Türk Biyokimya Dergisi [Turkish Journal of Biochemistry–Turk J Biochem] 2012; 37 (1) ; 54–61.
doi: 10.5505/tjb.2012.44153

Research Article [Araştırma Makalesi]

Yayın tarihi 30 Mart, 2012 © TurkJBiochem.com

[Published online 30 March, 2012]

Classification of apoptosis proteins by discriminant analysis

[Apoptotik proteinlerin diskriminant analizi aracılığıyla sınıflandırılması]

Cagin Kandemir-Cavas [1],
Efendi Nasibov[1,2]

[1] Department of Computer Science, Faculty of Sciences, Dokuz Eylül University, Tinaztepe Campus, 35160 Izmir, Turkey
[2]Institute of Cybernetics, Azerbaijan National Academy of Sciences, 9, F.Agayev str., AZ-1141 Baku, Azerbaijan

Yazışma Adresi
[Correspondence Address]

Assistant Prof. Dr. Cagin Kandemir-Cavas,

Department of Computer Science, Faculty of Sciences, Dokuz Eylül University, Tinaztepe Campus, 35160 Izmir, Turkey.
Tel: +90 232 3019512
Fax: +90 232 4534265,
E-mail: cagin.kandemir@deu.edu.tr

Registered: 30 September 2011; Accepted: 30 December 2011

[Kayıt Tarihi 30 Eylül 2011; Kabul Tarihi : 30 Aralık 2011]

ABSTRACT

Objective: Prediction of the locations of apoptosis proteins which are divided into four locations as cytoplasmic proteins, plasma membrane-bound proteins, mitochondrial inner-outer proteins and other proteins; that have different biological function in each location.

Methods: In this paper, we have encoded protein sequences as amino acid composition, thereby protein sequences are expressed via high dimensional structure, and consequently this case causes computational complexity. Principal component analysis has been used to reduce the dimension of apoptosis protein sequences. After this preprocessing step, apoptosis proteins are classified by linear discriminant analysis and fuzzy linear discriminant analysis.

Results: The overall prediction accuracies for linear discriminant analysis and fuzzy linear discriminant analysis are obtained as 16.3% and 80.2%, respectively. Apoptosis proteins assigned as testing data set represent overlapping data structure, therefore linear discriminant analysis yields unsuccessful result. Since fuzzy logic is appropriate for classification and clustering of overlapping data, fuzzy linear discriminant analysis, the integration of fuzzy logic and linear discriminant analysis, gives satisfying prediction accuracy rates.

Conclusion: It can be argued that the association of fuzzy approach with other classical methods can yield higher and more robust prediction accuracy rates for the classification problems of apoptosis proteins.

Key Words: apoptosis proteins, amino acid composition, bioinformatics, principal component analysis, fuzzy linear discriminant analysis, linear discriminant analysis

Conflict of interest: Authors have no conflict of interest.

ÖZET

Amaç: Sitoplazmik proteinler, plazma membranına bağlı proteinler, mitokondri iç-dış proteinleri ve diğer proteinler olarak dört bölgeye ayrılan ve her bir bölgede farklı biyolojik işlevlere sahip olan apoptotik proteinleri doğrusal diskriminant yöntemleri aracılığıyla sınıflandırmak.

Yöntem: Bu çalışmada, protein sekansları amino asit kompozisyonu olarak kodlanmakta, böylelikle protein sekansları yüksek boyutlu yapı aracılığıyla ifade edilmekte, ve sonucunda bu durum hesapsal karmaşıklığa neden olmaktadır. Apoptotik protein sekanslarının boyutunu azaltmak amacıyla temel bileşenler analizi uygulanmıştır. Verilere uygulanan bu önişlem adımından sonra, apoptotik proteinler doğrusal diskriminant analizi ve bulanık doğrusal diskriminant analizi ile sınıflandırılmıştır.

Bulgular: Doğrusal diskriminant analizi ve bulanık doğrusal diskriminant analizi yöntemleri için doğru sınıflama yüzdeleri sırasıyla %16.3 ve %80.2 olarak elde edilmiştir. Test verisi olan apoptotik proteinler üstüste örtüşen veri yapısına sahip oldukları için doğrusal diskriminant analizi başarısız bir sonuç vermiştir. Bulanık mantığın üstüste örtüşen verilerin sınıflanması ve kümelenmesi için daha uygun olması nedeniyle, bulanık mantığın ve doğrusal diskriminant analizinin birleşimi olan bulanık doğrusal diskriminant analizi memnun edici doğruluk oranları vermiştir.

