ANALYZING DNA SEQUENCES USING CLUSTERING ALGORITHM

TAHA TALEB RAGHEB ALHERSH

UNIVERSITY UTARA MALAYSIA
2009

ANALYZING DNA SEQUENCES USING CLUSTERING ALGORITHM

A thesis submitted to college Arts & Sciences
in partial fulfillment of the requirement for the degree
Master of Science (Intelligent Systems)
University of Utara Malaysia

By
Taha Alhersh

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Master of Science in IT degree from University Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor or, in their absence by the Academic Dean College of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to University Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to

Dean (Academic) College of Art and Sciences
University Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman.

ABSTRACT

Data mining gives a bright prospective in DNA sequences analysis through its concepts and techniques. This study carries out exploratory data analysis method to cluster DNA sequences. Feature vectors have been developed to map the DNA sequences to a twelve-dimensional vector in the space. Lysozyme, Myoglobin and Rhodopsin protein families have been tested in this space. The results of DNA sequences comparison among homologous sequences give close distances between their characterization vectors which are easily distinguishable from non-homologous in experiment it with a fixed DNA sequence size that does not exceed the maximum length of the shortest DNA sequence. Global comparison for multiple DNA sequences simultaneously presented in the genomic space is the main advantage of this work by applying direct comparison of the corresponding characteristic vectors distances. The novelty of this work is that for the new DNA sequence, there is no need to compare the new DNA sequence with the whole DNA sequences length, just the comparison focused on a fixed number of all the sequences in a way that does not exceed the maximum length of the new DNA sequence. In other words, parts of the DNA sequence can identify the functionality of the DNA sequence, and make it clustered with its family members.

ACKNOWLEDGEMENT

I would like to start with the words that any job will not be complete without, so I will say: “In the Name of Allah, the Beneficent, the Most Merciful”. All the thanks to Allah that pave the way for me to obtain my master degree.

My sincere appreciation goes to my supervisor AP Fadzilah Bt Siraj for her supervision, guidance, advice, knowledge and word of encouragement during this study, I’ll always be thankful to you, Terimah kasih!

Special thanks to all my lecturers Dr. Ahmad Otoom, Dr. Yuhanis binti Yusof, Dr. Norita Md Norwawi, Dr. Fawziah Ahmad, Mrs. Nur Azzah Abu Bakar, Dr. Shaidah Jusoh, Mr. Zhamri Che Ani and other UUM staff.

I must acknowledge the immeasurable contributions of my friends and collegues who have shown great love and care during my study especially Qais, Abdullateef, Hossam, Ali, big Thank you!

This acknowledgment won’t be complete without my family. Profound gratitude goes to my parents Taleb and Shifa who have never failed to give me the best in life. May Allah reward your efforts! To my brothers Faisal, Firas, Fadi, Abdullrahman and Abdullraheem, you never stopped in encouraging me, special thanks and appreciation to my brother Fadi for his continuous support. My family, you will be always in my heart, love you all.

DEDICATION

To my parents Taleb and Shifa, and to my brothers.

TABLE OF CONTENT

PERMISSION TO USE i

ACKNOWLEDGEMENT iii

DEDICATION iv

TABLE OF CONTENT v

LIST OF TABLES viii

LIST OF FIGURES ix

LIST OF ABBREVIATION xi

LIST OF APPENDICES xiii

INTRODUCTION 1

1.1 Problem Statement 5

1.2 Research Question...6

1.3 Research Objectives 6

1.4 Scope and Limitation 6

1.5 Chapters Overview..7

LITERATURE REVIEW...8

2.1 Data Mining 8

2.1.1 Predictive Model 9

2.1.2 Descriptive Model...10

2.2 Clustering Techniques...11

2.2.1 Hierarchical Clustering Algorithms 14

2.2.2 Partitional Algorithms..16

2.2.3 Graph-Theoretic Clustering 18

2.2.4 Expectation-Maximization (EM) Algorithm..20

2.2.5 Fuzzy C-Means Clustering Algorithm 21

2.2.6 Spectral Clustering...23

2.2.7 Kohonen Networks...24

2.3 Applications of Clustering 26

2.3.1 Bioinformatics..26

2.3.2 Image Segmentation...30

2.3.3 Information Retrieval..31

2.4 General Taxonomy of DNA Sequences..32

2.5 DNA Sequences Representations...33

2.5.1 Graphical Representation of DNA Sequences 34

2.5.2 Numerical Representation of DNA Sequences...37

2.6 Conclusion 38

METHODOLOGY..39

3.1 Introduction...39

3.2 Research Methodology..40

RESULTS AND DISCUSSION 48

4.1 DNA Sequences (Data)...49

4.2 The Experiment..51

4.2.1 Whole Size Sequence Experiment 54

4.2.2 Fixed Size Sequence Experiment..55

4.3 Discussion 58

CONCLUSION AND FUTURE WORK 62

5.1 Conclusion 62

5.2 Future Work 63

REFERENCES..64

Appendix..69

LIST OF TABLES

Table 3.1: Number of nucleotides (A, T, C, G) in the sequence GTGGGTGGTT 41

Table 3.2: Total distance of nucleotides (A, T, C, G) in the sequence GTGGGTGGTT 43

Table 3.3: The distribution of nucleotides (A, T, C, G) in the sequence GTGGGTGGTT 44

Table 4.1: Number of Nucleotides in each family member 50

Table 4.2: Whole sequence comparison results 54

Table 4.3: Results for the first 100 nucleotides 55

Table 4.4: Results for the first 200 nucleotides 56

Table 4.5: Results for the first 300 nucleotides 56

Table 4.6: Results for the first 400 nucleotides 57

Table 4.7: Results for the first 500 nucleotides 58

Table 4.8: Distances inside and outside families in different experiments 59

LIST OF FIGURES

Fig. 1.1: Gene bank growth 2

Fig. 1.2: DNA 3

Fig. 2.1: Clustering Technique 12

Fig. 2.2: Hierarchical Clusters 13

Fig. 2.3: Overlapping Clusters 13

Fig. 2.4: Taxonomy of clustering approaches 13

Fig. 2.5: Minimal spanning tree for clustering 19

Fig. 2.6: : General Taxonomy for Species 32

Fig. 2.7: characteristic curve of the sequence TGGTGCACCTGACTCCTGA 35

Fig 2.8: Graphical representation of the sequence ATGGTGCACC 36

Fig.2.9: A weak-H/strong-H bond graph 37

Fig.3.1: Methodology 40

Fig. 3.2: Original DNA sequence 41

Fig.3.3: The position of nucleotide in the sequence GTGGGTGGTT 42

Fig 3.4: : Bos taurus DNA sequence before and after characterization 45

Fig. 4.1: The total number of Nucleotides in each sequence 51

Fig. 4.2: DNA sequences comparison program 52

Fig. 4.3: How feature vectors stored in the database 52

Fig. 4.4: Pivot table form for the output 53

Fig. 4.5: All experiment results 60

LIST OF ABBREVIATION

A	Adenine
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
C	Cytosine
COI	Cytochrome “C” Oxidase I
D_A	The distribution of A in the DNA sequence
D_C	The distribution of C in the DNA sequence
D_G	The distribution of G in the DNA sequence
DNA	Deoxyribonucleic Acid
D_T	The distribution of T in the DNA sequence
EM	Expectation-Maximization
FCM	5 Fuzzy C-Means
G	Guanine
GBS	Global Bio-identification System
ILP	Inductive Logic Programming
IR	Information Retrieval
KDD	Knowledge Discovery in Database
KNIES	Kohonen Incorporating Explicit Statistics
LCC	Library of Congress Classification
LVQ	Learning Vector Quantization
MST	Minimal Spanning Tree

n_A	Number of instances A in the DNA sequence
n_C	Number of instances C in the DNA sequence
NCBI	National Center for Biotechnology Information
n_G	Number of instances G in the DNA sequence
NLP	Natural Language Processing
n_T	Number of instances T in the DNA sequence
PEs	Processing Elements
RNA	Ribonucleic Acid
SOM	Self Organizing Map
SVMs	Support Vector Machines
T	Thymine
T_A	The total distances of A from the origin of DNA sequence
T_C	The total distances of C from the origin of DNA sequence
T_G	The total distances of G from the origin of DNA sequence
TIS	Translation Initiation Sites
TSP	Travelling Salesman Problem
TSPLIB	Travelling Salesman Problem Library
TSS	Transition Split Site
T_T	The total distances of T from the origin of DNA sequence
VQ	Vector Quantization

LIST OF APPENDICES

Appendix

CHAPTER ONE

INTRODUCTION

This chapter introduces a brief description of this study. A general overview of the field of this work, problem statement, the objective and the scope of this study has been presented.

In the last few decades the rapid development of technology reflects to the number of biological data which has been growing in an exponential curve, from Gene Bank (www.ncbi.nlm.nih.gov) site the growth falls down in Fig.1.1. GenBank in 1982 had only 606 sequences with 680,338 bp (base pairs). In year 1992, GenBank contained 78,608 sequences with 101,008,486 bp. By the end of year 2002, GenBank had 22,318,883 sequences with 28,507,990,166 bp. This number had almost doubled in only two years. By the end of year 2008, GenBank had 98,868,465 sequences with 99,116,431,942 bp. Efficient and highly computational tools are needed to analyze the massive amount of data that contains rich information.

Data mining is the science of extracting useful information from large data sets or databases. This new discipline lies at the intersection of statistics, machine learning, artificial intelligence and other areas. The tasks of data mining include exploratory data analysis, descriptive modeling, predictive modeling, patterns and rules recognition etc. Compared to the traditional data analysis methods, the concepts and tools of data mining provide new prospective in the analysis of huge amount of biological sequences. DNA sequences clustering have been an issue in clustering analysis.

