A mixture of physicochemical and evolutionary-based feature extraction approaches for protein fold recognition

Abdollah Dehzangi*

Institute for Integrated and Intelligent Systems (IIIS),
Griffith University,
Brisbane, Australia
and
National ICT Australia (NICTA),
Brisbane, Australia
Email: a.dehzangi@griffith.edu.au
Email: abdollah.dehzangi@griffithuni.edu.au
*Corresponding author

Alok Sharma

Institute for Integrated and Intelligent Systems (IIIS),
Griffith University,
Brisbane, Australia
and
School of Engineering and Physics,
Suva, Fiji
Email: a.sharma@griffith.edu.au

James Lyons and Kuldip K. Paliwal

School of Engineering,
Griffith University,
Brisbane, Australia
Email: james.lyons@griffithuni.edu.au
Email: k.paliwal@griffith.edu.au

Abdul Sattar

Institute for Integrated and Intelligent Systems (IIIS),
Griffith University,
Brisbane, Australia
and
National ICT Australia (NICTA),
Brisbane, Australia
Email: a.sattar@griffith.edu.au

Abstract: Recent advancement in the pattern recognition field stimulates enormous interest in Protein Fold Recognition (PFR). PFR is considered as a crucial step towards protein structure prediction and drug design. Despite all the recent achievements, the PFR still remains as an unsolved issue in biological science and its prediction accuracy still remains unsatisfactory. Furthermore, the impact of using a wide range of physicochemical-based attributes on the PFR has not been adequately explored. In this study, we propose a novel mixture of physicochemical and evolutionary-based feature extraction methods based on the concepts of segmented distribution and density. We also explore the impact of 55 different physicochemical-based attributes on the PFR. Our results show that by providing more local discriminatory information as well as obtaining benefit from both physicochemical and evolutionary-based features simultaneously, we can enhance the protein fold prediction accuracy up to 5% better than previously reported results found in the literature.

Keywords: protein fold recognition; feature selection; mixture of feature extraction models; segmented-based distribution; segmented-based density; evolutionary-based features; physicochemical-based features.

Reference to this paper should be made as follows: Dehzangi, A., Sharma, A., Lyons, J., Paliwal, K.K. and Sattar, A. (2015) 'A mixture of physicochemical and evolutionary-based feature extraction approaches for protein fold recognition', *Int. J. Data Mining and Bioinformatics*, Vol. 11, No. 1, pp.115–138.

Biographical notes: Abdollah Dehzangi received the BSc degree in Computer Engineering-Hardware from Shiraz University, Iran in 2007 and Master degree in the area of Bioinformatics from Multi Media University (MMU), Cyberjaya, Malaysia, in 2011. Since 2011, he is pursuing the PhD degree in Bioinformatics at Griffith University Brisbane, Australia. He is also a researcher in National ICT Australia (NICTA). His research interests include bioinformatics, protein fold and structural class prediction problems, data mining, statistical learning theory, and pattern recognition.

Alok Sharma received the PhD degree in the area of Pattern Recognition from Griffith University, Brisbane, Australia, in 2006. He was with the University of Tokyo, Japan (2010–2012) as a research fellow. He is an A/Prof. at the USP and an Adjunct A/Prof. at the Institute for Integrated and Intelligent Systems (IIIS), Griffith University. He participated in various projects carried out in conjunction with Motorola (Sydney), Auslog Pty., Ltd. (Brisbane), CRC Micro Technology (Brisbane), the French Embassy (Suva) and JSPS (Japan). His research interests include pattern recognition, computer security, human cancer classification and protein fold and structural class prediction problems.

James Lyons received a BEng degree with Honours and a BIT from Griffith University Brisbane, Australia in 2007. He is now pursuing a PhD degree in Robust Automatic Speech and Speaker Recognition at Griffith University Brisbane, Australia. His research interests include automatic speech and speaker recognition, bioinformatics, protein fold and structural class prediction problems and pattern recognition.

Kuldip Paliwal has worked at a number of organisations including Norwegian Institute of Technology, and AT&T Shannon Laboratories, New Jersey, USA. Since 1993, he has been a Professor at Griffith University, in the School of Engineering. His current research interests include speech recognition, speech enhancement, face recognition, bioinformatics, protein fold and structural class

prediction problems. He has served the IEEE Signal Processing Society's Neural Networks Technical Committee as a founding member and the Speech Processing Technical Committee. He was an Associate Editor of the *IEEE Transactions on Speech and Audio Processing* and also the *IEEE Signal Processing Letters*.

Abdul Sattar is the founding Director of the Institute for Integrated and Intelligent Systems and a Professor of Computer Science and Artificial Intelligence at Griffith University. He is also a Research Leader at National ICT Australia (NICTA) Queensland Research Lab (QRL), where he has held the positions of QRL Education Director (2006–2008) and Leader of the Smart Applications for Emergencies (SAFE) project (2005–2008), and is currently leading the QRL node of NICTA's largest project, Advanced Technologies for Optimisation and Modelling In Constraints (ATOMIC). His research interests include knowledge representation and reasoning, constraint satisfaction, intelligent scheduling, temporal databases, and bioinformatics.

1 Introduction

Proteins can be categorised into finite number of groups called fold based on their major tertiary structure. It has been shown that proteins in the same fold share similar functionality. Therefore, being able to accurately classify a protein into its appropriate fold is considered as an important step towards protein structure prediction and drug design. In biological terminology, this problem is defined as *Protein Fold Recognition* (*PFR*). Despite recent advancements, PFR still remains an unsolved issue for bioinformatics and biological science especially for low homology protein sequences. In pattern recognition perspective, PFR is viewed as solving a multi-class classification task. The PFR procedure includes feature extraction, attribute selection and classification of proteins. During the past two decades, a wide range of classification techniques have been proposed and used successfully for the PFR (Ding and Dubchak, 2001; Damoulas and Girolami, 2008; Deschavanne and Tuffery, 2009; Nanni et al., 2010; Dehzangi and Karamizadeh, 2011; Kavousi et al., 2011; Geng et al., 2012; Dehzangi and Sattar, 2013; Habibi et al., 2013; Hsieh et al., 2013; Sharma et al., 2013a; Zangooei and Jalili, 2013). However, the most significant enhancements for this problem have been achieved by using attribute selection and feature extraction rather than relying on the classification techniques (Chen and Kurgan, 2007; Shamim et al., 2007; Dong et al., 2009; Ghanty and Pal, 2009; Shen and Chou, 2009; Chou, 2011; Haddow et al., 2011; Yang and Chen, 2011; Dehzangi and Phon-Amnuaisuk, 2011; Dehzangi et al., 2013a; Dehzangi et al., 2013b; Sharma et al., 2013b). Features that have been used for PFR can be generally categorised into three groups namely, *sequential-based*, *physicochemical-based* and *evolutionary-based* features. Early studies relied mainly on sequential-based (also called compositional-based) features for PFR. These features are extracted based on the alphabetic sequence of the proteins. Sequential-based features provide crucial information about the arrangement of the amino acids in a protein sequence. However, sequential-based features have two main drawbacks. First, when the sequential similarity is low then the recognition performance is low. Second, they do not incorporate any physicochemical-based information. Therefore, later studies shifted their focus to use

features extracted from the physicochemical-based attributes to address these two issues. Physicochemical-based attributes refer to physical, chemical and physicochemical properties of the amino acids and proteins such as hydrophobicity or polarity. The features extracted from the physicochemical attributes have an advantage that they do not depend on sequential similarity. Hence, the discriminatory information of these features is not affected even when the sequential similarity is low (Ghanty and Pal, 2009). The Ding and Dubchak study (2001) brought a tremendous attention to the physicochemical-based features. They used five popular physicochemical-based attributes namely, *normalised frequency of α-helix, hydrophobicity, polarity, polarisability and van der Waals volume* for feature extraction and significantly outperformed previous studies. Their extracted features have also been widely used in the later studies (Shen and Chou, 2009; Dehzangi et al., 2010b, Dehzangi et al., 2010c; Kavousi et al., 2011; Yang et al., 2011; Chmielnicki and Stapor, 2012).

Despite the promising results achieved by using physicochemical-based features, the effect of using a wide range of physicochemical-based attributes on PFR has not been adequately explored. To the best of our knowledge, the only study that explored the impact of a wide range of physicochemical-based attributes has been conducted by Gromiha (2005). They explored the impact of 49 different physicochemical-based attributes on the folding process. However, they merely extracted the global density based on each attributes for each protein (total 49 features were extracted from 49 attributes). Their extracted features do not reveal adequately the potential local discriminatory information of the attributes.

