# 基于集成学习的离子通道药物靶点预测

谢倩倩[1] 李订芳[1] 章 文[2]
(武汉大学数学与统计学院 武汉 430072)[1] (武汉大学深圳研究院 深圳 518057)[2]

**摘 要** 新药研制成功的关键在于药物靶点的发现和准确定位。在已知的药物靶点中,离子通道蛋白是一类广受欢迎的靶点,它与免疫系统、心血管等疾病密切相关。对于靶点的发现,传统生物方法成本高、耗时久。因此,探讨了基于机器学习的离子通道蛋白药物靶点的挖掘,以加快药物靶点发现过程,节约经费。由于药物靶点相关序列的长度不一致,考虑了蛋白质序列编码的13种特征,它们能将不等长的蛋白质序列转化成等长序列。通过数值实验筛选能够较好地区分靶点和非靶点的特征子集,并采用集成学习的方法整合特征得到预测模型。通过与已有工作的比较表明,提出的集成模型能得到较高的准确率,具有很好的应用前景。

**关键词** 离子通道,随机森林,药物靶点,分类器,集成学习
**中图法分类号** TP3-05 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.4.035

## Predicting Potential Drug Targets from Ion Channel Proteins Based on Ensemble Learning

XIE Qian-qian[1] LI Ding-fang[1] ZHANG Wen[2]
(School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China)[1]
(Research Institute of Shenzhen, Wuhan University, Shenzhen 518057, China)[2]

**Abstract** The identification of molecular targets is a critical step in the discovery and development process of new drugs. Among large known drug targets, ion channel proteins are the most attractive drug targets, which are closely linked to some diseases such as cardiovascular and central nervous systems. Traditional biological methods have the characteristics of high-cost and time-consuming in mining drug targets. Our work discussed the mining of potential ion channel drug targets based on random forests, which is aimed at speeding up the discovery process of drug targets and saving money. Since the lengths of sequences related to drug targets are diverse, thirteen types of protein encoding features were considered which can transform the protein sequences with distinct lengths into the sequences with same lengths in our study. A feature subset which has better performance in the division between drug targets and non-targets was chosen by numerical experiments and the ensemble learning was introduced to attain prediction models. Our study attains high accuracy by comparison to the developed methods, which plays the critical roles in the mining of new drug targets.

**Keywords** Ion channel, Random forests, Drug targets, Classifiers, Ensemble learning

## 1 引言

众所周知,新药研制成功的关键在于药物靶点的发现和准确定位[1]。好的药物不仅高效而且副作用小,其在很大程度上得益于药物靶点的恰当选取[2]。近年来,很多人致力于药物的研究及其发展,但是只有很少的药物靶点应用于临床[3]。因此,发掘更多的潜在药物靶点对于药物的设计和发现是至关重要的。

大多数的药物靶点集中在4类蛋白质家族:酶、G蛋白偶联受体、离子通道、核受体[3,4]。其中,离子通道是一个多样化的跨膜蛋白质家族,穿过离子通道的离子流为膜兴奋性、信号转导和神经传递提供了条件,因此离子通道蛋白在神经系统、心血管等的功能中起着不可替代的作用。另外,很多研究者[5]发现,离子通道蛋白是一类广受欢迎的药物靶点,而且与此相关的疾病很多,最典型的就是甲型流感[6-8]。M2质子通道是治愈禽流感的关键靶点,以其作为靶点的金刚烷胺和金刚乙胺等药物很早就用来对抗甲型流感[7,9]。因此,更深入地研究离子通道,挖掘其潜在的药物靶点,对于新药的研发有着重要的意义。

在药物靶点的研究中,很多研究者已经做出了巨大努力。文献[2]结合蛋白质的基本序列、二级结构和亚细胞定位等3种特征,利用SVM来预测离子通道中潜在的药物靶点。部

到稿日期:2014-04-13 返修日期:2014-07-23 本文受国家自然科学基金(61271337,61103126),教育部博士点基金(20100141120049),湖北省自然科学基金(2011CDB454),深圳市战略新兴产业发展专项资金项目(JCYJ20130401160028781)资助。
谢倩倩(1988—),女,硕士生,主要研究方向为机器学习、生物信息;李订芳(1966—),男,博士,教授,主要研究方向为计算流体力学、科学与工程计算软件、计算机应用;章 文(1981—),男,博士,副教授,主要研究方向为数据挖掘、机器学习、生物信息,E-mail:zhangwen@whu.edu.cn(通信作者)。

