# Exploring Potential Discriminatory Information Embedded in PSSM to Enhance Protein Structural Class Prediction Accuracy

Abdollah Dehzangi[1,2], Kuldip Paliwal[1], James Lyons[1], Alok Sharma[3], and Abdul Sattar[1,2]

[1] Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Brisbane, Australia
[2] National ICT Australia (NICTA), Brisbane, Australia
[3] University of the South Pacific, Fiji
{a.dehzangi,k.paliwal,j.lyons,a.sattar}@griffith.edu.au, sharma_al@usp.ac.fj

**Abstract.** Determining the structural class of a given protein can provide important information about its functionality and its general tertiary structure. In the last two decades, the protein structural class prediction problem has attracted tremendous attention and its prediction accuracy has been significantly improved. Features extracted from the *Position Specific Scoring Matrix (PSSM)* have played an important role to achieve this enhancement. However, this information has not been adequately explored since the protein structural class prediction accuracy relying on PSSM for feature extraction still remains limited. In this study, to explore this potential, we propose segmentation-based feature extraction technique based on the concepts of amino acids' distribution and auto covariance. By applying a *Support Vector Machine (SVM)* to our extracted features, we enhance protein structural class prediction accuracy up to 16% over similar studies found in the literature. We achieve over 90% and 80% prediction accuracies for 25PDB and 1189 benchmarks respectively by solely relying on the PSSM for feature extraction.

**Keywords:** Protein Structural Class Prediction Problem, Feature Extraction, Segmented distribution, Segmented Auto Covariance, Support Vector Machine (SVM).

## 1 Introduction

Protein structural class prediction problem is defined as assigning a given protein to one of four structural classes namely all-$\alpha$, all-$\beta$, $\alpha + \beta$, and $\alpha/\beta$ [1]. Protein structural class prediction can provide important information about the functionality of proteins as well as their general tertiary structure. Despite all the efforts that have been made to find a fast computational approach to solve this problem, especially for low homologous protein sequences, it still remains unsolved for computational biology and bioinformatics [2–4].

A. Ngom et al. (Eds.): PRIB 2013, LNBI 7986, pp. 208–219, 2013.

During the last two decades, a wide range of classification techniques have been proposed to tackle the protein structural class prediction problem such as, *Support Vector Machine (SVM)* [5–8], *Artificial Neural Network (ANN)* [9, 10], *Meta Classifiers* [11, 12], and ensembles of classifiers [13–15]. Among the proposed classification techniques used to tackle this problem, SVM has attained the best results [7, 16–18]. Similarly, a wide range of features have been proposed and used to reveal more discriminatory information for this task [5, 16, 19]. More significant improvement for protein structural class prediction accuracy has come from the new features being introduced rather than the classification technique being used for this task [16, 17, 20].

The first group of features that significantly enhanced the protein structural class prediction accuracy were extracted from the evolutionary information embedded in the *Position Specific Scoring Matrix (PSSM)* [21]. Latter on, several feature extraction techniques were proposed to explore the potential local and global discriminatory information embedded in PSSM to tackle this problem such as composition of the amino acids [8], pseudo amino acid composition [2], dipeptide composition [8], and auto covariance [17]. However, the discriminatory information embedded in PSSM has not been adequately explored since the prediction accuracy relying on these features remains limited. Further enhancement for the protein structural class prediction accuracy has been achieved by relying on the structural information extracted [7, 16] from the predicted secondary structure of proteins using PSIPRED [22]. despite a wide range of feature extraction techniques being explored [5, 7, 8, 20], the protein structural class prediction accuracy relying on structural information has not been improved adequately since the study of Mizianty and Kurgan in 2009 [16]. This highlights the need for novel feature extraction techniques relying on the alternative sources for feature extraction.

In this study, we propose two segmented feature extraction techniques based on the concepts of distribution and auto covariance methods to explore local discriminatory information embedded in the PSSM. We also use the concept of occurrence of the amino acids to explore global discriminatory information embedded in PSSM rather than composition of the amino acids that has been widely used for this task to capture the information regarding the length of the protein sequence [16, 17]. By applying SVM to our extracted features we achieve over 90% and 80% protein structural class prediction accuracies for 25PDB and 1189 benchmarks respectively. We enhance the protein structural class prediction accuracy for up to 16% compared to smilar studies which have used PSSM for feature extraction.