Recent studies shifted the focus to explore evolutionary-based features for PFR and significant improvement in prediction accuracy has been observed (Shamim et al., 2007; Kurgan et al., 2008; Dong et al., 2009; Shen and Chou, 2009; Yang and Chen, 2011; Dehzangi et al., 2014). Evolutionary-based features are extracted based on the probability of substitution of amino acids through their evolution process. To extract these features, instead of using original protein sequence, sequential-based features are extracted mainly from the *Position Specific Scoring Matrix* (*PSSM*) calculated from PSIBLAST (Altschul et al., 1997). However, similar to the sequential-based features, evolutionary-based features dramatically lose their discriminatory information when the sequential similarity rate is low. They also do not incorporate physicochemical-based information. To include physicochemical-based information, predicted secondary structure using PSIPRED machine (Jones, 1999) were used (Shen and Chou, 2009; Yang and Chen, 2011). PSIPRED predicts the secondary structure of the proteins with over 80% prediction accuracy. Using PSIPRED helped improving the PFR performance, however, due to the limited secondary structure prediction accuracy of PSIPRED, it does not provide reliable and adequate information for solving this problem (especially for the prediction accuracy over 80% (Nguyen and Rajapakse, 2007; Ghanty and Pal, 2009; Yang et al., 2011)).

In this study, to explore the highlighted insufficiencies and in order to enhance the protein fold prediction accuracy, a mixture of evolutionary-based and physicochemical-based feature extraction method is proposed. We first transform the protein sequences using evolutionary-based information obtained from the PSSM and then extract the physicochemical-based features from it. This approach enables us to obtain benefit from both the evolutionary-based and physicochemical-based information simultaneously. Therefore, intuitively we are able to provide more information for PFR. Thereafter, we explore 55 different physicochemical-based attributes and propose two novel feature

extraction methods (based on segmented density and segmented distribution) for PFR. To study the effectiveness of the proposed schemes, four classifiers namely, *AdaBoost.M1*, *Random Forest*, *Naive Bayes* and *Support Vector Machine (SVM)* are used. We first, analyse the prediction accuracy of the explored attributes considering the proposed feature extraction methods. Based on this analysis, we obtain eight different feature sets consisting of extracted features from the selected attributes. Then we concatenate two feature groups extracted based on the concept of composition and auto-covariance of the amino acids obtained directly from the PSSM to add more sequential-based and evolutionary-based information to our extracted features. As a result, these features capture more sequential, physicochemical and evolutionary information simultaneously. Finally, we explore the performance of the employed classifiers on these feature sets and found that SVM attains the best results. Then by using SVM, we show the protein fold prediction accuracy to be 5% better than previously reported results found in the literature.

2 Benchmarks and physicochemical-based attributes

In this study, we use two popular benchmarks to evaluate and assess the generality and performance of our proposed methods. To be able to directly compare our results with the state-of-the-art approaches found in the literature, we use the extended version of the Ding and Dubchak (2001) (DD) benchmark called EDD. We extract EDD from 1.75 *Structural Classification of Proteins (SCOP)* (Murzin et al., 1995) consisting of 3418 proteins with less than 40% sequential similarity belonging to the same 27-folds used in the DD benchmark. Note that due to the large number of duplications in the DD benchmark and its inconsistencies with new version of the SCOP (Dehzangi and Phon-Amnuaisuk, 2011; Yang and Chen, 2011), this benchmark is not explored in this study. In addition, we used the TG benchmark introduced by Taguchi and Gromiha (2007). This benchmark consists of 1612 proteins with less than 25% sequential similarity belonging to 30 different folds that was extracted from 1.73 SCOP.

We also study 55 different physicochemical-based attributes as listed in Table 1 and explored their effectiveness on PFR. In Table 1, column three shows the attributes' names and column one shows their corresponding numbers. From here onwards, we use the numbers to define the corresponding attributes. These attributes are taken from the APDbase (Mathura and Kolippakkam, 2005), the AAindex (Kawashima et al., 2008) and Gromiha's study (2005). The list and the numerical valued (normalised) calculated for each physicochemical-based attribute with respect to its references is provided in the Supplementary Material. In this experimental study, the aim is to explore the potential of each attribute to enhance the PFR performance. We also aim to address the issue of multi-referencing by finding the best reference for a specific attribute (given a feature extraction method). For instance, hydrophobicity has attributes number 1, 8, 9 and 11 (Table 1) and polarity has attribute numbers 13 and 55 (Table 1) and it is not clear which one would perform the best given the feature extraction method. To the best of our knowledge this issue has not been studied adequately in the literature which will be described in Section 5. We have also provided the supplementary data showing the normalised and real values assigned to each amino acid based on the studied attributes.

Table 1 Names and number of the explored attributes in this study

No.	*Reference*	*Attributes*
1	(Casari and Sippl, 1992)	Structure derived hydrophobicity value
2	(Charton and Charton, 1982)	Polarisability
3	(Chou and Fasman, 1978)	Normalised frequency of α-helix
4	(Chou and Fasman, 1978)	Normalised frequency of β-strand
5	(Chou and Fasman, 1978)	Normalised frequency of β turn
6	(Cowan and Whittaker, 1990)	Hydrophobicity at ph 7.5 by HPLC
7	(Dawson, 1972)	Size
8	(Eisenberg et al., 1984)	Consensus normalised hydrophobicity scale
9	(Engelman et al., 1986)	Hyd. index base on helix in membrane
10	(Nelson and Cox, 2008)	Molecular weight
11	(Fauchere and Pliska, 1983)	Hydrophobic parameter
12	(Fauchere et al., 1988)	Van Der Waals volume
13	(Grantham, 1974)	Polarity (driven from amino acids)
14	(Nelson and Cox, 2008)	Volume
15	(Ponnuswamy et al., 1980)	Compressibility
16	(Gromiha, 2005)	Average long range contact energy
17	(Gromiha, 2005)	Average medium range contact energy
18	(Gromiha, 2005)	Long range non bounded energy
19	(Gromiha, 2005)	Mean RMS fluctuational displacement
20	(Gromiha, 2005)	Refractive index
21	(Gromiha, 2005)	Solvent accessible reduction
22	(Gromiha, 2005)	Total non bounded energy
23	(Gromiha, 2005)	Unfolding entropy change of hydration
24	(Gromiha, 2005)	Unfolding hydration heat capacity change
25	(Guo et al., 1986)	Retention coefficient (retention times PH= 7.0)
26	(Guy, 1985)	Amino acids partition energy
27	(Nelson and Cox, 2008)	PKa-COOH
28	(Hopp and Woods, 1981)	Hyd. value (driven from free amino acids)
29	(Hutches, 2010)	Absolute entropy
30	(Hutches, 2010)	Entropy of formation
31	(Janin, 1979)	Buried and accessible molar fraction ratio
32	(Janin, 1979)	Energy of transfer from inside to outside
33	(Karplus and Schulz, 1985)	Flexibility for one rigid residue
34	(Krigbaum and Komoriya, 1979)	Side chain interaction parameter
35	(Krigbaum and Komoriya, 1979)	Side chain volume
36	(Kyte and Doolitle, 1982)	Hydropathy index
37	(Manavalan and Ponnuswamy, 1978)	Average surrounding hydrophobicity
38	(Meirovitch et al., 1980)	Average reduced distance for side chain

Table 1 Names and number of the explored attributes in this study (continued)

No.	*Reference*	*Attributes*
39	(Meirovitch et al., 1980)	Side chain orientation angle
40	(Miyazawa and Jernigan, 1985)	Ave number of nearest neighbor in chain
41	(Miyazawa and Jernigan, 1985)	Average Volume of surrounding residues
42	(Miyazawa and Jernigan, 1985)	Hyd. scale (contact energy derived from 3D data)
43	(Fauchere and Pliska, 1983)	Partition coefficient
44	(Ponnuswamy et al., 1980)	Average gain in surrounding hydrophobicity
45	(Ponnuswamy et al., 1980)	Surrounding hydrophobicity in α-helix
46	(Ponnuswamy et al., 1980)	Surrounding hydrophobicity in β-sheet
47	(Ponnuswamy et al., 1980)	Surrounding hydrophobicity in β turn
48	(Ponnuswamy et al., 1980)	Surrounding hydrophobicity in folded form
49	(Ponnuswamy et al., 1980)	Average number of surrounding residues
50	(Rao and Argos, 1986)	Membrane buried helix parameter
51	(Rose et al., 1985)	Mean fractional area loss (f)
52	(Vihinen et al., 1994)	Flexibility
53	(Wolfenden et al., 1981)	Hydration potential (transfer vapor to water at ph 7.0)
54	(Zimmerman et al., 1968)	Bulkiness
55	(Zimmerman et al., 1968)	Polarity (driven from amino acids in proteins)

3 Feature extraction approaches

In this study, we aim to propose a mixture of feature extraction model that incorporate sequential-based, physicochemical-based and evolutionary-based information simultaneously. In the proposed mixture model, we first find the consensus sequences from the original protein by using PSSM. This step embeds sequential-based and evolutionary-based information. Then we extract physicochemical-based features from this consensus sequence. In this way, we obtain information from all the three sequential-based, physicochemical-based and evolutionary-based features. As it will be shown later, this approach provides more discriminatory information compared to the conventional method.

