International Journal of Intelligent Information Systems
2019; 8(1): 26-42
http://www.sciencepublishinggroup.com/j/ijiis
doi: 10.11648/j.ijiis.20190801.15
ISSN: 2328-7675 (Print); ISSN: 2328-7683 (Online)

Discriminant Analysis for the Eigenvalues of Variance Covariance Matrix of FFT Scaling of DNA Sequences: An Empirical Study of Some Organisms

Salah Hamza Abid[1, *], Jinan Hamza Farhood[2]

[1]Mathematics Department, Education College, Al-Mustansiriya University, Baghdad, Iraq

[2]Mathematics Department, Education College, Babylon University, Babil, Iraq

Email address:

abidsalah@uomustansiriyah.edu.iq (S. H. Abid)

[*]Corresponding author

To cite this article:

Salah Hamza Abid, Jinan Hamza Farhood. Discriminant Analysis for the Eigenvalues of Variance Covariance Matrix of FFT Scaling of DNA Sequences: An Empirical Study of Some Organisms. *International Journal of Intelligent Information Systems*. Vol. 8, No. 1, 2019, pp. 26-42. doi: 10.11648/j.ijiis.20190801.15

Received: January 19, 2019; **Accepted**: March 11, 2019; **Published**: March 27, 2019

Abstract: Many studies discussed different numerical representations of DNA sequences. One naive approach for exploring the nature of a DNA sequence is to assign numerical values (or scales) to the nucleotides and then proceed with standard time series methods. The analysis will depend actually on the particular assignment of numerical values.Discriminant analysis aims to examine the dependence of one qualitative (classification) variable from several quantitative variables according to number of variations of qualitative variable we can distinction. Actually, there is a discriminant analysis for two or more groups. The essential work of discriminant analysis is to get the optimal assigning rules that will minimize the likelihood of incorrect classification of elements. In this paper, we discussed the discriminant analysis of the first, second, third and fourth eigenvalues of variance covariance matrix of Fast Fourier Transform (FFT) for numerical values representation of DNA sequences of five organisms, Human, E. coli, Rat, Wheat and Grasshopper. The analysis is based on three methods (All Variables, Forward Selection and Backward Selection) of discrimination. Functions have been reached whereby discrimination is made among organisms under consideration. Empirical studies are conducted to show the value of our point of view and the applications based on. Therefore, we recommended that, other empirical studies should be done for other organisms and statistical methods by using the point of view adopted here. Also, aspects stated here must be used in an applied manner for DNA sequences discrimination.

Keywords: FFT Scaling, DNA, Classification, Discriminant Analysis (DA), All Variables, Forward Selection, Backward Selection, Wilks-Lambda, Eigenvalue

1. Introduction

Discriminant analysis is a multivariate statistical analysis method that serves to set up a model to predict group memberships. The model consists of discriminant functions that appear based on a linear combination of predictive variables that provide the best discrimination between groups. These functions are derived from a sample whose group memberships are known. Afterward, they could be applied to new individuals or units with measures related to the same variables and unknown group memberships. Thus, although discriminant analysis is not frequently used in behavioral sciences because its assumptions are not always easy to meet, it is a conceptually and mathematically powerful multivarite statistical method. Therefore, a description and illustration of the discriminant analysis method may help increase its use [1].

In different areas of applications the term "discriminant analysis" has come to imply distinct meanings, uses, roles, etc. In the fields of learning, psychology, guidance, and others, it has been used for prediction [2-4]; in the study of classroom instruction it has been used as avariable reduction technique [5]; and in various fields it has been used as an adjunct to MANOVA [6]. In this sense, discriminant analysis

as a general research technique can be very useful in the investigation of various aspects of a multivariate research problem. Tatsuoka and Tiedeman [7] emphasized the multiphasic character of discriminant analysis in the early 1950s: (a) the establishment of significant group-differences, (b) the study and 'explanation' of these differences, and finally (c) the utilization of multivariate information from the samples studied in classifying a future individual known to belong to one of the groups represented. Essentially these same three problems related to discriminatory analysis.

Originally developed in 1936 by R. A. Fisher [8, 9], Discriminant Analysis is a classic method of classification that has stood the test of time. Discriminant analysis often produces models whose accuracy approaches (and occasionally exceeds) more complex modern methods. Discriminant analysis can be used only for classification (i.e., with a categorical target variable), not for regression. The target variable may have two or more categorical data.

Discriminant analysis is a powerful statistical pattern recognition method which has been applied to many DNA sequence motif finding problems. On other words, discriminant analysis is widely used in biological analyzes, including DNA analysis. Some of the relevant scientific literatures are as follows.

Solovyev and Salamov [10], introduced a complex of new programs for promoter, 3'-processing, splice sites, coding exons and gene structure identification in genomic DNA of several model species. The human gene structure prediction program FGENEH, exon prediction - FEXH and splice site prediction - HSPL have been modified for sequence analysis of Drosophila (FGENED, FEXD and DSPL), C. elegance (FGENEN, FEXN and NSPL), Yeast (FEXY and YSPL) and Plant (FGENEA, FEXA and ASPL) genomic sequences. They recomputed all frequency and discriminant function parameters for these organisms and adjusted organism specific minimal intron lengths. An accuracy of coding region prediction for these programs is similar with the observed accuracy of FEXH and FGENEH. They have developed FEXHB and FGENEHB programs combining pattern recognition features and information about similarity of predicted exons with known sequences in protein databases. These programs have approximately 10% higher average accuracy of coding region recognition. Two new programs for human promoter site prediction (TSSG and TSSW) have been developed which use Ghosh [11] and Wingender [12] data bases functional motifs, respectively. POLYAH program was designed for prediction of 3"-processing regions in human genes and CDSB program was developed for bacterial gene prediction. They have developed a new approach to predict multiple genes based on double dynamic programming, that is very important for analysis of long genomic DNA fragments generated by genome sequencing projects.

Since the identification of functional motifs in a DNA sequence is fundamentally a statistical pattern recognition problem. Discriminant analysis is widely used for solving such problems. Zhang [13], described two basic parametric methods: LDA (linear discriminant analysis) and QDA (quadracic discriminant analysis). He demonstrated their usage in recognition of splice sites and exons in the human genome.

Dudoit et al. [14] compared the performance of difierent discrimination methods for the classification of tumors based on gene expression data. These methods included: nearest neighbor classifiers, linear discriminant analysis, and classifcation trees. They also considered recent machine learning approaches such as bagging and boosting. They investigated the use of prediction votes to assess the confidence of each prediction. The methods are applied to datasets from three recently published cancer gene expression studies.

Kwon et al. [15] finded the causal relationship between several tumors and the gene-expression data by sequentially using the stepwise discriminant analysis method (SDA) and Bayesian decision theory (BDT). Eighty-five samples containing four tumor classes are used in this study. The classes are neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (BL) and the Ewing family of tumor (EWS). SDA is used to select critical genes for accurate classification of 4 tumors from original 2308 genes. With the selected genes, Bayesian classifier is made, which minimizes the misclassification rate. As a result, the classification performance increased to 100% and 9 new genes that have relation with the development of the tumors is found additionally.

Liu et al. [16] analyzed various functional regions of the human genome based on sequence fea- tures, including word frequency, dinucleotide relative abundance, and base-base correlation. They analyzed the human chromosome 22 and classified the upstream, exon, intron, downstream, and intergenic regions by principal component analysis and discriminant analysis of these features. The results show that they could classify the functional regions of genome based on sequence feature and discriminant analysis.

Guo et al. [17] in the same year, presented a modified version of linear discriminant analysis, called "shrunken centroids regularized discriminant analysis" (SCRDA). The SCRDA method is specially designed for classification problems in high dimension low sample size situations, for example, microarray data. Through both simulated data and real life data, it is shown that this method performed very well in multivariate classification problems, often outperforms the PAM method and can be as competitive as the SVM classifiers. It is also suitable for feature elimination purpose and can be used as gene selection method.

Jombart et al. [18], proposed the discriminant analysis of principal components (DAPC), a multivariate method designed to identify and describe clusters of genetically related individuals. When group priors are lacking, DAPC uses sequential K-means and model selection to infer genetic clusters. They evaluated the performance of our method using simulated data, which were also analyzed using STRUCTURE as a benchmark. Additionally, they illustrated the method by analyzing microsatellite polymorphism in

worldwide human populations and hemagglutinin gene sequence variation in seasonal influenza.

It is well known that outliers are present in virtually every data set in any application domain, and classical discriminant analysis methods (including linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA)) do not work well if the data set has outliers. In order to overcome the difficulty, Jin and An [19] used the robust statistical method. They choosed four different coding characters as discriminant variables and an approving result is presented by the method of robust discriminant analysis.

Libbrecht et al. [20], provided an overview of machine learning applications for the analysis of genome sequencing data sets, including the annotation of sequence elements and epigenetic, proteomic or metabolomic data. They introduced considerations and recurrent challenges in the application of supervised, semi-supervised and unsupervised machine learning methods, as well as of generative and discriminative modelling approaches. They provided general guidelines to assist in the selection of these machine learning methods and their practical application for the analysis of genetic and genomic data sets.

Corvelo et al. [21], introduced taxMaps, a highly efficient, sensitive, and fully scalable taxonomic classification tool. Using a combination of simulated and real metagenomics data sets, they demonstrate that taxMaps is more sensitive and more precise than widely used taxonomic classifiers and is capable of delivering classification accuracy comparable to that of BLASTN, but at up to three orders of magnitude less computational cost.

2. DNA Sequence

In the process of developing the technology, many possible interesting adaptations became apparent: One of the most interesting directions was the use of the technology in the analysis of long DNA sequences. A benefit of the techniques was that it combined rigorous statistical analysis with modern computer power to quickly search for diagnostic patterns within long DNA sequences. Briefly, a DNA strand can be viewed as a long string of linked nucleotides. Each nucleotide is composed of a nitrogenous base, a five carbon sugar, and a phosphate group. There are four different bases that can be grouped by size, the pyrimidines, thymine (T) and cytosine (C), and the purines, adenine (A) and guanine (G). The nucleotides are linked together by a backbone of alternating sugar and phosphate groups with the $5^{/}$ carbon of one sugar linked to the $3^{/}$ carbon of the next, giving the string direction. DNA molecules occur naturally as a double helix composed of polynucleotide strands with the bases facing inward. The two strands are complementary, so it is sufficient to represent a DNA molecule by a sequence of bases on a single strand; refer to Figure 1. Thus, a strand of DNA can be represented as a sequence $\{X_t; t = 1,2,...,n\}$ of letters, termed base pairs (*bp*), from the finite alphabet $\{A,C,G,T\}$ .1 The order of the nucleotides contains the genetic information specific to the organism. Expression of information stored in these molecules is a complex multistage process. One important task is to translate the information stored in the protein-coding sequences (CDS) of the DNA (Polovinkina et al. (2016) [22]).

