

基于序列的 G 蛋白偶联受体-药物相互作用预测研究

丁林松¹ 郑宇杰²

(南京理工大学计算机科学与工程学院 南京 210094)¹

(中国电子科技集团公司第28研究所 南京 210007)²

摘要 准确预测 G 蛋白质偶联受体(GPCR)是否与药物(Drug)相互作用是新药开发的关键步骤之一。从时间和费用方面来说,通过生物实验的方法来确定 GPCR-Drug 是否相互作用的代价是昂贵的。因此,直接从蛋白质序列出发预测 GPCR-Drug 的相互作用具有重要的意义。提出了一种基于序列的 GPCR-Drug 相互作用预测方法:从蛋白质序列抽取进化信息特征;对药物抽取指纹特征;基于上述两种特征,使用基于证据理论的 K 近邻算法进行分类预测。在标准数据集上的实验结果表明了所述方法的有效性。

关键词 G 蛋白质偶联受体,药物,特征抽取,预测

中图分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.8.016

Research on Sequence-based Predictor for GPCR-Drug Interaction Prediction

DING Lin-song¹ ZHENG Yu-jie²

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)¹

(The 28th Research Institute, China Electronics Technology Group Corporation, Nanjing 210007, China)²

Abstract Accurately identifying whether a G-protein-coupled receptor(GPCR) will interact with a drug is a crucial step in drug discovery. However, experimentally determining the interactions between GPCR and drug is time-consuming and expensive. Hence, developing automated prediction methods for GPCT-Drug interaction prediction solely from protein sequence is in urgent need. In this study, a new sequence-based method for GPCT-Drug interaction prediction was proposed. Evolutionary information feature from protein sequence and footprint feature from drug were combined to form discriminative feature. And the optimized evidence-theoretic K-nearest neighbor(OET-KNN) prediction algorithm was taken as classifier. Experimental results demonstrate that the proposed method achieves good performance and can act as complementary predictor to existing methods.

Keywords G-protein-coupled receptor, Drug, Feature extraction, Prediction

1 引言

G 蛋白偶联受体(G Protein-Coupled Receptors, GPCRs)是一大类膜蛋白受体的统称,因能结合 G 蛋白质并调节其活性而得名。

与 G 蛋白偶联受体相关的 G 蛋白在受体信号传递中至少有以下 4 种作用:(1)G 蛋白在偶联情况下直接介导细胞间信号转导;(2)通过控制受体的位置和自身传递功能,调节受体信号;(3)作为桥梁,将受体与不同的效应器蛋白进行物理连接;(4)作为受体构象的一个变构调节器,可以调节受体药理或者其他方面的功能^[1]。在当前最畅销的上市药物中,近一半直接或间接通过 GPCR 发挥药效^[2]。G 蛋白偶联受体已经被证明是临床广泛应用的重要药物靶点,作用于 G 蛋白偶联受体的药物对疼痛认知障碍、高血压、胃溃疡、鼻炎、哮喘等各类疾病均具有良好的治疗作用,为此很多学术机构和制药厂家已经投入了大量的努力来研究 G 蛋白偶联受体。

在新药开发的过程中,新的药物靶标的识别是药物开发

最为关键的一步^[3]。为了分析和预测蛋白质-药物相互作用,已经研究出了一些预测方法。其中最常用的几种方法有对接模拟^[4]、文献文本挖掘^[5]、结合化学结构、基因序列和 3D 结构信息^[6]等。在生物信息学领域中,机器学习和数据挖掘方法已经得到了广泛的应用。许多研究人员已经做了很多努力,并开发出了高效的方法和软件来研究与药物相关的各种生物问题,比如 HIV 蛋白酶切割位点预测、GPCR(G 蛋白偶联受体)类型的识别、蛋白质信号肽预测、蛋白质亚细胞定位预测、蛋白酶类型鉴定、膜蛋白类型预测以及一系列相关蛋白质属性预测方法^[7]。

本文提出了一种基于序列的 GPCR-Drug 相互作用预测方法:从蛋白质序列抽取进化信息特征;对药物抽取指纹特征;基于上述两种特征,使用基于证据理论的 K 近邻算法进行分类预测^[8]。