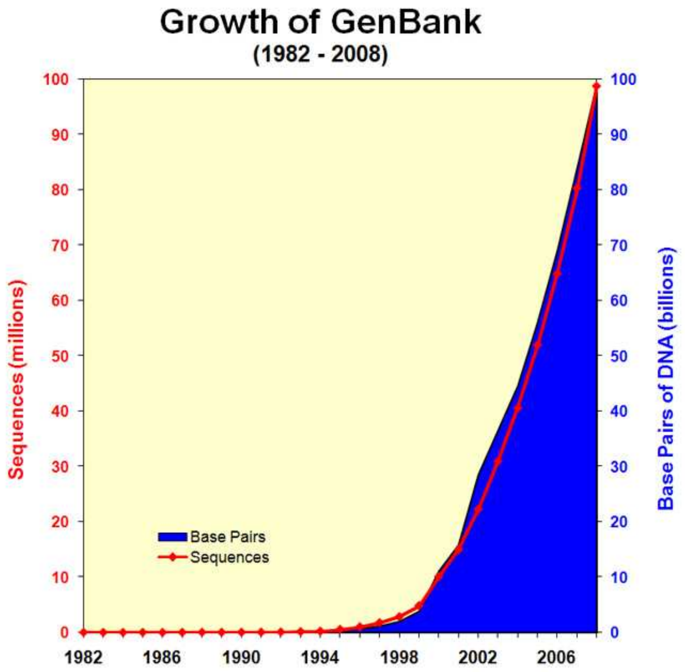


Figure 1.1: GeneBank source :
http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

A DNA is a long and un-branched polymer chain in double helix shape, composed of only four types of deoxyribonucleotides bases which are: adenine (A), cytosine(C), guanine (G), and thymine (T). The nucleotides are linked together by covalent phosphodiester bonds that join the 5' (five prime) carbon of one deoxyribose group to the 3’ (three prime) carbon of the next. The four kinds of bases are attached to this repetitive sugar-phosphate backbone chain as shown in Fig.1.2.

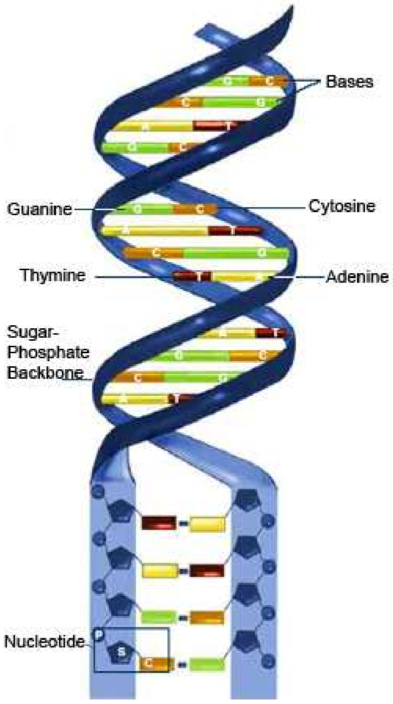


Figure 1.2: DNA
U.S. Department Of Health And Human Services (2006)

The two long chains of a DNA molecule are held together by complementary base pairs. Three hydrogen bonds form between G and C, and two hydrogen bonds exist between A and T. The base-pairing mechanism is the basis for DNA replication.

A DNA sequence or genetic sequence is a succession of letters constituent nucleotides listed from the 5'- to 3'- terminus representing the primary structure of DNA molecule or strand, which hold the information as described by the central dogma of molecular biology. Prior to discussion on applications of DNA sequences, several terminologies related to DNA are defined. For example, Genome is a complete set of DNA (Deoxyribonucleic Acid) for an organisms, and the DNA arranged into 23 pairs of DNA molecule called chromosomes, and each chromosome contain many genes, DNA molecule has millions of bases or nucleotides, these nucleotides sequences or base sequences has the information of making proteins encoded in it. A nucleotide is made up of one phosphate group linked to a pentose sugar, which is linked to one of 4 types of nitrogenous organic bases symbolized by the four letters A, C, G, and T. The rules that govern the correspondence of the base/ nucleotide sequences for DNA and RNA (Ribonucleic Acid) to the amino acids or proteins are known as Genetic Code. Sequence Alignment is the process of locating regions that are equivalent to increase the similarity of these sequences.

Each strand in the DNA complement the other, so an adenine (A) on one strand is always facing a thymine (T) (and vice versa), and cytosine (C) is always facing a guanine (G). When the sequence of nucleotides along one strand is known, automatically the sequence on the other one can be deduced. The double strand in helix structure of DNA makes the definition of a DNA sequence vague. Despite the convention of reading the nucleotides from the 5’ end toward the 3’ end, writing down the top or the bottom sequence. For convince they are both equally valid sequences by turning this page upside down. Thus, at each location, a DNA molecule corresponds to two different sequences, related by this reverse-and complement operation.

Various researchers have worked on clustering DNA analysis, some focused on local similarity while others make it global. One of the techniques that have already been implemented is Spectral Clustering.

1.1 Problem Statement

Each DNA sequence has its own functions. Once biologists come up with a new sequence, it is important to compare it with the previous existing sequences to know its functionality and category. Some of the most popular and effective methods for comparing sequences are BLAST and FASTA, but these methods have weakness. Though there is an extended version of these tools to deal with multiple sequence alignments. The weaknesses are:

- It can compare just two sequences at the same time and provide the similarity between them.
- It uses alpha representation of the sequences, which will add more burdens on the system and take a lot of memory space.

The area of DNA research is still considered at an infant stage. Therefore there are many sub-areas in DNA research that can be explored. One such area to focus on is DNA sequences representation, and how information and knowledge could be extracted from these sequences. To uncover the hidden information within DNA sequences, data mining approach can be employed. For example Liu *et al*. (2006) used Euclidian distance between the corresponding characterizations of DNA sequences to make clustering. For the purpose of this study, clustering technique has been chosen to be used since the DNA sequences represent unsupervised type of data.

1.2 Research Question

The research questions can be formulated as:

(a) How to identify suitable numerical representation of DNA sequence?

(b) How to evaluate DNA sequences features using clustering techniques?

1.3 Research Objectives

The research objectives are specified as:

(a) To identify suitable numerical representation of DNA sequence.

(b) To evaluate DNA sequences features using clustering techniques.

1.4 Scope and Limitation

The scope of this study will focus on some DNA sequences from the following families (Lysozyme, Myoglobin and Rhodopsin), and data mining that will be used in this study only uses clustering technique. The limitation of this study that it is concerned in DNA sequences; this study can be extended to other families of DNA sequences and can be implemented on protein sequences.

1.5 Chapters Overview

This section will provide a general overview for each chapter. This study falls into five chapters; Introduction, Literature Review, Methodology, Results and Discussion and Conclusion.

By starting with the introduction, an overall idea of this study will be gathered in the readers' mind. Explaining some terminologies that have been used in this study will make it easy to understand this work.

From the literature (Chapter TWO), the main concept of data mining and it applications has been clarified specially in clustering DNA sequences. Chapter THREE presents the methodology that has been applied in this study, which has been adopted from Liu *et al.* (2006).

The experiments applied in this study can be found in Chapter FOUR, there is two main experiments; one used the whole DNA sequence to produce results of clustering, and the other one is to have a fixed number of the DNA sequence in a way does not exceed the maximum length of the shortest DNA sequence. A conclusion and future work are presented in Chapter FIVE.

CHAPTER TWO

LITERATURE REVIEW

This chapter presents a general view of the data mining and its clustering techniques as well as some general applications for clustering, and ideas from previous researchers on using clustering algorithms especially in Bioinformatics. For the purpose of clustering DNA sequences the taxonomy of DNA sequences and DNA sequences representations have been presented.

2.1 Data Mining

Data mining is one of the steps in Knowledge Discovery in Database (KDD) process that consists of applying data analysis and discovery algorithms to produce a particular enumeration of patterns (or models) over the whole data (Fayyad *et al.*, 1996). The desired outcome of data mining activities is to discover knowledge that is not explicit in the data, and to put that knowledge to use (Ayre, 2006). Data mining also can be defined as the process of selection, exploration, and modeling of large quantities of data to discover regulations or relations that are unknown with the purpose of obtaining clear and useful results. Data mining is divided into two models, predictive (Supervised) and descriptive (Unsupervised) models.

2.1.1 Predictive Model

This model describes one or more dependent variables that are related to all of the independent variables; asymmetrical or direct methods can be assigned to the predictive models. This would be done by searching for rules of classification or prediction based on the data. Predictive modeling falls into category of supervised learning; hence, one variable is clearly labeled as Target variable Y and can be explained as a function of other variables X. By determining the nature of the target, classification model can be defined if Y is discrete variable, regression model, or continuous one. Typical methods of predictive modeling are classification, regression and time series analysis.

Classification, is the task of learning a target function f that maps each attribute set x to one of the predefined class labels y, also it can be defined as, assigning objects to one of several predefined categories (Tan *et al.*, 2006), Classification problems aim to identify the characteristics that indicate the group to which each case belongs

Regression is a predictive modeling technique using the value of one of a pair of correlated variables in order to predict the value of the second, where the target variable to be estimated is continuous, the goal of regression is to find a target function that can fit the input data with minimum error.

Time series forecasting predicts unknown future values based on a time-varying series of predictors. Like regression, it uses known results to guide its predictions. Models must take into account the distinctive properties of time, especially the hierarchy of periods (including such varied definitions as the five- or seven-day work week, the thirteen-“month” year, etc.), seasonality, calendar effects such as holidays, date arithmetic, and special considerations such as how much of the past is relevant.

2.1.2 Descriptive Model

Groups of data can be described more briefly in the descriptive models; these models can be named: symmetrical, unsupervised or indirect methods. A general description of the data is important but summaries also are important starting point and need more exploring. Models of data can be found through the descriptive models, so the aim is to describe not to predict. As a result, descriptive models are used in the setting of unsupervised learning. Typical methods of descriptive modeling are summarization, association rules, sequence discovery, and clustering. Data mining uses several types of analytical software such as statistical, machine learning, and neural network. In general, Classes are grouped into clusters, association rules, sequential patterns and summarization.

Clustering can be defined as the process of partitioning as set of data / (objects) in a set of consequential sub-classes, called clusters. Some of the data mining approaches which use clustering are database segmentation, predictive modeling, and visualization of large databases. Segmentation is a clustering method to segment databases into homogeneous groups, predictive modeling is statistical method of data analysis usually involves hypothesis testing of a model the analyst already has in mind. Visualization is the visualized representation of clusters in large databases in order to aid human analysts in identifying groups and subgroups that have similar characteristics (Jain *et al.*, 1999). Also descriptive models can be evaluated along the dimensions of predictive accuracy, novelty, utility, and understandability of the fitted model (Abonyi & Feil, 2005). The next section focuses on the clustering techniques.

2.2 Clustering Techniques

Clustering techniques can be considered as a part of the undirected data mining tools, the goal of the undirected data mining is to discover structures in the data as a whole. There is no prediction for the target variable, because there is not, so the distinction between independent and dependant variables will not be included. A cluster is a collection of objects, which are similar between them and dissimilar to the objects belonging to other clusters. Furthermore Clustering is seeks to identify a finite set of categories or clusters to describe data. Fayyad *et al.* (1996) defined that the categories can be mutually exclusive and exhaustive or consist of a richer representation, such as hierarchical or overlapping categories. Some times in loading databases, objects are physically allocated close to each other, and then it said that these objects have been clustered (Visnick, 2003) Fig. 2.1 illustrates how clustering works.

