Improved Protein Secondary Structure Prediction Using a Intelligent HSVM Method with a New Encoding Scheme

Haifeng Sui, Wu Qu*, Bingru Yan, LiJun Wang

School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, P.R. China,
hfsui@chinaums.com
*quwu.ustb@gmail.com

Abstract

Prediction of protein secondary structures is an important problem in bioinformatics and has many applications. Successful secondary structure predictions provide a starting point for direct tertiary structure modelling, and also can significantly improve sequence analysis and sequence-structure threading for aiding in structure and function determination. Now many secondary structure prediction methods routinely achieve accuracy (Q3) of about 70%. We believe this accuracy could be further improved by using a hybrid method as essential part of the prediction process. In this article, a hybrid SVM(HSVM) has been used to predict protein secondary structure based on the method of combining physicochemical properties of amino acid residues with position-specific scoring matrices (PSSM) containing evolutionary information. Secondary structure can be predicted at significantly increased accuracy. Using a knowledge discovery theory based on inner cognitive mechanism (KDTICM) method, we have developed a gradually enhanced, multi-layered prediction system to predict protein secondary structure, compound pyramid model (CPM). The results are found to be superior to those produced by other methods with blind test dataset from the CASP9 meeting, including the popular psipred method according to Q3 and SOV99 accuracy. The results show that our method has strong generalization ability.

The CPM website is accessible at http://kdd.ustb.edu.cn/protein_web/.

Keywords: Protein Secondary Structure Prediction, HSVM, Encoding Scheme

1. Introduction

In the field of biochemistry and structural biology, secondary structure is the general three-dimensional form of local segments of biopolymers such as proteins and nucleic acids (DNA/RNA). In proteins, the secondary structure is defined by patterns of hydrogen bonds between backbone amide and carboxyl groups, where the DSSP [1] definition of a hydrogen bond is used. Protein secondary structure prediction is to predict protein secondary structure based only on its sequence, where each amino acid is assigned a structure state, helix (H), strand (E) or coil (C). The prediction of protein secondary structure remains very significant, even if the methods have already comparatively been mature and perfect and development of the algorithms is a far less active area than a decade ago. One of the reasons for this decline in activity is that most of the competing methods have converged on a similar level of performance beyond which they have been unable to improve, and possibly because the level of performance that they achieve is, by bioinformatics standards, exceptionally excellent. This is reflected in the fact that the Critical Assessment of Techniques for Protein Structure Prediction (CASP) [2] competition for protein structure prediction ceased to assess this as an official category to whom correspondence should be addressed some years ago, as has the EVA[3] continuous benchmarking project. Accurate prediction of secondary structure elements from an amino acid sequence remains very useful to biologists in its own right, but it is worth pointing out that it plays an important role in tertiary structure prediction as it can be used to generate templates for tertiary structure prediction, which in contrast is far from solved and continues to be a highly active area of research. Fischer and Eisenberg [4] improved the tertiary structure prediction accuracy from 59.0% to 71.0% by using PHD [5] to predict secondary structures. In Yang and Wang’s paper [6], the tertiary structure prediction accuracy was reduced from 79.0% to 71.9% after switching off the secondary structure prediction in
the prediction procedure. McGuffin and Jones [7] reported that the predicted secondary structure information definitely contributes to a better performance for tertiary structure prediction.

Many approaches have been successfully applied to the prediction of protein secondary structure, such as neural networks [8][9][10][43][44], hidden Markov models[11], support vector machines(SVMs)[12][45], data mining[13][14], dynamic programming[15] and so on. Using a knowledge discovery theory based on inner cognitive mechanism (KDTICM) [16] method, we have developed a gradually enhanced, multi-layered prediction system to predict protein secondary structure, compound pyramid model (CPM). The CPM is composed of four layers of intelligent interface that modified knowledge discovery in databases (KDD*) process [16], integrate hybrid SVM (HSVM), hybrid back propagation neural network (HBP) and so on.

The article is organized as follows: we describe HSVM training dataset, testing data sets, evaluation of prediction accuracy and coding scheme in section2. Section3 provides a detailed introduction to the hybrid SVM method and the compound pyramid model. In section 4, we present the results of numerical experiments and our conclusions. In the final section, we discuss the issue of future research directions.