## 2 Benchmarks

In this study, two popular benchmarks that have been widely used for the protein structural class prediction problem namely, 25PDB and 1189 benchmarks are used. The 25PDB benchmark was introduced in [19] consists of 1673 proteins with less than 25% sequential similarities (the homology range between 22%

and 45%). This benchmark was extracted from 25% PDBSELECTED which includes high resolution protein sequences in the *Protein Data Bank (PDB)* [23]. Therefore, this benchmark is considered as a reliable representative of proteins in the twilight zone (proteins with the sequence similarities between 20% to 45%). Hence, this benchmark is employed in this study as the main source to investigate the performance of our proposed techniques.

The 1189 benchmark is a popular benchmark that has been widely used in the literature. This benchmark was introduced by [3] consisted of 1189 proteins. However, 97 proteins were dropped from this benchmark in later studies [19] to address further correction of *Structural Classification of Proteins (SCOP)* [24]. As the result, current version of this benchmark consists of 1092 proteins with less than 40% sequential similarities. Dissimilar to 25PDB, this benchmark includes proteins with low resolutions as well. Therefore, despite higher sequential similarity among proteins in this benchmark, lower prediction accuracies have been reported in the literature for this benchmark compared to 25PDB using similar approaches [5, 7, 8]. This benchmark is mainly used in this study to compare our results directly with previously reported results as well as tuning the classification and feature extraction parameters while 25PDB benchmark is not used at all in the tuning step.

## 3 Feature Extraction Method

Since our proposed features are all extracted directly from PSSM, we need to first produce this matrix. To calculate PSSM, PSI-BLAST [21] is applied for both 25PDB and 1189 benchmarks (using NCBI's non redundant (NR) database while its cut off value (E) is set to 0.001). PSSM provides the substitution probability of a given amino acid based on its position in a protein sequence with all 20 amino acids. It consists of two L $\times$ 20 matrices (where L is the length of protein sequence and 20 columns are representatives of 20 amino acids). The first matrix provides the log-odds of the amino acids substitution probabilities and it is called PSSM_cons while the second matrix provides normalized substitution probability and it is called PSSM_probs. Since PSSM_cons has been widely used in the literature for feature extraction [16, 17], it is also adopted in this study.

To explore potential local and global discriminatory information embedded in PSSM, four feature groups are proposed and used in this study. These feature groups are, consensus sequence-based occurrence of the amino acids (AAO), semi occurrence of the amino acids (PSSM-AAO), segmented distribution (PSSM-SD), and segmented auto covariance (PSSM-SAC). The first two feature groups are proposed to reveal global discriminatory information while the remaining two methods are proposed to reveal local discriminatory information embedded in PSSM. These four feature extraction methods are explained in detail in the following subsections.

### 3.1 Consensus Sequence-Based Occurrence (AAO)

To extract global discriminatory information embedded in PSSM, we first extract the occurrence of the amino acids feature group from the consensus sequence derived from PSSM. In the protein consensus sequence, amino acids along the original protein sequence $(O_1, O_2, ..., O_L)$ are replaced with the corresponding amino acids with the maximum substitution probabilities in PSSM $(C_1, C_2, ..., C_L)$. This is done in the following two steps. In the first step, the index of the amino acid with the highest substitution probability (based on its position in the protein sequence) is calculated as follows:

$$I_i = \text{argmax}\{P_{ij} : 1 \leq j \leq 20\},\ 1 \leq i \leq L, \tag{1}$$

where $P_{ij}$ is the substitution probability of the amino acid at location $i$ with the $j^{th}$ amino acid in PSSM_cons. In the second step, we replace the amino acid at $i^{th}$ location of original protein sequence by the $I_{i^{th}}$ amino acid to form the consensus sequence. After calculating the consensus sequence, we count the number of occurrence of each amino acid (for all 20 amino acids) along the consensus sequence and return the corresponding values. Therefore, a feature group consisting of 20 features is calculated. The occurrence feature group as the global descriptor of the proteins is used in this study since it maintains the information regarding the length of protein sequence which is discarded using the composition feature group (occurrence of amino acids divided by the length of the protein sequence (AAC) [16]).

