# Identification of Essential Residues from Protein Secondary Structure Prediction and Applications

Hsin-Nan Lin, Ting-Yi Sung and Wen-Lian Hsu
*Institute of Information Science*
*Academia Sinica, Taiwan*

## 1 Introduction

Proteins can perform functions when they fold into proper three-dimensional structures. However, since determining the structure of a protein through wet-lab experiments can be time-consuming and labor-intensive, computational approaches are preferable. To characterize the structural topology of proteins, Linderstrøm-Lang proposed the concept of a protein structure hierarchy with four levels: primary, secondary, tertiary, and quaternary. In the hierarchy, protein secondary structure (PSS) plays an important role in analyzing and modeling protein structures, since it represents the local conformation of amino acids into regular structures. There are three basic secondary structure elements (SSEs): α-helices (H), β-strands (E), and coils (C). Many researchers employ PSS as a feature to predict the tertiary structure (Fischer et al., 2001, Gong & Rose, 2005, Meiler & Baker, 2003, Rost, 2001), function (Aydin et al., 2006, Eisner et al., 2005, Ferre & King, 2006, Laskowski et al., 2005), or subcellular localization (Nair & Rost, 2003, Nair & Rost, 2005, Su et al., 2007) of proteins. It is noteworthy that, among the various features used to predict protein function, such as amino acid composition, disorder patterns, and signal peptides, PSS makes the largest contribution (Lobley et al., 2007). Moreover it has been suggested that secondary structure alone may be sufficient for accurate prediction of a protein's tertiary structure (Przytycka et al., 1999).

The majority of methods based on predicted secondary structure use the entire sequence for the prediction of this feature. However, individual residues in a given protein are not equally important; some are essential for folding, structural stability, and function of the protein, whereas others can be readily replaced (Capra & Singh, 2007). Moreover, many proteins contain unstructured regions which do not adopt well-ordered 3D structures (Schlessinger et al., 2007), making their secondary structure status unpredictable. The goals of this work include the identification of *essential residues* within a given protein sequence and their potential use for improvement of performance of other protein prediction problems. In particular, we explored whether it is more advantageous to use structural information using only the predicted essential residues or the entire sequence for prediction of PSS.

In a previous work on PSS prediction (Lin et al., 2005), we proposed a method called PROSP, which utilizes a sequence-structure knowledge base to predict a query protein's secondary structure. The knowledge base consists of sequence fragments, each of which is associated with a corresponding structure profile. The structure profile is a position specific scoring matrix that indicates the frequency of each SSE at each position. The average $Q_3$ score of PROSP is approximately 75%.

In this study we have modified the knowledge base and introduced a new prediction method, call SymPred, to identify of the essential residues. Further we applied SymPred for the prediction of the PSS, and developed a protein function classification method, called ProtoPred. We used ProtoPred to have explored the role of essential residues in protein function predictions, enzyme/non-enzyme classification, and prediction of unstructured regions of proteins.

The major differences between SymPred and PROSP are as follows. First, the constitutions of the knowledge bases are different. Second, the scoring systems of SymPred and PROSP are different. Third, unlike PROSP, SymPred allows inexact matching. Our experiment results show that SymPred can achieve 81.0% $Q_3$ accuracy on a non-redundant dataset, which represents a 5.9% performance improvement over PROSP.

Current PSS prediction methods can be classified into two categories: template-based methods and sequence profile-based methods (Bondugula & Xu, 2007). Template-based methods use protein sequences of known secondary structures as templates, and predict PSS by finding alignments between a query sequence and sequences in the template pool. The nearest-neighbor method belongs to this category. It uses a database of proteins with known structures to predict the structure of a query protein by finding nearest neighbors in the database. By contrast, sequence profile-based methods (or machine learning methods) generate learning models to classify sequence profiles into different patterns. In this category, Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) are the most widely used machine learning algorithms (Ceroni et al., 2003, Cheng et al., 2005, Jones, 1999, Karplus et al., 1998, Kim & Park, 2003, Rost & Sander, 2000, Ward et al., 2003). Template-based methods are highly accurate if there is a sequence similarity above a predefined threshold between the query and some of the templates; otherwise, sequence profile-based methods are more reliable. However, the latter may under-utilize the structural information in the training set when the query protein has some sequence similarity to a template in the training set (Bondugula & Xu, 2007). An approach that combines the strengths of both types of methods is required for generating reliable predictions despite the query sequence is similar or dissimilar to the templates in the training set.

There are significant differences between SymPred and other methods in the two categories. First, in contrast to template-based methods, SymPred does not generate a sequence alignment between the query protein and the template proteins. Instead, it finds templates by using local sequence similarities and their possible variations. Second, SymPred is not a machine learning based approach. Moreover, it does not use a sequence profile, so it cannot be classified into the second category. However, like machine learning-based approaches, SymPred can capture local sequence similarities and generate reliable predictions as well as the essential residues of protein sequences. The experiment results on the two latest independent test sets (*EVA_Set1* and *EVA_Set2*) show that, in terms of $Q_3$ accuracy, SymPred outperforms other existing methods by 1.4% to 5.4%.

## 2 Methods

### 2.1 Structure Conservation

It is well known that a protein structure is encoded and determined by its amino acid sequence. However, the folding process is perhaps the most fundamental unresolved question in biology (Alexander et al., 2007). Fortunately, biologists have found that two proteins with a sequence identity above 40% may have a similar structure and function. The high degree of robustness of the structure with respect to the sequence variation shows that the structure is more conserved than the sequence.

In evolutionary biology, protein sequences that derive from a common ancestor can be traced on the basis of sequence similarity. Such sequences are referred to homologous proteins. In fact, the most successful approaches for predicting protein structures involve the detection of homologous proteins of known structures. These methods rely on the observation that the number of folds appears to be limited and homologous proteins adopt remarkably similar structures (Kelley & Sternberg, 2009). However, some homologous proteins are difficult to identify due to the low sequence identity and they are referred to remotely homologs. For example, the sequence identity between proteins *1aab* and *1j46* is only 16.7% but they are structurally homologous and classified into the same family (*HMG-box*) in the SCOP classification. In such a case, it is difficult to discover the homologous relationship using sequence comparison methods. Profile-profile alignment methods (Pietrokovski, 1996, Rychlewski et al., 2000, Sadreyev & Grishin, 2003, Przybylski & Rost, 2007, Yona & Levitt, 2002) are capable of identifying remote homology; nevertheless, they are relatively slow.