2. Related Materials

2.1. Training dataset

For development of the HSVM method, 462 proteins from the CB513[17] dataset were selected, removing all entries that were shorter than 30 residues, and those from families that contained few sequences and so did not generate valid PSI-Search alignment profiles. Removing the poor homology sequences may extend the overall average accuracy of any prediction method. However, all the prediction methods studied here were tested on the same set of multiple sequence alignments, which was not used in training the methods. As a consequence, unlike in earlier work, a direct comparison of performance between methods was possible. The 462 training proteins were selected by a stringent definition of sequence similarity. As such, these proteins may be split to generate training and testing sets for prediction, with minimal concern that the test and training sets will be contaminated with proteins of similar sequence. In other words, the HSVM approach was developed through 7-fold cross validated training on a sequence and structure non-redundant dataset derived from the CB513 reduced dataset at the super-family level.

2.2. Testing Data Sets

We used three different datasets to develop and test our novel method:

1. RS126: The original set of 126 non-homologous protein chains proposed by Rost and Sander [18]. No two proteins in the dataset have pair-wise sequence similarity greater than 25% for lengths greater than 80 residues. The dataset contains a total of 24,395 amino acids (this number has varied slightly over the years due to changes and corrections in the PDB) with 32% α-helix, 21% β-sheet and 47% loop.

2. CB513: A dataset of 513 protein chains developed by Cuff and Barton [17] with the aim of evaluating and improving protein secondary structure prediction methods. It is, perhaps, one of the most used independent dataset in this field.

3. CASP9: CASP (Critical Assessment of Techniques for Protein Structure Prediction) is a community-wide experiment for protein structure prediction taking place every two years since 1994. Protein structure prediction refers to the problem of predicting the three-dimensional structure of a protein from its sequence. CASP provides users of structure prediction servers with an opportunity to assess the quality of the various methods and servers. It also provides the research community with an assessment of the state of the art in this field. The meeting facilitates large-scale experiments to assess protein structure prediction methods. Given the complexity of the CPM used for the final predictions (Figure.1), it was essential that the prediction method was “blind tested” on a new set of proteins. The CASP9 proteins would have made for a good blind test. However, at the time this work was completed, the CASP9
experiment was several months old and the alignments would have contained new sequences. In addition, we extracted 102 protein chains studied at the CASP9 meeting from the public web page of the Protein Structure Prediction Center http://predictioncenter.org/.

2.3. Evaluation of Prediction Accuracy

Protein secondary structure can be assigned from experimentally determined tertiary structures by algorithms such as DSSP, STRIDE, or DEFINES. We used the most widely used secondary structure definition, DSSP [1]. DSSP distinguishes 8 categories of secondary structure: H (Alpha-helix), G (3-helix), I (5-helix), E (extended-strand), B (isolated-strand), T (turn), S (bend), and coil (“C”). These 8 structure classes are typically reduced into 3 classes. There are four main methods to perform this reduction process. (1) DSSP: H, G to H; E, B to E; all other states to C ; (2) DSSP: H to H; E to E; all other states to C; (3) DSSP: H, G, I to H; E to E; all other states to C; and (4) DSSP: H, G to H; E to E; all other states to C. In this article, we adopt the strictest method (1), which usually results in lower prediction accuracy than other definitions. Our structure prediction methods are therefore evaluated over three states: alpha-helix, extended-strand and coil (or loop).

Several different measures can be used to assess secondary-structure prediction methods, the most common being Q3 [18] and SOV99 [19]. The Q3 score defines accuracy as the percentage of correctly identified states, as in (1). This measure depends on only three states (helix/strand/coil), hence the name Q3.

\[
Q_3 = \frac{\sum_{i \in \{H,E,C\}} n_i \times 100}{\sum_{i \in \{H,E,C\}} n_{i}}
\]  

(1)

The segment overlap measure (SOV99) evaluates of secondary structure prediction methods by the correct identification of secondary structure segments rather than individual residues. SOV is calculated as:

\[
SOV = 100 \times \left[ \frac{1}{N_{ov}} \sum_{i \in \{H,E,C\}} \frac{\min \text{ov}(s_i, s_i) + \delta(s_i, s_i)}{\max \text{ov}(s_i, s_i)} \times len(s_i) \right]
\]  