分研究者[10,11]基于序列同源性和结构域来分析已知的药物靶点，并将其应用于挖掘新靶点，也有研究者[12-14]基于蛋白质的3D结构来研究可以与类药物的化合物结合的绑定区域。Monica Campillos等[15]基于副作用相似性预测潜在的药物靶点。Li Qingliang等[1]利用蛋白质的几种简单序列特性，通过建立SVM模型来预测潜在的药物靶点。另外，Wang Yinying等[16]基于蛋白质域预测药物靶点。还有很多研究者也对药物靶点的挖掘做出了贡献[1,17-21]。

目前，用于挖掘新的药物靶点的生物方法不仅需要高额的科研经费，而且比较耗时。而数值计算方法可以方便快捷地处理一些大数据，将其应用到生命科学领域，快速高效地分析这些数据的生物学意义，对于挖掘潜在的药物靶点有着很大的帮助。针对离子通道药物靶点的预测，本文考察了13种可以将不等长的蛋白质序列转化为等长序列的特征，通过随机森林的方法建立分类器，筛选了能够较好区分靶点和非靶点序列的特征子集，并运用集成学习方法得到高精度的集成预测模型。

## 2 方法

### 2.1 数据集

数据集1采用Drugbank数据库[22]中已验证的1515个人类药物靶点作为正样本集。对于负样本集的构造，由于还没有专门的非药物靶点数据库，本文采用文献[23]中的人类非药物靶点作为负样本集。由于文献[23]采用的负样本集中的部分样本在后期的研究中被验证为正样本，本文将这部分样本剔除，最终得到了3204个负样本，将其作为本文数据集1的负样本集。数据集2的构造是建立在离子通道数据库中329个离子通道蛋白(http://www.ionchannels.org/database.php)的基础上的。其中，有145个离子通道蛋白是Drugbank数据库中已验证的药物靶点，将其作为数据集2的正样本集，另外16个非药物靶点作为负样本集。两个数据集的详细信息如表1所列。数据集1基于全基因组靶点蛋白进行预测，通过该数据集建立的分类模型能够实现全基因组范围内对离子通道蛋白潜在药物靶点的预测；数据集2基于所有的离子通道蛋白进行预测，通过该数据集建立的分类模型可实现在离子通道蛋白范围内有针对性地预测其潜在的药物靶点。这两个数据集之间的关系如图1所示。

表1 实验中的两个数据集

| 样本类别 \ 数据集 | 数据集1 | 数据集2 |
|---|---|---|
| 正样本数 | 1515 | 145 |
| 负样本数 | 3402 | 16 |

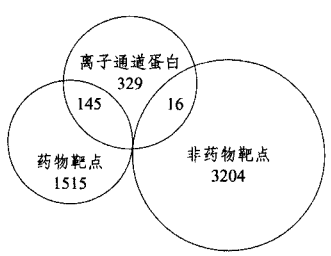



图1 人类临床验证的药物靶点蛋白、筛选得到的非药物靶点和离子通道蛋白三者之间的关系

### 2.2 蛋白质特征

蛋白质(protein)是生命的物质基础，没有蛋白质就没有生命。在本文的研究中，由于不同的蛋白质序列具有不同的长度，传统的机器学习方法无法处理这种不等长的序列，因此需要先对这些蛋白质序列进行处理[24]，将其转化成等长序列。在蛋白质的众多特征中，文中提取了以下的13种能够将不等长的蛋白质序列转化成等长序列的特征来评估蛋白质的特性[25]：氨基酸组成(Amino Acid Composition，AAC)、二肽组成(Dipeptide Composition，DC)、成分(Composition，CTDC)、转化(Transition，CTDT)、分布(Distribution，CTDD)、规范化Moreau-Broto自相关系数(Normalized Moreau-Broto Autocorrelation，MoreauBroto)、Moran自相关系数(Moran Autocorrelation，Moran)、Geary自相关系数(Geary Autocorrelation，Geary)、联合三元组描述符(Conjoint Triad，CTriad)、序列耦合数量(Sequence-order-coupling Number，SOCN)、准序列描述符(Quasi-sequence-order Descriptors，QSO)、伪氨基酸组成(Pseudo-Amino Acid Composition，PAAC)、两性分子伪氨基酸组成(Amphiphilic Pseudo-Amino Acid Composition，APAAC)，其具体维度如表2所列。

