Novel weighted amino acid composition for prediction of protein structural classes within the context of multi-sensor data fusion approach

Huseyin Seker

Abstract— **Prediction of structural classes of proteins is one of the most important but challenging research problems in computational biology and mainly based on amino acid sequence of the proteins. However, most of the predictive features based on the sequences don't consider natural amino acid scales, which have been shown to play an important role in characterising the proteins in other studies. Therefore, this paper aims to present development of a novel weighted amino acid composition features based on the amino acid scales and multi-sensor data fusion strategies for reliable and accurate prediction of the structural classes of the proteins. The approach is further developed applying principal component analysis in each weighted amino acid composition features, which then leades to a locally optimized multi-sensor data fusion model. This pilot study presents promising results that show that the methods improve predictive accuracy by 1 to 10% compared to previously studied methods for the same data set. The approach taken is also shown to be not only effective, but also, in particular, more informative as it fuses information obtained from natural amino acid index scales that help better understand nature of such proteins.**

Keywords— Regularized Discriminant Classifier, Principal Component Analysis, Majority Voting, Weighted Amino Acid Composition, Amino Acid Scales.

I. Introduction

One of the difficult problems in biology is to try to find a solution to computational prediction of still unknown information about a protein or a family of proteins. The protein structure classification and its computational prediction by means of a set of descriptive attributes are typical but challenging research problem in this context [1].

The function of a protein is highly correlated to its three-dimensional structures. Information about such 3-D structure therefore plays a central role in not only understanding a protein's function but also classifying protein's functional families. There are mainly four different structural classes of proteins defined: (1) all-alpha, (2) all-beta, (3) alpha/beta, and (4) alpha+beta. The first group of the proteins represents protein structures that consist of mainly alpha-helices whereas the second group, namely all-beta group, characterizes a group of proteins with mainly beta-strands. The last two groups are composed of both alpha-helices and beta-strands [2].

Prediction of structural classes is generally based on amino acid sequence of a protein that belongs to one of these four classes as there are millions of proteins that have been identified with unknown structures but known amino acid (AA) sequences [1, 3]. Therefore, a set of some useful descriptive features that can be extracted from AA sequences of proteins are quite useful, can be easily used to characterize such proteins and consequently to predict structural classes of a protein using various computational predictive techniques.

There are two main steps in computational prediction of structural classes of a protein: (1) extraction of characteristic features and (2) classification. Regarding the classification step, various traditional and state-of-the-art techniques have successfully been implemented [4]. However, as in pattern recognition problem, the most difficult problem turned out to be extraction of the most informative and reliable set of sequence-driven features in line with some other biological features.

There have been a number of methods developed to extract a set of informative features from AA sequence of proteins [5]. Among these feature sets, a traditional but one of the widely used approaches is amino acid composition feature set (AAC) that is shown to be very effective and yield higher predictive accuracy in many cases compared to various sequence-driven feature sets [6]. The AAC takes a normalized number of each of the natural twenty amino acids into consideration. However, such approach does not consider natural weight of each amino acid but assume that each amino acid has an equal weight of one. Therefore, it should be suggested that such weights are quite useful in terms of proteins' biological behavior and can help better characterize and understand nature of the proteins as well as predict functions and structural classes of the proteins more accurately. Such weights are called AA index values or scales, some of which were experimentally measured in biological labs [7] whereas some were derived from the experimentally measured scales [8].

It can be suggested that these index values or weights can be replaced with the traditional weight value of one and generate a novel set of sequence-driven features, which can be called "weighted amino acid composition features" (WAAC) developed in this paper.

Manuscript received July 8, 2008.

H. Seker is with Bio-Health Informatics Research Group, Centre for Computational Intelligence, School of Computing, De Montfort University, Leicester, LE1 9BH, UK (e-mail: hseker@dmu.ac.uk).

