Journal of Theoretical Biology 262 (2010) 750–756

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Predict potential drug targets from the ion channel proteins based on SVM

Chen Huang [a,1], Ruijie Zhang [a,*,1], Zhiqiang Chen [a,1], Yongshuai Jiang [a], Zhenwei Shang [a], Peng Sun [a], Xuehong Zhang [a], Xia Li [a,b,**]

[a] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China
[b] Biomedical Engineering Institute of CUMS, Beijing 100054, China

ARTICLE INFO

Article history:
Received 6 August 2009
Received in revised form
4 November 2009
Accepted 4 November 2009
Available online 10 November 2009

Keywords:
Two-stage prediction
Secondary structure
Subcellular localization
Feature selection
Cross-validation

ABSTRACT

The identification of molecular targets is a critical step in the drug discovery and development process. Ion channel proteins represent highly attractive drug targets implicated in a diverse range of disorders, in particular in the cardiovascular and central nervous systems. Due to the limits of experimental technique and low-throughput nature of patch-clamp electrophysiology, they remain a target class waiting to be exploited. In our study, we combined three types of protein features, primary sequence, secondary structure and subcellular localization to predict potential drug targets from ion channel proteins applying classical support vector machine (SVM) method. In addition, our prediction comprised two stages. In stage 1, we predicted ion channel target proteins based on whole-genome target protein characteristics. Firstly, we performed feature selection by Mann-Whitney U test, then made predictions to identify potential ion channel targets by SVM and designed a new evaluating indicator Q to prioritize results. In stage 2, we made a prediction based on known ion channel target protein characteristics. Genetic algorithm was used to select features and SVM was used to predict ion channel targets. Then, we integrated results of two stages, and found that five ion channel proteins appeared in both prediction results including CGMP-gated cation channel beta subunit and Gamma-aminobutyric acid receptor subunit alpha-5, etc., and four of which were relative to some nerve diseases. It suggests that these five proteins are potential targets for drug discovery and our prediction strategies are effective.

1. Introduction

So far, some people have exerted great efforts on drug research and development (Li and Lai, 2007). Effective drugs have the characteristics of excellent efficacy and low side effects, in large part because of the appropriate choice of pharmacological targets. It is well known that mining drug target plays a great role in modern drug discovery, but only a few drug targets have been identified for clinically using drugs (Drews, 2000; Overington et al., 2006). Therefore, it is important to find out more potential targets for drug design and discovery.

According to the current knowledge, most of drug targets fell into approximately 130 protein families, including enzymes, transporters, G-protein-coupled receptors (Chou, 2005; Lin et al., 2009; Xiao et al., 2009), ion channel, nuclear receptors, etc., more than 50% of drug targets were located on only four key protein families: GPCRs, nuclear receptors, ligand-gated ion channels (7.9%) and voltage-gated ion channels (5.5%) (Hopkins and Groom, 2002; Overington et al., 2006). Further, some researchers (Chou, 2004; Dunlop et al., 2008) have reported that the ion channel proteins are a highly attractive targets for pharmaceutical research. Specifically, the M2 proton channel of influenza A is a crucially important target for curing influenza (Du et al., 2009; Huang et al., 2008; Pielak et al., 2009), and some drugs targeted by the M2 channel such as amantadine and rimantadine, has been used against influenza A viruses for many years(Huang et al., 2008; Schnell and Chou, 2008). Wang et al. (2009) have made an in depth study of the interactions of adamantane-based drugs with the M2 proton channel from the H1N1 swine virus, and expected to provide useful structural insights for developing effective drugs against the new swine flu virus. Ion channels are a diverse family of transmembrane proteins, and formed by the aggregation of subunits into a cylindrical configuration whose function is lower than the free energy required for ions to traverse the plasma membrane. Ions flux through ion channel proteins provides the conditions for membrane excitability, signal

* Corresponding author. Fax: +86 045186615922.
** Corresponding author at: College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China. Tel.: +86 045186650721-106; fax: +86 045186615922.
E-mail addresses: zhangruijie2009@yahoo.com.cn (R. Zhang), lixia6@yahoo.com (X. Li).
[1] Joint First Authors.

0022-5193/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2009.11.002

transduction and neurotransmission, so ion channel proteins play an important role in functions of neurons, cardiac and muscle cells. All these indicate that ion channel proteins have huge potential for drug discovery, and it is of great significance to mine new drug targets from ion-channel proteins.

