[Article]

doi: 10.3866/PKU.WHXB201305171

www.whxb.pku.edu.cn

基于机器学习方法的丙型肝炎病毒聚合酶NS5B非核苷抑制剂的定量构效关系研究

丛湧¹ 薛英^{1,2,*}⁽¹⁾四川大学化学学院, 教育部绿色化学与技术重点实验室, 成都 610064;⁽²⁾西华大学四川省先进科学计算重点实验室, 成都 610039)

摘要: 对89个苯并异噻唑和苯并噻嗪类丙型肝炎病毒(HCV) NS5B聚合酶非核苷抑制剂进行了定量构效关系(QSAR)研究. 采用遗传算法组合偏最小二乘(GA-PLS)和线性逐步回归分析(LSRA)两种特征选择方法选择最优描述符子集, 然后建立多元线性回归和偏最小二乘线性回归模型. 并首次尝试使用遗传算法耦合支持向量机方法(GA-SVM)对两种特征选择方法所选的描述符子集分别建立非线性支持向量机回归模型. 三种机器学习方法所建模型均得到比较满意的预测效果. 采用LSRA所选的6个描述符建立的三个QSAR模型对于测试集的相关系数为0.958–0.962, GA-SVM法给出最好的预测精度(0.962). 采用GA-PLS所选的7个描述符建立的三个QSAR模型对于测试集的相关系数为0.918–0.960, 偏最小二乘回归模型的结果最好(0.960). 本工作提供了一种有效的方法来预测丙型肝炎病毒抑制剂的生物活性, 该方法也可以扩展到其他类似的定量构效关系研究领域.

关键词: 丙型肝炎病毒NS5B聚合酶; 非核苷抑制剂; 线性逐步回归分析; 偏最小二乘法; 遗传算法; 支持向量机

中图分类号: O647

Quantitative Structure–Activity Relationship Study of the Non-Nucleoside Inhibitors of HCV NS5B Polymerase by Machine Learning Methods

CONG Yong¹ XUE Ying^{1,2,*}

⁽¹⁾College of Chemistry, Key Laboratory of Green Chemistry and Technology, Ministry of Education, Sichuan University, Chengdu 610064, P. R. China; ⁽²⁾Key Laboratory of Advanced Scientific Computation, Xihua University, Chengdu 610039, P. R. China)

Abstract: The quantitative structure–activity relationship (QSAR) approach was used to predict the activity of two different scaffolds (benzothiazole and benzothiazine) of 89 non-nucleoside inhibitors of hepatitis c virus (HCV) NS5B polymerase. Two selection methods, linear stepwise regression analysis (LSRA) and genetic algorithm-partial least squares (GA-PLS), were used to select appropriate descriptor subsets for QSAR modeling with linear models. The genetic algorithm-support vector machine (GA-SVM) approach was first used to build nonlinear models with six LSRA- and seven GA-PLS-selected descriptors. Three QSAR models built with the six LSRA-selected descriptors gave correlation coefficients of 0.958–0.962 for the training set. GA-SVM provided the highest prediction accuracy of the models of 0.962. Three QSAR models built with the seven GA-PLS-selected descriptors gave correlation coefficients of 0.918–0.960 for the training set, of which the partial least squares (PLS) model was the best (0.960). The

Received: February 1, 2013; Revised: May 17, 2013; Published on Web: May 17, 2013.

*Corresponding author. Email: yxue@scu.edu.cn; Tel: +86-28-85418330.

The project was supported by the National Natural Science Foundation of China (21173151) and Open Research Fund of Key Laboratory of Advanced Scientific Computation, Xihua University, China (szjj2011-029).

国家自然科学基金(21173151)和西华大学先进科学计算省重点实验室开放基金(szjj2011-029)资助项目

© Editorial office of *Acta Physico-Chimica Sinica*

investigated models gave satisfactory prediction results and can be extended to other QSAR studies.

Key Words: HCV NS5B polymerase; Non-nucleoside Inhibitor; Linear stepwise regression analysis; Partial least square; Genetic algorithm; Support vector machine

1 引言

丙型肝炎病毒(HCV)是通过血液传染的非甲、非乙型肝炎(non-A, non-B viral hepatitis)的主要致病因子. 全世界约有1.7亿慢性丙肝病毒感染者, 每年有35万余人死于与丙肝相关的肝脏疾病.^{1,2} 由于缺少特异有效的药物和疫苗, 目前临床上采用 α -2-聚乙二醇化干扰素(pegylated interferon- α -2a)与利巴韦林(ribavirin)联合用药治疗丙型肝炎, 其治疗有效率仅为50%, 治疗周期较长, 不良反应发生频率高且比较严重.³ HCV的高度变异性使治疗和预防面临巨大困难, 急需研发有效的抗HCV药物, 以补充完善目前现有的治疗方法. 近年来, 寻找HCV特定靶向抗病毒治疗药物(STAT-C)是抗HCV研究的重要方向, 这些特定靶向的靶点包括NS3/NS4A蛋白酶/解旋酶和非结构蛋白5B(NS5B)RNA依赖的聚合酶(NS5B RNA-dependent RNA polymerase)等, 其中以丙型肝炎病毒NS5B RNA聚合酶为靶标的抗HCV药物研究颇受关注.⁴

NS5B是一种病毒编码的RNA依赖性RNA聚合酶, 可以调控丙型肝炎病毒RNA模板(-)链的合成及(+)链基因组RNA的再生,⁵ 在丙型肝炎病毒复制进程中必不可少. 研究发现通过对NS5B的抑制可以有效阻断丙肝病毒复制, 以NS5B为靶点的抑制剂设计、合成及生物测试越来越成为该领域的研究热点, 一系列具有不同骨架结构的核苷型(NIs)、非核苷型(NNIs) NS5B抑制剂被不断的合成及报道,⁶ 其中包括苯并咪唑,⁷ 吡啶,⁸ 噻吩,⁹ 吡啶酮衍生物,¹⁰ 苯并噻二嗪¹¹等类型抑制剂. 然而, 这些研究主要集中在合成新的抑制剂化合物及测试其生物抑制活性, 实验周期较长且资金消耗巨大. 因此, 在早期药物研发过程中, 为了缩短药物开发周期和节约成本, 我们急需大力发展计算机模拟技术对小分子抑制剂生物活性做出初步定性和定量的预测, 并从建立的数据统计模型中挖掘大分子蛋白靶点与小分子配体的可能作用机理, 进一步指导和辅助新型抗HCV药物分子的设计和发现.