Each index value set of AA and consequently WAAC feature set can be treated as an independent information source for proteins and their structure and functions, and then be fused to create a multi-sensor model for protein characterization and representation. Such a multi-sensor system is then expected to yield an improved prediction and help better understand the nature of such proteins as the multi-sensor data fusion approaches have successfully been developed and used in many different areas such as automated target recognition and medical diagnosis [9, 10].

This paper in the following sections presents novel sequence-driven features based on a set of natural AA index values, namely, weighted amino acid composition features (WAAC). It also demonstrates a multi-sensor data fusion strategy to fuse information generated by each of these WAAC features that can be obtained from independent information sources which are formed by the natural AA index values. In addition, principal component analysis, which is applied to each WAAC separately, is also briefly described as it is used to construct a locally optimized multi-sensor data fusion model. The final section discusses results obtained through empirical studies on the benchmark data set and provides some recommendations in the light of the findings.

II. Methods

In this section, development of novel amino acid composition features is described. In addition, regularized discriminant classifier, which is the predictive method used to construct a multi-sensor data fusion approach is presented. Principal component analysis, which is applied to each WAAC separately, is also briefly described as it is used to construct a locally optimized multi-sensor data fusion model. Brief information about the fused models is also provided.

A. Weighted amino acid composition features

A typical protein consists of an amino acid sequence of a number of natural 20 amino acids in many different orders (Fig.1). Based on this sequence, there have been methods developed to drive descriptive features that can characterise the protein. Among these feature sets, a traditional but one of the widely used approaches is amino acid composition feature set (AAC) that is shown to be very effective and yield comparable predictive accuracy compared to many different sequence-driven features [6]. The AAC takes a normalized number of each of the natural twenty amino acids into consideration. Therefore, number of AAC features is just twenty.

The traditional AAC feature set does not consider natural weight of each amino acid but assume that each amino acid has an equal weight of one. Therefore, it should be suggested that such weights are quite useful in terms of proteins' biological behavior and may help better understand nature of the proteins as well as predict functions and structural classes of the proteins. Such weights are called AA index values or scales, some of which were experimentally measured in biological labs whereas some were obtained using computational statistical methods. It should be noted that the latter sets are mainly based on the experimentally measured AA index values.

DACEQAAIQCVESACESLCTEGEDRTGCYMYIYSN CPPYV

Fig. 1. A typical protein sequence of 40 amino acids. This protein belongs to all-alpha group and its protein data bank ID is 1ERC

It can be suggested that such index values or weights can be replaced with the traditional weight value of one and generate a novel set of sequence-driven features, which can be called "weighted amino acid composition features" (WAAC) and can be defined for amino acid A as

$$WAAC(A)= n(A)\boldsymbol{x}w(A)/[n(A)\boldsymbol{x}w(A)+n(C)\boldsymbol{x}w(C)+\ldots+n(Y)\boldsymbol{x}w(Y)]$$

where WAAC(A) is the weighted composition of amino acid A, n(A), n(C), …and n(Y) are number of these natural twenty amino acids in a sequence, and w(A), w(C), …and w(Y) are the index values of these natural twenty amino acids. WAAC can be computed for other amino acids (C to Y) in a similar way.

It can be seen that this expression is general as this forms the traditional amino acid composition feature set (AAC) when w is set to 1.

There are currently over 500 sets of amino acid index values listed, some of which have been used in different applications for protein characterization [7]. This study takes two sets of the index values into consideration. Details of them can be seen in Tables I and II. The first set of the index values presented in Table I that consists of six different sets of amino acid index values has recently been highlighted in the extraction of pseudo-amino acid composition features [11], and can be used for extraction of WAAC for any type of proteins. The second set listed in Table II is actually application-oriented scales and consists of five different sets of such index values specifically obtained to characterise proteins that belong to the four structural classes. They are normalized hydrophobicity scales for alpha, beta, alpha+beta and alpha/beta proteins. The last one is normalized average hydrophobicity scales [12].