A common problem in analyzing long DNA sequence data is in identifying CDS that are dispersed throughout the sequence and separated by regions of noncoding (which makes up most of the DNA). Another problem of interest that we will address here is that of matching two DNA sequences, say X_{1t} and X_{2t} . The background behind the problem is discussed in detail in the study by Waterman and Vingron [23]. For example, every new DNA or protein sequence is compared with one or more sequence databases to find similar or homologous sequences that have already been studied, and there are numerous examples of important discoveries resulting from these database searches.

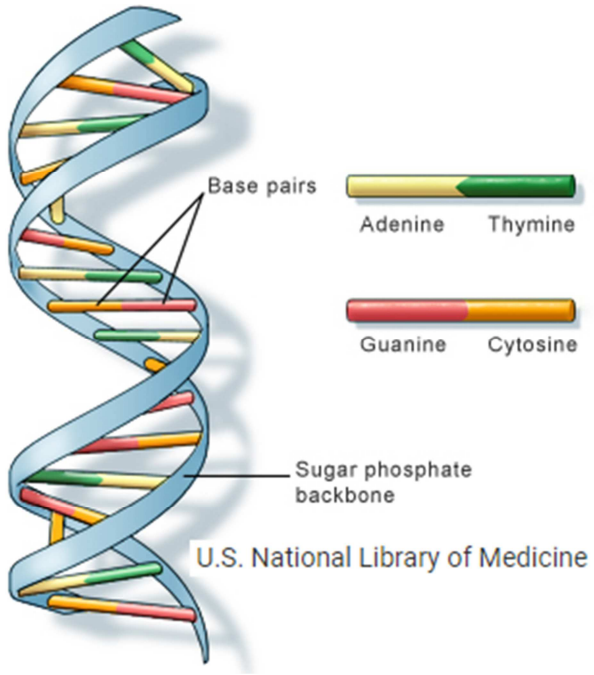


Figure 1. *The general structure of DNA and its bases.*

One naive approach for exploring the nature of a DNA sequence is to assign numerical values (or scales) to the nucleotides and then proceed with standard time series methods. It is clear, however, that the analysis will depend on the particular assignment of numerical values. Consider the artificial sequence ACGTACGTACGT... Then, setting A = G = 0 and C = T = 1, yields the numerical sequence 010101010101..., or one cycle every two base pairs (i.e., a frequency of oscillation of $\omega = 1/2$ Cycle/bp, or a period of oscillation of length $1/\omega = 2$ bp=cycle). Another interesting scaling is A = 1, C = 2, G = 3, and T = 4, which results in the sequence 123412341234..., or one cycle every four bp $(\omega = 1/4)$. In this example, both scalings of the nucleotides are interesting and bring out different properties of the sequence. It is clear, then, that one does not want to focus on only one scaling. Instead, the focus should be on finding all possible scalings that bring our interesting features of the

data. Rather than choose values arbitrarily, the spectral envelope approach selects scales that help emphasize any periodic feature that exists in a DNA sequence of virtually any length in a quick and automated fashion. In addition, the technique can determine whether a sequence is merely a random assignment of letters [22]).

Fourier analysis has been applied successfully in DNA analysis; McLachlan and Stewart [24] and Eisenberg et al. [25] studied the periodicity in proteins using Fourier analysis.

Stoffer et al. [26] proposed the spectral envelope as a general technique for analyzing categorical-valued time series in the frequency domain. The basic technique is similar to the methods established by Tavar´e and Giddings [27] and Viari et al. [28], however, there are some differences. The main difference is that the spectral envelope methodology is developed in a statistical setting to allow the investigator to distinguish between significant results and those results that can be attributed to chance.

The article authored by Marhon and Kremer [29], partitions the identification of protein-coding regions into four discrete steps. Based on this partitioning, digital signal processing DSP techniques can be easily described and compared based on their unique implementations of the processing steps. They compared the approaches, and discussed strengths and weaknesses of each in the context of different applications. Their work provides an accessible introduction and comparative review of DSP methods for the identification of protein-coding regions. Additionally, by breaking down the approaches into four steps, they suggested new combinations that may be worthy of future studies. A new methodology for the analysis of DNA/RNA and protein sequences is presented by Bajic [30]. It is based on a combined application of spectral analysis and artificial neural networks for extraction of common spectral characterization of a group of sequences that have the same or similar biological functions. The method does not rely on homology comparison and provides a novel insight into the inherent structural features of a functional group of biological sequences. The nature of the method allows possible applications to a number of relevant problems such as recognition of membership of a particular sequence to a specific functional group or localization of an unknown sequence of a specific functional group within a longer sequence. The results are of general nature and represent an attempt to introduce a new methodology to the field of biocomputing. Fourier transform infrared (FTIR) spectroscopy has been considered by Han et al. [31] as a powerful tool for analysing the characteristics of DNA sequence. This work investigated the key factors in FTIR spectroscopic analysis of DNA and explored the influence of FTIR acquisition parameters, including FTIR sampling techniques, pretreatment temperature, and sample concentration, on calf thymus DNA. The results showed that the FTIR sampling techniques had a significant influence on the spectral characteristics, spectral quality, and sampling efficiency. Ruiz et al. [32] proposed a novel approach for performing cluster analysis of DNA sequences that is based on the use of Genomic signal processing GSP methods and the K-means algorithm. We also propose a visualization method that facilitates the easy inspection and analysis of the results and possible hidden behaviors. Our results support the feasibility of employing the proposed method to find and easily visualize interesting features of sets of DNA data. A novel clustering method is proposed by Hoang et al. [33] to classify genes and genomes. For a given DNA sequence, a binary indicator sequence of each nucleotide is constructed, and Discrete Fourier Transform is applied on these four sequences to attain respective power spectra. Mathematical moments are built from these spectra, and multidimensional vectors of real numbers are constructed from these moments. Cluster analysis is then performed in order to determine the evolutionary relationship between DNA sequences. The novelty of this method is that sequences with different lengths can be compared easily via the use of power spectra and moments. Experimental results on various datasets show that the proposed method provides an efficient tool to classify genes and genomes. It not only gives comparable results but also is remarkably faster than other multiple sequence alignment and alignment-free methods. One challenge of GSP is how to minimize the error of detection of the protein coding region in a specified DNA sequence with a minimum processing time. Since the type of numerical representation of a DNA sequence extremely affects the prediction accuracy and precision, by this study Mabrouk [34] aimed to compare different DNA numerical representations by measuring the sensitivity, specificity, correlation coefficient (CC) and the processing time for the protein coding region detection. The proposed technique based on digital filters was used to read-out the period 3 components and to eliminate the unwanted noise from DNA sequence. This method applied to 20 human genes demonstrated that the maximum accuracy and minimum processing time are for the 2-bit binary representation method comparing to the other used representation methods. Results suggest that using 2-bit binary representation method significantly enhanced the accuracy of detection and efficiency of the prediction of coding regions using digital filters. Identification and analysis of hidden features of coding and non-coding regions of DNA sequence is a challenging problem in the area of genomics. The objective of the paper authored by Roy and Barman [35] is to estimate and compare spectral content of coding and non-coding segments of DNA sequence both by Parametric and Nonparametric methods. Consequently an attempt has been made so that some hidden internal properties of the DNA sequence can be brought into light in order to identify coding regions from non-coding ones. In this approach the DNA sequence from various Homo Sapien genes have been identified for sample test and assigned numerical values based on weak-strong hydrogen bonding (WSHB) before application of digital signal analysis techniques. The statistical methodology applied for computation of Spectral content are simple and the Spectrum plots obtained show satisfactory results. Spectral analysis can be applied to study

base-base correlation in DNA sequences. A key role is played by the mapping between nucleotides and real/complex numbers. In 2006, Galleani and Garello [36] presented a new approach where the mapping is not kept fixed: it is allowed to vary aiming to minimize the spectrum entropy, thus detecting the main hidden periodicities. The new technique is first introduced and discussed through a number of case studies, then extended to encompass time-frequency analysis.

For analyzing periodicities in categorical valued time series, the concept of the spectral envelope was introduced by Stoffer et al. [37] as a computationally simple and general statistical methodology for the harmonic analysis and scaling of non-numeric sequences. However, the spectral envelope methodology is computationally fast and simple because it is based on the fast Fourier transform and is nonparametric (i.e., it is model independent). This makes the methodology ideal for the analysis of long DNA sequences. Fourier analysis has been used in the analysis of correlated data (time series) since the turn of the century. Of fundamental interest in the use of Fourier techniques is the discovery of hidden periodicities or regularities in the data. Although Fourier analysis and related signal processing are well established in the physical sciences and engineering, they have only recently been applied in molecular biology. Since a DNA sequence can be regarded as a categorical-valued time series it is of interest to discover ways in which time series methodologies based on Fourier (or spectral) analysis can be applied to discover patterns in a long DNA sequence or similar patterns in two long sequences. Actually, the spectral envelope is an extension of spectral analysis when the data are categorical valued such as DNA sequences.

An algorithm for estimating the spectral envelope and the optimal scalings given a particular DNA sequence with alphabe $\xi = \{b_1, b_2, ..., b_{r+1}\}$, is as follows [26].

(1) Given a DNA sequence of length *n*, from the $r \times 1$ vectors $Y_t, t = 1, 2, ..., n$; namely, for $j = 1, 2, ..., r, Y_t = e_j$ if $X_t = b_j$ where e_j is a $r \times 1$ vector with a 1 in the *j*th position as zeros elsewhere, and $Y_t = 0$ if $X_t = b_{j+1}$.

(2) Calculate the Fast Fourier Transform FFT of the data, $d(j/n) = \sum_{t=1}^{n} Y_t \exp(-2\pi i t j/n) \Big/ \sqrt{n}$.

Note that $d(j/n)$ is a $r \times 1$ complex-valued vector. Calculate the periodogram,

$\tilde{f}(j/n) = d(j/n) d^*(j/n)$, for $j = 1, 2, ..., [n/2]$, and retain only the real part, say $f^{\sim re}(j/n)$.

(3) Smooth the real part of the periodogram as preferred to obtain $f^{\sim re}(j/n)$, a consistent estimator of the real part of the spectral matrix.

(4) Calculate the $r \times r$ variance–covariance matrix of the data, $S = \sum_{t=1}^{n} (Y_t - \bar{Y})(Y_t - \bar{Y})' \Big/ n$, where $\bar{Y}$ is the sample mean of the data.

