DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DISCRIMINANT ANALYSIS USING KERNEL AND FUZZY APPROACHES

by

Selen YILDIRIM

October, 2014

İZMİR

DISCRIMINANT ANALYSIS USING KERNEL AND FUZZY APPROACHES

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of Science in
Statistics, Statistics Program

by
Selen YILDIRIM

October, 2014
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"DISCRIMINANT ANALYSIS USING KERNEL AND FUZZY APPROACHES"** completed by **SELEN YILDIRIM** under supervision of **ASSOC. DR. ÇAĞIN KANDEMİR ÇAVAŞ** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Çağın KANDEMİR ÇAVAŞ

Supervisor

Prof. Dr. Efendi Nasibov

(Jury Member)

Asst. Prof. Dr. Celal Cem Sarıoğlu

(Jury Member)

Prof.Dr. Ayşe OKUR
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to express my special appreciation and thanks to my advisor Assoc. Prof. Dr. Çağın KANDEMİR ÇAVAŞ, for the continuous support of my Master study, for her patience, motivation and immence knowledge. Her guidance and help played a very important role in not only this study but also all my research.

Besides my advisor, I would like to thank Prof. Dr. Efendi NASİBOV due to the opportunities provided me in the Departmant of Computer Science.

I would like to express my heartfelt thanks to my parents Ünal YILDIRIM, Hatice YILDIRIM, Seçil YILDIRIM and all my family for their constant love plus moral and material support. They encouraged me to strive towards my goal and supported me in everything I decide to do.

Lastly, I would like to thank Mert KÜÇÜK for his continuous support, his patient and moral during my thesis process.

Selen YILDIRIM

DISCRIMINANT ANALYSIS USING KERNEL AND FUZZY APPROACHES

ABSTRACT

Proteins are molecular structures that are formed by chains of amino acids with a variety of shapes, size and chemical properties. Protein functions are determined by according to their protein structures. Different types of protein sequences have their specific biochemical function. For this reason, classification of protein sequences is of great importance in terms of identifying their tasks in the biological cycle and understanding the functions of proteins. The purpose of this thesis is to classify the ordered and disordered proteins. Intrinsically disordered proteins formed as a result of deformations happen in the tertiary strucure of a protein. Disordered proteins can be identified by using experimantal methods such as Electron microscopy, nuclear magnetic resonance (NMR) spectroscopy and protein crystallography. However, experimental classification is costly and time-consuming because of great amount of raw sequences. Therefore, computational solutions have become a useful tool for their analysis. For this purpose, Fisher linear discriminant analysis, fuzzy discriminant analysis and kernel discriminant analysis methods were applied to the ordered disordered proteins. The overall prediction accuracies obtained from jackknife validity technique for Fisher's linear discriminant analysis, Fuzzy linear discriminant analysis and kernel discriminant analysis are found as 77.1930 percent, 87.7193 percent and 100 percent, respectively. According to the results, it can be strongly referred that the method of kernel discriminant analysis can be a very robust method to identify whether a protein has a disordered or ordered structure.

Keywords: Bioinformatics, classification, ordered-disordered proteins, fuzzy discriminant analysis, kernel discriminant analysis.

DİSKRİMİNANT ANALİZİNDE KERNEL VE BULANIK YAKLAŞIMLARININ KULLANILMASI

ÖZ

Proteinler çeşitli şekillerde, boyutlarda ve kimyasal özelliklerde amino asit zincirlerinden oluşan moleküler yapılardır. Protein fonksiyonları proteinlerin yapılarına göre belirlenir. Farklı çeşitlerdeki protein sekanslarının kendilerine özgü biyokimyasal fonksiyonları vardır. Bu nedenle, protein sekanslarını sınıflamak biyolojik döngüde onların görevlerini tanımlamak ve proteinlerin fonksiyonlarını anlamak açısından büyük önem taşır. Bu tezde düzenli ve düzensiz proteinlerin sınıflandırılması amaçlanmıştır. Kendiliğinden düzensiz proteinler, bir proteinin tersiyer yapısında meydana gelen bozulma sonucu oluşurlar. Düzensiz proteinler elektron mikroskobu, nükleer manyetik rezonans spekstroskopisi ve protein kristalleştirme yöntemleri kullanılarak deneysel olarak tanımlanabilir. Ancak deneysel sınıflama ham sekansların büyük miktarda olmasından dolayı yüksek maaliyetlidir ve zaman alır. Bu nedenle bilgisayar tabanlı çözümler analiz için kullanışlı bir araç olmaktadır. Bu amaç doğrultusunda Fisher'ın doğrusal diskriminant analizi, bulanık diskriminant analizi ve kernel diskriminant analizi yöntemleri düzenli ve düzensiz proteinlere uygunlanmıştır. Fisher'ın doğrusal diskriminant analizi, bulanık diskriminant analizi ve kernel diskriminant analizi için yeniden yerine koyma geçerlilik yönteminden elde edilen tüm doğru sınıflama yüzdeleri sırasıyla yüzde 77.1930, yüzde 87.7193 ve yüzde 100 olarak bulunmuştur. Sonuçlara dayanılarak, kernel diskriminant analizi yönteminin bir proteinin düzenli ve ya düzensiz yapıya sahip olup olmadığını tanımlamakta çok güçlü bir method olduğu yargısına varılır.

Anahtar kelimeler: Biyoinformatik, sınıflama, düzenli-düzensiz proteinler, bulanık diskriminant analizi, kernel diskriminant analizi

CONTENTS **Page**

THESIS EXAMINATION RESULT FORM ii
ACKNOWLEDGEMENTS iii
ABSTRACT iv
ÖZ v
LIST OF FIGURES viii
LIST OF TABLES ix

CHAPTER ONE - INTRODUCTION 1

CHAPTER TWO – PRINCIPLE COMPONENT ANALYSIS 4

2.1 Principle Component Analysis 4
2.2 Kernel Principle Component Analysis 6

CHAPTER THREE – DISCRIMINANT ANALYSIS 8

3.1 Fisher's Linear Discriminant Analysis 8
3.1.1 Wilk's Lambda Test Statistic 9
3.1.2 Computing the Fisher Linear Discriminant Analysis Coefficients 10
3.2 Validity of Discriminant Analysis 11
3.2.1 Resubstitution Method 11
3.2.2 Jackknife Method 12
3.3 Fuzzy Linear Discriminant Analysis 12
3.3.1 Fuzzy Set Theory 12
3.3.2 Fuzzy C-Means 14
3.3.3 Computing the Fuzzy Linear Discriminant Analysis Coefficients 14
3.4 Kernel Discriminant Analysis 17
3.4.1 Kernel Trick 18
3.4.2 Kernel Functions 19
3.4.2.1 Linear Kernel 19

3.4.2.2 Polynomial Kernel 19
3.4.2.3 Gaussian Kernel 19
3.4.2.4 Anova Kernel 20
3.4.2.5 Exponential Kernel 21
3.4.2.6 Laplacian Kernel 21
3.4.3 Computing the Kernel Discriminant Analysis Coefficients 21

CHAPTER FOUR - BIOINFORMATICS 23

CHAPTER FIVE – DATA SET AND EXPERIMENTAL RESULTS 26

5.1 Data Set 26
5.2 Experimental Results 29

CHAPTER SIX - CONCLUSIONS 37

REFERENCES 39

LIST OF FIGURES

Page

Figure 2.1 Graphical representation of principle components 5
Figure 3.1 Visual representation of Fisher's linear discriminant analysis 8
Figure 3.2 Visual scheme of kernel trick 17
Figure 3.3 Visual shape of mapping funcion 18
Figure 4.1 The diagram of the contents of the bioinformatics 24
Figure 4.2 The list of amino-acids 25
Figure 4.3 The FASTA format of a protein sequence 25
Figure 5.1 Basic figure of (a) disordered; (b) ordered proteins 26
Figure 5.2 Visual respresentation of the data set after digitized 28
Figure 5.3 Scree plot of ordered proteins 28
Figure 5.4 Scree plot of disordered proteins 29
Figure 5.5 Scatter plot of discriminant scores of Fisher's LDA for ordered proteins 30
Figure 5.6 Scatter plot of discriminant scores of Fisher's LDA for disordered proteins 31
Figure 5.7 Scatter plot of discriminant score of FLDA 32
Figure 5.8 Distribution of the data 34
Figure 5.9 Scatter plot of discriminant scores of KDA 34

LIST OF TABLES

Page

Table 3.1 Example of discriminant scores table 11
Table 5.1 Coefficients of Fisher' LDA discriminant function 29
Table 5.2 Coefficients of FLDA discriminant function 31
Table 5.3 Coefficients of KDA discriminant function 33
Table 5.4 Table of resubstitution validity technique and accuracy rates 35
Table 5.5 Jackknife method validity results and accuracy rates 36

CHAPTER ONE
INTRODUCTION

Pattern analysis is the process of finding basic relations in a set of data, and also is the basic step of many disciplines (Shawe-Taylor & Cristianini, 2004). The goal of the pattern recognition is to describe, recognize, classify, and group patterns of interest (Bishop, 1995). Briefly, pattern is a quantitative or structural description of an object such as handwritten cursive words, speech signals, odor signals, fingerprint images, animal footprints, human faces or any type of measurements that need to be classified (Duda, Hart & Stork, 2001). In addittion, pattern analysis have been used in bioinformatics field in the recent years. Bioinformatics obtains solutions for the biological problems via computer science, statistics, informatics, applied mathematics.

Some of the scope of bioinformatics are as follows:

- Sequence analysis
- Computational evolutionary biology
- Measuring biodiversity
- Analysis of gene expression
- Analysis of protein expression
- Modelling biological systems
- Finding similar sequences in the database with a query sequence.
- Finding similar protein structures in the database with a query protein structure.
- Finding sequences in the databank that correspond to similar structures with a query protein structure

Proteins are large molecular structures consisting of chains of amino acids with a variety of shapes, size and chemical properties. Different types of protein sequences have specific biochemical function. Therefore, classification of protein sequences is very important process to identify their structure. Unlike, classifying proteins experimentally computational methods are easy to apply and less costly.

Computational methods are useful tools for protein classification. There are various computational methods to predict protein class in literature. One of the multivariate statistical models to make the classification is the linear discriminant analysis that is mostly used in bioinformatics (Wang et al., 2013; Li et al., 2010; Huang et al., 2012). Discriminant analysis can be extended with many other structures such as fuzzy, kernel, quadratic, etc. In this thesis, fuzzy and kernel discriminant analysis are investigated. In literature, fuzzy discriminant analysis and non-linear based discriminant analysis are also used in bioinformatics problems (Cadena et al., 2007; Khashei et al., 2012; Lin, 2008; Diaf, Boufama & Benlamri, 2013; Harrison & Pasupa, 2005).

