A STUDY ON THE EFFECT OF USING PHYSICO-CHEMICAL FEATURES IN PROTEIN SECONDARY STRUCTURE PREDICTION

G. L. JAYAVARDHANA RAMA, M. PALANISWAMI
Dept of Electrical and Electronics Engineering,
The University of Melbourne, Parkville,
Victoria - 3010, Australia
jrgl@ee.unimelb.edu.au and swami@ee.unimelb.edu.au

DANIEL LAI
Dept of Electrical and Computer Systems Engineering,
Monash University, Clayton,
Victoria - 3168, Australia
daniel.lai@eng.monash.edu.au.

MICHAEL W. PARKER
St. Vincent’s Institute of Medical Research,
9, Princes Street, Fitzroy,
Victoria - 3065, Australia.
m Parker@svi.edu.au

Protein structure prediction is a powerful tool in today’s drug design industry as well as in the molecular modeling stage of x-ray crystallography research. This paper proposes a redefined encoding scheme based on the combination of Chou-Fasman parameters, physico-chemical parameters and position specific scoring matrix for protein secondary structure prediction. A new method of calculating the reliability index based on the number of votes and the SVM decision value is also proposed and it has been shown to assist design of better filters. The proposed features are then tested on the RS186 and CBS15 datasets and shown to give better cross-validation results compared to the existing techniques.

1. Introduction

The dependence on experimental methods of protein structure prediction may not yield protein structures fast enough to keep up with the requirement of today’s drug design industry. With the availability of abundant
proteomic data, it has been shown that it is possible to predict the structure through machine learning techniques. As the prediction of tertiary structure from protein sequence is a very difficult task, the problem is usually sub-divided into secondary structure prediction and super secondary structure prediction leading to tertiary structure. This paper concentrates on secondary structure prediction using position specific scoring matrix and physico-chemical properties as features. Secondary structure prediction is based on prediction of the 1-D structure from the sequence of aminoacid residues in the target protein. Several methods have been proposed to find the secondary structure including PHD, PROF-King, PSI-Pred, JPred and SAMT99-Sec.

Recently, significant work has been done on secondary structure prediction using Support Vector Machines. Hua and Sun used SVMs and profiles of the multiple alignments and reported Q3 score as 73.5% on the CB513 dataset. In 2003 Ward et. al reported 77% with PSI-BLAST profiles on a small set of proteins. In the same year Kim and Park reported an accuracy of 76.6% on the CB513 dataset using PSI-BLAST Position Specific Scoring Matrix (PSSM). Nguyen and Rajapakse reported a highest accuracy of 72.8% on RS126 dataset using a two stage SVM. Guo et. al used dual layered SVM with profiles and reported a highest accuracy of 75.2% on CB513 dataset.

In this paper we make use of the Chou-Fasman parameters, physicochemical properties including Kyte-Dolittle Hydrophobicity, Grantham Polarity and Rigidity of Proline and compare it with existing techniques which mainly use only position specific scoring matrix (PSSM) obtained from PSI-BLAST. We investigate the performance when the PSSMs are used with physico-chemical properties as features. We propose a new method to calculate the Reliability Index (RI) based on the number of votes each class receives in combination with the decision value of the SVM classifier. We suggest an improvement to the tertiary classifier proposed by Hu et. al by calculating the posterior probability of SVM decision value.

2. Methods

Non-homologous CB513 and RS126 datasets were used for experiments as these are the most commonly used datasets in the literature. The secondary structure definitions used in our experiments were based on the DSSP algorithm. The 8 to 3 state reduction method used was H to H, E to E and all others to C where H stands for α Helix, E for β Strand and
C for Coil.

We used six parameters derived from physico-chemical properties and probability of occurrence of amino acids in each state. Chou-Fasman conformational parameters \(^ {19}\) (3 parameters), Kyte-Dolittle Hydrophobicity scale \(^ {8}\), Grantham Polarity \(^ {20}\) and Presence of Proline in the vicinity (1 parameter each) were used as the features in this set. Kyte-Dolittle hydrophobicity values and Grantham Polarity values were taken from the Protscale \(^ {8}\) website. The last parameter in the set is used to represent the information of rigidity \(R_i\) due to Proline residues. If a Proline residue is present at a particular position, \(R_i\) is given by 1 otherwise 0. We call this \(D2C-PC\). In the rest of the paper, the term physico-chemical refers to the six features from D2C-PC.