Sonuç: Apoptotik proteinlerin sınıflama problemleri için, bulanık yaklaşımın diğer klasik yöntemlerle birleşimi, daha yüksek ve daha güçlü doğruluk oranlarının elde edilmesine imkan sağlayabileceği iddia edilmektedir.

Anahtar kelimeler: apoptotik proteinler, amino asit kompozisyonu, biyoinformatik, temel bileşenler analizi, bulanık doğrusal diskriminant analizi, doğrusal diskriminant analizi

Çıkar çatışması: Yazarların çıkar çatışması yoktur.

Introduction

As a consequence of on-going genome project, the amount of biological data has increased rapidly. Therefore the necessity of using advanced computational tools to analyze the data prevails in the field of bioinformatics. The principle purpose of protein bioinformatics is to contribute both to the comprehension of the metabolic defects in the organism and to the improvement of the drug discovery studies. Nowadays, expensive and time-consuming experimental activities such as classification and clustering of proteins can be performed in a short time and with lower cost by computer-based applications [1-15].

One of the most significant protein classes for metabolic system is the apoptosis proteins which maintain the development of the organism. Apoptosis proteins have an efficient role in the process of apoptosis. Apoptosis is a programmed mechanism that cells have self-eradicated under normal physiologic conditions; which performs homeostasis in an organism [16]. Apoptosis proteins are divided into four locations, namely cytoplasmic proteins, plasma membrane-bound proteins, mitochondrial inner and outer proteins, and other proteins. Proteins in each class have different biological functions. Therefore, the determination of the locations of apoptosis proteins can give a clue about in which part of apoptosis they function. For this purpose, the classification of the location for the apoptosis proteins has been predicted by using mathematical and computational techniques. In literature, Zhou and Doctor [17] use the covariant discriminant analysis for their 98 apoptosis proteins extracted from SWISS-PROT [18]. Prediction accuracy for mitochondrial inner-outer and the others proteins, and the overall accuracy are 30.8%, 25.0%, and 72.5%, respectively. Chen and Li [19] encode the apoptosis proteins using the local composition of the amino acid pairs and the hydropathy distribution; and classify the proteins with the diversity measure. The performance of their method is 82.7%.

In the literature, proteins are encoded as vectorial expressions in order to integrate proteins into the computer-based applications. One of the most popular expression methods is the amino acid composition. With this method the set of protein sequences are obtained as high dimensional data. However, computer-based algorithms do not perform well and the quality of prediction accuracy rates of classification or clustering decreases due to the high dimensional effect. Therefore, in order to deal with this computational complexity and to make fast and efficient prediction, dimension reduction techniques have been developed. The principle component analysis (PCA), which is one of these techniques, is used in this paper to reduce the dimensions of the encoded protein sequences. After the dimension reduction, the structure of apoptosis proteins has been characterized as an overlapped feature. In addition, the linear discriminant analysis (LDA) and the fuzzy linear discriminant analysis (FLDA) are used to classify the apoptosis proteins. The results illustrate that the performance of LDA is lower than FLDA for apoptosis protein sequences. Since the FLDA fuzzifies the LDA, it gives better and more robust results. It can be argued that FLDA is a successful method in prediction of the subcellular location of apoptosis proteins.

In this current paper, apoptosis protein sequences chosen as testing and training dataset in each class and encoding scheme are presented in Section 2. Algorithm steps of PCA, LDA and FLDA are elaborated in Sections 3.1, 3.2 and 3.3, respectively. Since the derivation of membership degrees is the key step of FLDA, fuzzy c-means algorithm (FCM) is detailed in Section 3.3.1. Finally, the prediction performances of LDA and FLDA for all three locations of testing dataset are conferred in Section 4.

Materials and Methods

Dataset

The testing dataset is obtained from Zhou and Doctor [17] and the dataset generated by Chen and Li [19] is used as the training dataset. Proteins in those datasets were extracted from SWISS-PROT data bank [18]. Each subcellular location category of the datasets is shown in Table 1. In this paper, due to the application of the two different datasets, we select their three common locations, namely, 1) cytoplasmic, (2) plasma membrane-bound, (3) mitochondrial.

Table 1. Number of apoptosis protein sequences in subcellular locations in testing and training datasets.

Dataset	Subcellular localization	Number of sequences
Test	Cytoplasm	43
	Membrane	30
	Mitochondria	13
Training	Cytoplasm	112
	Membrane	55
	Mitochondria	34

Since a protein sequence is composed by amino acid chains and twenty kinds of amino acids exist in the form of proteins, it is possible to express the protein sequence, x, as a composition of amino acids defined by the following vector [20]:

$$F(x) = \left[f_1(x), f_2(x), f_3(x), \ldots, f_{20}(x) \right] \quad (1)$$

The amino acid frequencies were calculated as in Eq. (2). The percentage of the amino acid residues i in a protein x is defined by:

$$f_i(x) = 100 \times \frac{n_i}{N} \qquad i = 1,2,\ldots,20 \quad (2)$$

where n_i is the number of amino acid i and N is the total number of amino acid residues in the protein sequence [20].