(5) For each $\omega = j/n, j = 1, 2, ..., [n/2]$, determine the largest eigenvalue and the corresponding eigenvector of the matrix $2S^{-1/2} f^{\sim re}(\omega_j) S^{-1/2} \Big/ n$.

(6) The sample spectral envelope $\hat{\lambda}(\omega_j)$ is the eigenvalue obtained in the previous step.

(7) The optimal sample scaling is $\hat{\beta}(\omega_j) = S^{-1/2} v(\omega_j)$, where $v(\omega_j)$ the eigenvector obtained in the previous step.

In this paper, we discussed the discriminant analysis of the first, second, third and fourth eigenvalues of variance covariance matrix of Fast Fourier Transform (FFT) for numerical values representation of DNA sequences of five organisms, Human, E. coli, Rat, Wheat and Grasshopper. The analysis is based on three methods (All Variables, Forward Selection and Backward Selection) of discriminating. It should be noted that it is the first time that the variance covariance matrix eigenvalues of Fast Fourier Transform (FFT) for numerical values representation of DNA sequences, is used in an analysis like this and related analyzes.

3. Discriminant Analysis

Discriminant analysis aims to examine the dependence of one qualitative (classification) variable from several quantitative variables according to number of variations of qualitative variable we can distinction [38]). Actually, there is a discriminant analysis for two or more groups. The essential work of discriminant analysis is to get the optimal assigning rules that will minimize the likelihood of incorrect classification of elements. Every element is distinguished by some aspects which reflect its properties. This means that corresponding to measured characteristics; the examined elements are realizations of the random vector $X = (X_1, X_2, ..., X_n)$. The process starts with an analysis of group of elements in which is known relation to a specific group and also values of the random variables. The analysis result of the training set is to determine discriminant function that define the likelihood of classification of new unclassified element to certain group according to measured values. $x = (x_1, x_2, ..., x_m)$ of its characteristics [39].

Two basic aims of discriminant analysis are stated by Stankovičová and Vojtková [38], the first aim is to find appropriate statistical way to distinguish between groups (Descriptive or analytical). The second aim is to include new statistical unit (object) that is recognized by a vector of k features to one of the based groups (Classification).

4. Discriminant Analysis: Aims and Assumptions

Discriminant analysis aims is offered by Meloun et al. [40].

(1) Define whether there are significant statistical differences among profiles of the average score of

discriminators for two or more pre-defined classes.

(2) Define which of the discriminator is reflected the most in differential profiles of average score of two or more classes.

(3) Define procedures to involve objects into classes according to their score in discriminators set.

(4) Define the number of dimensions compilation of discrimination among classes created by a discriminators set.

Assumptions of discrimination model

(1) Multivariate normal distribution

conduct tests of significance of individual discriminatory variables and discriminatory functions are needful to assure this assumption. If the data is not distributed as multi-dimensional normal, then the results of classification are inaccurate. Moreover, the classification total error is not violated by Lack of performance of normal assumption because the classification error in one group may be overestimated and in the other group underestimated. [39]

(2) At least two groups must be there, with each case belonging to only one group so that the groups are independent and collectively exhaustive.

(3) Each group must be well defined and clearly distinguished from any one of groups.

(4) Before collecting the data, the groups should be well defined [41].

(5) Equality of variance-covariance within group.

(6) The covariance matrix within each group should be equal. Equality Test of Covariance Matrices can be used to verify it. When in doubt, try re-running the analyses using the Quadratic method, or by adding more observations or excluding one or two groups.

(7) Low multicollinearity of the variables

When high multicollinearity among two or more variables is present, the discriminant function coefficients will not reliably predict group membership. We can use the pooled within-groups correlation matrix to detect multicollinearity [42].

5. Practical and Computational Steps for Discrimination and Classification

In this section, we will introduce the discrimination and classification from practical and computational aspect.

5.1. Discrimination Among Several Populations

Suppose that we have p of populations, from the first population a sample $X_1, X_2, \ldots, X_{n_1}$ is drown, from the second population a sample $X_{n_1+1}, X_{n_1+2}, \ldots, X_{n_1+n_2}$ is drown and so on from the *pth* population a sample $X_{n_1+\ldots+n_{p-1}+1}, \ldots, X_{n_T}$, where $n_1+n_2+\ldots+n_p=n_T$. Let $\bar{X}_j$ be the sample mean for the population j, $j=1,\ldots,p$, and $\bar{X}=\sum_{i=1}^{n_T} X_i \Big/ n_T$. Then the sample between matrix is,

$$B=\sum_{j=1}^{p} n_j(\bar{X}_j-\bar{X})(\bar{X}_j-\bar{X})^t$$

Thus,

$$a^tBa=\sum_{j=1}^{p} n_j a^t\left(\bar{X}_j-\bar{X}\right)\left(\bar{X}_j-\bar{X}\right)^t a=\sum_{j=1}^{p} n_j\left(a^t\bar{X}_j-a^t\bar{X}\right)\left(\bar{X}_j^t a-\bar{X}^t a\right)=\sum_{j=1}^{p} n_j\left(\bar{Y}_j-\bar{Y}\right)^2, \tag{1}$$

$Y_i=a^tX_i, i=1,\ldots,n_T$, $\bar{Y}_j$ is the mean for the j'th population, $j=1,\ldots,p$.

The sample within group matrix is

$$W=\sum_{i=1}^{n_1}\left(X_i-\bar{X}_1\right)\left(X_i-\bar{X}_1\right)^t+\sum_{i=n_1+1}^{n_1+n_2}\left(X_i-\bar{X}_2\right)\left(X_i-\bar{X}_2\right)^t+\cdots+\sum_{i=n_1+\cdots n_{g-1}+1}^{n_T}\left(X_i-\bar{X}_g\right)\left(X_i-\bar{X}_g\right)^t \tag{2}$$

Thus,

$$a^tWa=\sum_{i=1}^{n_1} a^t\left(X_i-\bar{X}_1\right)\left(X_i-\bar{X}_1\right)^t a+\cdots+\sum_{i=n_1+\ldots+n_{g-1}+1}^{n_T} a^t\left(X_i-\bar{X}_p\right)\left(X_i-\bar{X}_p\right)^t a$$

$$=\sum_{i=1}^{n_1}\left(Y_i-\bar{Y}_1\right)^2+\sum_{i=n_1+1}^{n_1+n_2}\left(Y_i-\bar{Y}_2\right)^2+\cdots+\sum_{i=n_1+\ldots+n_{g-1}+1}^{n_T}\left(Y_i-\bar{Y}_g\right)^2 \tag{3}$$

The pooled estimate based on $Y_1, Y_2, \ldots, Y_{n_T}$ is

$$\frac{a^tWa}{n_T-p}=\frac{\sum_{i=1}^{n_1}\left(Y_i-\bar{Y}_1\right)^2+\sum_{i=n_1+1}^{n_1+n_2}\left(Y_i-\bar{Y}_2\right)^2+\ldots+\sum_{i=n_1+\ldots+n_{p-1}+1}^{n_T}\left(Y_i-\bar{Y}_p\right)^2}{n_T-g} \tag{4}$$

The pooled estimate based on $X_1, X_2, \ldots, X_{n_T}$ is

$$S_{pooled} = \frac{W}{n_T - p} = \frac{\sum_{i=1}^{n_1}\left(X_i - \bar{X}_1\right)\left(X_i - \bar{X}_1\right)^t + \cdots + \sum_{i=n_1+\ldots+n_{p-1}+1}^{n_T}\left(X_i - \bar{X}_g\right)\left(X_i - \bar{X}_g\right)^t}{n_T - p} \tag{5}$$

Now we will present Fisher's linear discriminant method for several populations. Fisher's discriminant method for several populations is as follows steps,

Find the vector $\hat{a}_1$ maximizing the separation function

$$S(a) = \frac{a^t B a}{a^t W a} = \frac{\sum_{j=1}^{p} n_j\left(\bar{Y}_j - \bar{Y}\right)^2}{\sum_{i=1}^{n_1}\left(Y_i - \bar{Y}_1\right)^2 + \sum_{i=n_1+1}^{n_1+n_2}\left(Y_i - \bar{Y}_2\right)^2 + \cdots + \sum_{i=n_1+\cdots+n_{g-1}+1}^{n_T}\left(Y_i - \bar{Y}_g\right)^2}, \tag{6}$$

subject to $\hat{a}_1^t S_{pooled}\hat{a}_1 = 1$.

The linear combination $\hat{a}_1^t X$ is called the sample first discriminant.

Find the vector $\hat{a}_2$ maximizing the separation function $S(a)$ subject to $\hat{a}_2^t S_{pooled}\hat{a}_2 = 1$ and $\hat{a}_2^t S_{pooled}\hat{a}_1 = 0$. So on,

Find the vector $\hat{a}_s$ maximizing the separation function $S(a)$ subject to $\hat{a}_s^t S_{pooled}\hat{a}_s = 1$ and $\hat{a}_s^t S_{pooled}\hat{a}_l = 0, l < s.$

Note that, $\hat{a}_j^t S_{pooled}\hat{a}_j$ is the estimate of $Var(\hat{a}_j^t X), j = 1,\ldots,s$ and $\hat{a}_j^t S_{pooled}\hat{a}_l, j \neq l.$ is the estimate of $Cov(\hat{a}_j^t X, \hat{a}_l^t X), j \neq l.$

The condition $\hat{a}_j^t S_{pooled}\hat{a}_l = 0$ is like to the condition given in the principal component analysis.

Axiomatically, $S(a)$ measures the difference among the transformed means reverberated by $\sum_{j=1}^{p} n_j\left(\bar{Y}_j - \bar{Y}\right)^2$ close to the random variation of the transformed data reverberated by $\sum_{i=1}^{n_1}\left(Y_i - \bar{Y}_1\right)^2 + \sum_{i=n_1+1}^{n_1+n_2}\left(Y_i - \bar{Y}_2\right)^2 + \cdots + \sum_{i=n_1+\ldots+n_{g-1}+1}^{n_T}\left(Y_i - \bar{Y}_k\right)^2$. As the transformed observations $Y_1, Y_2, \ldots, Y_{n_1}(population \cdot 1)$, $Y_{n_1+1}, Y_{n_1+2}, \ldots, Y_{n_1+n_2}(population \cdot 2), \ldots, Y_{n_1+\ldots+n_{p-1}+1}, \ldots, Y_{n_T}(population \cdot p)$ are separated, $\sum_{j=1}^{p} n_j\left(\bar{Y}_j - \bar{Y}\right)^2$ should be large even as the random variation of the transformed data is taken into account.