At present, a number of experimental and computational tools have been developed to identify new drug targets (An et al., 2004; Guimera et al., 2007; Hajduk et al., 2005a, b; Kinnings et al., 2009; Mullner et al., 1998; Russ and Lampel, 2005; Xie et al., 2009). Some researchers have analyzed known drug targets based on sequence homology and domain-containing to find new targets from target family members (Hopkins and Groom, 2002; Russ and Lampel, 2005), and others have searched binding sites that might bind to drug-like compounds on the protein surface based on 3D structures (Hajduk et al., 2005b; Kinnings et al., 2009; Xie et al., 2009). However, these kinds of methods are limited by the number of proteins with known 3D structure, only approximately 15% of human proteins have known 3D structures (Berman et al., 2000). Recently, Li and Lai (2007) and Han et al. (2007) predicted druggable proteins successfully by SVM methods, meanwhile the method had been widely used for predicting anticancer genes (Bao and Sun, 2002), proteins in families of high target concentrations (Bhardwaj et al., 2005; Cai et al., 2004a), various functional and structural classes proteins (Han et al., 2006). In our study, we applied the SVM method to make a two-stage prediction by combining three types of protein features. In stage 1, the whole-genome target protein characteristics were used to predict potential ion channel target proteins. In stage 2, we searched for potential ion channel targets based on known ion channel target protein characteristics.

2. Method

2.1. Datasets

We constructed two datasets to make a two-stage prediction based on whole-genome target protein characteristics and known ion channel target protein characteristics. The relationship of these datasets is illustrated in Fig. 1.

2.1.1. Dataset 1

The 1268 approved human target proteins stored in DrugBank database (2.5 versions) composed the positive sample set. Because there is still no protein which has been identified as non-durg target now, we followed the way in the previous works (Bakheet and Doig, 2009; Li and Lai, 2007) to establish the negative set. Firstly, we eliminated the 1268 drug targets and all members of their relevant 6009 families from Pfam database (Pfam 23.0 July 2008, 10,340 families), then we removed the high sequence similarity proteins to avoid feature information redundancy caused by high sequence similarity proteins (Bakheet and Doig, 2009). Finally, we got 7252 non-drug targets. To avoid a bias causing by different size of two sample sets, we extracted 1268 non-drug targets from 7252 non-drug targets randomly as the negative sample set. Furthermore, the number of negative samples is less than 1/5 of the total negative samples, so it may not represent all negative samples. To solve this problem, we randomly picked 1268 proteins from 7252 non-drug targets 100 times and constructed 100 negative sets. At last, we defined 100 sample sets, each of which was made up of the same positive set and one of 100 negative sets.

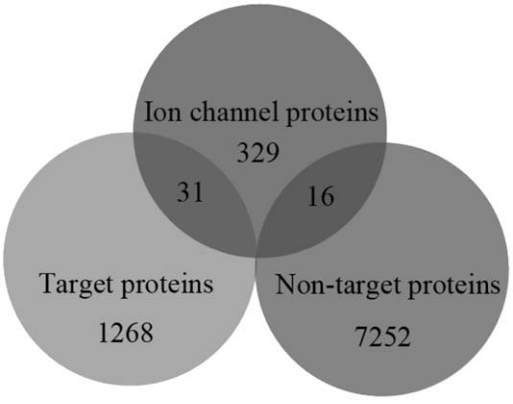

Fig. 1. The relationship of different datasets.

Training set: For each of 100 sample sets, 10-fold cross-validation was applied when training the model. We divided each of the sample sets into 10 folds, 9/10 of the positive set combined with 9/10 of negative set were used to construct the training set.

Testing set: The rest of 1/10 of positive set and 1/10 of negative set were used as testing set.

Prediction set: We took 329 human ion channel proteins as prediction set from the Ion Channel Database (http://www.ionchannels.org/database.php), the detail information provided in the supplementary file S1.

2.1.2. Dataset 2

We found out 31 ion channel proteins from 1268 approved target proteins, which we defined as known ion channel target proteins (positive sample set). Simultaneously, 16 ion channel proteins from 7252 non-target proteins were found, and we defined them as known ion channel non-target proteins (negative sample set).

Training set: Similar to dataset 1, we applied 5-fold cross-validation to train the model. 80% of the positive set, combined with 80% of negative set were considered as the training set.

Testing set: The remaining of 20% positive and negative set constituted the testing set.

Prediction set: Different from dataset 1, we used the 282 ion channel proteins as prediction dataset removing known ion channel target proteins and known ion channel non-target proteins.

2.2. Features of the protein

2.2.1. Protein primary sequence features

The UniprotKB/SwissProt datafile release 56.0 (22 July 2008) from Uniprot 14.0 was used for protein sequence information. The features of protein sequence were selected according to the previous works (Ding and Dubchak, 2001; Dubchak et al., 1995; Han et al., 2007; Li and Lai, 2007). First of all, we got the first 20 features by the percentage composition of the amino acid residues. Secondly, four physicochemical properties were used in our study including hydrophobicity, polarity, polarizability and normalized van der waals volume (Table 1). For each of physicochemical properties, amino acids could be divided into

Table 1
Features of proteins sequence.

Dimension	Properties
20	Composition of the 20 amino acid residues
21	Hydrophobicity
21	Polarity
21	Polarizability
21	Normalized van der Waals volume