All sequences are encoded as 1-by-20 vector which explains the frequencies of each amino acid. In order to transform all sequences to amino acid composition, a short programme is written in Matlab R2007a.

Principal Component Analysis (PCA)

In our study, each protein is expressed as 1-by-20 vector. Therefore, the dimension of such protein data should be reduced in order to deal with the computational complexity. For this purpose, the principal component analysis (PCA) is applied to reduce the high-dimension of the input protein sequence data.

PCA is the initial stage of extensive biological studies [21-24]. The basic aim of the PCA is to explain the total system variability by a small number *k* of the principal components while *p* components explain the total variability [25].

Steps of the PCA are as follows,

Arrange the data $\mathbf{S} = \left[\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}\right]$ as a matrix comprising of *N* variables (columns) and each has *M* dimensions (rows) such as $\mathbf{x_i} = \left(x_1, x_2, \ldots, x_M\right)^T$.

Calculate the mean vector $\boldsymbol{\mu}$ along each dimension $\{1,2,\ldots,M\}$ as,

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x_i} \,. \tag{3}$$

Find the centered data vectors, where $i = 1,2,\ldots,N$, $\mu = \left(\mu_1, \mu_2, \ldots, \mu_M\right)^T$.

$$\mathbf{d_i} = \mathbf{x_i} - \boldsymbol{\mu}. \tag{4}$$

Find the covariance matrix as

$$\mathbf{C} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{d_i}\mathbf{d_i^T} \,. \tag{5}$$

where $\mathbf{d} = \left(\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_M}\right)^T$.

Find the eigenvectors and eigenvalues of the covariance matrix via

$$\mathbf{v^{-1}Cv} = d \tag{6}$$

where **v** and *d* denote eigenvectors and eigenvalues, respectively.

Sort the eigenvectors in decreasing order according to the corresponding eigenvalues.

Transform eigenvectors and arrange them to form the row-vectors of the transformation matrix **T**.

Project a new data **A** into the eigenspace as follows,

$$\mathbf{Y} = \mathbf{T}\left(\mathbf{a} - \boldsymbol{\mu}\right) \tag{7}$$

where $\mathbf{a} = \left(a_1, a_2, \ldots, a_M\right)^T$ and $\mathbf{y} = \left(y_1, y_2, \ldots, y_N, 0, \ldots, 0\right)$.

The identification of the numbers of components is usually rather hard task for noisy data matrices. A component which corresponds to an approximately zero eigenvalues can be assumed negligible. An appropriate number of principal components can be determined by a scree plot in which the eigenvalues with decreasing order are illustrated versus their component number. One must search for the significant "knee" in this plot. The number of components at which the "knee" is observed indicates the appropriate number of principal components. If the eigenvalues are relatively small, the loss of information due to dimension reduction does not occur [25].

Linear Discriminant Analysis (LDA)

LDA was first introduced by R. A. Fisher [26] who tried to find a good projection line in order to obtain well-separated classes. With this respect, the method separates two classes of objects by maximizing the ratio of between-class variance to within-class variance in any particular dataset as follows.

$$J\left(\mathbf{A}\right) = \max_{\mathbf{A}} \frac{\mathbf{A}^T\mathbf{S}_B\mathbf{A}}{\mathbf{A}^T\mathbf{S}_W\mathbf{A}} \tag{8}$$

where $J\left(\mathbf{A}\right)$ is called as objective function and $\mathbf{S}_B$ and $\mathbf{S}_W$ denote the between- and within- class scatter matrix, respectively. Their definitions are as follows.

$$\mathbf{S}_B = \sum_c \left(\boldsymbol{\mu}_c - \overline{\mathbf{x}}\right)\left(\boldsymbol{\mu}_c - \overline{\mathbf{x}}\right)^T \,, \tag{9}$$

$$\mathbf{S}_W = \sum_c \sum_{i \in c} \left(\mathbf{x}_i - \boldsymbol{\mu}_c\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_c\right)^T \tag{10}$$

where c is the number of classes, $\boldsymbol{\mu}_c$ is the mean of the data in each class, $\overline{\mathbf{x}}$ is the overall mean of the data, $\mathbf{x}_i$. According to Eq. (8), one must find a linear transform matrix **A**, which maximizes the objective function besides maximizing the projected class means and minimizing the class variances [25]. The vector of coefficients of **A** is given by the eigenvectors, **v,** which corresponds to the nonzero eigenvalues λ of the $\mathbf{S}_W^{-1}\mathbf{S}_B$ matrix. The normalized eigenvectors **v** are obtained by solving the generalized eigenvalues problem via