(2)

where: \( s_i \) and \( s_j \) are the observed and predicted secondary structure segments; \( len(s_i) \) is the number of residues in the segment \( s_i \); \( \min \text{ov}(s_i, s_j) \) is the length of actual overlap of \( s_i \) and \( s_j \), i.e. the extent for which both segments have residues in state \( i \); \( \max \text{ov}(s_i, s_j) \) is the length of the total extent for which either of the segments \( s_1 \) or \( s_2 \) has a residue in state \( i \); \( \delta(s_i, s_j) \) is the integer value defined as being equal to:

\[
\delta(s_i, s_j) = \min \left( \frac{\max \text{ov}(s_i, s_j) - \min \text{ov}(s_i, s_j)}{\min \text{ov}(s_i, s_j) \times \int(len(s_i)) \times 2 \times \int(len(s_j)) \times 2} \right)
\]  

(3)

2.4. Multiple Sequence Alignment Profiles and Coding Scheme

Prediction from a multiple alignment profiles [20] of protein sequences rather than from a single sequence has long been recognized as a way to improve prediction accuracy. The theoretical basis is that the most reliable way to predict protein secondary structure is by similarity (homology) to a protein of known structure. During evolution, residues with similar physicochemical properties are conserved if they are important to the fold or function of the protein. The sequence alignment of homologous proteins accords with their structural alignment and aligned residues usually have similar secondary structures. Multiple-sequence alignment profiles [21] [22] can be generated using the PSI-Search tool. Eight representative physicochemical properties [23] of amino acids are considered to encode each residue and correlative information is examined in relation to the formation of protein secondary structure. Among the eight physicochemical properties, hydrophobic, hydrogen bond and charge properties are regarded as having relatively large impact on protein secondary structure [24].

Position-specific scoring matrices (PSSM) [25] were constructed from PSI-Search against a non-redundant sequences (NR) database together with evolutionary information from the hydrophobic, hydrogen bond and charge physicochemical properties of each residue and used as input data for the
hybrid SVM (HSVM) approach. We applied the SEG routine [26] to mask out low-complexity sequence regions, coiled coil regions and transmembrane spans. For PSI-Search, an E-value threshold for inclusion of 0.001 and three iterations were applied to search the NR database. All alignments are filtered for redundancy at 75% sequence identity, since this was found to be optimal in earlier work [20]. The position-specific scoring matrix has 20*N elements, where N is the length of the target sequence and each element represents the log-likelihood of a particular residue substitution based on a weighted average of BLOSUM62 [27] matrix scores for a given alignment position in the template. Profile matrix elements in the range [–7, 7] were scaled to the [0, 1].

As in the PHD [5] coding scheme, a sliding window [10] [18] method was used to consider a contiguous sequence of amino acids. In order to allow windows to extend over the N- and C-termini, an additional 21st unit was appended for each residue. The encoding scheme of the hybrid SVM includes the PSSM and the three above-mentioned physicochemical properties. Thus, each residue is encoded by a feature vector of dimension 24*w, where w is the sliding window size. The window is shifted residue by residue through the protein chain.

2.5. KDD* Theoretical Basis

Association analysis, as a data mining method, has been widely applied in biology, finance, telecommunications et al. For general data mining methods, association analysis has relatively better ability of induction, and its result is more interpretable [28]. As a novel association analysis data mining model, KDD* process model, which is based on the double bases cooperation mechanism [29] [30], considers the knowledge discovery system as cognition system, and investigates the process of knowledge discovery from the viewpoint of cognitive psychology. The key point of its research is cognitive activeness of knowledge discovery. The propositions of the heuristic coordinator and the maintaining coordinator, which simulate two features in cognitive psychology respectively, makes automatic discovery of the shortage of knowledge and the real-time maintenance of knowledgebase real. By establishing the specifically equivalent relations between the database and the knowledgebase, from a specific aspect, the double bases cooperation mechanism exposits the potential essences, regulars and complexity, and optimizes the process of knowledge discovery.