A second set containing position specific scoring matrix (PSSMs) generated by PSI-BLAST \(^ {25}\) using non-redundant (NR) database was used. \(pfilt\) \(^ {5}\) was used to filter the low complexity regions, coiled-coil regions and transmembrane helices before subjecting to PSI-BLAST. After getting the PSSM, a window of length \(w\) was considered around every residue and this is used as a feature for the classifier. PSSM have 20 \(\times\) \(L\) elements where \(L\) is the length of the protein chain. We used the following function to scale the profile values from the range \((-7.7)\) to the range \((0,1)\) \(^ {15}\).

\[
g(x) = \begin{cases} 
0.0 & \text{if } x \leq -5 \\
0.5 + 0.1x & \text{if } -5 < x < 5 \\
1.0 & \text{if } x \geq 5
\end{cases}
\]

where \(x\) is the value of the PSSM matrix. All the values within the window of length \(w\) were considered \(^ {12}\). The final feature length for each residue of this set is \(w\times 20\). We call this \(D2C-PSSM\). The third set comprises of combination of D2C-PC and D2C-PSSM as feature vector and we call this \(D2C-PCPSSM\).

**Support Vector Machines (SVM)**

The Support Vector Machines (SVM) developed by Vapnik \(^ {26}\) has been shown to be a powerful supervised learning tool for binary classification problems. The data to be classified is formally written as

\[
\Theta = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}
\]

\[
x_i \in \mathbb{R}^m, \quad y_i \in \{-1, 1\}
\]

\(^{a}\)http://au.expasy.org/tools/prot-scale.html
The SVM formulation defines a boundary separating two classes in the form of a linear hyperplane in data space where the distance between the boundaries of the two classes and the hyperplane is known as the margin. The idea is further extended for data that is not linearly separable where it is first mapped via a nonlinear function to a possibly higher dimension feature space. The nonlinear function is never explicitly used in the calculation. We note that maximizing the margin of the hyperplane in either space is equivalent to maximizing the distance between the class boundaries. For space constraint, details about SVM formulation has not been included in this paper. We use implementation by Lai et al. \cite{16}, namely the D2CSVM\textsuperscript{b} software based on the heuristic framework \cite{16} for all our experiments. More details about the algorithm can be found in the paper by Lai et al. \cite{16}. The SVM classifier decision surface is given by

$$f(x) = \text{sign} \left( \sum_{i=1}^{n} \alpha_i y_i k(x, x_i) + b \right)$$

(3)

3. Classification

Prediction of secondary structure by the proposed technique reduces to a three class \((H, E, C)\) pattern recognition problem. The idea is to construct six classifiers which include three one vs one classifiers \((H/E, E/C, C/H)\) and three one vs rest classifiers \((H/H, E/E, C/C)\). To adjust the free parameters of binary SVMs, we selected a small set of proteins containing about 20 chains and performed cross validation with different sets of parameters. Based on these experiments, we selected Radial Basis Function (RBF) kernel with \(C = 2.5\) and \(\sigma = 4\) for all six classifiers. We designed three one vs one classifiers \(H/E, E/C\) and \(C/H\) and three one vs rest classifiers \(H/H, E/E, C/C\). An ensemble of classifiers used by Hua and Sun \cite{24} and SVM.Represent method proposed by Hu et al. \cite{9} are combined using our voting procedure. The ensemble of classifiers used is as shown in Figure 1. The classifier which gives absolute maximum value amongst \(H/E, E/C, C/H\) classifiers is used for decision making. However, the output of the SVM classifier is uncalibrated and it is not wise to use it directly for comparison. We can convert the output of SVM to a posterior probability \cite{10,12}. We use Platt’s method to moderate the uncalibrated output to posterior probabilities \(P_i\) which would range between 0 and 1. Finally the classifier is chosen according to eq. 4. \(V_i\) is incremented based on

\textsuperscript{b}http://www-personal.monash.edu.au/~dlai/
classification of the chosen classifier

\[
\arg \max_{i \in \{\text{BE, BC, CH}\}} |P_i - 0.5|
\]  

(4)  

**Reliability Index and Post Processing**

The reliability index (RI) we propose is based on the highest vote the class gets as well as the posterior probability of the one vs rest classifier. The vote \( V_i \) any class can get is in the range \((0,4)\). As discussed earlier, the posterior probability of the winning class \( P \) is in the range \((0,1)\). We define RI by eq. 5.

\[
RI(k) = (0.5 \times V_{ki}/4) + (0.5 \times P_{ki})
\]

(5)

where \( k \) is the residue number and \( i \) represents the winning classifier.

### 4. Evaluation Methods

We use standard \( Q_3 \) accuracy, SOV \(^1\) and Matthew’s correlation coefficients for comparing the proposed technique with the existing results in literature. The procedure described by Rost and Sander \(^2\) was used for calculation of \( Q_3 \) accuracies and Matthew’s correlation coefficients. \( Q_3 \) was calculated as follows:

\[
Q_3 = \frac{\sum A_{ij}}{b} \times 100 \quad \text{where,}
\]

\( A_{ij} = \text{Number of residues predicted to be structure } j \text{ and observed in type } i \)

\( b = \text{Total number of residues in database} \)

Matthew’s Correlation Coefficient \( C_i \) was calculated using

\[
C_i = \frac{p_i u_i - n_i o_i}{\sqrt{(p_i + u_i)(n_i + o_i)(n_i + u_i)(p_i + o_i)}}
\]

