基于 SVM 的药物靶点预测方法及其应用 *

尚振伟 李 晋 姜永帅 张明明 吕洪超 张瑞杰△

(哈尔滨医科大学生物信息科学与技术学院 黑龙江 哈尔滨 150000)

摘要 目的：基于已知药物靶点和潜在药物靶点蛋白的一级结构相似性，结合 SVM 技术研究新的有效的药物靶点预测方法。方法：构造训练样本集，提取蛋白质序列的一级结构特征，进行数据预处理，选择最优核函数，优化参数并进行特征选择，训练最优预测模型，检验模型的预测效果。以 G 蛋白偶联受体家族的蛋白质为预测集，应用建立的最优分类模型对其进行潜在药物靶点挖掘。结果：基于 SVM 所建立的最优分类模型预测的平均准确率为 81.03%。应用最优分类器对构造的 G 蛋白预测集进行预测，结果发现预测排位在前 20 的蛋白质中有多个与疾病相关。特别的，其中有两个 G 蛋白在治疗靶点数据库(TTD)中显示已作为临床试验的药物靶点。结论：基于 SVM 和蛋白质序列特征的药物靶点预测方法是有效的，应用该方法预测出的潜在药物靶点能够为发现新的药靶提供参考。

关键词：支持向量机(SVM)；药物靶点预测；G 蛋白偶联受体(GPCR)

中图分类号：Q-33，R91 **文献标识码**：A **文章编号**：1673-6273(2012)20-3943-04

A Method of Drug Target Prediction Based on SVM and its Application*

SHANG Zhen-wei, LI Jin, JIANG Yong-shuai, ZHANG Ming-ming, LV Hong-chao, ZHANG Rui-jie△

(College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150000, China)

ABSTRACT Objective: Combined with the SVM technology to study a new and effective method of drug target prediction based on the protein primary structure similarity of known drug targets and potential drug targets. Methods: To construct the training set, extract the primary structure characteristics of protein sequences, and preprocess the data, then select the optimal kernel function and parameters, finally do feature selection, and train the best forecasting model, test its effect. Apply the optimal prediction model on the prediction set that composed of GPCRS for potential drug targets mining. Results: The average prediction accuracy of the optimal classification model that based on SVM is 81.03%. Applying the optimal classification model on the prediction set of G-proteins, then we find that some of the proteins that ranking top 20 of the prediction results are related with some certain diseases. Particularly, there are two G-proteins seem as clinical trials drug targets in the therapeutic target database (TTD). Conclusions: This drug target prediction method based on SVM and protein sequence features is effective, and the application of this method to predict potential drug targets can provide a valuable reference for the discovery of new drug targets.

Key words: Support vector machine (SVM); Prediction of drug target; G protein coupled receptor (GPCR)

Chinese Library Classification(CLC): Q-33，R91 Document code: A

Article ID:1673-6273(2012)20-3943-04

前言

药物分子大多通过与人体内靶标分子的相互作用产生疗效，因此靶点选择是药物研发中十分关键的一个步骤。新的药物作用靶点一旦被发现，往往成为一系列新药发现的突破口，所以寻找药物作用靶点成为当今创新药物研究激烈竞争的焦点。随着生物信息学的发展，药物靶点预测成为可能，这不仅大大缩短了药物研发的时间，减少了药物研发的费用也降低了在研究早期阶段由于靶点错误定位给新药研发带来损失的可能性。

支持向量机(SVM)是由 Vapnik 领导的 AT&TBell 实验室研究小组在 1963 年提出的一种基于统计学习理论的模式识别方法，主要应用于模式识别领域，是一种非常有潜力的分类技术[1]。在 20 世纪 90 年代的中后期 SVM 技术得到了全面深入的发展，现在已经成为机器学习和数据挖掘领域的重要工具，被广泛的应用在生物信息学，文本和手写识别等领域[2]。由于药物作用靶点蛋白与非药靶蛋白性质上存在着差异[3]，而药物作用靶点之间具有特征相似性，因此可以通过 SVM 对潜在的药物靶点进行挖掘，为新药的研发提供前期参考依据。