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{v} = \lambda\mathbf{v} \,. \tag{11}$$

Then the discriminates are

$$\mathbf{y} = \mathbf{v}^T\mathbf{x} \,. \tag{12}$$

Thereby the projected means of classes are

$$\overline{\mathbf{y}} = \mathbf{v}^T\overline{\mathbf{x}}_j \text{ for } j \in c \,. \tag{13}$$

Then **x** can be allocated to a class j which minimizes the following Euclidean distance

$$\left\|\mathbf{v}^T\left(\mathbf{x} - \overline{\mathbf{x}}_j\right)\right\| = \left\|\mathbf{y} - \overline{\mathbf{y}}_j\right\| \,. \tag{14}$$

3.3 Fuzzy aspects

A data point belongs to only one cluster in the crisp clustering technique. However, this case is quite impossible

in practice. Fuzzy set theory implies the representation of vagueness in everyday life. Then, the terms of belonging or not belonging have a flexible nature in fuzzy set theory. Fuzziness provides data points which may belong to more than one cluster. Thus, the membership degree of a data point is defined as a degree of compatibility or similarity value to the corresponding fuzzy set. Therefore, the membership value of data points to clusters can range between 0 to 1 [27].

Let A be a fuzzy set of the universe X and, $\mu_A(x)$ is the membership degree of an element $x \in X$ to the fuzzy set A. Then

$$\sum_{x \in X} \mu_A(x) \tag{15}$$

is the fuzzy cardinality of the fuzzy set A. The concepts of fuzzy sets theory such as membership degree, fuzzy cardinality,… etc. help to describe the characteristics of complex or ill-defined systems that are not possible to express with precise mathematical analysis.

3.3.1 Fuzzy c-means (FCM) algorithm

The critical part of the FLDA is to assign the membership degrees of each data point to the classes. Chen et al. [28] propose a K-nearest neighbor (KNN) rule to obtain the fuzzy membership degrees. Since FCM is more exploratory for fuzziness of overlapping data, in our study FCM algorithm is used to obtain membership degrees of apoptosis protein sequences.

The fuzzy clustering provides opportunity to express data points which belong to more than one cluster by giving membership degrees in $[0,1]$. FCM is also based on this assumption and aims to minimize the following objective function.

$$J\left(\mathbf{x}, \mathbf{U}_f, C\right) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \left\| x_j - v_i \right\|^2 \tag{16}$$

where x_j is the *j*th data, v_i is the d-dimension center of the cluster i, u_{ij} is the degree of membership of x_j to the cluster i, $\|.\|$ is any norm expressing the distance between x_j and v_i, and m is fuzzifier or weighting exponent greater than 1 [29]. The greater the value of *m*, the greater the fuzziness of clustering, in other words the generated clusters have an overlapping structure while m increases. For the Eq. (16), Bezdek [30] proposes to select m = 2. Besides, the other studies related to FCM put forward the optimal value of *m* as 2 [31]. Therefore, in this paper, the value of *m* is specified as 2.

In addition $\mathbf{U}_f = \left(\mathbf{u}_1, \ldots, \mathbf{u}_n\right)^T$ where $\mathbf{u}_j = \left(u_{1j}, \ldots, u_{cj}\right)^T$ is a $c \times n$ matrix that denotes the fuzzy partition matrix, $\mathbf{x} = \left\{\mathbf{x}_1, \ldots, \mathbf{x}_n\right\}$ is the set of given data objects and $C = \left\{C_1, \ldots, C_c\right\}$ is the set of cluster prototypes.

The following constraints

$$\sum_{j=1}^{n} u_{ij} > 0, \quad \forall i \in \left\{1, \ldots, c\right\} \tag{17}$$

$$\sum_{i=1}^{c} u_{ij} = 1, \quad \forall j \in \left\{1, \ldots, n\right\} \tag{18}$$

have to be satisfied when the objective function $J\left(\mathbf{x}, \mathbf{U}_f, C\right)$ is to be minimized.

The steps of FCM algorithm are as follows,

Choose any cluster prototype $C = \left\{C_1, \ldots, C_c\right\}$ for the values of membership values.

Calculate the membership values according to the following formula,

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| \mathbf{x}_j - v_i \right\|}{\left\| \mathbf{x}_j - v_k \right\|} \right)^{\frac{2}{m-1}}} \tag{19}$$

with the existence of a datum $\mathbf{x}_j$ with zero distance to some $v_{i_1}, v_{i_2}, \ldots, v_{i_t}$ class centers, $u_{ij} = 1/t$ for $i \in \{i_1, \ldots i_t\}$ and $u_{ij} = 0$ for $i \quad \{i_1, \ldots i_t\}$.