3.1 Consensus sequence extraction

In this study, PSSM obtained by running PSIBLAST on our employed benchmarks (using NCBI's non-redundant (NR) protein database and the cutoff E-value set to 0.001). PSSM consists of two $L \times 20$ (L is the length of a protein) matrices namely, PSSM_cons and PSSM_prob matrices. PSSM_cons and PSSM_Prob matrices respectively give the log-odds and normalised probability of the substitution score of an amino acid with other amino acids depending upon their positions. In the previous studies, the consensus sequence ($C_1,C_2,C_3,\ldots,C_L$) was extracted in the way that an amino acid in the original sequence ($O_1,O_2,O_3,\ldots,O_L$) was replaced by the amino acid that had the highest substitution score in the PSSM_cons matrix mainly. Let S_{ij} be the substitution score of an

amino acid *i* with an amino acid *j* in the PSSM_cons, then the index I_i of the amino acid with the highest substitution score in the PSSM for the *i*-th amino acid in a protein sequence is given as follows:

$$I_i = argmax\{S_{ij} : 1 \le j \le 20\}, 1 \le i \le L \quad (1)$$

The original amino acid in a protein sequence is then replaced by the I_i^{th} amino acid (as they are ordered in the PSSM). By exploring the PSSM_cons matrix, we observed that for the case when there is an unknown amino acid in the original protein sequence, the results of the PSSM_cons matrix for this specific unknown protein is a row with all the elements are equal to –1. Therefore, relying solely on this matrix does not provide any information about these unknown amino acids. On the other hand, for a given protein, if a similar sequence is detected in the in NR database, the substitution score of these unknown amino acids change in the PSSM_prob matrix (if no hit is found, the probabilities will be equal to zero which rarely occurs.). Therefore, by considering the PSSM_prob instead of the PSSM_cons, we can effectively address the issue of unknown amino acids. Furthermore, due to the case that the PSSM_cons consists of log-odds of the substitution score, it is more probable to find multiple values in this matrix as the maximum value which reduces the chance of choosing the best candidate for the replacement in the consensus sequence (conventionally, it was chosen randomly among the maximum values). On the other hand, the occurrence of multiple maximum substitution score for an amino acid based on its position is much less frequent (more accurately distinguished due to the better precision) in the PSSM_prob (normalised probability of substitution score).

Therefore, based on our observation we propose a novel approach for consensus sequence extraction. In this approach, for each amino acid, we first check the PSSM_prob. In the case that a unique maximum is found in this matrix, it will be replaced the amino acids in the original sequence. Otherwise, we refer to the PSSM_cons to check weather there is a unique maximum can be found. In case that there is a unique maximum is in PSSM_cons, it will be replaced the amino acid in the sequence. Otherwise, the first maximum in the PSSM_prob will be replaced within the original sequence. To address the issue of unknown amino acids, we merely rely on the PSSM_prob. In case that a maximum probability of the substitution score spotted in this matrix, the nominated amino acids is replaced the unknown amino acid in the original sequence. In the case that the PSSM_prob is zero (no hits found in the NR dataset), the unknown amino acid is transferred to the consensus sequence, unchanged. Using our proposed approach, we reduce the number of unknown amino acids in the consensus sequence for the EDD benchmark from 360 to 2 (two unknown amino acids in one protein).

As a result, we are able to produce a more precise and accurate consensus sequence compared to the previously used method. We also address the issue of unknown amino acids using evolutionary-based information that have not been addressed adequately in the previous studies. In continuation, it will be shown that the consensus sequence extracted using the proposed method in this study also provides more discriminatory information compared to the previous methods used for this task.

3.2 Physicochemical-based feature extraction method

As it was discussed earlier in the introduction section, despite the importance of the physicochemical-based features, the impact of a wide range of the physicochemical-based attributes has not been explored adequately for the PFR. In most of the cases, either a few popular attributes were used (Ding and Dubchak, 2001; Shamim et al., 2007; Dehzangi and Phon-Amnuaisuk, 2011) or when a wider range of attributes are explored, the adopted feature extraction method did not explore the potential local discriminatory information of the attributes adequately. For example, the global density feature based on each attribute employed by Gromiha (2005) captures only global information. In this paper, we use more sophisticated features to capture local information from the physicochemical-based attribute sequence. Two such feature extraction methods are proposed here based on the concepts of segmented density and segmented distribution to provide more local discriminatory information for the PFR. Proposed approaches aim to provide better understanding about the studied features for this problem.

3.2.1 Segmented density

This method is mainly proposed to add more local discriminatory information based on the density of a given attribute. In this approach, we first transform the original protein sequence to the protein consensus sequence using PSSM. Then, we assign numerical values to the amino acids along the protein consensus sequence based on a given physicochemical attribute (R_1,R_2,R_3,…,R_L). transform the original protein sequence to consensus sequence by using PSSM and then we assign numerical values to the amino acids based on a given physicochemical-based attribute. Then the sequence is segmented (using 5% segmentation factor) by dividing it equally into 20 segments. This value of segmentation factor is considered in this study because the shortest protein in our benchmarks consists of 23 amino acids and to have at least one amino acid in each segment. Choosing 5% segmentation factor also showed better performance than using 10% and 25% segmentation factors which were explored in our previous studies (Dehzangi and Phon-Amnuaisuk, 2011; Dehzangi et al., 2013a) and also were experimentally investigated by the authors and its results are provided and attached to this article as the Supplementary Material. The expression for segmented density can be given as follows:

$$D_{seg\_density} = \frac{\sum_{i=1}^{M} R_i}{M} \tag{2}$$

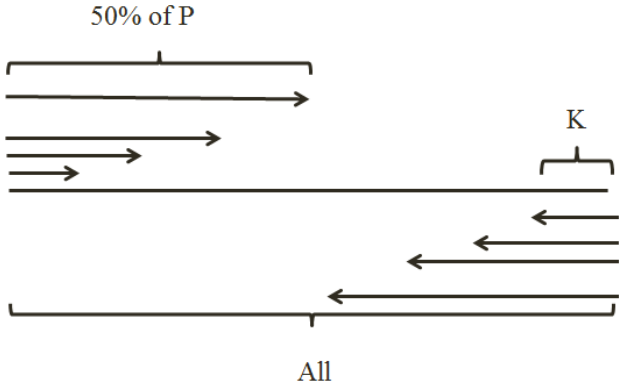
where M ($=L \times (5/100)$) is the length of each segment. This gives a set of 20 segmented-density features. To this set, we add the global density feature to make a final set of 21 (= 20 + 1) density features. The expression for global density is given as follows:

$$D_{glob\_density} = \frac{\sum_{i=1}^{L} R_i}{L} \tag{3}$$

3.2.2 Segmented distribution

As mentioned earlier, in the segmented density approach, the segments of a given protein sequence are of equal lengths and each segment is represented by a density feature given in equation (2). In this section, we propose another segmentation method where segments of a protein sequence are of unequal lengths and each segment is represented by a distribution feature which is computed as follows. We first compute the total sum of attribute values over the protein sequence which is equal to $T=\sum_{i=1}^{L} R_i$. Then, we start from the left hand side of the protein sequence and compute the partial sum of attribute values for the first I amino acids which is given by $P=\sum_{i=1}^{I} R_i$. Using the distribution factor K (which is a parameter investigated in this study), we find out the maximum value $I_{max}^{(1)}$ of index I such that partial sum P is less than or equal to K% of total sum T. Thus we can say that the first $I_{max}^{(1)}$ amino acids of the protein sequence contribute to K% of the total sum T. We use $I_{max}^{(1)}$ to define the ending location of the first segment, while its beginning point is taken to be 1. The distribution feature of this segment is given by $I_{max}^{(1)}/L$. In a similar manner, we find out the number of first $I_{max}^{(2)}, I_{max}^{(3)}, ..., I_{max}^{(50/K)}$ amino acids of the protein sequence that contribute to $2K$%,$3K$%,…,50% of T, respectively. Indices $I_{max}^{(2)}, I_{max}^{(3)}, ..., I_{max}^{(50/K)}$ are used to define the ending locations of segments 2,3,...,50/K, respectively; while the beginning location of all these segments remains to be 1. The distribution features for these segments are computed as $I_{max}^{(i)}/L, i=2,3,...,50/K$. Note that we have thus computed 50/K distribution features by processing the protein sequence from the left to the right direction. We repeat this process from right to left direction to get another set of 50/K features. Thus, the total of 2 × (50/K) = 100/K distribution features are computed in this study. This procedure is shown in Figure 1. The distribution factor *(K)* is a parameter which is determined here experimentally. For this, three values of K (5, 10 and 25) are investigated.