目前，药物已经成功研发出来的最常见的药物靶点蛋白包括蛋白酶、激酶、G 蛋白偶联受体(GPCR)、核激素受体等。在药物开发过程中，GPCR 和酶是蛋白质中的主要药物靶点。本文提出了一种基于 SVM 的药物靶点预测方法，并应用该方法预测了潜在的 G 蛋白药靶，发现预测结果中确实有一些 G 蛋白

* 基金项目：国家自然科学基金(G81172842,G61170154)

作者简介：尚振伟(1983-)，女，硕士，助教，主要研究方向：药物靶点挖掘. E-mail: dabaiweiwei@163.com

△通讯作者：张瑞杰，电话：0451-86620941. E-mail: zhangruijie2002@yahoo.com.cn

(收稿日期：2012-02-10 接受日期：2012-03-05)

参与了某些疾病的病理生理学过程，并且有一些蛋白已被研究作为某种疾病的临床试验药物靶点，说明该药靶预测方法是有效的，其预测结果能够为新药的研发提供有价值的参考。

1 材料与方法

1.1 数据收集和预处理

首先从 Drugbank 数据库[4](DrugBank Version 2.5)中得到 1309 个人类药物靶点蛋白。然后从 Pfam 数据库中得到的 9627 个人类蛋白家族中去除 1309 个药物靶点蛋白以及它们所处在的家族，共 2136 个家族，在余下的蛋白家族中我们共得到 24347 个非药物靶点蛋白。由于目前没有研究证明某个蛋白不是药物靶点，所以值得注意的是在非药靶集合中的蛋白不能证明它一定是非药物靶点，而只是目前现有的知识还没有发现药物靶点和药物靶点家族出现在该数据集中。实际上有可能在其中发现新颖的药物靶点或靶点家族，但是相对于其他蛋白家族中来说出现的几率较低。从 UniProt KB/Swiss-Prot(56.0)数据库提取上述蛋白质的序列信息。

1.1.1 提取蛋白质特征数据 为了获取蛋白质序列的初级特征，我们应用了一个在线的计算蛋白质结构和物理化学特征的软件—PROFEAT[5]。利用该在线软件由蛋白质的序列信息我们可以提取出蛋白质的 1497 个特征值，这些特征属于 7 个特征组。

1.1.2 特征数据的标准化 由于特征种类的不同，要处理的特征值处在不同的数值范围。在代数函数中，大的特征比小特征更具有影响力，但在 SVM 中特征数值的大小并不能反映它们具有的重要程度。因此，我们运用均值和方差的估计值对特征数据做归一化使不同特征数值处于相似的范围。

1.2 构造样本集

在这里我们采用如下的方式构建训练集合和检验集合。

训练样本集 将 1309 个药物靶点作为阳性样本，由于有研究表明当训练集采用 1:1 的比例时，可以避免大样本的偏倚性[2]，从而得到相对较好的预测结果。因此可从 24347 个非药物靶点蛋白中随机抽取 1309 个蛋白作为阴性样本与 1309 个阳性样本构成一个训练集合。考虑到如此抽取的阴性样本数量只占非靶点蛋白集合总数目的 1/20。为了解决这个问题，使得预测结果更加可靠，我们从 24347 个阴性样本当中随机抽样 100 次，得到 100 个阴性样本集，将它们分别与 1309 个阳性样本组合，产生 100 个训练集合。这样构成的 100 个训练集合得到的预测结果必然会有所不同，通过综合考虑各个预测结果，可以避免由于阴性样本的不均衡性而使得预测结果不稳定的缺陷。

检验样本集 对于每一个训练样本，我们都采用了十倍交叉证实，所以每个训练样本都被分为 10 份，其中每一份都要作为检验集合。

1.3 建立最优 SVM 分类预测模型

在这里我们应用 2.89 版的 libsvm 软件建立分类预测模型。

1.3.1 核函数和最优参数的选择 SVM 的关键在于核函数，采用不同的核函数将导致不同的 SVM 算法[1]。对 100 个训练样本我们分别用 SVM 中的 linear kernel，polynomial kernel and radial basis function(RBF)以及 sigmoid 四种核函数来进行训练，通过比较四种核函数在默认参数下的分类效能，找出一个最优的核函数。然后优化该核函数的 c，g 两个参数使 SVM 分类效能达到最大。

