Prediction of the deleterious nsSNPs in ABCB transporters

Yanhong Li\textsuperscript{a}, Yonghua Wang\textsuperscript{a}, Yan Li\textsuperscript{b}, Ling Yang\textsuperscript{a,*}

\textsuperscript{a} Laboratory of Pharmaceutical Resource Discovery, Dalian Institute of Chemical Physics, The Chinese Academy of Sciences, \#457 Zhongshan Road, Dalian 116023, China
\textsuperscript{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.

© 2006 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

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 (ΔΔG) 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 mediate 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
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/PGPIO), the first human ABC transporter cloned, could transport several hundreds of 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 pharmacogenetical 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.

2.1.1. Training sets.

I. Deleterious variations dataset: 540 nsSNPs were collected from 30 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 [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 α occurring at position i (fα) in MSA. The fα 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 ≤3.25 is considered as reasonable accuracy and the corresponding 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>
<thead>
<tr>
<th>Table 1</th>
<th>Distribution of nsSNPs in ABCB transporters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Member</strong></td>
<td><strong>Length(Aa)</strong></td>
</tr>
<tr>
<td>ABCB1</td>
<td>1280</td>
</tr>
<tr>
<td>ABCB2</td>
<td>808</td>
</tr>
<tr>
<td>ABCB3</td>
<td>686</td>
</tr>
<tr>
<td>ABCB4</td>
<td>1279</td>
</tr>
<tr>
<td>ABCB5</td>
<td>812</td>
</tr>
<tr>
<td>ABCB6</td>
<td>842</td>
</tr>
<tr>
<td>ABCB7</td>
<td>752</td>
</tr>
<tr>
<td>ABCB8</td>
<td>718</td>
</tr>
<tr>
<td>ABCB9</td>
<td>766</td>
</tr>
<tr>
<td>ABCB10</td>
<td>738</td>
</tr>
<tr>
<td>ABCB11</td>
<td>1321</td>
</tr>
</tbody>
</table>

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.

Table 2

The families of the 138 transporters in the training sets

<table>
<thead>
<tr>
<th>Human transporter family</th>
<th>Deleterious training set</th>
<th>Neutral training set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potassium channels</td>
<td>Members</td>
<td>Members</td>
</tr>
<tr>
<td>Calcium channels</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Annexins</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Sodium channels</td>
<td>24</td>
<td>42</td>
</tr>
<tr>
<td>Solute carriers</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>ATPase</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Amino acid transporters</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Others</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>
tolerant or intolerant for every position in the proteins. The SIFT version 2.0 was used for the analyses in this work [15].

II. Site entropy. It was used for measuring the sequence variability and calculated based on the entropy-based equation given as follows [17]:

\[ C^i(i) = \frac{\sum_{x=1}^{20} f_x(i) \ln f_x(i)}{\ln 20} \]

where \( C^i(i) \) is the entropy with the reverse sign at position \( i \), \( f_x(i) \) represents the frequency of the amino acid \( x \) at \( i \)th position obtained from MSA within ortholog sequences. We have developed a program to calculate Site Entropy according to this formula using C language.

2.2.2. Structural features.

I. Relative solvent accessibility (RSA). The RSA of amino acid residue \( i \) was defined as the ratio of its solvent-exposed surface area observed in a given structure (SA) and the maximum achievable solvent exposed surface area for this amino acid (MSA):

\[ RSA = \frac{SA_i}{MSA_i} \times 100\% \]

RSA can hence adopt values from 0% to 100%, with 0% corresponding to a fully buried and 100% to a fully accessible residue, respectively. In this study, the real value of RSA parameter was obtained for each residue using the threshold of 25% [30].

II. Solvent accessible surface area (ASA). The solvent accessible surface is defined as the locus of the centre of a probe sphere (representing the solvent molecule) as it rolls over the van der Waals' surface of the protein. A two-stage support vector regression (SVR) approach is proposed to predict the real values of ASA from the position specific scoring matrices generated from the PSI-BLAST profiles by using the MSA information [12].

2.2.3. Sequence-based features.

I. Protein stability changes: \( \Delta \Delta G \). A direct prediction of the value of \( \Delta \Delta G \) can be used to infer the directions of mutations. A statement of whether the protein stability is predicted to be increased or decreased by the mutation, and a confidence. If the \( \Delta \Delta G \) is positive, the mutation increases stability and is classified as a positive example. If \( \Delta \Delta G \) is negative, the mutation is destabilizing and is classified as a negative example. A score near 0 means unchanged stability. A score near -1 means high confidence in decreased stability. A score near +1 means high confidence in increased stability [16].