2 数据来源和研究方法

2.1 数据集

数据集 S 表示为

到稿日期:2014-06-01 返修日期:2014-08-25 本文受国家自然科学基金(61373062),江苏省自然科学基金(BK20141403)资助。

丁林松(1989—),男,硕士生,主要研究方向为生物信息学,E-mail:dlsynbmj@163.com.cn;郑宇杰(1976—),男,博士生,主要研究方向为特征提取、模式识别。

$$S = S^+ \cup S^- \quad (1)$$

其中, S^+ 是包含彼此发生相互作用的 GPCR-Drug 对的正样本集, S^- 是彼此没有发生相互作用的 GPCR-Drug 对的负样本集, 符号 $\cup$ 表示集合理论中的并集。正如 KEGG (<http://kegg.jp/kegg/>) 数据库里所定义的, GPCR-Drug 彼此发生相互作用指的是, GPCR-Drug 对中 G 蛋白偶联受体和药物两个部分在药物靶位中彼此相互作用。为了客观地评价本文所提方法, 采用文献[8]的数据集: 正样本集 S^+ 包含了 620 个彼此发生相互作用的 GPCR-Drug 对; 负样本集 S^- 包含了 1240 个彼此没有发生相互作用的 GPCR-Drug 对。

2.2 特征提取

2.2.1 药物的分子指纹提取

本研究采用文献[8]中所述的二维药物分子指纹编码来表示药物分子的特征。首先, 对于数据集中每个药物分子都可以通过它的编码从 KEGG 数据库中获得一个对应的 MOL 文件, 该文件中包含了药物的化学结构信息; 接着, 用化学工具软件 OpenBabel^[10] 将 MOL 文件转换成 2D 的分子指纹文件格式。OpenBabel 软件可以从 <http://openbabel.org/> 网站下载。目前, OpenBabel 软件的最新版本可以生成 4 种类型指纹数据: FP2、FP3、FP4 和 MACCS, 本研究采用 FP2 指纹类型, 它依据药物中所有线性和环形结构来识别小分子片段, 并且用哈希函数将它们映射成一个位字符串。该字符串就是所谓的药物指纹。FP2 编码是长度为 256 位十六进制字符串或 256 位的矢量, 其分量的值可以用 0 到 15 的整数表示。假设 V_1 是 256 位矢量的第一个矢量, V_2 为第二个矢量等。因此 256 位矢量可以转化为一个数字信号。为了能够发现药物分子指纹数值的本质属性, 采用离散傅里叶变换将其转化为频域值^[8]:

$$X_k = \sum_{m=1}^{256} V_m \exp[-j(\frac{2\pi m}{256}k)], k=1, 2, \dots, 256 \quad (2)$$

其中, j 为虚数单位, X_k 是一个复数, 它的模可以由下式计算得到:

$$A_k = [\text{real}(X_k) + \text{imag}(X_k)]^{\frac{1}{2}} \quad (3)$$

其中, $\text{real}(X_k)$ 为 X_k 的实数部分, $\text{imag}(X_k)$ 为 X_k 的虚数部分。至此, 可以得到其傅里叶频谱如下:

$$\{A_1, A_2, \dots, A_k\} \quad (4)$$

傅里叶频谱数据包含了大量关于数字信号的信息, 因此可以用来反映一个药物分子的某种特征。这样, 一个药物分子的特征可以用一个 256 维的向量 d 来进行表示:

$$d = [A_1 \ A_2 \ \dots \ A_j \ \dots \ A_{256}]^T \quad (5)$$

其中, A_j 在式(4)中已定义, T 是矩阵的转置操作。

2.2.2 GPCR 序列 PSSM 特征提取

位置特异性得分矩阵 (Position Specific Scoring Matrix, PSSM) 已经被证明是蛋白质结构与功能预测的有效特征来源^[11]。对于一个含有 L 个氨基酸残基的蛋白质序列 P , 首先使用 PSI-BLAST 程序得到一个对应的 L 行、20 列 PSSM 矩阵; 然后, 使用式(6)将 PSSM 矩阵中的每个元素规范化到 (0, 1) 区间:

$$f(x) = \frac{1}{1 + \exp(-x)} \quad (6)$$

规范化后的 PSSM 矩阵记为 $P_{psm} = (p_{i,j})_{L \times 20}$ 。

一个蛋白质的 PSSM 组成成分特征可以用一个 20 维的特征向量来表示:

$$F_{psm} = (\bar{p}_1, \bar{p}_2, \dots, \bar{p}_{20}) \quad (7)$$

其中:

$$\bar{p}_j = \frac{1}{L} \sum_{i=1}^L p_{i,j} \quad (8)$$

F_{psm} 中的 $\bar{p}_j$ 表示在演化过程中, 蛋白质序列中所有的氨基酸残基突变成第 j 种氨基酸类型的平均得分。虽然 F_{psm} 可以反映蛋白质的部分进化信息, 但是关于蛋白质中氨基酸之间的顺序信息完全丢失了。为了弥补这一缺陷, 可以对 PSSM 做自适应协方差变换 (ACT)^[12], 来提取隐藏在其中的序列顺序信息。更具体地说, 按照下式定义一个标量 θ_j^g :

$$\theta_j^g = \frac{1}{L-g} \sum_{i=1}^{L-g} (p_{i,j} - \bar{p}_j) \cdot (p_{i+g,j} - \bar{p}_j) \quad (9)$$

θ_j^g 表示的是在第 j ($1 \leq j \leq 20$) 列上的 g 阶相关因子; 然后, PSSM 的 g 阶 ACT 特征可以用下式来表示:

$$\theta^g = (\theta_1^g, \theta_2^g, \dots, \theta_{20}^g)^T \quad (10)$$

假设 G 为 g 所能取的最大值, 亦即: $g=1, 2, \dots, G$ 。那么蛋白质的 PSSM_ACT 特征就可以由式(7)和 G 个 ACT 特征向量组合而得到:

$$X_{PSSM_ACT}^G = \begin{bmatrix} F_{psm} \\ \theta^1 \\ \vdots \\ \theta^g \\ \vdots \\ \theta^G \end{bmatrix} \quad (11)$$

本研究中, G 设定为 6。因此一条蛋白质序列的 PSSM_ACT 特征向量的维数为 $20 + G \times 20 = 140$ 。

2.2.3 G 蛋白偶联受体-药物对的特征表示

至此, 一个 GPCR-Drug 的特征可以用一个 396 维的向量表示:

$$f = \begin{pmatrix} X_{PSSM_ACT}^G \\ w \cdot d \end{pmatrix} \quad (12)$$

其中, w 为权重。实验结果表明, 当权重取值为 0.3 时, 能够得到最优结果。

2.3 基于证据理论的 K 近邻算法

基于证据理论的 K 近邻算法 (OET-KNN) 在很多生物信息问题中得到了应用。本文中也使用 OET-KNN 作为分类器。下面简述 OET-KNN 算法:

首先, 定义类别:

$$\varphi = \begin{cases} 1, & \text{相互作用} \\ 0, & \text{不相互作用} \end{cases} \quad (13)$$

用 $S(N) = \{(f_1, \varphi_1), (f_2, \varphi_2), \dots, (f_N, \varphi_N)\}$ 表示一个具有 N 个 GPCR-Drug 对的训练数据集, 其中 f_i 表示第 i 个样本的特征向量 (参见式(12)), $\varphi_i \in \{0, 1\}$ 是第 i 个样本的类别。

对于一个测试 GPCR-Drug 对, 假设其特征可使用 f_{query} 表示。

首先, 可以从训练样本集 $S(N)$ 中找到该测试样本的 K 个近邻, 不妨用 S_k 表示。那么, 对于任一 $f_j \in S_k$, 如果它的类别是 φ , 就相当于有一个证据来支持 f_{query} 的类别也属于 φ , 并且, 该证据的信度可以使用下式来表示^[13]:

$$E(f_{query} | f_j, \varphi) = 0.95 \exp[-C_\varphi \cdot D(f_{query}, f_j)] \cdot \delta(\varphi_j, \varphi) \quad (14)$$