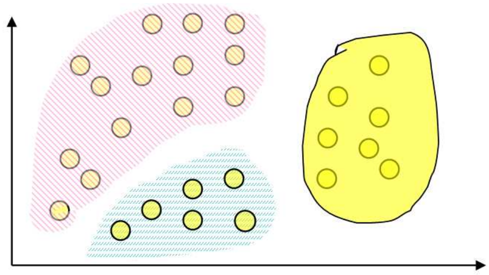


Figure 2.1: Clustering Technique

Two criteria have to be satisfied in order to use the clustering techniques for combining observed examples into clusters, namely

- each group or cluster is homogeneous; examples that belong to the same group are similar to each other.
- each group or cluster should be different from other clusters, that is, examples that belong to one cluster should be different from the examples of other clusters.

Depending on the clustering technique, different ways are used to express clusters:

- Identified clusters may be exclusive, so that any example belongs to only one cluster.
- Overlapping can be happen; an example may belong to several clusters.
- They may be probabilistic, whereby an example belongs to each cluster with a certain probability.
- Clusters might have hierarchical structure, having crude division of examples at highest level of hierarchy, which is then refined to sub-clusters at lower levels. Fig. 2.2 and Fig. 2.3 show an example for hierarchical clusters, and overlapping clusters, respectively.

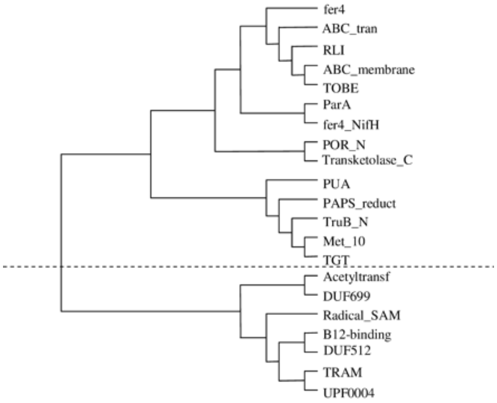


Figure 2.2: Hierarchical Clusters

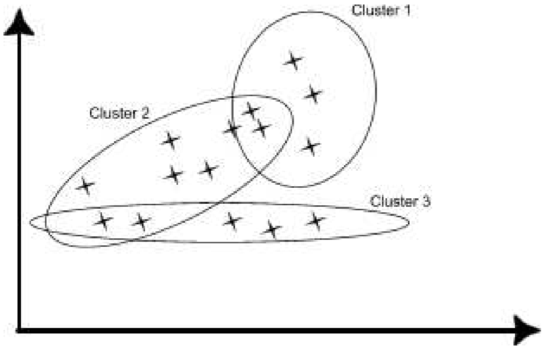


Figure 2.3: Overlapping Clusters

Furthermore different approaches to clustering data can be described with the help of the hierarchy as shown in Fig. 2.4.

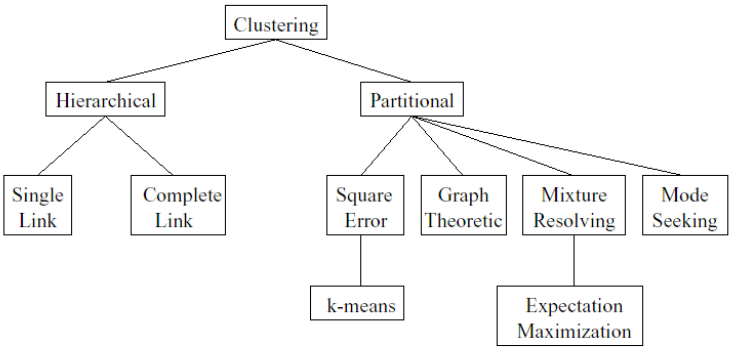


Figure 2.4: Taxonomy of clustering approaches (Jain *et al.*, 1999)

The algorithms of the hierarchical clustering produces a nested series of partitions and that based on criterion for splitting or merging clusters based on similarity. To identify the partition that optimizes a clustering criterion (usually local) partitional clustering algorithms have to be implemented.

2.2.1 Hierarchical Clustering Algorithms

A hierarchical clustering can be defined as a sequence of similarity partitions in which each partition is nested into the next partition in the sequence (Irene, 1999), and be represented in dendrogram, and this can be broken at different levels to produce different clustering's of the data. Different levels of abstraction might be represented in building a cluster of hierarchical structure. Most of hierarchical clustering algorithms are variants of the single-link, complete-link and minimum-variance algorithms. For these, the most popular are the single-link and complete-link algorithms. These two algorithms differ in the way they characterize the similarity between pairs of clusters. The clustering technique that works well on datasets that contains non-isotropic, chain like, well-separated, and concentric clusters is the single-link clustering algorithm. K-means algorithm as a typical partitional algorithm works well only datasets that are isotropic clusters. But partitional algorithms typically have lower space complexities and time than hierarchical algorithms. Lv *et al.* (2006) used hierarchical clustering to analyze 3D model database and improve the retrieval performance. Their proposed algorithm stops automatically by utilizing outlier information and adopts the concept of core group to reduce the influence of parameter on the clustering results.

Single Link Clustering

One of the simplest agglomerative hierarchical clustering methods is single linkage, also known as the nearest neighbor technique. The defining feature of the method is that distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group are considered (Tan *et al.*, 2006).

Complete Link Clustering

The complete linkage, also called farthest neighbor, clustering method is the opposite of single linkage. Distance between groups is now defined as the distance between the most distant pair of objects, one from each group (Jain *et al.*, 1999).

Basic Algorithm for Hierarchical Clustering

The basic rules for agglomerative hierarchical clustering are:

1. *Derive vector representation for each entity (i.e. gene expression values for each experiment make up the vector elements for a specific gene).*
2. *Compare every entity with all other entities by calculating a distance. Input that distance into a matrix. Calculation of the distance depends on:*
    a. *The linkage method being implemented.*
    b. *The method of calculation of actual distances.*

3. *Group closest two entities (or clusters) together (which make a new cluster) and go back to step 2, counting the new cluster as a single entity, until all entities are contained within one big cluster.*

2.2.2 Partitional Algorithms

Jain *et al.* (1999) said that partitional clustering algorithm obtains a single partition of the data instead of a clustering structure, such as the dendrogram produced by a hierarchical technique. Partitional methods have advantages in applications involving large datasets for which the construction of a dendrogram is computationally prohibitive. A problem accompanying the use of a partitional algorithm is the choice of the number of desired output clusters. Additionally, the partitional technique usually produce clusters by optimizing a criterion function defined either locally (on a subset of the patterns) or globally (defined over all the patterns). Combinatorial search of the set of possible labeling for an optimum value of criterion is clearly computationally prohibitive. In practice, therefore, the algorithm is typically run multiple times with different starting states, and the best configuration obtained from all the runs is used as the output clustering.

Squared Error Algorithms

The most intuitive and frequently used criterion function in partitional clustering techniques is the squared error criterion, which tends to work well with isolated and compact clusters. The squared error for a clustering L of a pattern set X (containing K clusters) is:

$$e^2(\mathscr{X}, \mathscr{L}) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \|\mathbf{x}_i^{(j)} - \mathbf{c}_j\|^2$$

Where $x_i^{(j)}$ is the ith pattern belonging to the jth cluster and c_j is the centroid of the jth cluster.

k-Means Clustering Algorithm

The k-means is the simplest and most commonly used algorithm employing a squared error criterion. It starts with random initial partition and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centers until a convergence criterion is met. The k-means algorithm is popular because it is easy to implement, and its time complexity is O (n), where n is the number of patterns. Moreover k-means would work well even on problems with large datasets. A major problem with this algorithm is that it is sensitive to the selection of the initial partition and may converge to a local minimum of the criterion function value if the initial partition is not properly chosen, and the user needs to specify the number of clusters in advance (Erban & Moldovan, 2006).

The steps of the k-means algorithm are given below:

1. *Select randomly k points (it can be also examples) to be the seeds for the centroids of k clusters.*

2. *Assign each example to the centroid closest to the example, forming in this way k exclusive clusters of examples.*
3. *Calculate new centroids of the clusters. For that purpose average all attribute values of the examples belonging to the same cluster (centroid).*
4. *Check if the cluster centroids have changed their "coordinates". If yes, start again form step 2. If not, cluster detection is finished and all examples have their cluster memberships defined.*

Zhang *et al.* (2004) used k-means to get stuck at locally optimal points for high dimensional data. The proposed algorithm combines Genetic Algorithms and k-means algorithm together for improving the search ability of the k-means algorithm. Also k-means clustering where used by Ng *et al.* (2006), to improve watershed segmentation algorithm making use of automated threshold on the gradient magnitude map and post-segmentation merging on the initial partitions to reduce the number of false edges and over-segmentation. By comparing the number of partitions in the segmentation maps of 50 images, they show that their proposed methodology produced segmentation maps which have 92% fewer partitions than the segmentation maps produced by the conventional watershed algorithm.

2.2.3 Graph-Theoretic Clustering

The best-known graph-theoretic divisive clustering algorithm is based on construction of the minimal spanning tree (MST) of the data, and then deleting the MST edges with the largest lengths to generate clusters. Fig. 2.5 depicts the MST obtained from nine two dimensional points. By breaking the link labeled CD with a length of 6 units (the edge with the maximum Euclidean length), two clusters ({A, B, C} and {D, E, F, G, H, I}) are obtained. The second cluster can be further divided into two clusters by breaking the edge EF, which has the length of 4.5 units.

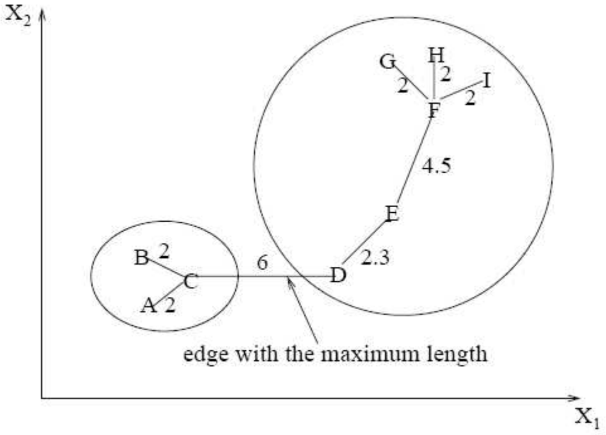


Figure 2.5: Using minimal spanning tree for clustering

The hierarchical approaches are also related to graph-theoretic clustering. Single-link cluster are sub-graphs of the minimum spanning tree of the data which are also the connected components. Complete-link cluster is maximal complete sub-graphs, and related to the node color ability of graphs.