In this study, rather than using global sequence similarities we compile a knowledge base by using local sequence similarities to find useful proteins as templates for structure prediction. We demonstrate that local sequence similarity also possesses structure conservation and it is helpful to predict protein structure.

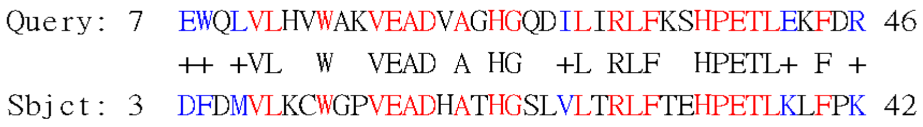
## 2.2 Local Sequence Similarity and Similar Peptides

The local sequence similarity between a pair of sequences represents a sequence similarity that arises within a part of the two sequences, which can be identified by using alignment algorithms such as that of Smith and Waterman (Smith & Waterman, 1981). However, it can not guarantee a structural relationship in a local sequence similarity (Sternberg & Islam, 1990). To assess whether a local sequence similarity implies a homologous relationship, one should estimate the significance of the alignment. Statistically, when performing a PSI-BLAST search (Altschul et al., 1997) for a given query protein, an e-value of 0.001 generally produces a safe searching and signifies sequence homology (Jones & Swindells, 2002). Therefore, we assume that two sequence fragments probably have a homologous relationship if they can be aligned by PSI-BLAST with an e-value of 0.001 at least. Based on the assumption, we can further generate *similar peptides* from a significant alignment between two sequences.

Given a query protein sequence *p*, PSI-BLAST would probably generate a large number of significant local pairwise alignments called *high-scoring segment pairs* (HSPs) between *p* and its similar proteins *sp*. Figure 1 shows an example of HSP with an e-value of 0.001. In the alignment, the identical residues are labeled with letters and conserved substitutions are labeled with plus symbols. The sequence identity between the two sequence fragments in this example is 50% (=20/40). We use a sliding window of size n to scan each HSP and define the similar peptides for *p*. Given an n-gram pattern *w* in *p*, the *similar peptide* of *w* is defined as the other n-gram pattern *sw* in *sp* within the sliding window that is aligned with *w* and both of them are without any gaps. Take the sequence alignment in Figure 1 as an example. The *Sbjct* sequence is a similar protein to the *Query* sequence; therefore, DFDM is deemed similar to the peptide EWQL if the window size is 4, and FDMV is deemed similar to the next peptide WQLV. Based on the observation of high robustness of structures, if the *Query* is of known structure and the *Sbjct* is of unknown structure, we assume that each similar peptide *sw* adopts the same structure as its corresponding peptide *w*; i.e., *sw* inherits the structure of *w*.

Moreover, different similar peptides *sw* for a given peptide *w* should have different similarity scores to *w*. To estimate the similarity between *w* and *sw*, we calculate the *similarity level* according to the number of amino acid pairs that are *interchangeable*. If two amino acids are aligned in a sequence alignment, they are said to be *interchangeable* if they have a positive score in BLOSUM62. Since in this study a peptide is defined as an n-gram pattern, the range of the similarity level between the components of a peptide pair is from 0 to *n*. For example, the similarity level between DFDM and EWQL is 3, and that between FDMV and WQLV is also 3.

In this study, we compile a knowledge base, called *SPKB*, which stands for *Similar Peptide Knowledge Base*. In this knowledge base, we collect a large amount of similar peptides for each protein sequence of known structure in a dataset. We transfer the structural information to the corresponding similar peptides and use the information as the knowledge of sequence to structure for PSS prediction.



```
Query: 7   EWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDR 46
           ++ +VL  W  VEAD A HG  +L RLF  HPETL+ F +
Sbjct: 3   DFDMVLKCWGPVEADHATHGSLVLTRLFTEHPETLKLFPK 42
```

**Figure 1:** A local sequence alignment derived by BLAST. The identical residues are labeled with letters and conserved substitutions are labeled with plus symbols. The alignment in this example shows that the sequence fragment from position 7 to position 46 of the query sequence is very similar to that from position 3 to position 42 in the subject sequence.

## 2.3 Construction of the Knowledge Base SPKB

Given a query sequence, we use PSI-BLAST to generate a number of significant alignments, from which we acquire possible sequence variations. In general, the similar protein sequences (i.e., the Sbjct sequences) reported by PSI-BLAST share highly similar sequence identities (between 25% and 100%) with the query, which implies that the sequences may have similar structures. Therefore, we identify similar peptides in those sequences.

Using a dataset of protein sequences with known secondary structures, we construct a knowledge base, called *SPKB*. The dataset used to construct *SPKB* is described later in the Result section. For each protein *p* in the dataset, we first extract n-gram patterns from its original sequence using a sliding window of size *n*. Each n-gram pattern, as well as the corresponding SSEs of the successive *n* residues, the protein source *p*, and the similarity level (here, the similarity level is *n*), are stored as an entry in *SPKB*. A protein source *p* represents the structural information provider. We then use PSI-BLAST to generate a number of similar protein sequences. Specifically, to find similar sequences, we perform a PSI-BLAST search of the NCBInr database with parameters *j*=3, *b*=500, and *e*=0.001 for each protein *p* in the dataset. Since the NCBInr database only contains protein sequence information, each similar peptide inherits the SSEs of its corresponding word in *p*. A PSI-BLAST search for a specific query protein *p* generates a number of local pairwise sequence alignments between *p* and its similar proteins. Statistically, an e-value of 0.001 generally produces a safe search and signifies sequence homology (Jones & Swindells, 2002). Similarly, each similar peptide and its inherited structure, the protein source *p*, and the similarity level are stored as an entry in *SPKB*.