结构-活性关系研究(SAR)已经成功应用于抗HCV抑制剂的发现. 在这些方法当中, 定量结构-活

性关系(QSAR)研究分子结构与其所表达生物活性之间的相关性, 在广泛实践中已被证明是一种非常有效的计算机辅助药物设计工具. 与其他方法相比, QSAR方法的优点在于其建立的数据统计模型可以直观指示在生物抑制活性确定中发挥重要作用的小分子抑制剂结构性因素; 构建模型所需的分子描述符独立于任何实验条件, 可以通过小分子抑制剂三维结构直接计算得出; 通过QSAR方法得到的构效关系能够提供非常有用的信息, 这些信息可以进一步指导及辅助基于配体(ligand-based)和基于受体(receptor-based)的药物设计. 然而, 遗憾的是, 基于机器学习的二维定量构效关系研究并不能明确揭示生物大分子蛋白受体与小分子抑制剂配体结合的三维构象; 在这样的二维定量构效关系研究中, 研究人员经常不会考虑样本分子集三维构象的叠合和取向. 基于大分子靶蛋白和小分子抑制剂三维构象的3D-QSAR仍然是定量构效关系研究的发展趋势, 也是我们工作组未来的研究重点. 传统的化学信息学方法多采用多元线性回归(MLR)、启发式方法(HM)、主成分回归(PCR)和偏最小二乘方法(PLS)等线性定量构效关系方法建模. 近年来, 支持向量机(SVM)和径向基函数网络(RBF)等非线性回归方法在多样性分子结构样本集中对药效学、药代动力学和毒理学性质不断显示其优异的预测性能,^{12,13} 正越来越多的受到关注. Melagraki等¹⁴对98个苯并噻二嗪衍生物抗HCV抑制剂进行了定量构效研究, 通过采用线性逐步回归特征消除选择方法(ES-SWR)从ChemSar和Topix软件计算的分子描述符集中筛选出5个重要的分子描述符建立QSAR模型. 他们的QSAR模型对训练集的平方相关系数(R^2)仅为0.74, 对验证集的 R^2 为0.81. Li研究小组¹⁵运用最佳多元线性回归方法(BMLR)选择建模最优描述符子集, 并采用多元线性回归、径向基函数神经网络(RBFNN)和支持向量机方法建立线性和非线性QSAR模型对118个苯并噻二嗪衍生物抗HCV抑制剂进行了定量构效关系研究,¹⁵ 其中非线性RBFNN和SVM模型给出更为准确和理想的预测结果, 两种非线性模型对于训练集的 R^2 分别为0.850和

0.875, 对于测试集的 R^2 分别为 0.893 和 0.854. 虽然苯并噻二嗪衍生物抗 HCV 抑制剂的定量构效关系研究已经取得了一些有成效的结果, 然而如何选择合适的特征选择和建模机器学习方法仍然是影响 QSAR 模型预测能力的两个关键因素. 本工作研究了 89 个最新报道的苯并异噻唑(benzoisothiazole)和苯并噻嗪(benzothiazine)类抗 HCV 抑制剂的定量构效关系. 我们尝试使用遗传算法组合偏最小二乘(GA-PLS)和线性逐步回归分析(LSRA)特征选择方法选择最优描述符子集. 对每种特征选择方法选择的描述符子集分别采用多元线性回归、偏最小二乘、遗传算法组合支持向量机(GA-SVM)三种方法用训练集建立 QSAR 模型, 并用这些模型预测了测试集中化合物的 NS5B 抑制活性.

2 材料与方法

2.1 丙型肝炎病毒聚合酶 NS5B 抑制剂数据集的选取

本工作从最近发表的文献中总共收集了 89 个苯并异噻唑^{16,17}和苯并噻嗪¹⁸⁻²⁰类丙型肝炎病毒聚合酶 NS5B 抑制剂. 这些化合物的 IC_{50} 值($1-868000 \text{ nmol} \cdot \text{L}^{-1}$)均用相同的实验测量方法在相同的实验条件下测定, 被测抑制剂中大部分是高效 NS5B 抑制剂. 首先使用 ChemDraw 软件绘制每个小分子抑制剂的二维结构, 随后通过 CORINA 软件将小分子二维平面结构转换成三维结构, 再采用量子化学 AM1 计算方法优化每一个抑制剂分子的三维构型, 然后对计算结果进行手动检查以确保每个优化分子都生成正确的手性结构且没有重复. 为了便于研究, 我们将抑制剂的 IC_{50} 值转换成 pIC_{50} 值($\lg(10^6/IC_{50})$)作为 QSAR 模型的因变量. 根据化合物的结构与化学性质在化学空间中的相似性和分布,²¹ 将所有的化合物分为训练集和测试集两大类. 训练集(含 45 个小分子抑制剂)训练和优化回归模型, 测试集(含 44 个小分子抑制剂)评价回归模型的预测能力.

2.2 分子描述符的计算

分子描述符在定量构效关系研究中经常被用于定量描述分子的结构和物理化学特性. 在我们的研究中, 通过手动方法从相关文献中找出一千多个分子描述符, 并根据化合物拓扑结构、电子结构和几何结构等性质, 剔除明显冗余和与预测药物性质不相关的描述符,²² 最终筛选出 189 个与化合物性质密切相关的分子描述符(见表 1). 其中包括 18 个简

单分子性质描述符(如分子量、可旋转的键数), 27 个分子连接性和形状描述符(如分子连接性指数和分子卡伯形状指数), 97 个电拓扑态分子描述符(如电子拓扑态指数), 22 个量子化学性质分子描述符(如原子电荷和分子的偶极矩)和 25 个分子几何特性描述符(如溶剂可及表面积和疏水区域). 我们应用实验室自编的分子描述符计算程序, 根据 AM1 方法优化过后的化合物三维结构计算所有的分子描述符. 为了减少建模中描述符之间高度自相关带来的多重共线性干扰, 在特征选择之前对这 189 个分子描述符集做了预处理, 步骤如下: (1) 移除在 90% 的样本分子中具有相同数值的描述符; (2) 移除相对标准偏差小于 0.05 的描述符; (3) 对于 Pearson 相关系数超过 0.95 的一对描述符, 留下与生物活性相关性较高的描述符, 剔除另外一个.²³ 经过上述预处理, 我们最终保留了 85 个分子描述符进行下一步的特征选择.

2.3 特征选择方法

2.3.1 逐步回归分析法

我们借助 SPSS 软件自带的逐步回归分析程序选择最优描述符子集, 逐步回归分析法中每步有两个过程即引进变量和剔除变量, 且引进变量和剔除变量均需作 F 检验后方可继续进行, 故又称为双重检验回归分析法. 其具体步骤如下: (1) 引入变量, 引入变量的原则是未引进变量中偏回归平方和最大者并经过 F 显著性检验, 若显著则引进, 否则终止. (2) 剔除变量, 剔除原则是在引进的自变量中偏回归平方和最小者, 并经过 F 检验不显著, 则剔除. (3) 终止条件即最优条件, 再无显著自变量引进, 也没有不显著自变量可以剔除.

2.3.2 遗传算法组合偏最小二乘方法

我们借助 MATLAB 偏最小二乘-遗传算法工具箱²⁴实现最优描述符子集的选取. GA-PLS 是一种基于遗传算法的优化工具,^{25,26} 其算法过程描述如下: (1) 定义和编码染色体; (2) 种群的初始化; (3) 评价每个染色体的适应度; (4) 保护染色体; (5) 保留最好的染色体; (6) 对种群进行交叉和变异遗传操作; (7) 若满足终止条件停止程序, 否则转入步骤 3. 本实验采用了 GA-PLS 工具箱里的三个函数, 它们分别是 GAPLSOPT(dataset, 1), GAPLSOPT(dataset, 2) 和 GAPLS(dataset, the number of evaluation, precision).

