Investigation into the role of sequence-driven-features for prediction of protein structural classes

Sundeep Singh Nanuwa and Huseyin Seker

Abstract – **There have been a number of techniques developed for the prediction of protein structural classes, however, they show various degrees of accuracies over different assessment procedures and, in particular, the role of sequence-driven-features (SDF) not rigorously investigated. Therefore, the aim of this study is to carry out the largest comprehensive and consistent investigation on approximately 1500 protein sequence-driven-features that form 65 subsets in order to develop a robust predictive model and identify how well these feature(s) are at predicting protein structural classes. For evaluation of the features, two high quality 40% (or less) homology datasets that contain over 7000 protein sequences were extracted from proteomic databases. As a predictive technique, an optimum K-Nearest Neighbour Classifier, namely multiple-K-NN (MKNN) was developed, which not only records MKNN results, but also a predictive accuracy for each K nearest neighbourhood for K=1 to 11. In order to make the analyses consistent, three different cross-validation test procedures, 10-fold, leave-one-out and independent set, were used for all data sets and methods implemented. Over 5000 individual predictive results obtained, no firm consensus found on which features are highly associated with protein structural classes. However, interestingly, the best subsets of the features are found to be traditional AAC (48.62%) for 10-fold and (50.09%) for LOO, and dipeptide composition (85.91%) for independent set. The results appear to suggest that the AAC features are one of the best two subsets over 65 different subsets. Interestingly, in particular, with pseudo-amino-acid composition (PseAAC), unlike other research results presented in the literature, this investigation finds that there is no statistical improvement obtained from the sequence-order effect aspect (lamda) of PseAAC, which averaged 39.15%. The results also suggest that most of its predictive power comes from the AAC part that averaged at 46.84%, and the overall average predictive accuracy for PseAAC is 47.86%. This information appears to suggest that this feature set, which is claimed to better capture sequence order, yields almost no improvement and can be considered a redundant and noisy feature set. It should be noted that overall outcome of this comprehensive study sheds light not only in structural class prediction, but also other proteomic studies.**

I. Introduction

PROTEIN prediction is one of the most difficult and important fields within proteomics, mainly because the thousands of conformational changes in a protein makes it difficult to predict how it will fold into its secondary or tertiary structure. Computational biology has a huge impact in this field because of relatively inexpensive computational power that has enabled vast amounts of data to be analysed relatively quickly.

A protein is a biological molecule that carries out a specific function within the body, knowing and incorporating the structural class information can (1) improve prediction accuracy of secondary and tertiary structure prediction [1-4] and more significantly, (2) to bridge the gap between verified and unverified protein structures. The number of unverified protein structures is over 6 million [Release 39.0 of 22-July-2008 UniProtKB/TrEMBL] very different from how many have been verified, as of 12-August-2008 there are 52402 structures in Protein Data Bank (PDB). Levitt and Chothia [4] developed the standard for protein structural classes used in this study, which consists of four main types of protein structural classes: -

1. All-Alpha (α) - proteins with only small amount of strands
2. All-Beta (β) - proteins with only small amount of helices
3. Alpha / Beta (α / β) - proteins that include both helices and strands and where strands are mostly parallel
4. Alpha + Beta (α + β) - proteins with both helices and strands and where strands are mostly anti-parallel

There is substantial progress in protein structural class prediction [5-19], some of these studies use selected sequence features i.e. amino acid composition (AAC) only, which may not include crucial physiochemical properties and/or using poor quality datasets with low number of sequences at higher homology and all combined with inconsistent methodologies to arrive at often boosted results. The approach this project is taking, which is unique within the field, is to use approximately 1500 protein features extracted from the web server ProFEAT [20] as it is more of an interest to examine how additional and combination of features predict protein structural classes. Analysing these features using the predictive model multiple-k-nearest neighbourhood (MKNN) classifier and finally gathering results with three-test procedures. With the abovementioned approach, the projected outcomes are (1) identifying which of these sequence-driven-features is good at predicting protein structural classes and (2) a study that has used a consistent and comprehensive methodology throughout.

Manuscript received July 5, 2008.

Authors are with the Bio-Health Informatics Research Team at the Centre for Computational Intelligence, School of Computing, De Montfort University, Leicester, UK, LE1 9BH (e-mail: ssn@dmu.ac.uk, hseker@dmu.ac.uk)

II. Materials and Methods

Based on a comprehensive experimental investigation this paper aims to address factors related to datasets, sequence representation and statistical test procedures.