Linear discriminant analysis (LDA) developed in 1936 by R. A. Fisher. This development was largely based on the methodology of multiple regression. Discriminant analysis (DA) is only used for the data classification which is one of the most widely used classification method in pattern recognition. It determines the differences between classes and it tries to find best classification equation by using data structure. DA aims to maximize the ratio of between class variance and within class variance in order to find the best equation which can represent the classes.

Fisher's linear discriminant analysis (Fisher's LDA) has some assumptions such as normality condition, equality of variance-covariance matrix and it consists of the principle of the crisp set theory. The principle of the crisp set theory says that although the overlapped points have different properties, they are assigned to the same classes. Because of these assumptions, Fisher's LDA is not sensitive for the overlapped points. For all these reasons, the alternative discriminant analysis techniques were improved.

One of the most widely used discriminant analysis techique is fuzzy linear discriminant analysis (FLDA). As mentioned before, Fisher's LDA is not sensitive for the overlapped points. On the contrary, FLDA is sensitive for overlapped points in the data set. In FLDA, overlapped points can not always belong to the same classes. Each point has different degree of membership to the each class and the data

can be classified according to their degrees of membership. In order to determine the degrees of memberships the fuzzy set theory is used. The detailed fuzzy set theory will be given in the cheapter 3, section 3.3.1.

Fisher's LDA has some assumptions as mentioned before. The most important assumption is normality condition. In real world, this assumption may not always be provided. So, it is convenient to use non-linear methods. But, Fisher's LDA and FLDA are not good at solving non-linear problems. Therefore, the kernel discriminant analysis (KDA) was improved. KDA is a non-linear DA techique and it has no assumptions unlike Fisher's LDA. The distribution of the data set is not important to use the kernel method. KDA can be applied on both linear or non-linear data set.

Classification of ordered-disordered proteins is a very important problem in bioinformatics. Many disease-based proteins have disordered structure while the ordered proteins are not related with diseases mostly. In this thesis, three discriminant analysis methods are used in order to classify the protein sequences as "ordered" or "disordered" structures by mathematically and algorithmically. In the second chapter, the method of principle component analysis is explained which is used for dimensionality reduction in this thesis. Chapter three refers three discriminant analysis methods such as Fisher linear, fuzzy and kernel. Chapter four includes general informations about bioinformatics. The data set and experimental results are mentioned in cheapter five and chapter six includes conclusions.

CHAPTER TWO
PRINCIPLE COMPONENT ANALYSIS

2.1 Principle Component Analysis

Principle Component Analysis (PCA) is an algorithm that determines an orthogonal transformation to a coordinate system where the variance among the data points is maximized along each (new) coordinate axis (Jeong, Ziemkiewicz, Fisher, Ribarsky & Chang). These axes are often called principle components. Therefore, PCA is the transformation method. This transformation helps us to dimension reduction without losing any information.

PCA solves the eigenvalue problem of varicance-covariance matrix (Σ) of the data set.

$$\Sigma v = \lambda v \tag{2.1}$$

The first principle component is the eigenvector $\vec{v}$ which corresponds to the maximum eigenvalue of Σ. The second principle component is the eigenvector which corresponds to the subsequent biggest eigenvalue. This process carries on up to the least non-zero eigenvalue.

The maximum eigenvalue means the greatest statistical variability in the data points. And the second one means the second greatest statistical variability in the data set.

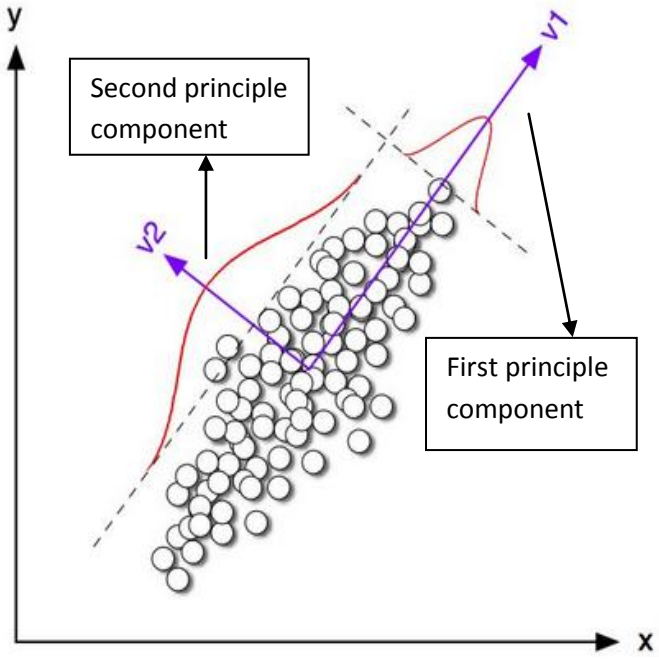


Figure 2.1 Graphical representation of principle components

The data is centered by subtracting the mean value. Indeed, since the aritmetic mean vector of the data is:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{2.2}$$

Then we have the new data set:

$$\tilde{x} = x_i - \bar{x} \tag{2.3}$$

$$\sum_{i=1}^{n} \tilde{x}_i = \sum_{i=1}^{n} \left(x_i - \bar{x} \right) = \sum_{i=1}^{n} x_i - n\bar{x} = 0 \tag{2.4}$$

Let $\tilde{X} = \left[\tilde{x}_1 \quad \tilde{x}_2 \quad \ldots \quad \tilde{x}_n \right]$ be the $p \times n$ matrix with p-dimensional column vectors $\tilde{x}_i$.

$$\tilde{X} = \left[\tilde{x}_1 \quad \tilde{x}_2 \quad \ldots \quad \tilde{x}_n \right] \tag{2.5}$$

Its variance-covariance matrix can be computed by the following formula:

$$\operatorname{cov}\left(\tilde{x}_1^{\,j} \quad \tilde{x}_2^{\,j} \quad \ldots \quad \tilde{x}_n^{\,j}\right) = \tilde{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \ldots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{n1}^2 & \ldots & \ldots & \sigma_{nn}^2 \end{bmatrix}_{nxn} \tag{2.6}$$

The principle components are the eigenvectors of $\tilde{\Sigma}$;

$$\tilde{\Sigma}v = \lambda v \Rightarrow \left(\lambda I_n - \tilde{\Sigma}\right)\vec{v} = 0 \tag{2.7}$$

$$\det(\lambda I - \tilde{\Sigma}) = 0 \tag{2.8}$$

By the definition of $\tilde{\Sigma}$, it is clear that it is a real symetric matrix. Then by the spectral theorem, $\tilde{\Sigma}$ is a diagonalizable matrix and all of its eigenvalues are real. Assume, λ_i represents the i th largest non-zero eigenvalue of $\tilde{\Sigma}$.

$$\lambda_1 > \lambda_2 > \ldots > \lambda_m$$

Since the covariance matrix is positive definite, then all of its eigenvalues are positive. Let $\vec{v}_i$ be the corresponding eigenvector to λ_i where $i = 1,2,\ldots,m$:

$$V = \left[v_1, v_2, \ldots, v_m\right] \tag{2.9}$$

We can transform the data by the following formula:

$$\varphi = \tilde{x}V^T \tag{2.10}$$

PCA reduces the data sets dimensions by decreasing the number of features. Thus, the PCA transformation is not aim to classify the data set, it aims to classify the features (Malhi & Gao, 2004).

2.2 Kernel Principle Component Analysis

Kernel PCA was developed for non-linear feature extraction (kernelpca) has similar to 'PCA' algorithm. The only difference is the transformation between input space and the new feature space is non-linear (Scholköpf et al., 1998). Non-linear

transformation can be costly for the purposes of memory usage. Because, in Kernel PCA, data mapped into a high-dimensional feature space F and PCA computed in the feature space F. The use of kernel ‘trick’ solves the memory problem by using ‘dot product’ (Schölkopf et al., 1998). ‘Kernel’ trick has been discussed in detail in the section 3.4.1.

CHAPTER THREE
DISCRIMINANT ANALYSIS

3.1 Fisher Linear Discriminant Analysis

Fisher's LDA is one of the most popular techniques for pattern recognition and dimension reduction (Yang, Zhang, Yong & Yang, 2005). Fisher's LDA maximizes the between-class scatter matrix while minimizing the within-class scatter matrix in any particular data set. In this manner, it tries to find best discriminant function to represent each class.

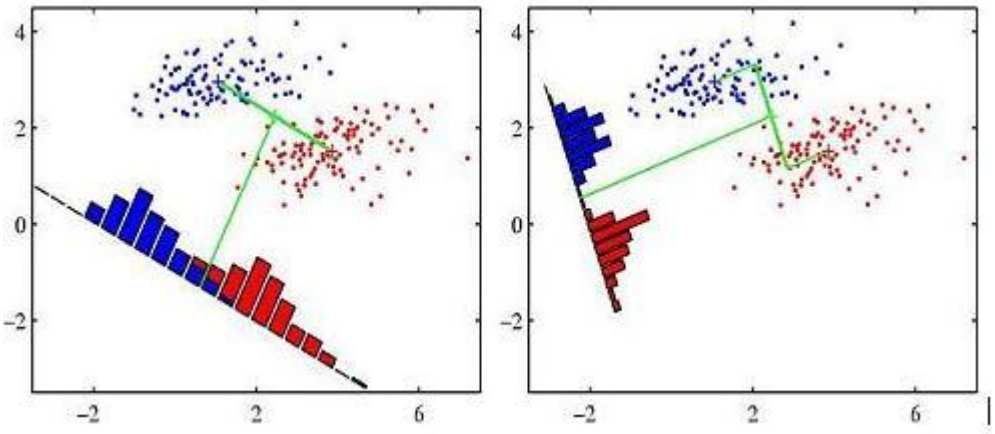


Figure 3.1 Illustration of Fisher's linear discriminant analysis

Fisher's LDA is almost similar to PCA. The fundamental difference between Fisher's LDA and PCA is that PCA is mostly involved in feature classification and Fisher's LDA is mostly involved in data classification. The other difference between Fisher's LDA and PCA is that, the location of original data set changes in PCA when data is transformed to a different space whereas in Fisher's LDA, the location of original data set doesn't change.

Fisher linear discriminant analysis has some assumptions (Alpar, 2011)

- The independent variables are normally distributed
- The equality of covariance matrices for classes
- There is no multi-related problem between independent variables

- There are no outlier values in the data set.

One of the most important step about Fisher's LDA is the selection of appropriate variables to represent the classes. Selection of variables is of great importance because discriminant analysis depends on discriminotary power of variables. So, if the selected variables represent the classes in the best way, discriminant analysis will give good results. Wilk's lambda test statistic is used for measuring the discriminotary power of the variables.

