	hydroxypregnenolone	−5.000	29	19-norprogesterone	−6.817
8	cortisone	−6.892	19	progesterone	−7.380	30	2a-methylcortisol	−7.688
9	dehydroepiandrosterone	−5.000	20	hydroxyprogesterone	−7.740	31	2a-methyl-9a-fluorocortisol	−5.797
10	deoxycorticosterone	−7.653	21	testosterone	−6.724			
11	deoxycortisol	−7.881	22	prednisolone	−7.512			

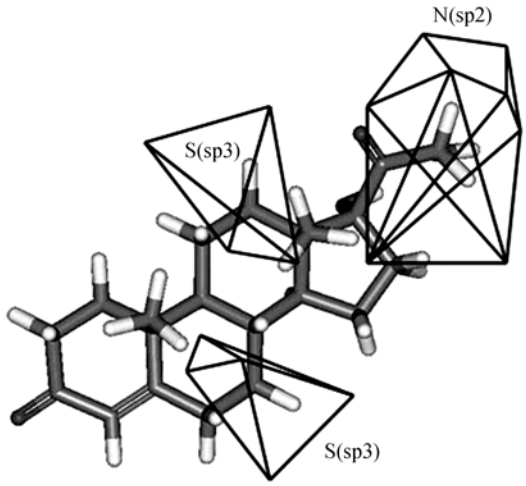

图 1 硫(sp^3)和氮(sp^2)原子在对应甾环周围的分布情况

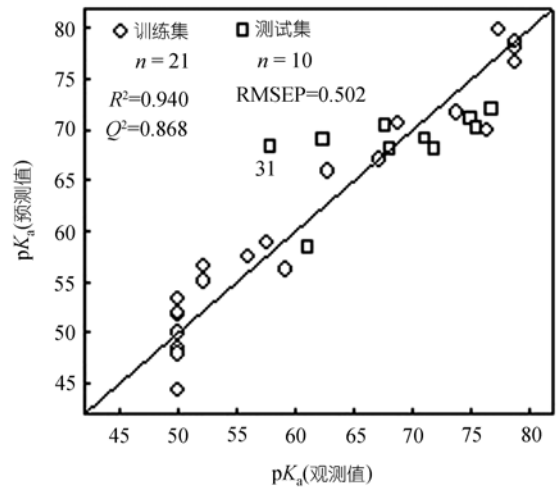

图 2 甾体化合物 CoMPIA 模型计算值与实验观测值的相关性

表 3 当今流行的分子表征方法和 CoMPIA 对 21 个甾体化合物的建模结果比较

编号	方法	A	R^2	Q^2	编号	方法	A	R^2	Q^2	编号	方法	A	R^2	Q^2
1	CoMFA[3]	2	0.897	0.662	6	MEDV[19]	5	0.881	0.793	11	CoMMA[5]	3	0.856	0.828
2	PWPLS[15]	1	0.847	0.792	7	MSM[20]	2	–	0.820	12	CoMSA[6]	1	0.921	0.880
3	EEVA[16]	3	0.970	0.840	8	TQSAR[21]	6	0.939	0.832	13	RSM[23]	1	0.702	0.646
4	MFTA[17]	2	0.920	0.715	9	MTD[22]	4	0.930	0.700	14	PARM[24]	1	0.834	0.806
5	MS-WHIM[18]	2	–	0.631	10	CoMSIA[4]	4	0.937	0.665	15	CoMPIA	4	0.940	0.868

表 4 测试集 10 个甾族化合物的部分文献报道及 CoMPIA 预测结果

编号	观测活性[13]	CoMFA[3]	MS-WHIM[18]	MEDV[19]	MSM[20]	TQSAR[21]	RSM[23]	PARM[24]	CoMPIA
22	−7.512	−8.084	−7.300	−8.166	−7.453	−7.237	−7.417	−7.449	−7.116
23	−7.553	−7.666	−8.332	−7.553	−7.022	−7.879	−7.646	−8.037	−7.008
24	−6.779	−6.538	−6.821	−7.652	−6.939	−6.648	−6.647	−6.601	−7.042
25	−7.200	−7.804	−7.445	−7.765	−7.146	−7.809	−6.905	−6.015	−6.803
26	−6.144	−6.396	−6.121	−5.659	−5.908	−6.832	−6.164	−6.246	−5.844
27	−6.247	−7.346	−6.901	−6.666	−7.046	−7.318	−6.841	−5.742	−6.903
28	−7.120	−7.010	−6.532	−7.340	−6.569	−7.363	−6.623	−6.925	−6.914
29	−6.817	−6.864	−6.838	−5.740	−6.850	−7.540	−6.550	−6.100	−6.807
30	−7.688	−7.970	−7.860	−7.642	−7.539	−7.628	−6.948	−6.108	−7.208
31	−5.797	−8.005	−7.491	−6.845	−7.457	−7.537	−6.116	−5.991	−6.830
RMSEP		0.837	0.662	0.650	0.380	0.640	0.762	0.709	0.502

仅有第 31 号样本预测结果的正向偏差较大，经分析表明该分子是惟一一个含有氟原子的化合物，因为它结构上的特殊性使得模型对其活性预测结果偏大.

