Introduction

Cytochrome P450 (CYP450) is one of the most functionally diverse superfamilies of heme-thiolate proteins [1]. CYP450s have been found in a great number of organisms, such as bacteria, animals and plants [2]. The name CYP450s refers to the intense spectral absorption peak at 450 nm when their reduction states are combined with carbon monoxide [3]. As strong oxidizing agents, CYP450s play decisive roles in much of the oxidative metabolism of various endogenous and xenobiotic substrates [4]. In humans, CYP450s are of central importance to drug interactions and inter-individual variability in phase I drug metabolism [5]. They are involved in the metabolism of many marketed drugs and are responsible for major elimination pathways of drug clearance [4]. In recent years, human CYP450s have been widely studied. More than 50 human CYP450 genes have been identified with genetic polymorphisms. However, only a minority of the known human CYP450 genes can encode proteins, such as CYP1, CYP2 and CYP3 families, that are likely to work on drug metabolism [6].

Single nucleotide polymorphisms (SNPs) are single nucleotide mutations in the genome sequence with mutation rates higher than 1% of the population. SNPs make up about 90% of all human genetic variation [7, 8]. Human CYP450 genes have minor variation in genetic sequence, leading to nu-
cleotide changes and polymorphisms [9]. SNPs constitute a great proportion of genetic polymorphisms of human CYP450, and play crucial roles in causing individual and population differences in response to diseases, viruses, toxins, drugs, etc. [8]. In other words, SNPs in different CYP450 genes have different effects [10]. SNPs are valuable for predicting individual response to drugs as well as designing personalized drugs with greater efficacy and safety [11]. SNPs occur in both coding regions (cSNPs) and non-coding regions. cSNPs code for proteins and are most likely to affect gene function [12]. cSNPs can be classified into synonymous SNPs and non-synonymous SNPs (nsSNPs). Compared with synonymous SNPs, nsSNPs can lead to amino acid mutations, which may cause changes in the structure and function of proteins and a high incidence of disease [13]. In this study, we focus on nsSNPs. Unlike general mutation prediction methods based on DNA sequences [14–16], our prediction method is based on protein sequences of the human CYP450 nsSNPs.

Many studies have identified that various computational approaches, such as among others molecular docking [17–19], molecular dynamics [20, 21], structural bioinformatics [22–24], predicting drug-target interaction [25], protein subcellular location prediction [26, 27], predicting metabolic stability [28], protein cleavage site prediction [29, 30], and signal peptide prediction [31], can provide quite useful information and insights for drug development. Mutation prediction is also a major part of computational biology and hence widely welcomed by the science community. There have been a great number of successful applications of mutation prediction, e.g., for influenza viruses [32–37], metabolism enzymes [38–43], hemoglobin [44, 45], and tumor proteins [46, 47].

The support vector machine (SVM) is one of the most popular machine learning techniques used for mutation prediction in recent studies. Many successful applications of SVM have identified an association between mutations and the susceptibility to complex diseases, such as breast cancer or heart disease [14–16]. The efficiency and accuracy of SVM classification is largely limited by two factors, feature selection and parameter setting [48]. Highly redundant features increase the computational complexity of modeling and decrease the mutation prediction accuracy. Parameter optimization plays a decisive part in model generation [49]. These two factors are of interactive influence to a large extent and needed to be studied simultaneously.

Most studies, however, focus on either one or the other factor separately. For feature selection, the reported algorithms contain, for example, a genetic algorithm (GA) [50], simulated annealing (SA) [51], or a recursive feature eliminations (RFE) [52]. For parameter optimization, algorithms such as GA [49] and a grid search strategy [53] have been proposed. Unfortunately, few studies have made simultaneous consideration of both factors. In this study, we developed a GA-SVM program to optimize feature selection and parameter setting simultaneously. In GA-SVM, an adaptive genetic algorithm is utilized for feature selection and the grid search strategy in LIBSVM is used for parameter optimization. The results demonstrate that our GA-SVM model can select features for a predictive model as well as achieve higher mutation prediction accuracy for human CYP450 nsSNPs.