表2 13种蛋白质特征的维度

| 特征 | AAC | DC | CTDC | CTDT | CTDD | MoreauBroto | Moran |
|---|---|---|---|---|---|---|---|
| 维度 | 20 | 400 | 21 | 21 | 105 | 40 | 40 |
| 特征 | Geary | CTraid | SOCN | QSO | PAAC | APAAC | |
| 维度 | 40 | 343 | 10 | 50 | 25 | 30 | |

### 2.3 基于随机森林和重抽样的模型构建

本文使用的两个数据集中正负样本的数量相差很大，为了避免样本不均衡导致的偏差，在随机森林分类器的构建过程中采用重抽样的方式建立几个均衡数据集。文中采取的重抽样方法是有放回地随机抽取样本。本文在构建随机森林分类器过程中采用了五折交叉验证，重抽样方法应用在每一折交叉验证中训练集的构建上。对于数据集1，五折交叉验证的每一折验证中，4/5的正样本和4/5的负样本共同构成训练集，此时正样本集个数为303个(1515 * 4/5)，负样本集个数为2721或是2722个(3402 * 4/5)。每次从该负样本集中随机抽取303个负样本组成新的负样本集，与此时的正样本集构建一个子分类器。对于数据集2，五折交叉验证的每一折验证中，训练集中的正样本个数为116个(145 * 4/5)，负样本个数为12或是13个(16 * 4/5)。每次从116个正样本中随机抽取12或是13个正样本组成新的正样本集(正样本的个数与此时负样本的个数保持一致)，与负样本集构建一个子分类器。重抽样产生的多个子分类器构成不均衡样本分类的基础分类器，并用于集成学习。

### 2.4 集成学习

集成学习是指在对新的样本进行分类时，把若干个单个分类器集成起来，通过对多个分类器的分类结果进行某种组合来决定最终的分类，以取得比单个分类器更好的性能。本文综合考虑了蛋白质的13种特征，并筛选了其中若干种较好的特征。若直接结合这些特征的特征向量进行训练预测，由于维度、冗余性较高，可能会影响最终模型的精度，因此引入集成学习的思想。针对蛋白质的每一种筛选特征，分别构建一个随机森林的基础分类器，而在每个基础分类器的内部，采用重抽样的方法构建多个子分类器，子分类器结果的集成值

作为每个基础分类器的结果。使用筛选得到的特征子集进行集成学习的最终结果是所有基础分类器结果的平均值。其流程如图 2 所示。

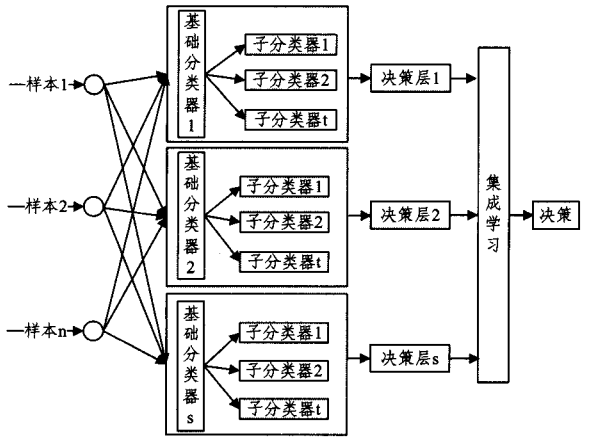



图 2　流程

## 3　实验与分析

### 3.1　实验设置

无论对于数据集 1 还是数据集 2，本文均采用五折交叉验证的方法进行预测，将数据集随机划分为 5 份，每次以其中的 4 份作为训练集，余下的 1 份作为测试集。对于五折交叉验证的每一次验证，在训练集的内部采用重抽样方法构建 10 个不同的子分类器，10 个子分类器的集成结果作为测试集的结果。另外，本文采用 ROC 值作为分类器好坏的评判标准。通过不断变化分类的阈值，可以得到一条表示错分为正类样本的比重与错分为负类样本的比重两者之间的变化曲线（即 ROC 曲线），这些比重越小表明分类器的分类性能越好。虽然 ROC 曲线很好地反映了分类器的性能表现，但是却无法准确地比较两条 ROC 曲线，因此选择将其量化，采用 ROC 曲线下的面积（即 AUC），AUC 值越大表明分类性能越好，文中涉及到的 ROC 值均是指 AUC 值。由 13 个特征建立的 13 个随机森林分类器的分类表现各异，根据它们的 ROC 值，选取较高的 ROC 值对应的特征子集再进行集成学习，以求得到更好的分类表现。