B. Regularized discriminant classifier

Although there have been many different methods developed for not only prediction of classes of protein structures but also construction of multi-sensor data fusion models including support vector machines-based classifiers, this paper considers a simple but effective method in order to show novelty, applicability and robustness of the approach taken. It should be noted that one of the factors that heavily affects power of a classifier is a set of robust predictive features and therefore even a simple classifier should be expected to perform better under such feature set. It is why such a simple and well-known, but effective classifier was adapted in this preliminary study.

General purpose of a discriminant analysis is to assign an object or a set of objects to one of several classes using a set of descriptive features (e.g., WAAC). Linear and Quadratic discriminant analysis are the two methods widely adapted for such purpose. Classification rules based on these two methods rely on reliable estimates of eigenvalues by correcting the eigenvalue distortion in a sample covariance matrix. Therefore, a regularization approach was proposed to combine Linear and Quadratic discriminant classifiers through regularization of a covariance matrix as

$$\Sigma(a,b) = (1-b)\Sigma(a) + \frac{b}{p} tr[\Sigma(a)]I$$

where Σ represents covariance matrix, a and b are regularization parameters, p is the number of features (dimension), and I is the identify matrix. Values of the two regularization parameters that take a value between 0 and 1 can be chosen to minimize a misclassification rate. It can be seen that the expression above is general as it forms a linear discriminant classifier when a is set to 0 whereas b=0 will produce a quadratic discriminant classifier [13].

The approach briefly described above is called regularized discriminate classifier (RDC). Further details can be obtained from [13].

C. Multi-sensor data fusion model

The multi-sensor data fusion model developed for this study is depicted in Fig. 2.

The model uses WAAC obtained for each index value set as an independent information source and inputs to each RDC. There are different successful formulations developed to fuse outputs of classifiers, in this case, the RDCs. However, it is shown that majority voting is found to be one of the best and, in particular, stable methods [14]. Due to its simplicity and successful applications, the multi-sensor data fusion model in this study fuses the outputs of the RDCs using the majority voting.

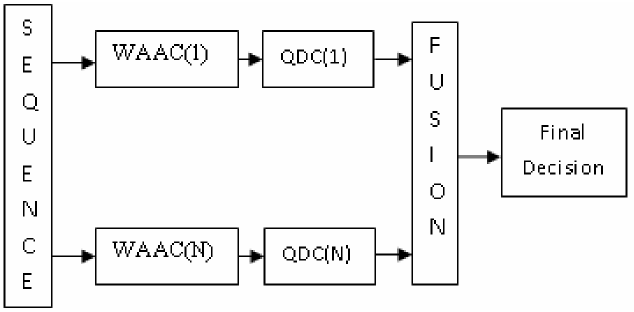

Fig. 2. Multi-sensor data fusion model with WAAC and RDC for N number of Information Sources (AA Index values)

D. Principal component analysis and locally optimized multi-sensor data fusion model

Principal components analysis (PCA) is a quantitatively rigorous method for reducing dimension of the features. The method generates a new set of features, called principal components. Each principal component is a linear combination of the original features. All the principal components are orthogonal to each other, so it is assumed that there is no redundant information. Finally, these principal components then form a new but reduced feature set.

Number of the principal components that needs to be selected is based on variance explained by the corresponding principal component. Total variance is generally taken into consideration for selection of optimum number of these components, which then form new features. For example, 99% variability is used for this study. Further details about PCA can be found in [15].

As dimension of each WAAC feature set is separately reduced using the PCA, the output of each PCA model forms a new but reduced feature set, which will then fed to RDCs. This model is called “locally optimized multi-sensor data fusion model” as redundant information is removed from each information source by means of the PCA. For this PCA-based model, a block that performs PCA operation will be included to the model in Fig. 2. between WAAC and RDC. As in the previous model, majority voting is also selected to combine output of each PCA-based analysis.

