Prediction of Thermophilic Protein with Pseudo Amino Acid Composition: An Approach from Combined Feature Selection and Reduction

De Wang, Liang Yang, Zhengqi Fu and Jingbo Xia*

College of Science, Huazhong Agricultural University, Wuhan, P. R. of China

Abstract: Prediction of thermophilic and mesophilic protein plays a crucial role in both biochemistry and bioengineering. In this study, a different mode of pseudo amino acid composition (PseAAC) was proposed to formulate the protein samples by integrating the amino acid composition, the physiochemical features, as well as the composition transition and distribution features, where each of the protein samples was represented by a numerical vector through the sequence-based approach. Using the support vector machine algorithm, an accurate and reliable classifier was constructed to predict the thermophilic and mesophilic proteins. Moreover, three feature reduction algorithms were obtained for locating the most vital features and reducing the size of feature space. Among the three feature reduction algorithms, the genetic algorithm performed best. Finally, with the reduced features extracted from the genetic algorithm, it was observed that for the selected dataset the new classifier achieved a high accuracy of 95.93% with the Matthews correlation coefficient of 0.9187.

Keywords: Amino acid composition, classifier, feature reduction, genetic algorithm, support vector machine.

1. INTRODUCTION

Temperature of the environment is crucial to the cellular lives. According to the optimum growth temperature (OGT), organisms are categorized into: mesophilies, which live at temperatures below 50°C, and thermophilies, OGT of which is 50°C or above [1]. Protein thermostability is an important issue in basic research and it also throws light upon the mechanism of protein folding and the relationship of protein structure to function, which in turn contributes to the design of thermostable proteins [1] and high temperature biocatalysts [2].

By using features based upon sequence information, including the amino acid composition (AAC) and dipeptide composition, sequence-based approach has been widely used [3-8]. Gromiha [1] found the charged and hydrophobic occurred frequently in thermophiles, while Zhang [3] found the dipeptide composition contains more information than AAC does. Furthermore, Zhou et al. [4] summarized an increase in hydrophobicity and charged residues and a decrease in uncharged polar residues. In addition, Wu [5] proposed that (E+F+M+R)/residue and charged/non-charged residues were critical to the thermostability.

To avoid losing many important information hidden in protein sequences, the pseudo amino acid composition (PseAAC) was proposed [6, 7] to replace the simple amino acid composition (AAC) for representing the sample of a protein. For a summary about its development and applications, such as how to use the concept of Chou’s PseAAC to develop 16 different forms of PseAAC, including those that are able to incorporate the functional domain information, GO (gene ontology) information, cellular automation image information, sequential evolution information, among many others, see two recent comprehensive reviews [8, 9]. Ever since the concept of PseAAC was introduced, various PseAAC approaches have been stimulated to deal with different problems in proteins and protein-related systems [6-40]. Here, we are to consider a different approach of PseAAC in hope to further strengthen the power of PseAAC.

Lin [41] predicted thermophilic proteins with selected AAC and achieved 93.8% for thermophilic and 92.7% for mesophilic proteins. Even slight change of mutations could affect the protein thermostability [42].

Meanwhile, physical chemical features (PC) are discovered to be important features, including extinction coefficient, molecular weight, aliphatic index, grand average of hydropathy, the length of protein chain [43]. As an analogue approach, Inna [44] proposed a composition transition and distribution (CTD) descriptor for the prediction of protein folding class and got an average accuracy of 71.7% and 95% in predicting positive and negative proteins.

In all, these features are all considered to be significant factors in protein thermostability. Though feature extraction methods have got a lot of success in function gene prediction, feature reduction is still an important issue. Actually, the dimension of the biological data is often very high and redundancy in data and interaction between them will decrease the discriminating accuracy. By feature reduction, higher accuracy can be obtained [41].

Meanwhile, taking methodology into account, another focus is put on the choice of discrimination tool. Actually, support vector machine (SVM) now is a popular and prevailing classifier to discriminate positive and negative samples, which was also used in predicting membrane protein type [30], protein subcellular location [14, 45], protein structural
class [10, 13, 26, 29], apoptosis protein location [11], membrane protein type [12], enzyme subfamily classes [15], specificity of GalNAc-transferase [31], HIV protease cleavage sites in protein [34], protein signal sequences and their cleavage sites [35], catalytic triads of serine hydrolases [37], etc.

According to a recent review [9], to develop a useful predictor, the following things were often needed to be considered: (1) benchmark dataset construction or selection, (2) protein sample formulation, (3) operating algorithm (or engine), (4) anticipated accuracy, and (5) web-server establishment.