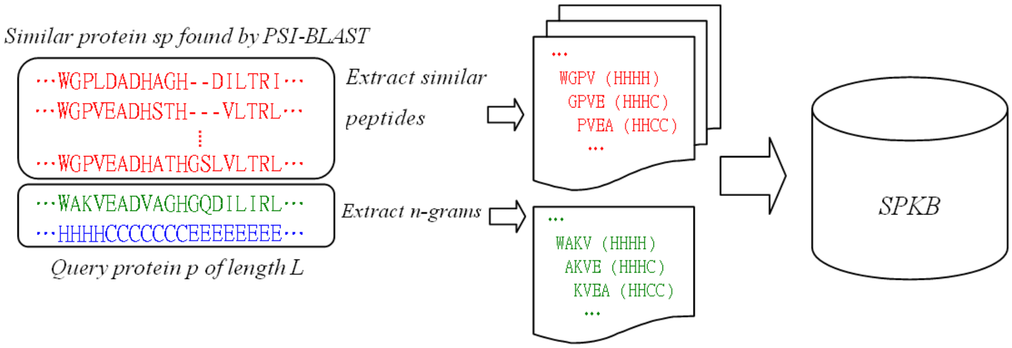

Similar protein sp found by PSI-BLAST
···WGPLDADHAGH--DILTRI···
···WGPVEADHSTH---VLTRL···
⋮
···WGPVEADHATHGSLVLTRL···
···WAKVEADVAGHGQDILIRL···
···HHHHCCCCCCCEEEEEEEEE···
Query protein p of length L
Extract similar peptides
Extract n-grams
···
WGPV (HHHH)
GPVE (HHHC)
PVEA (HHCC)
···
···
WAKV (HHHH)
AKVE (HHHC)
KVEA (HHCC)
···
SPKB

**Figure 2:** The procedure used to extract n-grams and similar peptides for a query protein *p*. The procedure used to extract n-grams and their similar peptides for a given query protein *p* (assuming the window size *n* is 4). We use a sliding window to screen the query sequence and all the similar protein sequences found by PSI-BLAST and extract all n-gram peptides. Each n-gram is associated with a piece of structural information of the region from which it is extracted. The protein source of all the extracted n-gram is the query protein *p,* since all the structural information is derived from *p*.

Figure 2 shows the procedure used to extract similar peptides for a query protein *p*. We use a sliding window to screen the query sequence, as well as all the similar protein sequences found by PSI-BLAST, and extract all n-grams. The query protein *p* is the protein source of all the extracted n-gram patterns. Each pattern is associated with a piece of structural information according to the region from which it is extracted. For example, WGPV is a similar peptide of WAKV. Since it is from a similar protein of unknown structure, it is associated with a piece of structural information of WAKV, which is HHHH.

Note that a similar peptide may appear in more than one similar protein when all similar protein sequences are screened. We cluster identical n-gram patterns together and store the frequency in the similar peptide entry. Table 1 shows an example of a similar peptide entry in *SPKB*. In the example, WGPV is a similar peptide of proteins *A*, *B* and *C*, since it is extracted from the similar proteins of *A*, *B* and *C*. The similar peptide inherits the corresponding structural information of its source, and we can derive the corresponding similarity levels and frequencies via the extraction procedure. For example, the similarity level of WGPV in terms of protein source *A* is 3 and the frequency is 7. This implies that WGPV has 3 interchangeable amino acids with the corresponding protein word of *A* and it appears 7 times among the similar proteins of *A* found in the PSI-BLAST search result.

| **Similar peptide: WGPV** | | | |
|---|---|---|---|
| Protein Source | Secondary Structure | Similarity Level | Frequency |
| *A* | HHHH | 3 | 7 |
| *B* | HHCH | 4 | 11 |
| *C* | CHHH | 2 | 3 |

**Table 1:** An example of a similar peptide entry in *SPKB* (assuming the window size *n* is 4). WGPV is a similar peptide of proteins *A*, *B* and *C*, since it is extracted from the similar proteins of *A*, *B* and *C*. We transfer the structural information of protein sources to the corresponding similar peptides, and calculate the corresponding similarity levels and frequencies. For example, the similarity level of WGPV in terms of protein source *A* is 3 and the frequency is 7.

## 2.4 SymPred: A PSS Predictor Based On SPKB

### 2.4.1 Preprocessing

Given a target protein *t*, whose secondary structure is unknown and to be predicted, we perform a PSI-BLAST search on *t* to compile a set of n-gram patterns containing its original n-grams and similar peptides. The procedure is similar to the construction of *SPKB*. We also calculate the frequency and similarity level of each n-gram in the peptide set.

### 2.4.2 The Scoring Function

Each peptide *w* in the set is used to match against similar peptides in *SPKB*, and the structural information of each protein source in the matched entry is used to vote for the secondary structure of *t*. To differentiate the effectiveness of matched entries, we design a scoring function based on the protein sources in the matched entries and the sum of the weighted scores on the associated structures determines the predicted structure.

Since we use the structural information of protein sources in the matched entries for structure prediction, we define the scoring function based on its similarity level and frequency recorded in the dictionary for the following observation. *The similarity level represents the degree of similarity between a protein n-gram pattern and its similar peptide, and the frequency represents the degree of sequence conservation in the protein's evolution.* Intuitively, the greater the similarity between two peptides, the closer they are in terms of evolution; likewise, the more frequently a peptide appears in a group of similar proteins, the more conserved it is in terms of evolution.

To define the scoring function, we consider the similarity level and the frequency of the peptide in the peptide set of *t*, denoted by $Sim_t$ and $freq_t$ respectively, as well as those of a protein source *i* in its matched entry, denoted by $Sim_i$ and $freq_i$ respectively. Note that $sim_t$ and $freq_t$ are obtained in the preprocessing stage. To measure the effectiveness of the structural information of the protein source *i*, we define the voting score $s_i$ as $min(freq_t, freq_i)\times(1+min(Sim_t, Sim_i))$. The structural information provided by *i* will be highly effective if: 1) *w* is very similar to the corresponding peptides of *t* and *i*; and 2) *w* is well conserved among the similar proteins of *t* and *i*.