III. THE DATA SET AND INFORMATION FOR THE ANALYSIS

A. The Data Set

The data set, namely the 1189 data set, used for this study was obtained from [4] where only protein IDs were listed. A MATLAB program was then developed to extract sequences for each protein, and their corresponding weighted amino acid composition (WAAC) features were then calculated. The final data set consists of 223 all-alpha, 294 all-beta, 241 alpha+beta, and finally 334 alpha/beta proteins. Sequence length (total number of amino acids in a sequence) in the data set was found to be between 30 and 842.

As mentioned before, the data set studied in this paper was previously used as a benchmark data set. For example, Kurgan and Homaeian presented the most comprehensive analysis of the data set using eight different classifiers and various sequence-driven features in [4]. They reported that the logistic regression, one of the widely used statistical classifiers, yielded the highest predictive accuracy of 53.9% . Interestingly, it should also be noted that the second best classifier was found to be the support vector machine classifier that resulted in a predictive accuracy of 52.3%. In addition to the eight classifiers implemented in [4], Wang and Yuan adapted Bayes classification approach that yielded a predictive accuracy of 53.8% [2], which is higher than that of the support vector machine but similar to that of the logistic regression.

B. *General information for the analysis*

The index values listed in Tables I and II were separately analysed in addition to the traditional AAC features. All possible combinations of these sets were also investigated in order to find the best possible fusion model(s) for each table. Therefore, for the models in Table I, 127 different combinations were studied whereas 63 different combinations were investigated for the models in Table II.

Regarding optimisation of the RDC, the regularisation parameters (a and b) were scaled between 0 and 1 by 0.1 interval, which resulted in a study of 121 different pairs of the values. This was carried out in order to find the most optimum values of the parameters for the given problems and consequently reach the least misclassification rate.

In order to make the comparison with other studies for the same data set consistent, jack-knife cross-validation approach was also adapted.

TABLE I
THE HYDROPHOBICITY, HYDROPHILICITY, MASS, pK1(ALPHA-COOH), pK2(NH3) AND pI (at 25^{o}C) VALUES FOR W**AAC**

Amino acid	Hydrophobicity	Hydrophilicity	Mass	pK1(a-CO2H)	pK2(NH3)	pI(at 25^{o}C)
A	0.62	-0.5	15	2.35	9.87	6.11
C	0.29	-1	47	1.71	10.78	5.02
D	-0.9	3	59	1.88	9.6	2.98
E	-0.74	3	73	2.19	9.67	3.08
F	1.19	-2.5	91	2.58	9.24	5.91
G	0.48	0	1	2.34	9.6	6.06
H	-0.4	-0.5	82	1.78	8.97	7.64
I	1.38	-1.8	57	2.32	9.76	6.04
K	-1.5	3	73	2.2	8.9	9.47
L	1.06	-1.8	57	2.36	9.6	6.04
M	0.64	-1.3	75	2.28	9.21	5.74
N	-0.78	0.2	58	2.18	9.09	10.76
P	0.12	0	42	1.99	10.6	6.3
Q	-0.85	0.2	72	2.17	9.13	5.65
R	-2.53	3	101	2.18	9.09	10.76
S	-0.18	0.3	31	2.21	9.15	5.68
T	-0.05	-0.4	45	2.15	9.12	5.6
V	1.08	-1.5	43	2.29	9.74	6.02
W	0.81	-3.4	130	2.38	9.39	5.88
Y	0.26	-2.3	107	2.2	9.11	5.63

IV. Results and Discussions

There were 15367 (121x127) and 7623 (121x63) different analyses carried out for the models in Tables I and II, respectively, for the multi-sensor data fusion method with and without PCA. It should be noted that traditional AAC feature set was also included to the models.

The highest predictive accuracy results are presented in Table III that also gives optimum values of the regularization parameters as well as the models that yielded the highest accuracy.