Calculate the new cluster prototypes $C = \left\{C_1, \ldots, C_c\right\}$

$$C_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} \mathbf{x}_j}{\sum_{j=1}^{n} u_{ij}^{m}}, i = 1, .., c \tag{20}$$

Iterate the steps 2 and 3 until the memberships or cluster centers for consecutive iterations differ by more than a threshold ε (a termination criterion).

With FCM, the centroid of a cluster is calculated as being the mean of all points, weighted by their degree of belonging to the cluster. The degree of being in a certain cluster is related to the inverse of the distance to the cluster [31]. But LDA takes into account the equal importance for all data [28]. Hereby the general structure of the FLDA model is obtained by modifying the LDA approach in terms of the identification of membership value for each data point to the related classes.

FLDA tries to maximize the following objective function,

$$J\left(\mathbf{B}\right) = \max_{\mathbf{B}} \frac{\mathbf{B}^T \mathbf{S}_B \mathbf{B}}{\mathbf{B}^T \mathbf{S}_W \mathbf{B}} \tag{21}$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ denote the between- and within-class scatter matrix, respectively, and

$$\mathbf{S}_B = \sum_c \left(U_c . \left((\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^T \right) \right) \quad (22)$$

$$\mathbf{S}_W = \sum_c \sum_j \left((\mathbf{x}_j - \mathbf{m}_c)(x_j - \mathbf{m}_c)^T \, u_{cj} \right) \quad (23)$$

where u_{cj} is the membership degree of data point $\mathbf{x}_j$ to the class c, U_c denotes the fuzzy cardinality of the class *c*, $\mathbf{m}_c$ is the fuzzy mean vector of a class *c*, and $\mathbf{m}$ is the fuzzy mean vector of the whole data points that is calculated as follows.

$$U_c = \sum_j u_{cj}, \quad (24)$$

$$\mathbf{m}_c = \sum_j u_{cj} \mathbf{x}_j \Big/ U_c \quad , \quad (25)$$

$$\mathbf{m} = \sum_c \sum_j u_{cj} \mathbf{x}_j \Big/ U_T \quad (26)$$

where U_T is the fuzzy cardinality of whole data set which is calculated as

$$U_T = \sum_c U_c \, . \quad (27)$$

As in LDA, the objective function can be maximized by obtaining **B** matrix given by the eigenvectors **v** which corresponds to nonzero eigenvalues λ of the $\mathbf{S}_W^{-1}\mathbf{S}_B$ matrix. The generalized eigenvalues problem is solved as in Eqs. (11)–(14).

Prediction Performance

Overall accuracy and sub-class accuracy are measured as

$$\text{Overall accuracy} = \frac{\sum_{i=1}^{c} \text{TP}_i}{\text{N}} \quad (28)$$

$$\text{Subclass accuracy} = \frac{\text{TP}_i}{\text{n}_i}, \quad (29)$$

respectively, where TP_i (true positives) is the number of correctly predicted proteins in location i, N is the total number of sequences, and n_i is the total number of sequences existing in location i.

Results and Discussion

An independent dataset, the jackknife test and the re-substitution test have been used by many researchers as the statistical prediction methods. In this study, the dimensions of testing and training datasets are denoted in Table 1 and the performance of the prediction for the LDA and FLDA methods in testing (independent) dataset is summarized in Table 2.

In this paper, for encoding of amino acids, the algorithms of PCA, LDA and FLDA are coded in Matlab 2007a. Also LDA and FLDA are used to discriminate apoptosis proteins. Before the analysis of LDA and FLDA, the dimension of the apoptosis protein sequences has been reduced by PCA. In order to identify appropriate number of components, scree plots of training and testing dataset are obtained as seen in Fig. 1 and Fig. 2, respectively. It can be seen that the significant knee is observed at mostly about the 2nd component value in the figures. That is the component values after 2 are all relatively small. Therefore protein sequences expressed by 1-by-20 can be reduced as 1-by-2 vector.

The overall prediction accuracies for LDA and FLDA are 16.3%and 80.2%, respectively, as illustrated in Table 2. This result shows that for apoptosis protein data (one of the classical methods) LDA can not yield high classification accuracy. As seen in Fig. 3, apoptosis proteins sequences chosen as testing data which are reduced to 2-D by PCA has overlapping data structure, especially, in mitochondrial and membrane protein sequences, therefore LDA is unsuccessful in classifying the given apoptosis protein data. However, these overlapping data points have been detected by FLDA and their negative effects on the classification algorithm is reduced.