Take the similar peptide WGPV in Table 1 as an example. If WGPV is a similar peptide of *t* (assuming $freq_t$ is 5 and $Sim_t$ is 4), then the voting score of the structural information provided by protein source *A* is $min(5, 7)\times(1+min(4, 3)) = 5\times(1+3) = 20$. Similarly, the voting score provided by protein source *B* is $min(5, 11)\times(1+min(4, 4)) = 5\times(1+4) = 25$, and the score provided by protein source *C* is $min(5, 3)\times(1+min(4, 2)) = 3\times(1+2) = 9$. The structural information provided by protein source *B* has highest score in this matched entry and therefore has the most effect on the prediction.

### 2.4.3 Structure Determination

The final structure prediction of the target protein *t* is determined by summing the voting scores of all the protein sources in the matched entries. Specifically, for each amino acid in a protein *t,* we associate three variables, $H(x)$, $E(x)$, and $C(x)$ which correspond to the total voting scores for the amino acid *x* to have structures H, E, and C, respectively. For example, assume that the above similar peptide WGPV is aligned with the residues of protein *t* starting at position 11. Then, protein *A's* contribution to the voting score of $H(11)$, $H(12)$, $H(13)$, and $H(14)$ will be 20. Similarly, protein *B* will contribute a voting score of 25 to $H(11)$, $H(12)$, $C(13)$, and $H(14)$; and protein *C* will contribute a voting score of 9 to $C(11)$, $H(12)$, $H(13)$, and $H(14)$. The structure at *x* is predicted to be *H*, *E* or *C* depending on $max(H(x), E(x), C(x))$. When two or more variables have the same highest voting score, C has a higher priority than H, and H has a higher priority than E.

### 2.4.4 Confidence Level

A confidence measure on a prediction for each residue is important to a PSS predictor, because it reflects the reliability on the output of the predictor. To evaluate the prediction confidence on each amino acid *x*, we calculate a *confidence level* to measure the reliability of the prediction. The confidence level on amino acid *x* is defined as follows.

$$ConLvl(x) = 10 \times \frac{H(x) + E(x) + C(x)}{\sum_{i,t} \left\{ \frac{(freq_t + freq_i)}{2} \times max\left[ 1, \frac{(Sim_t + Sim_i)}{2} \right] \right\}}$$

The range of *ConLvl*(*x*) is forced to be between 0 and 9 by rounding down. In the Results section, we will analyze the correlation coefficient between the confidence level and the average $Q_3$ score, and compare with that of PSIPRED.

## 2.5 SymPsiPred: A Secondary Structure Meta-predictor

SymPred is different from sequence profile-based methods, such as PSIPRED, which is currently the most popular PSS prediction tool. PSIPRED achieved the top average $Q_3$ score of 80.6% in the 20 methods evaluated in the CASP4 competition. SymPred and PSIPRED use totally different features and methodologies to predict the secondary structure of a query protein. Specifically, SymPred relies on synonymous words, which represent local similarities among protein sequences and their homologies; however, PSIPRED relies on a position specific scoring matrix (PSSM) generated by PSI-BLAST, which is a condensed representation of a group of aligned sequences. Furthermore, SymPred constructs a protein-dependent synonymous dictionary for inquiries about structural information. In contrast, PSIPRED builds a learning model based on a two-stage neural network to classify sequence profiles into a vector space; thus, it is a probabilistic model of structural types.

To combine the results derived by the two methods, we compare the prediction confidence level of each residue from each method and return the structure with the higher confidence. Since SymPred and PSIPRED use different measures for the confidence levels, we transform their confidence levels into $Q_3$ scores. For each method, we generate an accuracy table showing the average $Q_3$ score for each confidence level, i.e., we use the average $Q_3$ score of an SSE to reflect the prediction confidence.

For example, suppose SymPred predicts that a residue in a target sequence has structure *H* with a confidence level of 6, PSIPRED predicts that the residue has structure *E* with a confidence level of 6, and the corresponding $Q_3$ scores in the accuracy tables are 77.6% and 64.6% respectively. In this case, SymPsiPred would predict the residue as *H*.

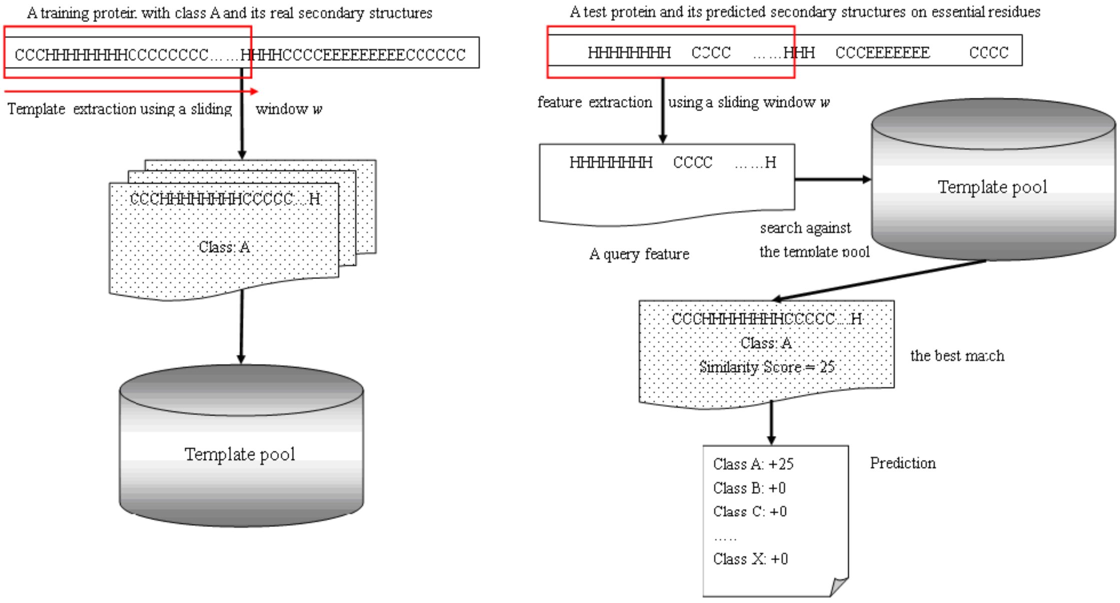
## 2.6 Essential Residues

Since the confidence level measures the ratio of voting scores a residue *x* gets to the summation of the normalization factors, it reflects the degree of sequence conservation in protein evolution. We use the confidence levels representing the degrees of importance of residues in determining the structure and function of a protein sequence.