Akosy and Haralick (1999), used graph-theoretic approach for image retrieval by formulating the database search as a graph clustering problem by using a constraint that retrieved images should be consistent with each other (close in the feature space) as will as being individually similar (close) to the query image. Graph-theoretic techniques where adopted by Schenker (2003) for performing data mining on web documents which utilize graph representations of document content. Because the graphs are more robust than typical vector representations as they can model structural information that is usually lost when converting the original web content to a vector representation.

2.2.4 Expectation-Maximization (EM) Algorithm

The EM algorithm has become a popular tool in statistical estimation problems involving incomplete data or in problems which can be posed in a similar form, such as mixture estimation. The EM algorithm has also been used in various motion estimation frameworks and variant of it have been used in multi-frame super resolution restoration methods which combine motion estimation along the lines (Borman, 2009).

EM algorithm used when data is only partially observable, unsupervised clustering (target value unobservable) or supervised learning (some instance attributes unobservable). Furthermore EM produce begins with an initial estimate of the parameter vector and iteratively rescores the patterns against the mixture density produced by the parameter vector. The rescored patterns are then used to update the parameter estimate. In a clustering context, the scores of the patterns (which essentially measure their likelihood of being drawn from particular components of the mixture) can be viewed as hints at the class of the pattern. Those patterns, placed (by their scores) in a particular component, would therefore be viewed as belonging to the same cluster.

Ansari and Viswanathan (1992), used EM algorithm to estimate the unknown jammer parameters and hence obtain a decision on the binary signal based on the estimated likelihood functions. Simulation results show that at low signal-to-thermal noise ratio and high jammer power, the EM detector performs significantly better than the hard limiter and somewhat better than the soft limiter. Also EM algorithm was implemented to joint depth estimation and segmentation from multi-view images is presented. The distribution of the luminance of each image pixel is modeled as a random variable, which is approximated by a "mixture of Gaussians model". After recovering 3D motion, a reference images segmented into a fixed number of regions, each characterized by a distinct affine depth model with 3 parameters (Grammalidis *et al.*, 2002).

2.2.5 Fuzzy C-Means Clustering Algorithm

Traditional clustering approaches generate partitions; in a partition, each pattern belongs to one and only one cluster. Hence, the clusters in a hard clustering are disjoint. Fuzzy clustering extends this notion to associate each pattern with every cluster using a membership function. FCM employs fuzzy partitioning such that a data point (method) can belong to all groups with different membership degrees between 0 and 1. The output of such algorithms is clustering, but not a partition.

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method developed by Dunn in 1973 and improved by Bezdek in 1981 is frequently used in pattern recognition. A high level partitional fuzzy clustering algorithm as below:

1. *Select initial fuzzy partition of the N objects into K clusters by selecting the N × K membership matrix U. An element* u_{ij} *of this matrix represents the grade of membership of object* x_i *in cluster* c_j*. Typically,* u_{ij} *[0,1].*
2. *Using U, find the value of fuzzy criterion function, e.g., a weighted squared error criterion function, associated with the corresponding partition. One possible fuzzy criterion is:*

   $$E^2(\mathscr{X}, \mathbf{U}) = \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ij}\|\mathbf{x}_i - \mathbf{c}_k\|^2, \quad \textit{Where,} \quad \mathbf{c}_k = \sum_{i=1}^{N} u_{ik}\mathbf{x}_i$$

   is the kth fuzzy cluster center. Reassign patterns to clusters to reduce this criterion function value and recomputed U.
3. *Repeat step 2 until entries in U do not change significantly.*

Carvalho (2006) used fuzzy c-means clustering algorithm for symbolic interval data based on adaptive and non-adaptive Euclidean distance, the proposed method furnish a partition of the input data and a corresponding prototype (a vector of intervals) for each class by optimization and adequacy criterion which is based on adaptive and non-adaptive Euclidean distance between vectors of intervals, after that the evaluation of this method has been carried out. The accuracy of the results furnished by these clustering methods were assessed by the corrected Rand index considering synthetic interval datasets in the framework of a Monte Carlo experience and application with real dataset.

2.2.6 Spectral Clustering

Spectral clustering is one of the most popular modern clustering algorithms. It is simple to implement, can be solved efficiently by standard linear algebra software. Spectral clustering refers to a class of techniques which rely on the Eigen structure of a similarity matrix to partition points into disjoint clusters with points in the same cluster having high similarity and points in different clusters having low similarity (Bach & Jordan, 2003).

The learning algorithm for spectral algorithm is the following:

Input: *Similarity matrix S, number of clusters K.*

- *Compute transition matrix P by (2).*
- *Compute v^1, …,v^k the eigenvectors corresponding to the k largest eigenvalues of P.*
- *Clusters the rows of V = [v^1, …,v^k] as points in R^k by using K-means.*

Output: *Clustering C.*

Spectral clustering used to support cases where the entries in the affinity matrix are costly to compute, this method is incremental – the spectral clustering algorithm is applied to the affinity matrix after each row/column is added – which makes it possible to inspect the clusters as new data points are added. The method is well suited to the problem of appearance-based, online topological mapping for mobile robots (Valgren *et al.*, 2007). Moreover Weiming and Zhang (2007) used spectral clustering algorithm to propose a hierarchical corner detection framework in which the mean shift is embedded.

In the corner cell extraction stage, several atomic corner cells are obtained by spectral clustering.

2.2.7 Kohonen Networks

The Kohonen network was invented by Teuvo Kohonen (1981), and is closely modeled on the way that certain parts of the brain are known to work. The basic idea behind the kohonen network is to setup a structure of interconnected processing units "neurons" which compete for the signal. Kohonen networks make the basic assumption that clusters, or classes, are formed from pattern that share common features and groups similar patterns together. Teuvo Kohonen has been invented a variety of networks. The phrase "Kohonen network" most often refers to one of the following three types of networks:

- VQ: Vector Quantization is a competitive network that can be viewed as unsupervised density estimates or auto-associators.
- SOM: Self Organizing Map is a competitive network that provides a "topological" mapping from the input space to the clusters.
- LVQ: Learning Vector Quantization is a competitive network for supervised classification.

The objective of Kohonen Networks is to map input vectors (patterns) of arbitrary dimension N onto a discrete map with 1 or 2 dimensions. Patterns close to one another in the input space should be close to one another in the map: they should be topologically ordered. A Kohonen Network is composed of a grid of output units and N input units. The input pattern is fed to each output unit. The input lines to each output unit are weighted. These weighted are initialized to small random numbers.

This type of neural network is known as an unsupervised network. Clustering techniques apply when there is no classes to be predicted but rather when the instances are to be divided into natural groups. These clusters presumably reflect some mechanism that causes some instances to bear a stronger resemblance to one another that they do to the remaining instances. Clustering naturally requires different techniques to the classification and association learning methods. The kohonen SOM is fully connected, single layer linear network. The output generally is organized in a one or two dimensional arrangement of Processing Elements (PEs) in a line of elements, so each element only has two neighbors (the preceding and the following PE). A one dimensional SOM can be thought of as a string of PEs, where each PE is restricted to be near its two neighbors. When SOM adapts to an input of higher dimensions, it must stretch and curl itself to cover the input space. Basically, the clustering with the Kohonen Network consists of three layers: the input layer, where the data are introduced to the network, the hidden layer, where the data are processed and the output layer, where the results for given are produced.

Arasa *et al*. (1999) introduced a new self organizing neural network, the Kohonen Incorporating Explicit Statistics (KNIES) that is based on Kohonen's Self-Organization Map (SOM). The results of the study showed that the new scheme has been used to solve the Euclidean Travelling Salesman Problem (TSP). Moreover, it has been indicate that NN the most accurate strategy for the TSP currently reported from TSPLIB (A Traveling Salesman Problem Library).

2.3 Applications of Clustering

Clustering algorithms have been used in a large variety of applications including image segmentation, information retrieval, fault diagnosis, condition monitoring and bioinformatics. However further discussion on the applications of clustering is focused on Bioinformatics.

2.3.1 Bioinformatics

The rapidly inflation of biological data, on a way it seems to be exponential, the increasing demand on getting information from such huge data, require to use intelligence techniques to speed up the process of preparing the information through the deep observation inside data and show the relations that can be found between it. Data mining can be applied in this case. For example Deoxyribonucleic acid (DNA) the molecule that our genes produced from is made from proteins which in its turn made from amino acids and there is typically 100 to 500 different amino acids produce protein sequence.

Zien *et al*. (2000) succeeded in determining the protein sequences that are included within nucleotides sequences, knowing where is the start point for the encoding regions for that protein which is called translation initiation sites (TIS), and it's a classification problem that can be handled through support vector machines (SVMs) using the suitable kernel technique. Cancer disease is a significant research field. Tumor types have to be critically classified to be diagnosed, and for cure discovery. Many classification algorithms can be applied to the problem of cancer classification such as decision tree, linear discrimination analysis, nearest-neighbor analysis, and SVMs. All the previous algorithms face a major problem which is the high dimensionality of input space to express the gene, which increase the computational cost. The identification of the marker genes is challenging edge facing the researchers to discriminate tumors for cancer diagnosis (Hu & Pan, 2007).

A new approach has been presented by Graham *et al.* (2003) for drug design process, which accelerates the chemical evaluation phase through parallel inductive logic search for pharmacophores, the new system applies the concept of data partitioning on a loosely coupled collection of parallel inductive logic searches. One of the key design features of this system is its portability and the ease of parallelizing sequential inductive logic programming (ILP) systems, the system shows ease of parallelizing sequential ILP systems based on the concept of data partitioning on a loosely coupled collection of parallel inductive searches.

Distributed processing components using on a workflow model, for co-ordination the execution is the base of Discovery Net system for mixed data and text mining. a flexible infrastructure has been designed to allow end users like biologists to construct their own text mining applications easily, a new form of text mining proceeds by using a generic pipeline that takes in text documents, performs any number of text pre-processing operations (cleaning, NLP parsing, regular expression operations, etc), followed by coding the features of the documents in vector form where counts are recorded for user-defined (Ghanem *et al.*, 2005).