GAPLSOPT(dataset, 1)函数测试样本数据集是否适用 GA-PLS 方法选取特征, 根据 GA-PLS 软件

表1 所用的分子描述符
Table 1 Molecular descriptors used in this work

Descriptor class	Number of descriptor in class	Descriptor
simple molecular properties	18	molecular weight, numbers of rings, rotatable bonds, H-bond donors, and H-bond acceptors, element counts
molecular connectivity and shape	28	molecular connectivity indices, valence molecular connectivity indices, molecular shape Kappa indices, Kappa alpha indices, flexibility index
electrotopological state	97	electrotopological state indices, and atom type
quantum chemical properties	31	electrotopological state indices, Wiener index, centric index, Altenburg index, Balaban index, Harary number, Schultz index, PetitJohn R2 index, PetitJohn D2 index, mean distance index, PetitJohn I2 index, information Wiener, Balaban rmsd index, graph distance index
geometrical properties	25	polarizability index, hydrogen bond acceptor basicity(covalent HBAB), hydrogen bond donor acidity (covalent HBDA), molecular dipole moment, absolute hardness, softness, ionization potential, electron affinity, chemical potential, electronegativity index, electrophilicity index, most positive charge on H, C, N, O atoms, most negative charge on H, C, N, O atoms, most positive and negative charge in a molecule, LSum of squares of charges on H,C,N,O and all atoms, mean of positive charges, mean of negative charges, mean absolute charge, relative positive charge, relative negative charge
		length vectors (longest distance, longest third atom, 4 th atom), molecular van der Waals volume, solvent accessible surface area, molecular surface area, van der Waals surface area, polar molecular surface area, sum of solvent accessible surface areas of positively charged atoms, sum of solvent accessible surface areas of negatively charged atoms, sum of charge weighted solvent accessible surface areas of positively charged atoms, sum of charge weighted solvent accessible surface areas of negatively charged atoms, sum of van der Waals surface areas of positively charged atoms, sum of van der Waals surface areas of negatively charged atoms, sum of charge weighted van der Waals surface areas of positively charged atoms, sum of charge weighted van der Waals surface areas of negatively charged atoms, molecular rugosity, molecular globularity, hydrophilic region, hydrophobic region, capacity factor, hydrophilic-hydrophobic balance, hydrophilic integrity moment, hydrophobic integrity moment, amphiphilic moment

设计者介绍, 如果 GAPLSOPT 函数对于样本数据集的测试输出结果在 0 到 5 之间, 使用 GA-PLS 方法对数据集进行特征选择是安全稳定的. 图 1 中显示了本工作 89 个小分子抑制剂对应的 85 个分子描述符样本数据集的 GAPLSOPT(dataset, 1) 输出结果, 样本数据集的随机测试结果在 0 到 4.9468 之间, 这说明采用 GA-PLS 方法对该数据集进行特征选择是可靠的. 为了避免 GA-PLS 方法在训练过程中产生过拟合, 我们使用 GAPLSOPT(dataset, 2) 函数估计 GAPLS 函数所需的最优评价次数(the number of evaluation)参数, 如图 2 所示, GAPLSOPT(dataset, 2) 的差异曲线在评价次数为 115 处有全局最大值, 因此, 评价次数被置为 115 作为 GAPLS 函数的最优控制参数. 经过上述准备工作, 我们运行 GAPLS 函数对数据集进行特征选择. 为了减少随机误差, 我们重复了 10 次 GAPLS 实验得到平均结果. 图 3 显示了交叉验证响应和每个描述符的被选择频率, 在 85 个分子描述符中, GAPLS 函数最终选出 7 个分子描述

符用于下一步的 QSAR 建模.

2.4 遗传算法组合支持向量机方法

支持向量机是基于结构风险最小化(SRM)理论的机器学习方法, 其两分类理论经过多年发展已日趋成熟. 通过引入 ε -不敏感损失函数, 支持向量机也

图1 GAPLSOPT(1)函数测试结果

Fig.1 GAPLSOPT(1) test results

图2 GAPLSOPT(2)差异曲线
Fig.2 GAPLSOPT(2) difference curve

可以扩展用来解决小样本数据集的回归问题. 支持向量机经由核函数 $K(x_i, x_j)$ 将输入空间的 X 向量映射到高维希尔伯特空间 H , 其常用的核函数包括线性、多项式、径向基和 S 形等函数形式, 其中径向基函数由于良好的非线性映像能力已在诸多领域得到了广泛应用.²⁷ 本研究使用高斯径向基函数(RBF)构建支持向量机回归模型. 高斯核函数具体表示如下:

$$K(x_i, x_j) = \exp(-\|x_j - x_i\|^2 / 2\sigma^2) \quad (1)$$

我们在 $K(x_i, x)$ 特征空间构造决策函数(方程2)作为 ε -支持向量机的最优解:

$$f(x) = \sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x) + b \quad (2)$$

其中 b 为方程的偏置项.

支持向量回归机的泛化性能取决于最优正则化参数 C , 不敏感参数 ε 和RBF核函数宽度 σ 的选取.²⁸ 正则化参数 C 对回归函数的复杂性和泛化能力

进行折衷. 在确定的数据子空间中, 参数 C 取得太小, 则对样本数据中超出 ε 不敏感带的样本惩罚就越小, 使训练误差变大, 系统的泛化能力变差, 会出现“欠学习”现象; C 取得太大, 相应 $\|\omega\|^2/2$ 的权重就小, 系统的泛化能力变差, 会出现“过学习”现象. 不敏感参数 ε 控制着 ε 不敏感带的宽度, 影响着支持向量的数目. ε 值选得太小, 回归估计精度高, 但支持向量数目增多, ε 选的太大, 回归估计精度降低, 支持向量数目减少, 支持向量机的稀疏性大. RBF核函数宽度 σ 反映了训练样本数据的分布或范围特性, 它确定了局部领域的宽度, 较大的 σ 意味着较低的方差.

遗传算法是模拟达尔文生物进化论自然选择和遗传学机理的生物进化过程的计算模型, 是一种通过模拟自然进化过程搜索最优化解的方法.²⁹ 本工作中我们采用遗传算法同时对支持向量回归机的三个参数(正则化参数 C , 不敏感参数 ε 和RBF核函数宽度参数 σ)进行优化. 新型进化计算框体—Python环境下分布式进化算法(DEAP)³⁰被用来实现SVM参数优化的遗传算法框架, Libsvm程序³¹被用于构建 ε -SVM回归模型, 通过引入Python多路处理模块(The multi-processing techniques in Python), 我们实现了GA-SVM算法的并发, 大大提高了算法的计算速度. 在遗传算法中, 染色体使用二进制编码, 每个染色体由 C : (C_i , $i=1-nc$), ε : (ε_j , $j=1-n\varepsilon$)和 σ : (σ_k , $k=1-n\sigma$)三部分组成, C_i 表示染色体中表征正则化参数 C 二进制位串中第 i 位的数值, σ_k 表示染色体中表征RBF核函数宽度参数 σ 二进制位串中第 j 位的数值, ε_k 表示染色体中表征不敏感参数 ε 二进制位串中第 k 位的数值, nc 、 $n\sigma$ 和 $n\varepsilon$ 分别表示染色体中表征 C 、 σ 和 ε 的二进制位串的长度(在本工作中 $nc=10$, $n\sigma=10$, $n\varepsilon=10$). 根据解码公式(5), 这三个二进制编码位串被分别转换成正则化参数 C 、不敏感参数 ε 和RBF核函数宽度参数 σ 的实数值.

$$R = 2^{\frac{\min_R + \frac{\max_R - \min_R}{2^l - 1} \times d}{2}} \quad (3)$$

其中 R 表示二进制位串所表征相应参数的真实值, $\max_R$ 和 $\min_R$ 为相应参数最大和最小指数幂取值范围(参数取值在 $2^{\min_R}$ 和 $2^{\max_R}$ 之间), d 为二进制位串转十进制数值, l 为二进制位串的长度. 我们设置 C 的取值范围为 2^{-10} – 2^{15} ; ε 的取值范围为: 2^{-10} – 2^{10} ; σ 的取值范围为 2^{-10} – 2^8 ; n 重交叉验证均方误差(MSECV)

图3 GAPLS函数描述符选择频率估计
Fig.3 Selected frequency figure by GAPLS function
Ten repetitions were executed to obtain an average result.