2 Materials and methods

2.1 Dataset preparation

2.1.1 Mutation site selection

The human CYP450 nsSNPs and their amino acid mutations were sourced from the Human Cytochrome P450 Allele Nomenclature Committee (http://www.cypalleles.ki.se). Protein sequences of the human CYP450 superfamily were downloaded from NCBI protein database (http://www.ncbi.nlm.nih.gov/protein/). Among all families of human CYP450 superfamily, the CYP2 family has the largest number of nsSNPs, with 13 subfamilies. We therefore extracted the data for the CYP2 family and took four subfamilies (CYP2W1, CYP2E1, CYP2R1 and CYP2S1) with fewer than six nsSNPs. Finally, 188 amino acid sequences of nsSNPs were obtained and labeled as positive examples. The local window has a length of 3 and is centered at the mutation site with one amino acid residue in both sides.

2.1.2 Conservative site selection

Conservative sites of the CYP2 family were obtained by multiple sequence alignment using the CLUSTAL X program with the default parameters. The nine sequences used in the alignment were from the same subfamilies of CYP2 family. We chose only the conservative sites with the conservative neighboring amino acids on both sides. Thus, we obtained 207 conservative amino acid sequences as negative examples.

We randomly selected 40 positive and 40 negative examples as the testing dataset. The training dataset included the remaining 148 positive and 167 negative examples. For details about training and testing data, see Supporting information online.
2.2 Feature selection

2.2.1 AAindex database

2.2.1.1 Feature preprocession

The AAindex database [54] in the GenomeNet Database Resources is the source for feature selection for the algorithm (http://www.genome.jp/dbget/aaindex.html). It consists of 544 amino acid features, which are the various attributes of amino acids and are represented as numerical values [55].

These 544 features were preprocessed in two steps. Firstly, the 531 features with complete values were retained. In the second step, we calculated the correlation coefficients between the different combinations of features. We only retained the features whose absolute values for the correlation coefficients were smaller than 0.8, removing highly correlated features. We finally had 147 features to make an initial feature set.

2.2.1.2 Vector selection

For each three-amino acid protein sequence x₁x₂x₃ in the dataset, the vector was defined as:

\[ X = (\Sigma f_{1j}, \Sigma f_{2j}, \Sigma f_{3j}, \ldots, \Sigma f_{nj}, \Sigma f_{i1}, \Sigma f_{i2}, \Sigma f_{i3})^T \]  

(1)

where \( n \) is the number of features selected through the GA algorithm, \( x_j \) is the amino acid at the mutation or conservative site, \( f_{ij} \) is the jth feature of the amino acid \( x_i \). The vector dimension is \( n+3 \). More details are shown in Table 1. The first \( n \) elements correspond to the \( n \) properties of the amino acid sequence and the last three elements represent the sequence-order information.

Considering the mutation sequence VGF, for example, the amino acid G is the mutation site. If \( n \) is set to 147, which means that all 147 features are selected, the vector can be calculated (see Table 2).

Thus, the vector can be represented as:

\[ X = (12.58, 3.41, 3.53, 1.77 \ldots 3.346, 987.8787, 689.41, 1072.234)^T \]  

(2)

with 150 dimensions.

This process is the same for conservative amino acid sequences.

2.2.2 Pseudo amino acid composition

The pseudo amino acid composition (PseAAC) [56] can represent a protein sequence while avoiding losing the sequence-order information. There are two discrete models of PseAAC, the parallel-correlation type (type 1) and series-correlation type (type 2). Each type contains six physicochemical properties of amino acids: hydrophobicity, hydrophilicity, mass, pK₁ (alpha-COOH), pK₂ (NH₃) and pI (at 25°C). Further information about PseAAC can be found on the website (http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/#).

For parameter \( \lambda \), we can get \( 20+\lambda \) discrete values to represent a protein sequence. The first 20 values are the frequencies of 20 amino acids, reflecting the effect of the amino acid composition. The last \( \lambda \) values represent physicochemical properties and are an indicator of the sequence order [57].

When the parameter \( \lambda \) is set to 1 and 21 values are obtained for each protein sequence. We used this value, which represents the property of the protein sequence. When the weight factor is set to 0.5, six properties of each type can be selected separately. Thus, our samples can be represented with 12 dimension vectors as:

\[ X = (x_1, x_2, x_3 \ldots x_{10}, x_{11}, x_{12})^T \]  

(3)

where \( x_1 \) to \( x_6 \) are six properties of type 1 and \( x_7 \) to \( x_{12} \) are six properties of type 2.