Clustering DNA Sequences

Clustering techniques can be implemented to clustering DNA sequences. The prior to clustering the sequences of DNA, the DNA needs to be transformed into numeric sequence. Next, the distribution of the nucleotides must be identified. FitzGerald *et al.* (2004) determined the distribution of all sequences ranging from 2-mers to 8-mers, in addition to identified the clusters for Transition Split Site (TSS), and finally they identified DNA sequences that cluster in promoters.

Liu *et al*. (2006), has used numeric characterization, through the number of A, T, G, and C nucleic bases in DNA sequence the total distance of each nucleotide from the origin (0, 0), and the distribution of each nucleotide along the DNA sequence. Once the vector is being produced, the Euclidian distance between each characterization vector will be measured to identify the clusters and a sensitivity analysis is conducted.

DNA curvature excess profile technique were used to reduce a comprehensively big text file (genome) to a numerical vector contains 801 real positive numbers smaller than 0.5.

Two widespread clustering methods: k-means and Partitioning Around Medoids (PAM) were used to cluster 205 complete prokaryotic genomes. The results obtained by k-means algorithm application seem to possess better biological relevance. K-means algorithm was applied to cluster genomes using curvature excess distributions upstream of the starts of genes. Optimal growth temperature, genome size and the A + T composition are the main factors influencing curvature distribution in promoter regions of the prokaryotes (Kozobay-Avrahama *et al.*, 2008).

DNA splice site adjacent sequences have remarkable conservative feature and has much genetic information. 2796 donor sequences of human being have been chose as the experimental data set to cluster DNA sequences using DBSCAN and analyzing the clustered results to mine the regulation in each cluster. In order to improve the applicability of the algorithm, dissimilarity definition methods were used. The frequencies of ‘‘T+C” (A+C) and the di-base bias are identified. This helps to predict the functions of the sequences in each cluster, and it will be also helpful to mine more biological knowledge from the clustering results (Zhang *et al.*, 2008).

Based on orthologous gene property conservation profiles Bolshoy and Volkovich (2008) introduced an unsupervised genome clustering strategy of taxonomic analysis based on an information bottleneck method, supposing that n genomes have been used to construct a genome tree. They define an orthologous gene property conservation profile of a gene x as an n-component vector of zeros and ratios, this will reflects on an evolutionary conservation history of a property p across the n species. In their study, Bacteria and Archaea clusters showed a clear separation and clustering of relatively close species.

2.3.2 Image Segmentation

The segmentation of image(s) presented to an image analysis system is a critically dependent on the scene to be sensed, the imaging geometry, configuration, and sensor used to transducer the sense into a digital image, and ultimately the desired output (goal), of the system. And image segmentation is typically defined as an exhaustive partitioning of an input image into regions, each of which is considered to be homogenous with respect to some image property of interest (e.g., intensity, color, or texture).

The goal of clustering was to obtain a sequence of hyperellipsoid clusters starting with cluster centers positioned at maximum density locations in the pattern space, and growing clusters about these centers until the test for goodness of fit is violated. An agglomerative clustering algorithm was applied to solve the problem of unsupervised leaning of clusters of coefficient vectors for two image models that correspond to image segments. The algorithm proceeds by obtaining vectors of coefficients of least-squares fits to the data in M disjoint image windows (Silverman & Cooper, 1988). Two neural networks have been designed to perform pattern clustering when combined. A two layer network operates on multidimensional histogram of the data to identify prototypes which are used to classify the input patterns into clusters. These prototypes are fed to the classification network, another two-layer network operating on the histogram of the input data, but are trained to have different weights from the prototype selection network. In both networks, the histogram of the image is used to weight the contributions of patterns neighboring the one under consideration to the location of prototypes or the ultimate classification; as such, it is likely to be more robust when compared to techniques which assume an underlying parametric density function for the pattern classes. This architecture was tested on gray-scale and color segmentation (Vinod *et al.*, 1994).

2.3.3 Information Retrieval

Information retrieval (IR) is concerned with automatic storage and retrieval of documents, many university libraries use IR systems to provide access to books, journals, and other documents. Libraries use the Library of Congress Classification (LCC) scheme for efficient storage and retrieval of books.

The clustering problem can be stated as follows; given a collection B of books, its need to obtain a set of clusters. Jain and Dubes (1988) used a proximity dendrogram, using the complete link agglomerative clustering algorithm for the collection of 100 books. Seven clusters are obtained by the threshold (T) value of 0.12. It is well known that different values for T might give different clustering. This threshold value is chosen because the gap in the dendogram between the levels at which six and seven clusters are formed is the largest. An examination of the subject areas of the books in these clusters revealed that the clusters obtained are indeed meaningful. Each of these clusters is represented using a list of string and frequency pairs, where the frequency represents the number of books in the cluster that is presented in the string.

2.4 General Taxonomy of DNA Sequences

DNA sequences in general have two types of cellular architecture: Prokaryotic (Bacteria and Archaea) and these two named as Prokaryotes and the other type is Eukaryotic. Bacteria and Archaea are unicellular, Eukaryotes are either unicellular (e.g. yeast) or multi cellular (e.g. mammals) as shown in Fig. 2.6, a general sampling of DNA sequences has helped establish the diversity of life and allowed researchers to analyze evolutionary relationships within groups in detail (Stoeckle, 2003).

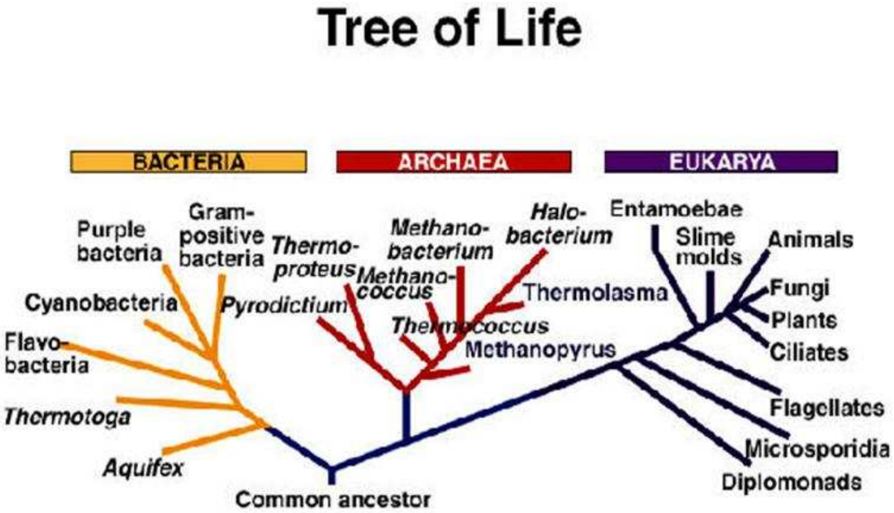


Figure 2.6: General Taxonomy for Species
https://eapbiofield.wikispaces.com/FRF+PR9

Hebert *et al.* (2003) discussed in their paper that many mammalian genes can be organized into gene families consisting of a number of genes with similar sequences, and the DNA extracted from small tissue samples using the Isoquick protocol. DNA sequence should contain more than enough information to resolve 10 million or even 100 million species, there is no universal DNA bar code gene, no single gene that is conserved in all domains of life and exhibits enough sequence divergence for species discrimination, there may be a need for a standalone; curate database to supplement GenBank, the stand-alone database would be designed to integrate sequence data with specimen and taxonomic information.

A cytochrome “c” oxidase I (COI) database could serve as the basis for a global bio-identification system (GBS) for animals. Implementation on this scale will require the establishment of a new genomics database. While GenBank aims for comprehensive coverage of genomic diversity, the GBS database would aim for comprehensive taxonomic coverage of just a single gene. The creation of the GBS will be a large undertaking and will require close bonds between molecular biologists and taxonomists. DNA-based species identification offers enormous potential benefits for the biological scientific community, educators, and the interested public. It will help open the treasury of biological knowledge and increase community interest in conservation biology and understanding of evolution.

2.5 DNA Sequences Representations

It is possible to represent DNA sequence with numeric or graphs for easy analysis, these numeric or graph representations can be in the form or real numbers or complex numbers depending on the further analysis required for the clustering.

2.5.1 Graphical Representation of DNA Sequences

As mentioned in the previous chapter, there is a large volume of DNA sequences, and for the purpose of analyzing it in a mathematically way it will be challenging. For a simple way to view, compare and sort DNA sequences the graphical representation of DNA sequences gives an operative way. The main goal of graphical representation of DNA sequences is; to show the similarity and the difference in the gene structure in an easy way visually (Gates, 1985).

In order to have a unique graphical representation for the DNA sequences, it is required that the graphical representation has no degeneracy. Many efforts have been made to avoid the degeneracy caused by overlapping and crossing paths itself. One of the examples of the degeneracy is the high dimensional graphical representations of the DNA sequences. For more straight visual display, 2-D graphical representation gives that with less computation and drawing tools.

Song and Tang (2005) create a new 2-D graphical representation of DNA sequences based on chemical structure of bases. They reduced the DNA primary sequence into some characteristic curves. Each characteristic curve may be regarded as a coarse grained description of the DNA primary sequence, which avoids overlapping and crossing of the curve, reflects the distribution of different base pairs. This approach is accompanied with an arbitrary decision in assigning to the different types of bases different geometrically non-equivalent graphical choices. The graphical representation results in a numerical characterization of a DNA sequence by the leading eigenvalues of M/M, L /L matrices associated with the DNA sequences, only six out of 12 possible graphs are shown in Fig. 2.7.