对 100 个训练样本，每次训练我们都采用十倍交叉证实，将各核函数下 100 个训练样本十倍交叉验证的准确率 (Qa)取均值作为评价相应核函数分类效能的指标。

$$Qa = (TP + TN)/(TP + TN + FP + FN)$$

其中 TP, TN, FP 和 FN 分别代表真阳性，真阴性，假阳性和假阴性。在四种核函数及默认参数下准确率 Qa 的均值如表 1 所示。

表 1 四种核函数分类准确率均值表

Table1 Average classification accuracy of four kinds of nuclear function

Nuclear function	Linear kernel	Polynomial kernel	Radial basis function (RBF)	Sigmoid
Average classification Accuracy	72.47%	58.17%	79.40%	74.57%

可见对于这 100 个训练样本 RBF 核函数具有最优的分类效果。接下来对 RBF 核函数的参数 c 和 g 进行优化。在 RBF 核函数及其最优参数下，100 个训练样本的十倍交叉证实分类准确率均值为 79.88%与参数优化前相比分类器的分类效能有所提高。

1.3.2 特征选择 研究表明一些噪音特征能使得 SVM 平面更为复杂，从而降低分类效能。为了得到一个理想的 SVM 模型，从而提高其分类效能，我们有必要对特征进行筛选。在这里我们用秩和检验去掉特征集中不显著的特征。具体做法是对每一个训练集进行一次秩和检验，显著性水平为 0.05，比较 100 个检验结果，将特征集中不显著的次数在 50 次以上的特征去掉(P 值大于 0.05 的)。

利用特征选择后的 100 个训练样本对 SVM 进行训练，由于训练样本不同，训练得到的 SVM 模型也随之不同，由此可得到 100 个最优的 SVM 分类模型。

2 预测结果分析

通过特征选择共去掉 402 个特征。在 RBF 核函数及优化参数下特征选择之前，即基于原始特征集十倍交叉验证得到的分类准确率均值为 79.88%，特征选择后预测的平均准确率提高到了 81.03%。利用得到的最优分类模型对同一个靶点进行预测，将得到 100 个不同的预测结果，预测结果返回被预测蛋白距离超平面的距离，且当预测结果是正值时，表示预测该蛋白质是药物靶点，否则当预测结果是负值时，表示预测该蛋白

质不是药物靶点。将得到的 100 个不同的预测结果加和取平均值作为综合评价预测结果的测度。

3 G 蛋白药物靶点预测应用

从 GPCRDB 数据库上能得到 632 个人类 G 蛋白。其中有 96 个 G 蛋白是已知的药物靶点，将其从这 632 个人类 G 蛋白中去除，对余下的 536 个 G 蛋白做特征提取和相应的数据预处理，将其作为预测集合，用我们训练好的 100 个最优 SVM 分类器对其进行分类预测，在这里我们对预测集中各蛋白的 100 次预测结果分别计算其中每个蛋白到相应的最优超平面的距离之和，记为 D，把该值作为综合评价预测结果的测度。根据支持向量机的原理，距离超平面距离最大的样本，被正确分类的可靠性越大，因此依据 D 值将预测集中的蛋白按照由大到小的顺序排列，排在前面的蛋白具有更大的潜力成为候选药物靶点。排在前 20 位的 G 蛋白如表 2 所示。

表 2 G 蛋白药物靶点预测结果

Table 2 Prediction results of GPCR drug targets

Rank	D value	Swiss-prot ID	Protein name	Gene encoding	Swiss-prot ID of known drug targets with the same gene encoding
1	139.2367	A8K1R9	cDNA FLJ78161	similar(GRM2)	Q14416
2	137.7233	O15303	Metabotropic glutamate receptor 6	GRM6	
3	137.3721	A8K0F9	cDNA FLJ78562	similar(GRM4)	Q14833
4	130.7463	Q96RG9	M3 muscarinic cholinergic receptor	CHRM3	P20309
5	130.3831	A8K2D2	cDNA FLJ75348	similar(GRM8)	O00222
6	129.7981	A8K5P7	cDNA FLJ75523	GRM5	P41594
7	122.3493	B0ZBD3	Adrenergic, alpha-1A-, receptor variant 3	ADRA1A	P35348
8	119.6493	B0UXY8	Gamma-aminobutyric acid (GABA) B receptor, 1	GABBR1	Q9UBS5
9	113.7434	Q712M9	Serotonin receptor 5-HT4	HTR4	Q13639
10	113.1264	A1L441	G protein-coupled receptor 128	GPR128	
11	112.7781	P49190	Parathyroid hormone 2 receptor	PTH2R	
12	111.1167	Q5CZ57	Prostaglandin E receptor 3 (Subtype EP3), isoform CRA_i	EP3-I	
13	109.2679	Q53EM0	Gamma-aminobutyric acid (GABA) B receptor 1 isoform b variant	GABBR1	Q9UBS5
14	107.803	Q59HC2	Glutamate receptor, metabotropic 1 variant	GRM1	Q13255
15	107.254	Q59EH9	Dopamine receptor D2 isoform long variant	DRD2	P14416
16	107.2304	P49146	Neuropeptide Y receptor type 2	NPY2R	
17	106.932	Q59FW2	G protein-coupled receptor 63 variant	GPCR63	
18	106.5431	Q05AH0	Uncharacterized protein	FSHR	P23945
19	106.201	B1ALU3	Latrophilin 2	LPHN2	
20	103.3057	Q7Z5R9	HRH2 protein	HRH2	P25021