Table 1. Vector calculation of protein sequence x₁x₂x₃

<table>
<thead>
<tr>
<th>Feature number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>……</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x₁</td>
<td>f₁₁</td>
<td>f₂₁</td>
<td>f₃₁</td>
<td>f₄₁</td>
<td>……</td>
<td>fₙ₁</td>
</tr>
<tr>
<td>x₂</td>
<td>f₁₂</td>
<td>f₂₂</td>
<td>f₃₂</td>
<td>f₄₂</td>
<td>……</td>
<td>fₙ₂</td>
</tr>
<tr>
<td>x₃</td>
<td>f₁₃</td>
<td>f₂₃</td>
<td>f₃₃</td>
<td>f₄₃</td>
<td>……</td>
<td>fₙ₃</td>
</tr>
</tbody>
</table>

where \( \Sigma f_{ij} \) is the sum of the \( i \)th feature of the \( j \)th amino acid.

Table 2. Vector calculation of mutation sequence VGF

<table>
<thead>
<tr>
<th>Feature number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>……</th>
<th>147</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid</td>
<td>ANDN920101</td>
<td>ARGP820101</td>
<td>ARGP820102</td>
<td>BEGF750101</td>
<td>DIGM050101</td>
<td>Σ</td>
</tr>
<tr>
<td>V</td>
<td>3.95</td>
<td>1.32</td>
<td>1.08</td>
<td>0.82</td>
<td>……</td>
<td>1.131</td>
</tr>
<tr>
<td>G</td>
<td>3.97</td>
<td>0.07</td>
<td>0.49</td>
<td>0.35</td>
<td>……</td>
<td>1.346</td>
</tr>
<tr>
<td>F</td>
<td>4.66</td>
<td>2.02</td>
<td>1.96</td>
<td>0.60</td>
<td>……</td>
<td>0.869</td>
</tr>
<tr>
<td>Σ</td>
<td>12.58</td>
<td>3.41</td>
<td>3.53</td>
<td>1.77</td>
<td>……</td>
<td>3.346</td>
</tr>
</tbody>
</table>
2.3 GA-SVM for mutation prediction of human CYP450 nsSNPs

2.3.1 SVM for parameter optimization and modeling
SVM is a useful tool for pattern recognition, and is based on the structural risk minimization principle and a statistical learning approach [58]. It maps data into a feature space of high dimensions by kernel functions and seeks an optimal separating hyperplane maximizing the margin in this dimensional space [59]. In this study, LIBSVM software [60], developed by Chang and Lin at the National Taiwan University (http://www.csie.ntu.edu.tw/~cjlin/libsvm), was used for data classification. We applied a testing tool (easy.py) with Radial Basis Function kernel (RBF kernel) throughout the whole prediction process (Scheme 1). The steps of the process are:

1) Scaling the training data. Each attribute in the training data is linearly scaled to avoid large numeric ranges of the feature values [53]. All the attribute values are scaled into the range (-1, 1).

2) Cross-validation. The two parameters c and g of the RBF kernel are optimized using a fivefold cross-validation and C-SVM classification. The (c,g) pair with the best cross-validation accuracy (CV rate) is picked.

3) Training. The scaled training data are trained with the best (c,g) pair and C-SVM classification to obtain optimal model.

4) Scaling the testing data.

5) Testing. The scaled testing data are predicted with the optimal model to get prediction accuracy. Evaluation measures, sensitivity and specificity, are also calculated at the same time in this study.