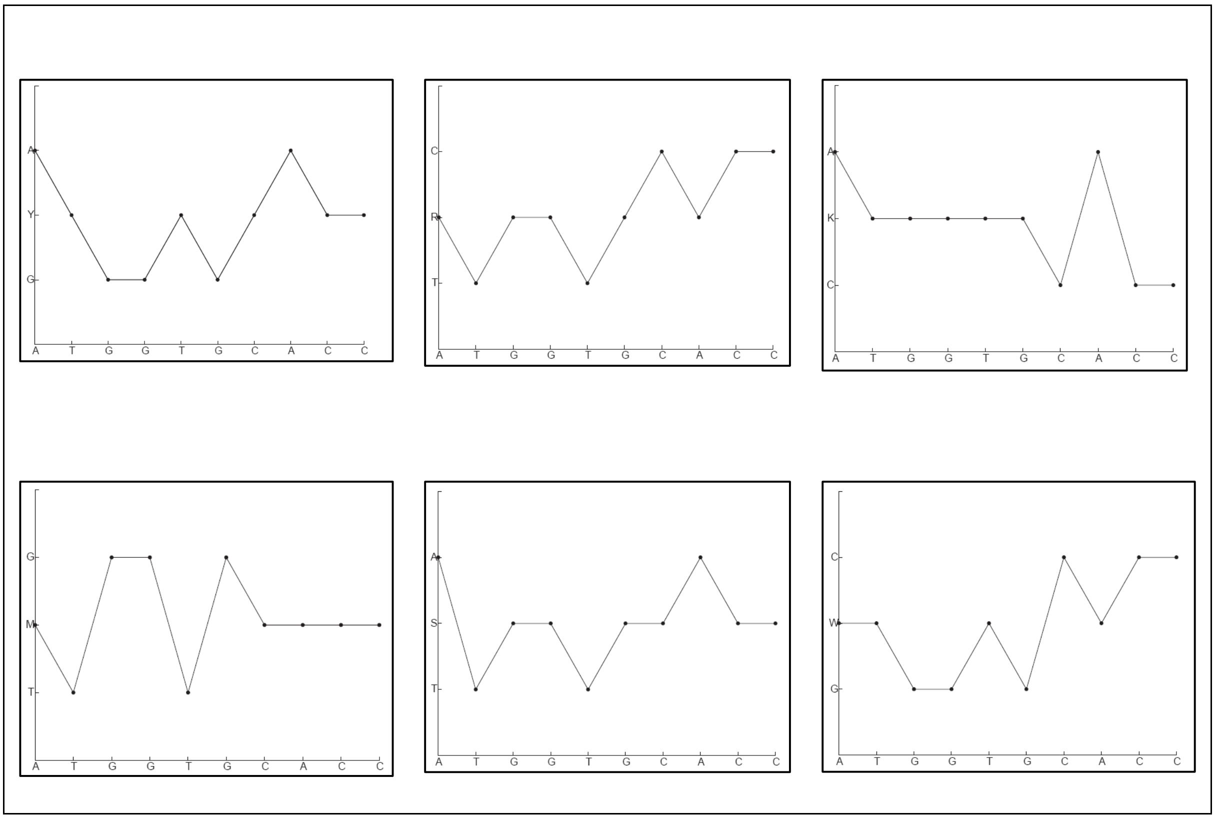


Figure: 2.7: 2-D characteristic curve of the sequence TGGTGCACCTGACTCCTGA (Song & Tang, 2005).

Randi *et al.* (2002) transferred data from a DNA sequence to its mathematical representation presented in a 2-D graphical representation that preserve and avoid the loss of information on sequential adjacency of nucleotides and allow numerical characterization. Zigzag curve illustrates DNA sequence that will smooth the progress of quantitative comparisons of DNA sequences Fig 2.8. Associating the four nucleotide: A, T, C and G with the four horizontal lines, the consecutive bases along the horizontal axes are placed at unit displacement. Also 2D numeric representation using 2 Cartesian co-ordinate system where A, G, C and T are represented with a unit vector has been used by Wan and Johnson (2002).

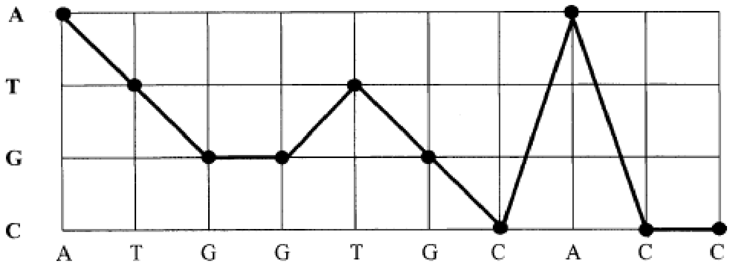


Figure 2.8: Graphical representation of the sequence ATGGTGCACC. (Randi et al., 2002)

Guo and andy (2002) introduced a method to reduce the degeneracies of the DNA sequence representation, so that there are considerably less overlaps in the graphs. DNA sequence of four nucleotides A, T, C and G and have the length n can be considered as a successive vector sequence of length n containing the four vectors corresponding to A, T, C, G and used in 2-D graphical representation of the DNA sequence. The new descriptors of DNA sequences give a good numerical characterization of DNA sequences, which have lower degeneracy.

2.5.2 Numerical Representation of DNA Sequences

It is possible to represent DNA sequence with numeric for easy analysis, these numeric can be in the form or real numbers or complex numbers depending on the further analysis required for the clustering. Real number representation has been implemented to represent the DNA sequence. For example, Kauer and Blocker (2003), assigned numbers to nucleotide such as an instance A=1, T=2, C=3 and G =4, based on the assumption that A<T<C<G. Complex numbers has been used by Anastassiou (2000) to represent DNA sequences. DNA sequences are converted to vectors of complex number by assigning nucleotide with complex Number and the corresponding nucleotides with a complex conjugate.

Based on digital signal method, Qi and Qi (2009) propose a new representation of DNA primary sequence. It is very difficult to find DNA characteristic vector particularly when the sequence is very long. To deal with the situation, Qi and Qi used signal theory to characterize DNA sequences. By constructing a weak-H/strong-H bond graph in the first quadrant of the Cartesian coordinate system, the unit digital signals representing four nucleotides A, T, G and C can be represented as in Fig.2.9.

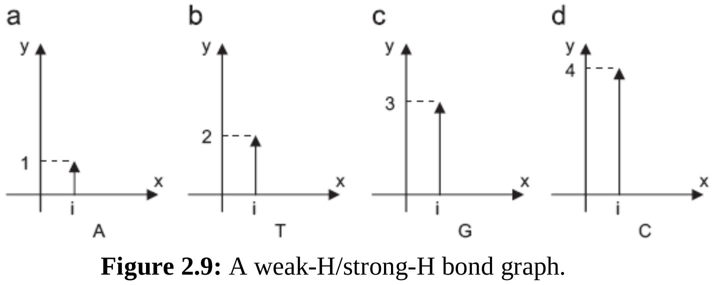


Figure 2.9: A weak-H/strong-H bond graph.
(Qi & Qi ,2009)

Four nucleotides, A, T, C and G have been represented as four two-component vectors (Huang *et al.*, 2009). Each vector contains a constant (equal to 1) and different angles between A, T, C and G nucleotides and the x-axis. By comparing the corresponding curve of nucleotides, a new measure of similarity and dissimilarity was proposed. This conveniently discovers the evolutionary relationship among various DNA sequences by observing the graphical representations.

2.6 Conclusion

This chapter presented data mining concept and the clustering techniques which is part of the descriptive models. Clustering techniques have been categorized into two main categories; hierarchical and partitional, each of which has different clustering algorithm. From previous literature, representing the DNA sequences in a characterization vectors enables the use of Euclidean distance between the corresponding vectors for the DNA sequences, for the similarity between the DNA sequences. This methodology as proposed by Liu *et al.* (2006) has been adopted in this work. The next chapter discusses the methodology in more details.

CHAPTER THREE

METHODOLOGY

In this chapter, a twelve-dimensional vector is associated with each DNA sequence. A new genomic geometry will be produced in a twelve-dimensional space. The length of the vectors have to be the same in order to utilize vectors to characterize the DNA sequence, regardless the difference of the original sequences that are in alpha representation.

3.1 Introduction

Finding similarities between the new genes and the previous sequenced genes with known functions through sequence comparison will help to discover the function of the new sequenced gene. BLAST and FASTA is the most popular tools used in sequence comparison. However, BLAST and FASTA only compares two sequences at one time and does not provide a global picture of the comparison of all genes simultaneously. Consequently it will be advantageous to put all the genes in a fixed Euclidean space so that a global view will be produced for all genes comparison.

BLAST and FASTA tools cannot compare more than two DNA sequences simultaneously, also they are using the alpha representation of the DNA sequences, so the methodology which is adopted in this work does not compare the DNA sequences in its alpha representation but it converts it into a numerical representation which is characterization vector, this will make it easy to compare more than two DNA sequence simultaneously in a genomic space. Hence, no need to compare the results of this study with BLAST and FASTA tools because of the differences in the why of representing the DNA sequence and the way of comparing more than two DNA sequences.

3.2 Research Methodology

In this study, the methodology from Liu *et al*. (2006) as shown in Fig.3.1 is adopted.

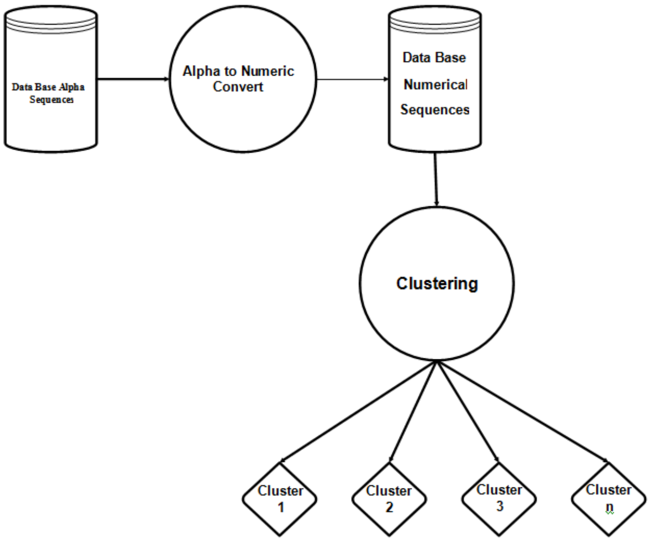
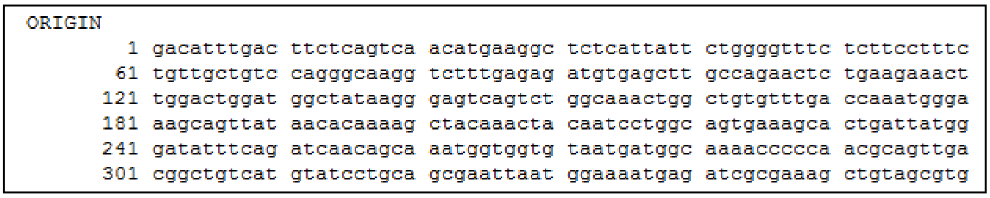


Figure 3.1: Methodology (Liu *et al*., 2006)

Experiments were conducted based on the following steps:

STEP 1: Alpha DNA sequences were obtained from Gene Bank from the following site:

http://www.ncbi.nlm.nih.gov/. An example of the data can be seen in Fig. 3.2.



```
ORIGIN
        1 gacatttgac ttctcagtca acatgaaggc tctcattatt ctggggtttc tcttcctttc
       61 tgttgctgtc cagggcaagg tctttgagag atgtgagctt gccagaactc tgaagaaact
      121 tggactggat ggctataagg gagtcagtct ggcaaactgg ctgtgtttga ccaaatggga
      181 aagcagttat aacacaaaag ctacaaacta caatcctggc agtgaaagca ctgattatgg
      241 gatatttcag atcaacagca aatggtggtg taatgatggc aaaacccca acgcagttga
      301 cggctgtcat gtatcctgca gcgaattaat ggaaaatgag atcgcgaaag ctgtagcgtg
```

Figure 3.2: Original DNA sequence

STEP 2: The alpha DNA sequences were converted into 12 parameter characterization vector and this include:

(a) The first four parameters of the characterization vector contain n_A, n_T, n_C, and n_G and can be defined as following:

i. n_A: Total number of nucleotide A in the DNA sequence.

ii. n_T: Total number of nucleotide T in the DNA sequence.

iii. n_C: Total number of nucleotide C in the DNA sequence.

iv. n_G: Total number of nucleotide G in the DNA sequence.