Figure 1 Segmented distribution-based feature extraction method



To this set of *100/K* distribution features, we add the global density feature to provide more global information. Therefore, we have a total of 1 + (100/K) features. Thus there will be 21, 11 and 6 features for K = 5, 10 and 25, respectively.

The main contributions of the segmented-based density and segmented-based distribution feature extraction methods can be highlighted in two following points. First, it provides more information based on a given attribute based on the concept of segmentation. Therefore, it is able to provides better opportunity to more appropriately explore the potential discriminatory information of the studied attributes. It is also able to add more local discriminatory information to the previously used methods (e.g., global density used by Gromiha (2005)). Second, different to previous studies, instead of calculating the distribution from one side or categorising amino acids into subgroups based on a given attribute (Ding and Dubchak, 2001; Deschavanne and Tuffery, 2009), we calculate the distribution from both sides of the proteins. In this way, instead of providing cumulative distribution in the rear side of a protein, we calculate the distribution from both sides that shifts the emphasis of the distribution calculation to the sides of the proteins. This modification is made due to two main reasons. First, the cumulative distribution reduces the general impact of the distribution feature. Second, due to the flexibility of the ends of a protein, they can play more crucial role on the folding process. Avoiding categorisation of the amino acids into different groups also helps to maintain the information that might be lost through simplification (alphabet reduction (Ding and Dubchak, 2001; Deschavanne and Tuffery, 2009)).

3.3 Evolutionary-based features

In addition to our extracted attributes based on the mixture of the physicochemical and evolutionary-based features, we extract two feature groups directly from the PSSM namely, Semi-AC and PSSM-AC. The aim of extracting these two feature groups is to provide more information based on the sequential and evolutionary-based information. Using these feature groups also enable us to explore the discriminatory information of all three categories of features simultaneously (sequential, physicochemical and evolutionary-based features). In the first method, we calculate semi-composition of the amino acids. It is called semi-composition (semi_AC) because instead of using the protein sequence directly to calculate the percentage of the occurrence of each amino acid, we calculate the average of the substitution score for each amino acids directly from the PSSM. The Semi_AC is calculated as follows:

$$C_i = \frac{1}{L}\sum_{i=1}^{L} S_{ij}, (j = 1, ..., 20) \tag{4}$$

This feature is able to provide more discriminatory information compared to the consensus sequence in the sense of extracting the amino acids composition feature group (Liu et al., 2012). The second sequential-based feature is extracted based on the concept of PSSM-AC which was successfully used for the PFR (Dong et al., 2009). For each amino acids, PSSM-AC gives the auto-covariance of the substitution score of a given amino acid with other amino acids with at most T distance factor (in this study, T is set to 10 as it was shown as the most effective distance factor for the PSSM-AC by Dong et al. (2009)). PSSM-AC is calculated as follows:

$$PSSM - AC_{F,j} = \frac{1}{L-F}\sum_{i=1}^{L-F}(S_{i,j} - S_{ave,j})(S_{i+F,j} - S_{ave,j}), \quad (j = 1, ..., 20 \& F = 1, ..., T) \tag{5}$$

where $S_{ave,\ j}$ is the average of substitution scores of the amino acid i with other amino acids along the protein sequence. Therefore, in total $20 \times T$ features are calculated in this feature group.

4 Classification techniques

In this study, to explore the performance of the proposed approaches, several classification techniques such as, AdaBoost.M1, Naive Bayes, Random Forest and SVM are used.The employed classifiers are selected based on their popularity (Chen and Kurgan, 2007; Dehzangi et al., 2010b), their diversity in learning techniques (Dietterich, 2000; Dehzangi et al., 2010a; Dehzangi and Karamizadeh, 2011) and their performances in the previous studies found in the literature (Chen and Kurgan, 2007; Gromiha, 2009; Jain et al., 2009; Dehzangi et al., 2010b). These methods are briefly described as follows.

Naive bayes: As a kind of a baysian-based learner is considered as one of the simplest classifiers yet attained promising results for many different tasks including PFR. Naive bayes is based on the naive assumption of independency of the employed features from each other to calculate the posterior probability (Dehzangi et al., 2010a). In most of the real world problems, its assumptions is not valid, however, Naive Bayes has showed promising performance. It also provides important information about the correlation of the employed features. The highest results (55.3%) found in the literature for the TG benchmark was achieved using this classifier (Gromiha, 2009). We have used Naive Bayes classifier implemented in the WEKA toolbox which is designed for multi-class classification tasks (as it was used by Dehzangi et al. (2009) and Gromiha (2009)).

AdaBoost.M1: Multi-class Adaptive boosting (AdaBoost.Ma) is an extension of the AdaBoost method introduced by Freund and Schapire (1995) for multi-class classification task. It is considered as the best-of-the-shelf meta-classifier (aims at producing a strong classifier by boosting a weak learner using different approaches) and attained promising results for different tasks as well as the PFR (Freund and Schapire, 1996; Dehzangi et al., 2010a). The main idea of the AdaBoost.M1 is to sequentially (in *Iter* iterations) apply a base learner (also called weak learner which refers to a classifier that at least performs better than random guess) on the bootstrap samples of data, adjust the weights of misclassified samples and enhances the performance in each step (Freund and Schapire, 1996). In this study, Adaboost.M1 implemented in WEKA is used (Witten and Frank, 2005). The C4.5 decision tree is used as its base learner and the number of iterations is set to 100 (*Iter* = 100) (this is shown as the best parameter for this algorithm for the PFR (Dehzangi et al., 2010a; Dehzangi and Phon-Amnuaisuk, 2011).

Random forest: Is also considered as a kind of meta-learner which recently attracted tremendous attention specifically for the PFR (Jain et al., 2009; Dehzangi et al., 2010b). Unlike Adaboost.M1, Random Forest is based on the bagging (Breiman, 2001). It applies a base learner independently on B different bootstrap samples of data using randomly selected subsets of features. Despite its simplicity, it has showed significant potential to encourage diversity which is considered as an important factor with profound impact on the performance of the meta-classifiers (Dietterich, 2000). Random Forest have been successfully used for the PFR and its similar studies and outperformed most of the classifiers used for this task (Jain et al., 2009; Dehzangi et al., 2010b). In this study, for the Random Forest (implemented in WEKA and designed for multi-class classification task) the number of iteration is set to 100 (k = 100) and random tree based on the gain ratio is used as its base learner.

Support vector machine: Introduced by Vapnik (1999) is considered as the state-of-the-art classification technique which also attained the best results for the PFR (Dong et al., 2009; Yang and Chen, 2011). It aims at minimising the classification error by finding the *Maximal Marginal Hyperplane (MMH)* based on the concept of support vector theory. To find the appropriate support vector, it transforms the input data to the higher dimension using the concept of the kernel function. *Polynomial* and *Radial Base Function (RBF)* kernels are considered as the best kernels used for the SVM classifier to tackle the PFR. Due to the importance of this classifier as well as its promising performance, three different SVM-based classifiers are used to explore our proposed approaches. We use SVM with *Sequential Minimal Optimisation (SMO)* (as a kind of polynomial kernel (implemented in WEKA)) in two different experiments in which in the first experiment, its kernel degree is set to one and in the second experiment, its kernel is set to three (p = 1 and 3). We also use SVM using RBF kernel implemented in LIBSVM (Chang and Lin, 2011) with its parameters (γ and C) are optimised using SVMgrid algorithm which is also implemented in the LIBSVM. For all these three SVM-based classifiers, we have used one-versus-one approach to adopt this classifier for multi-class classification task.