To study the effectiveness of essential residues, we developed a general prediction method, called ProtoPred, which only uses the secondary structural information as the single feature for general proteome prediction problems, such as function prediction and enzyme/non-enzyme classification. The confidence levels are used as weights to indicate the degrees of importance of residues when finding protein templates for the prediction.

## 2.7 ProtoPred: A Prototype of Prediction Method

Figure 3 shows the main algorithm of ProtoPred. ProtoPred is a simple template based method for general prediction problems. It is a standard query-template alignment algorithm that is used frequently in homology modeling or threading methods (Laskowski et al., 2005, Pandey et al., 2006, Frenkel-Morgenstern et al., 2005).


A training protein with class A and its real secondary structures
CCCHHHHHHHHCCCCCCCC……HHHHCCCCEEEEEEEEEECCCCCC
Template extraction using a sliding window w
CCCHHHHHHHHCCCCC…H
Class: A
Template pool
A test protein and its predicted secondary structures on essential residues
HHHHHHHH CCCC ……HHH CCCEEEEEEE CCCC
feature extraction using a sliding window w
HHHHHHHH CCCC ……H
A query feature
search against the template pool
Template pool
CCCHHHHHHHHCCCCC…H
Class: A
Similarity Score = 25
the best match
Class A: +25
Class B: +0
Class C: +0
…..
Class X: +0
Prediction

(a) Template Extraction (b) ProtoPred's Prediction Procedure

**Figure 3:** The main algorithm of ProtoPred. (a) Template extraction (b) the prediction procedure.

For the training of ProtoPred, we used a sliding window of size *w* to extract the real secondary structure fragments from each of the training proteins. Each structure fragment carried the related information from its origin, such as function labels or protein classes. These fragments were treated as templates for predictions. For test phase we used the same sliding window to extract the predicted secondary structure fragments from the target protein. Each structure fragment (denoted as *s*) was used to search against the template pool. We compared the similarities between *s* and each template *t* in the template pool. The similarity was estimated as follows.

For each position $x$ (from 1 to $w$) if $s(x)$ was identical to $t(x)$, then $t$ would get a weighted score from $s$, i.e., the confidence level of $s(x)$. Each $s$ selects the best template $t$ with the highest sum of weighted scores (denoted as $Sum_{ws}$). If the best template $t$ was labeled as class $A$, then the target protein would get a score of $Sum_{ws}$ for class $A$. Finally, the target protein would be predicted as the class with the highest score.

# 3 Results

## 3.1 Datasets and Experiment Design

We downloaded all the protein files in the DSSP database and compiled subsets of protein chains based on the thresholds of different sequence identities. Then, we generated two datasets, called *DsspNr-25* and *DsspNr-60* with the PSI-CD-HIT program (Li & Godzik, 2006). *DsspNr-25* contains 8297 protein chains, whose mutual sequence identities are below 25%; *DsspNr-60* contains 12975 protein chains, whose mutual sequence identities are below 60%. We use *DsspNr-25* as the validation set to evaluate the performance of SymPred and SymPsiPred, and use *DsspNr-60* as the template proteins to construct the *SPKB*. To predict a target protein $t$ in the test set, we ignore the structural information of protein $p$ in the template pool if $p$ and $t$ share at least 25% of the sequence identity, which we calculate with the ClustalW tool. To evaluate our methods, we use the $Q_3$ score to measure the performance of PSS prediction methods. In our approach, the $Q_3$ score is calculated as follows:

$$Q_3 = \frac{\text{number of residues correctly predicted}}{\text{number of all residues}} \times 100\%$$

We measure the $Q_3$ score for each protein in *DsspNr-25* and take the average Q3 score as the final result to estimate the performance of SymPred and SymPsiPred.

## 3.2 Performance Evaluation of SymPred and SymPsiPred on DsspNr-25

Table 2 shows the prediction performance of SymPred, PSIPRED, PROSP, and SymPsiPred on the *DsspNr-25* dataset. SymPred achieved $Q_3$ of 81.0% and SOV of 76.0%, outperforming PROSP by 5.9% in $Q_3$ and 7.3% in SOV. The meta-predictor, SymPsiPred which integrates the prediction power of SymPred and PSIPRED, achieved a further improvement on $Q_3$ of 83.9% on *DsspNr-25*. This result demonstrates that SymPsiPred can combines the strengths of the two methods and thus yield much more accurate predictions.

| ***DsspNr-25*** | $Q_3$ | $Q_3$**Ho** | $Q_3$**Eo** | $Q_3$**Co** | **sov** | **sovH** | **sovE** | **sovC** |
|---|---|---|---|---|---|---|---|---|
| SymPred | 81.0 | 84.3 | 71.6 | 77.7 | 76.0 | 82.5 | 76.9 | 70.7 |
| PSIPRED | 80.1 | 78.8 | 68.8 | 78.3 | 76.9 | 79.2 | 74.4 | 72.2 |
| SymPsiPred | 83.9 | 81.5 | 75.8 | 83.9 | 80.2 | 82.3 | 80.3 | 76.5 |
| PROSP | 75.1 | 79.7 | 67.6 | 71.3 | 68.7 | 77.0 | 73.0 | 63.4 |

**Table 2:** The prediction performance of SymPred, PSIPRED, PROSP, and SymPsiPred. *Q3*Ho (*Q3*Eo and *Q3*Co, respectively) represents correctly predicted helix (strand and coil, respectively) residues (percentage of helix observed). sovH/E/C values are the specific SOV accuracies of the predicted helix, strand and coil, respectively.