### 3.2 Semi Occurrence (PSSM-AAO)

This feature group is directly extracted from the PSSM. It is called semi occurrence because it is not calculated in the similar manner to the occurrence feature group as it was explained in previous subsection. Instead, it is produced by summation of the substitution score of a given amino acid with all the amino acids along the protein sequence which is calculated as follows:

$$PSSM\text{-}AAO_j = \sum_{i=1}^{L} P_{ij}\ ,\ (j = 1, ..., 20). \tag{2}$$

This feature group is able to provide important global discriminatory information about the substitution probability of the amino acids [17]. Different to composition of the amino acid extracted from PSSM (which is called PSSM-AAC in [17]), PSSM-AAO maintains the information regarding to the length of protein sequence. In PSSM-AAC the the summation of substitution probabilities of the amino acids are divided by the length of protein sequence.

### 3.3 Segmented Distribution (PSSM-SD)

This method is specifically proposed to add more local discriminatory information about how the amino acids, based on their substitution probability with each

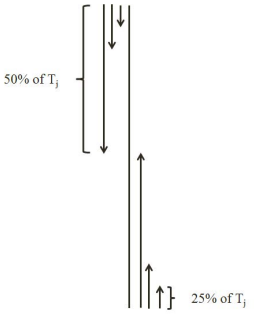



**Fig. 1.** The segmentation method used to extract PSSM-SD feature group

other (extracted from PSSM), are distributed along the protein sequence. We propose this segmentation method in the manner where segments of a protein sequence are of unequal lengths and each segment is represented by a distribution feature which is computed as follows. First, for the $j^{th}$ column in the PSSM, we calculate the total substitution probability $T_j = \sum_{i=1}^{L} P_{ij}$. Then, starting from the first row of PSSM, we calculate the partial sum $S_1$ of the substitution probabilities of the first $i$ amino acids until reaching to *25%* of the total sum $S_1 = \sum_{i=1}^{I_j^1} P_{ij}$. Using the distribution factor $F = 25\%$, we calculate the $I_j^1$. The $I_j^1$ corresponds to the number of the amino acids such that the summation of their substitution probabilities is less than or equal to the $F = 25\%$ of $(T_j)$. Similarly, we calculate the partial sum of the first $i$ amino acids (starting from the first row of PSSM) until reaching $2 \times F = 50\%$ of the total sum $S_2 = \sum_{i=1}^{I_j^2} P_{ij}$ and calculate the $I_j^2$ corresponding to the number of amino acids such that the summation of their substitution probabilities is less than or equal to $F = 50\%$ of the total $T_j$.

We repeat the same process beginning from the last row of the PSSM for the $j^{th}$ column. We calculate the partial sum of the substitution probability of the first $i$ amino acids until reaching $F = 25\%$ and $2 \times F = 50\%$ of the total sum which are $S_3 = \sum_{i=1}^{I_j^3} P_{ij}$ and $S_4 = \sum_{i=1}^{I_j^4} P_{ij}$ respectively and calculate the $I_j^3$ and $I_j^4$. $I_j^3$ and $I_j^4$ correspond to the number of amino acids such that the summation of their substitution probability is less than or equal to $F$ and $2 \times F$ of $T_j$ respectively (starting from the last row of PSSM). In this manner we extract four segmented distribution features for each column in PSSM. The method used to calculate PSSM-SD is shown in Figure 1. We repeat the same process for all 20 columns corresponding to 20 amino acids in PSSM and extract 80 features in total in this feature group $(4 \times 20 = 80)$. Note that $F = 25\%$ is adopted in this study due it s better performance compared to use of $F = 10\%$ and $F = 5\%$ explored experimentally by the authors. In the other word, using four segments is sufficient for providing adequate local discriminatory information compared to the use of 10 or 20 segments.