作为遗传算法的适应度函数, 定义如下:

$$\text{MSE} = \frac{\sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2}{n_{\text{test}}} \quad (4)$$

其中, y_i 是实验真实值, $\hat{y}_i$ 是模型的计算值, n_{test} 是 n 重交叉验证集的样本化合物数目. 遗传算法的选择、交叉和变异操作, 我们通过调用 DEAP 内置函数——selTournament (individuals, k , tournsize)、cxTwo Points (ind1, ind2) 和 mutFlipBit (individual, indpb) 来实现, 设置种群规模为 500, 总共迭代 100 代, 交叉率和变异率分别设置为 0.6 和 0.2. 当某代种群中 95% 以上的个体得到相同的均方误差适应度数值停止算法, 作为 GA-SVM 的停机准则.

3 结果和讨论

3.1 两种特征选择方法选出的最优描述符子集及讨论

逐步回归分析法共选出 6 个分子描述符, 其中包括 2 个简单分子性质描述符、1 个分子连接性和形状描述符、2 个量子化学性质描述符和 1 个分子几何特性描述符; 遗传算法组合偏最小二乘方法共选出 7 个分子描述符, 其中包括 3 个简单分子性质描述符、1 个分子连接性和形状描述符、1 个电拓扑态分子描述符、2 个量子化学性质描述符和 1 个分子几何特性描述符. 如表 2 所示, 两种特征选择方法同时选出 4 个相同的分子描述符, 分别为分子中杂原子数目(nhet)、氧原子数目(noxy)、分子最大负电荷($A_{Q,\min}$)和亲水性指数(Hiwpl).

起初苯并噻嗪类抑制剂的构性关系研究集中在苯并噻嗪环的 C-7 位,¹⁸ 通过在该位置引入取代基来改善抑制活性. 研究发现在苯并噻嗪环的 C-7 位

引入极性基团如羟基或甲基磺酰胺基会提高抑制剂对酶 NS5B 的抑制效力. RNA 聚合酶 NS5B 和其抑制剂的复合物晶体结构进一步证实, 在 1,3-二巯基功能基团上的负电荷易在环上发生离域与聚合酶 NS5B 骨架上 Tyr448 残基的 N-H 基团及两个保守水分子反生静电相互作用; 与此同时, 苯并噻嗪环中的苯环正好与 Phe193 残基基面对面接触, 苯环之间发生 π - π 堆积作用. 在 C-7 位引入甲基磺酰胺基的苯并噻嗪类抑制剂中, 砒氧基团与桥接 Ser556 残基和甲基磺酰胺基团的结构水分子形成氢键静电相互作用; 磺酰胺基团也会与 NS5B 的 Asn291 残基形成静电相互作用; 氮磺酰胺基团中酸性的 N-H 与 NS5B Asp318 残基产生更强有力的氢键相互作用力, 磺酰胺基团的这种独特的三点式相互作用有助于解释抑制剂中该极性基团的引入对于抑制活性的改善. 构性关系还发现在苯并噻嗪环的 C-2 和 C-5 位掺入吸电子基团降低环电子密度, 可以有效改善胆汁的转运识别, 并同时减少代谢负产物的生成.²⁰

在所选描述符中, nhct (分子中所含 N、O 和 S 等杂原子数目)、noxy (分子中 O 原子数目) 和 nsulph (分子中 S 原子数目) 这三个简单性质描述符间接表征了抑制剂分子中极性基团(如羟基或甲基磺酰胺基等)与 NS5B 发生静电相互作用及形成氢键的能力; 亲水性指数(Hiwpl)描述抑制剂分子与 NS5B 酶的亲水疏水相互作用; 两个量子化学描述符, 分子最大负电荷($A_{Q,\min}$)和分子平均负电荷(Mnc)可能与苯并噻嗪环的电子离域化及 C-2、C-5 位上吸电基团的引入有关; S(1) (羟基 H 原子电拓扑态指数)和 $Q_{H,\max}$ (分子中氢原子上的最大正电荷)给出氢键给体的质子提供能力. 从上面的讨论可以看出, 两种方法所选择的描述符能够反映抑制剂分子的电

表 2 LSRA 和 GA-PLS 特征选择方法选取的分子描述符

Table 2 Molecular descriptors selected from the LSRA and GA-PLS feature selection methods

Molecular descriptor selected by LSRA ^a	Molecular descriptor selected by GA-PLS ^b	Description
nhct	nhct	count of hetero atoms
noxy	noxy	count of O atoms
χ^p		valence molecular connectivity Chi indices for path order 3
$A_{Q,\min}$	$A_{Q,\min}$	most negative charge in a molecule
Mnc		mean of negative charges
Hiwpl	Hiwpl	hydrophilic integrity moment
	nsulph	count of S atoms
	S(1)	atom-type H estate sum for —OH
	$Q_{H,\max}$	most positive charge on H atoms

^asix molecular descriptors selected from statistical product and service solutions (SPSS) linear stepwise regression analysis; ^bseven molecular descriptors selected from GA-PLS method

荷分布、氢键相互作用、分子连接性、亲水疏水相互作用等性质。

3.2 逐步回归分析法所选描述符的建模回归结果

我们采用逐步回归分析法所选的6个最优描述符分别建立多元线性、偏最小二乘和支持向量机回归模型, 训练集(含45个化合物)用来训练和优化模型, 测试集(含44个小分子抑制剂)评价回归模型的预测能力。训练得到的多元线性回归方程为: $pIC_{50} = 0.329 \times nh_{et} + 0.383 \times noxy - 0.167 \times \chi_p^v - 29.524 \times A_{Q,min} + 23.903 \times Mnc + 0.051 \times Hiwpl - 5.588$, $R^2 = 0.940$, $R^2_{adjust} = 0.931$, $s^2 = 0.144$, $F = 99.473$, $N = 45$, R^2_{adjust} 为调整相关系数平方, s^2 为回归分析的标准偏差平方, 即回归分析方差; F 为回归分析的显著性检验。图S1 (Supporting Information)为多元线性回归模型对于训练集和测试集的实验与预测 pIC_{50} 数值对比图, 模型对于训练集、测试集及整个数据集的均方误差(MSE)分别为0.121、0.122和0.122, 相应的相关系数分别为0.970、0.958和0.965。偏最小二乘方法抽选出3个主成分, 训练得到的回归方程为: $pIC_{50} = 0.2153 \times nh_{et} + 0.3652 \times noxy - 0.1463 \times \chi_p^v - 25.6605 \times A_{Q,min} - 1.8054 \times Mnc + 0.1519 \times Hiwpl - 7.3150$, $R^2 = 0.931$, $R^2_{adjust} = 0.925$, $s^2 = 0.166$, $F = 85.454$, $N = 45$, 图S2为偏最小二乘回归模型对于训练集和测试集的实验与预测 pIC_{50} 数值对比图, 模型对于训练集、测试集及整个数据集的MSE分别为0.140、0.122和0.131, 相应的相关系数分别为0.965、0.958和0.961。由于影响NS5B酶抑制剂活性的分子特征极其复杂, 并不是所有的分子描述符都与活性存在绝对的线性关系, 为了与上述线性回归建模方法做比较, 基于相同的最优描述符子集, 我们使用遗传算法组合支持向量机方法建立非线性QSAR模型。首先, 我们在训练集上使用GA-SVM方法同时优化SVM模型的三个参数, 当SVM模型参数取值为: $C = 7.7387$, $\sigma = 0.3546$, $\varepsilon = 0.04664$ 时, 回归模型给出最好的留一法交叉验证误差(MSECV=0.135304); 然后, 我们采用这组优化参数建立SVM模型。图S3为SVM回归模型对于训练集和测试集的实验与预测 pIC_{50} 数值对比图, 模型对于训练集、测试集及整个数据集的MSE分别为0.113、0.108和0.111, 相应的相关系数分别为0.972、0.962和0.968。在上述三种方法建立的QSAR模型中, 支持向量机给出最好的回归模型, 而多元线性回归给出预测效果最好的线性回归模型, 三种机器学习模型对抑制剂活性的预测结果详见表S1。