表 2 中排在第 11 位的甲状旁腺激素受体 2 (PTH2R)在大脑和胰腺中表达丰富，甲状旁腺激素(PTH)可能通过 PTH2R 在多个生理系统中发挥着影响。PTH2R 对胰腺的功能具有重要作用，而该受体蛋白在神经细胞中的存在表明了它可能作为神经递质受体。在治疗靶点数据库中(TTD Version 1.0.11)，我们发现该受体蛋白已经被作为临床试验药物靶点，用于治疗绝经，骨质疏松症和牛皮癣。排在第 16 位的神经肽 Y2 受体(NPY2R)，有多篇文献指出 NPY2R 与肥胖有关联[6-9]，也有文献指出 NPY2R 基因可能是男性 2 型糖尿病的候选基因[10]，更有文献明确指出 NPY2R 是神经元母细胞瘤的治疗靶点[11]。另外我们发现该受体是 TTD 上的临床试验药物靶点，用于治疗肥胖症。

表 2 中排在第 2 位的谷氨酸受体 6，有文献指出其编码基因 GRM6 变异会导致先天性静止性夜盲症[12-14]，并且在高度近视的患者身上发现了 GRM6 基因的 3 个变体，这表明 GRM6 变异可能引发高度近视[15]。排在表 2 中第 12 位的前列腺 E 受体 3，这种受体可能有许多生物功能，其中涉及消化，神经系统，肾脏重吸收，子宫收缩活动，有报道 EP3 基因是提高化疗敏感性和抑制抗化疗乳腺癌和肿瘤转移的重要靶[16-18]，也有文献指出 EP3 为动脉粥样硬化疾病的一个新的靶[19]。

排在第 1，3，5 位上的 G 蛋白所对应的编码基因尚不明确，但根据 NCBI 上的数据，它们对应的编码基因有极大可能分别是基因 GRM2,GRM4,GRM8，其真实性有待进一步的证实。但是我们发现其可能对应的基因均为已知药物靶点的编码

基因。表 2 中列出了排在前 20 位的 G 蛋白中与已知药物靶点蛋白具有相同编码基因或者近似相同编码基因的 G 蛋白，共有 13 个，由于由相同基因编码的蛋白质其序列具有很大的相似性，导致提取的特征具有较大相似性，所以通过 SVM 分类器预测必然排位靠前，这与我们得到的结果一致。另外当蛋白质具有相似的结构时，其功能可能也是相近的。因此表 2 中所给出的此类 G 蛋白是值得进行研究的。特别是可以作为与其具有相同编码基因的已知药物靶点的替代靶点，研究新药的开发。

4 讨论

本研究基于已知药物靶点和潜在药物靶点蛋白的一级结构相似性，利用 SVM 技术建立最优分类模型，并以 GPCR 家族的蛋白质为预测集，应用我们训练的分类器对其进行潜在药物靶点预测，结果发现排位在前 20 的蛋白质中有多个蛋白质与疾病相关。特别的，其中有两个 G 蛋白在治疗靶点数据库中(TTD)显示已作为临床试验的药物靶点。另外有 13 个蛋白质与已知的药物靶点蛋白具有相同的编码基因。这说明我们采用的基于支持向量机和蛋白质序列特征的药物靶点预测方法用于新药物靶点预测是有效的，应用该方法预测出的潜在药物靶点能够为发现新的药靶提供参考。