Among these results, the highest predictive accuracy achieved was 54.6% that the 7-set model in Table I yielded. This result is higher than those presented in previous studies [2, 4]. The results in Table III include traditional AAC features and WAAC features for Hydrophobicity, pK1(a-CO2H) and pK2(NH3). The same accuracy for the 7-set model was also obtained from the PCA-based model, but with WAAC features obtained from only Hydrophilicity and pK1(a-CO2H). This result appears to indicate that PCA was effectively used and helped reduce the dimension significantly by not only selecting a fewer number of the index sets from Table I but also reducing the number of Hydrophilicity and pK1(a-CO2H)-based WAAC features. This also means that less number of classifiers was used. Close observation showed that number of the principal components varied between 1 and 10 throughout the analysis.

The model developed using the six sets listed in Table II showed interesting results. Although the highest accuracy of the fusion model was 53.4% slightly less than that of the 7-set model, the locally optimized fusion model that is obtained using the PCA yielded a better result achieving 53.7% accuracy. This result further supports effective use of PCA for the construction of a locally optimized multi-sensor fusion model. Both results were obtained by using WAAC features from only CIDH920102 that characterizes the beta proteins.

As far as the regularization parameters a and b used to optimize the discriminate classifier are concerned, the results in Table III appear to suggest that the optimum values of the parameters "a" should be kept at 0 or 0.1, and "b" at 0.5 or 0.4 for these two different methods in order to achieve optimal classifier and fusion models described in this study.

TABLE II

AMINO ACID INDEX VALUES USED FOR WAAC: NORMALIZED HYDROPHOBICITY SCALES FOR ALPHA (CIDH920101), BETA (CIDH920102), ALPHA+BETA (CIDH920103) AND ALPHA/BETA (CIDH920104) PROTEINS. THE LAST ONE (CIDH920105) IS NORMALIZED AVERAGE HYDROPHOBICITY SCALES.

Amino acid	CIDH920101	CIDH920102	CIDH920103	CIDH920104	CIDH920105
A	-0.45	-0.08	0.36	0.17	0.02
C	-0.24	-0.09	-0.52	-0.7	-0.42
D	-0.2	-0.7	-0.9	-0.9	-0.77
E	-1.52	-0.71	-1.09	-1.05	-1.04
F	0.79	0.76	0.7	1.24	0.77
G	-0.99	-0.4	-1.05	-1.2	-1.1
H	-0.8	-1.31	-0.83	-1.19	-1.14
I	-1	-0.84	-0.82	-0.57	-0.8
K	1.07	0.43	0.16	-0.25	0.26
L	0.76	1.39	2.17	2.06	1.81
M	1.29	1.24	1.18	0.96	1.14
N	-0.36	-0.09	-0.56	-0.62	-0.41
P	1.37	1.27	1.21	0.6	1
Q	1.48	1.53	1.01	1.29	1.35
R	-0.12	-0.01	-0.06	-0.21	-0.09
S	-0.98	-0.93	-0.6	-0.83	-0.97
T	-0.7	-0.59	-1.2	-0.62	-0.77
V	1.38	2.25	1.31	1.51	1.71
W	1.49	1.53	1.05	0.66	1.11
Y	1.26	1.09	1.21	1.21	1.13

Compared to the results obtained from previous studies (53.9%), our results are shown to be higher. Interestingly, such a simple fusion model constructed using the regularized discriminant analysis yielded better than those of the support vector machine that has been widely adapted in computational biology. It should also be worth noting that the models presented in this paper use quite less number of the descriptive features.

V.CONCLUSION

The paper presents a novel multi-sensor data fusion approach for prediction of the structural classes of proteins. This is based on the novel weighted amino acid index scales. This is also further improved by incorporating with PCA. Effectiveness of the methods is demonstrated by applying to the benchmark 1189 data set.

It should be noted that the methods presented in the paper is general and applicable to other proteo-informatics applications.