3.1.1 Wilk's Lambda Test Statistic

The Wilk's lambda test statistic determines the discriminatory power of variables. It means that the Wilk's lambda finds which variables maximize the between class variance and minimize the within class variance. It computes the ratio of within class scatter matrix and total variance scatter matrix. It is expected that this ratio approaches to zero. This means, there are differences between classes and as a result the selected variables can discriminate the groups. The Wilk's lambda test statistic can be tested by using F or Chi-square distribution (Alpar, 2011).

$$T = B + W \tag{3.1}$$

where;

$T =$ Sum of squares

$B =$ Sum of between class squares

$W =$ Sum of with-in class squares

$$B = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \tag{3.2}$$

$$W = \sum_{i=1}^{k} S_i (n_i - 1) \tag{3.3}$$

where;

$k =$ Number of groups

$\bar{x}_i =$ The mean vector of i^{th} group

$\bar{x} =$ The mean vector of whole data

$n_i =$ The number of observations of i^{th} group

$S_i =$ The variance-covariance matrix of i^{th} group

The wilk's lambda test (Λ) statistic can be computed by the following formula:

$$\Lambda = \frac{|W|}{|B+W|} \tag{3.4}$$

3.1.2 Computing the Fisher Linear Discriminant Analysis Coefficients

Different classes have the different discriminant functions in the Fisher's LDA.

Suppose that we have *k* groups and *p* variables and the Fisher's discriminant function was given the following equation for group *j*;

$$C_j = c_{j0} + c_{j1}x_1 + c_{j2}x_2 + ... + c_{jp}x_p \tag{3.5}$$

where

$C_j =$ The discriminant score of class j

$c_{jp} =$ The discriminant coefficients

$x_j =$ Indepentent variables

$c_{j0} =$ Constant

The coefficients of classification functions is found as the following formula;

$$C_i = S_{Wg}^{-1} \cdot \bar{x}_i \tag{3.6}$$

$$S_W = \sum_{i=1}^{k} (n_i - 1) S_i \tag{3.7}$$

S_W is the within class scatter matrix where S_i is variance-covariance matrix of class *j*.

$$S_{Wg} = \frac{\sum_{i=1}^{k}(n_i - 1)S_i}{\sum_{i=1}^{k}(n_i - 1)} = \frac{S_W}{(n-k)} \tag{3.8}$$

S_{Wg} is the common variance-covariance matrix.

$$c_{i0} = -\frac{1}{2} C_i \cdot \bar{x}_i \tag{3.9}$$

c_{i0} is the constant of discriminant function of class *j*.

As a result, the discriminant analysis scores can be computed as the following formula:

$$C_i = C_{0i} + \sum_{i=1}^{p} C_{ji}.x_i \tag{3.10}$$

Table 3.1 Example of discriminant scores table

Observation	Class 1	Class 2
$\lfloor x_1 \quad ... \quad x_p \rfloor$	Score 1	Score 2

The new observation discriminant scores are computed by according to the discriminant function of each class. Thus, if Score 1>Score 2, the observation will be assigned to class 1, otherwise the observation will be assigned to class 2. However, it must be controlled that the classification process is performed correctly or not. In this thesis, two validity techniques are investigated:

1. Resubstitution Method
2. Jack-Knife Method

3.2 Validity of Discriminant Analysis

3.2.1 Resubstitution Method

In resubstitution method, the classification model is computed by using all data and validity is tested on the same data by using the model. Resubstitution method

may not be good indicator for the future performance since the training and testing data are the same, the learning proces fails because of the memorizing of the classifier and eventually gives optimistic results.

3.2.2 Jack-Knife Method

The jack-knife or “leave one out” technique is a cross-validation technique. It is a iterative process. The first discriminant function is estimated from the whole data and then each element is removed from the data set respectively. The discriminant function is calculated from the remaining data. This calculation is called a *partial estimate* (Abdi & Williams, 2010). The function value of the removed data is obtained from the calculated discriminant function. This iteration continues for the whole data. The classes are expected for each removed observation. The jackknife technique is superior comparing the traditional techniques for assesing the external validity of statistical results of discriminant analysis (Daniel, 1989).

3.3 Fuzzy Linear Discriminant Analysis

Fisher’s LDA is one of the most popular classification technique for data classification. However, it can not always do successfully classification because of its mathematical substructure. Fisher’s LDA seperates the data sets along an axis and it may give inappropriate results even if the data sets have just a bit overlapped points (Chen et al., 1999). So as to improve the efficiency of Fisher’s LDA, some methods should be used to reduce the effect of overlapped points in the data set. Fuzzy set theory may be helpful to solve the such problem.

3.3.1 Fuzzy Set Theory

Fuzzy sets was introduced by Zadeh in 1965. Some of application fields of this theory are in artificial intelligence, computer science, decision theory, operations research and pattern recognition.

In crisp set theory, objects can not belong to more than one class. In other words, it has crisp boundaries (Chen et al., 1999). Each object either belong to or not belong to the set (Zimmermann, 2001). In crisp set theory, all elements are enumarated and countable:

$$A = \{a_1, \quad a_2, \quad a_3, \quad ..., \quad a_n\}$$

If *A* is a subset of the universal set *X*, the set of *A* can be symbolized as $\mu_A(X)$ and it can be written by the following definition:

$$\mu_A(X) = \begin{cases} 1 & if \quad a_n \in A \\ 0 & if \quad a_n \notin A \end{cases}$$

The set containing 0 and 1. This can be generalized by the following function (Klir & Folger, 1988):

$$\mu_A : X \mapsto \{0,1\}$$

This function can be generalized for the each element of universal set and is called as the membership degree of these elements in the set (Hong & Lee, 1995). Thus, A is a crisp set and elements of A can not belong to another set. We can discriminate all elements if it is element of A or not. However, this is not the case in daily life. We can not always reach definite conclusions or only numerical value in our measurements. Mostly, we have to specify a range. Values in the range that we specify are not always belong to a class. Therefore, the fuzzy sets are useful analysis techniques in such cases. In fuzzy sets, each element have degree of membership between [0,1] and each element can belongs to more than one class. In this situation, the defined function is called as 'membership function' and the set is called as 'fuzzy set'.

Let X be a universal set. The membership function is always defined by the following definition:

$$\mu : X \to [0,1]$$

According to the definition, in fuzzy sets, elements have degrees of membership between [0,1]. We can also integrate "fuzzy" to the classification techniques. The following section is related to one of the fuzzy clustering technique.

3.3.2 Fuzzy C-means

Fuzzy c-means (FCM) is one of the most commonly used clustering method. It assigns degrees of membership to each observation by using fuzzy memberships. FCM is a clustring method which allows to data can belong to more than one cluster. This method developed by Dunn, in 1973. After, It was improved by Bezdek in 1981. The FCM algorithm assigns degrees of membership to each observation. Each point has different degree of membership to each class. For instance, in crisp set theory, objects just belong to one cluster as {0,1} however, in fuzzy set theory, objects have degrees of membership between [0,1] (Kandemir-Cavas & Nasibov, 2012).

The following steps give us the FCM algorthm:

i. Assign values to $w_{i,j}$ (degrees of membership)

$$w_{i,j} = \frac{\left(1 \Big/ dist(x_i, c_j)^2 \right)^{1/p-1}}{\sum_{q=1}^{k} \left(1 / dist(x_i, c_q)^2 \right)^{1/p-1}} \tag{3.11}$$

ii. Repeat

iii. Compute the centroid of each cluster

$$c_j = \frac{\sum_{i=1}^{n} w_{i,j}{}^{p} x_i}{\sum_{i=1}^{n} w_{i,j}{}^{p}} \tag{3.12}$$

iv. Update $w_{i,j}$

v. If the centroids don't change, stop algorithm.

FLDA is computed using the degrees of membership.

3.3.3 Computing the Fuzzy Linear Discriminant Analysis Coefficients

The crisp DA method is known to compute under same assumptions such as normality, equality of variance-covariance matrix, class distribution etc (Pop &

Sarbu, 2013). In addition, in ordinary DA method, overlapped points must belong to the same classes because it uses the Euclidean Distance to compute the coefficients of discriminant function. However, in FLDA, observations can belong to more than one cluster. FLDA has the same aim with Fisher's LDA method. It aims to maximize the ratio of between class variance and within class variance and tries to find best discriminating function to represent the classes. However, there is a basic difference between two discriminant analysis methods. Fisher's LDA assigns the overlapped points to the same classes. Whereas, in FLDA the overlapped points can not belong to the same classes.

FLDA is computed by using the degrees of membership.
Fuzzy amount of element in fuzzy group is computed as,

$$N(w_i) = \sum_{j=1}^{n} \mu_{w_i}(x_j) \tag{3.13}$$

and the fuzzy total amount of element in fuzzy groups is as follows,

$$N_T = \sum_{i=1}^{m} N(w_i) \tag{3.14}$$

In addition, the fuzzy mean vector of each class is defined as,

$$M(w_i) = \frac{\sum_{j=1}^{n} \mu_{w_i}(x_j) X_j}{N(w_i)} \tag{3.15}$$

and the fuzzy mean vector of whole data set is computed as,

$$M = \frac{\left(\sum_{i=1}^{m} \sum_{j=1}^{n} \mu_{w_i}(x_j) X_j \right)}{N_T} \tag{3.16}$$

Then, the total fuzzy within class covariance matrix and the fuzzy between-class covariance matrix is given by

$$S_T = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_j - M(w_i)\right)\times\left(X_j - M(w_i)\right)^T \times \mu_{w_i}(x_j) \tag{3.17}$$

and,

$$S_B = \left(\sum_{i=1}^{m} N(w_i)\times\left(M(w_i) - M\right)\times\left(M(w_i) - M\right)^T\right) \tag{3.18}$$

respectively.

The basic problem is to find a matrix which separates the classes well and at the same time maximizes the following ratio.

$$\eta = \frac{A^T S_B A}{A^T S_T A} \tag{3.19}$$

This is also eigenvalue problem. *A* is an eigenvector which is corresponding the maximum eigenvalue. The vector of *A* gives us coefficients of discriminant function.

$$A^* = [A_1^* \mid A_2^* \mid \ldots \mid A_{c-1}^*] = \arg\max\left\{\left|\frac{A' S_B A}{A' S_T A}\right|\right\} \Rightarrow (S_B - \lambda_i S_T)A_i = 0 \tag{3.20}$$

$$Y_i^* = x_i^{'} \cdot A^* \tag{3.21}$$

Y_i^* is the vector of discriminant scores. Discriminant scores are computed for each class using the obtained fuzzy discriminant function. It is expected that each discriminant scores for the classes must be different from each other. For example, discriminant scores of class 1 must be greater than class 2 or discriminant scores of class 2 must be greater than class 1. Thus, classes can be seperated each other according to their discriminant scores.