3.3 遗传算法组合偏最小二乘法所选描述符的建模回归结果

我们采用GA-PLS所选的7个最优描述符子集分别建立偏最小二乘、多元线性和支持向量机回归模型。训练得到的多元线性回归方程为: $pIC_{50} = 0.207 \times nh_{et} + 0.295 \times nsulph + 0.304 \times noxy - 0.035 \times S(1) + 36.165 \times Q_{H,Max} - 4.742 \times A_{Q,min} + 0.115 \times Hiwpl - 6.864$, $R^2 = 0.935$, $R^2_{adjust} = 0.923$, $s^2 = 0.159$, $F = 76.566$, $N = 45$, 图S4为多元线性回归模型对于训练集和测试集的实验与预测 pIC_{50} 数值对比图, 模型对于训练集、测试集及整个数据集的MSE分别为0.131、0.243和0.186, 相应的相关系数分别为0.967、0.918和0.946。偏最小二乘方法抽选出3个主成分, 训练得到的回归方程为: $pIC_{50} = 0.1667 \times nh_{et} + 0.5267 \times nsulph + 0.2915 \times noxy + 0.0220 \times S(1) + 8.1236 \times Q_{H,Max} - 8.7093 \times A_{Q,min} + 0.0827 \times Hiwpl - 3.7535$, $R^2 = 0.933$, $R^2_{adjust} = 0.920$, $s^2 = 0.163$, $F = 73.606$, $N = 45$, 图S5为偏最小二乘回归模型对于训练集和测试集的实验与预测 pIC_{50} 数值对比图。模型对于训练集、测试集及整个数据集的MSE分别为0.134、0.114和0.124, 相应的相关系数分别为0.966、0.960和0.964。基于相同的最优描述符子集, 使用遗传算法组合支持向量机方法建立非线性QSAR模型, 当SVM模型参数取值为: $C = 11.620579$, $\sigma = 0.182558$, $\varepsilon = 0.224924$ 时, 回归模型给出最好的留一法交叉验证误差(MSECV=0.145978), 图S6为SVM回归模型对于训练集和测试集的实验与预测 pIC_{50} 数值对比图, 模型对于训练集、测试集及整个数据集的MSE分别为0.123、0.118和0.121, 相应的相关系数分别为0.970、0.958和0.965。在上述三种方法建立的QSAR模型中, 支持向量机回归模型对于训练集和整个数据集给出最好的预测结果; 两种线性回归模型中, 偏最小二乘模型对测试集和整个数据集给出最好的预测结果, 而多元线性回归模型对于训练集给出最好的预测效果, 三种机器学习模型对抑制剂活性的预测结果详见表S2。

4 结 论

本工作研究了89个苯并异噻唑和苯并噻嗪类抗HCV抑制剂的定量构效关系。线性逐步回归分析和遗传算法组合偏最小二乘方法被用来选取最优描述符子集, 对于如上两种特征选择方法所选描述符, 我们分别采用多元线性回归、偏最小二乘、遗传算法组合支持向量机方法建模并得到了比较满意

的预测结果. 在采用 LSRA 所选描述符建立的三个 QSAR 模型中, 支持向量机给出最好的回归模型, 多元线性回归给出预测效果最好的线性模型; 在采用 GA-PLS 所选描述符建立的三个 QSAR 模型中, 支持向量机同样给出最好的回归模型, 而偏最小二乘回归给出效果最好的线性模型. 研究结果表明, 非线性支持向量机方法的建模回归效果好于传统的多元线性回归和偏最小二乘方法; 采用 LSRA 和 GA-PLS 特征选择方法所选描述符建立的模型都得到比较满意的回归效果, 这表明两种特征选择方法都能筛选合适的描述符子集用于 QSAR 建模; 多元线性回归和偏最小二乘方法建立的线性 QSAR 方程可以直观指示在生物抑制活性确定中发挥重要作用的小分子结构和物理化学特征信息, 这些信息可以进一步指导及辅助基于配体和受体的抗 HCV 药物设计.

Supporting Information: The plots of experimental vs computational $\lg(10^9/IC_{50})$ values are given in Figs.S1–S6. The information of the investigated dataset is provided in Tables S1 and S2. This information is available free of charge via the internet at <http://www.whxb.pku.edu.cn>.

References

- (1) Choo, Q. L.; Weiner, A. J.; Overby, L. R.; Bradley, D. W.; Houghton, M. *Science* **1989**, *244*, 359. doi: 10.1126/science.2523562
- (2) (a) Lauer, G. M.; Walker, B. D. *N. Engl. J. Med.* **2001**, *345*, 41. doi: 10.1056/NEJM200107053450107
(b) Di Bisceglie, A. M. *Lancet* **1998**, *351*, 351.
(c) Alter, M. J.; Kruszon-Moran, D.; Nainan, O. V.; McQuillan, G. M.; Gao, F.; Moyer, L. A.; Kaslow, R. A.; Margolis, H. S. *N. Engl. J. Med.* **1999**, *341*, 556.
- (3) Manns, M. P.; McHutchison, J. G.; Gordon, S. C.; Rustgi, V. K.; Shiffman, M.; Reindollar, R.; Goodman, Z. D.; Koury, K.; Ling, M. H.; Albrecht, J. K. *Lancet* **2002**, *347*, 975.
- (4) (a) Koch, U.; Narjes, F. *Curr. Top. Med. Chem.* **2007**, *7*, 1302. doi: 10.2174/156802607781212211
(b) Rönn, R.; Sandström, A. *Curr. Top. Med. Chem.* **2008**, *8*, 533.
(c) Zapf, C. W.; Bloom, J. D.; Levin, J. I. *Ann. Rep. Med. Chem.* **2007**, *42*, 281.
- (5) Appel, N.; Schaller, T.; Penin, F.; Bartenschlager, R. *J. Biol. Chem.* **2006**, *281*, 9833. doi: 10.1074/jbc.R500026200
- (6) Ni, Z. J.; Wagman, A. S. *Curr. Opin. Drug Discov. Dev.* **2004**, *7*, 446.
- (7) Beaulieu, P. L.; Bos, M.; Bousquet, Y.; Fazal, G.; Gauthier, J.; Gillard, J.; Goulet, S.; LaPlante, S.; Poupart, M. A.; Lefebvre, S.; McKercher, G.; Pellerin, C.; Austel, V.; Kukolj, G. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 119. doi: 10.1016/j.bmcl.2003.10.023
- (8) Stansfield, I.; Ercolani, C.; Mackay, A.; Conte, I.; Pompei, M.; Koch, U.; Gennari, N.; Giuliano, C.; Rowley, M.; Narjes, F. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 627. doi: 10.1016/j.bmcl.2008.12.068
- (9) Louise-May, S.; Yang, W.; Nie, X.; Liu, D.; Deshpande, M. S.; Phadke, A. S.; Huang, M.; Agarwal, A. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 3905. doi: 10.1016/j.bmcl.2007.04.103
- (10) Stankiewicz-Drogon, A.; Palchykovska, L. G.; Kostina, V. G.; Alexeeva, I. V.; Shved, A. D.; Boguszevska-Chachulska, A. M. *Bioorg. Med. Chem.* **2008**, *16*, 8846. doi: 10.1016/j.bmc.2008.08.074
- (11) Bosse, T. D.; Larson, D. P.; Wagner, R.; Hutchinson, D. K.; Rockway, T. W.; Kati, W. M.; Liu, Y.; Masse, S.; Middleton, T.; Mo, H.; Montgomery, D.; Jiang, W.; Koev, G.; Kempf, D. J.; Molla, A. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 568. doi: 10.1016/j.bmcl.2007.11.088
- (12) Lü, W. J.; Chen, Y. L.; Ma, W. P.; Zhang, X. Y.; Luan, F.; Liu, M. C.; Chen, X. G.; Hu, Z. D. *Euro. J. Med. Chem.* **2008**, *43*, 569. doi: 10.1016/j.ejmech.2007.04.011
- (13) Luan, F.; Liu, H. T.; Ma, W. P.; Fan, B. T. *Euro. J. Med. Chem.* **2008**, *43*, 43. doi: 10.1016/j.ejmech.2007.03.002
- (14) Melagraki, G.; Afantitis, A.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O. *Bioorg. Med. Chem.* **2007**, *15*, 7237. doi: 10.1016/j.bmc.2007.08.036
- (15) Su, L.; Li, L.; Li, Y.; Zhang, X.; Huang, X.; Zhai, H. *Med. Chem. Res.* **2012**, *21*, 2079. doi: 10.1007/s00044-011-9734-x
- (16) de Vicente, J.; Hendricks, R. T.; Smith, D. B.; Fell, J. B.; Fischer, J.; Spencer, S. R.; Stengel, P. J.; Mohr, P.; Robinson, J. E.; Blake, J. F.; Hilgenkamp, R. K.; Yee, C.; Adjabeng, G.; Elworthy, T. R.; Li, J.; Wang, B.; Bamberg, J. T.; Harris, S. F.; Wong, A.; Leveque, V. J. P.; Najera, I.; Pogam, S. L.; Rajyaguru, S.; Ao-Ieong, G.; Alexandrova, L.; Larrabee, S.; Brandl, M.; Briggs, A.; Sukhtankar, S.; Farrell, R. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 5652. doi: 10.1016/j.bmcl.2009.08.022
- (17) Hendricks, R. T.; Spencer, S. R.; Blake, J. F.; Fell, J. B.; Fischer, J.; Stengel, P. J.; Leveque, V. J. P.; Pogam, S. L.; Rajyaguru, S.; Najera, I.; Swallow, S. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 410. doi: 10.1016/j.bmcl.2008.11.060
- (18) de Vicente, J.; Hendricks, R. T.; Smith, D. B.; Fell, J. B.; Fischer, J.; Spencer, S. R.; Stengel, P. J.; Mohr, P.; Robinson, J. E.; Blake, J. F.; Hilgenkamp, R. K.; Yee, C.; Adjabeng, G.; Elworthy, T. R.; Tracy, J.; Chin, E.; Li, J.; Wang, B.; Bamberg, J. T.; Stephenson, R.; Oshiro, C.; Harris, S. F.; Ghate, M.; Leveque, V.; Najera, I.; Pogam, S. L.; Rajyaguru, S.; Ao-Ieong, G.; Alexandrova, L.; Larrabee, S.; Brandl, M.; Briggs, A.; Sukhtankar, S.; Farrell, R.; Xu, B. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 3642. doi: 10.1016/j.bmcl.2009.05.004
- (19) Hendricks, R. T.; Fell, J. B.; Blake, J. F.; Fischer, J. P.; Robinson, J. E.; Spencer, S. R.; Stengel, P. J.; Bernacki, A. L.;