3. Compound pyramid model

For non-trivial problems, such as protein secondary structure prediction, general single-method model and simple combinations of prediction models failed to return satisfactory prediction results. Our compound pyramid model (CPM)[31] adopts a gradually refining, multi-hierarchical configuration, in which the different layers focus on independent functions. As a result, this model achieves comparatively higher prediction accuracies. Its configuration is shown in Figure 1. The model combines the hybrid SVM (HSVM), the hybrid back propagation neural network (HBP), the structural association classifier (SAC) [31] [32], the attribute association classifier (AAC) [31] [32] and optimization layer (OL).
Figure 1. The Compound Pyramid Model (Comprehensive analysis layer integrates HSVM method and HBP method, then generates two kinds of result, one of which is denoted as ① and will be directly sent to result database, the other of which is denoted as ② and need further determination by SAC and AAC module, finally get the results by the optimization layer.)

The compound pyramid model is composed of 4 layers, each of which has an independent function but is also closely coordinated. The 4 layers comprise a comprehensive analysis layer, a kernel judgment layer, an assistant judgment layer, and an optimization layer.

In the comprehensive analysis layer, this layer combines the prediction results of the HSVM method and the HBP method, which is based on unanimity rule. When unanimous, every module is of same mind and acting together as one in a given situation. That is to say, the comprehensive analysis layer generates two kinds of result, one of which is denoted as "unanimity" and will be directly sent to final result database, the other of which is denoted as "ambiguous" and need further determination by SAC and AAC module. The final results are calculated by cause-effect relationship in the optimization layer. The HSVM module and HBP module are based on both the structure of the sequence and the physicochemical properties of the individual amino acids.

In the kernel judgment layer, this layer applies KDD* process and M algorithm to mining the RS126 structure database for getting structural classification rules, which is based on KDTICM theory. Then, we construct the SAC module to predict protein secondary structure by structural classification rules. The SAC module is in the core layer of classification, which takes on the classification of data that is hard to judge in the comprehensive analysis layer.

Under the theory environment above, the assistant judgment layer applies KDD* model and M algorithm to mining the RS126 structure and physicochemical properties database for getting attribute classification rules. Then, we construct the AAC module to predict protein secondary structure by attribute classification rules. The AAC module is located in the assistant layer of the CPM. Via association analysis of the physicochemical properties of the amino acids, a refinement rule dataset is created to predict data that the lower layers can not identify.

Under the theory environment above, the optimization layer applies indeterminacy causal induction auto-reasoning mechanism to analysis CB513 structure and physicochemical properties database for getting cause-effect relationship between the sequence and the physicochemical properties of the individual amino acids.

4. Methods

4.1. Structure of the Hybrid SVM Prediction Subsystem

A two-level SVM structure was used in the prediction subsystem [42]. Its configuration is shown in Figure 2. The first level is an SVM tertiary classifier that classifies each residue of each sequence into
the 3 secondary structure classes (H, E, or C). The outputs represent the probability that the residue belongs to that class. Since consecutive patterns are correlated (e.g. a helix contains at least 4 consecutive patterns, and a sheet contains at least 3 consecutive patterns), the second-level SVM tertiary classifier filtered successive outputs from the first layer. The target outputs of the secondary layer are the same as for the first layer. As with the first-level SVM, the secondary layer also uses a SVM tertiary classifier, with each residue assigned to the class with the largest output value.

4.2. Design of the Tertiary Classifier

Before devising the tertiary classifier, we first constructed several SVM binary classifiers, including three one-versus-rest classifiers named H/~H, E/~E and C/~C, and three classifiers named H/E, E/C and C/H which distinguish the sample between each of two states. For example, the classifier H/E is constructed using training samples that possess helices and sheets and classifies the testing sample as being either helix or sheet.

The goal of our machine learning approach to secondary structure prediction is to devise a tertiary classifier with good prediction performance. Thus, the next step is to design a tertiary classifier using the trained binary classifier described above. There are many ways to design a tertiary classifier based on binary classifiers. In the compound pyramid model, we achieved and improved the method proposed by Hua and Sun [33]. This method is based on three one-versus-rest binary classifiers (H/~H, E/~E, C/~C) and three one-versus-one binary classifiers (E/C, C/H, H/E). Three cascade tertiary classifiers, SVM-TREE1 (H/~H, E/C), SVM-TREE2 (E/~E, C/H), and SVM-TREE3(C/~C, H/E), were constructed from pairs of binary classifiers, as shown in Figure 3. For the SVM-MAX-D tertiary classifier, the class for a testing sample is assigned as that corresponding to the largest positive distance to the optimal separating hyper-plane among the SVM-TREE1, SVM-TREE2, and SVM-TREE3 classifiers. The SVM vote classifier combines all six binary classifiers using a simple voting principle: the testing sample will be predicted to be in state i if the majority of the six binary classifiers assign it as state i. The SVM jury uses the jury technique to combine all the results of the tertiary classifiers discussed above.