2.2.4. Physicochemical properties of amino acids. All changes of the physicochemical properties and their absolute values were analyzed using both the RPLS and LDA methods. The relevant calculation programs were compiled by C language.

I. Volume calculations (\( \Delta V \)). The average standard residue volumes were collected from Gerstein et al [31]. The changes in volume, \( \Delta V \), were evaluated using the following expressions for substitutions:

\[ \Delta V = V_{\text{wild-type}} - V_{\text{variant}} \]

Here, \( V \) is the volume of each amino acid. \( V_{\text{variant}} \) refers to the volume pertaining to amino acid causing substitution and \( V_{\text{wild-type}} \) refers to the volume of the amino acid in the native protein.

II. Extended state ASA calculations. The extended state ASA was calculated using ECEPP/2 algorithm (the empirical conformational energy program for peptides algorithm) with dihedral angles given in an extended tripeptide Ala-X-Ala/Glu-x-Glu conformation [32].

III. \( \Delta \Delta G \), calculations. \( \Delta \Delta G \) (kcal/mol) of hydration was computed at 298 K for the N-acetyl-N'-methylamides of the 20 naturally occurring amino acids [33].

IV. Amino acid scales (Swiss-Prot) calculations [28]. An amino acid scale is defined as a numerical value assigned to each type of amino acid. In this work, the scales used in our calculations were as follows (each scale was represented by its abbreviation in the bracket):

Changes in hydrophobicity (CH); average area buried on transfer from standard state to folded protein (AAB); free energy of transfer from inside to outside of a globular protein (FE); hydration potential at 25 °C (HP); hydrophobicity scale based on free energy of transfer (kcal/mol) (HE); hydrophobicity (HH); hydrophobicity scale (pi); (HS); normalized consensus hydrophobicity scale (NCH); optimized matching hydrophobicity (OMH); refractivity; average flexibility index (AFI); recognition factors (RF); bulkiness; molecular weight of each amino acid (MW); PG and PZ represent two types of polarity parameter according to Zimmermann and Grantham, respectively.

2.2.5. Other features.

I. DIspro: ordered or disordered [34]: The residue is ordered if a score <0.5. Otherwise, the residue is disordered if a score >0.5. The DIspro score was used to judge a nsSNPs to be deleterious or neutral.

II. DIspro: disulfide bridges [34]: If a mutation is located in the disulfide bridges, it would be thought to be deleterious.

III. Secondary structure: disease-associated nsSNPs also have a slightly different secondary structure propensity, which tend to occur at \( \beta \)-sheet sites [35].

VI. The mutations located in those places described as binding or active sites in the Swiss-Prot database are more likely to be deleterious [5,8].

2.3. Mathematical tools

A recently developed algorithm combining partial least squares (PLS) and ridge penalized logistic regression was used. The procedure is an integration of the regularization step (ridge penalty) and the dimension-reduction step (PLS).

As a detailed implementation has been given elsewhere [36], here, we only give a brief description of this method. RPLS divides in two steps:

1. \( (Z^o, W^o) \to \text{RRLS}(y, X, \lambda) \);
2. \( \text{RPLS} \to \text{WPLS}(z^o, X, W^o, \kappa). \)

Let \( \lambda \) be some positive real constant and \( \kappa \) be some positive integer. RPLS depends on the two parameters, \( \lambda \) and \( \kappa \) is determined at the end of Step 1, as minimizing the BIC criterion (see the Ridge penalty section), and thus independently of \( \kappa \). The predictive performance of the resulting classification rule is illustrated on the datasets in this work. The question of classification in high dimensional setting can be effectively addressed by means of this method.

As a comparison to the RPLS, the LDA method was also applied into these datasets. The stepwise selection method has been used in building the LDA model. The aim of stepwise selection is to look alternately out several crucial features for differentiating two classes among all parameters. The coefficients of features in the regression equation were calculated by LDA, and they can somewhat describe the importance of the features in the LDA model. Building regression equation is the process to evaluate the parameters (partial regression coefficient) in the regression model. According to the absolute value of the coefficients (\( b_1, b_2, \ldots, b_p \)) before the dependent variables, it is easy to know the contributions of each parameters in the optimized LDA equation:

\[ y_i = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p \]

An internally developed C language programs have been applied for this work.