A. Datasets

Two datasets are used, first dataset 1189 obtained from [21] is chosen because of it's prior use in [21-23], the protein sequences are based on Structural Classification of Proteins (SCOP) version 1.67 February 2005 and homology is at 40%, it is considered to be a standard and benched marked dataset [11], using standard datasets ensures consistency in the wider field for future research and more importantly to keep in line with past studies. Second dataset is named astral40 and is constructed using the latest astral database tool release version 1.71 July 2007 [24], this was selected for it's easy access to a large set of proteins. The sample sizes of these datasets are not too low and it includes only high quality low-homologous proteins. A recent use of a similar astral dataset in a past study is version 1.63 [25], it is ideal as it contains a large amount of protein sequences at the chosen homology rate of 40%.

June 2006 saw a release of a new sequence-driven-feature named pseudo-amino-acid-composition (PseAAC) to ProFEAT web server [20], to analyse PseAAC, protein sequences length needed to be a minimum of 30 or more residues long. A bespoke programming function checked for sequences that had less than 30 residues and removed them from the datasets. At time of construction, no sequences with less than 30 residues were included in astral40 due to prior knowledge of the new sequence-driven-feature.

The final set of sequences for 1189 dataset is 223 all-α, 292 all-β, 331 α/β and 240 $\alpha+\beta$ and for astral40 dataset is 1446 all-α, 1728 all-β, 2065 α/β and 1850 $\alpha+\beta$.

B. Sequence representation

1) Prediction based on sequence-driven-features:

Sequence-driven-features are highly useful for distinguishing and representing proteins of different structural classes, function and interactions profiles, which is essential for the successful application of statistical learning methods in predicting different aspects of proteins. It is important to know which of these features predict proteins and more so, the features that are not so good, from this, we can see which features play an important role and are highly useful for distinguishing protein structural classes.

2) Extraction of sequence-driven-features

Protein Features (ProFEAT) [20] is a web server for computing commonly used structural and physicochemical features of proteins and peptides from amino acid sequences. It computes eight feature groups composed of 11 sub-features that include 53 descriptors and 1497 descriptor values. The computed feature groups include (1) amino acid composition, (2) dipeptide composition, (3) normalized-moreau–broto autocorrelation, (4) moran autocorrelation, (5) geary auto-correlation, (6) composition, transition and distribution (7) sequence-order and (8) PseAAC. Sequences from our datasets inputted in to the web server with the output result being a set of numerical vectors, which then applied to a statistical model for evaluation.

C. Predictive models

1) Multiple-K-Nearest Neighbourhood (MKNN)

MKNN calculates the predictive accuracy for each k-model and multiple models that yielded highest accuracy, multiple models is the result of combining strongest k-models; which is achieved by removing lower predicted k-models and re-analysing using highest resulted k-models, this investigation used eleven models. The method that determines which k-models selected is on a voting technique. M-KNN can analyse predictive accuracy results using each of the cross-validation test procedures i.e. 10-fold, LOO and independent set. KNN aspect of the algorithm tries to classify new patterns into their class membership by comparing features of unknown new patterns with features of known patterns, which already been classified. It is particularly useful in situations when distributions of the patterns and categories are unknown – such as protein structural classes [26].

D. Statistical test procedures

1) 10-fold

Each dataset divided into 10 folds i.e. 10% sequences per fold, MKNN repeated 11 times (as 11 models are used) and each time, one fold is testing data and the remaining nine folds are training data, this, being the main advantage for this test procedure – for speed. The disadvantage is the algorithm has to run k-times, hence, the larger datasets took considerable time to analyse, for example on the largest size dataset against all features [7052*1497] matrix, took eight hours to complete, however, it's relatively quicker running analysis on smaller datasets.

2) Leave-one-out (jack-knife)

Also known as jack-knife, during the process, both testing and training datasets are open and a protein will in turn move from training to testing, is analysed and then moved back to training. This is computationally demanding and resourceful technique, analysis time goes over eight hours on a smaller [5155*400] matrix (the same time on a [7052*1497] with 10-fold). Because of this process, the only advantage is that it applies thorough testing to each protein sample.