其中, $D(f_{query}, f_j)$ 表示的是 f_{query} 和 f_j 之间的距离, 参数 C_φ 是与类别 φ 相关的权重, 并且:

$$\delta(\varphi_i, \varphi) = \begin{cases} 1, & \text{if } \varphi_i = \varphi \\ 0, & \text{if } \varphi_i \neq \varphi \end{cases} \quad (15)$$

本研究使用参考文献[8]中的优化方法对 C_φ 进行优化处理, 得出 $C_\varphi: C_1=10.1424$ 和 $C_0=8.7296$ 。

那么, f_{query} 属于类别 φ 的总的信度则由 S_k 中 K 个样本来决定, 可以表示为:

$$E(f_{query}, \varphi) = 1 - \prod_{j=1}^K (1 - E(f_{query} | f_j, \varphi)) \quad (16)$$

下面进一步将信度归一化:

$$E(f_{query}, \varphi) = \frac{E(f_{query}, \varphi)}{E(f_{query}, \varphi=1) + E(f_{query}, \varphi=0)} \quad (17)$$

归一化后的信度就作为测试样本 f_{query} 属于类别 φ 的倾向性。由于本文主要是判定一个测试样本中 GPCR 和 Drug 是否相互作用, 因此

$$E(f_{query}, \varphi=1) \quad (18)$$

就是测试样本 f_{query} 为相互作用类别的倾向性。定义一个阈值 T , 当 $E(f_{query}, \varphi=1) > T$ 时, 就判定 f_{query} 是相互作用的。阈值 T 的确定将在第 3 节中进一步阐述。

3 实验结果和讨论

3.1 评价指标

使用特异性 (Spe)、灵敏度 (Sen)、总体预测准确度 (Acc) 和马氏相关系数 (MCC) 来评估所提出的方法的性能。这 4 个指标定义如下:

$$Spe = \frac{TN}{TN + TP} \quad (19)$$

$$Sen = \frac{TP}{TP + FN} \quad (20)$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (21)$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (22)$$

其中, FP 、 TP 、 FN 、 TN 分别为假阳性值、真阳性值、假阴性值、真阴性值。

使用 ROC (Receiver Operating Characteristic) 和 AUC (Area under Curve) 来评估预测模型的综合预测性能。ROC 曲线是用来反映灵敏度和特异性连续变量的综合指标。通过绘制 ROC 曲线揭示敏感性和特异性的相互关系, 能够直观地衡量预测结果的好坏。AUC 则是体现一个预测模型的综合预测性能; AUC 的值越大, 表明预测模型的总体预测性能越高。

3.2 交叉验证

本研究使用留一法来验证所提方法的有效性; 假设数据集有 N 个样本, 每次取出一个样本作为测试样本, 剩余的 $N-1$ 个样本为训练样本; 重复 N 次, 直到每个样本均被测试过; 依据 N 次测试的结果, 可以结算得到上述评估指标。

3.3 实验结果

正如 2.3 小节所述, 阈值 T 的选取对于模型的预测精度至关重要。为此, 使用留一法计算出每个样本属于相互作用类别的倾向性, 然后, 选取一个阈值 T , 使得预测结果的 MCC

最大。实验结果表明, 当 $T=0.35$ 时, 预测的 MCC 达到最大, 也就是 0.6218。表 1 列出了 $T=0.35$ 时, Sen 、 Spe 和 Acc 的值, 以及与其他方法的对比。

表 1 GPCR-Drug 预测的准确率

Performance evaluation	本文方法	文献[14]方法
Sen	71.77%	N/A
Spe	89.27%	N/A
Acc	83.44%	78.49%
MCC	62.18%	N/A

图 1 给出了相应的 ROC 曲线。从表 1 中可以看出, 数据集 S 整体的准确率可以达到 83.44%。相比之下, 在参考文献[14]中 He 等人提出的结合 6 种生物化学和物理化学特性来表示 GPCR-Drug 对的准确率仅仅是 78.49%。表 1 中的对比结果表明, 采用二维的分子指纹来表示药物特征, 采用 PSSM_ACT 来表示 GPCR 的特征, 对于 G 蛋白偶联受体-药物相互作用预测是有效的。

