Three-State Protein Stability Prediction from Sequence-Based Features

Jose A. Guevara-Coto¹, Charles E. Schwartz², and Liangjiang Wang¹

¹Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA
²J.C. Self Research Institute of Human Genetics, Greenwood Genetic Center, Greenwood, SC 29646, USA

Abstract - Amino acid substitutions can have significant and deleterious effects on proteins. Prediction of the effects of substitutions on protein stability has been explored, but many studies make use of structure-based features, which are not available for all proteins. In this study, we have developed a sequence-based SVM model for three-state protein stability prediction. This model used features extracted from the primary sequence, and feature selection identified the most informative feature set for model construction. We evaluated this model with an independent test dataset, and obtained the accuracy of 70.52% with 61.20% sensitivity and 79.84% specificity. Our results suggest that sequence features contain sufficient information for accurate prediction of three-state protein stability changes caused by amino acid substitutions.

Keywords: Amino acid substitutions, three-state protein stability prediction, sequence features, support vector machines

1 Introduction

Disease-causing sequence changes have been identified for many human genes [1]. The type of changes varies in nature, affecting multiple mechanisms from RNA processing to post-translational modifications. However, it has been identified that the most common effect of these changes is on protein stability [1–3]. An analysis of human single nucleotide polymorphisms (SNPs) revealed that ~80% of disease-causing amino acid substitutions affect protein stability [4]. In humans, ~60% of single nucleotide missense mutations in coding regions account for monogenic diseases based on the Human Gene Mutation Database [1]. Thus, understanding the effects of sequence variations on proteins is an important task.

Variations in the amino acid sequence can impact the physicochemical characteristics of a polypeptide. These sequence changes may alter the biochemical properties of the protein, resulting in modified structural characteristics with deleterious effects. The notion of predicting the effect of amino acid substitutions on protein stability has been previously explored [5, 6]. However, our approach, similar to [7, 8] and more recently [9], proposes to focus on sequence-based features and avoid using any features based on the protein structure.

The development of a three-state protein stability predictor was previously reported [10]. However, the approach made use of structural features which require a known protein structure. This can be a limiting factor as not all proteins have structural information available. Although computational methods are available for protein structure prediction [11], they are not as accurate in resolution as the experimental determination of a protein structure. The limitation of previous works in the use of structural features may be circumvented by encoding the instances with features derived completely from amino acid sequence.

In this study, we have developed a new three-state protein stability predictor based on sequence features. The importance of the sequence features in model performance was examined using two feature selection methods capable of reducing and ranking variables, random forests and recursive feature elimination. Our three-state protein stability predictor based on Support Vector Machines (SVMs) showed performance comparable to the currently available methods using structural information. The results suggest that sequence features can provide useful information for predicting the effect of amino acid substitutions on protein stability.

2 Methods

2.1 Data acquisition

The protein dataset was derived from that used by Capriotti et al [10] and later modified by Folkman et al [12]. We modified this dataset to contain 68 unique protein entries whose sequences were obtained in FASTA format from the Protein Data Bank (PDB)
The sequences were subsequently subjected to redundancy reduction with a threshold of 85% sequence similarity using BlastClust. This assured the uniqueness of each sequence and eliminated the presence of multiple chains for a single entry. This process also resulted in the condensation of small sequences into larger clusters. The resulting dataset contained 1,332 non-redundant instances (henceforth s1332) with information about the PDB identifier, the wild-type position, the mutated amino acid, the substituted position, and the free energy change ($\Delta\Delta G$), which was used to determine the class label of the instance. In this study, we had three classes or states: decreased stability (DS), increased stability (IS), and no significant change (NC). The $\Delta\Delta G$ thresholds used in this study were as follows: DS if the effect of an amino acid substitution on the protein stability change $\Delta\Delta G \leq -0.5$ kcal/mol, IS if $\Delta\Delta G \geq 0.5$ kcal/mol, or NC if $-0.5$ kcal/mol $\leq \Delta\Delta G \leq 0.5$ kcal/mol. For building an independent test dataset, we selected the s238 sequence set, and subjected it to the same process as the s1332 dataset. This test dataset contained unique entries that were not included in the training dataset s1332, and consisted of variants representing the three different states.