3.4 Kernel Discriminant Analysis

Non-linear problems often encountered in daily life. Kernel discriminant analysis (KDA) is a non-linear discriminant analysis technique in pattern recognition. The kernel method was introduced to solve non-linear problems and it is used to increase the efficiency of the linear discriminant analysis (Peng et al., 2013). As it is mentioned in section 3.1, Fisher's LDA has some assumptions such as data set have normal distribution, the variance-covariance matrixes must be equal...etc. On the other hand, kernel method does not have the aforementioned assumptions. The distribution of the data set is not an important asuumption in the Kernel method. KDA can be applied on either linear or non-linear data set.

In KDA, data is mapped into the feature space F and DA is computed in the feature space F. The input data mapped into a high-dimensional feature space F.

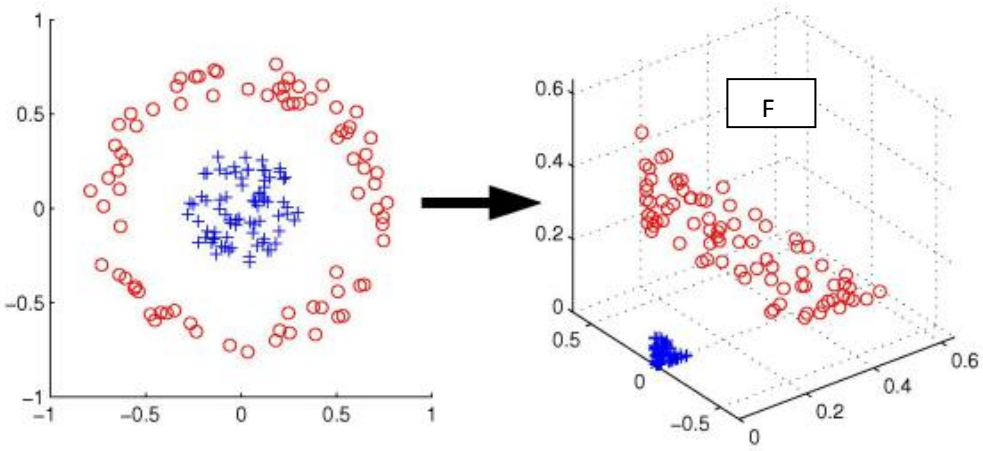


Figure 3.2 Visual scheme of Kernel trick

The mapping function Φ can be denoted as:

$$x = (x_1, x_2, ..., x_n) \mapsto \Phi(x) = (\Phi_1(x), ..., \Phi_j(x), ...)$$
$$\langle x_i . x_j \rangle \rightarrow \langle \Phi(x_i) . \Phi(x_j) \rangle = k(x_i, x_j) \quad (3.22)$$

where x is the set of inputs and $k(x_i, x_j)$ is the kernel matrix of x.

KDA uses ‘Kernel Trick’ .

3.4.1 Kernel Trick

Kernel trick provides a bridge from linearity to non-linearity as seen in Figure 3.3. ‘ϕ’ is a mapping function from data to feature space F.

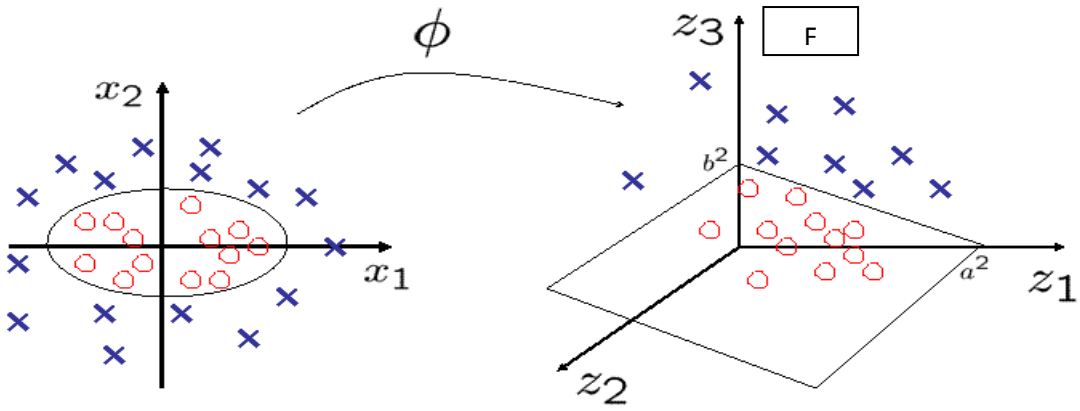


Figure 3.3 Visual shape of mapping function

If the data can not be seperated linearly, it can be transformed high-dimensional feature space. After the transformation, data can be seperated in high-dimensional feature space F, as seen in Figure 3.3. This situation is called ‘Kernel trick’. ‘Kernel trick’ is a mathematical tool which depends on the ‘dot product’ between two vectors. If ‘dot product’ used, ‘Kernel’ function used for the transformation. This means, linear algorithm is transformed into non-linear algorithm.

Kernel functions are continous, symmetric and positive-definite functions. For this reason, kernel matrices have not negative eigenvalue. It can be used so many kernel functions in analysis and each kernel function has different properties. Because of this reason, selection of kernel function is the very significant step in the analysis. Different kernel functions produce different results.

3.4.2 Kernel Functions

Some important kernel functions are given below.

$k(x, y)$ refers to the kernel function where $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\}$ are the inputs data.

3.4.2.1 Linear Kernel

Linear kernel is the most basic kernel function. Linear kernel function gives the similar results with the application not applied kernel transformation. The linear kernel function given below:

$$k(x, y) = x^T y + c \tag{3.23}$$

where *x* and *y* are the input values of the data.

3.4.2.2 Polynomial Kernel

Polinomial kernel is a widely used method in kernelized models. Polynomial kernel is convenient for normalized data and it is given as below.

$$k(x, y) = (\alpha \cdot x^T y + c)^d \tag{3.24}$$

where *x* and *y* are the input values of the data and $k(x, y)$ is the kernel matrix of the input values.

3.4.2.3 Gaussian Kernel

Gaussian kernel is the most commonly used method in pattern recognition. In this study, we used 'Gaussian Kernel' function for the data transformation. The 'Gaussian Kernel' function is given below:

$$k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{3.25}$$

where *x* and *y* are the input values of the data, $k(x, y)$ is the kernel matix of input values and σ is the parameter which depends on researchers.

The parameter "σ" plays an important role in kernel performance. It should be attentively chosen for problem. The parameter σ controls the flexibility of the kernel (Shawe-Taylor & Cristianini, 2004). If the parameter "σ" is selected extremely large, the transformation will lose its non-linearity power. On the other hand, if the parameter "σ" selected fairly low, the transformation lose its regularity.

The parameter "σ" determined by the analyst. The value of "σ" never be 0. It always selected as positive, i.e., $\sigma > 0$.

According to the 'Gaussian kernel' formula (3.25);

$k(x, x) = \exp(0) = 1$

obtained by using the formula (Shawe-Taylor & Cristianini, 2004).

3.4.2.4 Anova Kernel

Anova kernel is a radial basis function as 'Gaussian' kernel function. It has a good perform on multidimensional regression problems (Holfmann et al., 2013). The Anova kernel function given as below formula:

$$k(x, y) = \sum_{k=1}^{n} \exp\left(-\sigma\left(x^k - y^k\right)^2\right)^d \tag{3.26}$$

where *x* and *y* are the input data and "σ" is the parameter which is determined by the researchers.

3.4.2.5 Exponential Kernel

The exponential kernel related to the Gaussian kernel and it is radial basis kernel function. The exponential kernel function is given below formula:

$$k(x, y) = \exp\left(-\frac{\|x - y\|}{2\sigma^2}\right) \tag{3.27}$$

where *x* and *y* are the input data and "σ" is the parameter which is determined by the researchers.

3.4.2.6 Laplacian Kernel

The laplacian kernel is exactly same with the exponential kernel. The only difference between laplacian and exponential kernel is sensitivity. The laplacian kernel less sensitive for changes in the sigma parameter and it is given by the following formula:

$$k(x, y) = \exp\left(-\frac{\|x - y\|}{\sigma}\right) \tag{3.28}$$

where *x* and *y* are the input data and "σ" is the parameter which is determined by the researchers.

3.4.3 Computing the Coefficients of Kernel Discriminant Function

It was mentioned before that in KDA, data mapped into non-linear feature space F and discriminant analysis is computed on the transformed data. As a first step, data is transformed into non-linear feature space F by using the Gaussian kernel function.

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \tag{3.29}$$

$$x = (x_1, x_2, ..., x_n) \mapsto \Phi(x) = (\Phi_1(x), ..., \Phi_j(x), ...)$$
$$\langle x_i . x_j \rangle \mapsto \langle \Phi(x_i).\Phi(x_j) \rangle = k(x_i, x_j)$$

This transformation Φ will allow to compute the discriminant analysis in the feature space *F*.

The mean vector of class i is,

$$m_i^{\Phi} = \frac{1}{l_i} \sum_{j=1}^{l_i} \Phi(x_j^i) \tag{3.30}$$

and the between-class scatter matrix (Zhu & Song, 2010);

$$S_B^{\Phi} = (m_1^{\Phi} - m_2^{\Phi})^T (m_1^{\Phi} - m_2^{\Phi}) \tag{3.31}$$

Similarly, the within-class scatter matrix is given by;

$$S_W^{\Phi} = \sum_{c=1}^{C} K_c (I - L_c) K_c^T \tag{3.32}$$

where K_c is the kernel matrix belongs to class c, l_c is the number of samples class c and L_c is the matrix that is dimensions $l_c x l_c$ and its elements $1/l_c$. I is the idendity matrix, $l_c x l_c$ (Wang et al., 2009).

KDA has the same goal as Fisher's LDA analysis and FLDA. KDA tries to find best discriminant function which respresents to classes. It maximizes the ratio of between-class scatter matrix and within-class scatter matrix as seen following formula:

$$J(w) = \frac{w^T S_B^{\Phi} w}{w^T S_W^{\Phi} w} \tag{3.33}$$

This is also an eigenvalue problem. Here, "w" is an eigenvector corresponding to the maximum eigenvalue. The vector of "w" gives us the coefficients of discriminant function of KDA. Discriminant scores are computed for each class using the obtained kernel discriminant function. It is expected that each discriminant scores for the classes must be different from each other.

CHAPTER FOUR
BIOINFORMATICS

Bioinformatics formed by the combination of biology and computational science such as statistics, applied mathematics, informatics and computer science. It finds solutions for biological problems via the aforementioned scientific fields.