3) Independent set

Independent test procedure uses two separate datasets, testing and training, these contain only unique protein sequences; this means there are no two identical protein sequences between each dataset. This test procedure gives higher results when training samples are larger than testing dataset. The construction of independent datasets involved

applying the Euclidean distance function, which loops between two matrices (i.e. testing and training datasets) and calculates the distance between each data point, if it comes out to zero it means a duplicate protein data is found, the protein code is established and located within the ProFEAT data and removed.

III. Results And Discussion

This section will primarily focus on which features ranked in the top and bottom 10 per test procedure, as around 390 analyses took place, which resulted in thousands of individual results. Top 10 results for dataset 1189 and astral40 across each test procedure are in table I and II, respectively.

A. Assessment of 10-fold test procedure

Figure 1 shows a graphical view of how spread out the data values is for each dataset; column numbers 1-2 relate to each dataset 1189 and astral40 respectively.

Dataset 1189 results do not have extreme high or low values; the results are quite consistent and compact throughout all the features. Astral40 dataset results have a lower median value and generally seem more robust, this could be the result of using larger quantity of protein sequences.

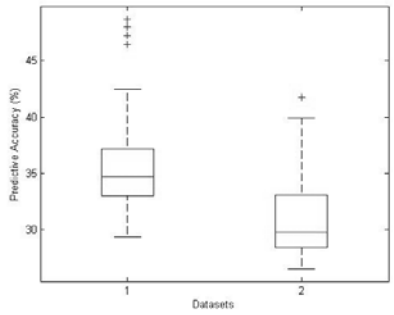

Figure. 1. Boxplot for 10-fold test procedure

1) Top 10 features

AAC is ranked 1st with the highest rate at 48.62% using 1189 dataset, which is not consistent using astral40 dataset as it's AAC ranked 3rd at 39.35%. AAC and dipeptide composition ranks 3rd and 5th with astral40, in simplistic form, both of those features are quite similar and more likely to appear closer using astral40 than 1189 dataset, because of the dataset size. Dipeptide composition does not appear using 1189 dataset, which is representative in larger datasets. Interestingly the results of analysing all features ranked 1st using astral40 at 41.76% and ranked 5th at 42.47% with 1189 dataset, very little difference in accuracy, but big difference in rank, this strongly shows that there are features that reduce predictive accuracy. PseAAC results, looking at both datasets, the AAC part of PseAAC ranks higher than PseAAC as a whole feature group, this illustrates that the lamda part has no improvement to PseAAC feature group. Secondary structure sub-feature from the composition feature group ranked 9th and 7th at 39.42% and 35.73% with 1189 and astral40 respectively.

2) Bottom 10-features

The bottom 10 features varied in order between the two datasets. The range of values between the bottom 10 are smaller with astral40 than 1189, 0.7% and 2.9% respectively, consequently, larger datasets produces results that are more robust. Most of the features ranked are sub-features from the autocorrelation feature groups. Common features between the datasets are, polarizability, relative mutability and, free energy in water, these are weaker features that (1) bring down the overall results when analysing using all features and (2) poor at individually predicting protein structural classes. Lamda aspect to PseAAC ranks 36th at 34.99% in comparative terms it is better than 35 other features, however, it still shows it does not have any statistical improvement.

B. Assessment of leave-one-out (loo) test procedure

Figure 2 shows 1189 is more compact and contains less extreme values indicating consistent range of results across all features. Astral40 has a longer range of lower values than 1189 indicating that again dataset size has a control on predictive accuracies. Overall, the results are slightly higher than 10-fold, however there are some variations in selected features.

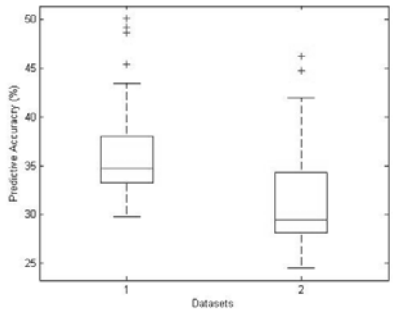

Figure. 2. Boxplot for test procedure

1) Top 10 features

Highest predicted feature using 1189 dataset is AAC at 50.09%, which further enforces AAC as a strong feature; however, it does not infer anything new. AAC part to PseAAC ranks 2nd and PseAAC 3rd using 1189 dataset, however, PseAAC ranks 1st at 46.20% and lamda aspect of PseAAC ranks 9th at 37.37% using astral40, even though it appears in the top 10 features, it shows that majority of the power for PseAAC comes from AAC part and not so much from lamda. AAC aspect of PseAAC and AAC are ranked 3rd and 4th at 41.92% and 41.84% respectively using astral40, shows both have similar strength. Secondary structure sub-feature from composition feature group ranks 9th using 1189 dataset.