2.4. Model assessments

Matthew’s correlation coefficient (MCC), a rigorous statistical marker, was used to evaluate the prediction accuracy in our study, given by

\[ \text{MCC} = \sqrt{\frac{TP \times TN - FP \times FN}{(TP + FN)(TN + FP)(TP + FN)(TP + FP)}} \]

The MCC ranges from –1 to 1, where TP is the deleterious number of true positives, TN is the neutral number of true negatives, FP is the neutral number of false positives and FN is the deleterious number of false negatives. Moreover, we also evaluated our results using balanced error rate (BER) [37]:

\[ \text{BER} = \left( \frac{1}{2} \times \frac{FN}{TP + FN} \right) + \left( \frac{1}{2} \times \frac{FP}{TN + FP} \right) \]

Moreover, the effectiveness of our models is measured in the terms of sensitivity, specificity, error rate and accuracy [38]:

- Sensitivity is the percentage of relevant point mutations identified by TP/(TP + FN).
- Specificity is the percentage of correct point mutations validated by TN/(TN + FP).
- Error rate is the percentage of wrong decisions made by (FP + FN)/total extracted.
Accuracy is the percentage of correct decisions made by (TP + TN)/total extracted. It is clear, therefore, that any single number that represents the predictive power of the methods should account for all of the possibilities listed above.

3. Results

3.1. RPLS and LDA models

RPLS and LDA methods were employed to build several models for the 762 nsSNPs of the datasets, which were collected from all of the transporters without 3D structures in Swiss-Prot. In the optimized models, the RPLS analyses revealed that 500 of totally 540 known deleterious nsSNPs and 146 of totally 222 known neutral nsSNPs were correctly predicted in the training sets, respectively. In the testing sets, RPLS analyses revealed that the 21 (total = 22) known deleterious nsSNPs and 20 (total = 29) known neutral nsSNPs were correctly predicted, respectively. For the LDA method, the TP and TN were 479 and 163 for the same training sets. The TP and TN were 19 and 22 using LDA method for the same testing sets as shown in Table 3a.

3.2. Assessment of the models

In the RPLS model, the sensitivity, specificity, accuracy, MCC and BER were 92.6%, 65.8%, 84.8%, 0.617 and 0.208 for the training sets, respectively. Correspondingly, they were 95.5%, 69.0%, 80.4%, 0.648 and 0.178 for the testing sets, respectively. On the other hand, the sensitivity, specificity, accuracy, MCC and BER for the LDA model were 88.7%, 73.4%, 84.3%, 0.620 and 0.189, respectively, in the training sets. Correspondingly, they were 86.4%, 75.9%, 80.4%, 0.616 and 0.189, respectively, for the testing sets as shown in Table 3b.

3.3. ABCB transporters

Among the 56 known deleterious or neutral nsSNPs in different ABCB transporters, several nsSNPs were consistently predicted as deleterious by two models, which were consistent with the experimental results in the reports [39–41]. And they are S704R, G705R, G706S/D in ABCB2 (TAP1), A565T, T665A in ABCB3 (TAP2), and I498T in ABCB11 (ABCB). In addition, three other nsSNPs, the G185V in ABCB1, V284L and G1004D in ABCB11, which have been still unclassified in the Swiss-Prot [28], were consistently predicted as deleterious in this work. Interestingly, five nsSNPs, whose diseases-causing phenotype was always a disputed issue [42–45], were predicted as deleterious in our study as the K433M, A893T and K1076M in ABCB1, the R150K in ABCB4 and the Y343H in ABCB11. Because of the different experimental techniques used, the criteria for labeling a residue as a site of deleterious mutations were somewhat different. For both the RPLS and LDA models, 28 nsSNPs, distributed in the ABCB transporters with the exception of ABCB7, have been consistently evaluated to be deleterious as shown in Table 4.

3.4. Variations of amino acids

Fig. 1 shows the ratio of the top-eight ranked amino acids of variations in the datasets. Those amino acids were R, G, P, L, N, and S.