5 Results and discussion

In order to evaluate the performance of our proposed methods, we carried out the experiments in two parts. In the first part, we partition the data into training set (having 3/5 of data) and test set (having 2/5 of data) to simulate the condition of previously proposed approaches to study the impact of the physicochemical-based feature extraction method (Ding and Dubchak, 2001) and in the second part, we use tenfold cross-validation procedure on the employed datasets for an exhaustive run and to compare our results with the best results reported in the literature for PFR.

5.1 Part one

In the first part,our aim is to study the impact of the proposed consensus sequence extraction method, extracting physicochemical-based features from the consensus sequence rather than original protein sequence and analyse our proposed feature extraction methods (segmented density and segmented distribution).

5.1.1 The impact of the proposed consensus sequence extraction method

In this section, we evaluate the performance (in terms of classification accuracy) of the proposed consensus sequence extraction method. To do this, we first extract features by computing frequency of amino acid composition and frequency of amino acid occurrence (Taguchi and Gromiha, 2007) from the following three cases: (1) original protein sequence; (2) the consensus sequence derived using the conventional way (Yang and Chen, 2011) and (3) the consensus sequence derived by using our proposed way. We call the extraction of features from case 1, 2 and 3 as Methods I, II and III respectively. Next, we compare the classification accuracies of the extracted features using the four classifiers (Adaboost.M1, Random Forest, SVM, Naive Bayes). The classification accuracies on two benchmarks (EDD and TG) are shown in Table 2. It can be seen from

Table 2 that Method III outperforms other methods consistently for all the for classifiers on both of the benchmark datasets. Therefore, our proposed method for the consensus sequence extraction method is used for the rest of this study.

Table 2 Comparison of the achieved classification accuracy (%) using Adaboos.M1 (Ada), Random Forest (RF), Naive Bayes (NB) and SVM (using SMO and $p = 1$) to evaluate the proposed consensus sequence (Method III) extraction method compared to use of original sequence (Method I) as well as previously used method (Method II)

Dataset	*Method*	*Composition of the amino acids*				*Occurrence of the amino acids*			
		Ada	*RF*	*SVM*	*NB*	*Ada*	*RF*	*SVM*	*NB*
*EDD	Method I	35.9	36.6	32.4	34.9	43.0	42.4	41.2	34.4
	Method II	47.5	46.7	42.2	44.0	55.0	55.4	48.2	41.5
	Method III	50.7	50.4	44.4	45.5	55.8	56.7	48.9	42.6
*TG	Method I	32.7	34.4	31.6	29.3	34.9	36.4	33.6	30.1
	Method II	37.4	38.0	34.7	37.1	45.3	44.6	38.6	33.3
	Method III	41.0	41.3	36.3	39.6	47.0	47.2	38.8	34.4

5.1.2 The impact of the proposed mixture model

In this section, we analyse the proposed segmented distribution and segmented density feature extraction methods. For segmented distribution factor K which are 5%, 10% and 25% and we use $K = 5\%$ for segmented density. We transform the original protein sequence to the consensus sequence using our proposed approach (as discussed in Sections 3.1 and 5.1.1) to compare the effectiveness of features extracted from these two type of sequences. The dimensionality of a feature vector extracted from an original/consensus protein sequence by segmented distribution using $K = 5\%$ is 21, using $K = 10\%$ is 11 and using $K = 25\%$ is 5 and by segmented density using $K = 5\%$ is 21. Therefore, four different ways of extracting features will give four sets of feature vectors. These feature vectors are then processed through a classifier to get classification accuracy. There are four classifiers used in this study (Adaboost.M1, Random Forest, SVM and Naive Bayes). Note that there are 55 attributes and therefore we will get 55 values of classification accuracies for each method of extracting features and for each classifier. These classification accuracies are calculated for the original protein sequences as well as for the consensus sequences. Since these values are quite extensive to show here in the paper, we present them as the Supplementary Material.

For presentation, here we first compute the average and maximum of the classification accuracy over 55 attributes by using the consensus sequence for a given feature extraction approach. Similarly, we then compute the values by using the original protein sequence. We then subtract the average classification accuracy obtained by using the original protein sequence from the average classification accuracy obtained by using the consensus sequence. In a similar manner, we subtract the maximum classification accuracies obtained by using the original protein sequence and consensus sequence. The results are shown in Table 3. The prediction accuracy achieved using all four employed classifiers show that extracting features from the consensus protein sequence consistently performs better than using the original protein sequence for feature extraction. Despite the small number of extracted features (5, 11 and 21 features) the average and maximum prediction performance are significantly increased. The enhancement achieves by using

the proposed mixture of feature extraction method also suggests a new approach to obtain benefit from the evolutionary information and to provide crucial information about the impact of the physicochemical-based attributes on the PFR, simultaneously.

Table 3 Comparison of the achieved results (%) using explored classifiers to evaluate mixture model for feature extraction

Adaboost.M1				
EDD				
Average	2.0	2.4	2.7	3.0
Maximum	1.4	3.7	3.2	4.3
TG				
Average	0.8	1.4	1.3	1.1
Maximum	1.9	3.5	1.6	0.6
Random Forest				
EDD				
Average	2.1	2.6	2.60	2.8
Maximum	3.6	3.8	2.7	2.7
TG				
Average	1.2	1.2	1.2	1.3
Maximum	0.2	2.7	1.7	1.6
SVM				
EDD				
Average	2.0	1.9	2.0	2.2
Maximum	2.4	0.5	1.1	2.1
TG				
Average	1.3	0.9	1.0	1.4
Maximum	1.4	1.0	2.2	1.6
Naive Bayes				
EDD				
Average	1.6	1.8	1.9	2.6
Maximum	1.3	1.8	1.5	3.1
TG				
Average	0.5	0.5	0.4	1.3
Maximum	0.2	1.7	1.6	3.2

Notes: In each case, the enhancement of the average values (average mixture – average original) and maximum values (maximum mixture – maximum original) are shown. The maximum refers to the highest results achieved for a specific attributes. From left to right, column shows the performance for segmented distribution using 25%, 10% and 5% distribution factor and 5% segmented density, respectively.

5.1.3 The impact of the segmentation factor on the segmented-based distribution method

In this section we study the impact of segmentation factor on the segmented-based distribution method in the following two steps. First, for a given classifier, we calculate the average and maximum classification accuracies as it was done in the previous subsection. Then for a given classifier, we subtract the maximum and average values calculated using segmented-based distribution with K = 25% feature extraction method from the average and maximum values calculated using segmented-based distribution with K = 10% as well as K = 5%. As it is shown in Table.4, by trivially increasing the number of extracted features by adjusting segmentation factor from 25% to 5% and from 25% to 10%, the enhancements is noticeable. Note that the performance of Naive Bayes did not improve due to increase in correlation of the extracted features. Therefore, the enhancement of the other three classifiers are shown and compared in this subsection. As it is shown in Table.4, the enhancement for the Random Forest and the Adaboost.M1 classifiers using 5% distribution factor compared to 25% is higher than 25% and 10% while for the SVM classifier it is not significant. This phenomenon emphasises the impact of each approach based on the classifier being used.

Table 4 Comparison of the achieved results (%) using Adaboos.M1, Random Forest and SVM (using SMO and p = 1) to evaluate the enhancement achieved considering the segmentation-based distribution approach

EDD		
AdaBoost.M1	*From 25% to 5%*	*From 25% to 10%*
Average	8.3	5.7
Maximum	11.3	10.2
Random Forest	*From 25% to 5%*	*From 25% to 10%*
Average	7.8	5.6
Maximum	11.5	9.3
SVM	*From 25% to 5%*	*From 25% to 10%*
Average	3.8	2.1
Maximum	7.1	6.5
TG		
AdaBoost.M1	*From 25% to 5%*	*From 25% to 10%*
Average	6.6	4.3
Maximum	10.3	7.9
Random Forest	*From 25% to 5%*	*From 25% to 10%*
Average	6.9	4.8
Maximum	12.2	7.3
SVM	*From 25% to 5%*	*From 25% to 10%*
Average	3.6	2.1
Maximum	6.9	7.1

5.2 Part two

In the second part and final stage, we aim to tackle the state-of-the-art approaches used for the PFR. Therefore, we have explored the impact of the proposed approaches for the employed benchmarks (EDD and TG) using tenfold cross-validation.