### 3.4 Segmented Auto Covariance (PSSM-SAC)

The concept of auto covariance has been widely used in the literature to capture local discriminatory information and has attained better results compared to similar methods used for this task such as dipeptide composition [8, 17]. Pseudo amino acid composition based features are good examples of these types of features [2, 4]. These features have been computed using the whole protein sequence as a single entity for feature extraction. Therefore, they could not adequately explore the local sequence order information embedded in protein sequence [17]. In the present study, we extend the concept of segmented distribution features as described in the previous subsection to compute the auto covariance features from the segmented protein sequence. This is done to enforce local discriminatory information extracted from PSSM.

To extract this feature group, we calculate the auto covariance of the substitution probability of the amino acids using $K$ as the distance factor for each segment of proteins generated using segmented distribution in the following manner. Starting from the first row of PSSM, for the $j^{th}$ column of PSSM, we calculate $K$ auto covariance features for the first $I_j^1$. Similarly, we calculate auto covariance for the first $I_j^2$ amino acids. Then starting from the last row of PSSM for the $j^{th}$ column of PSSM, We repeat the same process for $I_j^3$, and $I_j^4$ ($I_j^1$, $I_j^2$, $I_j^3$, and $I_j^4$ are calculated from the previous subsection). This process is repeated for all 20 columns of PSSM and corresponding features are calculated as follows:

$$\text{PSSM-seg}_{n,m,j} = \frac{1}{(I_j^n - m)} \sum_{i=1}^{I_j^n - m} (P_{i,j} - P_{ave,j}) \times (P_{(i+m),j} - P_{ave,j}),$$
$$(n = 1, ..., 4 \ \& \ m = 1, ..., K \ \& \ j = 1, ..., 20), \quad (3)$$

where, $P_{ave,j}$ is the average substitution probability for the $j^{th}$ column in PSSM. Note that $2 \times K$ auto covariance coefficients are computed in this manner by analyzing PSSM in the downward direction and $2 \times K$ auto covariance coefficients are computed in this manner by analyzing PSSM in the upward direction ($4 \times K$ features in total). We also compute the global auto covariance coefficient ($K$ features) of PSSM as follows:

$$\text{PSSM-AC}_{m,j} = \frac{1}{(L - m)} \sum_{i=1}^{L-m} (P_{i,j} - P_{ave,j}) \times (P_{(i+m),j} - P_{ave,j}),$$
$$(m = 1, ..., K \ \& \ j = 1, ..., 20). \quad (4)$$

Thus, we have extracted a total of ( $2K + 2K + K = 5K$) auto covariance features in this manner (for the $j^{th}$ column of the PSSM). Therefore, for all 20 columns of the PSSM, segmented auto covariance of substitution probability of the amino acids are extracted and combined to build the corresponding feature group which will be referred to as PSSM-SAC (PSSM-seg + PSSM-AC which consists of $20 \times (5K)$) features in total).

## 4 Support Vector Machine

SVM was introduced by [25] to find the *Maximum Margin Hyper-plane (MMH)* based on the concept of the support vector theory to minimize classification error. It transforms the input data to higher dimension using the kernel function to be able to find support vectors (for nonlinear cases). The classification of some known points in input space $\mathbf{x}_i$ is $y_i$ which is defined to be either -1 or +1. If $x'$ is a point in input space with unknown classification then:

$$y' = sign\left(\sum_{i=1}^{n} a_i y_i K(\mathbf{x}_i, \mathbf{x}') + b\right), \tag{5}$$

where $y'$ is the predicted class of point $\mathbf{x}'$. The function *K()* is the kernel function; $n$ is the number of support vectors and $a_i$ are adjustable weights and $b$ is the bias. This classier is considered as the state-of-the-art classification techniques in the pattern recognition and attained the best results for the protein structural class prediction problem [7, 16, 17]. In this study, SVM classifier implemented in the LIBSVM (C-SVC type) toolbox using *Radial Basis Function (RBF)* as its kernel is used [26]. The $\gamma$ in addition to the regularization parameter $C$ (which also called the soft margin parameter) of the RBF kernel are optimized using grid search algorithm implemented in the LIBSVM package.