Two Important results provide ways to obtain the discriminants,

Let $e_1, e_2, \ldots, e_s$ be the orthonormal eigenvector of $W^{-1/2} B W^{-1/2}$ corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s > 0.$ Then, $\hat{a}_j = S_{pooled}^{-1/2} e_j, j = 1,\ldots,s,$ where $S_{pooled}^{-1/2} S_{pooled}^{-1/2} = S_{pooled}^{-1}$.

Let $e_1, e_2, \ldots, e_s$ be the eigenvectors of $W^{-1}B$ corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s > 0.$ Then, $\hat{a}_j, j = 1,\ldots,s,$ are the scaled eigenvectors satisfying $\hat{a}_j{}^t S_{pooled}\hat{a}_j = 1$. That is, $\hat{a}_j = \frac{e_j}{\sqrt{e_j^t S_{pooled} e_j}}$

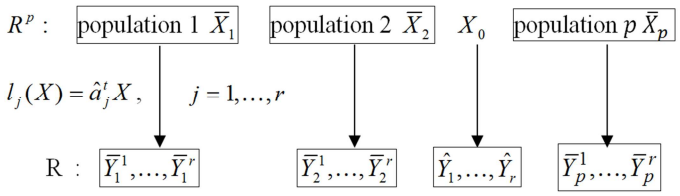
5.2. Classification for Several Populations

Fisher's classification procedure according to the first $r \leq s$ sample discriminants is to assign observation X_0 to the first population if

$$\sum_{j=1}^{r}\left(\hat{Y}_j - \bar{Y}_l^j\right)^2 = \sum_{j=1}^{r}\left[\hat{a}_j^t\left(X_0 - \bar{X}_l\right)\right]^2 \leq \sum_{j=1}^{r}\left[\hat{a}_j^t\left(X_0 - \bar{X}_i\right)\right]^2 = \sum_{j=1}^{r}\left(\hat{Y}_j - \bar{Y}_i^j\right)^2, i \neq l,$$

Where, $\hat{Y}_j = \hat{a}_j^t X_0, \bar{Y}_i^j = \hat{a}_j^t \bar{X}_i, j = 1,\ldots,r; i = 1,\ldots p$

Intuition of Fisher's method,



$\sum_{j=1}^{r}\left(\hat{Y}_j-\overline{Y}_1^j\right)^2$: represent the total square distance between the transformed X_0 ($\hat{Y}_1,\ldots,\hat{Y}_r$) and the transformed mean of the first population ($\overline{Y}_1^1,\ldots,\overline{Y}_1^r$).

$\sum_{j=1}^{r}\left(\hat{Y}_j-\overline{Y}_2^j\right)^2$: represent the total square distance between the transformed X_0 ($\hat{Y}_1,\ldots,\hat{Y}_r$) and the transformed mean of the second population ($\overline{Y}_2^1,\ldots,\overline{Y}_2^r$) and so on,

$\sum_{j=1}^{r}\left(\hat{Y}_j-\overline{Y}_k^j\right)^2$: represent the total square distance between the transformed ($\hat{Y}_1,\ldots,\hat{Y}_r$) X_0 and the transformed mean of the *pth* population ($\overline{Y}_p^1,\ldots,\overline{Y}_p^r$).

$\Rightarrow \sum_{j=1}^{r}\left(\hat{Y}_j-\overline{Y}_l^j\right)^2 \le \sum_{j=1}^{r}\left(\hat{Y}_j-\overline{Y}_i\right)^2, i\neq l,$ imply the total distance between the transformed X_0 and the transformed mean of the first population is smaller than the one between the one between the transformed X_0 and the transformed mean of the other populations. In another meaning, X_0 is closer to the first population than to the other populations. Therefore, X_0 is assigned to the first population.

6. Stepwise Discriminant Analysis

In stepwise discriminant analysis, large number of variables are entered, then with a series of steps, we are selected variables which discriminate the best and from them is created discriminant function. We can identify by some criteria how the stepwise discriminant analysis seeks at chosen of these variables (Stankovičová and Vojtková (2007) [38]).

(1) *Forward selection*: Variables come in into the discriminant function progressively and constantly is chosen the one that has the paramount profit in terms of discrimination. If this benefit is not statistically significant, no new variable enter into the function.

(2) *Backward selection*: here we get in all variables In the discriminant function and gradually are outcasted those whose removal does not case a statistically significant decrease rate of discrimination. When any other throw away would intend significant decrease in discrimination between groups, Then this process is completed.

(3) *Stepwise selection*: This chosen is mixing of the two past procedures. Here, enter new variables by degrees into discriminant function and it is always selection one with the utmost assist in terms of discrimination, while in every step is confirmed the possibility whether the variable would be eliminated and if eliminated variable does not have significant effect on decrease rate of discrimination [39].

Whatever, these procedures attain same outcomes but stipulation is that the input data have to be mutually uncorrelated. Otherwise if the correlation between input variables is significant, it is approperate to take Stepwise selection, where initially selected variable may be excluded in further steps because it is only correlation of other variables in the model. Criteria for making decision about enter of variable into the model or its elimination from the model avail following statistics. [40]

Wilks Lambda (λ)

The ratio of intra – group variability to the total variability represent Wilks λ statistic. At every step is chosen the variable that satisfies the minimum value of this statistic. The significance of changes of Wilks criteria after discriminators submitting into the model or abstraction from the model is based on F test criterion. The value of F for change of Wilks criteria while adding discriminator into the model so that the model includes p discriminators is calculated as follows,

$$F_{zmeny}=\frac{n-g-p}{g-1}\left[\frac{1-\lambda_{p+1}}{\frac{\lambda_p}{\frac{\lambda_{p+1}}{\lambda_p}}}\right] \tag{7}$$

Where p represent the number of discriminators in the model, n represent the total number of objects, g represent the number of classes, and λ_p and λ_{p+1} represent Wilks criterion before and after adding discriminators to the model respectively.

Härdle and Simar (2012) [43], derived Wilks lambda as follows,

$$\Psi=\frac{SS_{withn-groups}}{SS_{total}} \tag{8}$$

So the smaller value of Ψ implies to more doubt upon the null hypothesis.

Determination the amount of variance in the grouping variable is interpreted by predictor variables by subtracting Ψ from one [41].

7. An Empirical Study

The following algorithm steps is performed to achieve our aims.

Generate the DNA sequence for five organisms, Human, E. coli, Rat, Wheat and Grasshopper with corresponding information in table 1.

Table 1. *Relative proportions (%) of Bases in DNA.*

Organisms	A	T	G	C
Human	30.9	29.4	19.9	19.8
E. coli	26.0	23.9	24.9	25.2
Rat	28.6	28.4	21.4	21.5
Wheat	27.3	27.1	22.7	22.8
Grasshopper	29.3	29.3	20.5	20.7

The sequence size is n=500 and run size is k=205.

Transform DNA sequence to numerical values by setting one to the base that appears and zero to the other bases.

Transform the sequence of numerical values to the corresponding FFT values.

Calculate the eigenvalues of variance covariance matrix for each run results, and then we get 205 fourth order vectors of eigenvalues for each organism. Each vector contains the four eigenvalues, rank from the largest one to the smallest.

All Variables, Forward Selection and Backward Selection methods of discrimination have been applied of the first, second, third and fourth variance- covariance matrix eigenvalues of Fast Fourier Transform (FFT) for numerical values representation of DNA sequences of five organisms, Human, E. coli, Rat, Wheat and Grasshopper. It should be noted that it is the first time that the variance covariance matrix eigenvalues of Fast Fourier Transform (FFT) for numerical values representation of DNA sequences, is used in an analysis like this and related analyzes.

For convenient, in the following discussions, we will refer to the organism by the first letter of his name.

7.1. Results and Discussion

Table 2. *Eigenvalues and canonical correlation.*

Discriminant Function	Eigenvalue	Relative Percentage	Canonical Correlation
1	1.21447	75.87	0.74056
2	0.357763	22.35	0.51332
3	0.0262985	1.64	0.16008
4	0.0021347	0.13	0.04615

The three methods (All Variables, Forward Selection and Backward Selection) methods of discriminating are designed to develop a set of discriminating functions which can help predict cf based on the values of other quantitative variables. 1017 cases were used to develop a model to discriminate among the 5 levels of cf. Using a stepwise selection algorithm, it was determined that 4 variables were significant predictors of cf. That is, 4 predictor variables were entered. The 3 discriminating functions with P-values less than 0.05 are statistically significant at the 95.0% confidence level.

Table 3. *Wilks lambda and P-value.*

Functions Derived	Wilks Lambda	Chi-Square	DF	P-Value
1	0.323375	1141.9250	16	0.0000
2	0.716104	337.7694	9	0.0000
3	0.9723	28.4141	4	0.0000
4	0.99787	2.1569	1	0.1419

When we use the forward selection method for the stepwise regression, we consider the following:

F-to-enter: 4.0
F-to-remove: 4.0
Step 0:
0 variables in the model.
Step 1:
Adding variable 3 with F-to-enter = 189.851
1 variables in the model.
Wilk's lambda = 0.571298 Approximate F = 189.851 with P-value = 0.0000
Step 2:
Adding variable 4 with F-to-enter = 65.601
2 variables in the model.
Wilk's lambda = 0.453574 Approximate F = 122.54 with P-value = 0.0000
Step 3:
Adding variable 2 with F-to-enter = 10.9635
3 variables in the model.
Wilk's lambda = 0.434699 Approximate F = 82.424 with P-value = 0.0000
Step 4:
Adding variable 1 with F-to-enter = 86.8387
4 variables in the model.
Wilk's lambda = 0.323375 Approximate F = 86.1478 with P-value = 0.0000
Final model selected.

and when we use the backward selection method for the stepwise regression, we consider the following:

F-to-enter: 4.0
F-to-remove: 4.0
Step 0:
4 variables in the model.
Wilk's lambda = 0.323375 Approximate F = 86.1478 with P-value = 0.0000
Final model selected.

The following Classification Function Coefficients for cf shows the functions used to classify observations. There is a function for each of the 5 levels of cf. For example, the function used for the first level of cf is

-263305. + 1054.3*1 + 1051.03*2 + 1052.79*3 + 1054.87*4

These functions are used to predict which level of cf new observations belong to.

Table 4. *Classification Function Coefficients for cf of each of five organisms.*

	e	*g*	*h*	*r*	*w*
1	1054.3	1052.06	1054.39	1053.32	1052.99
2	1051.03	1048.78	1051.12	1049.96	1049.65
3	1052.79	1050.27	1052.53	1051.6	1051.38
4	1054.87	1052.4	1054.71	1053.72	1053.49
CONSTANT	-263305	-262129	-263284	-262760	-262621

The classification function coefficients for cf in table 4, shows the functions used to classify observations. There is a function for each of the 5 levels of cf. For example, the function used for the first level of cf is

-263305. + 1054.3*1 + 1051.03*2 + 1052.79*3 + 1054.87*4

These functions are used to predict which level of cf new observations belong to.