### 3.2　实验结果

#### 3.2.1　两个数据集的结果

对于数据集 1，当使用全部的 13 个特征分别进行随机森林分类器的训练时，其对应的 ROC 如表 3 所列。从表 3 可以看出，单个特征进行训练时，所有特征的 ROC 值均在0.700以上；除了 CTraid 和 SOCN 这两种特征，其余的 11 种特征的 ROC 值均在 0.7200 以上，且最大的 ROC 值（也仅有 0.7876）是由特征 PAAC 进行训练取得的。同样地，数据集 2 中 13 个特征的 ROC 值如表 4 所列，仅有 4 种特征的 ROC 值在 0.7000 以上，且 MoreauBroto、CTDC 和 Geary 这 3 种特征的 ROC 值均在 0.7300 以上，其中最大的 ROC 值为 0.7603。此时默认的分类阈值均为 0.500。

表 3　数据集 1 中 13 个特征的 ROC 值

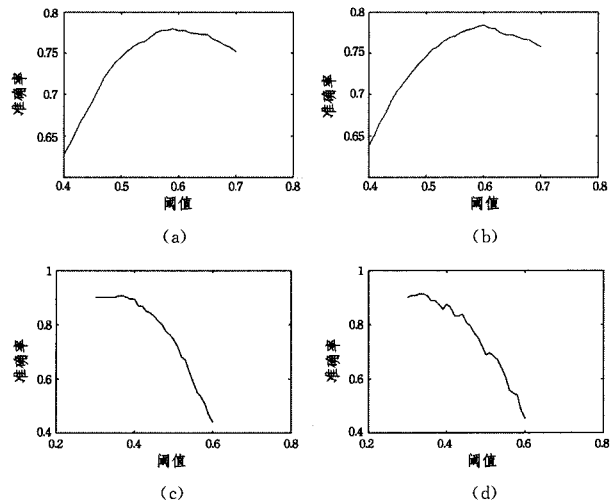
| 特征 | AAC | DC | CTDC | CTDT | CTDD | MoreauBroto | Moran |
|---|---|---|---|---|---|---|---|
| ROC | 0.7670 | 0.7746 | 0.7525 | 0.7426 | 0.7280 | 0.7400 | 0.7296 |
| 特征 | Geary | CTraid | SOCN | QSO | PAAC | APAAC | |
| ROC | 0.7281 | 0.7191 | 0.7070 | 0.7758 | 0.7876 | 0.7793 | |

表 4　数据集 2 中 13 个特征的 ROC 值

| 特征 | AAC | DC | CTDC | CTDT | CTDD | MoreauBroto | Moran |
|---|---|---|---|---|---|---|---|
| ROC | 0.5782 | 0.6862 | 0.7466 | 0.6787 | 0.6403 | 0.7603 | 0.7188 |
| 特征 | Geary | CTraid | SOCN | QSO | PAAC | APAAC | |
| ROC | 0.7321 | 0.5950 | 0.6431 | 0.5922 | 0.6388 | 0.6019 | |

#### 3.2.2　集成学习的结果

对数据集 1 的 13 个基础分类器得到的结果进行集成学习，得到的 ROC 值为 0.83155，明显高于单个分类器的 ROC 值（单个分类器的最大 ROC 值为 0.7876，分类阈值默认为 0.500），但其准确率仅有 0.7455。图 3(a)显示了阈值与准确率之间的变化趋势，从图中可以看出，准确率最大可以达到 0.7796（此时阈值为 0.59）。此结果表明，使用全部的特征进行预测并不一定能得到最好的预测效果，因此考虑去除部分贡献较小的特征。表 3 显示了 13 种特征分别对应的 ROC 值，选取具有较大 ROC 值的几个特征进行集成。通过多次实验表明，当选取 ROC 值较高的前 11 个特征构成的特征子集进行集成学习，即选取 PAAC、APAAC、QSO、DC、ACC、CTDC、CTDT、MoreauBroto、Moran、Geary 和 CTDD 时，能够得到最大的 ROC 值 0.83413，与之前使用全部特征进行预测相比，其更为高效，且得到了最大的准确率0.7841（分类阈值为 0.60），高于使用全部特征进行预测的准确率。此时，阈值与其相应准确率之间的关系如图 3(b)所示。