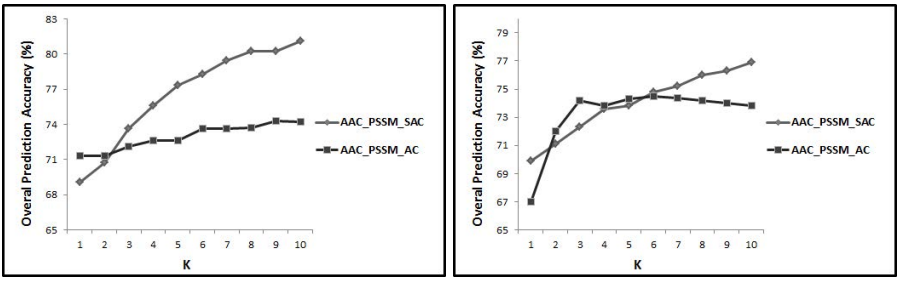
## 5 Results and Discussion

We first explore the effectiveness of the segmented auto covariance (PSSM-SAC) method compared to global auto covariance (PSSM-AC) used in [17]. PSSM-AC was used to explore local discriminatory information embedded in PSSM and attained the best results for this task. Then, one by one, we add the rest of the feature groups extracted in this study and explore their impact on the protein structural class prediction accuracy, separately. Finally, we compare the results reported in this study with the similar studies found in the literature for the protein structural class prediction problem. To evaluate the performance of our proposed methods and to be able to directly compare our results with previously studies, we adopt Jackknife cross validation as it was widely used for this task in the literature [16, 17, 19]

### 5.1 The Effectiveness of PSSM-SAC versus PSSM-AC

To investigate the effectiveness of PSSM-SAC compared to PSSM-AC we first reproduce the experiments conducted in [17]. In this experiment, PSSM-AC in combination of PSSM-AAC was used as the input feature group (called AAC-PSSM-AC) for different values of K (between 1 and 10) using an SVM classifier. We similarly combine the PSSM-SAC with PSSM-AAC (called AAC-PSSM-SAC) to be able to directly compare these two feature groups with respect to different values of distance factor $K$ between 1 and 10 (using an SVM as it

was used in [17]). The results achieved for 25PDB and 1189 are respectively shown in Figure 2.a and Figure 2.b. As it is shown in these figures, increasing the K value, AAC-PSSM-SAC significantly outperform AAC-PSSM-AC. Using K = 10 we achieve up to 81.1% and 76.9% prediction accuracies respectively for 25PDB and 1189 benchmarks. This highlights the effectiveness of PSSM-SAC to extract local discriminatory information based on the concept of auto covariance from the PSSM. Note that our results using solely AAC-PSSM-SAC enhances the protein structural class prediction accuracy for up to 6% and 2.3% for 25PDB and 1189 benchmarks respectively compared to the best results found in the literature relying on PSSM for feature extraction. In continuation, we replaced PSSM-AAC with PSSM-AAO which enhances the protein structural class prediction accuracy for all 10 values of K between 0.5% and 2% (when increasing K from 1 to 10, the impact of AAO is reduced from almost 2% to 0.5%) which shows the effectiveness of using AAO compared to AAC. Therefore, for the rest of this study, AAO is used instead of AAC. We then use grid search algorithm on 1189 to optimize SVM parameters (C and $\gamma$) for AAO-PSSM-AC (where K = 10) to avoid over tuning. 25PDB also was not used at all for this task. The optimal values achieved for C and $\gamma$ are respectively 500 and 0.05 which are used for the rest of this study.