### 3.3 Evaluation of the confidence level

Figure 4 shows the utility of our confidence level and PSIPRED's confidence level in judging the prediction accuracy of each residue in the test set. The statistics are based on more than 2 million residues. The correlation coefficient between the confidence levels and $Q_3$ scores for SymPred is 0.992, and that for PSIPRED is 0.976. Thus, both methods provide strong confidence measures for the output. We observe that a confidence level of 7 or above reported by SymPred is attributed to 53% of the residues with more than 81% of the $Q_3$ accuracy which is comparable to the confidence level of 8 or above reported by PSIPRED. Furthermore, it can be observed that the prediction of SymPred is more reliable when the confidence levels of both methods are low. For example, the average $Q_3$ score of SymPred for the confidence level of 6 is 77.6%, whereas that of PSIPRED is 64.6%.

### 3.4 Performance Comparison with Existing Methods on EVA Benchmark Datasets

EVA test sets usually serve as benchmarks of protein secondary structure predictors, particular for CASP competitions. Only proteins without significant sequence identity to previously known PDB proteins were used to test on different existing methods. We chose two latest EVA sequence-unique subsets of the PDB, called *EVA_Set1* (protein list: http://cubic.bioc.columbia.edu/eva/sec/set_com1.html) and *EVA_Set2* (protein list: http://cubic.bioc.columbia.edu/eva/sec/set_com6.html), the former containing 80 proteins tested on the most number of methods and the latter with the maximum number of proteins (212 proteins). The two datasets serve as independent test sets for performance comparison of SymPred with other existing methods.

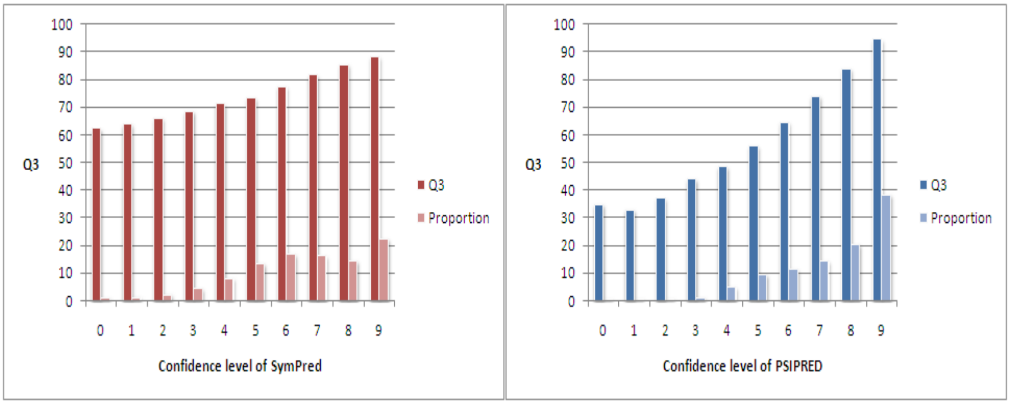


**Figure 4:** Comparison of the $Q_3$ scores and confidence levels derived from SymPred and PSIPRED. The correlation coefficient between the confidence levels and $Q_3$ scores for SymPred is 0.992, and that for PSIPRED is 0.976.

For fair comparison, when predicting the secondary structure of each target protein in an independent set, SymPred discarded the structural information of all proteins sharing at least 25% of the sequence identity with the target protein in the template pool, i.e., SymPred used in the template pool the structural information of proteins sharing no more than 25% sequence identity with the target protein.

Table 3 shows the experiment result on the two benchmark datasets, *EVA_Set1* and *EVA_Set2*. It shows that SymPred achieves $Q_3$ accuracies of 78.8% (SOV=76.4%) and 79.2% (SOV=76.0%), outperforming existing state-of-the-art methods by 1.4% to 5.4%. It can be observed that SymPred performs better than each single predictor on most of performance measurements.

| ***EVA_Set1*** | $Q_3$ | ***ERRsig*** $Q_3$ | **sov** | ***ERRsig*sov** | **sovH** | **sovE** | **sovC** |
|---|---|---|---|---|---|---|---|
| SymPred | 78.8 | ±1.4 | 76.4 | ±1.9 | 85.0 | 76.5 | 70.4 |
| SAM-T99sec | 77.2 | ±1.2 | 74.6 | ±1.5 | 80.9 | 72.5 | 71.2 |
| PSIPRED | 76.8 | ±1.4 | 75.4 | ±2.0 | 82.1 | 72.3 | 69.2 |
| PROFsec | 75.5 | ±1.4 | 74.9 | ±1.9 | 78.3 | 75.9 | 71.3 |
| PHDpsi | 73.4 | ±1.4 | 69.5 | ±1.9 | 73.7 | 73.9 | 65.2 |

| ***EVA_Set2*** | $Q_3$ | ***ERRsig*** $Q_3$ | **sov** | ***ERRsig*sov** | **sovH** | **sovE** | **sovC** |
|---|---|---|---|---|---|---|---|
| SymPred | 79.2 | ±0.9 | 76.0 | ±1.2 | 85.1 | 77.7 | 71.3 |
| PSIPRED | 77.8 | ±0.8 | 75.4 | ±1.1 | 80.6 | 72.6 | 70.4 |
| PROFsec | 76.7 | ±0.8 | 74.8 | ±1.1 | 79.2 | 76.2 | 71.8 |
| PHDpsi | 75.0 | ±0.8 | 70.9 | ±1.2 | 77.0 | 72.4 | 67.0 |

**Table 3:** The prediction performance of different methods on the EVA benchmark datasets. sovH/E/C values are the specific SOV accuracies of the predicted helix, strand and coil, respectively. The prediction results of other methods on EVA_Set1 and EVA_Set2 are reported at http://cubic.bioc.columbia.edu/eva/sec/common3.html.

## 3.5 Results on Other Applications Using Essential Residues

Although our prediction method and the meta-predictor can generate highly accurate predictions, we are more concerned about the efficacy of the essential residues in various applications that use PSS as a feature for predictions. We demonstrated the efficacy of essential residues by applying ProtoPred with essential residues as input features to protein function prediction and enzyme/non-enzyme classification.