(a)和(b)针对数据集 1，其中(a)是 13 个特征用于预测，(b)是 11 个特征用于预测；(c)和(d)针对数据集 2，其中(c)是 13 个特征用于预测，(d)是 3 个特征用于预测

图 3

类似地，对于数据集 2，使用全部的 13 个基础分类器进行集成学习时，其 ROC 值仅为 0.7373，准确率仅有 0.7453（阈值默认为 0.500）。图 3(c)反映了阈值与准确率之间的变化关系，从图中可以看出，最大的准确率为 0.9068（阈值为 0.36），虽然已经达到较高的准确率，但是其相应的 ROC 值比较小，低于单个基础分类器的最大 ROC 值，此结果说明所有基础分类器的简单集成并不一定能产生更好的结果，因此考虑剔除部分产生负影响力的特征，使其得到最好的结果。表 4 显示了数据集 2 的 13 种特征对应的 ROC 值。通过多次实验测试，当选取前 3 个具有较高 ROC 值的特征（MoeauBroto、CTDC、Geary）构建的特征子集进行集成学习时，可以得到最大的 ROC 值 0.8192，且能得到最大的准确率 0.9130（阈值为 0.33），高于全部特征用于集成学习的准确率。图 3(d)反映

了使用3种特征构成的特征子集进行集成学习时的阈值与准确率之间的关系。

3.2.3 与已有方法的比较

涂白等[26]采用支持向量机方法(SVM)预测离子通道蛋白,同时也采用了另外的两种方法(线性判别分析和InterPro与GO映射规则,简记为LDA和InterPro2Go)进行预测。涂白等采用留一法交叉验证后的敏感度分别为95.9%(SVM)、92.3%(LDA)、87.0%(InterPro2Go),特异度分别为98.3%(SVM)、97.2%(LDA)、97.4%(InterPro2Go);本文中的数据集2采用五折交叉验证后的敏感度和特异度分别为91.19%和1,其特异度远高于SVM、LDA和InterPro2Go,数据集1采用五折交叉验证后的敏感度和特异度略低于涂白等人的结果。文献[2]曾对离子通道蛋白的潜在药物靶点进行预测,它采用支持向量机建立分类模型,分别使用了3种不同的核函数(线性核函数、多项式核函数、径向内积核函数)训练分类器。文献[2]同样使用了两个不同的数据集进行预测,一个是基于全基因组的靶点预测(数据集1),一个是基于离子通道蛋白的预测(数据集2)。对于数据集1,文献[2]采用十折交叉验证后的准确率分别为:0.8513(径向内积核函数)、0.7693(线性核函数)、0.7691(多项式核函数),而本文中的准确率0.7841仅次于以径向内积核函数进行训练的分类器。对于数据集2,文献[2]采用五折交叉验证后的准确率为:0.8380(径向内积核函数)、0.8667(线性核函数)、0.9381(多项式核函数),而本文中的准确率0.9130仅次于以多项式核函数作为训练的分类器。另外,文献[2]在采用支持向量机建立分类模型的过程中进行了相应的参数优化,而且本文与文献[2]中所采用的数据集也有一定的差异,其原因在于文献[2]的负样本集中的部分样本在后期研究中不断被验证为正样本。因此,考虑到以上几点,本文的研究对于挖掘潜在的药物靶点具有一定的应用价值。

**结束语** 本文综合考虑蛋白质的13种特征,将不等长的蛋白质序列转化成等长序列,建立随机森林分类器,并应用集成学习的思想构建高精度的分类模型,预测离子通道蛋白中潜在的药物靶点。实验结果表明,采取部分特征进行集成学习的效果优于使用全部的特征,而且高效。另外,通过与已有工作的比较,文中采用的方法也能得到较高的准确率,这对于挖掘潜在的药物靶点具有重要的指导意义。