As there are many more index scales defined, it is necessary to investigate which sets of these values play a better role in characterising these proteins within the context of the weighted AAC and consequently predicting classes of such proteins. This should be further investigated using various data sets in order to show robustness of the methods. This may also lead to identify appropriate and different set(s) of amino acid scales for protein structures and functions. Further research is being carried out in this direction.

TABLE III
PREDICTIVE ACCURACY RESULTS

Multi-sensor data fusion model			
Models	Optimal values of the regularisation parameters (a , b)	Fusion Model	Accuracy
The 7-set model (Table I)	(0.1 , 0.5)	Traditional AAC and WAAC feature sets for Hydrophobicity, pK1(a-CO2H) and pK2(NH3)	54.6%
The 6-set model (Table II)	(0.0 , 0.5)	WAAC features for CIDH920102	53.4%
Locally optimised multi-sensor data fusion model with PCA			
The 7-set model (Table I)	(0.1 , 0.4)	WAAC features for Hydrophilicity and pK1(a-CO2H)	54.6%
The 6-set model (Table II)	(0.1 , 0.4)	WAAC features for CIDH920102	53.7%

REFERENCES

[1] H. Lin and Q.Z. Li, “Using Pseudo Amino Acid Composition to Predict Protein Structural Class: Approached by Incorporating 400 Dipeptide Components”, Journal of Computational Chemistry, Vol. 28, pp: 1463-1466, 2007.

[2] Z.-X.Wang and Z.Yuan, “How good is the prediction of protein structural class by the component-coupled method?”, Proteins, Vol 38, pp: 165–175, 2000.

[3] L. Kurgan and K. Chen, “Prediction of protein structural class for the twilight zone sequences”, Biochemical and Biophysical Research Communications, Vol. 357, pp: 453-460, 2007.

[4] L. Kurgan and L. Homaeian, “Prediction of structural classes for protein sequences and domains–impact for prediction algorithms, sequence representation and homology, and test procedures on accuracies”, Pattern Recognition, Vol.39, No.12, pp:2323-2343, 2006.

[5] Z.R.Li *et al.*, “PROFEAT: A Web server for computing structural and physicochemical features of proteins and peptides from Amino Acid Sequence”, Nucleic Acids Research Journal, Vo. 34, pp: W32-7, July 2006.

[6] S.A.K. Ong *et. al.*, “Efficacy of different protein descriptors in predicting protein functional families”, BMC Bioinformatics, Vol 8, pp: 300-314, August 2007.

[7] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, “AAindex: amino acid index database, progress report 2008”, Nucleic Acid**s** Research, Vol 36(suppl_1), pp: D202 - D205, January 2008.

[8] L.Kurgan, W. Stach, and J. Ruan, “Novel scales based on hydrophobicity indices for secondary protein structure”, Journal of Theoretical Biology, Vol 248, pp: 354-366, 2007.

[9] N.C. Oza and K. Tumer, “Classifier ensembles: Select real-world applications”, Information Fusion, Vol 9, pp:4-20, January 2008.

[10] D.L. Hall and J. Llinas, “An Introduction to Mutlisensor Data Fusion”, Proceedings of the IEEE, Vol. 85, No. 1, pp: 6-23, 1997.

[11] Hong-Bin Shen and Kuo-Chen Chou, “PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition”, Analytical Biochemistry, Vol. 373, pp: 386-388, 2008.

[12] H. Cid, M. Bunster, M. Canales and F. Gazitua, “Hydrophobicity and structural classes in proteins”, Journal of Protein Engineering, Vol 5, pp: 373-375, 1992.

[13] J.H. Friedman, “Regularized Discriminant Analysis”, Journal of the American Statistical Association, Vol 84, No. 405, pp: 165-175, March 1989.

[14] L.I. Kuncheva, *Combining Pattern Classifiers: methods and algorithms*, A Wiley-Interscience publication, Canada, 2004.

[15] I.T. Jolliffe, *Principal Component Analysis*, 2nd Edition, Springer, 2002.