5.2.1 Comparison of the results achieved in this study with the best results found in the literature

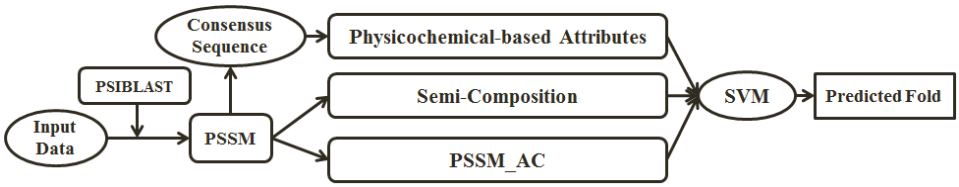
In this part, we have extracted eight different feature sets consisting of a combination of features extracted from different attributes using our proposed feature extraction methods in the following two steps. We first study the performance of a given classifier, based on the employed feature extraction method (explored on the TG benchmark). Then, based on each classifier, two feature set are constructed such that each feature set consists of features extracted using a similar feature extraction method and attained the best results for the studied classifier (eight combinations in total). These feature sets have been constructed in the manner that maintains the number of employed features remains small. In the following paragraph, attributes as well as feature extraction method used to build each of our eight feature sets are explained. For simplicity, we refer to each attribute by its number as in Table 1.

The first and second combinations are extracted respectively based on the performance of the Adaboost.M1 classifier on the segmented-based distribution (with K = 10%) (attribute numbers: 3, 4, 5, 14, 17, 26, 28, 30, 33, 41, 48 = 121 features) and the segmented-based density (with K = 5%) feature extraction methods (attributes numbers: 1, 3, 4, 20, 54, 55 = 126 features). The third and fourth are extracted based on the performances of the Random Forest classifier on the segmented-based density (with K = 5%) (1, 3, 16, 17, 41, 55 = 126 features) and the segmented-based distribution (with K = 10%) (3, 4, 5, 14, 16, 17, 26, 28, 30, 41, 44, 48 = 132 features) feature extraction approaches. The fifth and sixth combinations are extracted based on the performances of the SVM classifier on the segmented-based distribution (with K = 25%) (1, 3, 4, 5, 17, 27, 29, 30, 31, 33, 35, 37, 38, 39, 40, 41, 44, 47, 48, 55 = 100 features) and the segmented-based distribution (with K = 5%) (3, 5, 15,17, 30, 41, 44 = 147 features) feature extraction methods. Finally, the seventh and eighth are extracted based on the performances of the Naive Bayes classifier on the segmented-based distribution (with K = 25%) (1, 3, 4, 5, 14, 16, 17, 27, 29, 30, 31, 32, 33, 37, 38, 39, 40, 41, 44, 47, 48, 55 = 110 features) and the segmented-based density (with K = 5%) (3, 16, 17, 24, 33, 42 = 126 features) feature extraction methods.

It is important to highlight that most of the attributes used to construct these feature sets have not been used or adequately explored for the PFR. These attributes individually outperform most of the popular attributes used to tackle this problem (e.g., average long range contact energy (16), total non bounded energy (22) and mean fractional area loss (51)). In many cases, even for similar attributes, usage of references that have not been used for this problem (e.g., hydrophobicity scale extracted from the contact energy derived from the 3D data (42) for the hydrophobicity attribute) compared to popular references (consensus normalised hydrophobicity scale (9)) also showed better performance which highlights the demand for revision of current physicochemical-based attribute selection approaches. In continuation, the Semi_AC, PSSM_AC feature groups as well as the length of the amino acids feature (which attained good results in previous studies (Ghanty and Pal, 2009)) are added (221 features in total) to each extracted

combination of feature groups (which will be referred as comb_1 to comb_8). As it was shown in previous studies (Dong et al., 2009; Yang and Chen, 2011; Chmielnicki and Stapor, 2012; Dehzangi et al., 2013b), the SVM classifier attains the best results using evolutionary-based features extracted from PSSM and outperforms other classifiers used for the PFR. Therefore, in this part, we only report the results attained using SVM. The overall architecture of the proposed model is shown in Figure.2. In continuation, three SVM-based classifiers applied to the input feature vectors (SVM using SMO kernel with $p = 1$ and $p = 3$, as well as the SVM using the RBF kernel which its parameters optimised in the LIBSVM).

Figure 2 The overall architecture of the proposed approach



Among the employed SVM-based classifiers, SVM using SMO ($p = 1$) attains similar or in some cases slightly better results compared to the other SVM-based classifiers employed in this study which emphasises on the effectiveness of the input feature vector (discriminatory information provided by the employed features reduces the dependency of the performance on a more complex kernel). We also duplicate the experiments of Dong et al. (2009) which outperformed other methods used for the PFR (Shamim et al., 2007; Deschavanne and Tuffery, 2009; Kavousi et al., 2011; Yang and Chen, 2011). We have also extracted the 49-D feature vector introduced by Gromiha (2005) and added the Semi_AC, the PSSM_AC and the length of the protein sequence (221 features) to it. Then we study the effectiveness of this feature set compared to the combination of features extracted in this study using similar classification technique (SVM with SMO ($p = 1$)). The best results achieved in this study, compared to the state-of-the-art results reported in previous studies are shown in Table.5.

As it is shown in Table.5, using the EDD and TG benchmarks, we significantly outperformed the results achieved by reproducing the Dong et al. (2009) results. We enhance the protein fold prediction accuracy by 5% and 4.8% achieving up to 83.1% and 63.7% for the EDD and TG benchmarks. These results are achieved using fewer features compared to the 4000 features used in the Dong et al. study (2009). We also compare our results with the results achieve using 49-D features extracted from the original protein sequence as well as the consensus sequence. As we can see, we significantly outperform these results (over 23% and 17% for the EDD and TG benchmarks respectively) as well which emphasises on the effectiveness of the proposed segmented-density and distribution-based feature extraction methods. In conclusion, suing our proposed approaches, we achieve several goals such as: improving the consensus sequence extraction method as well as addressing the issue of unknown proteins; enhancing discriminatory information based on the concept of the physicochemical-based features proposing a segmented-based approach; exploring a mixture model that simultaneously obtains benefit from the evolutionary-based and the physicochemical-based features; and finally, enhancing the protein fold prediction accuracy over than previously reported results found in the literature.

Table 5 The best results (%) achieved in this study compared to the best results achieved for the PFR

Study	*Features (No. of features)*	*Method*	*EDD*	*TG*
(Taguchi and Gromiha, 2007)	AAO original sequence (20)	LDA	46.9	36.3
(Taguchi and Gromiha, 2007)	AAC original sequence (20)	LDA	40.9	32.0
(Dehzangi and Phon-Amnuaisuk, 2011)	Physicochemical (219)	SVM	52.8	41.9
(Ding and Dubchak, 2001) Physicochemical (125)	Physicochemical (125)	SVM	50.1	39.5
(Ghanty and Pal, 2009)	Bi-gram (400)	SVM	75.2	52.7
(Ghanty and Pal, 2009)	Tri-gram (8000)	SVM	71.0	49.4
(Shamim et al., 2007)	Combination of bi-gram features (2400)	SVM	69.9	55.0
(Deschavanne and Tu_ery, 2009)	PSIPRED and PSSM-based features (242)	SVM	77.5	57.1
(Gromiha, 2009)	Threading (-)	Naive Bayes	70.3	55.3
(Dong et al., 2009)	ACCFold_ACC (4000)	SVM	78.1	58.9
(Dong et al., 2009)	ACCFold_AC (200)	SVM	76.2	56.4
This study	Comb_1 (342)	SVM	82.9	63.1
This study	Comb_5 (321)	SVM	82.8	63.5
This study	Comb_7 (331)	SVM	83.1	63.7
This study	Original sequence (49+221)	SVM	44.7	35.7
This study	Consensus sequence (49+221)	SVM	59.7	45.9

6 Conclusion

In this study, to enhance the protein fold prediction accuracy as well as providing more information about the impact of physicochemical-based attributes on the folding process, several approaches were proposed. In the first step, a new enhanced consensus sequence extraction method was proposed that enhanced the protein fold prediction accuracy compared to the previously used methods. It also addressed the issue of unknown proteins using evolutionary-based information. In the second step, we proposed two different feature extraction methods based on the concepts of segmented-distribution and density to provide more local discriminatory information. In the third step, we explored a wide range of physicochemical-based attributes (55 attributes) using the proposed feature extraction methods and four best-of-the-shelf classifiers were used for this task namely, Random Forest, Adaboost.M1, Naive Bayes and SVM. In the next step, we proposed a mixture of feature extraction approach to extract physicochemical-based features from the transformed protein sequence using evolutionary-based information. Our results suggested a new approach to explore physicochemical-based attributes in conjunction with evolutionary-based information for PFR. In the final step, by using a combination of a wide range of attributes that mostly have not been adequately explored in previous studies combined with sequential-based features, we achieved up to 83.1% and 63.7% prediction accuracy for the EDD and the TG benchmarks respectively, over 5% and 4.8% better than previously reported results found in the literature for these two benchmarks.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J. (1997) 'Gapped blast and psi-bast: a new generation of protein database search programs', *Nucleic Acids Research*, Vol. 25, No. 17, pp.3389–3402.

Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, No. 1, pp.5–32.

Casari, G. and Sippl, M.J. (1992) 'Structure-derived hydrophobic potential: Hydrophobic potential derived from x-ray structures of globular proteins is able to identify native folds', *Journal of Molecular Biology*, Vol. 224, No. 3, pp.725–732.

Chang, C.C. and Lin, C.J. (2011) 'LIBSVM: a library for support vector machines', *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp.27:1–27:27.

Charton, M. and Charton, B.I. (1982) 'The structural dependence of amino acid hydrophobicity parameters', *Journal of Theoretical Biology*, Vol. 99, No. 4, pp.629–644.

Chen, K. and Kurgan, L.A. (2007) 'Pfres: protein fold classification by using evolutionary information and predicted secondary structure', *Bioinformatics*, Vol. 23, No. 21, pp.2843–2850.

Chmielnicki, W. and Stapor, K. (2012) 'A hybrid discriminative-generative approach to protein fold recognition', *Neurocomputing*, Vol. 75, No. 1, pp.194–198.

Chou, K.C. (2011) 'Some remarks on protein attribute prediction and pseudo amino acid composition', *Journal of Theoretical Biology*, 273(1, pp.236–247.

Chou, P.Y. and Fasman, G.D. (1978) 'Empirical predictions of protein conformation', *Annual Review of Biochemistry*, Vol. 47, No. 1, pp.251–276.

Cowan, R. and Whittaker, R.G. (1990) 'Hydrophobicity indices for amino acid residues as determined by high-performance liquid chromatography', *Peptide Research*, Vol. 3, No. 2, pp.75–80.

Damoulas, T. and Girolami, M. (2008) 'Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection', *Bioinformatics*, Vol. 24, No. 10, pp.1264–1270.

Dawson, D.M. (1972) *The Biochemical Genetics of Man*, in Brock, D.J.H. and Mayo, O. (Eds), Academic Press, New York, NY, USA.

Dehzangi, A. and Karamizadeh, S. (2011) 'Solving protein fold prediction problem using fusion of heterogeneous classifiers', *INFORMATION, An International Interdisciplinary Journal*, Vol. 14, No. 11, pp.3611–3622.

Dehzangi, A. and Phon-Amnuaisuk, S. (2011) 'Fold prediction problem: the application of new physical and physicochemical-based features', *Protein and Peptide Letters*, Vol. 18, No. 2, pp.174–185.

Dehzangi, A. and Sattar, A. (2013) 'Ensemble of diversely trained support vector machines for protein fold recognition', *Proceedings of the 5th Asian Conference on Intelligent Information and Database Systems, ACIIDS05*, 18–20 March, Kuala Lumpur, Malaysia, pp.335–344.

Dehzangi, A., Phon-Amnuaisuk, S., Ng, K.H. and Mohandesi, E. (2009) 'Protein fold prediction problem using ensemble of classifiers', *Proceedings of the 16th International Conference on Neural Information Processing: Part II, ICONIP'09*, 1–5 December, Bangkok, Thailand, pp.503–511.

Dehzangi, A., Phon-Amnuaisuk, S. and Dehzangi, O. (2010a) 'Enhancing protein fold prediction accuracy by using ensemble of different classifiers', *Australian Journal of Intelligent Information Processing Systems*, Vol. 26, No. 4, pp.32–40.

Dehzangi, A., Phon-Amnuaisuk, S. and Dehzangi, O. (2010b) 'Using random forest for protein fold prediction problem: an empirical study', *Journal of Information Science and Engineering*, Vol. 26, No. 6, pp.1941–1956.

Dehzangi, A., Phon-Amnuaisuk, S., Manafi, M. and Safa, S. (2010c) 'Using rotation forest for protein fold prediction problem: an empirical study', *Proceedings of the 8th European conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, EvoBIO'10*, 7–9 April, Istanbul, Turkey, pp.217–227.

Dehzangi, A., Paliwal, K.K., Sharma, A., Dehzangi, O. and Sattar, A. (2013a) 'A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem', *IEEE Transaction on Computational Biology and Bioinformatics (TCBB)*, Vol. 10, No. 3, pp.564–575.

Dehzangi, A., Paliwal, K.K., Lyons, J., Sharma, A. and Sattar, A. (2013b) 'Enhancing protein fold prediction accuracy using evolutionary and structural features', *Proceeding of the 8th IAPR International Conference on Pattern Recognition in Bioinformatics, PRIB*, 17–20 June, Nice, France, pp.196–207.

Dehzangi, A., Paliwal, K.K., Lyons, J., Sharma, A. and Sattar, A. (2014) 'Proposing a highly accurate protein structural class predictor using segmentation-based features', *BMC Genomics*, Vol. 15, No. 1, p.S2.

Deschavanne, P. and Tuffery, P. (2009) 'Enhanced protein fold recognition using a structural alphabet', *Proteins: Structure, Function and Bioinformatics*, Vol. 76, No. 1, pp.129–137.

Dietterich, T.G. (2000) 'Ensemble methods in machine learning', *Proceedings of the 1st International Workshop on Multiple Classifier Systems, MCS'00*, London, UK, pp.1–15.

Ding, C. and Dubchak, I. (2001) 'Multi-class protein fold recognition using support vector machines and neural networks', *Bioinformatics*, Vol. 17, No. 4, pp.349–358.

Dong, Q., Zhou, S. and Guan, G. (2009) 'A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation', *Bioinformatics*, Vol. 25, No. 20, pp.2655–2662.

Eisenberg, D., Schwarz, E., Komaromy, M. and Wall, R. (1984) 'Analysis of membrane and surface protein sequences with the hydrophobic moment plot', *Journal of Molecular Biology*, Vol. 179, No. 1, pp.125–142.

Engelman, D.M., Steitz, T.A. and Goldman, A. (1986) 'Identifying nonpolar transbilayer helices in amino acid sequence of membrane proteins', *Annual Review of Biophysics and Biophysical Chemistry*, Vol. 15, pp.321–353.

Fauchere, J.L. and Pliska, V. (1983) 'Hydrophobic parameters pi of amino-acid side chains from the partitioning of n-acetyl-amino-acid amides', *European Journal of Medicinal Chemistry*, Vol. 18, pp.369–375.

Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V. (1988) 'Amino acid side chain parameters for correlation studies in biology and pharmacology', *International Journal of Peptide and Protein Research*, Vol. 32, No. 4, pp.269–278.

Freund, Y. and Schapire, R.E. (1995) 'A decision-theoretic generalization of on-line learning and an application to boosting', *Proceedings of the 2nd European Conference on Computational Learning Theory*, 13–15 March, Barcelona, Spain, pp.23–37.

Freund, Y. and Schapire, R.E. (1996) 'Experiments with a new boosting algorithm', *International Conference on Machine Learning*, 3–6 July, Bari, Italy, pp.148–156.

Geng, X., Guan, J., Dong, Q. and Zhou, S. (2012) 'An improved genetic algorithm for statistical potential function design and protein structure prediction', *International Journal of Data Mining and Bioinformatics*, Vol. 6,No. 2, pp.162–177.

Ghanty, P. and Pal, N.R. (2009) 'Prediction of protein folds: extraction of new features, dimensionality reduction and fusion of heterogeneous classifiers', *IEEE Transactions on NanoBioscience*, Vol. 8, No. 1, pp.100–110.

Grantham, R. (1974) 'Amino acid difference formula to help explain protein evolution', *Science*, Vol. 185, No. 4154, pp.862–864.

Gromiha, M.M. (2005) 'A statistical model for predicting protein folding rates from amino acid sequence with structural class information', *Journal of Chemical Information and Modeling*, Vol. 45, No. 2, pp.494–501.