Figure 2. The two-level architecture of the hybrid SVM (HSVM) in the compound pyramid model.

The HSVM subsystem includes three parts: system input data (physicochemical properties of amino acid residues and position-specific scoring matrix, PSSM), the first level SVM, and the secondary level SVM. System inputs are transformed into a series of 24*13 dimensional vectors using the slide-window method and are then imported into the first level SVM. The first-level SVM outputs a large number of 3-dimensional vectors, each representing the probability that a residue belongs to a particular class. Using
the slide-window method, the outputs of the first-level SVM are transformed into a series of 4*13 dimensional vectors, which are used as the inputs of the second-level SVM. The final decisions are based on the outputs of the second-level SVM.

![Figure 3](image)

Figure 3. The structure of tertiary classifiers: (a) SVM_TREE1, (b) SVM_TREE2 and (c) SVM_TREE3. Each tertiary classifier is made up of two cascaded binary classifiers. For example, SVM_TREE1 will classify a sample as helix (H) if the output of the first binary classifier, H/~H, is larger than 0; otherwise the secondary classifier, E/C, will be used. If the output of E/C is larger than 0, the sample will be classified as sheet (E), otherwise coil (C) will be assigned.

4.3. Parameter Optimization of the Prediction System

It is much simpler to construct a system of SVM binary classifier than a neural network (NN). The choice of an appropriate structure for a NN is dependent on the skill of the designer. However, here we only need to select a kernel function and a regularization parameter C to train the SVM. Our substantial tests show that the RBF (radial basis function) kernel, defined as,

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

is appropriate for complex classification problems. Once the kernel function is selected, the parameter of the kernel function $\gamma$ and the regularization parameter $C$ are optimized based on the process of the previous studies [12] [34]. Namely, for the proper choice of $C$ value and $r$ value of RBF kernel $\exp(-\gamma \|x_i - x_j\|^2)$, the previous studies tested different $r$ and upper bound values of $C$ over their own data sets and selected the pairs which show the best accuracy[12][34]. Similarly, in this study, different $r$ and $C$ pairs are tested to find out the optimum parameter values. Finally, we devise the SVM classifiers with the parameter of the kernel function $\gamma = 0.05$, the regularization parameter $C = 1.0$.

5. Experiment & Conclusion

To test the CPM approach experimentally, we selected the RS126, CB513 and CASP9 datasets, using the standard Q3 and SOV99 measures to evaluate our results. The results for each layer of the CPM are shown in Table 1 and Table 2. The final experimental results are shown in Table 3, Figure 4, Figure 5, Figure 7 and Figure 8.

After an exhaustive search over Internet, we obtained results for the RS126 and CB513 data sets from 6 alternative secondary structure prediction servers for comparison with our CPM prediction server. The details of the prediction servers are shown in Table 4 and the results are summarized in Table 5 and Figure 6. In addition, we selected the results from the 4 best servers for assessment using the CASP9 dataset; these results are shown in Table 6, Figure 7 and Figure 8.

Because non-homologous datasets were selected, homogenous HBP analysis method were not able to make high accuracy predictions of secondary structure. As Table 1 and Table 2 show, in the comprehensive analysis layer, our hybrid SVM (HSVM) approach made independent predictions for more than 55% of all amino acid residues with accuracies of up to 90%. This feature assures the accuracy of the whole model to a large extent.
TABLE 1. Each layer prediction accuracy and scale of CPM on the RS126 data set

<table>
<thead>
<tr>
<th>Module</th>
<th>Accuracy</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehensive Analysis Layer</td>
<td>HSVM</td>
<td>19248/24806 = 77.59%</td>
</tr>
<tr>
<td></td>
<td>HBP</td>
<td>19879/24806 = 80.14%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16937/18646 = 90.83%</td>
</tr>
<tr>
<td>Kernel Judgment Layer</td>
<td></td>
<td>3885/6053 = 64.18%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6053/24806 = 24.40%</td>
</tr>
<tr>
<td>Assistant Judgment Layer</td>
<td>15/107  = 14.02%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>107/24806 = 0.43%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>20731/24806 = 83.57%</td>
</tr>
</tbody>
</table>