2) Bottom 10-features

Bottom 10 features are largely from autocorrelation

feature groups. Sub-features common between the two datasets are, free energy in water, relative mutability & steric parameter, which are all from the three different autocorrelation feature groups, here is a clear indication that these features are not so good at predicting protein structural classes using this test procedure. Sub features of composition, transition and distribution appear within the bottom 10 features quite highly, however, when combined together as a feature group and analysed it appears in the top 10.

C. Assessment of independent test procedure

Figure 3 column numbers 1-2 relates to each dataset 1189vAstral40 and Astral40v1189 respectively. Dataset 1189vAstral40 is based on 1189 as the training set, results acquired are consistent per feature, i.e. there is not a vast difference between lowest and highest extreme values. However, dataset Astral40v1189, astral40 is the training set, achieves higher values, this is because the training dataset has trained the algorithm with more samples than to test with, thus, able to predict the testing dataset far better.

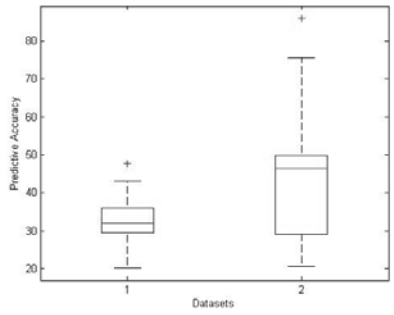

Figure. 3. Boxplot for independent test procedure

1) Top 10 features

PseAAC as a whole feature group ranks 1st and 5th at 47.58% and 63.17% than the AAC aspect of PseAAC, which ranks 3rd and 6th at 42.77% and 60.04%, using 1189vAstral40 and Astral40v1189 respectively. Conventional AAC ranks lower at 2nd and 7th at 43.09% and 59.58%, more importantly lamda aspect of PseAAC ranks 6th and 11th at 39.68% and 51.75%, 1189vAstral40 and Astral40v1189 respectively, proving further PseAAC power comes from the AAC part not lamda of PseAAC. Independent test procedures selected fewer sub-features, hydrophobicity scale, relative mutability and residue accessibility surface area in tripeptide. Overall, majority selected features form the main feature group's such as dipeptide composition and all three autocorrelation groups.

2) Bottom 10 features

Majority of features are from composition, transition and distribution feature groups and sub-features from sequence-order-coupling number feature group, which is quite interesting as it shows a consistent selection of weaker features across both datasets. The results for the well-trained dataset Astral40v1189 is relatively lower and similar to 1189vAstral40, nothing above 30%, so, not only consistent feature selection but very similar range of accuracies.

D. Assessment of the sequence-driven-features

Each statistical test procedure and dataset claim different sets of features are better or worse, therefore, relying on one single method may mislead in precisely identifying features, below is a summary of each feature groups: -

1) All features

Analysing all the available features with certain test procedures has a negative outcome because it includes the less predictive ones; in particular, it reduces the result when used with independent test procedure, however, 10-fold and LOO test procedures have better accuracies. It yielded better results than other single feature groups, however, it is more useful to look at which feature(s) or sub feature primarily make up the result.

2) Amino acid composition

AAC is consistent and important feature to appear in the top 10, the highest achieved is 59.48% with the independent test procedure using Astral40v1189 dataset, the lowest result obtained is 39.35% with 10-fold and Astral40 dataset, it's likely that dataset 1189 is picked carefully to represent protein structural classes, as its accuracies compared to astral40 are higher. However, astral40 is overall more representative being the latest release and largest dataset of protein samples.

3) Dipeptide composition

Dipeptide composition is very similar to AAC in terms of its physiochemical properties; however, with 10-fold it was one of the weaker features appearing bottom 10 using 1189 but appeared once in the top 10 using Astral40v1189 dataset.

4) Normalized moreau-borto autocorrelation

Using astral40 and 10-fold test procedure produced the lowest set of results for this feature group. Astral40 with LOO produced the lowest result for the sub-feature free energy in water, with independent test procedure using Astral40v1189 dataset worked well with this feature group.