### 2.2 Sequence-based features

The instances in the s1332 dataset were encoded with a total of 31 sequence features using the R package “Peptides” [14]. This study used various physicochemical and biochemical features as defined by ExPASy in ProtParam (http://web.expasy.org/protparam/protparam-doc.html). These features include molecular mass, amino acid composition, charge, and aliphatic index, defined as the relative volume occupied by aliphatic side chains (Alanine, Valine, Isoleucine, and Leucine). Other features include: the Kidera factors, which comprise 10 features derived from 188 physical properties; three different hydrophobicity indices, obtained with three different scales (Fasman, Bull, Chothia); instability index, based on dipeptide composition; and the Boman index, which indicates the potential of a peptide to bind to the membrane or other proteins.

### 2.3 Model construction

In this study, the multiclass protein stability model was constructed using Support Vector Machines (SVMs) with sequence-based features. SVMs were shown to produce well-performing models for protein stability prediction in previous studies [7,8]. In its essence, an SVM model defines a separation hyperplane that divides the space into two distinct halves. Based on the sign given by the $f(x)$, a point will be assigned to a given side of the hyperplane. If $f(x)>0$, a point will be assigned to the positive side of the hyperplane. The soft margin can increase the performance of the classifier when compared to hard margins, and this is achieved by allowing misclassification of some points [15]. The use of soft margins in SVMs comes as a response to the fact that not all data is linearly separable, which is especially true with biological data [15]. The third component refers to kernels, which can be used to make the calculation process more efficient, especially in feature spaces of high dimensionality. The radial basis function (RBF) kernel was used to construct the SVM model in this study. RBF is one of the most widely used kernel functions in model development and often performs well on biological data [15].

The caret and e1071 packages [16,17] available in R were used to construct the SVM model. We examined multiple options for model construction, including multiclass SVM models, kernel functions, and performance metrics. The Receiver Operating Characteristic (ROC) curve, specifically the area under the curve (ROC-AUC), was calculated using the package pROC [18]. The SVM model was also compared with two different Random Forest (RF) classifiers, a single RF and an ensemble of RFs. The RF learning algorithm was previously shown to perform well for protein stability prediction [19].

### 2.4 Feature selection

Feature or variable selection for a classification problem can have two main objectives: (1) to identify highly important variables that are related to the response variable, and (2) to reduce the feature space to improve the prediction of the class label [20]. An efficient feature selection method can not only achieve these objectives, but may also combine important components such as determining importance thresholds, variable ranking and stepwise introduction of variables into the feature set [21, 22]. In this study, we implemented a Random Forest (RF) based feature ranking method [21]. RFs are built upon decision trees, with every node being a condition on a variable. We also tested the recursive feature elimination method [22], in which the lowest 20% features were eliminated in each round to determine the most important sequence-based features.

### 2.5 Model performance evaluation

Model performance was evaluated using the R package performanceEstimation [23]. We used the
tenfold cross-validation method. Briefly, this method consists of partitioning the data into different folds where one-fold is used for testing and the remaining folds are used for training. We also used the independent test dataset s238 to evaluate model performance. The following metrics were used in this study to measure model performance:

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

The accuracy provides the information regarding the true positives (TP) and true negatives (TN) in the total of the dataset, which also includes the identified false positives (FP) and false negatives (FN). Sensitivity or true positive rate refers to the proportion of positive instances that were properly identified in each class, whereas specificity or true negative rate is the number of negative instances identified as such. The Matthews Correlation Coefficient (MCC) was also used in this study to measure the performance of multi-class models. For a model with three classes (A, B and C), the values for TP, TN, FP and FN can be calculated as described previously [24]:

\[ TP = TP_A + TP_B + TP_C \]  
\[ TN = TN_A + TN_B + TN_C \]  
\[ FP = FP_A + FP_B + FP_C \]  
\[ FN = FN_A + FN_B + FN_C \]

3 Results

3.1 Prediction of three-state protein stability changes

To develop an accurate model for three-state protein stability prediction, we tested two widely used machine learning algorithms, Support Vector Machine (SVM) and Random Forest (RF). The SVM and RF models were constructed using 31 sequence features. As shown in Table 1, the SVM method outperformed the RF learning algorithm for predicting protein stability changes. The SVM model that was fine-tuned with the training parameters (C = 15, \( gamma = 0.40 \)) achieved higher performance measures than the RF models in the tenfold cross-validation. In this study, we constructed two RF models: a single RF model and an ensemble comprising of three RFs (RF-E). It was previously shown that an ensemble could result in improved model performance [25]. However, the RF ensemble achieved similar performance measures as the single RF model for protein stability prediction (Table 1). We thus selected the SVM model for further analyses.