Quantitative tools are essential in modern biology. Bioinformatics is the field of quantitative analysis of information regarding to biological macromolecules with the help of computers (Nasibov & Kandemir-Cavas, 2009). Moleculer biology and computer scince provided bioinformatics to develop as a field. Thus, bioinformatics is a multidiscipliner area.

Bioinformatics involves the technology that uses computers for storage, retrieval, manipulation and distribution of information related to biological macromolecules such as DNA, RNA and proteins (Xiong, 2006).

The basic goal of bioinformatics is to better understand living cells and to predict its moleculer functions (Nasibov & Kandemir-Cavas, 2008; 2011). For this reason, bioinformatics research can generate new visions about cells.

Bioinformatics consist of three basic subfields such as sequence analysis, structure analysis and functional analysis.

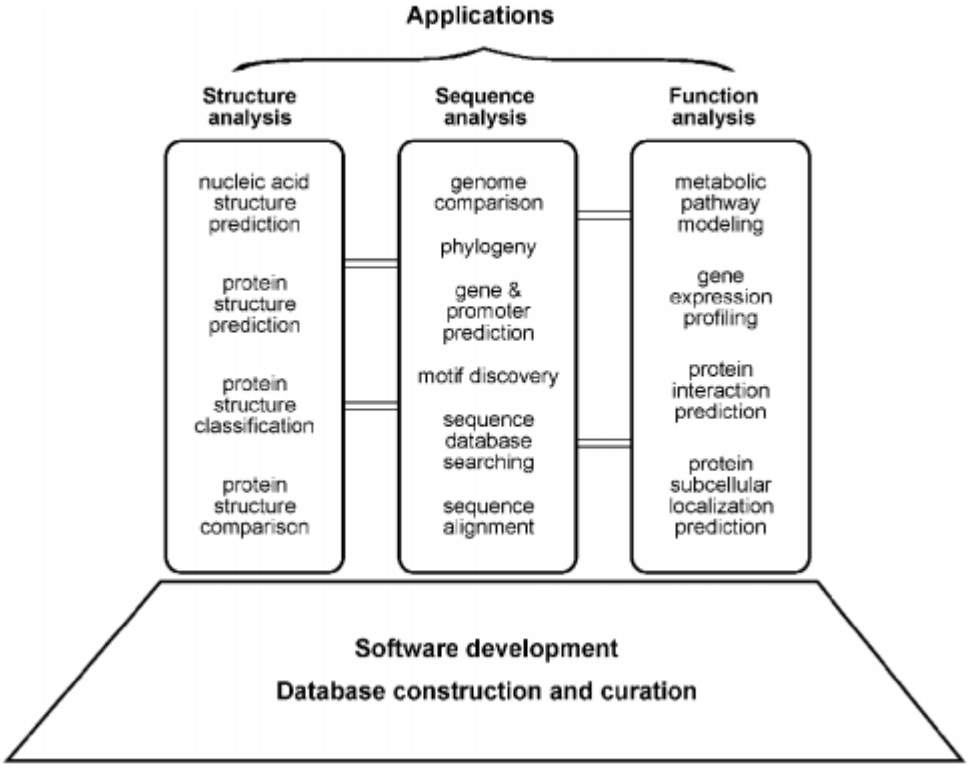


Figure 4.1 The diagram of the contents of the bioinformatics (Xiong, 2006)

According to the three basic subfields of bioinformatics, the scope of bioinformatics can be sorted as follows:

- Sequence analysis
- Computational evolutionary biology
- Measuring biodiversity
- Analysis of gene expression
- Analysis of protein expression
- Modelling biological systems
- Finding similar sequences in the database with a query sequence.
- Finding similar protein structures in the database with a query protein structure.
- Finding sequences in the databank that correspond to similar structures with a query protein structure (Lesk, 2005).

Proteins are large moleculer structures and they are formed by the combination of aminoacids. There are 20 different types of aminoacids as seen in the following figure:

G – Glycine – Gly	T – Threonine –Thr
A – Alanine – Ala	N – Asparagine – Asn
P – Proline – Pro	Q – Glutamine – Glu
V – Valine – Val	H – Histidine – His
I – Isoleucine – Ile	Y – Tyrosine – Tyr
L – Leucine – Leu	W – Trytophan – Trp
F – Phenylalanine – Phe	D – Aspartic acid – Asp
M – Methionine – Met	E – Glutamic acid – Glu
S – Serine – Ser	K – Lysine – Lys
C – Cysteine – Cys	R – Arginine – Arg

Figure 4.2 List of amino-acids

Protein sequences are composed by assembled amino acids. This is also called by *primary structure*. In bioinformatics, these representation is called FASTA format. FASTA format is the general format in the world for protein data.

The following figure is an example of protein sequences.

```
>Q9NQX6|ZN331_HUMAN Zinc finger protein 331 - Homo sapiens (Human).
MAQGLVTFADVAIDFSQEEWACLNSAQRDLYWDVMLENYSNLVSLDLESAYENKSLPTEK
NIHEIRASKRNSDRRSKSLGRNWICEGTLERPQRSRGRYVNQMIINYVKRPATREGTPPR
THQRHHKENSFECKDCGKAFSRGYQLSQHQKIHTGEKPYECKECKKAFRWGNQLTQHQKI
HTGEKPYECKDCGKAFRWGSSLVIHKRIHTGEKPYECKDCGKAFRRGDELTQHQRFHTGE
KDYECKDCGKTFSRVYKLIQHKRIHSGEKPYECKDCGKAFICGSSLIQHKRIHTGEKPYE
CQECGKAFTRVNYLTQHQKIHTGEKPHECKECGKAFRWGSSLVKHERIHTGEKPYKCTEC
GKAFNCGYHLTQHERIHTGETPYKCKECGKAFIYGSSLVKHERIHTGVKPYGCTECGKSF
SHGHQLTQHQKTHSGAKSYECKECGKACNHLNHLREHQRIHNS
```

Figure 4.3 The FASTA format of a protein sequence

CHAPTER FIVE
DATA SET AND EXPERIMENTAL RESULTS

This chapter gives the general information about data set and its obtaining process.

5.1 Data Set

The purpose of this thesis is to classify the ordered and disordered proteins. Classification of ordered-disordered protein is a very challenging problem in bioinformatics because disordered proteins is related with many diseases (Babu, Lee, Groot & Gsponer, 2011).

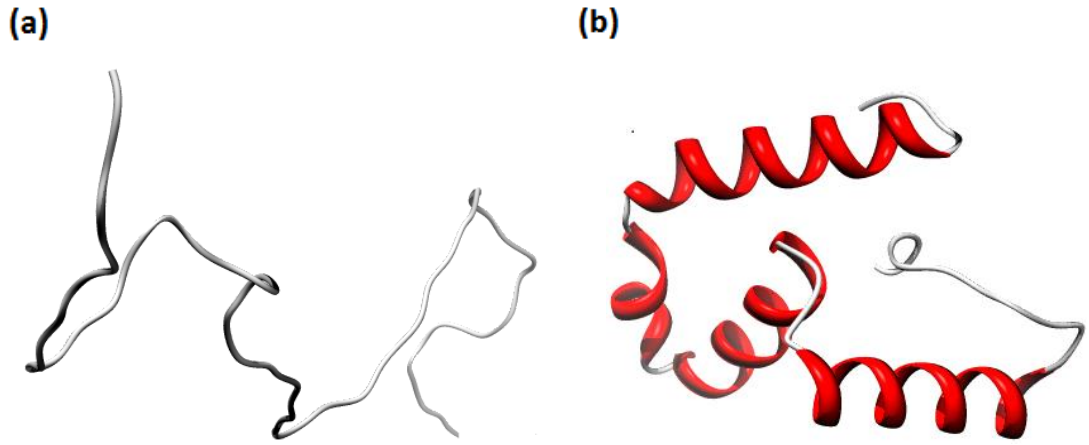


Figure 5.1 Basic figure of (a) disordered; (b) ordered proteins

Primary structure is composed by amino acid sequence, secondary structure is called by the assignment of helices and sheets, tertiary structure is the assembly and interactions of the helices and sheets, quaternary structure is the assembly of the proteins composed of more than one subunit (Mika, 2003; Desk, 2001). Intrinsically disordered proteins (IDPs) formed as a result of deformations in the tertiary strucure of a protein (Dunker et al, 2001). IDPs are frequently seen in macromolecular structures and they can be helped to recognize complex regulatory processes (Barbar & Nyarko, 2014).

The sequences of the disordered and ordered proteins were extracted from Disprot (Vucetic, 2005) and Expasy (www.expasy.ch), respectively. Protein data set obtained as “FASTA format” from protein data bank.

The protein dataset are transformed its amino acid composition by:

$$F(x) = \left[f_1(x), f_2(x), f_3(x), \ldots, f_{20}(x) \right] \tag{5.1}$$

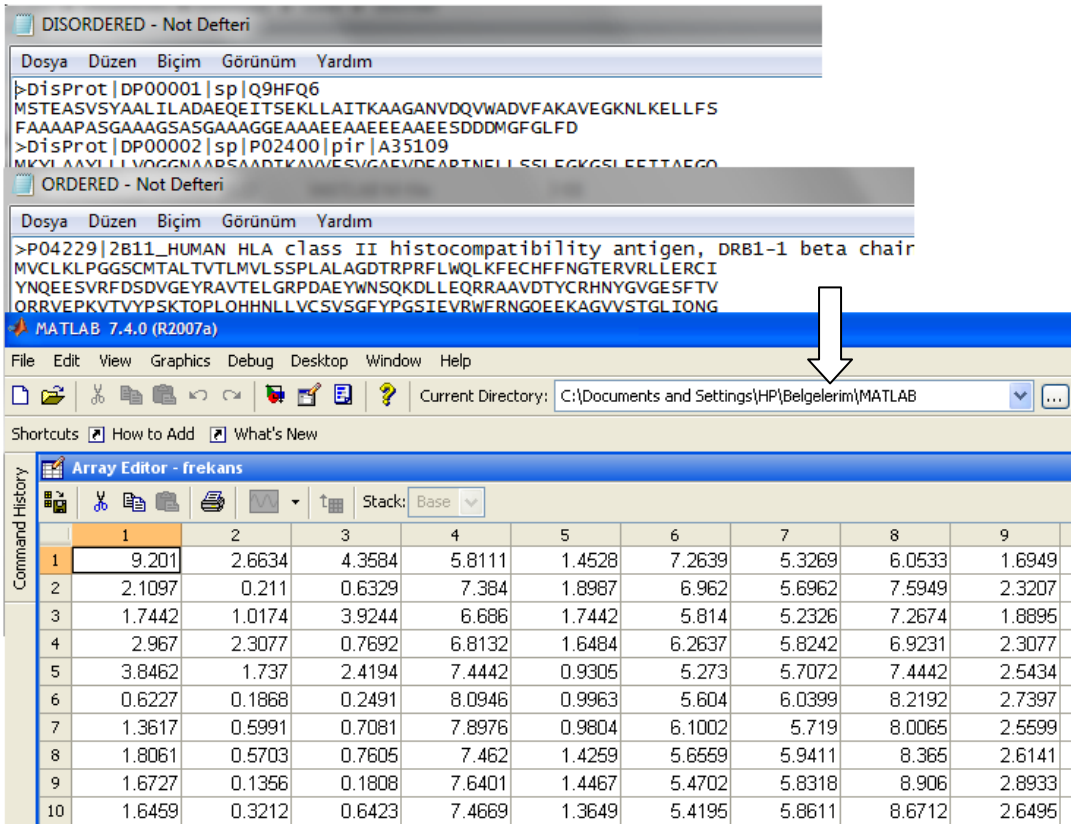
where $F(x)$ is the vector contained each frequency of 20 amino acid types. Therefore, the amino acid frequencies were calculated as follows. The percentage of the amino acid residue *i* in a protein is defined by:

$$f_i(x) = 100 \times \frac{n_i}{N} \qquad i = 1,2,\ldots,20 \tag{5.2}$$

where n_i is the frequency of amino acid *i* and *N* is the number of amino acid residues in the protein *x* (Cedano et al., 1997).