Furthermore, although Chen and Li [19] used to local compositions of twin amino acids to encode the protein sequences, in this paper, their dataset was encoded

Table 2. Overall accuracy and sub-class accuracy achieved in testing dataset for LDA and FLDA

Subcellular location	LDA sub-location accuracy	LDA overall accuracy	FLDA sub-location accuracy	FLDA overall accuracy
Cytoplasm	50.0%		97.5%	
Membrane	0.23%	**16.3%**	72.5%	**80.2%**
Mitochondria	29.3%		16.7%	

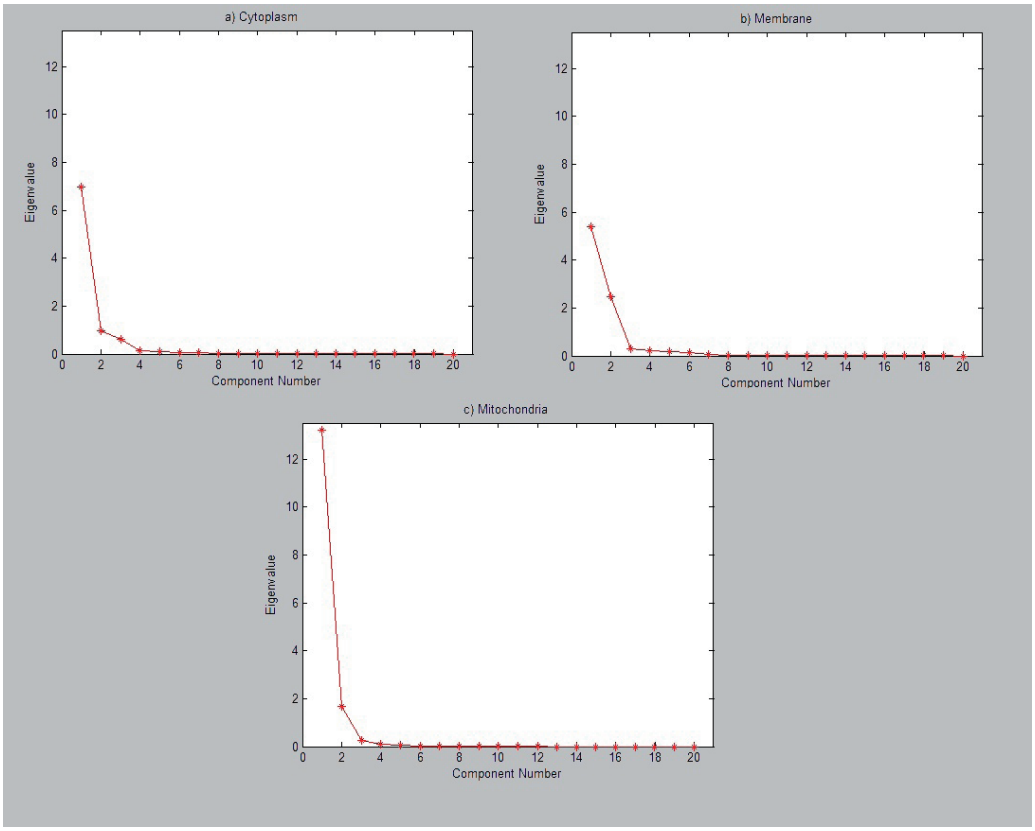

Figure 1. Scree plots of a) cytoplasm, b) membrane and c) mitochondria locations in testing dataset.