In this paper, a different mode of PseAAC was proposed and a SVM-based classifier is constructed for predicting the thermophilic proteins. Three kinds of protein features are extracted, including AAC, CTD and physic chemical features. Furthermore, combinatorial experiments are designed and carried out by using these features. The accuracy and MCC reaches 95.69% and 0.9143, respectively. For the purpose of improving discriminating capability, feature reduction was carried out using three different strategies: filter method, relief algorithm, genetic algorithm (GA). After that, 30 vital features are found, and new classifiers based on the reduced features get better performance. The accuracy and Matthews correlation coefficient (MCC) reaches 95.93% and 0.9187, respectively, which show the reliability of the algorithm in feature extraction and feature reduction.

2. MATERIALS AND METHODS

2.1. Data Set

The original data set contains 8648 thermophilic and 209 mesophilic protein sequences retrieved from NCBI (http://www.ncbi.nlm.nih.gov/). It is noted that the motive of this research is not to try distinguishing all of the thermophilic and mesophilic proteins, but to discuss the effectiveness of the combined or reduced numerical feature corresponding to thermophilic protein. Henceforth, only part of protein sequences are chosen as the experimental data, i.e., 209 thermophilic protein and 209 mesophilic ones are chosen from the dataset so as to keep the balance of the training data.

To remove the homologous sequences from the benchmark dataset, a cutoff threshold of 25% was imposed in [46, 49-51] to exclude those proteins from the benchmark datasets that have equal to or greater than 25% sequence identity to any other in a same subset. However, in this study we did not use such a stringent criterion because the currently available data do not allow us to do so. Otherwise, the numbers of proteins for some subsets would be too few to have statistical significance.

2.2. Feature Extraction

Prior to statistics learning, each protein should be converted into a numerical vector which reflects sufficient biological information. To achieve this goal, three kinds of different features are extracted as follows.

The first kind of features is based on AAC. Here, the feature vector is denoted as \( x_i = (x_{i1}, x_{i2}, \ldots, x_{i20}) \), where \( x_i(i = 1, 2, \ldots, 20) \) is the composition of each amino acid. In addition, dipeptide compositions map to the other 400 dimensions, where \( X_2 = (x_{11}, x_{12}, \ldots, x_{400}) \). Thus, a 420 dimension vector is retrieved.

The second kind of features, consisting of four physic chemical (PC) features, is calculated by a protein analysis and identification tools on the ExPASy server (http://expasy.org/tools/protparam.html). The following four PC features are considered to be the vital properties which might affect the protein thermophility.

1. Extinction Coefficient:

   \[ \text{Value} = N(Y) \cdot \text{Ext}(Y) + N(W) \cdot \text{Ext}(W) + N(C) \cdot \text{Ext}(C) \]  

   where \( N(Y), N(W), N(C) \) represent the number of Tyr, Trp, and Cystine in the protein, respectively. Here, Ext(Y) = 1490, Ext(W) = 5500, Ext(C) = 125.

2. Molecular weight:

   \[ \text{MW} = \sum_{i=1}^{20} w_i n_i, \]  

   where \( w_i \) represents the relative molecular weight of an amino acid, \( n_i \) represents the number of the specified amino acid.

3. Aliphatic index:

   \[ \text{Value} = X(A) + a \cdot X(V) + b \cdot (X(I) + X(L)), \]  

   where \( X(A), X(V), X(I), X(L) \) are mole percent (100 x mole fraction) of alanine, valine, insoleucine, and leucine. The coefficients \( a \) and \( b \) are the relative volume of valine side chain (\( a = 2.9 \)) and of L/I side chains (\( b = 3.9 \)) to the side chain of alanine.

4. Grand Average of Hydrophathy (GRAVY):

   \[ \text{Value} = \sum_{i=1}^{20} N_i V_i / L, \]  

   where \( N_i \) and \( V_i \) represent the composition and hydrophy values of each amino acid in the protein respectively and \( L \) is the length of the protein.

   The third ones are CTD features, which mean the global protein sequence descriptors including composition (C), transition (T), and distribution (D) [44]. First, 20 kinds of amino acids are divided into the following three groups based on hydrophobicity: Arg, Lys, Glu, Asp, Gln, and Asn as polar; Gly, Ala, Ser, Thr, Pro, His, and Tyr as neutral; and Cys, Val, Leu, Ile, Met, Phe, and Trp as hydrophobic. The three descriptors were used to reflect features of protein, i.e., C represents the percent compositions of polar, neutral, and hydrophobic residues in the protein; T characterizes the transition frequency in which one group is followed by another or inversely; and D represents the position of the first, 25%, 50%, 75%, and 100% amino acid of the three groups, so that D has five numbers for each group. By this way, a 21-
dimension vector \((CTD_1, CTD_2, \ldots, CTD_d)\) is formed by the three descriptors for each given protein sequence.