### Table 3a
Analysis results of LDA and RPLS

<table>
<thead>
<tr>
<th>Sets</th>
<th>Methods</th>
<th>TN</th>
<th>TP</th>
<th>FN</th>
<th>FP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>Training</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPLS</td>
<td></td>
<td>146</td>
<td>500</td>
<td>40</td>
<td>76</td>
<td>762</td>
</tr>
<tr>
<td>LDA</td>
<td></td>
<td>163</td>
<td>479</td>
<td>61</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>b.</td>
<td>Testing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPLS</td>
<td></td>
<td>20</td>
<td>21</td>
<td>1</td>
<td>9</td>
<td>51</td>
</tr>
<tr>
<td>LDA</td>
<td></td>
<td>22</td>
<td>19</td>
<td>3</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

LDA: linear discriminant analysis; RPLS: ridge partial least squares. TN (true negative): a neutral match that correctly recognized as neutral nsSNP by discriminators; TP (true positive): a deleterious match correctly recognized as deleterious nsSNP by discriminators; FN (false negative): a deleterious match that incorrectly recognized as neutral nsSNP by discriminators; and FP (false positive): a neutral match that incorrectly recognized as deleterious nsSNP by discriminators, respectively.

<table>
<thead>
<tr>
<th>Table 3b</th>
<th>Analysis results of LDA and RPLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sets</td>
<td>Methods</td>
</tr>
<tr>
<td>b.</td>
<td>Training</td>
</tr>
<tr>
<td>RPLS</td>
<td></td>
</tr>
<tr>
<td>LDA</td>
<td></td>
</tr>
<tr>
<td>Testing</td>
<td></td>
</tr>
<tr>
<td>RPLS</td>
<td></td>
</tr>
<tr>
<td>LDA</td>
<td></td>
</tr>
</tbody>
</table>

Sc, Sp, Er and Ac mean the sensitivity, specificity, error rate and accuracy, respectively. MCC means the Matthew’s correlation coefficient and BER means the balanced error rate.
replacements [15]. Therefore, it is feasible to predict the deletions. SNPs might be expected to produce the least conservative and informative predictors [37]. In this work, a new method, depending on two factors: the superior machine learning approach [4].

4.1. Performance of LDA and RPLS

In the transporters applying RPLS and LDA methods, combined predicted structure features, sequence-based features were firstly calculated and used to predict deleterious nsSNPs using different mathematical methods. In this work, the parameters of position-specific phylogenetic features were firstly calculated and used to predict deleterious nsSNPs in the transporters applying RPLS and LDA methods, combining predicted structure features, sequence-based features and physicochemical properties of amino acids.

4.2. Candidate features

The RPLS is a combination method of partial least squares and ridge penalized logistic regression, therefore it seems to be difficult to directly find the crucial features among the original ones to build a model by only using RPLS. In contrast, LDA, appearing complementary to the RPLS, is tractable to the crucial features in the model by using the stepwise method. Through an analysis of LDA equation, it is easy to explain the impacts of different features on the LDA model. Here, based on our resulted LDA equation, the candidate features are described as follows:

\[
Y = 0.557 \times \text{SIFT} - 0.341 \times \text{Site entropy} + 0.225 \times \Delta \text{DG} - 0.445 \times \text{ASA} + 0.381 \times \text{RSA} + 0.386 \times \text{DISpro} - 0.136 \times \text{FE} + 0.126 \times \text{OMH} + 0.173 \times \text{PZ} + 0.341 \times \text{Bulkiness}
\]  

(1)

4.2.1. Position-specific phylogenetic features.

The phylogenetic estimators could quantify how well conserved a given amino acid is at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a protein [15]. The SIFT score and Site Entropy describe the conservation of an amino acid at a specific position in a sequence. In Eq. (1), the coefficients for the position-specific phylogenetic features are 0.557 and –0.341, respectively, indicating the importance of the predictors. These results also show the validity of SIFT in the prediction of deleterious nsSNPs, which is consistent with previous research [14,15].
4.2.2. Solvent accessibility. It is a key property of amino acid residues, important for both the structure maintaining and functioning of proteins. The residues whoever identified as buried using the solvent RSA are confidently deleterious [9]. Two algorithms for predicting the RSA and ASA were used in this study, which were significant improvements over previously published tools [12,30]. The coefficients of ASA and RSA in Eq. (1) are −0.445 and 0.381, respectively, revealing that they are both important for differentiating the deleterious variations from neutral ones. The absolute value of the ASA coefficient was larger than that of the RSA in our model, which might indicate that it is more meaningful to know the real ASA values than the types of the residues as buried or exposed (indicted by RSA) [47]. One possible reason is that ASA could directly reflect the degree to which the residues are in contact with the solvent molecules.