TABLE 2. Each layer Prediction accuracy and scale of CPM on the CB513 data set

<table>
<thead>
<tr>
<th>Module</th>
<th>Accuracy</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehensive Analysis Layer</td>
<td>HSVM</td>
<td>115543/146233 = 79.01%</td>
</tr>
<tr>
<td></td>
<td>HBP</td>
<td>123324/146233 = 84.33%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>105106/113638 = 92.49%</td>
</tr>
<tr>
<td>Kernel Judgment Layer</td>
<td></td>
<td>19961/32194 = 62.00%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32194/146233 = 22.02%</td>
</tr>
<tr>
<td>Assistant Judgment Layer</td>
<td>86/401  = 21.45%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>401/146233 = 0.27%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>122344/146233 = 83.66%</td>
</tr>
</tbody>
</table>

Figure 4. The line-symbol graph illustrating the distribution of accuracy (Q3 and SOV99) scores (%) as measured on the RS126 test set. The mean of Q3 is 83.57% and the mean of SOV99 is 79.06%. (Note that the SOV99 score is very sensitive to bias towards alpha helical, beta strand or coil structures and can therefore return very low values even if the Q3 score is high.)

Figure 5. The line-symbol graph illustrating the distribution of accuracy (Q3 and SOV99) scores (%) as measured on the CB513 test set. The mean of Q3 is 83.66% and the mean of SOV99 is 77.04%. (Note that the SOV99 score is very sensitive to bias towards alpha helical, beta strand or coil structures and can therefore return very low values even if the Q3 score is high.)

TABLE 3. CPM’s Prediction result

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Q3 (%)</th>
<th>SOV99 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS126</td>
<td>83.57</td>
<td>79.06</td>
</tr>
<tr>
<td>CB513</td>
<td>83.66</td>
<td>77.04</td>
</tr>
<tr>
<td>CASP9</td>
<td>80.06</td>
<td>77.46</td>
</tr>
</tbody>
</table>
Improved Protein Secondary Structure Prediction using an Intelligent HSVM Method with a New Encoding Scheme
Haifeng Sui, Wu Qu, Bingru Yan, LiJun Wang
International Journal of Advancements in Computing Technology, Volume 3, Number 3, April 2011

TABLE 4. The best six protein secondary structure prediction servers’ details

<table>
<thead>
<tr>
<th>PSSP-Server</th>
<th>Website</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSIPRED</td>
<td><a href="http://bioinf.cs.ucl.ac.uk/psipred/psiform.html">http://bioinf.cs.ucl.ac.uk/psipred/psiform.html</a></td>
<td>two-stage neural network[8]</td>
</tr>
<tr>
<td>SSPro</td>
<td><a href="http://download.igb.uci.edu/sspro4.html">http://download.igb.uci.edu/sspro4.html</a></td>
<td>BRNNs[35]</td>
</tr>
<tr>
<td>SAM-T02</td>
<td><a href="http://compbio.soe.ucsc.edu/HMM-apps/T02-query.html">http://compbio.soe.ucsc.edu/HMM-apps/T02-query.html</a></td>
<td>HMM[31]</td>
</tr>
<tr>
<td>Prof</td>
<td><a href="http://www.aber.ac.uk/~phiwww/prof/">http://www.aber.ac.uk/~phiwww/prof/</a></td>
<td>neural networks and linear</td>
</tr>
<tr>
<td></td>
<td></td>
<td>discrimination[36]</td>
</tr>
<tr>
<td>JPred</td>
<td><a href="http://www.compbio.dundee.ac.uk/www-jpred/">http://www.compbio.dundee.ac.uk/www-jpred/</a></td>
<td>Six different prediction methods[37]</td>
</tr>
</tbody>
</table>

TABLE 5. Secondary structure prediction accuracy of CPM versus that obtained with 9 different protein prediction results using RS126 and CB513 datasets. Note that results obtained on the test set use 8-to-3-state reduction method (1).