After the transformation process (see Figure 5.2), the data matrix, which has twenty columns, was obtained. The matrix has twenty columns because there are twenty different types of amino acid. Row values in each column give the percentage of aminoacids.

The obtained matrices have high dimensions and the comlexity increases in working with high-dimensional matrix. For this reason, PCA method is applied on the data set for dimension reduction.



	1	2	3	4	5	6	7	8	9
1	9.201	2.6634	4.3584	5.8111	1.4528	7.2639	5.3269	6.0533	1.6949
2	2.1097	0.211	0.6329	7.384	1.8987	6.962	5.6962	7.5949	2.3207
3	1.7442	1.0174	3.9244	6.686	1.7442	5.814	5.2326	7.2674	1.8895
4	2.967	2.3077	0.7692	6.8132	1.6484	6.2637	5.8242	6.9231	2.3077
5	3.8462	1.737	2.4194	7.4442	0.9305	5.273	5.7072	7.4442	2.5434
6	0.6227	0.1868	0.2491	8.0946	0.9963	5.604	6.0399	8.2192	2.7397
7	1.3617	0.5991	0.7081	7.8976	0.9804	6.1002	5.719	8.0065	2.5599
8	1.8061	0.5703	0.7605	7.462	1.4259	5.6559	5.9411	8.365	2.6141
9	1.6727	0.1356	0.1808	7.6401	1.4467	5.4702	5.8318	8.906	2.8933
10	1.6459	0.3212	0.6423	7.4669	1.3649	5.4195	5.8611	8.6712	2.6495

Figure 5.2 Visual representation the data set after digitized

The scree plots are given below (see Figure 5.3 and Figure 5.4). Analyzing the results, it is decided to work with two dimensions.

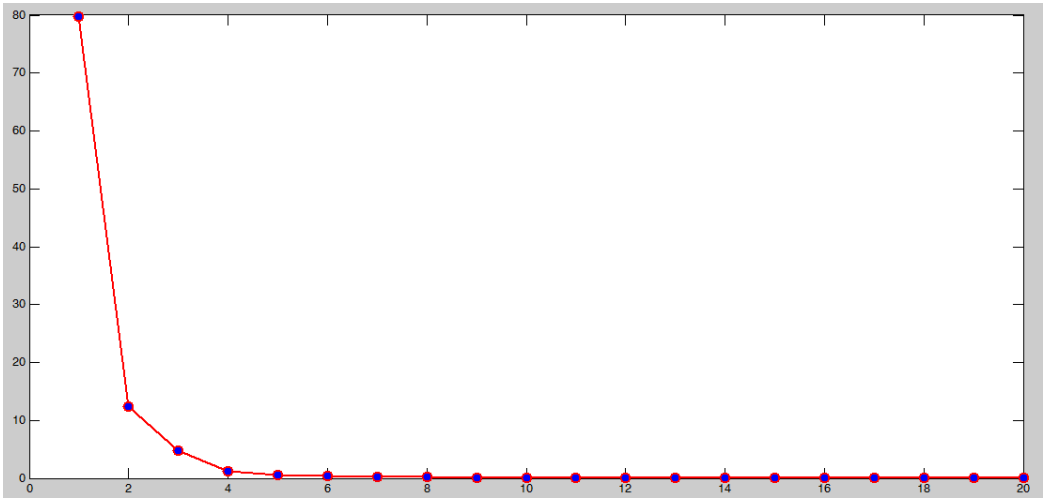


Figure 5.3 Scree plot of ordered proteins

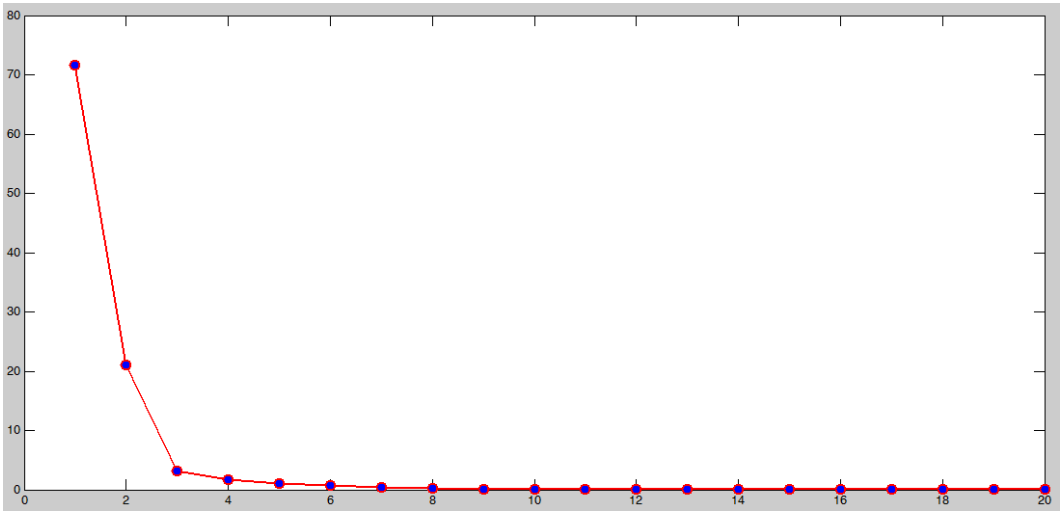


Figure 5.4 Scree plot of disordered proteins

5.2 Experimental Results

Data set is obtained from protein data bank, it is digitized and it dimensions are reducted. Three discriminant analysis methods, Fisher's LDA, FLDA and KDA are applied on protein data set and results are observed. Coefficients of discriminant functions of Fisher's LDA are given below in Table 5.1.

Table 5.1 Coefficients of Fisher linear discriminant analysis discriminant function

	Coefficients of discriminant function of ordered proteins	Coefficients of discriminant function of disordered proteins
x_1	0.2350	0.2132
x_2	0.0914	0.6323
Constant	-1.0336	-2.4891

Discriminant scores which are belonging to the Fisher's LDA are computed according to the discriminant functions given above. Scatter plots of discriminant scores are represented by Figure 5.4 and Figure 5.5. The points on the figures represent the ordered and disordered proteins discriminant scores. Circular symbols symbolize the ordered proteins discriminant scores and the star symbols symbolize disoredered proteins discriminant scores. Discriminant scores in Figure 5.4 were

obtained from the ordered proteins discriminant function as seen on the Table 5.1. For this reason, circular points must be greater than star points. However, as seen from the Figure 5.4 there are some overlapped discriminant scores. Discriminant scores in Figure 5.5 were obtained from the disordered proteins discriminant function as seen on the table 5.1. Therefore, star points must be greater than circular points. However, as seen from the Figure 5.5 there are some crossing discriminant scores. Briefly, it is clearly seen from the Figure 5.4 and Figure 5.5, there are misclassifications in Fisher's LDA. For this reason, FLDA is applied to the ordered-disordered proteins.

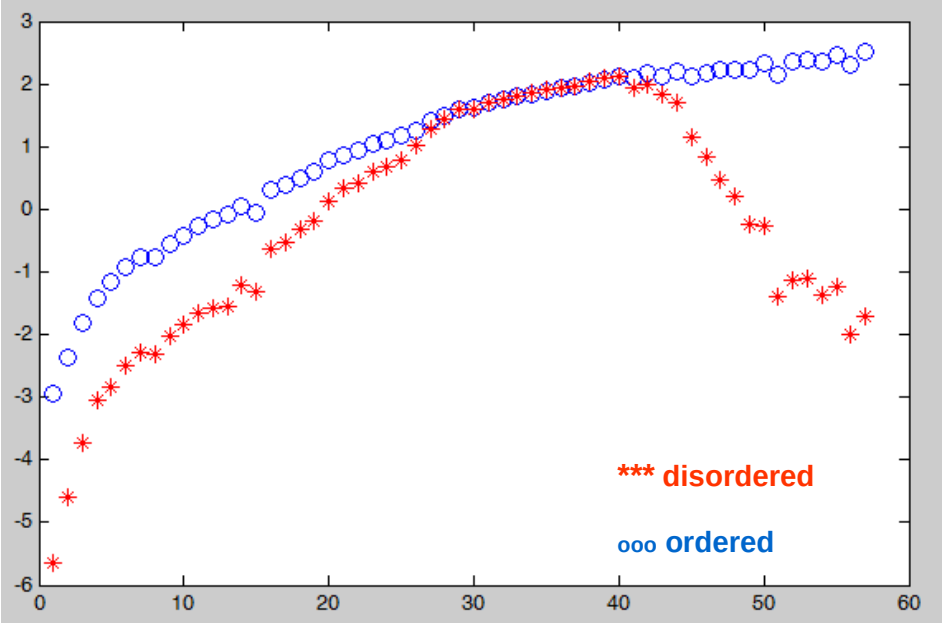


Figure 5.5 Scatter plot of discriminant scores of Fisher's LDA for the ordered proteins

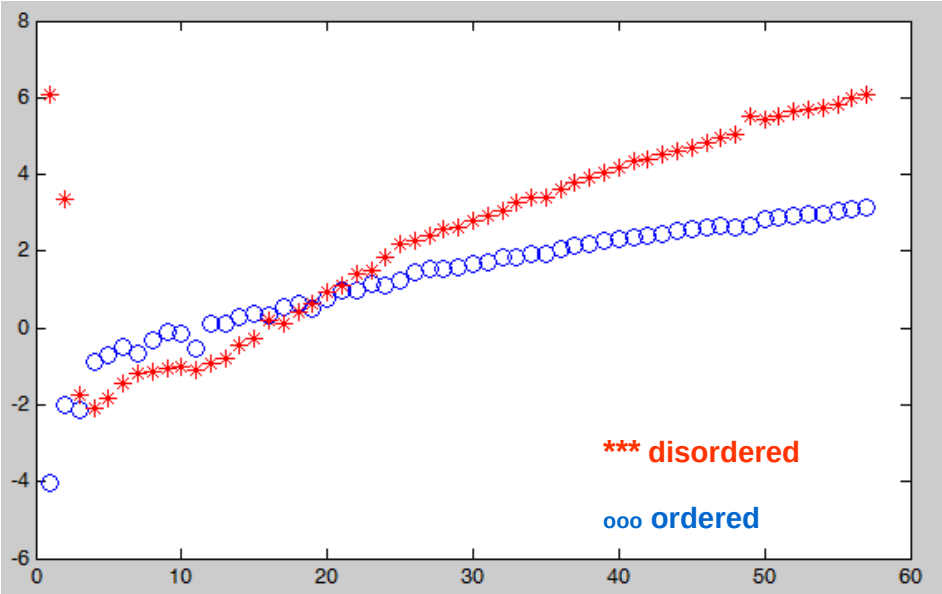


Figure 5.6 Scatter plot of discriminant scores of Fisher‘s LDA for the disordered proteins

The coefficients of discriminant function of FLDA are given the below Table 5.2. There is one discriminant function in FLDA. Discriminant scores are computed by using the obtained discriminant function.