- Leveque, V. J. P.; Pogam, S. L.; Rajyaguru, S.; Najera, I.; Josey, J. A.; Harris, J. R.; Swallow, S. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 3637. doi: 10.1016/j.bmcl.2009.04.119
- (20) de Vicente, J.; Hendricks, R. T.; Smith, D. B.; Fell, J. B.; Fischer, J.; Spencer, S. R.; Stengel, P. J.; Mohr, P.; Robinson, J. E.; Blake, J. F.; Hilgenkamp, R. K.; Yee, C.; Zhao, J.; Elworthy, T. R.; Tracy, J.; Chin, E.; Li, J.; Lui, A.; Wang, B.; Oshiro, C.; Harris, S. F.; Ghate, M.; Leveque, V. J. P.; Najera, I.; Pogam, S. L.; Rajyaguru, S.; Ao-leong, G.; Alexandrova, L.; Fitch, B.; Brandl, M.; Masjedizadeh, M.; Wua, S. Y.; de Keczer, S.; Voronin, T. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 5648. doi: 10.1016/j.bmcl.2009.08.023
- (21) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: New York, 2000.
- (22) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. *J. Chem. Inform. Comp. Sci.* **2004**, *44*, 1630. doi: 10.1021/ci049869h
- (23) Tan, N. X.; Rao, H. B.; Li, Z. R.; Li, X. Y. *SAR QSAR Environ. Res.* **2009**, *20*, 27. doi: 10.1080/10629360902724085
- (24) <http://www.models.kvl.dk/source/GAPLS/index.asp>, accessed June 2008.
- (25) Leardi, R.; Boggia, R.; Terrile, M. *J. Chemom.* **1992**, *6*, 267.
- (26) Leardi, R. *J. Chemom.* **1994**, *8*, 65.
- (27) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput. Chem.* **2001**, *26*, 5. doi: 10.1016/S0097-8485(01)00094-8
- (28) Cherkassky, V.; Ma, Y. Selection of Meta-parameters for Support Vector Regression. Proceedings of the International Conference on Artificial Neural Networks, Madrid, Spain, Aug 28–30, 2002.
- (29) Hao, M.; Li, Y.; Wang, Y.; Zhang, S. *Anal. Chim. Acta* **2011**, *690*, 53. doi: 10.1016/j.aca.2011.02.004
- (30) Rainville, F. M. D.; Fortin, F. A.; Gardner, M. A.; Parizeau, M.; Gagné, C. DEAP: A Python Framework for Evolutionary Algorithms. In *EvoSoft Workshop*, Companion Proc. of the Genetic and Evolutionary Computation Conference, July 07–11, 2012.
- (31) Chang, C. C.; Lin, C. J. LIBSVM: A Library for Support Vector Machines, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, accessed Jun 2008.

基于机器学习方法的丙型肝炎病毒聚合酶 NS5B 非核苷抑制剂的定量构效关系研究

丛湧¹ 薛英^{1,2,*}

(¹ 四川大学化学学院, 教育部绿色化学与技术重点实验室, 成都 610064;
² 西华大学四川省先进科学计算重点实验室, 成都 610039)

Quantitative Structure-Activity Relationship Study of the Non-Nucleoside Inhibitors of HCV NS5B Polymerase by Machine Learning Methods

CONG Yong¹ XUE Ying^{1,2,*}

(¹ College of Chemistry, Key Laboratory of Green Chemistry and Technology, Ministry of Education, Sichuan University, Chengdu 610064, P. R. China; ² Key Laboratory of Advanced Scientific Computation, Xihua University, Chengdu 610039, P. R. China)

*Corresponding author. Email: yxue@scu.edu.cn; Tel: +86-28-85418330.