4.2.3. Sequence-based features. \( \Delta \Delta G \): The sequence information can be used to effectively predict the changes of the protein stability for single site mutations [16]. In our model, the LDA analysis indicated that \( \Delta \Delta G \) was not predominant in classifying disease-causing ones from neutral variations (coefficient = 0.225).

4.2.4. Other features. DISpro uses a one-dimensional recursive neutral network (1D-RNN) to predict the probability whether the residues are disordered or not with threshold of 0.5. Although originally the DISpro scores were not used to differentiate the deleterious nsSNPs from the neutral ones, they were a good predictor in the models (coefficient = 0.386) surprisingly [34].

Bulkiness was defined as the ratio of the volume to the length of a side chain, providing a measure of the average cross section of the chain [48]. It was easy to see from Eq. (1) that the absolute value of the coefficient for Bulkiness was similar with the Site Entropy (0.341 and −0.341), which indicate that the Bulkiness is the same important as the RSA. Moreover, it was firstly testified that three amino acid scales, i.e., the polarity (PZ), OMH, and free energy (FE), were useful for improving the predictive power of the models. From above analyses, the ten features were optimized alternately using the LDA method.

Using RPLS and LDA models, 28 nsSNPs were always predicted to be deleterious based on the phylogenetic features in conjunction with the structure information and other simple amino acid features. As shown in Table 4, the 19 deleterious nsSNPs may occur in the helix sections, and nine of them may occur in the transmembranes, e.g., A374T in ABCB3, L293V in ABCB6, R304C in ABCB8. It was seen that the frequency of occurrence of the deleterious nsSNPs in transmembrane (TM) regions was larger than those of the extracellular and intracellular loop regions. It may be that the tiny changes of properties in the amino acids are inclined to induce the changes in TM structure [18]. Moreover, several deleterious nsSNPs may occur near the ATPase binding sections of ABCB transporters, e.g., G1063A in ABCB1; P576L in ABCB2; A1100T in ABCB4; S560C in ABCB5; R432T in the ABCB11. This subset of nsSNPs in ABCB transporters we predicted would be best candidate nsSNPs for further investigation of genotypen and phenotype, as well as their role on the pathogenesis of specific human conditions in experiments.

In addition, two facts about the variation occurring in the transporters were observed in our results. Firstly, the distributions of all 20 amino acids were found to be greatly different between deleterious and neutral variations as shown in Fig. 1. Interestingly, those results were different from the statistical results about the relative mutability of amino acids in the Swiss-Prot database [31], and the top-eight amino acids were N, S, D, E, A, T, I and M. Secondly, the amino acid composition of variations of the two training sets were also different. For example, the most variation rates in the neutral sets were from I to V (or from V to I), from A to T, and from D to N; however, they were the rates from L to P (or from P to L), from K to E, from M to T, and from R to W in the deleterious sets. Those results indicate that the two facts may be the specific characteristic only belonging to the transporters, which may be different from the general proteins.

In conclusion, we have firstly and systemically refined a series of features for analyzing the deleterious variations in the transporters and predicting the deleterious nsSNPs in the ABCB transporters using RPLS and LDA methods. The models built in this work would be applicable for predicting the deleterious nsSNPs of other transporters. Moreover, many indices, including the SIFT, ASA, DISpro, RSA, Site Entropy, Bulkiness, \( \Delta \Delta G \), Polarity, free energy and OMH, were demonstrated to be useful for identifying the deleterious variations from neutral ones in the LDA models. Twenty-eight possibly deleterious nsSNPs in ABCB subfamily were also identified, which would be helpful for further genotype–phenotype researches as well as the pharmacogenetics study of transporters.

Acknowledgements: This work was supported by the funds from the National Natural Science Foundation of China (30600666) and (30630075), DUT-DICP Scientific Research Foundation. We thank Dr. Zhe Chen for program (School of Electronic and Information Engineering Department of Electronic Engineering, Dalian University of Technology, China). We also thank Dr. Pauline C. Ng for SIFT analysis as well as responding to all our questions regarding SIFT.

References