<table>
<thead>
<tr>
<th>Method</th>
<th>Q3(RS126)</th>
<th>SOV99(RS126)</th>
<th>Q3(CB513)</th>
<th>SOV99(CB513)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref2[34]</td>
<td>76.10</td>
<td>78.80</td>
<td>76.60</td>
<td>73.50</td>
</tr>
<tr>
<td>Psipred</td>
<td>81.01</td>
<td>76.24</td>
<td>79.95</td>
<td>76.48</td>
</tr>
<tr>
<td>Prof</td>
<td>76.95</td>
<td>71.7</td>
<td>77.13</td>
<td>73.74</td>
</tr>
<tr>
<td>PHD Expert</td>
<td>76.92</td>
<td>72.57</td>
<td>77.61</td>
<td>74.98</td>
</tr>
<tr>
<td>SSPro</td>
<td>77.01</td>
<td>70.24</td>
<td>79.07</td>
<td>74.39</td>
</tr>
<tr>
<td>Jpred</td>
<td>73.82</td>
<td>66.55</td>
<td>73.37</td>
<td>68.03</td>
</tr>
<tr>
<td>SAM</td>
<td>78.81</td>
<td>73.3</td>
<td>78.17</td>
<td>74.01</td>
</tr>
<tr>
<td>Predator[38]</td>
<td>80.06</td>
<td>71.42</td>
<td>80.04</td>
<td>74.88</td>
</tr>
<tr>
<td>Ref[39]</td>
<td>81.65</td>
<td>70.67</td>
<td>80.99</td>
<td>73.37</td>
</tr>
<tr>
<td>CPM</td>
<td>83.57</td>
<td>79.06</td>
<td>83.66</td>
<td>77.04</td>
</tr>
</tbody>
</table>

Figure 6. Accuracy comparison with other research results on the RS126 and CB513 data set.
Figure 7. Histogram comparing the Q3 scores of CPM versus Psipred, Jpred, Bhairpred \(^{40}\) and APSSP2 \(^{41}\) for CASP9 reduced dataset of 102 proteins. The Q3 score is written at the top of each predictor's set of bars.

Figure 8. Histogram comparing the SOV99 scores of CPM versus Psipred, Jpred, Bhairpred \(^{40}\) and APSSP2 \(^{41}\) for CASP9 reduced dataset of 102 proteins. The SOV99 score is written at the top of each predictor's set of bars.

Table 6. Comparison with the results of the Psipred, the Jpred, the APSSP2, the Bhairpred, and our CPM on the CASP9 data set

<table>
<thead>
<tr>
<th>Method/Dataset</th>
<th>Q3 (%)</th>
<th>SOV99 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPM</td>
<td>80.62</td>
<td>77.46</td>
</tr>
<tr>
<td>Psipred</td>
<td>77.14</td>
<td>74.03</td>
</tr>
<tr>
<td>Jpred</td>
<td>75.26</td>
<td>71.9</td>
</tr>
<tr>
<td>Bhairpred</td>
<td>70.53</td>
<td>65.6</td>
</tr>
<tr>
<td>APSSP2</td>
<td>73.61</td>
<td>68.07</td>
</tr>
</tbody>
</table>

According to the results mentioned above, we conclude that the CPM prediction system, which adds HSVM, HBP, the SAC module, the AAC module and OL module, can be applied to predict the secondary structures of proteins in the RS126, CB513 and CASP9 datasets with very good classification accuracy. Meanwhile, these results also indicate that the performance of our algorithm is superior to those of other available protein secondary structure prediction servers, and has stronger generalization ability.

6. Future Research

There are open issues still ahead:
1. To improve the prediction accuracy by evaluating new protein secondary structure prediction methods in the Comprehensive Analysis Layer.
2. To improve the prediction speed by the algorithm parallelization.
3. To create a portal service to access existing web servers for protein second structure prediction. This service would provide users with a single point to access multiple servers.
4. To improve the prediction accuracy of protein tertiary structure by using the results of CPM protein second structure prediction.

7. References

Improved Protein Secondary Structure Prediction using a Intelligent HSVM Method with a New Encoding Scheme
Haifeng Sui, Wu Qu, Bingru Yan, LiJun Wang
International Journal of Advancements in Computing Technology, Volume 3, Number 3, April 2011


[40] Behairpred(http://www.imtech.res.in/raghava/bhairpred/)

[41] APSSP2(http://www.imtech.res.in/raghava/apssp2/).