### 2.3. Support Vector Machine

After the feature extraction procedure, protein sequence become numerical vector and information hidden behind the data could be found by training and testing phase in classification through machine learning. The mechanism of SVM is a classifier based on statistical learning. Training samples are mapped from the input space \(R^d\) into high dimensional space and a hyper plane is obtained. The trained samples are separated by the chosen hyper plane with maximum distance among different samples and the related hyper plain.

Radial basis function (RBF) is a typical kernel function with one parameter \(\gamma\):

\[
K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / \gamma^2}.
\]

For the selected kernel function, the learning task is to solving the following convex quadratic programming (QP) problem,

\[
\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j),
\]

Subject to:

\[
0 \leq \alpha_i \leq C, \sum_{i=1}^{n} \alpha_i y_i = 0,
\]

where labels, \(y_i = +1, -1\), stand for the positive label and negative label, respectively.

Tool package used here is LIBSVM, developed by Zhihen Lin. Implicit usage information of SVM could be found in [29, 30].

### 2.4. Feature Reduction Algorithm

In order to figure out the top features and improve the classifier accuracy, three feature reduction methods are applied, i.e., filter method, relief algorithm and genetic algorithm.

#### 2.4.1. Filter Method

A fisher criterion score \(F\) can be calculated using the following formula for each feature.

\[
F_i = \frac{\mu_i^+ - \mu_i^-}{\delta_i^+ + \delta_i^-},
\]

where \(\mu_i^+, \mu_i^-\) represent the mean value of the \(i_{th}\) feature in the positive samples and negative samples, respectively, and \(\delta_i^+, \delta_i^-\) represent the corresponding standardized deviation. The value of \(F_i\) reflects the importance of the \(i_{th}\) feature independently to the protein thermostability.

#### 2.4.2. Relief Algorithm

Relief algorithm is firstly proposed by Kira and Rendell [46]. The core idea of the algorithm is to iteratively update each feature’s weight based on their discriminating ability between the two patterns. In each iteration, the weight of the \(i_{th}\) feature is updated according to the following formula:

\[
w_i = w_i + \left| x' - NM' (x) \right| + \left| x' - NH' (x) \right|,
\]

where \(NH\) means the nearest hit and \(NM\) means the nearest miss.

#### 2.4.3. Genetic Algorithm

Genetic algorithm is a modern biological modelling intelligence optimization algorithm. Key issues involved are coding and decoding method and the definition of fitness value.

Coding and Decoding: Each individual is represented by a sequence \(x_1 x_2 \ldots x_n\), \(x_i \in \{0,1\}\), where \(x_i = 1\) denotes the \(i_{th}\) feature is counted in, while \(x_i = 0\) means not chosen, indicating a subsets of the feature space.

Fitness value: Each individual’s fitness value is evaluated by the following formula:

\[
f_i = \frac{k \cdot r_{ff}}{\sqrt{k + k(k-1) r_{ff}}}.
\]

where \(k\) represents for the size of the subsets, and \(r_{ff}\), \(r_{rf}\) stand for average feature-class correlation and average feature-feature inter correlation, respectively.

### 3. RESULT AND DISCUSSION

#### 3.1. Evaluation Criteria

In statistical prediction, the following three cross validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test [47]. However, as elucidated in [9, 48] and demonstrated by Eq.50 of [49] or Eqs.28-32 of [2], among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset. Specifically, each protein is singled out for testing and remaining proteins are used for training, and hence it has been increasingly used by investigators to examine the accuracy of various predictors (see, e.g., [10-22,47-59]).

The performance of each experiment is measured by the following benchmark criteria, sensitivity (Sens), specificity (Spec), accuracy (Accu) and Matthew’s correlation coefficient:

\[
\begin{align*}
\text{Sens} &= \frac{TP}{TP+FN}, \\
\text{Spec} &= \frac{TN}{TN+FP}, \\
\text{Accu} &= \frac{TP+TN}{TP+TN+FP+FN}, \\
\text{MCC} &= \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}.
\end{align*}
\]

where TP is true positive, FN represents false negatives, TN means true negatives, and FP is false positives.