Table 5.2 Coefficients of FLDA discriminant function

	Coefficients of fuzzy linear discriminant analysis discriminant function
x_1	0.6411
x_2	-0.7674

Figure 5.6 gives the FLDA discriminant scores. It is clearly seen from the following Figure 5.6, FLDA gives better results than Fisher’s LDA. However, there are still misclassification scores.

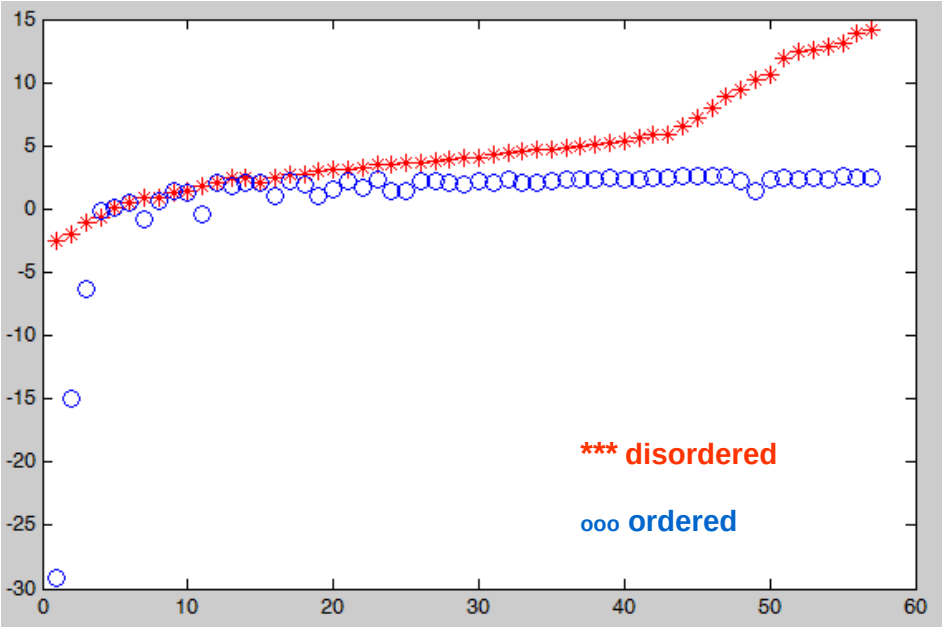


Figure 5.7 Scatter plot of discriminant scores of FLDA

FLDA has also same problem as Fisher's LDA. These discriminant analysis are good at solving the linear problem The following Figure 5.7 represents the distribution of data. It is clearly seen from the Figure 5.7, data does not normally distributed and data can not be seperated linearly. For this reason, non-linear DA techniques such as KDA is applied on ordered disordered protein data. The following Table 5.3 gives the kernel discriminant analysis coefficients. There are 57 coefficients in kernel discriminant analysis because data transformed into high-dimensional feature space by using gaussian kernel function.

Table 5.3 Coefficients of KDA discriminant function

Coefficients of Kernel Discriminant Analysis Discriminant Function			
1	0.00071199	30	-0.000495262
2	0.000641069	31	-0.075179056
3	0.000655825	32	0.046448063
4	-9.49E+09	33	0.036752933
5	0.001749375	34	-0.043517256
6	-0.003344918	35	0.216076749
7	-0.000341014	36	-0.278153774
8	0.004031257	37	0.328850539
9	0.001361174	38	-0.637363018
10	-0.004046758	39	0.546766578
11	0.000364356	40	-0.04502642
12	0.002333122	41	-0.165857211
13	0.001199789	42	0.017859984
14	-0.006446318	43	0.058553046
15	0.00307599	44	0.00022754
16	0.002023546	45	-0.024015997
17	0.00281908	46	0.00795462
18	-0.000927989	47	-0.006552292
19	-0.000404893	48	0.010867598
20	-0.006405983	49	-0.008033712
21	-0.008726186	50	0.00090919
22	0.031764955	51	0.001039273
23	-0.00178993	52	-0.003081938
24	-0.034074402	53	0.04458005
25	0.024871777	54	-0.050083311
26	-0.033826869	55	-0.009257559
27	0.081506386	56	0.019939198
28	-0.039816093	57	0.002325675
29	-0.003330383		

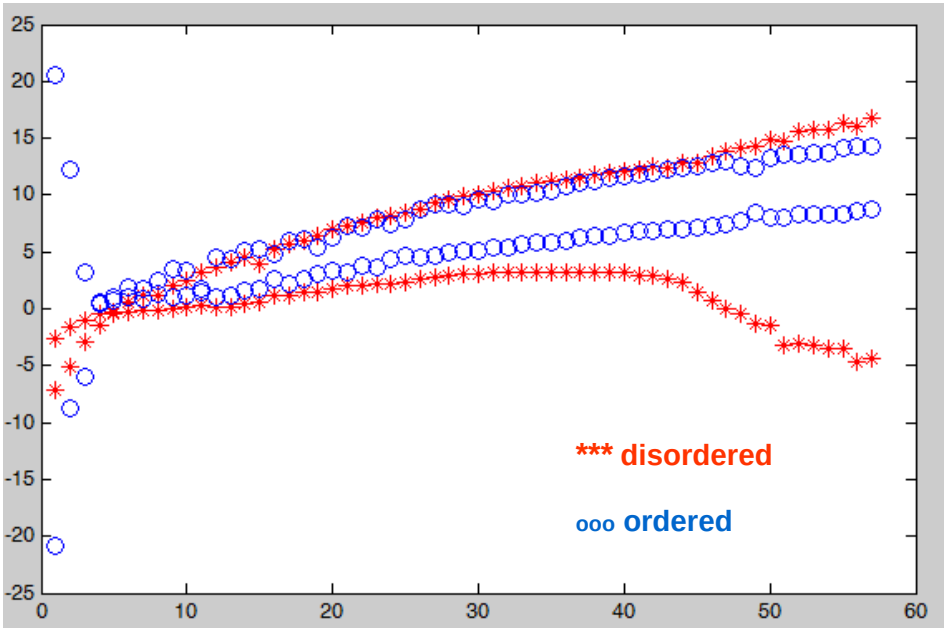


Figure 5.8 Distribution of the whole data

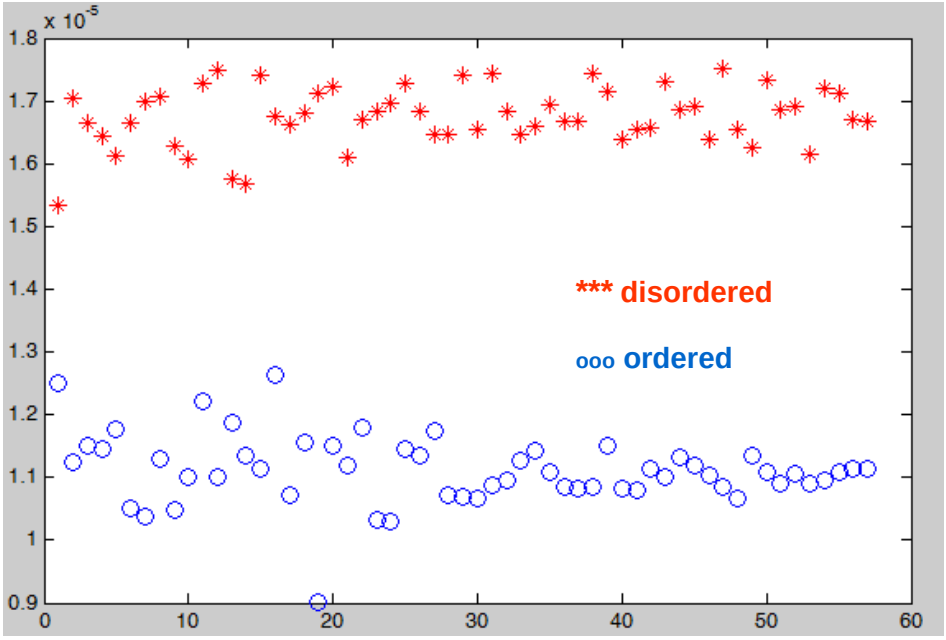


Figure 5.9 Scatter plot of discriminant scores of KDA

As obviously seen from the Figure 5.8, KDA can easily discriminate the data. Even though the data can be seperated linearly, KDA is appropriate to use. It is clearly seen from the obtained results, KDA gives best discriminating scheme for the ordered-disordered protein data compared with the two discriminant analysis method Fisher's LDA and FLDA.

As mentioned in chapter three, there are two validity techniques for discriminant analysis such as resubstitution method and jackknife method. Two validity techniques are used to measure the accuracy rate of three discriminant analysis methods. Table 5.4 gives the results of resubstitution validity technique.

Table 5.4 Table of resubstitution validity technique and accuracy rates

	Fisher Linear Discriminant Analysis		Fuzzy Linear Discriminant Analysis		Kernel Discriminant Analysis	
	True	False	True	False	True	False
Ordered	**48**	**9**	**54**	**3**	**57**	**0**
Disordered	**42**	**15**	**54**	**3**	**57**	**0**
TOTAL	**90**	**24**	**108**	**6**	**114**	**0**
Prediction Percentage	**78.95%**		**94.74%**		**100%**	

As clearly seen from the Table 5.4, the classification performance of KDA is error free. FLDA gives the better results than Fisher's LDA and KDA gives the best results by the resubstitution technique.