2.3.2 GA for feature selection
1) GA is an adaptive optimization search algorithm simulating the principles of Darwinian evolution and survival of the fittest in biological systems [48, 49]. In this study, GA was used for the feature selection. It begins with a population of individuals represented by chromosomes. The chromosome of each individual is encoded with a randomly generated binary string to express the condition of feature selection (0: not selected, 1: selected) [48]. A fitness function is assigned as the quality evaluation of the individuals in the population to obtain the optimal individual surviving to the next generation. Other individuals can become members of successive populations through being adjusted randomly by three operators – selection (i.e., the probability of an individual being selected into the next generation is in proportion to its fitness value), crossover and mutation [61, 62] (adaptive crossover and mutation functions were used in this study):

\[ P_c = \begin{cases} P_{c1} & f > f_{avg} \\ P_{c2} & f \leq f_{avg} \end{cases} \]  

\[ P_m = \begin{cases} P_{m1} & f > f_{avg} \\ P_{m2} & f \leq f_{avg} \end{cases} \]  

where \( P_c \) is the crossover rate of individuals, \( P_m \) is the mutation rate of individuals, \( f_{max} \) is the best fitness value in the population, \( f_{avg} \) is the average fitness value of all the individuals in the population, \( f' \) is the better fitness value of the two crossover candidates, \( f \) is the fitness value of the mutation candidate. Here two crossover rates were separately set to 0.9 (\( P_{c1} \)) and 0.6 (\( P_{c2} \)). Two mutation rates were set to 0.1 (\( P_{m1} \)) and 0.001 (\( P_{m2} \)).

In the functions, the individuals with small fitness values are of high crossover rate (\( P_{c1} \)) and mutation rate (\( P_{m1} \)). Those with higher fitness values can have variable crossover and mutation rates according to their fitness values. The crossover and mutation rates of the optimal individual are \( P_{c2} \) and \( P_{m2} \) separately, which means that the optimal individual can also participate in evolution process. At the same time, the elitist strategy was employed to protect the optimal individual. Adaptive crossover and mutation functions can both keep the diversity of the population and the astringency of GA in such a way as to avoid the local optimization.

2.3.3 GA-SVM and evaluation measures
In the GA-SVM system, two processes, feature selection and parameter optimization, work simultaneously. The procedure flow is shown in Scheme 2.
The fitness function is defined as:

\[
\text{Fitness} = 10^4 \times \text{CV}_{\text{rate}} + \text{Zeros} \tag{6}
\]

where CV_rate is the accuracy of fivefold cross-validation, Zeros is the number of features that are not selected. The weight of CV_rate is set to $10^4$.

Some measures are used for performing the evaluation of classification: sensitivity (Sen), specificity (Spe), and prediction accuracy (Acc) [63].

\[
\text{Sen} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{7}
\]

\[
\text{Spe} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \tag{8}
\]

\[
\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100\% \tag{9}
\]

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives.

The Weka (Waikato Environment for Knowledge Analysis) workbench [64, 65] was used in this study. It contains a collection of data analysis and predictive modeling tools, such as clustering, classification, and regression. Three common classification methods (RBF network, multi-layer perceptron, and logistic methods) were chosen for comparison with GA-SVM. Among the three methods, the logistic method is a typical linear classifier, and the other two are typical non-linear classification methods.

3 Results and discussion

3.1 Selection of features

The whole process flow is shown in Scheme 3. Features were preprocessed and the training and testing datasets were set up. The initial model was then constructed with a total 147 features using the easy.py of LIBSVM with RBF kernel. The prediction accuracy was 0.637 and the CV rate was 0.644 in the initial model.

In our study, the GA-SVM program was performed with different parameter values and subsets of features four times. First, the 147 features, as the initial feature set, were fed into the GA-SVM program; 100 individuals were set in each generation of GA. After 200 generations, features were decreased into a subset of 56 features and the CV rate was increased to 0.724. Details are presented in Table 3. Second, we started GA-SVM again with 56 features and got a smaller subset of 24 features with 300 generations. We then calculated another 300 generations with these 24 features. The CV rate was 0.733 and 12 features were selected. Finally, we performed the GA-SVM program with 12 features for verification and all 12 features were selected as the final subset of features.

![Scheme 2. Procedure flow of GA-SVM.](image)

![Scheme 3. The whole process flow.](image)
It can be seen from the Fig. 1 that, along with the decrease of features, the CV rate rises stably. Three peaks in the figure correspond to optimal feature subsets of four GA-SVMs (GA-SVM3 and GA-SVM4 correspond to the same peak).