(a) Comparison of the AAC_PSSM_AC and AAC_PSSM_SAC on 1189 benchmark

(b) Comparison of the AAC_PSSM_AC and AAC_PSSM_SAC on 25PDB benchmark

**Fig. 2.** Results achieved for AAC_PSSM_SAC and AAC_PSSM_AC with respect the value of K (Between 1 to 10) for 1189 and 25PDB benchmarks

### 5.2 The Effectiveness of PSSM-SD Feature Group

In continuation, we add the PSSM-SD feature group to the combination of PSSM-SAC and PSSM-AAO (AAO-PSSM-SAC) and study its impact for different values of K (between 1 and 10). The results achieved for 25PDB and 1189 benchmarks are shown in Figure 3. As we can see, by adding PSSM-SD, dissimilar to AAC-PSSM-SAC by increasing the value of K to 10, the prediction accuracy does not improve (it even slightly reduces). Therefore, adding PSSM-SD reduce the dependency to the value of K in PSSM-SAC to provide local

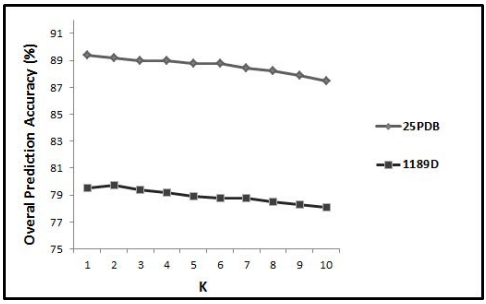



**Fig. 3.** The results achieved for combination of PSSM-AAO, PSSM-SAC, and PSSM-SD using SVM for different values of K (between 1 to 10) for 1189 and 25PDB benchmarks

information. In another word, we are able to increase the provided local information using PSSM-SD feature group and at the same time reduce the number of features. Using the combination of PSSM-AAO, PSSM-SAC, and PSSM-SD where K = 1 (20 + 100 + 80 = 200 features in total), we achieve up to 89.4% and 79.5% prediction accuracies for 25PDB and 1189 benchmarks respectively which are 15.3% and 4.9% better than the highest results reported for these benchmarks in the literature using features extracted from PSSM.

### 5.3 The Effectiveness of AAO Feature Group

In this Step, we add the AAO feature group to the combination of PSSM-AAO, PSSM-SAC (where K = 1), and PSSM-SD (20 + 20 + 100 + 80 = 220 features in total). By adding this feature group and applying SVM to these combination, we achieve up to 90.1% and 80.2% prediction accuracies respectively for 25PDB and 1189. These results are up to 16% and 5.6% respectively better than the best results reported for these two benchmarks using PSSM for feature extraction. It is important to highlight that these results are achieved using the same number of features used in [17] to achieve their best results for these two benchmark using PSSM for feature extraction. The results adding each feature group in each step is shown in Table.1. Note that in this table the impact of PSSM-SAC where K = 1 is shown while as it was explained in previous section, depend on the combination of feature groups being used, this impact has changed.

### 5.4 Performance Comparison with Existing Methods

In this section, the overall protein structural class prediction accuracy as well as prediction accuracy achieved for each structural class achieved by using the combination of our feature groups (PSSM-AAO + PSSM-SAC + PSSM-SD + AAO which will be referred as PSSM-S for simplicity) compared to previously reported results for this task are shown in Table 2 and Table 3. As we can see, we

**Table 1.** The impact of proposed feature extraction groups proposed in this study to enhance protein structural class prediction accuracy (in %)

| Combination of features | Classifier | 25PDB | 1189 |
|---|---|---|---|
| PSSM-AAO | SVM | 65.5 | 62.4 |
| PSSM-AAO + PSSM-SAC (K = 1) | SVM | 69.9 | 69.1 |
| PSSM-AAO + PSSM-SD | SVM | 87.1 | 76.4 |
| PSSM-AAO + PSSM-SAC (K = 1) + PSSM-SD | SVM | 89.4 | 79.5 |
| PSSM-AAO + PSSM-SAC (K = 1) + PSSM-SD + AAO | SVM | 90.1 | 80.2 |
| PSSM-AAO + PSSM-AC (K = 6) + PSSM-SD + AAO | SVM | 89.1 | 78.1 |