For example if a sequence A is: GTGGGTGGTT, and sequence B is: TGAAGCTGTT, the sequences will be used in the following parts, whose corresponding four parts for each sequence is shown in Table 3.1.

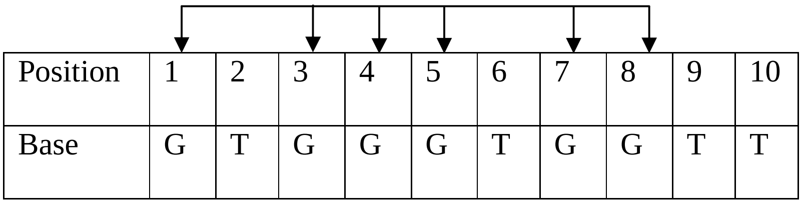
Table 3.1: Number of nucleotides (A, T, C, G) in the sequence GTGGGTGGTT

Parameter	Sequence A GTGGGTGGTT	Sequence B TGAAGCTGTT
n_A	0	2
n_T	4	4
n_C	0	1
n_G	6	3

(b) The second four parameters of the characterization vector are the total distance for each nucleotide base on the first nucleotide, or the origin (0,0) of the DNA sequence. It can be defined as:

$$T_i = \sum_{j=1}^{n_i} t_j \qquad (3.1)$$

Where, i = A, T, C, G. And t_j: is the distance from the first nucleotide to the j^{th} nucleotide in the DNA sequence. Therefore the set of the four parameters for the characterization vector dictated by DNA sequence are T_A, T_T, T_C, and T_G. Fig.3.3 shows the positions of G base in the sequence GTGGGTGGTT; hence the total number of distances for base G is the summation of the distances which is: 1 + 3 + 4 + 5 + 7 + 8 = 28, and so on for the rest of nucleotide bases.



Position	1	2	3	4	5	6	7	8	9	10
Base	G	T	G	G	G	T	G	G	T	T

Figure 3.3: The position of nucleotide in the sequence GTGGGTGGTT

The second four parts of the characterization vectors for the same sequences that have been used in the previous step is shown in Table 3.2.

Table 3.2: Total distance of nucleotides (A, T, C, G) in the sequence GTGGGTGGTT

Parameter	Sequence A GTGGGTGGTT	Sequence B TGAAGCTGTT
T_A	0	7
T_T	27	27
T_C	0	6
T_G	28	15

(c) The distribution of each nucleotide along the DNA sequence will produce the third four parameter of the characterization vector.

The distribution is defined as:

$$D_i = \sum_{j=1}^{n_i} \frac{\left(t_j - \mu_i\right)^2}{n_i} \tag{3.2}$$

Where i = A, T, C, G. And t_j: is the distance from the first nucleotide to the j^{th} nucleotide in the DNA sequence and

$$\mu_i = \frac{T_i}{n_i} \tag{3.3}$$

The four parameters for the third part will be D_A, D_T, D_C, and D_G.

For example $\mu_T = \frac{T_T}{n_T}$, this value will be used in main function to calculate the distribution D (3.2) of nucleotide T in the sequence, so the values of n_T and T_T can be obtained from Table 3.1 and Table 3.2 respectively, then $\mu_T = \frac{27}{4}$ = 6.75, so this value

$D_T = \sum_{j=1}^{n_T} \frac{\left(t_j - \mu_T\right)^2}{n_T}$ = 9.68, the distribution of all the nucleotides (A, T, C and G) for the same sequences above shown in Table 3.3.

Table 3.3: The distribution of nucleotides (A, T, C, G) in the sequence GTGGGTGGTT

Parameter	Sequence A GTGGGTGGTT	Sequence B TGAAGCTGTT
D_A	0	0.25
D_T	9.68	12.18
D_C	0	0
D_G	5.48	6

So the characterization vector that contains 12-dimensional information is:

< n_A , T_A , D_A , n_T , T_T , D_T , n_C , T_C , D_C , n_G , T_G , D_G >

STEP 3: Store the characterization vector that represents the DNA sequence into database.

STEP 4: Clustering the numeric representation of DNA sequences

(d) The distance between two characterization vectors has to be small to indicate that these two sequences are similar.

(e) Two characterization vectors that have large distance between them correspond to non-homologous DNA sequences.

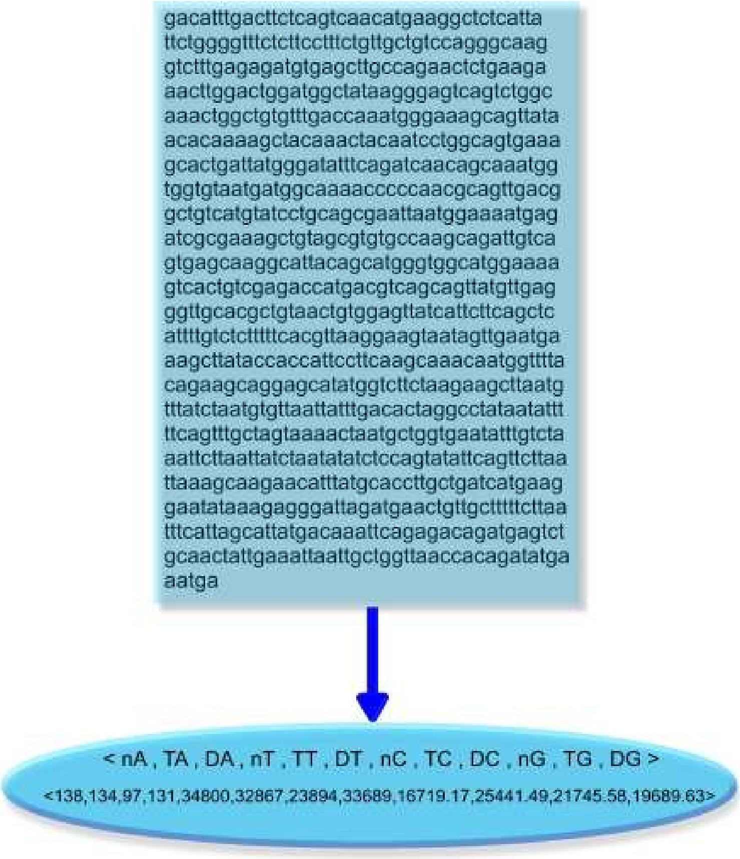
(f) The distance between two characterization vectors defined as:

$$L = \sqrt{\sum_j \sum_i \left(j_i - j_i^{'}\right)^2} \tag{3.4}$$

Where i = A, T, C, and G; j = n, T, D.

By applying the previous formula L (3.4) on the characterization vectors of the sequences A and B we will find that the distance between these two vectors is = 42.84.

Fig 3.4 shows an example of DNA sequence from Lysozyme family named “Bos Taurus” and the corresponding characterization vector for this DNA sequence.



```
gacatttgacttctcagtcaacatgaaggctctcatta
ttctggggtttctcttcctttctgttgctgtccagggcaag
gtctttgagagatgtgagcttgccagaactctgaaga
aacttggactggatggctataagggagtcagtctggc
aaactggctgtgtttgaccaaatgggaaagcagttata
acacaaaagctacaaactacaatcctggcagtgaaa
gcactgattatgggatatttcagatcaacagcaaatgg
tggtgtaatgatggcaaaacccccaacgcagttgacg
gctgtcatgtatcctgcagcgaattaatggaaaatgag
atcgcgaaagctgtagcgtgtgccaagcagattgtca
gtgagcaaggcattacagcatgggtggcatggaaaa
gtcactgtcgagaccatgacgtcagcagttatgttgag
ggttgcacgctgtaactgtggagttatcattcttcagctc
attttgtctctttttcacgttaaggaagtaatagttgaatga
aagcttataccaccattccttcaagcaaacaatggtttta
cagaagcaggagcatatggtcttctaagaagcttaatg
tttatctaatgtgttaattatttgacactaggcctataatattt
ttcagtttgctagtaaaactaatgctggtgaatatttgtcta
aattcttaattatctaatatatctccagtatattcagttcttaa
ttaaagcaagaacatttatgcaccttgctgatcatgaag
gaatataaagagggattagatgaactgttgctttttcttaa
tttcattagcattatgacaaattcagagacagatgagtct
gcaactattgaaattaattgctggttaaccacagatatga
aatga
```

< nA , TA , DA , nT , TT , DT , nC , TC , DC , nG , TG , DG >

<138,134,97,131,34800,32867,23894,33689,16719.17,25441.49,21745.58,19689.63>

Figure 3.4: Bos taurus DNA sequence before and after characterization

The steps of the methodology shows that the numerical characterization of the DNA sequence after obtaining the alpha representation of the DNA sequences which contain the four nucleotides A, T, C and G falls in three main parts, each part contain four parameters to produce the twelve dimensional vector. The first four parameters of the vector contains the total number of the nucleotides A, T, C and G which is represented by n_A, n_T, n_C and n_G respectively, using only these parameters cannot denote a specific DNA sequence, because two different DNA sequences can have exactly the same nucleotide contents. So more parameters are needed.

The second four numerical parameters are the total distance of each nucleotide bases to the origin (0,0). For example if there is two DNA sequences, the first one has two thymine nucleotides at the position 4 and 5, and the second one has two thymine nucleotides at position 6 and 8, both DNA sequences have two thymine bases, the total distance generated from these two cases are different. Therefore, it is a special characteristic to the sequence.

The characteristics of the four sets of total distances T_A, T_T, T_C, and T_G are dictated by the DNA sequence that reflect the information of how far each nucleic base is from the origin. The similarity between DNA sequences also can be reflected through the similarity between the total distances of the nucleotides from the origin. Though, the total distance of the nucleotides alone is not sufficient to denote the DNA sequence for comparison. So there is a need to other numerical parameters for further characterization of DNA sequence.