Table 5. *Discriminant Function Coefficients for cf of each of five organisms.*

	1	2	3	4
1	-1.42552	10.3929	-2.29605	-0.837972
2	-1.20084	9.37211	-0.758743	-0.578266
3	-1.85225	8.41815	-1.2011	-1.29646
4	-1.88923	9.5499	-1.61871	0.0991884

Unstandardized Coefficients

	1	2	3	4
1	-0.192768	1.4054	-0.310486	-0.113316
2	-0.183743	1.43404	-0.116096	-0.0884812
3	-0.306698	1.39389	-0.198879	-0.214669
4	-0.27933	1.41199	-0.239333	0.0146654
CONSTANT	117.068	-705.273	108.875	51.6442

The discriminant function coefficients for cf in table 5, shows the coefficients of the functions used to discriminate amongst the different levels of cf. Of particular interest are the standardized coefficients. The first standardized discriminating function is

-1.42552*1 - 1.20084*2 - 1.85225*3 - 1.88923*4

From the relative magnitude of the coefficients in the above equation, you can determine how the independent variables are being used to discriminate amongst the groups.

The following Classification Table 6 shows the results of using the derived discriminant functions to classify observations. It lists the two highest scores amongst the classification functions for each of the 1017 observations used to fit the model, as well as for any new observations. For example, row 1 scored highest for cf = e and second highest for cf = w. In fact, the true value of cf was e. Amongst the 1017 observations used to fit the model, 583 or 57.3255% were correctly classified. You can predict additional observations by adding new rows to the current data file, filling in values for each of the independent variables but leaving the cell for cf blank.

Table 6. *Classification Table of each of five organisms.*

Actual	**Group**	**Predicted**	**cf**			
cf	**Size**	**e**	**g**	**h**	**r**	**w**
e	203	168	0	1	29	5
		(-82.76%)	(0.00%)	(-0.49%)	(-14.29%)	(-2.46%)
g	203	9	115	41	22	16
		(-4.43%)	(-56.65%)	(-20.20%)	(-10.84%)	(-7.88%)
h	205	8	0	168	29	0
		(-3.90%)	(0.00%	(-81.95%)	(-14.15%)	(0.00%)
r	203	33	37	36	57	40
		(-16.26%)	(-18.23%)	(-17.73%)	(-28.08%)	(-19.70%)
w	203	68	14	6	40	75
		(-33.50%)	(-6.90%)	(-2.96%)	(-19.70%)	(-36.95%)

Percent of cases correctly classified: 57.33%

Table 7. *Group Centroids for cf of each of five organisms.*

Group	**1**	**2**	**3**	**4**
e	-1.52919	0.456441	0.172723	-0.0272532
g	1.09833	-0.772067	0.179081	-0.0154093
h	1.31789	0.874488	-0.0143922	0.0279504
r	-0.0312728	-0.132379	-0.246469	-0.0589958
w	-0.868747	-0.435099	-0.0908001	0.0734325

The group centroids for cf in table 7, shows the average values of each of the 4 discriminant functions for each of the 5 values of cf.

The following summary statistics by group in table 8, shows the averages and standard deviations of each independent variable for each level of cf.

Table 8. *Summary Statistics by Group of each of five organisms.*

cf	**e**	**G**	**h**	**r**	**w**	**TOTAL**
COUNTS	203	203	205	203	203	1017
MEANS						
1	139.002	152.004	154.26	148.687	143.425	147.489
2	129.323	138.523	141.248	132.763	129.461	134.277
3	121.549	110.033	107.429	114.702	118.538	114.436
4	110.127	98.3395	97.0624	103.335	107.925	103.345
STD. DEVIATIONS						
1	6.22749	7.05805	7.31945	9.22462	6.79524	9.26124
2	4.12905	7.57472	6.97772	7.73094	5.53362	8.12094
3	4.66854	7.38305	6.3962	6.47941	4.80889	7.97446
4	6.93967	6.62588	6.36103	7.53226	6.28599	8.48025

Table 9. *Pooled Within-Group Statistics for cf of each of five organisms.*

Within-Group Covariance Matrix

	1	2	3	4
1	54.686	-3.04941	-22.4781	-29.0527
2	-3.04941	42.7123	-18.3195	-21.103
3	-22.4781	-18.3195	36.4736	4.43221
4	-29.0527	-21.103	4.43221	45.7441

Within-Group Correlation Matrix

	1	2	3	4
1	1	-0.0630959	-0.503306	-0.580872
2	-0.0630959	1	-0.464138	-0.477419
3	-0.503306	-0.464138	1	0.108508
4	-0.580872	-0.477419	0.108508	1

In addition, the following pooled within-group statistics for cf in table 9, shows the estimated correlations between the independent variables within each group. The within group information from all of the groups has been pooled.

This can be seen in figure 2 down.

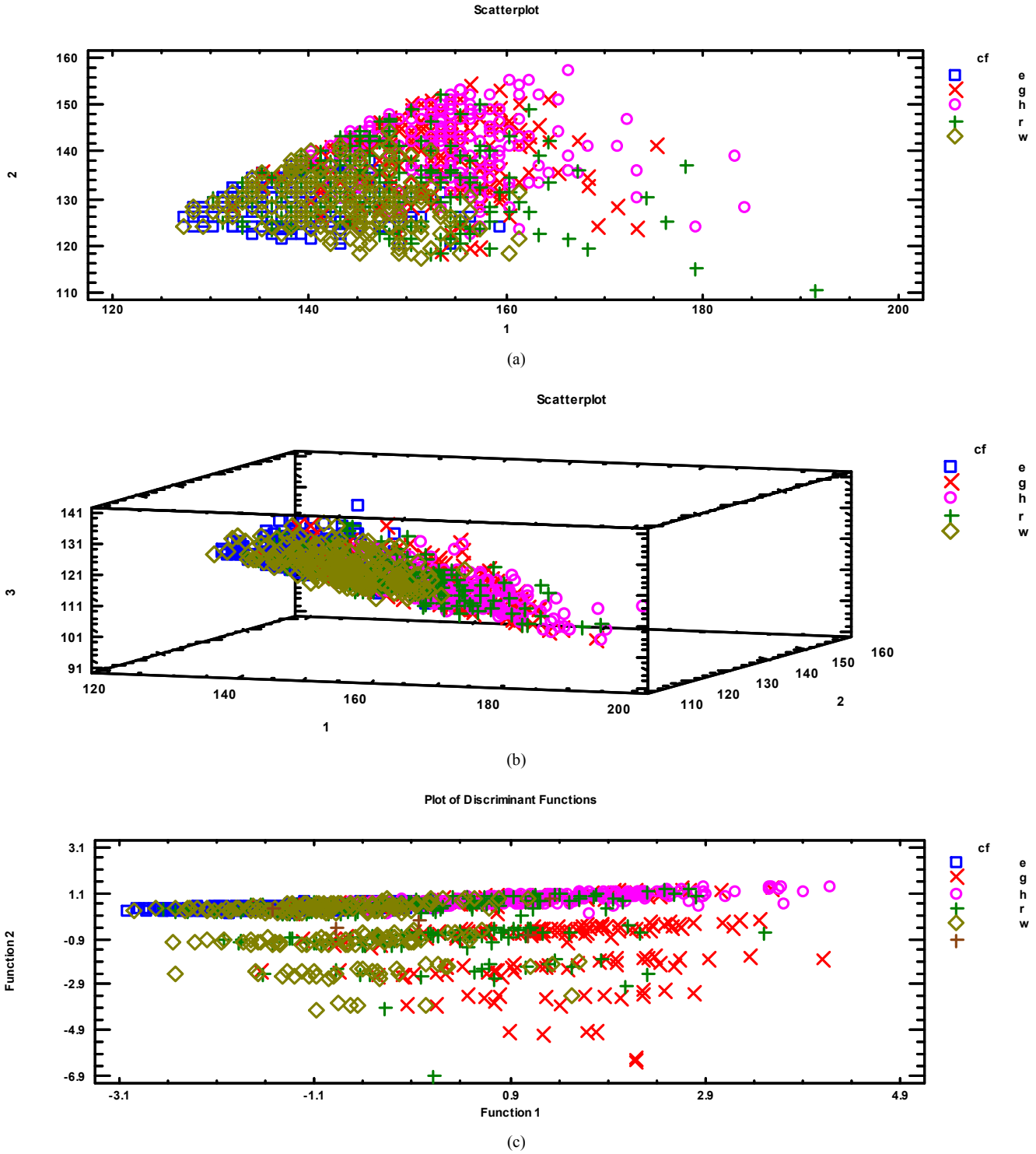


Figure 2. *Plots of: (a) Two dimension scatter plot, (b) Three dimension scatter plot, (c) Discriminant functions.*

8. Summary

Functions have been reached whereby a discrimination is made among organisms according to eigenvalues of variance covariance matrix of FFT for numerical values representation of DNA sequences, and then classify any other observation to any of organisms belong.

The methods used here are aimed to discriminant among different organisms using another point of view. This point of view is based on eigenvalues of variance covariance matrix of FFT for numerical values representation of DNA sequences. It should be noted that, it is the first time this point of view is used to achieve aims like ours.

Empirical studies are conducted to show the value of our

point of view and the applications based on. Therefore, we recommended that,

1. Other empirical studies should be done for other organisms and statistical methods by using the point of view adopted here.
2. Aspects stated here must be used in an applied manner for DNA sequences discrimination.