**Table 2.** Comparison of the results reported for the 25PDB benchmark (in percentage %)

| References | Method | All-$\alpha$ | All-$\beta$ | $\alpha$ / $\beta$ | $\alpha$ + $\beta$ | Overall |
|---|---|---|---|---|---|---|
| [19] | Logistic Regression | 69.1 | 61.6 | 60.1 | 38.3 | 57.1 |
| [27] | Specific Tri-peptides | 60.6 | 60.7 | 67.9 | 44.3 | 58.6 |
| [13] | LLSC-PRED | 75.2 | 67.5 | 62.1 | 44.0 | 62.2 |
| [13] | SVM | 77.4 | 66.4 | 61.3 | 45.4 | 62.7 |
| [14] | SSA | 92.6 | 83.7 | 80.5 | 65.9 | 81.5 |
| [28] | SCPRED | 92.6 | 80.1 | 74.0 | 71.0 | 62.7 |
| [29] | CWT-PCA-SVM | 76.5 | 67.3 | 66.8 | 45.8 | 64.0 |
| [18] | AATP | 81.9 | 74.7 | 75.1 | 55.8 | 71.7 |
| [8] | AADP-PSSM | 83.3 | 78.1 | 76.3 | 54.4 | 72.9 |
| [17] | AAC-PSSM-AC | 85.3 | 81.7 | 73.7 | 55.3 | 74.1 |
| This Study | PSSM-S | 93.8 | 92.8 | 92.6 | 81.7 | 90.1 |

**Table 3.** Comparison of the results reported for the 1189 benchmark (in percentage %)

| References | Method | All-$\alpha$ | All-$\beta$ | $\alpha$ / $\beta$ | $\alpha$ + $\beta$ | Overall |
|---|---|---|---|---|---|---|
| [3] | Bayes Classifier | 54.8 | 57.1 | 75.2 | 22.2 | 53.8 |
| [19] | Logistic Regression | 57.0 | 62.9 | 64.7 | 25.3 | 53.9 |
| [30] | FKNN | 48.9 | 59.5 | 81.7 | 26.6 | 56.9 |
| [27] | Specific Tri-peptides | - | - | - | - | 59.9 |
| [15] | IB1 | 65.3 | 67.7 | 79.9 | 40.7 | 64.7 |
| [31] | SVM | 75.8 | 75.2 | 82.6 | 31.8 | 67.6 |
| [18] | AATP | 72.7 | 85.4 | 82.9 | 42.7 | 72.6 |
| [8] | AADP-PSSM | 69.1 | 83.7 | 85.6 | 35.7 | 70.7 |
| [17] | AAC-PSSM-AC | 80.7 | 86.4 | 81.4 | 45.2 | 74.6 |
| This Study | PSSM-S | 93.3 | 85.1 | 77.6 | 65.6 | 80.2 |

not only significantly enhance the overall protein structural class prediction accuracy but also in most of the cases achieve better results for different structural classes. Relying solely on PSSM for feature extraction, we achieve over 90% and 80% prediction accuracies for 25PDB and 1189 benchmarks. It is important to highlight that we also achieved significantly higher results for 25PDB compared to studies which have used PSIPRED for feature extraction as well while it was relatively comparable for 1189 [7, 16].

## 6 Conclusion and Future Works

In this study, we proposed novel feature extraction methods to explore potential local and global discriminatory information embedded in PSSM for protein

structural class prediction problem. We proposed the concepts of segmented auto covariance and segmented distribution to extract this local information. We also employed the concept of occurrence to extract potential global discriminatory information directly from PSSM as well as the transformed protein sequence using PSSM. By applying SVM we showed the effectiveness of our proposed feature groups by enhancing protein structural class prediction accuracy for up to 16% and 5.6% for 25PDB and 1189 benchmarks respectively. We, for the first time, achieved over 90% and 80% (90.1% and 80.2%) protein structural class prediction accuracies for 25PDB and 1189 benchmarks respectively using PSSM for feature extraction. For our future work, we aim to study the effectiveness of structural information based on predicted secondary structure of proteins to enhance the protein structural class prediction accuracy, further.