Compared to the initial model, the optimal model of GA-SVM1 utilized fewer features with a higher CV rate, from 0.644 to 0.724. Among the four GA-SVM optimal models, as the capacity of feature subsets decreased (from 56 to 12 features), the CV rates were stable and were all greater than 0.700. Thus, the results demonstrate that our GA-SVM program plays an important role in feature selection and succeeds in cutting down features as well as improving the predictive model. Other evaluation measures, prediction accuracy, sensitivity and specificity, can be seen in Table 3.

### 3.2 Construction of the final model

GA-SVM4 was calculated using 300 generations with 12 features from GA-SVM3. As the CV rate rose, some predictive models with different feature subsets were obtained. For example, we obtained a predictive model using only 9 features and achieved a CV rate of 0.705. Similarly, there were also predictive models with 10 features and 11 features (Table 4). In the final model, 12 features were selected as the final feature subset with the highest CV rate of 0.733.

The CV rate of the final model was significantly higher than the initial model whose CV rate was only 0.644. At the same time, the final model used only 12 features, much smaller than the initial feature set with 147 features. Other evaluation measures including prediction accuracy, sensitivity and specificity, were stable. This result demonstrates that the final model has a better performance than the initial model.

To compare with the final model, we constructed three classification models with WEKA software, an RBF network model, a multi-layer perceptron model and a logistic model. The three methods are typical linear and non-linear classification methods. The first two belong to non-linear classification, as does the final model of the SVM method. The logistic method is typical for linear classification. The features, training/testing data and vectors used in these three models were the same as in the final GA-SVM model. As shown in Table 5, the final model had the highest CV rate, which demonstrates the good performance of the SVM classification method in mutation prediction. In addition, the model constructed with logistic method also had good results, indicating the superior performance of our final feature subset and the stability of SVM classification method.

In the final feature subset, four of the features are physicochemical properties of the amino acids, and seven features are structural properties (Table 6). Other researches have proved that diverse properties have different effects on mutation prediction [66]. Therefore, we analyzed the influences of physicochemical and structural properties, separately, in mutation prediction, as described in the next section.

### 3.3 Analysis of physicochemical and structural properties

To analyze the two kinds of properties, we constructed two models [PseAAC and structural (Str) model], both of which use easy.py of LIBSVM with
RBF kernel. PseAAC model was constructed with its vector represented by 12 physicochemical properties as mentioned above. Then, we constructed a model (Str model) using only structural properties. We selected features from optimal feature subset of GA-SVM2 and obtained 13 structure derived features in total 24 features. The training/testing datasets of the two models were the same as the GA-SVM final model. The comparison of the two models and GA-SVM final model is shown in Table 7.

As can be seen in Table 7, the specificity of PseAAC model was higher than other methods; however, the Str model achieved the highest sensitivity. Therefore, it can be expected that physicochemical properties may play an important role in recognizing conservative sites, while structural properties may have great influence on recognizing mutation sites.

The final GA-SVM model, utilizing both physicochemical and structural properties, produced the highest CV rate and had the best performance. This suggests that the predictive model performs better when considering physicochemical and structural properties simultaneously.

4 Conclusions

In this study, we successfully developed a GA-SVM program to deal with feature selection and parameter optimization simultaneously, to increase the efficiency and accuracy of the predictive model. The GA-SVM program was applied to the mutation prediction of human CYP450 nsSNPs and succeeded in decreasing the capacity of feature subsets, from initial 147 features to 12 features. The final mutation predictive model also had a better performance than many typical linear and non-linear classification models. In addition, we analyzed the influences of physicochemical and structural properties in mutation prediction, and suggest that both kinds of properties should be considered when mutation predictive model is constructed.

In this research we studied the mutation of human CYP450 nsSNPs in protein sequences. This study has implications for further research of human CYP450 cSNPs, such as discovering new nsSNPs, and finding the differences between nsSNPs and synonymous SNPs in the process of drug metabolism and disease occurrence.