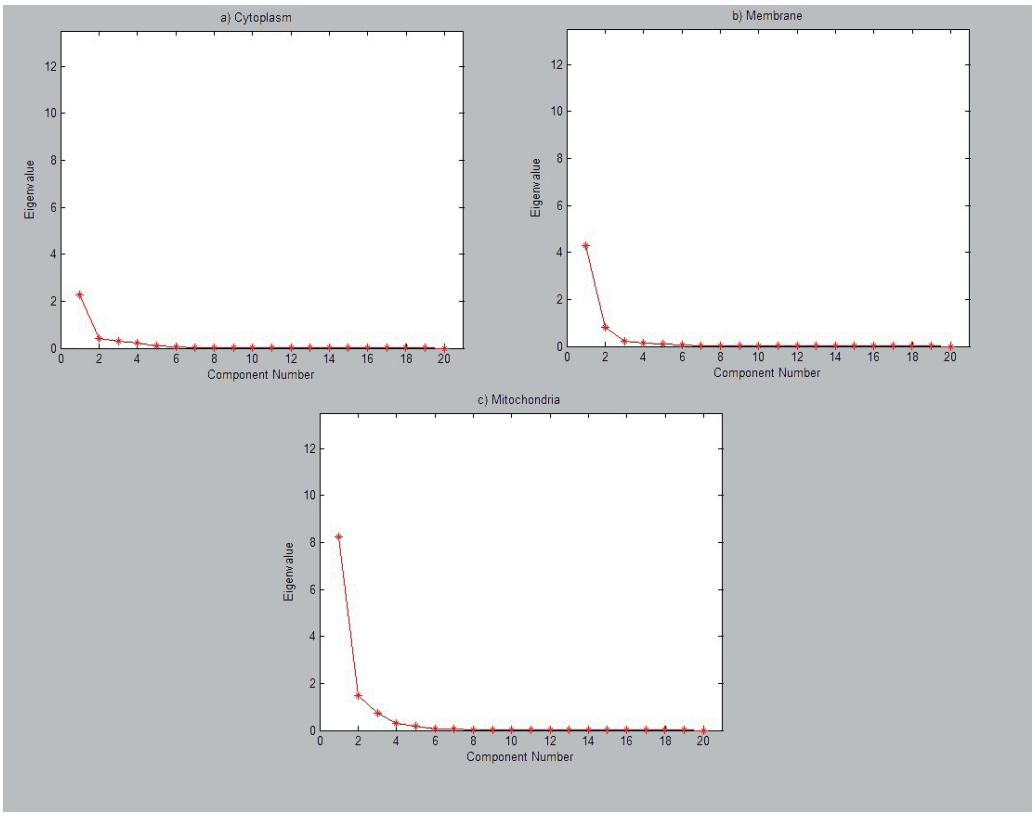

Figure 2. Scree plots of a) cytoplasm, b) membrane and c) mitochondria locations in training dataset.

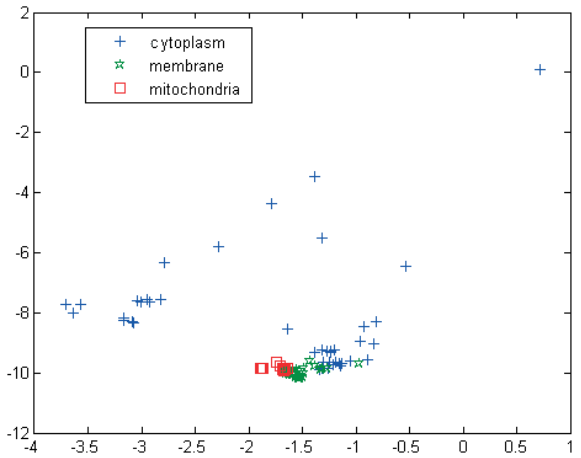

Figure 3. Overlapping data structure of the apoptosis proteins chosen as testing data (Scatter plot of the data)

by amino acid composition in order to make comparison with the results of FLDA. Then, their algorithm called the increment of diversity (ID), is performed to the same (PCA-used) dataset and the prediction accuracy rate is found as 45.3%. Thereby, FLDA yields higher prediction performance than LDA and ID.

Since a protein can be part of two or more clusters simultaneously, with partial or full membership in each cluster, the fuzzy logic is favorable to define some biological systems [32, 33].

Conclusions

In this paper, first of all, apoptosis proteins have been encoded as their amino acid compositions. The dimension of the encoded data has been reduced by the linear PCA in order to reduce the computational complexity. The classification of apoptosis protein sequences is performed via LDA and FLDA.

LDA is a classification method for the multi-class dataset where the classes are linearly separable. However, it yields insufficient results for the data points with overlapping classes. Since the fuzzy logic is based on managing the overlapping data structure, LDA modified by the fuzzy approach, also called FLDA, can overcome the difficulty of the feature separation for the data points in different classes. Thus, the performance of FLDA is superior to LDA for the selected datasets.

LDA and FLDA are the linear aspects for the classification problem as the future work, the nonlinear perspective of the relationships between proteins can be studied by addressing their phylogenetic structure.

Acknowledgements

We are grateful to referees for their valuable comments and suggestions on our manuscript.

Conflict of Interest: Authors have no conflict of interest.

References

[1] Horton P, Nakai K. (1997). Better prediction of protein cellular localization sites with the k- nearest neighbors classifier. ISMB 5:147–152.

[2] Reinhardt A, Hubbard T. (1998). Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res. 26: 2230–2236.

[3] Yuan Z. (1999). Prediction of protein subcellular locations using Markov chain models. FEBS Lett. 451:23–26.

[4] Hua S., Sun Z. (2001). Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17:721–728.

[5] Chou KC, Cai YD. (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. J. Biol. Chem. 277:45765–45769.

[6] Cai YD, Chou KC. (2003). Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochem. Biophys. Res. Commun. 305:407–411.