(7)

where, \( p_i = A_{ii} \)

\( u_i = \sum_{j \neq i} \sum_{k \neq i} A_{jk} \) for \( i \in \alpha, \beta, \gamma \)

\( o_i = \sum_{j \neq i} A_{ji} \)

\( u_i = \sum_{j \neq i} A_{ij} \)
We also use the Segment Overlap (SOV) Score proposed by Zemla et. al.\(^1\) (SOV99) which is defined in eq. 8

\[
SOV = 100 \times \left[ \frac{1}{N} \sum_{i \in \{H,E,C\}} \sum_{j(i)} \frac{\text{minov}(s_1, s_2) + s_i(s_1, s_2)}{\text{maxov}(s_1, s_2)} \times \text{len}(s_1) \right]
\]  (8)

where \(S(i)\) is the set of all overlapping pairs of segments \((s_1, s_2)\) in conformation state \(i\), len\((s_1)\) is the number of residues in segment \(s_1\), minov\((s_1, s_2)\) is the length of the actual overlap and maxov\((s_1, s_2)\) is the total extent of the segment.

5. Results and Discussion

We tested our method on the RS126 and CB513 datasets. Ten fold cross validation was performed for the RS126 dataset and Seven fold cross validation was performed on the CB513 dataset for all experiments. Three sets of features (D2C-PC, D2C-PSSM and D2C-PCPSSM) were used to evaluate RS126 dataset and D2C-PCPSSM feature for CB513 dataset. Results for datas sets RS126 and CB513 are tabulated in the following tables respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>(Q_3)</th>
<th>SOV</th>
<th>(Q_H)</th>
<th>(Q_E)</th>
<th>(Q_C)</th>
<th>(Q_H)</th>
<th>(Q_E)</th>
<th>(Q_C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim and Park</td>
<td>76.1</td>
<td>79.6</td>
<td>77.2</td>
<td>63.9</td>
<td>81.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nguyen and Rajapakse</td>
<td>72.8</td>
<td>66</td>
<td>66.1</td>
<td>57.8</td>
<td>81.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHD</td>
<td>72.5</td>
<td>74.8</td>
<td>74.5</td>
<td>59.7</td>
<td>61.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JPred</td>
<td>74.8</td>
<td>60.3</td>
<td>60.5</td>
<td>59.7</td>
<td>61.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2C-PC</td>
<td>74.6</td>
<td>70.18</td>
<td>70.29</td>
<td>65.03</td>
<td>79.28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2C-PSSM</td>
<td>76.9</td>
<td>75.2</td>
<td>74.56</td>
<td>68.2</td>
<td>79.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2C-PCPSSM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>(Q_3)</th>
<th>SOV</th>
<th>(Q_H)</th>
<th>(Q_E)</th>
<th>(Q_C)</th>
<th>(C_H)</th>
<th>(C_E)</th>
<th>(C_C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHD</td>
<td>70.8</td>
<td>73.5</td>
<td>72</td>
<td>66</td>
<td>72</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PSIPred</td>
<td>76.5</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>JNet(^2)</td>
<td>76.4</td>
<td>74.2</td>
<td>76.4</td>
<td>63.9</td>
<td>80.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hua and Sun(^2)</td>
<td>73.5</td>
<td>75</td>
<td>69</td>
<td>79</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kim and Park(^5)</td>
<td>76.6</td>
<td>80.1</td>
<td>78.1</td>
<td>65.6</td>
<td>81.6</td>
<td>0.68</td>
<td>0.60</td>
<td>0.56</td>
</tr>
<tr>
<td>Guo et. al.(^})</td>
<td>75.2</td>
<td>80</td>
<td>80.4</td>
<td>71.5</td>
<td>72.8</td>
<td>0.71</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>D2C-PCPSSM</td>
<td>77.9</td>
<td>76.17</td>
<td>77.6</td>
<td>69.8</td>
<td>81.1</td>
<td>0.67</td>
<td>0.65</td>
<td>0.6</td>
</tr>
</tbody>
</table>

\(^1\)SOV99
\(^2\)Results not for CB513 dataset
\(^3\)Matthew's correlation coefficients
Overall we have looked at several aspects of protein secondary structure prediction including the use of physico-chemical properties as features, fast trainable support vector machines, reliable tertiary classifier and calculation of reliability index. From the cross validation experiments it is clear that the use of physico-chemical parameters will improve the performance of secondary structure prediction. As a fair comparison we have experimented with PSSM alone as feature set as well as PSSM along with physico-chemical properties. We found that the improvement in accuracy was about 3% (on RS126 dataset) demonstrating the role played by the physico-chemical properties.

References


Acknowledgements: The authors would like to thank Prof. David Jones for kindly providing the pfam program. We are also grateful to NCBI for access to the *PSI-BLAST* program and the Barton Group for making *CB513* and *RS196* datasets available on the web.