3.2 Selection of relevant sequence features for model construction

Feature selection was conducted to determine the impact of sequence features in the model’s ability to discriminate the three protein stability states, and to potentially enhance the model performance. The first feature selection method ranked the importance of the variables (sequence features) using the Gini index [22,26]. The approach based on Boruta [26] sets the variable selection threshold based on the value of shadow attributes, which are shuffled copies of all attributes to create randomness, culminated in the reduction of the feature set from 31 to 12 features. The second method, recursive feature elimination, identified a set of 7 features. Interestingly, the two feature selection methods identified several common features, including the hydrophobicity indices of Fasman and Chothia, some of the Kidera factors, and the aliphatic index. The first method also identified additional features associated with the putative overall effect of amino acid substitutions on the peptide, such as the isoelectric point.

However, the SVM models constructed with the selected features did not show improved performance in the tenfold cross-validation (Table 2). The SVM model using all the 31 sequence features (SVM_Full) achieved higher accuracy, ROC-AUC and MCC than the two models after feature selection (SVM_12 and SVM_7). One possibility was that the model SVM_Full might be slightly overfitted. To examine this possibility, we further compared the model performance using an independent test dataset.

3.3 Model performance evaluation using an independent test dataset

When compared using an independent test dataset (s238), the SVM models constructed with the selected features appeared to give slightly better performance measures than the model with the full feature set (Table 3). In particular, the SVM model using the 12 selected features (SVM_12) achieved slightly better ROC-AUC and MCC than the other two models (SVM_Full and SVM_7). The results suggest
that the 12-feature set might be optimal for predicting three-state protein stability changes. Using the full set of 31 sequence features could cause model overfitting, whereas the 7-feature set might not provide sufficient information for prediction.

4 Conclusions

In this study, we have developed a new model for three-state prediction of protein stability changes caused by amino acid substitutions. The SVM model was built with sequence-based features. Feature selection identified 12 features for accurate protein stability prediction. We further validated the predictive performance of the model using an independent test dataset. Our results suggest that sequence features can provide sufficient information for predicting the effect of amino acid substitutions on protein stability.

Table 1. Performance of the SVM and RF models based on tenfold cross-validation. The models were constructed using 31 sequence features.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>MCC</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.8972</td>
<td>0.7038</td>
<td>0.8659</td>
<td>0.8411</td>
<td>0.8907</td>
</tr>
<tr>
<td>RF</td>
<td>0.7654</td>
<td>0.5360</td>
<td>0.7938</td>
<td>0.6907</td>
<td>0.8453</td>
</tr>
<tr>
<td>RF-E</td>
<td>0.7666</td>
<td>0.5163</td>
<td>0.7826</td>
<td>0.6814</td>
<td>0.8348</td>
</tr>
</tbody>
</table>

Table 2. Performance of the SVM models after feature selection based on tenfold cross-validation. SVM_Full was constructed using all the 31 sequence features; SVM_12 was constructed with 12 selected features; and SVM_7 was constructed with 7 selected features.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>MCC</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM_Full</td>
<td>0.8972</td>
<td>0.7038</td>
<td>0.8659</td>
<td>0.8411</td>
<td>0.8907</td>
</tr>
<tr>
<td>SVM_12</td>
<td>0.8827</td>
<td>0.6115</td>
<td>0.8204</td>
<td>0.7830</td>
<td>0.8578</td>
</tr>
<tr>
<td>SVM_7</td>
<td>0.8439</td>
<td>0.5529</td>
<td>0.7902</td>
<td>0.7445</td>
<td>0.8360</td>
</tr>
</tbody>
</table>

Table 3. Predictive performance of the SVM models on an independent test dataset. SVM_Full was constructed using all the 31 sequence features; SVM_12 was constructed with 12 selected features; and SVM_7 used 7 selected features.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>MCC</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM_Full</td>
<td>0.7423</td>
<td>0.3682</td>
<td>0.6647</td>
<td>0.5731</td>
<td>0.7563</td>
</tr>
<tr>
<td>SVM_12</td>
<td>0.7477</td>
<td>0.4555</td>
<td>0.7052</td>
<td>0.6120</td>
<td>0.7984</td>
</tr>
<tr>
<td>SVM_7</td>
<td>0.7367</td>
<td>0.4452</td>
<td>0.7092</td>
<td>0.6235</td>
<td>0.7950</td>
</tr>
</tbody>
</table>
5 References