[7] Huang Y, Li Y. (2004). Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20:121–128.

[8] Gao QB, Wang ZZ. (2005). Using nearest feature line and tunable nearest neighbor methods for prediction of protein subcellular locations. Comput. Biol. Chem. 29:388–392.

[9] Zhang T, Ding Y, Chou KC. (2006). Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence. Comput. Biol. Chem. 30:367–371.

[10] Nasibov E, Kandemir-Cavas C. (2008). Protein subcellular location prediction using optimally weighted fuzzy k-NN algorithm. Comput. Biol. Chem. 32:448–451.

[11] Chi SM. (2010). Prediction of protein subcellular localization by weighted gene ontology terms. Biochem. Biophys. Res. Commun. 399:402-405.

[12] Shi SP, Qiu JD, Sun XY, Huang JH, Huang SY, Suo SB, Liang RP, Zhang L. (2011). Identify submitochondria and subchloroplast locations with pseudo amino acid composition: Approach from the strategy of discrete wavelet transform feature extraction. Biochim. Biophys. Acta (BBA)–Molecular Cell Research, 1813:424–430.

[13] Nasibov E, Kandemir-Cavas C. (2009). Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction. Comput. Biol. Chem. 33:461–464.

[14] Fang Y, Ma D, Li M, Wen Z, Diao Y. (2010). Investigation of the proteins folding rates and their properties of amino acid networks. Chemometr. Intell. Lab. 101:123–129.

[15] Nasibov E, Kandemir-Cavas C. (2011). OWA-based linkage method in hierarchical clustering: Application on phylogenetic trees. Expert. Syst. Appl. 38:12684–12690.

[16] Kerr JFR, Wyllie AH, Currie AR. (1972). Apoptosis: A basic biological phenomenon with wideranging implications in tissue kinetics. Br. J. Cancer 26:239–257.

[17] Zhou GP, Doctor K. (2003). Subcellular location prediction of apoptosis proteins. Proteins Struct. Funct. Genet. 50:44–48.

[18] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. (2003). The SWISS PROT protein knowledgebase and its supplement TrEMBL. Nucleic Acids Res. 31:365–370.

[19] Chen YL, Li QZ. (2007). Prediction of the subcellular location of apoptosis proteins. J. Theor. Biol. 245:775–783.

[20] Cedano J, Aloy P, Pérez-Pons JA, Querol E. (1997). Relation be-

tween amino acid composition and cellular location of proteins. J. Mol. Biol. 266:594–600.

[21] Liu L, Zhang J, Chen B, Shao W. (2004). Principle component analysis in F/10 and G/11 xylanase. Biochem. Biophys. Res. Commun. 322:277–280.

[22] Bharanidharan D, Gautham N. (2006). Principal component analysis of DNA oligonucleotide structural data. Biochem. Biophys. Res. Commun. 340:1229-1237.

[23] Sarıyar B, Perk S, Akman U, Hortaçsu A. (2006). Monte Carlo sampling and principal component analysis of flux distributions yield topological and modular information on metabolic networks. J. Theor. Biol. 242:389-400.

[24] Tsai CY, Chiu CC. (2008). An efficient conserved region detection method for multiple protein sequences using principal component analysis and wavelet transform. Pattern Recogn. Lett. 29:616-628.

[25] Johnson RA, Wichern DW. (2007). Applied Multivariate Statistical Analysis, Pearson Education, New Jersey.

[26] Fisher RA. (1938). The statistical utilization of multiple measurements. Ann. Eugenics 8:376-386.

[27] Zadeh LA. (1965). Fuzzy sets. Inf. Control 8:338-353.

[28] Chen ZP, Jiang JH, Li Y, Liang YZ, Yu RQ. (1999). Fuzzy linear discriminant analysis for chemical data sets. Chemometr. Intell. Lab. 45:295–302.

[29] Mitra S, Acharya T. (2003). Data Mining, Multimedia, Soft Computing, and Bioinformatics, John Wiley and Sons, New Jersey.

[30] Bezdek JC. (1976). A physical interpretation of Fuzzy ISODATA, IEEE Trans. Syst., Man, Cybern. SMC-6:387–390.

[31] De Oliveira JV, Pedrycz W. (2007). Advances in Fuzzy Clustering and its Applications, John Wiley and Sons, West Sussex.

[32] Chang B, Halgamuge S. (2002). Protein Motif Extraction with Neuro-Fuzzy Optimisation. Bioinformatics 18:1804–1090.

[33] Dong X, Keller JM, Popescu M, Bondugula R. (2008). Applications of fuzzy logic in bioinformatics. Series on Advances in Bioinformatics and Computational Biology Vol 9, Imperial College Press, London.