## References

1. Chothia, C.: The nature of the accessible and buried surfaces in proteins. Journal of Molecular Biology 105(1), 1–12 (1976)
2. Chou, K.C.: Progress in protein structural class prediction and its impact to bioinformatics and proteomics. Current Protein and Peptide Science 6, 423–436 (2005)
3. Wang, Z.X., Yuan, Z.: How good is prediction of protein structural class by the component-coupled method? Proteins: Structure, Function, and Bioinformatics 38(2), 165–175 (2000)
4. Chou, K.C.: Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology 273(1), 236–247 (2011)
5. Yang, J.Y., Peng, Z.L., Chen, X.: Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. BMC Bioinformatics 11(suppl. 1), S9 (2010)
6. Li, Z.C., Zhou, X.B., Lin, Y.R., Zou, X.Y.: Prediction of protein structure class by coupling improved genetic algorithm and support vector machine. Amino Acids 35(3), 581–590 (2008)
7. Zhang, S., Ding, S., Wang, T.: High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure. Biochimie 93(4), 710–714 (2011)
8. Liu, T., Jia, C.: A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. Journal of Theoretical Biology 267(3), 272–275 (2010)
9. Jahandideh, S., Abdolmaleki, P., Jahandideh, M., Asadabadi, E.B.: Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophysical Chemistry 128(1), 87–93 (2007)
10. Jahandideh, S., Abdolmaleki, P., Jahandideh, M., Hayatshahi, S.H.S.: Novel hybrid method for the evaluation of parameters contributing in determination of protein structural classes. Journal of Theoretical Biology 244(2), 275–281 (2007)
11. Cai, Y.D., Feng, K., Lu, W., Chou, K.: Using logitboost classifier to predict protein structural classes. Theoretical Biollogy 238, 172–176 (2006)
12. Jain, P., Hirst, J.: Automatic structure classification of small proteins using random forest. BMC Bioinformatics 11(1), 364 (2010)
13. Kurgan, L.A., Chen, K.: Prediction of protein structural class for the twilight zone sequences. Biochemical and Biophysical Research Communications 357(2), 453–460 (2007)

14. Kurgan, L.A., Zhang, T., Zhang, H., Shen, S., Ruan, J.: Secondary structure-based assignment of the protein structural classes. Amino Acids 35, 551–564 (2008)
15. Chen, K., Kurgan, L.A., Ruan, J.: Prediction of protein structural class using novel evolutionary collocation-based sequence representation. Journal of Computational Chemistry 29(10), 1596–1604 (2008)
16. Mizianty, M., Kurgan, L.A.: Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. BMC Bioinformatics 10(1), 414 (2009)
17. Liu, T., Geng, X., Zheng, X., Li, R., Wang, J.: Accurate prediction of protein structural class using auto covariance transformation of psi-blast profiles. Amino Acids 42, 2243–2249 (2012)
18. Zhang, S., Ye, F., Yuan, X.: Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via pssm. Journal of Biomolecular Structure and Dynamics 29(6), 1138–1146 (2012)
19. Kurgan, L.A., Homaeian, L.: Prediction of structural classes for protein sequences and domains - impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recognition 39, 2323–2343 (2006)
20. Yang, J.Y., Peng, Z.L., Yu, Z.G., Zhang, R.J., Anh, V., Wang, D.: Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. Journal of Theoretical Biology 257(4), 618–626 (2009)
21. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., Lipman, D.J.: Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Research 17, 3389–3402 (1997)
22. Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 292(2), 195–202 (1999)
23. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. Nucleic Acids Research 28(1), 235–242 (2000)
24. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: Scop: A structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 247(4), 536–540 (1995)
25. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc. (1995)
26. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines (2001)
27. Costantini, S., Facchiano, A.M.: Prediction of the protein structural class by specific peptide frequencies. Biochimie 91(2), 226–229 (2009)
28. Kurgan, L.A., Cios, K.J., Chen, K.: Scpred: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. BMC Bioinformatics 9, 226 (2008)
29. Li, Z.C., Zhou, X.B., Dai, Z., Zou, X.Y.: Prediction of protein structural classes by chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis. Amino Acids 37, 415–425 (2009)
30. Zhang, T.L., Ding, Y.S., Chou, K.C.: Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern. Theoretical Biology 250, 186–193 (2008)
31. Chen, C., Zhou, X., Tian, Y., Zou, X., Cai, P.: Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Analytical Biochemistry 357(1), 116–121 (2006)