#### 3.2. Experiment with Full Features

In order to evaluate the efficiency of the extracted features, sole feature and combined features are used separately in the classifier. The performance is shown in Table 1.
Table 1. Experiment Comparison

<table>
<thead>
<tr>
<th>Features Used</th>
<th>Sens (%)</th>
<th>Spec (%)</th>
<th>MCC</th>
<th>Accu (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>74.16</td>
<td>75.60</td>
<td>0.4977</td>
<td>74.88</td>
</tr>
<tr>
<td>CTD</td>
<td>77.03</td>
<td>89.95</td>
<td>0.6755</td>
<td>83.49</td>
</tr>
<tr>
<td>AAC</td>
<td>94.74</td>
<td>96.18</td>
<td>0.9092</td>
<td>95.45</td>
</tr>
<tr>
<td>PC+CTD</td>
<td>75.12</td>
<td>90.91</td>
<td>0.6687</td>
<td>83.01</td>
</tr>
<tr>
<td>PC+AAC</td>
<td>94.74</td>
<td>96.65</td>
<td>0.9140</td>
<td>95.69</td>
</tr>
<tr>
<td>CTD+AAC</td>
<td>92.46</td>
<td>96.17</td>
<td>0.9045</td>
<td>95.22</td>
</tr>
<tr>
<td>PC+CTD+AAC</td>
<td>94.26</td>
<td>97.13</td>
<td>0.9143</td>
<td>95.69</td>
</tr>
</tbody>
</table>

Result in Table 1 shows that sole feature with AAC performs better than PC and CTD does. Furthermore, while take PC and CTD features into account, it also has a rather high accuracy, 74.88% and 83.49%, respectively, and MCC is 0.4977 and 0.6755. This confirms the potential capability of these features.

In general, accuracy and MCC will be well elevated if AAC is considered and combined with sole PC or CTD features. For instance, if combing PC and AAC feature, the accuracy and MCC of the classification increase to 95.69% and 0.9140. While combining CTD and AAC, the results are 95.22% and 0.9045, respectively, and both of which show the robust discriminating capability and the reliability of AAC. In addition, it also implies that the potential discriminating capability of PC and CTD features may be swallowed by the 420-dimension AAC features. Hence it is essential to reduce the feature dimension so as to relieve redundant features and further improve the classifier.

3.3. Experiment of Feature Reduction

In statistical and machine learning, over-fitting happens when there are too many parameters versus too few observations. In line with this, three feature reduction methods are carried out during the following experiment and 30 top important features are obtained. Features in detail are listed in Table 2.

Table 2 shows that AAC performs better than PC and CTD features. In detail, there exist merely four CTD features in the Top30 through filter method, i.e., CTD2-5. While in relief and genetic reduction, only one PC feature and two CTD features are listed in Top30, namely, CTD2, PC4 and CTD10. The result demonstrates that the significance of the AAC features is closely linked to proteins thermostability. This is in consistent with the former study of Zhang [3], who declared the dipeptide composition contains more information.

Table 3 shows that accuracy and MCC with features reduced by the genetic algorithm are much higher than other methods. In detail, MCC and accuracy reaches 0.9187 and 95.93% respectively, which is the best classifier. Besides, the vector has reduced from high dimensions to 30 dimensions and it is beneficial to accelerate our experiments contrast to the former one.