4 结论

本文将虚拟受体的概念和 CoMFA 思想有机地结合在一起，得到了一种新颖的 3D-QSAR 研究方法：比较分子/虚拟受体相互作用分析法. 该法通过定义受体原子探针以及利用遗传算法优化探针在药物分子周围的最佳分布方式来找寻模型的最优解，同时也间接构造了虚拟受体模式，从而有效地提高了 QSAR 模型的统计质量和可解释性. 文中使用 31 个经典甾体化合物作为研究对象，成功建立了该组样本的 3D-QSAR 模型，同时从探针原子分布情况来看

CoMPIA 优化得到的虚拟受体在一定程度上能够反映药物分子与真实受体的结合方式. 上述结论表明基于虚拟受体和遗传优化的 CoMPIA 方法在药物 QSAR 研究中具有可喜的应用前景.

致谢 本工作为化学生物传感与计量学国家重点实验室基金(批准号: 05-12-1)和重庆大学创新基金(批准号: 06-1-1)资助项目.

参 考 文 献

1 Hansch C, Fujita T. ρ-σ-π analysis: a method for the correlation of biological activity and chemical structure. J Am Chem Soc, 1964, 86(8): 1616—1626

2 Free S M, Wilson J B. A mathematical contribution to structure-activity studies. J Med Chem, 1964, 7(4): 395—399

3 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]

4 Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem, 1994, 37: 4130—4146 [DOI]

5 Silverman B D, Platt D E. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. J Med Chem, 1996, 39: 2129—2140 [DOI]

6 Polanski J, Walczak B. The comparative molecular surface analysis (CoMSA): a novel tool for molecular design. Comput & Chem, 2000, 24: 615—625 [DOI]

7 Hahn M. Receptor surface models. 1. Definition and construction. J Med Chem, 1995, 38(12): 2080—2090 [DOI]

8 Bondi A. van der Waals volumes and radii. J Phys Chem, 1964, 68(3): 441—451

9 Pei J, Wang Q, Zhou J, et al. Estimating protein-ligand binding free energy: atomic solvation parameters for partition coefficient and solvation free energy calculation. Proteins, 2004, 57: 651—664 [DOI]

10 Levitt M. Protein folding by restrained energy minimization and molecular dynamics. J Mol Biol, 1983, 170: 723—764

11 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]

12 Dunn J F, Nisula B C, Rodbard D. Transport of steroid hormones: binding of 21 endogeneous steroids to both testosterone-binding globulin and corticosteroid-binding globulin in human plasma. J Clin Endocrin Metab, 1981, 53: 58—68

13 Mickelson K E, Forsthoefel J, Westphal U. Steroid-protein interactions. Human corticosteroid binding globulin: some physicochemical properties and binding specificity. Biochemistry, 1981, 20: 6211—6218 [DOI]

14 Defaye G, Basset M, Monnier N, et al. Electron spin resonance study of human transcortin: thiol groups and binding site topography. Biochim Biophys Acta, 1980, 623(2): 280—294

15 Tominaga Y, Fujiwara I. Prediction-weighted partial least-squares regression method (PWPLS). 2. Application to CoMFA. J Chem Inf Comput Sci, 1997, 37: 1152—1157 [DOI]

16 Tuppurainen K, Viisas M, Laatikainen R, et al. Evaluation of a novel electronic eigenvalue (EEVA) molecular descriptor for QSAR/QSPR studies: validation using a benchmark steroid data set. J Chem Inf Comput Sci, 2002, 42: 607—613 [DOI]

17 Palyulin V A, Radchenko E V, Zefirov N S. Molecular field topology analysis method in QSAR studies of organic compounds. J Chem Inf Comput Sci, 2000, 40: 659—667 [DOI]

18 Bravi G, Gancia E, Mascagni P, et al. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D QSAR study in a series of steroids. J Comput-Aided Mol Des, 1997, 11: 79—92 [DOI]

19 Liu S S, Yin C S, Wang L S. Combined MEDV-GA-MLR method for QSAR of three panels of steroids, dipeptides, and COX-2 inhibitors. J Chem Inf Comput Sci, 2002, 42: 749—756 [DOI]

20 Good A C, Peterson S J, Richards W G. QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods. J Med Chem, 1993, 36: 2929—2937 [DOI]

21 Robert D, Amat L, Carbó-Dorca R. Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures: prediction of the corticosteroid-binding globulin binding affinity for a steroid family. J Chem Inf Comput Sci, 1999, 39: 333—344 [DOI]

22 Oprea T I, Ciubotariu D, Sulea T I, et al. Comparison of the minimal steric difference (MTD) and comparative molecular field analysis (CoMFA) methods for analysis of binding of steroids to carrier proteins. Quant Struct-Act Relat, 1993, 12: 21—26

23 Hahn M, Rogers D. Receptor surface models. 2. Application to quantitative structure-activity relationship studies. J Med Chem, 1995, 38(12): 2091—2102 [DOI]

24 Chen H, Zhou J, Xie G. PARM: a genetic evolved algorithm to predict bioactivity. J Chem Inf Comput Sci, 1998, 38: 243—250 [DOI]

(2005-10-10 收稿, 2006-03-14 接受)