### 3.5.1 Protein Function Prediction

The knowledge of protein functions is crucial to the understanding of biological process. Since the experimental procedures for protein function annotation are inherently low throughput, the accurate computational techniques for protein function prediction represent useful tools. Automated protein function prediction methods include direct homology-based and indirect subsequence/feature-based approaches. For the indirect subsequence-based approaches, often only specific subsequences are crucial for the protein to perform its function (Pandey et al., 2006). This motivated us to use the essential residues in the function predictions.

We downloaded the protein function labels from the Gene Ontology Annotation Database (goa_pdb) (Camon et al., 2004). Since we needed to compile a dataset whose protein sequences are not redundant (mutual sequence identity less than 25%) and each of them is of known secondary structure, we then made an intersection set of goa_pdb with *DsspNr-25*. The number of proteins is 2677 and the total number of distinct function labels is 1539. It is worth to note that the function labels contain all GO annotations for the 2677 proteins, including the function labels of biological process, molecular functions, and cellular components. For example, the function labels of protein 1ak6 are 3779 (molecular function: actin binding) and 5622 (cellular component: intracellular).

In this application, we focus on verifying the efficacy of different sources of PSS. These sources are the real secondary structures, the predicted secondary structures of SymPred, and the predicted secondary structure of PSIPRED. ProtoPred predicts the most specific function label among 1539 candidates for a target protein by using one of the sources of secondary structures rather than general functions. The prediction accuracy is 100% if the predicted function label belongs to the target protein, otherwise it is 0%. For example, if we predict 1ak6 as the function 3779 (or 5622) then the accuracy is 100%. The hierarchical structure of GO annotations is not exploited in our prediction method, though it could be used to improve prediction accuracy (Eisner et al., 2005).

ProtoPred extract structure fragments using a sliding window of size *w*. Table 4 shows the results for several different window sizes. It can be observed that ProtoPred's prediction using the predicted secondary structure of SymPred shows the highest accuracy for all studied window sizes (except the window size of 11 because it is too short to represent the uniqueness of structures for different function classes). For example, for the window size of 51, the prediction accuracies of ProtoPred using the features of real structure, PSIPRED's prediction, and SymPred's prediction are 49.8%, 35.4%, and 57.6% respectively. Notably, the $Q_3$ of PSIPRED and SymPred on this dataset are 80.3% and 81.1%. Although the performances of PSS prediction of the two methods are similar, the effectiveness is quite different. Moreover, the performance of ProtoPred with SymPred's prediction is also better than that of ProtoPred with real structure. A possible explanation for this discrepancy is that different structures within a protein do not have equal importance for its function. It shows that SymPred can identify the essential residues which are crucial for proteins to perform their functions. Structural identities of low relevance residues dilute the influence of major residues when using the real structure as the feature in the ProtoPred's prediction.

| Window Size | 11 | 21 | 31 | 41 | 51 | 61 | 71 |
|---|---|---|---|---|---|---|---|
| Real Structure | 21.0 | 21.1 | 31.5 | 45.5 | 49.8 | 51.8 | 53.0 |
| PSIPRED | 21.0 | 21.0 | 23.3 | 28.9 | 35.4 | 40.6 | 44.0 |
| SymPred | 21.0 | 21.5 | 39.4 | 53.8 | 57.6 | 58.3 | 59.1 |

**Table 4:** The accuracy (%) of function predictions using different structure sources and different window sizes.

### 3.5.1 Enzyme/non-enzyme classification

Many protein function prediction methods focus on only one specific type of functions (Borgwardt et al., 2005, Dobson & Doig, 2003). The problem of enzyme and non-enzyme classifications is a special case of function prediction. We do not have to predict a functional type but only to distinguish between enzyme and non-enzyme. In Dobson and Doig's study, they use multiple features such as secondary structure, amino acid propensities, and surface properties to do the binary classifications. They further divide the features into 52 sub-features and select 36 optimal sub-features for the SVM models to generate the classifier. The overall accuracies are 77.16% and 80.14% for the two different sizes of sub-features, respectively.

We download Dobson and Doig's dataset which contained 1076 proteins. Since SymPred's prediction is the most effective feature among different sources of PSS in the above protein function prediction, ProtoPred uses SymPred's prediction as the input feature for the problem of enzyme and non-enzyme classifications. ProtoPred achieves an overall accuracy 81.8%.

In this application, we only use the secondary structural information for enzyme/non-enzyme classification and achieve a better result. It suggests that the secondary structural information with the essential residue annotation may be sufficient to predict protein functions, which supports the conclusion of Przytycka et al (Przytycka et al., 1999).

## 4 Conclusions

In this paper, we have proposed an improved knowledge based approach called SymPred for PSS prediction. We have also presented a meta-predictor called SymPsiPred, which combines a knowledge based approach (SymPred) and a machine learning based approach (PSIPRED). Tests on a proteome-scale dataset of 8297 protein chains show that the overall average $Q_3$ accuracy of SymPred and SymPsiPred is 81.0% and 83.9% respectively. SymPred can be regarded as a special case of a template-based approach because it predicts PSS by finding template sequences based on local similarities, i.e., n-gram words.

Almost all other methods adopting the predicted secondary structure as a feature are based on the entire sequence; we demonstrate that it is advantageous to weight amino acids according to their importance in determining the structure and function of a protein sequence. We test two different problems to verify the efficacy of the predicted structures and their weighted scores. In the problem of protein function prediction and enzyme/non-enzyme classification, the performance of using the feature of SymPred's PSS prediction result shows that the structural information with essential residue annotation is more suitable for predicting specific protein functions.

## References

Alexander, P. A., He, Y., Chen, Y., Orban, J., & Bryan, P. N. (2007). The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci U S A, 104*(29), 11963-11968.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research, 25*(17), 3389-3402.

Aydin, Z., Altunbasak, Y., & Borodovsky, M. (2006). Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. *Bmc Bioinformatics, 7*, -.