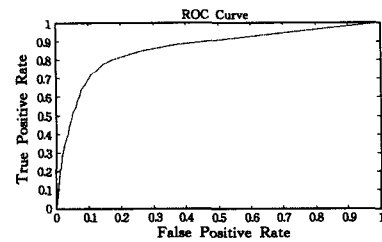


图 1 ROC 曲线

结束语 本文提出了一种从序列直接出发预测 GPCR 与 Drug 相互作用的预测方法; 使用傅里叶变换从 Drug 的指纹抽取其鉴别特征; 使用 PSI-Blast 从 GPCR 序列获得位置特异性得分矩阵 PSSM, 并在此基础上进一步抽取 PSSM_ACT 特征; 使用基于证据理论的 K 近邻算法进行分类预测。在标准数据集上的实验结果表明, 本文所述方法是有效的, 亦是对现有方法的有益补充。

参考文献

- [1] Tilley D G. G protein-dependent and G protein-independent signaling pathways and their impact on cardiac function [J]. Circ Res, 2011, 109(2): 217-230
- [2] Alexander S P, Mathie A, Peters J A. Guide to Receptors and Channels (GRAC) [J]. Br J Pharmacol, 2011, 164(suppl. 1): S1-324
- [3] Knowles J, Gromo G. A guide to drug discovery; Target selection in drug discovery [J]. Nat Rev Drug Discov, 2003, 2(1): 63-69
- [4] Chou K C. Review: Structural bioinformatics and its impact to biomedical science [J]. Current Medicinal Chemistry, 2004, 11(16): 2105-2134
- [5] OZhu S, Okuno Y, Tsujimoto G, et al. A probabilistic model for mining implicit 'chemical compound-gene' relations from literature [J]. Bioinformatics, 2005, 21(2): 245-251
- [6] Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces [J]. Bioinformatics, 2008, 24(13): 232-240

(下转第 111 页)

稳定节点以及“搭便车”节点与其他节点在网络中地位的差异性。角色越重要的节点在计算信任值时影响越大,角色越重要的节点允许下载量就越大。仿真表明,该模型可以有效地区分网络中的稳定节点和不稳定节点,有效地延长稳定节点在网络中的地位,同时减少“搭便车”节点以及伪善节点的不良影响,提高交易成功率。由于 MP2P 网络多是混合 P2P 的拓扑结构,且存在物理拓扑结构和逻辑拓扑不符、拓扑结构多变的问题,后续的主要研究方向是 MP2P 的安全拓扑结构。

参 考 文 献

- [1] 李致远,王汝传.一种移动 P2P 网络环境下的动态安全信任模型[J].电子学报,2012,40(1):1-7
Li Zhi-yuan, Wang Ru-chuan. A Dynamic Secure Trust Model for Mobile P2P Networks [J]. Chinese Journal of Electronics, 2012,40(1):1-7
- [2] 陈世平,王佳炳,裘慧奇.一种移动环境中的 P2P 网络信任模型[J].计算机应用,2009,29(10):2603-2606
Chen Shi-ping, Wang Jia-bing, Qiu Hui-qi. Trust Model for mobile P2P network [J]. Journal of Computer Application, 2009, 29(10):2603-2606
- [3] 任艳,任平安,吴振强,等.移动 P2P 网络中的多粒度信任模型[J].计算机工程与应用,2009,45(6):137-140
Ren Yan, Ren Ping-an, Wu Zhen-qiang, et al. Multiple granularity trust model in mobile P2P network [J]. Computer Engineering and Applications, 2009,45(6):137-140
- [4] 曹晓梅,朱海涛,沈何阳,等. MP2P 网下一种基于代理的安全反馈信任模型[J].计算机科学,2014,41(7):200-205
Cao Xiao-mei, Zhu Hai-tao, Shen He-yang, et al. Proxy-based Security-Feedback Trust Model in MP2P Network [J]. Computer Science, 2014,41(7):200-205
- [5] Zuo Cui-hua, Zhou Jian-fang, Feng Hong-cai. A Novel Multi-level Trust Model to Improve the Security of P2P Networks[C]//2010 3rd IEEE International Conference Computer Science and Information Technology (ICCSIT). Chengdu, IEEE, 2010: 100-104
- [6] Gao Ji-xun, Guo Yun-hong. Research on P2P Repeated Game Trust Model Based on Unknown Node Type[C]//2012 Interna-

tional Conference on Computer Science & Service System (CSSS). Nanjing: IEEE, 2012: 655-658