## 参 考 文 献


[1] Li Qing-liang, Lai Lu-hua. Prediction of potential drug targets based on simple sequence properties [J]. BMC Bioinformatics, 2007,8(1):353

[2] Huang Chen, Zhang Rui-jie, Chen Zhi-qiang, et al. Predict potential drug targets from the ion channel proteins based on SVM [J]. Journal of Theoretical Biology, 2010, 262:750-756

[3] Drews J. Drug discovery: a historical perspective [J]. Science, 2000, 287(5460):1960-1964

[4] Imming P, Sinning C, Merver A. Drugs, their targets and the nature and number of drug targets [J]. Nature Reviews Drug Discovery, 2006, 5(10):821-834

[5] Dunlop J, Bowlby M, Peri R, et al. Highthroughput electrophysiology: an emerging paradigm for ion-channel screening and physiology [J]. Nature Reviews Drug Discovery, 2008, 7(4): 358-368

[6] Du Qi-shi, Huang Ri-bo, Wang Cheng-hua, et al. Energetic analysis of the two controversial drug binding sites of the M2 proton channel in influenza A virus [J]. Journal of Theoretical Biology, 2009, 259(1):159-164

[7] Huang R B, Du Q S, Wang C H, et al. An in-depth analysis of the biological functional studies based on the NMR M2 channel structure of influenza A virus [J]. Biochemical and Biophysical Research Communications, 2008, 377(4):1243-1247

[8] Pielak R M, Schnell J R, Chou J J. Mechanism of drug inhibition and drug resistance of influenza A M2 channel [J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(5):7379-7384

[9] Schnell J R, Chou J J. Structure and mechanism of the M2 proton channel of influenza A virus [J]. Nature, 2008, 451(7178): 591-595

[10] Hopkins A L, Groom C R. The druggable genome[J]. Nature Reviews Drug Discovery, 2002, 1(9):727-730

[11] Russ A P, Lampel S. The druggable genome: an update [J]. Drug Discovery Today, 2005, 10(23/24):1607-1610

[12] Hajduk P J, Huth J R, Tse C. Predicting protein druggability [J]. Drug Discovery Today, 2005, 10(23/24):1675-1682

[13] Kinnings S L, Liu N, Buchmeier N, et al. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis [J]. PLoS Computational Biology, 2009, 5(7):e1000423

[14] Xie Li, Li Jerry, Xie Lei, et al. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors [J]. PLoS Computational Biology, 2009, 5(5):e1000387

[15] Campillos M, Kuhn M, Gavin A C, et al. Drug target identification using side-effect similarity [J]. Science, 2008, 321(5886): 263-266

[16] Wang Yin-ying, Nacher J C, Zhao Xing-ming. Predicting drug targets based on protein domains [J]. Molecular BioSystems, 2012, 8(5):1528-1534

[17] Han Lian-yi, Zheng Chan-juan, Xie Bin, et al. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness [J]. Drug Discovery Today, 2007, 12(7/8):304-313

[18] Bao Lei, Sun Zhi-rong. Identifying genes related to drug anticancer mechanisms using support vector machine [J]. FEBS Letters, 2002, 521(1-3):109-114

[19] Bhardwaj N, Langlois R E, Zhao Gui-jun, et al. Kernel-based machine learning protocol for predicting DNA-binding proteins [J]. Nucleic Acids Research, 2005, 33(20):6486-6493

[20] Cai C Z, Han L Y, Ji Z L, et al. Enzyme family classification by support vector machines [J]. Proteins, 2004, 55(1):66-76

[21] 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, 6(14):4023-4037

[22] Knox C, Law V, Jewison T, et al. DrugBank 3.0: A comprehensive resourcefor 'omics' research on drugs [J]. Nucleic Acids Research, 2011, 39:1035-1041

[23] Bakheet T M, Doig A J. Properties and identification of human protein drug targets[J]. Bioinformatics, 2009, 25(4):451-457

[24] 刘明吉,王秀峰,黄亚楼. 数据挖掘中的数据预处理[J]. 计算机科学,2000,4(27):54-57

[25] Nan Xiao, Cao Dong-sheng, Xu Qing-song, et al. protr: Protein Sequence Feature Extraction with R[OL]. http://CRAN.R-project.org/package=protr

[26] 涂白,毕然. 支持向量机方法预测离子通道白[J]. 计算机与数字工程,2007,35(10):8-10