Bondugula, R., & Xu, D. (2007). MUPRED: A tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. *Proteins-Structure Function and Bioinformatics, 66*(3), 664-670.

Borgwardt, K. M., Ong, C. S., Schonauer, S., Vishwanathan, S. V. N., Smola, A. J., & Kriegel, H. P. (2005). Protein function prediction via graph kernels. *Bioinformatics, 21*, I47-I56.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., et al. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research, 32*, D262-D266.

Capra, J. A., & Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics, 23*(15), 1875-1882.

Ceroni, A., Frasconi, P., Passerini, A., & Vullo, A. (2003). A combination of support vector machines and bidirectional recurrent neural networks for protein secondary structure prediction. *Ai(Asterisk)Ia 2003: Advances in Artificial Intelligence, Proceedings, 2829*, 142-153.

Cheng, H. T., Sen, T. Z., Kloczkowski, A., Margaritis, D., & Jernigan, R. L. (2005). Prediction of protein secondary structure by mining structural fragment database. *Polymer, 46*(12), 4314-4321.

Dobson, P. D., & Doig, A. J. (2003). Distinguishing Enzyme Structures from Non-enzymes Without Alignments. *Journal of Molecular Biology, 330*(4), 771-783.

Eisner, R., Poulin, B., Szafron, D., Lu, P., & Greiner, R. (2005). *Improving Protein Function Prediction using the Hierarchical Structure of the Gene Ontology.* Paper presented at the Computational Intelligence in Bioinformatics and Computational Biology, 2005. CIBCB '05. Proceedings of the 2005 IEEE Symposium on.

Ferre, S., & King, R. D. (2006). Finding motifs in protein secondary structure for use in function prediction. *Journal of Computational Biology, 13*(3), 719-731.

Fischer, D., Elofsson, A., Rychlewski, L., Pazos, F., Valencia, A., Rost, B., et al. (2001). CAFASP2: The second critical assessment of fully automated structure prediction methods. *Proteins-Structure Function and Genetics*, 171-183.

Frenkel-Morgenstern, M., Voet, H., & Pietrokovski, S. (2005). Enhanced statistics for local alignment of multiple alignments improves prediction of protein function and structure. *Bioinformatics, 21*(13), 2950-2956.

Gong, H. P., & Rose, G. D. (2005). Does secondary structure determine tertiary structure in proteins? *Proteins-Structure Function and Bioinformatics, 61*(2), 338-343.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology, 292*(2), 195-202.

Jones, D. T., & Swindells, M. B. (2002). Getting the most from PSI-BLAST. *Trends in Biochemical Sciences, 27*(3), 161-164.

Karplus, K., Barrett, C., & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics, 14*(10), 846-856.

Kelley, L. A., & Sternberg, M. J. E. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols, 4*(3), 363-371.

Kim, H., & Park, H. (2003). Protein secondary structure prediction based on an improved support vector machines approach. *Protein Engineering, 16*(8), 553-560.

Laskowski, R. A., Watson, J. D., & Thornton, J. M. (2005). ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Research, 33*, W89-W93.

Li, W. Z., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics, 22*(13), 1658-1659.

Lin, H. N., Chang, J. M., Wu, K. P., Sung, T. Y., & Hsu, W. L. (2005). HYPROSP II - A knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics, 21*(15), 3227-3233.

Lobley, A., Swindells, M. B., Orengo, C. A., & Jones, D. T. (2007). Inferring function using patterns of native disorder in proteins. *Plos Computational Biology, 3*(8), 1567-1579.

Meiler, J., & Baker, D. (2003). Coupled prediction of protein secondary and tertiary structure. *Proceedings of the National Academy of Sciences of the United States of America, 100*(21), 12105-12110.

Nair, R., & Rost, B. (2003). Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins-Structure Function and Genetics, 53*(4), 917-930.

Nair, R., & Rost, B. (2005). Mimicking cellular sorting improves prediction of subcellular localization. *Journal of Molecular Biology, 348*(1), 85-100.

Pandey, G., Kumar, V., & Steinbach, M. (2006). *Computational Approaches for Protein Function Prediction*: Department of Computer Science and Engineering, University of Minnesota, Twin Cities.

Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Research, 24*(19), 3836-3845.

Przybylski, D., & Rost, B. (2007). Consensus sequences improve PSI-BLAST through mimicking profile-profile alignments. *Nucleic Acids Research, 35*(7), 2238-2246.

Przytycka, T., Aurora, R., & Rose, G. D. (1999). A protein taxonomy based on secondary structure. *Nature Structural Biology, 6*(7), 672-682.

Rost, B. (2001). Review: Protein secondary structure prediction continues to rise. *Journal of Structural Biology, 134*(2-3), 204-218.

Rost, B., & Sander, C. (2000). Third generation prediction of secondary structure *Protein Structure Prediction: Methods and Protocols* (pp. 71-95): Humana Press.

Rychlewski, L., Jaroszewski, L., Li, W. Z., & Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science, 9*(2), 232-241.

Sadreyev, R., & Grishin, N. (2003). COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance. *Journal of Molecular Biology, 326*(1), 317-336.

Schlessinger, A., Punta, M., & Rost, B. (2007). Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics, 23*(18), 2376-2384.

Smith, T. F., & Waterman, M. S. (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology, 147*(1), 195-197.

Sternberg, M. J. E., & Islam, S. A. (1990). Local protein sequence similarity does not imply a structural relationship. *Protein Engineering Design and Selection, 4*(2), 125-131.

Su, E., Chiu, H.-S., Lo, A., Hwang, J.-K., Sung, T.-Y., & Hsu, W.-L. (2007). Protein subcellular localization prediction based on compartment-specific features and structure conservation. *BMC Bioinformatics, 8*(1), 330.

Ward, J. J., McGuffin, L. J., Buxton, B. F., & Jones, D. T. (2003). Secondary structure prediction with support vector machines. *Bioinformatics, 19*(13), 1650-1655.

Yona, G., & Levitt, M. (2002). Within the twilight zone: A sensitive profile-profile comparison tool based on information theory. *Journal of Molecular Biology, 315*(5), 1257-1275.