Ordered and disordered proteins are classified by using Fisher's LDA, FLDA and KDA. It is obviously seen that discriminating the data with linear transform is difficult. It is clearly seen from the Table 5.4, Fisher's LDA gives worse performance than FLDA and KDA according to the resubstitution method. It can be seen from the results in the Table 5.4, overall prediction rate obtained by Fisher's LDA, FLDA and KDA in identifying the disordered-ordered proteins are 78.95%, 94.74% and 100%, respectively.

Jackknife method is applied to the classification of ordered-disordered protein data and results are given the following table.

Table 5.5 Table of jack-knife validity technique and accuracy rates

	Fisher Linear Discriminant Analysis		Fuzzy Linear Discriminant Analysis		Kernel Discriminant Analysis	
	True	False	True	False	True	False
Ordered	**46**	**11**	**43**	**14**	**57**	**0**
Disordered	**42**	**15**	**57**	**0**	**57**	**0**
TOTAL	**88**	**26**	**100**	**14**	**114**	**0**
Prediction Percentage	**78.95%**		**94.74%**		**100%**	

It can be seen from the results in Table 5.5, overall prediction rate obtained by jackknife method by using the Fisher's LDA, FLDA and KDA techniques in identifying the disordered-ordered proteins are 77,1930%, 87,7193% and 100%, respectively. It is clearly seen from the accuracy rates in Table 5.5, Fisher's LDA gives worse performance than FLDA and KDA according to the Jackknife method. As a result, KDA can be a very robust method to identify whether a protein has a disordered and ordered structure considering the two validation techniques.

CHAPTER SIX
CONCLUSIONS

Pattern analysis is a data recognition process which is used in many fields. One of these fields is bioinformatics. In bioinformatics, proteins have different chemical properties and different types of protein sequences have different biochemical function. For this reason, classification of protein sequences is of great importance to determine their biochemical structures. The aim of this thesis is classify the ordered and disordered proteins by using Fisher's LDA, FLDA and KDA. Classification of ordered-disordered protein is a very complex and challenging problem in bioinformatics because disordered proteins is associated with many diseases.

The analysis part of the thesis is composed of three sections. Firstly, discriminant analysis method Fisher's linear discriminant analysis method is applied in order to classify the ordered-disordered proteins. According to the jackknife validity technique, the accuracy rate of Fisher's linear discriminant analysis is obtained as 77.1930%.

Secondly, fuzzy linear discriminant is applied as second discriminant analysis method. The accuracy rate for jackknife validity technique is obtained as 87.7193%.

Thirdly, the last discriminant analysis method in order to classify the ordered-disordered proteins is kernel discriminant analysis. Kernel discriminant analysis is a non-linear based discriminant analysis technique. The accuracy rate obtained from kernel discriminant analysis is found 100% for jackknife validity technique.

Fisher's linear discriminant analysis is not sensitive for the overlapped points compared with the fuzzy linear discriminant analysis. For this reason, fuzzy linear discriminant analysis gives better results than Fisher's linear discriminant analysis. However, Fisher's and fuzzy discriminant analysis are linear transform techniques. It is obviously seen that discriminate the data with linear transform is difficult. For this reason, impelemention of non-linear discriminant technique (Kernel) is found

approppriate. Jackknife validity technique is used to measure the performance of discriminant analysis. According to the jackknife validity techniques, Fisher's linear discriminant analysis gives worse performance than fuzzy linear and kernel discriminant analysis, and kernel discriminant analysis achieved the best accuracy rates.

It is clearly seen that kernel discriminant analysis gives the best discrimating scheme for the ordered-disordered protein data compared with the two discriminant analysis method such as Fisher's and fuzzy because of its convenient mathematical structure for linear and also nonlinear data scheme. According to the all validity results in Table 5.4 and Table 5.5, it can be strongly referred that kernel method can be a very robust method to identify whether a protein has a disordered or ordered structure.

These results have been presented in the following symposiums as oral and poster:

- 9^{th} Statistics Days Symposium 2014, May 10-14, Antalya, TURKEY
- Informatics for Health and Life Science 2014, October 19-22, Kuşadası, İzmir, TURKEY

REFERENCES

Abdi, H. & Williams, L., J. (2010). *Jackknife*. In N.J. Salkind (Ed), Encyclopedia of research design. Thousand Oaks: Sage.

Alpar, R. (2011).*Uygulamalı çok değişkenli istatistiksel yöntemler*. Ankara: Detay Yayıncılık.

Barbar, E. & Nyarko, A. (2014). Polybivalency and disordered proteins in ordering macromolecular assemblies. *Seminars in Cell & Developmental Biology, 1675*, 1-6.

Babu, M., M., Lee, R., Groot, N., S. & Gsponer, J. (2011). Intrinsically disordered proteins: regulation and disease. *Current Opinion in Structural Biology, 21*, 1–9.

Bezdek, J., Ehrlich, R. & Full, W. (1984). The fuzzy c-means clustering algorithm. *Computer & Geoscience, 10*, 191-203.

Bishop, C., M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Cedano, J., Aloy, P., Pérez-Pons, J. A. & Querol, E. (1997). Relation between amino acid composition and cellular location of proteins. *Journal of Molecular Biology, 266*, 594–600.

Chen, Z-P., Jiang, J-H., Li, Y., Liang, Y-Z. & Yu, R-Q. (1999). Fuzzy linear discriminant analysis for chemical data set. *Chemometrics Intelligent Laboratory Systems, 45*, 295-302.

Daniel, L., G. (1989). Use of the jackknife statistic to establish the external validity of discriminant analysis results. *Annual Meeting of the Southwest Educational Research Association.*

Diaf, A., Boufama, B. & Benlamri, R. (2013). Non-parametric Fisher's discriminant analysis with kernels for data classification. *Pattern Recognition Letters, 34,* 552-558.

Duda, R. O., Hart, P., E. & Stork D., G. (2001). *Pattern classification* (2nd Edition). Wiley: Wiley-Interscience Puplication.

Dunker, A. K., Lawson, J., D., Brown, C., J., Williams, R., M., Romero, P., Oh, J., S., et al. (2001). Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling, 19,* 26-59.

Hofmann, T., Schölkopf, B. & Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics, 36,* 1171–1220.

Hong, T. P. & Lee, C. Y. (1995). Induction of fuzzy rules and membership functions from training examples. *Fuzzy Sets and Systems, 84,* 33 -47.

Huang, H., Li, J. & Liu, J. (2012). Gene expression data classification based on improved semi-supervised local Fisher discriminant analysis. *Expert Systems with Application, 39,* 2314-2320.

Jeong, D. H., Ziemkiewicz, C., , Fisher, B., , Ribarsky, W. & Chang, R. (2009). iPCA: An Interactive System for PCA-based Visual Analytics. *Eurographics/ VGTC Symposium on Visualization, 28,* 3.

Kandemir, C. C., & Nasibov, E. (2012). Classification of apoptosis proteins by discriminant analysis, *Turkish Journal of Biochemistry, 37,* 54–61.

Khashei, M., Hamadani, A., Z. & Bijari, M. (2012). A fuzzy intelligent approach to the classification problem in gene expression data analysis. *Knowledge-Based System, 27,* 465-474.

Klir, G. J. & Folger, T. A. (1988). *Fuzzy sets, uncertainty, and information*. New Jersey: Prentice Hall.

Lesk, A. M. (2005). *Introduction to bioinformatics*, NewYork: Oxford University Press.

Li, B., Zheng, C., H., Huang, D., S., Zhang, L. & Han, K. (2010). Gene expression data classification using locally linear discriminant embedding. *Computers in Biology and Medicine, 40,* 802-810.

Lin, H. (2008). The modified Mahalanobis Discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *Journal of Theoretical Biology, 252,* 350-356.

Malhi, A. & Gao, R., X. (2004). PCA-based feature selection scheme for machine defect classification. *Transactions on Instrumentation and Measurement, 53,* 1517-1525.

Mitra, S. & Acharya, T. (2003). *Data mining: Multimedia, soft computing and bioinformatics*. New Jersey: Wiley

Nasibov, E. & Kandemir-Cavas, C. (2009). Efficiency analysis of KNN and minimum distance based classifiers in enzyme family prediction. *Computational Biology and Chemistry, 33,* 461-464.

Nasibov, E. & Kandemir-Cavas, C. (2011). OWA-based linkage method in hierarchical clustering: Application on phylogenetic trees. *Expert Systems with Applications, 38,* 12684-12690.

Nasibov, E. & Kandemir-Cavas, C. (2008). Protein subcellular location prediction using optimally weighted fuzzy k-NN algorithm. *Computational Biology and Chemistry, 32,* 448-451.

Peng, H., Li, B., Luo, X., Wang, J. & Zhang, Z. (2013). A learning-based audio watermarking scheme using kernel Fisher discriminant analysis. *Digital Signal Processing, 23*, 382–389.

Pop, H., F. & Sarbu, C. (2013). A new fuzzy discriminant analysis method. *MATCH: Communications in Mathematical and in Computer Chemistry, 69,* 391-412.

Rodríguez-Piñeiro, A. M., Rodríguez-Berrocal, F.J. & Páez de la Cadena, M. (2007). Improvements in the search for potential biomarkers by proteomics: Application of principal component and discriminant analyses for two-dimensional maps evaluation. *Journal of Chromatography B, 849,* 251-260.

Schölkopf, B., Smola, A. & Müller K., R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation, 10,* 1299-1319.

Shawe-Taylor, J. & Cristianini, N. (2004). *Kernel methods for pattern recognition.* New York: Cambridge University Press.

Yang, J., Zhang, D., Yong, X. & Yang, J. (2005). Two-dimensional discriminant transform for face recognition, *Pattern Recognition, 38,* 1125 – 1129.

Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L., M., et al. (2005). DisProt: a database of protein disorder. *Bioinformatics, 21,* 137-140.

Wang, C., Cao, L. & Miao, B. (2013). Optimal feature selection for sparse linear discriminant analysis and its applications in gene expression data. *Computational Statistics & Data Analysis, 66,* 140-149.

Wang, L., Yuan, Z., Chen, X. & Zhou, Z. (2009). Predicting membrane protein types with dimensionality reduction and kernel method. *Lecture Notes in Computer Science,5,* 71-74.

Xiong, J. (2006). *Essential bioinformatics*, New York: Cambridge University Press.

Zimmermann, H. J. (2001). *Fuzzy set theory and its applications* (4th ed.). Norwell: Kluwer Academic Puplishers.

Zhu, Z. & Song, Z. (2010). Fault diagnosis based on imbalance modified kernel Fisher discriminant analysis. *Digital Signal Processing, 23,* 382-389.