图 S1 实验与预测 pIC_{50} 数值对比图(使用线性逐步回归分析选择的描述符建立的多元线性回归模型，训练集为蓝色，测试集为红色)

Fig.S1 The experimental vs. computational $\log(10^9/IC_{50})$ (The MLR Model with the LSRA selected descriptors, Training set in blue, test set in red)

图 S2 实验与预测 pIC_{50} 数值对比图(使用线性逐步回归分析选择的描述符建立的偏最小二乘回归模型, 训练集为蓝色, 测试集为红色)

Fig.S2 The experimental vs. computational $\log(10^9/IC_{50})$ (The PLS Model with the LSRA selected descriptors, Training set in blue, test set in red)

图 S3 实验与预测 pIC_{50} 数值对比图(使用线性逐步回归分析选择的描述符建立的支持向量机回归模型, 训练集为蓝色, 测试集为红色)

Fig.S3 The experimental vs. computational $\log(10^9/IC_{50})$ (The SVM Model with the LSRA selected descriptors, Training set in blue, test set in red)

图 S4 实验与预测 pIC_{50} 数值对比图(使用遗传算法组合偏最小二乘选择的描述符建立的多元线性回归模型, 训练集为蓝色, 测试集为红色)

Fig.S4 The experimental vs. computational $\log(10^9/\text{IC}_{50})$ (The MLR Model with the GA-PLS selected descriptors, Training set in blue, test set in red)

图 S5 实验与预测 pIC_{50} 数值对比图(使用遗传算法组合偏最小二乘选择的描述符建立的偏最小二乘回归模型, 训练集为蓝色, 测试集为红色)

Fig.S5 The experimental vs. computational $\log(10^9/\text{IC}_{50})$ (The PLS Model with the GA-PLS selected descriptors, Training set in blue, test set in red)

图 S6 实验与预测 pIC_{50} 数值对比图(使用遗传算法组合偏最小二乘选择的描述符建立的支持向量机回归模型，训练集为蓝色，测试集为红色)

Fig.S6 The experimental vs. computational $\log(10^9/IC_{50})$ (The SVM Model with the GA-PLS selected descriptors, Training set in blue, test set in red)

表 S1 选用 LSRA 描述符子集建立的机器学习模型对抑制剂活性的预测结果

Table S1 The molecular structures of compounds and its actual and calculated HCV NS5B polymerase inhibitory activities with LSRA selected descriptors

Structure	Actual activity ($\log(10^9/IC_{50})$)	Calculated activity ^a (MLR)	Calculated Activity ^b (PLS)	Calculated activity ^c (SVM)
7.6580	7.7022	7.7791	7.70438
5.3470	5.4759	5.4281	5.30053
5.8540	5.5711	5.8512	5.80746
7.8540	8.1219	7.9914	7.9223

8.3010	7.9430	7.7603	7.95313
6.0610	6.7768	6.9102	6.95793
3.1660	3.3451	3.2421	3.23069
8.0000	8.4705	8.5682	8.04658

6.1370	6.3572	6.3919	6.53784
6.8300	6.1450	6.1434	6.0734
8.3010	8.0714	7.9642	8.04676
8.6990	8.1782	8.1488	8.21821

8.3010	8.2223	8.1835	8.25545
8.0000	8.2854	8.3458	8.23343
5.9590	5.6818	5.7730	5.90883
8.5230	8.5551	8.4726	8.4619
8.0460	8.4647	8.6072	8.54461
6.9750	7.7735	7.6193	7.83153

8.1550	7.7017	7.5550	7.89836
8.0000	7.3022	7.6836	7.48994
6.1340	6.1743	6.2745	6.18075
8.0970	7.9073	7.8765	7.87742

5.9590	6.0066	5.9509	5.99738
8.5230	8.1697	8.2130	8.32178
8.1550	8.3081	8.1032	8.10899
6.1430	5.8807	5.8673	5.98199
5.7450	5.7620	6.0737	5.78681

4.7100	5.3889	5.2994	5.24684
7.6020	7.4062	7.6346	7.62846
8.2220	8.6835	8.9527	8.46541
5.1610	5.3040	5.3260	5.20845
8.3980	8.2025	8.1493	8.35153

5.6580	5.5492	5.3470	5.23552
5.8240	5.9426	5.8249	5.99914
8.0000	8.2988	8.3202	8.04641
7.9210	7.8768	7.7255	7.96753
8.3010	8.0475	8.0118	8.21603

7.9210	7.6668	7.5626	7.86223
8.5230	8.0273	7.9625	8.24101
7.3870	7.5992	7.5917	7.79642
5.8860	5.5276	5.5109	5.77535

6.1590	6.2986	6.2800	6.25803
8.1550	7.9823	8.0046	8.20206
5.4200	5.7693	5.9645	6.30609
3.0610	2.9308	2.9964	3.10778

7.9590 ^d	7.7023	7.7741	7.70662
6.0000 ^d	5.3828	5.3053	5.28991
5.5690 ^d	5.4823	5.7337	5.78478
7.7960 ^d	8.1834	7.9403	7.99413

8.0000 ^d	8.0143	7.7631	8.03723
6.9870 ^d	7.6750	7.7798	7.67369
3.0000 ^d	3.5662	3.5470	3.50181
8.0000 ^d	7.9101	7.9486	8.15456

6.3870 ^d	6.4865	6.3740	6.57326
5.9210 ^d	6.2754	6.2952	6.23989
7.8860 ^d	8.0614	7.8163	8.02931
8.0460 ^d	8.1396	8.3098	8.14373

7.8860 ^d	8.0073	8.0002	8.03582
7.7700 ^d	7.7677	7.5303	7.96263
5.6200 ^d	5.7599	5.7304	5.93064
9.0000 ^d	8.3928	8.4004	8.35496
8.5230 ^d	8.5240	8.5344	8.42245

7.9590 ^d	8.5267	8.6704	8.2359
6.4560 ^d	6.2293	6.1333	6.37448
8.0000 ^d	7.9966	7.8436	7.93941
6.3980 ^d	6.1719	6.0321	6.15552

6.1430 ^d	5.8880	6.0841	5.83493
5.6380 ^d	5.5491	5.6768	5.78548
8.3980 ^d	7.9823	7.9435	8.26345
7.9210 ^d	7.9551	7.7148	7.9774
6.2150 ^d	5.9702	6.2450	6.25515

6.2600 ^d	5.6561	5.7043	5.59333
6.4710 ^d	6.1684	6.4654	6.36583
7.1670 ^d	7.9469	7.7895	8.00165
5.4090 ^d	5.6081	5.2704	5.32562

5.6580 ^d	5.7898	5.5921	5.49657
7.7450 ^d	7.5892	7.4189	7.82692
6.0810 ^d	5.7202	6.0727	5.89906
5.9210 ^d	6.3131	6.5973	6.53446

6.2760 ^d	6.3826	6.3433	6.56164
8.3980 ^d	8.3335	8.5072	8.30763
8.2220 ^d	8.1033	8.1674	8.14673
8.0970 ^d	7.7336	7.7920	8.11641

8.3980 ^d	8.3042	8.4641	8.28773
8.3010 ^d	7.6646	7.9031	8.04976
6.2370 ^d	5.8323	5.7143	5.57263
7.1610 ^d	6.5758	6.7039	6.81786

8.0970 ^d	7.7062	7.8299	8.09136
6.1550 ^d	6.1128	6.0687	6.12209

^a Predicted $\log(10^9/\text{IC}_{50})$ by Multiple Linear Regression. ^b Predicted $\log(10^9/\text{IC}_{50})$ by Partial Least Square. ^c Predicted $-\log(\text{IC}_{50})$ by Support Vector Machine. ^d Compounds in the testing set.