- [7] Hu Jian-lin, Wu Quan-yuan, Zhou Bin. FCTrust: A Robust and Efficient Feedback Credibility-Based Distributed P2P Trust Model[C]//The 9th International Conference for Young Computer Scientists(ICYC 2008). Hunan: IEEE, 2008: 1963-1968
- [8] Li Zhi-yuan, Bi Jun-lei. An Adaptive Trusted Tequest and Authorization Model for Mobile Peer-to-Peer Networks[C]//2013 IEEE 10th International Conference. Zhang Jiajie: IEEE, 2013: 1274-1280
- [9] Wang Jun-she, Li Xiao-long, Zhang Xun. Research of P2P Network Trust Model [C]//2013 Fifth International Conference on Intelligent Human-Machine Systems and Cybernetics. Hangzhou: IEEE, 2013: 70-73
- [10] Li Xiong, Liu Ling. PeerTrust: Supporting Reputation-based Trust for Peer-to-Peer Electronic Communities[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(7): 843-857
- [11] Yang Li, Zhang Yu-se, Xing Chang-you, et al. A Node Interest Similarity based P2P Trust Model[C]//ICCT. Nanjing: IEEE, 2010: 572-575
- [12] Xi Jing, Nie Peng-peng, Lu Jian-de. A P2P Trust Model Based on Dynamic Fuzzy Theory—BGTR [C] // 2008 International Symposium on Information Processing (ISIP). Moscow: IEEE, 2008: 187-191
- [13] Hu Jian-lin, Li Xiao-hua, Zhou Bin, et al. SECtrust: A Secure and Effective Distributed P2P Trust Model[C]//2010 Third International Symposium on Intelligent Information Technology and Security Informatics(IITSI). Jinggangshan: IEEE, 2010: 34-38
- [14] 安晓江,李大兴.基于角色的代理模型的实现[J].计算机工程, 2004,30(10):62-64
An Xiao-jiang, Li Da-xing. Realization of Role-based Dlegation Model [J]. Computer Engineering, 2004,30(10):62-64
- [15] 张华平,刘群.基于角色标注的中国人名自动识别研究[J].计算机学报,2004,27(1):85-91
Zhang Hua-ping, Liu Qun. Automatic Recognition of Chinese Personal Name Based on Role Tagging [J]. Chinese Journal of Computers, 2004, 27(1): 85-91

(上接第 77 页)

- [7] Chou K C, Shen H B. Review: recent advances in developing Web-servers for predicting protein attributes[J]. Natural Science, 2009,1(2):63-92
- [8] Zouhal L M, Denoeux T. An evidence-theoretic KNN rule with parameter optimization[J]. IEEE Transactions on Systems, Man and Cybernetics, 1998,28(2):263-271
- [9] Xuan Xiao, Jian-Liang Min, Pu Wang, et al. iGPCR-Drug: A Web Server for Predicting Interaction between GPCRs and Drugs in Cellular Networking [J]. PLoS ONE, 2013, 8(8): e72234
- [10] O'Boyle N M, Banck M, James C A, et al. Open Babel: An open chemical toolbox[J]. J Cheminform, 2011, 3(33)

- [11] Schaffer A A. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements[J]. Nucleic Acids Res, 2011, 29(14): 2994-3005
- [12] Liu T, et al. Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles[J]. Amino Acids, 2012, 42(6): 2243-2249
- [13] Denoeux T. A k-nearest neighbor classification rule based on Dempster-Shafer theory[J]. IEEE Transactions on Systems Man and Cybernetics, 1995, 25(5): 804-813
- [14] He Z, Zhang J, Shi XH, et al. Predicting drug-target interaction networks based on functional groups and biological features[J]. PLoS ONE, 2010, 5(3): e9603