Appendix

Table A1. *Actual and highest two groups for each observation.*

Row	Actual Group	Highest Group	Highest Value	Squared Distance	Prob.	2nd Highest Group	2nd Highest Value	Squared Distance	Prob.
1	e	e	263318	2.32255	0.7162	w	263317	4.54363	0.2359
2	e	e	263286	0.611535	0.5394	w	263285	2.06059	0.2614
3	e	e	263306	2.52435	0.7935	w	263305	5.50989	0.1783
4	e	e	263323	1.59725	0.5233	w	263323	2.50057	0.3331
5	e	e	263307	0.329547	0.5967	w	263307	1.79854	0.2863
6	e	e	263301	1.51209	0.6882	w	263299	3.65553	0.2356
7	e	*r	263311	1.38723	0.3531	w	263311	1.65234	0.3092
8	e	e	263287	0.756714	0.6499	w	263286	2.77045	0.2374
9	e	e	263296	0.280903	0.6548	w	263295	2.22558	0.2477
10	e	e	263304	0.303498	0.6662	w	263303	2.2509	0.2516
11	e	e	263322	3.24033	0.7671	w	263321	5.88307	0.2047
12	e	e	263294	1.85343	0.5933	w	263293	3.50248	0.2601
13	e	e	263323	3.82238	0.7813	w	263322	6.60322	0.1945
14	e	e	263316	0.694525	0.6521	w	263315	2.44216	0.2722
15	e	e	263306	0.570849	0.3741	w	263305	0.916633	0.3147
16	e	e	263301	1.14726	0.6362	w	263301	2.94301	0.2592
17	e	e	263306	2.33252	0.807	w	263304	5.46231	0.1688
18	e	e	263307	0.44653	0.6718	w	263306	2.40104	0.2528
19	e	e	263299	1.14912	0.5835	w	263298	2.62183	0.2794
20	e	e	263307	2.36445	0.741	w	263306	4.85805	0.213
21	e	e	263267	3.88883	0.5018	w	263266	5.41709	0.2337
22	e	e	263318	2.83505	0.7333	w	263317	5.19423	0.2254
23	e	e	263298	2.55544	0.546	w	263297	3.92338	0.2755
24	e	e	263310	0.230065	0.5876	w	263309	1.62754	0.2922
25	e	e	263282	1.14422	0.458	w	263281	2.30997	0.2557
26	e	e	263315	1.58014	0.6668	w	263314	3.45231	0.2615
27	e	e	263296	0.709799	0.6838	w	263295	2.85819	0.2336
28	e	e	263321	1.8433	0.7198	w	263320	4.07177	0.2362
29	e	e	263294	0.915504	0.7313	w	263293	3.41748	0.2093
30	e	e	263293	1.66903	0.3331	r	263292	2.11359	0.2667
31	e	e	263290	0.675837	0.4885	w	263289	1.84403	0.2724
32	e	e	263293	0.655956	0.6928	w	263292	2.88175	0.2277
33	e	e	263282	1.69514	0.555	w	263282	3.22829	0.2579
34	e	*r	263287	1.19922	0.3004	e	263287	1.26536	0.2906
35	e	e	263312	3.15163	0.7894	w	263310	6.06688	0.1838
36	e	*r	263294	0.766822	0.3471	w	263294	1.61339	0.2273
37	e	e	263277	2.24763	0.6094	w	263276	4.127	0.2381
38	e	e	263306	1.05394	0.6859	w	263305	3.13326	0.2425
39	e	e	263276	8.59582	0.4498	w	263276	9.66047	0.2641
40	e	e	263288	1.3139	0.6103	w	263288	3.05912	0.255
41	e	e	263322	1.29199	0.4453	w	263321	1.78103	0.3487
42	e	e	263287	2.3495	0.3946	w	263286	3.22796	0.2543
43	e	e	263289	1.80866	0.3141	r	263289	2.08926	0.273
44	e	*w	263335	3.603	0.3899	e	263335	3.86512	0.342
45	e	e	263322	1.51194	0.6756	w	263321	3.39272	0.2638
46	e	*r	263285	4.44316	0.401	w	263284	5.51328	0.2348
47	e	*w	263323	2.26932	0.3534	e	263323	2.27222	0.3529
48	e	e	263326	1.7479	0.4799	w	263326	2.37626	0.3505
49	e	*r	263290	2.23961	0.3633	e	263289	2.92666	0.2577
50	e	e	263287	1.19163	0.7155	w	263286	3.6258	0.2118
51	e	e	263307	0.585495	0.5783	w	263307	1.96943	0.2895
52	e	*r	263292	1.25049	0.3355	h	263292	1.70278	0.2676
53	e	*r	263310	1.33908	0.4061	w	263309	2.71064	0.2045
54	e	e	263288	1.05333	0.7146	w	263287	3.48134	0.2122

Row	Actual Group	Highest Group	Highest Value	Squared Distance	Prob.	2nd Highest Group	2nd Highest Value	Squared Distance	Prob.
55	e	*r	263296	1.45779	0.3339	e	263296	1.72554	0.292
56	e	*r	263298	2.5653	0.3045	w	263298	3.00824	0.244
57	e	e	263293	0.630536	0.569	w	263292	2.14739	0.2665
58	e	*r	263265	3.98211	0.3049	e	263265	4.13897	0.2819
59	e	e	263307	0.428474	0.5724	w	263306	1.75938	0.2942
60	e	e	263278	3.52538	0.3759	r	263277	4.0931	0.283
61	e	e	263298	2.05163	0.3766	w	263298	2.48003	0.304
62	e	e	263305	0.242478	0.4423	w	263304	0.935955	0.3127
63	e	e	263312	1.41194	0.3452	w	263312	1.54317	0.3233
64	e	e	263315	2.39604	0.7629	w	263313	5.03633	0.2038
65	e	*r	263318	2.55336	0.4871	w	263317	4.06385	0.2289
66	e	e	263319	0.79105	0.5421	w	263319	1.8374	0.3213
67	e	e	263305	0.194699	0.4709	w	263305	1.02627	0.3107
68	e	e	263289	0.77199	0.424	w	263288	1.60482	0.2796
69	e	*r	263290	2.38	0.3022	w	263290	3.14839	0.2058
70	e	e	263295	1.78155	0.4839	w	263295	2.76725	0.2956
71	e	e	263301	2.26304	0.5289	w	263300	3.49411	0.2858
72	e	e	263300	0.0594097	0.4986	w	263299	1.09936	0.2964
73	e	e	263327	1.51277	0.6014	w	263326	2.84391	0.3091
74	e	*r	263293	0.933136	0.3556	w	263292	1.61701	0.2526
75	e	e	263313	2.54073	0.7838	w	263312	5.39063	0.1885
76	e	e	263311	0.55169	0.4714	w	263311	1.31297	0.3222
77	e	e	263314	0.556485	0.5101	w	263313	1.4694	0.3232
78	e	e	263335	2.65766	0.566	w	263334	3.7089	0.3346
79	e	*r	263254	10.3034	0.346	e	263253	11.066	0.2363
80	e	*r	263314	2.5466	0.4736	w	263314	3.8423	0.2478
81	e	*r	263303	2.51479	0.3309	w	263302	3.19855	0.2351
82	e	e	263313	1.51124	0.4969	w	263312	2.40471	0.3179
83	e	*r	263337	5.57066	0.5084	w	263336	7.4	0.2037
84	e	e	263318	1.23402	0.6295	w	263317	2.83123	0.2833
85	e	e	263328	1.98278	0.6008	w	263327	3.30775	0.3097
86	e	e	263270	4.96895	0.6935	w	263269	7.41056	0.2046
87	e	e	263304	0.454412	0.5703	w	263303	1.83712	0.2857
88	e	e	263303	0.317217	0.4497	w	263303	1.07604	0.3077
89	e	e	263306	1.43259	0.7054	w	263305	3.65	0.2328
90	e	*r	263307	1.01706	0.373	w	263306	1.77899	0.2549
91	e	e	263282	1.632	0.6436	w	263281	3.65068	0.2346
92	e	e	263299	0.259963	0.5343	w	263299	1.50335	0.287
93	e	e	263270	7.14301	0.6126	w	263269	9.10526	0.2297
94	e	e	263313	2.24836	0.3934	w	263313	2.54358	0.3394
95	e	e	263309	0.908499	0.6704	w	263308	2.85096	0.2538
96	e	*h	263286	1.26569	0.4013	r	263285	1.70775	0.3217
97	e	*r	263344	5.70577	0.3943	w	263344	5.85367	0.3662
98	e	e	263302	1.51455	0.6587	w	263301	3.4484	0.2505
99	e	e	263290	0.781306	0.3825	w	263290	1.40453	0.2801
100	e	e	263280	3.0155	0.6984	w	263279	5.38804	0.2133
101	e	e	263310	0.303126	0.5281	w	263310	1.35291	0.3124
102	e	e	263309	0.750599	0.5432	w	263308	1.92519	0.3019
103	e	e	263297	1.41181	0.6171	w	263297	3.09412	0.2661
104	e	e	263273	5.0622	0.7285	w	263271	7.71463	0.1934
105	e	e	263311	0.437637	0.603	w	263311	1.90785	0.2891
106	e	e	263319	1.84284	0.5494	w	263318	2.94837	0.3161
107	e	*r	263296	1.68545	0.44	w	263295	3.39992	0.1867
108	e	*r	263286	0.875928	0.3475	w	263286	1.86788	0.2116
109	e	e	263265	3.51752	0.3167	r	263265	3.78759	0.2767
110	e	e	263313	2.36718	0.3414	w	263313	2.49253	0.3207
111	e	e	263334	2.9092	0.639	w	263333	4.44596	0.2964
112	e	e	263323	1.60184	0.6474	w	263322	3.27302	0.2807
113	e	e	263313	3.03356	0.769	w	263312	5.73921	0.1988
114	e	e	263285	1.43349	0.6271	w	263284	3.31531	0.2447
115	e	e	263314	3.68124	0.8099	w	263313	6.80732	0.1697
116	e	e	263274	4.63719	0.7651	w	263272	7.5658	0.1769
117	e	e	263321	2.25078	0.7364	w	263320	4.61735	0.2255