Table 4. Predictive models of GA-SVM

<table>
<thead>
<tr>
<th>Feature</th>
<th>CV rate</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>0.705</td>
<td>0.613</td>
<td>0.600</td>
<td>0.625</td>
</tr>
<tr>
<td>10</td>
<td>0.695</td>
<td>0.588</td>
<td>0.550</td>
<td>0.625</td>
</tr>
<tr>
<td>11</td>
<td>0.724</td>
<td>0.600</td>
<td>0.550</td>
<td>0.650</td>
</tr>
<tr>
<td>12</td>
<td>0.733</td>
<td>0.613</td>
<td>0.550</td>
<td>0.675</td>
</tr>
</tbody>
</table>

Table 5. Comparison with classification methods in WEKA

<table>
<thead>
<tr>
<th>Model</th>
<th>CV rate</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBF network model</td>
<td>0.5016</td>
<td>0.5625</td>
<td>0.4750</td>
<td>0.6500</td>
</tr>
<tr>
<td>Multilayer perceptron model</td>
<td>0.5905</td>
<td>0.5250</td>
<td>0.3750</td>
<td>0.6750</td>
</tr>
<tr>
<td>Logistic model</td>
<td>0.6095</td>
<td>0.6000</td>
<td>0.5750</td>
<td>0.6250</td>
</tr>
<tr>
<td>Final GA-SVM model</td>
<td>0.7330</td>
<td>0.6130</td>
<td>0.5500</td>
<td>0.6750</td>
</tr>
</tbody>
</table>

Table 6. Final feature subset

<table>
<thead>
<tr>
<th>Number</th>
<th>Code</th>
<th>Description (see <a href="http://www.genome.jp/aaindex">www.genome.jp/aaindex</a> for complete citations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BUNA790101</td>
<td>alpha-NH chemical shifts [67]</td>
</tr>
<tr>
<td>2</td>
<td>FASG760102</td>
<td>Melting point [68]</td>
</tr>
<tr>
<td>3</td>
<td>FAUJ880107</td>
<td>N.m.r. chemical shift of alpha-carbon [69]</td>
</tr>
<tr>
<td>4</td>
<td>WERD780103</td>
<td>Free energy change of alpha (Ri) to alpha (Rh) [70]</td>
</tr>
<tr>
<td>5</td>
<td>BURA740101</td>
<td>Normalized frequency of alpha-helix [71]</td>
</tr>
<tr>
<td>6</td>
<td>CHOP780206</td>
<td>Normalized frequency of N-terminal non helical region [72]</td>
</tr>
<tr>
<td>7</td>
<td>CHOP780207</td>
<td>Normalized frequency of C-terminal non helical region [72]</td>
</tr>
<tr>
<td>8</td>
<td>LEWP710101</td>
<td>Frequency of occurrence in beta-bends [73]</td>
</tr>
<tr>
<td>9</td>
<td>PRAM820101</td>
<td>Intercept in regression analysis [74]</td>
</tr>
<tr>
<td>10</td>
<td>RACS820101</td>
<td>Average relative fractional occurrence in A0(i) [75]</td>
</tr>
<tr>
<td>11</td>
<td>RICJ880101</td>
<td>Relative preference value at N&quot; [76]</td>
</tr>
<tr>
<td>12</td>
<td>WOLS870103</td>
<td>Principal property value z3 [77]</td>
</tr>
</tbody>
</table>

© 2011 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
We have several plans for future work. First, our model is still limited by the employed features since they are only selected from the AIndex database. More potential features will be incorporated into the current model. Second, we focused on nsSNPs of human CYP450 in this work, while the unstudied synonymous SNPs may play an indispensable role in the stability of mRNA and the translation of protein. We are going to move our focus onto synonymous SNPs. Finally, the feature selection and parameter optimization algorithm proposed in this study can be applied to other similar tasks, such as linear classification and support vector regression.

This work is supported by the grants from the Chinese National Comprehensive Technology Platforms For Innovative Drug R&D under the Contract no. 2009ZX9301-007, the Chinese National Science Foundation under the Contract no. 30870476, the National Basic Research Program of China (973 Program) (Contract no. 2011CB707500).

The authors have declared no conflict of interest.

5 References


Wu, G., Yan, S., Prediction of distributions of amino acids and amino acid pairs in human haemoglobin α-chain and its seven variants causing α-thalassaemia from their occurrences according to the random mechanism. Comp. Haematol. Int. 2000, 10, 80–84.


Witten, L. H., Eibe, F., Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2005.


Gao, S., Zhang, N., Duan, G. Y., Yang, Z. et al., Prediction of function changes associated with single-point protein mu-
tations using support vector machines (SVMs). Hum. Mutat. 2009, 30, 1161–1166.


