FEBS Letters 580 (2006) 6800–6806

Prediction of the deleterious nsSNPs in ABCB transporters

Yanhong Li[a], Yonghua Wang[a], Yan Li[b], Ling Yang[a,*]

[a] *Laboratory of Pharmaceutical Resource Discovery, Dalian Institute of Chemical Physics, The Chinese Academy of Sciences, #457 Zhongshan Road, Dalian 116023, China*
[b] *School of Chemical Engineering, Dalian University of Technology, #158 Zhongshan Road, Dalian 116012, China*

Received 11 September 2006; revised 2 November 2006; accepted 14 November 2006

Available online 27 November 2006

Edited by Robert B. Russell

Abstract The non-synonymous SNPs (nsSNPs) in coding regions, neutral or deleterious, could lead to the alteration of the function or structure of proteins. We have developed the computational models to analyze the deleterious nsSNPs in the transporters and predict ones in ABCB (ATP-binding cassette B) transporters of interest. The RPLS (ridge partial least square) and LDA (linear discriminant analysis) methods were applied to the problem, by training on a selection of datasets from a specified source, i.e., human transporters. The best combination of datasets and prediction attributes was ascertained. The prediction accuracy of the theoretical RPLS model for the training and testing sets is 84.8% and 80.4%, respectively (LDA: 84.3% and 80.4%), which indicates the models are reasonable and may be helpful for pharmacogenetics studies.

Keywords: Prediction; Deleterious nsSNPs; ABCB transporters; Computational

1. Introduction

Single nucleotide polymorphisms (SNPs), which are found every 200–300 bp, represent the most abundant class of genetic variations in the human genome [1]. Up to June 17, 2006, 24910873 SNPs have been deposited to public databases (NCBI dbSNP Build 126) [2]. Non-synonymous SNPs (nsSNPs), which cause the changes of amino acid residues in proteins, account for almost half of all DNA mutations and may be functionally neutral or deleterious [3]. The disease-causing variations may cause deleterious effects on proteins: they may inactivate the functional sites or interact sites of enzymes or impact the folding of proteins; they may significantly destabilize the stability of proteins, or change the solubility of proteins [4–6]. Moreover, mutation sites at the N and C termini (or even within domains) often lead to difficulties in the protein expression, purification and crystallization [7], and are hence diseases associated.

Discovering the deleterious mutations is the mainly task of pharmacogenomics and pharmacogenetics. It is well known that mining them from dbSNP database is a laborious project only by site directed mutagenesis experiments and gene knockout/knockin experiments with more and more nsSNPs data available. Therefore, a primary challenge currently is that how to accurately predict those potentially deleterious nsSNPs. Several groups have tried to evaluate the deleterious nsSNPs based on 3-dimensional (3D) structure information of proteins (or homologous structures) in silico. Karchin et al. considered that the strongest predicting signals in the lac repressor/lysozyme set were solvent accessibility and superfamily-level evolutionary conservation [8]. Sunyaev et al. and Chen et al. also indicated that the residue solvent accessibility, which could identify the buried residues, was confidently proposed as predictors of deleterious substitutions [5,9].

However, the theoretical prediction methods for deleterious nsSNPs are still in its infancy since the 3D structural information of most proteins are still unavailable [10–12]. Therefore, it is a consequentially trend to predict the deleterious variations of proteins using sequence-based and position-specific evolutionary information [5,13,14]. The homology-based algorithm, SIFT (Sorting Intolerant From Tolerant) developed by Pauline et al. [14,15], was used to predict the conservation indices of all 20 possible amino acids at a given position according the ortholog sequences and determine which nsSNPs would be intolerant variations. Some other methods based on Site Entropy calculations, relative stability changes ($\Delta\Delta G$) were also developed for predicting deleterious nsSNPs [14,16,17]. These methods based on protein sequence have been demonstrated that the accuracy is the same as other methods using tertiary structure information [17].

The relationships between the genotype and phenotype of nsSNPs in transporters have received a plenty of research attentions because of their prevalence in the drug responses and close association to many inherited diseases. Transporters could medicate a wide range of fundamental biological processes, such as the cell signaling, transport of membrane-impermeable molecules, cell–cell communication, cell adhesion and recognition [18,19]. The ATP-binding cassette B (ABCB/MDR/TAP) transporter subfamily includes 11 members and is unique in mammals in that it contains both the full and half transporters [20]. Both in vitro and in vivo studies have

*Corresponding author. Fax: +86 411 84676961.
E-mail address: yling@dicp.ac.cn (L. Yang).

Abbreviations: SNP, single nucleotide polymorphism; nsSNPs, non-synonymous SNPs; ABCB, ATP-binding cassette B; RPLS, ridge partial least square; LDA, linear discriminant analysis; 3D, 3-dimensional; SIFT, sorting intolerant from tolerant; $\Delta\Delta G$, relative stability changes; ADME/T, absorption, distribution, metabolism, excretion and toxicity; DDI, drug–drug interaction; SVR, support vector regression; RSA, relative solvent accessibility; ASA, accessible surface area; ECEPP, the empirical conformational energy program for peptides algorithm; MCC, Matthew's correlation coefficient; BER, balanced error rate; TM, transmembrane