Row	Actual Group	Highest Group	Highest Value	Squared Distance	Prob.	2nd Highest Group	2nd Highest Value	Squared Distance	Prob.
118	e	e	263327	2.08234	0.6515	w	263326	3.75474	0.2823
119	e	e	263284	1.08144	0.6586	w	263283	3.16669	0.2322
120	e	e	263295	1.4689	0.7596	w	263294	4.18787	0.1951
121	e	e	263292	2.66088	0.5316	w	263291	3.9277	0.2822
122	e	e	263295	1.4689	0.7596	w	263294	4.18787	0.1951
123	e	e	263280	4.33962	0.6461	w	263279	6.37089	0.234
124	e	e	263279	1.66312	0.5999	w	263279	3.47097	0.243
125	e	e	263315	0.51406	0.6389	w	263314	2.18896	0.2765
126	e	e	263272	2.88592	0.4742	w	263272	4.2058	0.2451
127	e	e	263296	0.876421	0.6251	w	263295	2.62408	0.2609
128	e	e	263310	0.643232	0.5546	w	263310	1.83727	0.3053
129	e	e	263299	0.556122	0.6216	w	263298	2.2853	0.2618
130	e	e	263285	1.58973	0.7231	w	263284	4.09542	0.2066
131	e	e	263296	1.60058	0.3508	w	263296	2.06845	0.2777
132	e	e	263241	20.3701	0.5256	r	263240	22.1704	0.2137
133	e	e	263298	0.704482	0.7102	w	263297	3.00424	0.2249
134	e	e	263293	0.756636	0.7127	w	263292	3.12052	0.2186
135	e	e	263329	1.84379	0.4881	w	263329	2.47955	0.3552
136	e	e	263335	2.74903	0.4855	w	263335	3.31449	0.366
137	e	e	263315	0.687011	0.474	w	263315	1.39033	0.3335
138	e	e	263282	2.298	0.6904	w	263281	4.59903	0.2185
139	e	e	263286	0.983223	0.3653	w	263286	1.66441	0.2599
140	e	e	263294	0.797287	0.5838	w	263293	2.33412	0.2707
141	e	e	263301	0.237512	0.4303	w	263300	0.929764	0.3044
142	e	e	263328	2.27045	0.6224	w	263328	3.73325	0.2995
143	e	e	263290	1.59418	0.5547	w	263289	3.10364	0.2608
144	e	e	263331	2.63401	0.6552	w	263330	4.30764	0.2838
145	e	e	263309	0.575169	0.5166	w	263308	1.61167	0.3077
146	e	e	263312	1.54306	0.7204	w	263311	3.83443	0.2291
147	e	*w	263308	2.05838	0.3004	e	263308	2.07262	0.2983
148	e	*r	263289	2.50691	0.3842	w	263289	3.39692	0.2462
149	e	e	263267	7.84914	0.5339	w	263266	9.46247	0.2383
150	e	e	263286	1.6261	0.7416	w	263285	4.26988	0.1977
151	e	*r	263326	2.89343	0.4008	w	263326	3.44268	0.3046
152	e	e	263291	0.5681	0.4781	w	263291	1.61176	0.2837
153	e	e	263320	1.494	0.7022	w	263319	3.58439	0.2469
154	e	e	263339	3.36943	0.4919	w	263339	3.93612	0.3705
155	e	e	263280	1.18674	0.5697	w	263279	2.85034	0.248
156	e	e	263317	2.90105	0.7857	w	263316	5.75218	0.1889
157	e	*r	263309	1.14256	0.3115	w	263309	1.15814	0.3091
158	e	e	263297	1.00421	0.4134	w	263297	1.63606	0.3014
159	e	*r	263320	2.00448	0.3552	w	263320	2.15024	0.3303
160	e	e	263265	4.21504	0.4395	r	263264	5.46907	0.2348
161	e	e	263314	1.49222	0.4482	w	263314	2.06361	0.3368
162	e	e	263287	0.624206	0.5635	w	263286	2.15243	0.2624
163	e	e	263290	1.52025	0.7625	w	263289	4.30334	0.1896
164	e	e	263321	3.92416	0.8004	w	263320	6.9146	0.1795
165	e	e	263321	2.25078	0.7364	w	263320	4.61735	0.2255
166	e	e	263289	2.07229	0.7631	w	263288	4.86155	0.1892
167	e	e	263301	1.96695	0.4293	w	263301	2.7125	0.2957
168	e	e	263302	2.19225	0.8052	w	263300	5.32082	0.1685
169	e	*r	263303	0.739077	0.3538	w	263303	1.28623	0.2691
170	e	e	263289	0.626525	0.4573	w	263289	1.65663	0.2732
171	e	e	263317	0.620179	0.5765	w	263316	1.88202	0.3068
172	e	e	263292	0.629164	0.6494	w	263291	2.57878	0.245
173	e	e	263309	0.910641	0.7099	w	263308	3.13544	0.2334
174	e	e	263297	0.674076	0.67	w	263296	2.69766	0.2436
175	e	e	263309	0.575169	0.5166	w	263308	1.61167	0.3077
176	e	*w	263323	2.00753	0.355	e	263323	2.20798	0.3211
177	e	*r	263288	1.98753	0.3037	e	263288	2.70256	0.2124
178	e	e	263304	0.303498	0.6662	w	263303	2.2509	0.2516
179	e	*r	263308	0.915298	0.3568	w	263308	1.32576	0.2906
180	e	e	263330	2.04731	0.6143	w	263329	3.44502	0.3054

Row	Actual Group	Highest Group	Highest Value	Squared Distance	Prob.	2nd Highest Group	2nd Highest Value	Squared Distance	Prob.
181	e	*w	263322	2.14979	0.3496	r	263322	2.37287	0.3127
182	e	e	263297	0.553969	0.6979	w	263295	2.78099	0.2292
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
913	w	w	262297	6.8555	0.4844	r	262297	7.2855	0.3907
914	w	*e	263300	0.558083	0.3707	w	263299	0.974156	0.3011
915	w	*e	263307	0.538307	0.3961	w	263307	0.95683	0.3213
916	w	w	262251	1.60299	0.521	e	262251	2.41744	0.3467
917	w	*e	263295	1.12119	0.5861	w	263294	2.70489	0.2655
918	w	*e	263306	0.352091	0.4347	w	263306	0.980218	0.3175
919	w	*r	263292	0.708351	0.3588	h	263291	1.60071	0.2297
920	w	w	262259	1.96504	0.4916	r	262258	2.87881	0.3113
921	w	*h	263269	1.95497	0.3553	r	263269	2.34554	0.2922
922	w	*g	261153	4.39359	0.8402	r	261151	8.78819	0.0933
923	w	w	262249	3.33382	0.4842	r	262248	4.30831	0.2975
924	w	w	262205	9.53868	0.4045	e	262204	10.1388	0.2996
925	w	w	261188	4.05228	0.4631	g	261187	5.05645	0.2803
926	w	*r	263291	7.06487	0.4345	h	263290	7.83502	0.2957
927	w	w	261189	5.66834	0.6711	r	261188	9.01728	0.1258
928	w	*e	263297	2.55968	0.3473	w	263297	2.85	0.3004
929	w	w	262218	4.30627	0.4042	r	262217	5.1983	0.2588
930	w	*g	260116	14.9224	0.4453	w	260116	15.0015	0.4281
931	w	*r	263318	2.50473	0.4484	w	263318	3.44026	0.2809
932	w	w	262207	6.26982	0.2986	r	262207	6.2701	0.2985
933	w	w	261169	7.16851	0.5817	g	261168	9.81915	0.1546
934	w	w	261190	4.86953	0.5106	g	261189	6.35082	0.2435
935	w	w	262267	1.34679	0.5714	e	262266	2.91804	0.2605
936	w	*r	262252	1.20763	0.3608	g	262252	1.33978	0.3377
937	w	w	262232	3.67251	0.3179	r	262232	3.98965	0.2713
938	w	*r	263319	1.75623	0.3467	w	263319	1.87595	0.3265
939	w	*r	263349	8.00885	0.5114	w	263348	9.01594	0.3091
940	w	w	262262	1.93686	0.4361	r	262262	2.54264	0.3222
941	w	*r	263275	3.01569	0.347	e	263275	3.68749	0.248
942	w	*e	263317	0.889507	0.4374	w	263317	1.38322	0.3417
943	w	*e	263318	0.678704	0.6223	w	263317	2.217	0.2884
944	w	*h	263283	0.805881	0.4234	r	263283	1.67768	0.2738
945	w	w	261206	4.29759	0.4637	r	261206	5.49854	0.2544
946	w	*e	263270	2.39781	0.4854	w	263269	3.84839	0.235
947	w	*e	263308	0.282265	0.4829	w	263308	1.12122	0.3174
948	w	w	262243	1.18312	0.4351	r	262242	2.15117	0.2681
949	w	*e	263301	1.01974	0.4638	w	263301	1.90927	0.2973
950	w	*e	263297	0.385144	0.6306	w	263296	2.18556	0.2563
951	w	*e	263287	1.29586	0.5649	w	263287	2.87684	0.2563
952	w	w	261219	5.00107	0.6165	r	261218	7.08509	0.2175
953	w	*e	263292	0.609023	0.5655	w	263291	2.07312	0.272
954	w	w	261202	5.30642	0.6736	r	261201	8.36247	0.1461
955	w	w	262266	1.77418	0.5669	e	262265	3.28624	0.2662
956	w	*r	263320	2.49712	0.3555	w	263320	2.77949	0.3087
957	w	*e	263316	1.64346	0.3576	w	263316	1.77593	0.3347
958	w	w	262258	2.77047	0.4559	r	262257	3.28935	0.3517
959	w	w	262243	0.836017	0.4542	r	262243	1.71723	0.2923
960	w	*e	263318	2.32255	0.7162	w	263317	4.54363	0.2359
961	w	*r	263312	1.91358	0.3948	w	263312	3.09637	0.2185
962	w	*r	262251	0.778578	0.3647	w	262251	1.08196	0.3134
963	w	w	263352	6.76886	0.4265	e	263352	7.51776	0.2933
964	w	w	261182	4.89619	0.5059	g	261181	6.32623	0.2475
965	w	w	262257	1.3866	0.5487	e	262256	3.0925	0.2338
966	w	*r	263299	1.26018	0.3104	e	263298	1.49431	0.2761
967	w	*e	263294	1.61637	0.5492	w	263294	2.94976	0.282
968	w	*e	263300	2.19279	0.8108	w	263298	5.39287	0.1637
969	w	*r	263337	4.98619	0.4929	w	263337	6.24057	0.2632
970	w	w	263338	4.68794	0.3891	r	263338	5.14101	0.3102

Row	Actual Group	Highest Group	Highest Value	Squared Distance	Prob.	2nd Highest Group	2nd Highest Value	Squared Distance	Prob.
971	w	*e	263288	0.583037	0.499	w	263287	1.82275	0.2685
972	w	*e	263339	3.49047	0.5158	w	263339	4.19525	0.3626
973	w	w	262258	1.5635	0.3871	r	262257	1.80911	0.3423
974	w	*r	263273	2.25524	0.2879	h	263273	2.39444	0.2685
975	w	w	261210	4.89524	0.596	r	261209	6.81651	0.2281
976	w	*h	263297	2.29402	0.3938	r	263297	2.87603	0.2943
977	w	*e	263314	1.15481	0.3625	w	263314	1.2996	0.3372
978	w	w	262228	2.07541	0.3517	r	262228	2.58998	0.2719
979	w	*r	263328	3.54526	0.3863	w	263328	3.78649	0.3424
980	w	*e	263310	0.303126	0.5281	w	263310	1.35291	0.3124
981	w	*r	263306	1.34984	0.315	w	263306	1.51078	0.2907
982	w	*r	263335	6.32257	0.5516	w	263334	8.00663	0.2377
983	w	*g	262224	1.15157	0.3611	r	262224	1.56673	0.2934
984	w	*e	263310	1.41283	0.4387	w	263310	1.983	0.3299
985	w	w	262242	2.4498	0.3368	r	262242	2.66287	0.3028
986	w	*g	261184	4.78087	0.5328	w	261183	6.24223	0.2566
987	w	w	262239	2.03947	0.495	e	262239	2.71362	0.3533
988	w	*r	262245	2.55275	0.4008	g	262245	2.86909	0.3422
989	w	*e	263257	7.09848	0.5506	w	263256	8.97815	0.2151
990	w	w	262266	1.07911	0.5679	e	262265	2.9327	0.2248
991	w	*e	263285	1.61155	0.7033	w	263284	3.97915	0.2153
992	w	*e	263299	3.50479	0.3405	r	263298	3.71411	0.3067
993	w	w	262249	0.284409	0.4861	r	262248	1.54317	0.259
994	w	*e	263275	4.11977	0.7209	w	263273	6.70068	0.1984
995	w	*e	263316	1.70664	0.6374	w	263315	3.3692	0.2776
996	w	w	262275	2.33474	0.4541	r	262274	2.78425	0.3627
997	w	*e	263311	0.391118	0.6103	w	263310	1.92053	0.2841
998	w	w	262261	4.10211	0.5076	e	262261	4.506	0.4148
999	w	w	261222	5.40593	0.7207	r	261220	8.4337	0.1586

* = incorrectly classified.