表 S2 选用 GA-PLS 描述符子集建立的机器学习模型对抑制剂活性的预测结果

Table S2 The molecular structures of compounds and its actual and calculated HCV NS5B polymerase inhibitory activities with GA-PLS selected descriptors

Structure	Actual activity ($\log(10^9/IC_{50})$)	Calculated activity ^a (MLR)	Calculated Activity ^b (PLS)	Calculated activity ^c (SVM)
7.6580	7.8743	7.9098	7.95188
5.3470	5.3728	5.3492	5.32284
5.8540	5.7573	5.7700	5.83432
7.8540	8.1044	8.1071	8.07864

8.3010	7.8917	7.9285	7.97115
6.0610	6.7663	6.6908	6.8189
3.1660	3.2476	3.2045	3.39128
8.0000	8.2579	8.3596	8.22528

6.1370	6.3560	6.2635	6.32661
6.8300	6.0919	6.0993	6.17818
8.3010	8.0797	8.1520	8.04664
8.6990	8.2243	8.2056	8.22366

	8.3010	8.2203	8.1891	8.21462
	8.0000	8.4261	8.3691	8.31075
	5.9590	5.8100	5.8289	5.88778
	8.5230	8.3944	8.3360	8.2978
	8.0460	8.4157	8.3994	8.32326
	6.9750	7.9066	7.9464	7.98873

8.1550	7.6935	7.7616	7.86199
8.0000	7.4105	7.2483	7.37549
6.1340	6.1990	6.1745	6.25855
8.0970	7.9151	7.9296	7.97869

5.9590	5.8356	5.8263	5.89146
8.5230	8.1427	8.1291	8.17498
8.1550	8.1806	8.1528	8.13157
6.1430	5.8251	5.8421	5.90044
5.7450	6.0567	5.9702	5.97013

4.7100	5.3903	5.4245	5.4111
7.6020	7.4082	7.2511	7.37688
8.2220	8.6964	8.6175	8.44411
5.1610	5.3778	5.3908	5.3862
8.3980	8.1266	8.1272	8.17265

5.6580	5.3670	5.3871	5.38175
5.8240	5.8101	5.8532	5.90513
8.0000	8.3926	8.3935	8.22537
7.9210	8.0459	8.0796	8.14158
8.3010	8.1305	8.1387	8.17791

7.9210	7.6919	7.7520	7.85424
8.5230	8.0105	8.0205	8.10607
7.3870	7.6983	7.7473	7.85248
5.8860	5.6009	5.6606	5.69918

6.1590	6.2163	6.1950	6.27748
8.1550	8.0087	8.0113	8.1008
5.4200	5.9866	6.0441	6.13207
3.0610	2.9953	3.0326	3.24172

7.9590 ^d	7.8738	7.9094	7.95143
6.0000 ^d	5.3724	5.3489	5.3226
5.5690 ^d	5.7612	5.7727	5.8371
7.7960 ^d	8.1041	8.1069	8.07842

8.0000 ^d	7.9253	7.9526	7.99856
6.9870 ^d	7.8712	7.9077	7.94936
3.0000 ^d	4.0459	3.4699	3.68881
8.0000 ^d	8.0152	8.0374	8.11579

6.3870 ^d	6.3549	6.2618	6.32531
5.9210 ^d	6.2150	6.1879	6.272
7.8860 ^d	7.8935	7.9340	7.97579
8.0460 ^d	8.2355	8.2039	8.22444

7.8860 ^d	8.0390	8.0606	8.10439
7.7700 ^d	7.7051	7.7764	7.87646
5.6200 ^d	5.8105	5.8292	5.88812
9.0000 ^d	8.3921	8.3334	8.29671
8.5230 ^d	8.4092	8.3370	8.29898

7.9590 ^d	8.6228	8.6098	8.41831
6.4560 ^d	4.9903	6.7661	6.34099
8.0000 ^d	7.8841	7.9135	7.95763
6.3980 ^d	6.0133	6.0250	6.09379

6.1430 ^d	6.0692	5.9723	5.97364
5.6380 ^d	5.6162	5.6498	5.69472
8.3980 ^d	7.9608	7.9815	8.06708
7.9210 ^d	7.8902	7.9372	7.97721
6.2150 ^d	6.1174	6.0390	6.0952

6.2600 ^d	5.7604	5.7391	5.72648
6.4710 ^d	6.2319	6.1982	6.26916
7.1670 ^d	7.9536	7.9761	8.02313
5.4090 ^d	5.4209	5.4769	5.45191

5.6580 ^d	5.5932	5.5510	5.53355
7.7450 ^d	7.7008	7.7684	7.86921
6.0810 ^d	5.9481	5.9469	6.02256
5.9210 ^d	6.3524	6.2910	6.35466

6.2760 ^d	6.3620	6.2820	6.33978
8.3980 ^d	8.4670	8.4457	8.34399
8.2220 ^d	8.1665	8.1794	8.20094
8.0970 ^d	7.8047	7.8403	7.97745

8.3980 ^d	8.4470	8.4694	8.35321
8.3010 ^d	7.8604	7.8626	7.99844
6.2370 ^d	5.7368	5.6832	5.68466
7.1610 ^d	5.4338	7.1728	6.81275

8.0970 ^d	7.8189	7.8420	7.98077
6.1550 ^d	6.0066	6.0123	6.08313

^a Predicted $\log(10^9/\text{IC}_{50})$ by Multiple Linear Regression. ^b Predicted $\log(10^9/\text{IC}_{50})$ by Partial Least Square. ^c Predicted $-\log(\text{IC}_{50})$ by Support Vector Machine. ^d Compounds in the testing set.

References:

- (1) de Vicente, J.; Hendricks, R. T.; Smith, D. B.; Fell, J. B.; Fischer, J.; Spencer, S. R.; Stengel, P. J.; Mohr, P.; Robinson, J. E.; Blake, J. F.; Hilgenkamp, R. K.; Yee, C.; Adjabeng, G.; Elworthy, T. R.; Li, J.; Wang, B.; Bamberg, J. T.; Harris, S. F.; Wong, A.; Leveque, V. J.; Najera, I.; Le Pogam, S.; Rajyaguru, S.; Ao-Ieong, G.; Alexandrova, L.; Larrabee, S.; Brandl, M.; Briggs, A.; Sukhtankar, S.; Farrell, R. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 5652.
- (2) Hendricks, R. T.; Spencer, S. R.; Blake, J. F.; Fell, J. B.; Fischer, J.; Stengel, P. J.; Leveque, V. J.; Le Pogam, S.; Rajyaguru, S.; Najera, I.; Swallow, S. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 410.
- (3) de Vicente, J.; Hendricks, R. T.; Smith, D. B.; Fell, J. B.; Fischer, J.; Spencer, S. R.; Stengel, P. J.; Mohr, P.; Robinson, J. E.; Blake, J. F.; Hilgenkamp, R. K.; Yee, C.; Adjabeng, G.; Elworthy, T. R.; Tracy, J.; Chin, E.; Li, J.; Wang, B.; Bamberg, J. T.; Stephenson, R.; Oshiro, C.; Harris, S. F.; Ghate, M.; Leveque, V.; Najera, I.; Le Pogam, S.; Rajyaguru, S.; Ao-Ieong, G.; Alexandrova, L.; Larrabee, S.; Brandl, M.; Briggs, A.; Sukhtankar, S.; Farrell, R.; Xu, B. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 3642.
- (4) Hendricks, R. T.; Fell, J. B.; Blake, J. F.; Fischer, J. P.; Robinson, J. E.; Spencer, S. R.; Stengel, P. J.; Bernacki, A. L.; Leveque, V. J.; Le Pogam, S.; Rajyaguru, S.; Najera, I.; Josey, J. A.; Harris, J. R.; Swallow, S. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 3637.
- (5) de Vicente, J.; Hendricks, R. T.; Smith, D. B.; Fell, J. B.; Fischer, J.; Spencer, S. R.; Stengel, P. J.; Mohr, P.; Robinson, J. E.; Blake, J. F.; Hilgenkamp, R. K.; Yee, C.; Zhao, J.; Elworthy, T. R.;

Tracy, J.; Chin, E.; Li, J.; Lui, A.; Wang, B.; Oshiro, C.; Harris, S. F.; Ghate, M.; Leveque, V. J.; Najera, I.; Le Pogam, S.; Rajyaguru, S.; Ao-Ieong, G.; Alexandrova, L.; Fitch, B.; Brandl, M.; Masjedizadeh, M.; Wu, S. Y.; de Keczer, S.; Voronin, T. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 5648.