5) Moran autocorrelation

Second autocorrelation feature best-suited independent test procedure, it was not a strong predictor for 10-fold using astral40 dataset, excluding sub-feature free energy in water which achieved the best result for this feature group.

6) Geary autocorrelation

The third autocorrelation feature group suited independent test procedure, not so well for 10-fold or LOO, most of its sub features appeared in the bottom 10 consistently.

7) Composition, Transition & Distribution

This is a strong feature group for all datasets using 10-fold and LOO assessment; in particular, the sub-feature secondary structure, which appeared in the top 10 consistently across 10-fold test procedure, however, this feature, achieves low results with independent test

procedure.

8) Sequence Order

Sequence-order feature group is the strongest across 10-fold and LOO test procedures, almost all its sub features appear using 1189 dataset. Results between sequence-order whole feature group, its sub features and sequence-order-coupling number are the same across all three test-procedures and datasets. Sub feature sequence-order-coupling number alone is better than using the whole feature group of sequence-order. Looking more closely, sub feature quasi-sequence-order-descriptors achieves majority of the highest results. Majority of representation appears in sub features.

9) Pseudo amino acid composition

PseAAC is a popular feature group used to predict protein structural classes and other aspects in other studies; it uses a weighted AAC combined with extra sequence-order information, which AAC alone does not contain, thus, supposedly to be better than conventional AAC. This study splits the feature group into two sub-features (1) AAC part of PseAAC and (2) lamda (sequence-order-information part), the reason was to analyse the separate parts to find out (1) where the predictive power comes from and (2) how well the lamda part predicts. The results are attention grabbing, AAC part of PseAAC did achieve higher result consistently using 10-fold and LOO than using the whole PseAAC feature group, as for the independent test procedure, the results were lower. The lamda aspect consistently came out lower than both PseAAC and conventional AAC, indicating lamda part of PseAAC has minimal (often statistically insignificant) influence on the accuracy and the feature group power comes from weighted AAC, this being a significant finding in the investigation.

E. Additional analyses on combination of features

Analysing combination of feature groups allows further investigation into which features groups when combined are better or worse at predicting protein structural classes than using individual feature groups. These are the five combination analyses: -

1. AAC + PseAAC

Across 10-fold/LOO, 1189 achieves slightly higher results at 48.81% and 50.09%, respectively; however, this is to do with analysing two sets of AAC compositions (1) the conventional AAC and (2) the AAC part of PseAAC, thus, the results higher because of that.

- AAC + Dipeptide composition

Little improvement made with independent test procedure using 1189vAstral40 by 2% over dipeptide feature group, no big improvements with 10-fold or LOO. AAC + Dipeptide composition was not as promising as expected, published results using various methods averaged 79% a lot higher than this investigation result which is between 35% - 45%, however, the published results used a low homology dataset with just 204 sequences.

- Dipeptide composition + PseAAC

Using independent dataset 1189vAstral40 does not achieve any higher results than individual feature groups, across 10-fold/LOO test procedure results were not improved expect for astral40/10-fold at 41.83%, increased by 7.51%. Compared to the overall results using dipeptide composition alone, the accuracies ranged between 27-37%, it confirms that the dipeptide composition is not a very strong feature to use.

- AAC + Dipeptide composition + PseAAC

Better improvement over dipeptide composition by 7% with independent test procedure using 1189vAstral40 smaller increase over AAC and PseAAC, indicating smaller testing datasets work better with this combination at 47.82%. 10-fold and LOO using 1189 dataset achieved lower results.

- AAC + Sequence-order

10-fold/LOO test procedures produced higher results, using astral40, same for the independent test procedure when 1189 is the training dataset; results are lower with astral40 dataset as testing.

F. Cross-validation test procedures

10-fold cross validation is less computationally demanding test procedure, e.g. between 10-25mins processing time per sub feature using 1189 dataset, whereas LOO would take between one to four hours to complete a single sub feature from the same dataset, there was no substantial improvements to the results using LOO over 10-fold. With independent test procedure, when the training dataset is astral40 the results achieved are the highest across majority of the features, due to the extensive training with 7000 protein sequences, when the smaller 1189 dataset is used majority of results obtained are mid-ranged across all the results.