0014-5793/$32.00
doi:10.1016/j.febslet.2006.11.047

revealed that some nsSNPs in ABCB transporters play a key role influencing the ADME/T processes (absorption, distribution, metabolism, excretion and toxicity) of a wide variety of drugs, and are also one reason to induce the drug–drug interaction (DDI) in humans [21]. ABCB1 (MDR1/PGY1), the first human ABC transporter cloned, could transport several hundreds drugs and confer cancer multidrug resistance [22]. The nsSNPs of ABCB4 and ABCB11, located in the liver, are mainly reasons for the deregulation of the hepatobiliary circulation and correlative diseases with the cholestasis [23]. The variations of ABCB2 (TAP1) and ABCB3 (TAP2) proteins could lead to immunodeficiency [24,25]. The variations of four half transporters, ABCB6, ABCB7, ABCB8, and ABCB10, localized in the mitochondria and involved in iron metabolism, could baffle the transport of Fe/S complex into cytoplasm [26]. ABCB5, a novel drug transporter and chemoresistance mediator, determines the membrane potential and regulates the cell fusion in the physiologic skin progenitor cells [27]. The ABCB9 half transporter, which is the closest homolog of the TAPs, has been localized to lysosomes [26]. The wealth of pharmcogenetical studies revealed that most common diseases clusters, such as the ulcerative colitis (UC), progressive familial intrahepatic cholestasis (PFIC) syndromes, systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), ankylosing spondylitis, sideroblastic anemia, insulin-dependent diabetes mellitus, cholestatic liver and so on, are partially responsible for the variations of ABCB transporters, more information as shown in (http://www.tcdb.org/disease_explore.php) [21–27]. With increasing knowledge of the properties of ABCB transporters now, it is feasible to predict the phenotype of an nsSNP from the genotype by in silico methods.

Deleterious nsSNPs analyses for the transporters have not been estimated computationally till now, although they have received great focus from experimental researchers. Therefore, in this work, the computational models were built to analyze the deleterious nsSNPs in the transporters, and were used to predict the deleterious ones in the ABCB subfamily. Up to our knowledge, it is still difficult to obtain the whole 3D structure information of most human transporters, including the ABCB transporters, thus resulting in the difficulties of building computational models based on their 3D structures. In order to overcome the barriers, we have developed sequence-based models combined with some predicted structure information for all transporters in the datasets. The testing sets including 121 nsSNPs of ABCB transporters and the training sets including 762 nsSNPs of other transporters were carefully built, and a ridge partial least square (RPLS) analysis derived tool has been applied to predict the disease-causing variations in the datasets. As a comparison to the RPLS, the linear discriminant analysis (LDA) method has also been used in building models.

2. Materials and methods

2.1. Datasets

All the transporter IDs were collected from the TCDB database with classification information (http://www.tcdb.org/hgnc_explore.php), the detailed description about polymorphism and protein sequence were obtained by the Swiss-Prot database [28] and NCBI human genome protein sequence [2].

The databases of Swiss-Prot sequence variants provide full information of classification about nsSNPs associated with a given Swiss-Prot entry (Release 49.1 of 21-Feb-2006) [28]. All the variants in the database are therefore labeled as disease, unclassified or polymorphism, respectively, which have been demonstrated by a variety of reports [28]. Mutations in transporters labeled as disease or polymorphism used in this work were collected from the Swiss-Prot database. The mapped nsSNP was kept where the amino acid was the same in both the Swiss-Prot protein sequence and the NCBI human genome protein sequence [2]. All the transporters applied in this work lack the whole 3D structure information, which is limited from 350 to 1500 amino acids in length. The length restriction of sequence in training sets is made to build a more reasonable dataset, since all the ABC transporters (testing sets) are relatively large proteins, ranging from 686 to 1321, as shown in Table 1.

Table 1
Distribution of nsSNPs in ABCB transporters

Member	Length(Aa)	Pro_ID	No. nsSNP
ABCB1	1280	2506118	28 (18)
ABCB2	808	9665248	16 (8)
ABCB3	686	549044	16 (3)
ABCB4	1279	126932	13 (9)
ABCB5	812	36413607	4 (0)
ABCB6	842	13123949	6 (0)
ABCB7	752	8928549	7 (5)
ABCB8	718	6005804	6 (1)
ABCB9	766	22095458	2 (1)
ABCB10	738	22095459	2 (1)
ABCB11	1321	12643301	21 (10)

No. nsSNP means the number of the nsSNPs in ABCB transporters. The number in each bracket refers to the number of neutral or deleterious nsSNPs already known according to Swiss-Prot or literatures.

2.1.1. Training sets.

I. *Deleterious variations dataset*: 540 nsSNPs were collected from 50 transporters of five families (Table 2). Deleterious variations were labeled as disease in the Swiss-Prot database.

II. *Neutral variations dataset*: 222 nsSNPs were collected from 88 transporters of eight families (Table 2). Neutral variations were labeled as polymorphism in Swiss-Prot database.

2.1.2. Testing sets.

One hundred and twenty-one nsSNPs in ABCB transporters were extracted from the above databases and literature [2,21–29]. In this dataset, the 56 nsSNPs have already been known as phenotypes, neutral and deleterious according to the literature information, as shown in Table 1.

2.2. Candidate features

2.2.1. Evolutionary-conservation features.

I. *SIFT score*. PSI-BLAST in SIFT was used to search against the EMBL non-redundant protein database for homologous sequences and to build a multiple sequence alignment (MSA). It could compute the frequency of the amino acid a occurring at position i (f_{ia}) in MSA. The f_{ia} is given as a score ranging from 0.0 to 1.0, and the nsSNP whose score is less than 0.05 is considered to be deleterious. A median sequence conservation score of $\leqslant 3.25$ is considered as reasonable accuracy and the correspondingly sequence diversity is adequate. In general, for the protein sequence, SIFT performs MSA until a median sequence conservation score for the sequence is reached at the default of 3.0 and whether a substitution with any of the other amino acids is

Table 2
The families of the 138 transporters in the training sets

Human transporter family	Deleterious training set	Neutral training set
	Members	Members
Potassium channels	10	11
Calcium channels	0	2
Annexins	0	1
Sodium channels	24	42
Solute carriers	8	17
ATPase	6	8
Amino acid transporters	2	4
Others	0	3