Table 2. Top 30 Features by Feature Reduction

<table>
<thead>
<tr>
<th>Vital Features</th>
<th>Filter</th>
<th>Relief</th>
<th>Genetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I</td>
<td>K</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>AQ</td>
<td>Q</td>
</tr>
<tr>
<td>3</td>
<td>CTD2</td>
<td>PG</td>
<td>I</td>
</tr>
<tr>
<td>4</td>
<td>CTD3</td>
<td>A</td>
<td>K</td>
</tr>
<tr>
<td>5</td>
<td>EI</td>
<td>Q</td>
<td>F</td>
</tr>
<tr>
<td>6</td>
<td>F</td>
<td>QM</td>
<td>Y</td>
</tr>
<tr>
<td>7</td>
<td>IV</td>
<td>QV</td>
<td>AA</td>
</tr>
<tr>
<td>8</td>
<td>K</td>
<td>PA</td>
<td>AD</td>
</tr>
<tr>
<td>9</td>
<td>Y</td>
<td>AM</td>
<td>AQ</td>
</tr>
<tr>
<td>10</td>
<td>YI</td>
<td>CTD2</td>
<td>AS</td>
</tr>
<tr>
<td>11</td>
<td>VY</td>
<td>TQ</td>
<td>RI</td>
</tr>
<tr>
<td>12</td>
<td>CTD4</td>
<td>GA</td>
<td>RK</td>
</tr>
<tr>
<td>13</td>
<td>II</td>
<td>QR</td>
<td>DA</td>
</tr>
<tr>
<td>14</td>
<td>AS</td>
<td>L</td>
<td>DQ</td>
</tr>
<tr>
<td>15</td>
<td>AQ</td>
<td>SQ</td>
<td>EE</td>
</tr>
<tr>
<td>16</td>
<td>EK</td>
<td>AA</td>
<td>EK</td>
</tr>
<tr>
<td>17</td>
<td>SA</td>
<td>AE</td>
<td>GQ</td>
</tr>
<tr>
<td>18</td>
<td>FL</td>
<td>AL</td>
<td>GI</td>
</tr>
<tr>
<td>19</td>
<td>KI</td>
<td>IK</td>
<td>GS</td>
</tr>
<tr>
<td>20</td>
<td>CTD5</td>
<td>GQ</td>
<td>IN</td>
</tr>
<tr>
<td>21</td>
<td>IY</td>
<td>LQ</td>
<td>IV</td>
</tr>
<tr>
<td>22</td>
<td>IK</td>
<td>TD</td>
<td>LY</td>
</tr>
<tr>
<td>23</td>
<td>AV</td>
<td>LP</td>
<td>MI</td>
</tr>
<tr>
<td>24</td>
<td>NE</td>
<td>AC</td>
<td>PA</td>
</tr>
<tr>
<td>25</td>
<td>AL</td>
<td>QP</td>
<td>SA</td>
</tr>
<tr>
<td>26</td>
<td>Q</td>
<td>DR</td>
<td>SQ</td>
</tr>
<tr>
<td>27</td>
<td>YV</td>
<td>DQ</td>
<td>TI</td>
</tr>
<tr>
<td>28</td>
<td>IL</td>
<td>PC4</td>
<td>YV</td>
</tr>
<tr>
<td>29</td>
<td>MI</td>
<td>LD</td>
<td>VY</td>
</tr>
<tr>
<td>30</td>
<td>GA</td>
<td>WH</td>
<td>CTD10</td>
</tr>
</tbody>
</table>

Table 3. Performance of Experiments Using Different Feature Reduction Method

<table>
<thead>
<tr>
<th>Methods</th>
<th>Sens (%)</th>
<th>Spec (%)</th>
<th>MCC</th>
<th>Accu (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter</td>
<td>91.39</td>
<td>93.78</td>
<td>0.8519</td>
<td>92.58</td>
</tr>
<tr>
<td>Relief</td>
<td>89.95</td>
<td>92.82</td>
<td>0.8281</td>
<td>91.39</td>
</tr>
<tr>
<td>Genetic</td>
<td>96.17</td>
<td>95.69</td>
<td>0.9187</td>
<td>95.93</td>
</tr>
</tbody>
</table>
The elevation of prediction capability owes to feature reduction which eliminates the redundancy of features and interaction between them and offers a promising method to deal with high dimensional data.

Figure 1 shows the process of genetic algorithm optimization. The x and y axis represent the iterations and current best solution, respectively. The curve reflects the change of the optimum dynamically.

![Figure 1. Optimizing process of genetic algorithm.](image)

4. CONCLUSION

Feature extraction plays an important role in discriminating proteins thermostability and exploring principles in sequence, structure and function. In this paper, the original data set of protein sequences retrieved from NCBI is converted to numeric vectors by feature extracting, and various experiments are conducted based on the vectors containing AAC, CTD and physicochemical features.

To eliminate the interrelation and interaction of different features, further experiments of feature reduction are designed by using filter, relief, genetic algorithm contrastively. Result implies that CTD and PC features are swallowed by high dimensional ACC features, and it also supports the reliability of sequence based approach.

In addition, the new classifier with vital Top30 features achieves better results. It shows that GA, as an intelligence algorithm, has superiority and great potential capacity in artificial intelligence phase. Also, the method developed in this paper is actually rooted in PseAAC.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors [48, 50, 58], we shall make efforts in our future work to provide a web-server for the method presented in this paper.

ACKNOWLEDGEMENT

This paper is supported by the Doctoral funding of HZAU, 52204-06031 and National Funding of Creative College Experiment, hzau091050429.

The authors would also like to thank the anonymous reviewers for many important and instructive suggestions.

REFERENCES


isolating and prediction of a novel sequence from Psidium guajava.