IV. CONCLUSIONS

Computational prediction of protein structural classes is vastly complex, thus, carrying uncertain results. The success of this investigation largely returned positive outcomes, (1) AAC is still the best feature to represent protein structural classes, (2) many individual sub-features have little minimal increase on predictive accuracies and (3) combination of datasets and test procedures influence the rank of features. It is a challenging problem to draw a single and concise conclusion when results are varied. Overall, there is no firm consensus which other feature-sets or sub-features are distinct at predicting protein structural classes. Highlighting the important finding from this study is the lamda part of PseAAC does not add any statistical value to PseAAC results; the power of this feature group comes from the weighted AAC part of PseAAC, as discussed early and hence should not be used blindly and requires further investigation, which is under way.

TABLE I
TOP 10 FEATURES ACROSS EACH TEST PROCEDURE FOR 1189

Feature	10-fold	Feature	LOO	Feature	Independent
1	48.62%	1	50.09%	8	47.58%
8.1	47.97%	8.1	49.17%	1	43.09%
8	47.23%	8	48.62%	8.1	42.77%
6a	46.41%	6a	45.42%	6a	40.97%
All	42.47%	All	43.41%	2	40.82%
6	40.54%	7b	41.30%	8.2	39.68%
7b	40.40%	7b.2	40.48%	3	39.50%
7b.2	39.80%	6	40.11%	3.1	38.38%
6a.6	39.42%	6a.6	39.84%	4	38.16%
6b	39.32%	7	39.10%	3.8	38.08%

Key (# of features):- All=All features(1497), 1=AAC(20), 2=Dipeptide composition(400), 3=Normalized Moreau-Borto Autocorrelation(240), 3.1=Hydrophobicity scale(30), 3.8=Relative mutability(30), 4=Moran Autocorrelation(240), 6=Composition, Transition & Distribution(147), 6a=Composition(21), 6a.6=Secondary structure(3), 6b=Transition(21), 7=Sequence Order(160), 7b=Quasi-sequence-order descriptors(100), 7b.2=Based on normalized Granthan chemical distance(50), 8=Pseudo amino acid composition(50), 8.1=AAC part of PseAAC(20), 8.2=Lamda part of AAC(30)

TABLE II
TOP 10 FEATURES ACROSS EACH TEST PROCEDURE FOR ASTRAL40

Feature	10-fold	Feature	LOO	Feature	Independent
All	41.76%	8	46.20%	2	85.91%
6a	39.92%	All	44.72%	4	75.51%
1	39.35%	8.1	41.92%	5	73.02%
8.1	39.15%	1	41.84%	3	68.23%
2	36.98%	6a	41.09%	8	63.17%
8.2	35.95%	2	38.67%	8.1	60.04%
6a.6	35.73%	4.1	37.61%	1	59.48%
6	34.93%	3	37.40%	5.1	52.67%
6a.7	34.72%	8.2	37.37%	3.1	52.12%
8	34.37%	6	37.06%	4.5	51.93%

Key (# of features):- All=All features(1497), 1=AAC(20), 2=Dipeptide composition(400), 3=Normalized Moreau-Borto Autocorrelation(240), 3.1=Hydrophobicity scale(30), 3.8=Relative mutability(30), 4=Moran Autocorrelation(240), 4.1=Hydrophobicity scale(30), 4.5=Residue accessibility surface area in Tripeptide(30), 5=Geary autocorrelation(240), 5.1=Hydrophobicity scale(30), 6=Composition, Transition & Distribution(147), 6a=Composition(21), 6a.6=Secondary structure(3), 6a.7=Solvent accessibility(3), 6b=Transition(21), 7=Sequence Order(160), 7b=Quasi-sequence-order descriptors(100), 7b.2=Based on normalized Granthan chemical distance(50), 8=Pseudo amino acid composition(50), 8.1=AAC part of PseAAC(20), 8.2=Lamda part of AAC(30)

REFERENCES

[1] Gromiha, M.M. and S. Selvaraj, *Protein secondary structure prediction in different structural classes.* Protein Engineering, 1998. **11**(4): p. 249-251.

[2] Chou, K.C. and C.T. Zhang, *Prediction of protein structural classes.* Critical Reviews in Biochemistry and Molecular Biology, 1995. **30**(4): p. 275-349.

[3] Bahar, I., et al., *Understanding the recognition of protein structural classes by amino acid composition.* Proteins: Structure, Function and Genetics, 1997. **29**(2): p. 172-185.

[4] Levitt, M. and C. Chothia, *Structural patterns in globular proteins.* Nature, 1976. **261**(5561): p. 552-558.