参考文献 (References)

[1] George P., Ré dei. Support Vector Machine (SVM) [M]. Encyclopedia of Gen etics, Genomics, Proteomics and Inf ormatics, 2008, Part 19: 4020-6754

[2] Han L, Cui J, Lin H, et al. Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity [J]. Proteomics, 2006, (7)6:4023-4037

[3] Li Q and Lai L.Prediction of potential drug targets based on simple sequence properties [J]. BMC Bioinformatics, 2007, 8:353-354

[4] Wishart DS, Knox C, Guo AC, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets [J]. Nucleic Acids Res, 2008, 1 (36):D901-906

[5] Bakheet TM and Doig AJ. Properties and identification of human protein drug targets [J]. Bioinformatics, 2009, 25:451-457

[6] Friedlander Y, Li G, Fornage M, et al. Candidate molecular pathway genes related to appetite regulatory neural network, adipocyte homeostasis and obesity: results from the CARDIA Study [J]. Ann Hum Genet, 2010, 74(5):387-398

[7] Takiguchi E, Fukano C, Kimura Y, et al. Variation in the 5'-flanking region of the neuropeptide Y2 receptor gene and metabolic parameters [J]. Metabolism, 2010, 59(11):1591-1596

[8] Zhang J, Wang HJ, Ma J, et al. Association between obesity and the polymorphism of neuropeptide Y2 receptor gene in children and adolescents [J]. Zhonghua Liu Xing Bing Xue Za Zhi, 2009, 30(7):695-698

[9] Elbers CC, de Kovel CG, van der Schouw YT, et al. Variants in neuropeptide Y receptor 1 and 5 are associated with nutrient-specific food intake and are under recent selection in Europeans [J]. PLoS One,2009, 4(9): 70-76

[10] Campbell CD, Lyon HN, Nemesh J, et al. Association studies of BMI and type 2 diabetes in the neuropeptide Y pathway: a possible role for NPY2R as a candidate gene for type 2 diabetes in men [J]. Diabetes, 2007, 56(5):1460-1467

[11] Lu C, Everhart L, Tilan J, et al. Neuropeptide Y and its Y2 receptor: potential targets in neuroblastoma therapy [J]. Oncogene, 2010, 29 (41):5630-5642

[12] O'Connor E, Allen LE, Bradshaw K, et al. Congenital stationary night blindness associated with mutations in GRM6 encoding glutamate receptor MGluR6 [J]. Br J Ophthalmo, 2006, 90(5):653-654

[13] Dryja TP, McGee TL, Berson EL, et al. Night blindness and abnormal cone electroretinogram ON responses in patients with mutations in the GRM6 gene encoding mGluR6 [J]. Proc Natl Acad Sci U S A, 2005, 102(13):4884-4889

[14] Zeitz C, Forster U, Neidhardt J, et al. Night blindness-associated mutations in the ligand-binding, cysteine-rich, and intracellular domains of the metabotropic glutamate receptor 6 abolish protein trafficking [J]. Hum Mutat, 2007, 28(8):771-780

[15] Xu X, Li S, Xiao X, et al. Sequence variations of GRM6 in patients with high myopia [J]. Mol Vis, 2009, 19(15):2094-2100

[16] Kang JH, Song KH, Jeong KC, et al. Involvement of Cox-2 in the metastatic potential of chemotherapy-resistant breast cancer cells [J]. BMC Cancer, 2011, 4:327-334

[17] Robertson FM, Simeone AM, Lucci A, et al. Differential regulation of the aggressive phenotype of inflammatory breast cancer cells by prostanoid receptors EP3 and EP4 [J]. Cancer, 2010, 116:2806-2814

[18] Amano H, Ito Y, Suzuki T, et al. Roles of a prostaglandin E-type receptor, EP3, in upregulation of matrix metalloproteinase-9 and vascular endothelial growth factor during enhancement of tumor metastasis [J]. Cancer Sci, 2009, 100(12):2318-2324

[19] Heptinstall S, Espinosa DI, Manolopoulos P, et al. DG-041 inhibits the EP3 prostanoid receptor-a new target for inhibition of platelet function in atherothrombotic disease [J]. Platelets, 2008, 19 (8): 605-613