Gromiha, M.M. (2009) 'Multiple contact network is a key determinant to protein folding rates', *Journal of Chemical Information and Modeling*, Vol. 49, No. 4, pp.1130–1135.

Guo, D., Mant, C.T., Taneja, A.K. and Hodges, R.S. (1986) 'Prediction of peptide retention times in reversed-phase high-performance liquid chromatography II. Correlation of observed and predicted peptide retention times factors and influencing the retention times of peptides', *Journal of Chromatography A*, Vol. 359, pp.519–532.

Guy, H.R. (1985) 'Amino acid side-chain partition energies and distribution of residues in soluble proteins', *Biophysical Journal*, Vol. 47, No. 1, pp.61–70.

Habibi, N., Saraee, M. and Korbekandi, H. (2013) 'Protein contact map prediction using committee machine approach', *International Journal of Data Mining and Bioinformatics*, Vol. 7, No. 4, pp.397–415.

Haddow, C., Perry, J., Durrant, M. and Faith, J. (2011) 'Predicting functional residues of protein sequence alignments as a feature selection task', *International Journal of Data Mining and Bioinformatics*, Vol. 5, No. 6, pp.691–705.

Hopp, T.P. and Woods, K.R. (1981) 'Prediction of protein antigenic determinants from amino acid sequences', *Proceedings of the National Academy of Sciences*, Vol. 78, No. 6, pp.3824–3828.

Hsieh, C.W., Hsu, H.H. and Pai, T.W. (2013) 'Protein crystallization prediction with adaboost', *International Journal of Data Mining and Bioinformatics*, Vol. 7, No. 2, pp.214–227.

Hutches, J.O. (2010) *Handbook of Biochemistry and Molecular Biology*, in Rogur, L., Lundblad and MacDonald, F.M. (Eds) 4th ed., CRC Press.

Jain, P., Garibaldi, J.M. and Hirst, J.D. (2009) 'Supervised machine learning algorithms for protein structure classification', *Computational Biology and Chemistry*, Vol. 33, No. 3, pp.216–223.

Janin, J. (1979) 'Surface and inside volumes in globular proteins', *Nature*, Vol. 277, Vol. 5696, pp.141–142.

Jones, D.T. (1999) 'Protein secondary structure prediction based on position-specific scoring matrices', *Journal of Molecular Biology*, Vol. 292, No. 2, pp.195–202.

Karplus, P.A. and Schulz, G.E. (1985) 'Prediction of chain flexibility in proteins', *Naturwissenschaften*, Vol. 72, No. 4, pp.212–213.

Kavousi, K., Moshiri, B., Sadeghi, M., Araabi, B.N. and Moosavi-Movahedi, A.A. (2011) 'A protein fold classifier formed by fusing different modes of pseudo amino acid composition via pssm', *Computational Biology and Chemistry*, Vol. 35, No. 1, pp.1–9.

Kawashima, S., Pokarowska, P.P.M., Kolinski, A., Katayama, T. and Kanehisa, M. (2008) 'Aaindex: amino acid index database, progress report', *Nucleic Acids*, Vol. 36, pp.D202–D205.

Krigbaum, W.R. and Komoriya, A. (1979) 'Local interactions as a structure determinant for protein molecules: Ii', *Biochimica et Biophysica Acta (BBA) - Protein Structure*, Vol. 576, No. 1, pp.204–228.

Kurgan, L.A., Cios, K.J. and Chen, K. (2008) 'Scpred: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences', *BMC Bioinformatics*, Vol. 9, p.226.

Kyte, J. and Doolitle, R.F. (1982) 'A simple method for displaying the hydropathic character of a protein', *Journal of Molecular Biology*, Vol. 157, No. 1, pp.105–132.

Liu, T., Geng, X., Zheng, X., Li, R. and Wang, J. (2012) 'Accurate prediction of protein structural class using auto covariance transformation of psi-blast profiles', *Amino Acids*, Vol. 42, No. 6, pp.2243–2249.

Manavalan, P. and Ponnuswamy, P.K. (1978) 'Hydrophobic character of amino acid residues in globular proteins', *Nature*, Vol. 275, No. 5681, pp.673–674.

Mathura, V.S. and Kolippakkam, D. (2005) 'Apdbase: amino acid physico-chemical properties database', *Bioinformation*, Vol. 12, No. 1, pp.2–4.

Meirovitch, H., Rackovsky, S. and Scheraga, H.A. (1980) 'Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids', *Macromolecules*, Vol. 13, No. 6, pp.1398–1405.

Miyazawa, S. and Jernigan, R.L. (1985) 'Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation', *Macromolecules*, Vol. 18, No. 3, pp.534–552.

Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) 'Scop: a structural classification of proteins database for the investigation of sequences and structures', *Journal of Molecular Biology*, Vol. 247, No. 4, pp.536–540.

Nanni, L., Brahnam, S. and Lumini, A. (2010) 'High performance set of pseaac and sequence based descriptors for protein classification', *Journal of Theoretical Biology*, Vol. 266, No. 1, pp.1–10.

Nelson, D.L. and Cox, M.M. (2008) *Lehninger Principles of Biochemistry*, 5th ed.

Nguyen, M.N. and Rajapakse, J.C. (2007) 'Prediction of protein secondary structure with two-stage multi-class svms', *International Journal of Data Mining and Bioinformatics*, Vol. 3, No. 1, pp.248–269.

Ponnuswamy, P.K., Prabhakaran, M. and Manavalan, P. (1980) 'Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins', *Biochimica et Biophysica Acta (BBA) - Protein Structure*, Vol. 623, No. 2, pp.301–316.

Rao, J.K.M. and Argos, P. (1986) 'A conformational preference parameter to predict helices in integral membrane proteins', *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, Vol. 869, No. 2, pp.197–214.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H. and Zehfus, M.H. (1985) 'Hydrophobicity of amino acid residues in globular proteins', *Science*, Vol. 229, No. 4716, pp.834–838.

Shamim, M.T.A., Anwaruddin, M. and Nagarajaram, H.A. (2007) 'Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs', *Bioinformatics*, Vol. 23, No. 24, pp.3320–3327.

Sharma, A., Lyons, J., Dehzangi, A. and Paliwal, K.K. (2013a) 'A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition', *Journal of Theoretical Biology*, Vol. 320, pp.41–46.

Sharma, A., Paliwal, K.K., Dehzangi, A., Lyons, J., Imoto, S. and Miyano, S. (2013b) 'A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition', *BMC Bioinformatics*, Vol. 14, p.233.

Shen, H.B. and Chou, K.C. (2009) 'Predicting protein fold pattern with functional domain and sequential evolution information', *Journal of Theoretical Biology*, Vol. 256,No. 3, pp.441–446.

Taguchi, Y.H. and Gromiha, M.M. (2007) 'Application of amino acid occurrence for discriminating different folding types of globular proteins', *BMC Bioinformatics*, Vol. 8, No. 1, p.404.

Vapnik, V.N. (1999) *The Nature of Statistical Learning Theory*, Springer-Verlag.

Vihinen, M., Torkkila, E. and Riikonen, P. (1994) 'Accuracy of protein flexibility predictions', *Proteins*, Vol. 19, No. 2, pp.141–149.

Witten, I. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann, San Francisco, CA, USA.

Wolfenden, R., Andersson, L., Cullis, P.M. and Southgate, C.C.B. (1981) 'Affinities of amino acid side chains for solvent water', *Biochemistry*, Vol. 20, No. 4, pp.849–855.

Yang, J.Y. and Chen, X. (2011) 'Improving taxonomy-based protein fold recognition by using global and local features', *Proteins: Structure, Function, and Bioinformatics*, Vol. 79, No. 7, pp.2053–2064.

Yang, T., Kecman, V., Cao, L., Zhang, C. and Huang, J.Z. (2011) 'Margin-based ensemble classifier for protein fold recognition', *Expert Systems with Applications*, Vol. 38, No. 10, pp.12348–12355.

Zangooei, M.H. and Jalili, S. (2013) 'Protein fold recognition with a two-layer method based on svm-sa, wp-nn and c4.5 (tlm-snc)', *International Journal of Data Mining and Bioinformatics*, Vol. 38, 8, No. 2, pp.203–223.

Zimmerman, J.M., Eliezer, N. and Simha, R. (1968) 'The characterization of amino acid sequences in proteins by statistical methods', *Journal of Theoretical Biology*, Vol. 21, No. 2, pp.170–201.