[5] Luo, R.Y., Z.P. Feng, and J.K. Liu, *Prediction of protein structural class by amino acid and polypeptide composition.* European Journal of Biochemistry, 2002. **269**(17): p. 4219-4225.

[6] Cai, Y.D., et al., *Support vector machines for predicting protein structural class.* BMC bioinformatics [electronic resource], 2001. **2**(1): p. 3.

[7] Jin, L., W. Fang, and H. Tang, *Prediction of protein structural classes by a new measure of information discrepancy.* Computational Biology and Chemistry, 2003. **27**(3): p. 373-380.

[8] Cai, Y.-D., et al., *Using LogitBoost classifier to predict protein structural classes.* Journal of Theoretical Biology, 2006. **238**(1): p. 172-176.

[9] Chen, C., et al., *Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network.* Analytical Biochemistry, 2006. **357**(1): p. 116-121.

[10] Du, Q.S., et al., *Amino Acid Principal Component Analysis (AAPCA) and its applications in protein structural class prediction.* Journal of Biomolecular Structure & Dynamics, 2006. **23**(6): p. 635-640.

[11] Kedarisetti, K.D., L. Kurgan, and S. Dick, *Classifier ensembles for protein structural class prediction with varying homology.* Biochemical and Biophysical Research Communications, 2006. **348**(3): p. 981-988.

[12] Xiao, X., et al., *Using pseudo amino acid composition to predict protein structural classes: Approached with complexity measure factor.* Journal of Computational Chemistry, 2006. **27**(4): p. 478-482.

[13] Ding, Y.S., T.L. Zhang, and K.C. Chou, *Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network.* Protein and Peptide Letters, 2007. **14**(8): p. 811-815.

[14] Jahandideh, S., et al., *Novel two-stage hybrid neural discriminant model for predicting proteins structural classes.* Biophysical Chemistry, 2007. **128**(1): p. 87-93.

[15] Kurgan, L. and K. Chen, *Prediction of protein structural class for the twilight zone sequences.* Biochemical and Biophysical Research Communications, 2007. **357**(2): p. 453-460.

[16] Lin, H. and Q.Z. Li, *Using pseudo amino acid composition to predict protein structural class: Approached by incorporating 400 dipeptide components.* Journal of Computational Chemistry, 2007. **28**(9): p. 1463-1466.

[17] [1] Yu, T., et al., Structural class tendency of polypeptide: A new conception in predicting protein structural class. Physica A: Statistical Mechanics and its Applications, 2007. 386(1): p. 581-589.

[18] Ke Chen, L.A.K.J.R., *Prediction of protein structural class using novel evolutionary collocation-based sequence representation.* Journal of Computational Chemistry, 2008. **9999**(9999): p. NA.

[19] Zhang, T.-L., Y.-S. Ding, and K.-C. Chou, *Prediction protein structural classes with pseudo-amino acid composition: Approximate entropy and hydrophobicity pattern.* Journal of Theoretical Biology, 2008. **250**(1): p. 186-193.

[20] Li, Z.R., et al., *PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence.* Nucleic Acids Research, 2006. **34**(WEB. SERV. ISS.).

[21] Kurgan, L.A. and L. Homaeian, *Prediction of structural classes for protein sequences and domains-Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy.* Pattern Recognition, 2006. **39**(12): p. 2323-2343.

[22] Wang, Z.X. and Z. Yuan, *How good is prediction of protein structural class by the component- coupled method?* Proteins: Structure, Function and Genetics, 2000. **38**(2): p. 165-175.

[23] Chou, K.C. and G.M. Maggiora, *Domain structural class prediction.* Protein Engineering, 1998. **11**(7): p. 523-538.

[24] Chandonia, J.M., et al., *The ASTRAL Compendium in 2004.* Nucleic Acids Research, 2004. **32**(DATABASE ISS.).

[25] Chou, K.C., *Progress in protein structural class prediction and its impact to bioinformatics and proteomics.* Current Protein and Peptide Science, 2005. **6**(5): p. 423-436.

[26] Zhang, Z.H., Z.H. Wang, and Y.X. Wang. Nearest neighbor algorithm for prediction of protein domain structural class. in Proceedings - Eighth International Conference on High-Performance Computing in Asia-Pacific Region, HPC Asia 2005. 2005. Beijing.