References

[1] Büyüköztürk, S. and Çokluk-Bökeoğlu, O. (2008) "Discriminant Function Analysis: Concept and Application", Eurasian Journal of Educational Research, 33, PP. 73-74.

[2] Alexakos, C. E. (1966) "Predictive Efficiency of two Multivariate Statistical Techniques in Comparison with Clinical Predictions", Journal of Educational Psychology, 57, PP. 297-306.

[3] Chastian, K. (1969) "Prediction of Success in Audio-lingual and Cognitive Classes", Lan-guage Learning, 19, PP. 27-39.

[4] Stahmann, R. F. (1969) "Predicting Graduation Major Field from Freshman Entrance Data", Journal of Counseling Psychology, Vol. 16, PP. 109-113.

[5] Anderson, G. J., Walberg, H. J., and Welch, W. W. (1969) "Curriculum Effects on the Social Climate of Learning: A new Representation of Discriminant Functions", American Educational Research Journal, No. 6, PP. 315-328.

[6] Saupe, J. L. (1965) "Factorial-design Multiple Discriminant Analysis: A description and An illustration", American Educational Research Journal, Vol. 2, PP. 175-184.

[7] Tatsuoka, M. M., and Tiedeman, D. V. (1954) "Discriminant Analysis. Review of Educational Research", No. 24, PP. 402-420.

[8] Fisher, R. A. (1936) "The Use of Multiple Measurements in Taxonomic Problems", Annals of Eugenics, 7.

[9] Fisher, R. A. (1938) "The Statistical Utilization of Multiple Measurements", Annals of Eugenics, Vol. 8, PP. 376-386.

[10] Solovyev, V. and Salamov, A. (1997) "The Gene-Finder Computer Tools for Analysis of Human and Model Organisms Genome Sequences", Department of Cell Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX, American Association for Artificial Intelligence (www.aaai.org), PP. 294- 302.

[11] Ghosh, D. (1993). "Status of the Transcription Factors Database (TFD)". Nucl. Acids Res., Vol. 21, PP. 3117-3118.

[12] Wingender, E. (1994) "Recognition of Regulatory Regions Genomic Sequences", J. Biotechnol. 35, PP. 273-280.

[13] Zhang, M. Q. (2000) "Discriminant Analysis and its Application in DNA Sequence Motif Recognition", Henry Stewart Publications 1467-5463, Briefings in Bionformatics, Vol. 1, No. 4.

[14] Dudoit, S., Fridlyand, J. and Speed, T. P. (2000) "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data", Department of Statistics, University of California, Berkeley, Berkeley, CA 94720-3860, sandrine@stat.berkeley.edu, PP. 1-43.

[15] Kwon, S., Chu, Y. H., Yi, H. S. and Han, C. (2001) "DNA Microarray Data Analysis for Cancer Classification Based on Stepwise Discriminant Analysis and Bayesian Decision Theory", Genome Informatics 12, PP. 252-254.

[16] Liu, Z. H., Jiao, D. and Sun, X. (2005) "Classifying Genomic Sequences by Sequence Feature Analysis", State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China, Geno. Prot. Bioinfo., Vol. 3, No. 4, PP. 201-205.

[17] Guo, Y., Hastie, T. and Tibshirani, R. (2005) "Regularized Discriminant Analysis and Its Application in Microarrays", Printed in Great Britain, Biostatistics, Vol. 1, No. 1, PP. 1–18.

[18] Jombart, T., Devillard, S. and Balloux, F. (2010) "Discriminant Analysis of Principal Components: A new Method for the Analysis of Genetically Structured Populations", Jombart et al. BMC Genetics, 11:94, http://www.biomedcentral.com/1471-2156/11/94, PP. 1-15.

[19] Jin, J. and An, J. (2011) "Robust Discriminant Analysis and its Application to Identify Protein Coding Regions of Rice Genes", Contents lists available at Science Direct, Mathematical Biosciences, journal homepage: www.elsevier.com/locate/mbs

[20] Libbrecht, M. W. and Noble, W. S. (2015) "Machine Learning Applications in Genetics and Genomics", Nature Reviews | Genetics, Vol. 16, PP. 321-332.

[21] Corvelo, A., Clarke, W. E., Robine, N. and Zody, M. C. (2018) "TaxMaps: Comprehensive and Highly Accurate Taxonomic Classification of Short-read Data in Reasonable Time", New York Genome Center, New York 10013, USA, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/18; www.genome.org, Genome Research, Vol. 28, PP. 751–758.

[22] Polovinkina, A., Krylova, I., Druzhkova, P., Ivanchenkoa, M., Meyerova, I., Zaikina, A., and Zolotykha, N. (2016) "Solving Problems of Clustering and Classification of Cancer Diseases Based on DNA Methylation", Data Pattern Recognition and Image Analysis, Vol. 26, No. 1, PP. 176–180.

[23] Waterman, M. and Vingron, M. (1994) "Sequence Comparison Significance and Poisson Approximation", Stat. Sci., Vol. 9, PP. 367–381.

[24] McLachlan, A. and Stewart, M. (1976) "The 14-fold Periodicity in Alpha-Tropomyosin and the Interaction with Actin ", J. Mol. Biol., Vol. 103, PP. 271–298.

[25] Eisenberg, D., Weiss, R. M., Terwillger, T. C., (1994) "The Hydrophobic Moment Detects Periodicity in Protein Hydrophobicity". Proc. Natl. Acad. Sci., Vol. 81, PP. 140–144.

[26] Stoffer, D. (2012) "Frequency Domain Techniques in the Analysis of DNA Sequences", Handbook of Statistics Volume 30, PP. 261-295.

[27] Tavar´e, S., Giddings, B. (1989) "Some Statistical Aspects of the Primary Structure of Nucleotide Sequences", In Waterman M. S. (Ed), Mathematical Methods for DNA Sequences. CRC Press, Boca Raton, Florida, PP. 117–131.

[28] Viari, A., Soldano, H. and Ollivier, E. (1990) "A Scale-independent Signal Processing Method for Sequence Analysis. Comput. Appl. Biosci., Vol. 6, PP. 71–80.

[29] Marhon, S. and Kremer, S. (2011) "Gene Prediction Based on DNA Spectral Analysis: A literature Review", J Comput Biol., Apr, Vol. 18, No. 4, 639-76.

[30] Bajic, V., Bajic, I. and Hide, W. (2000) "A new Method of Spectral Analysis of DNA/RNA and Protein sequences" Centre for Engineering Research.

[31] Han, Y., Han, L., Yao, Y., Li, Y. and Liu, X. (2018) "Key Factors in FTIR Spectroscopic Analysis of DNA: The Sampling Technique, Pretreatment Temperature and Sample Concentration", Analytical Methods, Issue Vol. 21, No. 10, PP. 2436-2443.

[32] Ruiz, G., Israel, Godínez, I., Ramos, S., Ruiz, S., Pérez, H. and Morales, J. (2018) "Genomic Signal Processing for DNA Sequence Clustering" PeerJ v. 6; DOI 10.7717/peerj.4264.

[33] Hoang, T., Yin, C., Zheng, H. Yu, C., Lucy He, R. and Yau, S. (2015) "A new Method to Cluster DNA Sequences Using Fourier Power Spectrum", J Theor Biol. 7; 372:135-45.

[34] Mabrouk, M. (2017) "Advanced Genomic Signal Processing Methods in DNA Mapping Schemes for Gene Prediction Using Digital Filters", American Journal of Signal Processing, Vol. 7, No. 1, PP. 12-24.

[35] Roy, M. and Barman, S. (2011) "Spectral Analysis of Coding and Non-coding Regions of a DNA Sequence by Parametric and Nonparametric Methods: A comparative Approach", Annals of Faculty Engineering Hunedoara– International Journal Of Engineering; Tome IX; Faccicule 3; PP. 57-62.

[36] Galleani, L. and Garello, R. (2006) "Spectral Analysis of DNA Sequences by Entropy Minimization", 14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September, PP. 4-8.

[37] Stoffer, D., Tyler, D. and McDougall, A. (1993) "Spectral Analysis for Categorical Time Series: Scaling and the Spectral Envelope"; Biometrika, Vol. 80, PP. 611–622.

[38] Stankovičová, I. and Vojtková, M. (2007) "Viacrozmerné Statistické Metódy s Aplikáciami", Bratislava, Iura Edition.

[39] Kočišová, K. and Mišanková, M. (2013) "Discriminant Analysis as A tool for Forecasting Company´s Financial Health", Contemporary Issues in Business, Management and Education, University of Žilina, Faculty of Operation and Economics of Transport and Communications, Department of Economics, Procedia - Social and Behavioral Sciences 110, PP. 1148-1157.

[40] Meloun, M., Militký, J., and Hill, M. (2005) "Počítačová Analýza Vícerozměrných Dát v Príkladech", Praha: Academia.

[41] Muhameed, A. S. and Saleh, A. M. (2014) "Classification of Some Iraqi Soils Using Discriminant Analysis", Dept. of Soil Sci. and Water Res. Agric. College – Univ. of Baghdad, IOSR Journal of Agriculture and Ve terinary Science (IOSR-JAVS), e-ISSN: 2319-2380, p-ISSN: 2319-2372., Vol. 7, Issue 1 Ver.

[42] Ayinla, A. S. and Adekunle, B. K. (2015) "An Overview and Application of Discriminant Analysis in Data Analysis", IOSR Journal of Mathematics (IOSR-JM), e-ISSN: 2278-5728, p-ISSN: 2319-765X, Volume 11, Issue 1 Ver. V, PP. 12-15.

[43] Härdle, W. K. and Simar, L. (2012) "Applied Multivariate Statistical Analysis", Sixth Edition. Copyrighted Material.