The distribution of each nucleotide along the DNA sequence is the third four parameters selected for the vector. If the distribution of each nucleotide base is different, DNA sequences cannot be similar even though they may have the same nucleotide contents and the same total distance measurement. Therefore, the information about distribution has also been included in the vector analysis.

As preceded above, each parameter of which the characterization vector is consists of is not sufficient to denote a specific DNA sequence. However, combining the parameters together to produce the characterization vector can be used to characterize the similarity between DNA sequences.

After obtaining the characterization vector for each DNA sequence, the similarity of different DNA sequences can be measured. The distance between vectors is used for the comparison, if the distances between two DNA sequences are small, then they are similar. Otherwise, large distance between the characterization vectors is expected for non-homologous DNA sequences.

The practical application of using the characterization vector for DNA sequence comparison is straightforward; the following chapter shows how to apply this method on different protein families.

CHAPTER FOUR

RESULTS AND DISCUSSION

This chapter presents the results that have been obtained from the conducted experiments. Two main experiments have been carried out; the first experiment has been applied on the whole sequence, the second experiment contains five stages in each stage just a part of the sequence considered in such way that does not exceed the maximum length of the shortest DNA sequence.

For the purpose of presenting the DNA sequence in a 12-dimensional feature vector that contains the following components; the number of instances for each nucleotides A, T, C and G that creates the first 4 parameters of the vector, the second 4 parameters of the vector is the total distances for each nucleotide from the origin of the sequence, the distribution of each nucleotide among the sequence will generate the last 4 parameters of the vector, a program in JAVA has been developed to generate the corresponding vectors for each sequence, and the results are stored in an Access database. This chapter shows how make DNA sequences comparison. The global comparison of gene structures is tested on Lysozyme, Myoglobin and Rhodopsin families (http://www.ncbi.nlm.nih.gov/).

The comparison will be in two ways on the DNA sequences. First the compare is made on the whole DNA sequences, while the other is to take part of the DNA sequences such that it does not exceed the maximum number of nucleotides for the shortest sequence.

4.1 DNA Sequences (Data)

A DNA sequence or genetic sequence is a succession of letters constituent nucleotides listed from the 5'- to 3'- terminus representing the primary structure of DNA molecule or strand, which hold the information as described by the central dogma of molecular biology. Prior to discussion on applications of DNA sequences, several terminologies related to DNA are defined for example, Genome is a complete set of DNA (Deoxyribonucleic Acid) for an organisms, and the DNA arranged into 23 pairs of DNA molecule called chromosomes, and each chromosome contain many genes, DNA molecule has millions of bases or nucleotides, these nucleotides sequences or base sequences has the information of making proteins encoded in it. A nucleotide is made up of one phosphate group linked to a pentose sugar, which is linked to one of 4 types of nitrogenous organic bases symbolized by the four letters A, C, G, and T. The rules that govern the correspondence of the base/ nucleotide sequences for DNA and RNA to the amino acids or proteins are known as Genetic Code. Sequence Alignment is the process of locating regions that are equivalent to increase the similarity of these sequences.

The number of nucleotides that have been used in this study for all the three protein families and its members is 10779 base pairs (bp), more information about the DNA sequences that have been used in this study will be found in Table 4.1. As mentioned earlier, the gene structures are classified into Lysozyme, Myoglobin and Rhodopsin families. Note that the structures are not of the same length although they are from the same gene structures. Due to this reason, the experiments were conducted by considering the length of the gene structures. In other words, gene structures were considered as a whole or fixed size sequences.

Table 4.1: Number of Nucleotides in each family member

Lysozyme	Bos taurus	915(bp)
	Homo sapiens	1487(bp)
Myoglobin	Danio rerio	1360(bp)
	Homo sapiens	1206(bp)
	Mus musculus	505(bp)
	Rattus norvegicus	1015(bp)
	Sus scrofa	1111(bp)
Rhodopsin	Homo sapiens	1620(bp)
	Rattus norvegicus	1560(bp)

The total number (nA, nT, nC and nG) of nucleotides A, T, C and G respectively can be graphically represented as shown in Fig. 4.1.

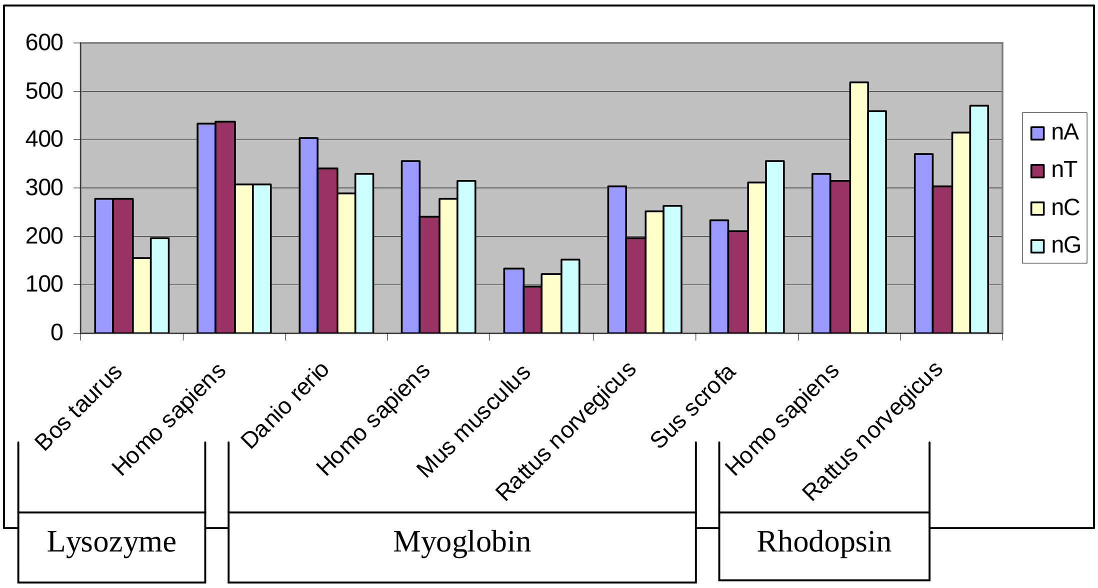
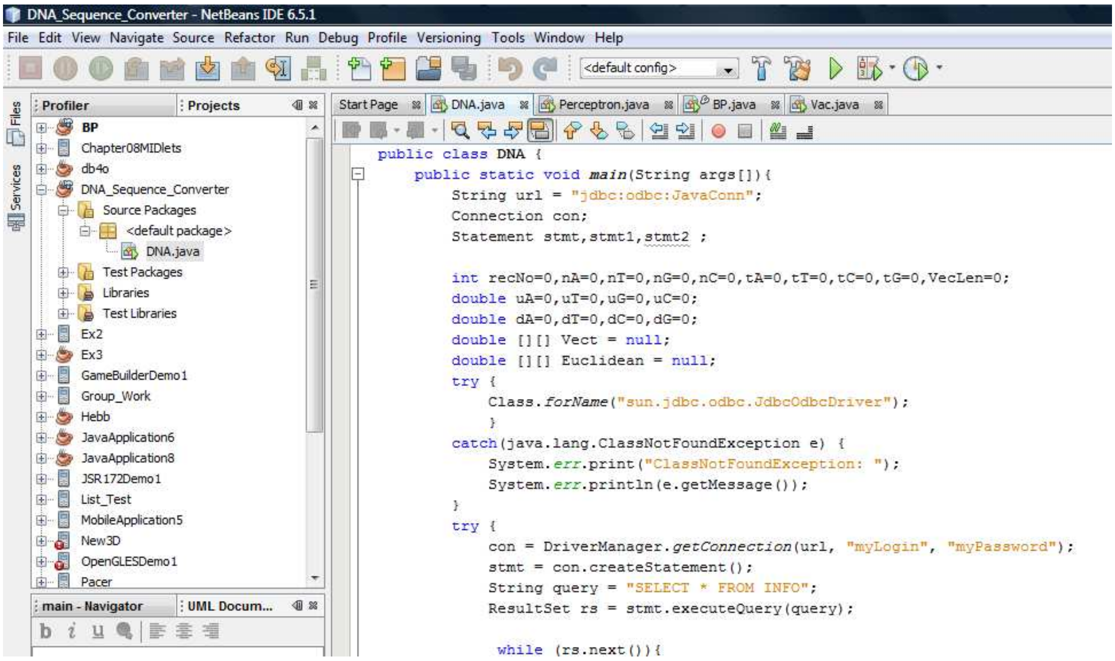


Figure 4.1: The total number of Nucleotides in each sequence

The DNA sequences are stored in an MS Access database, with its name and the name of the family that it belong to, and stored in alpha characters. This database will facilitate the process of retrieving and storing the feature vectors later.

4.2 The Experiment

After obtaining the DNA sequence from the NCBI web site, the DNA sequences was then stored in the database. A JAVA program has been designed to make the process of global comparison among all sequences Fig. 4.2.



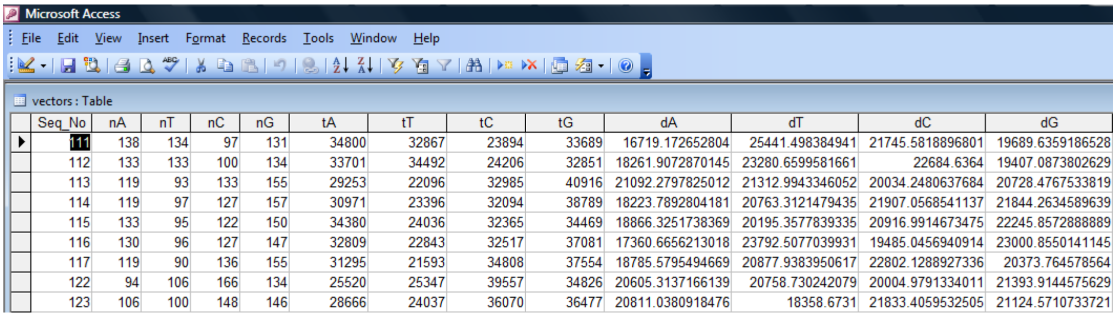
```
public class DNA {
    public static void main(String args[]){
        String url = "jdbc:odbc:JavaConn";
        Connection con;
        Statement stmt,stmt1,stmt2 ;

        int recNo=0,nA=0,nT=0,nG=0,nC=0,tA=0,tT=0,tC=0,tG=0,VecLen=0;
        double uA=0,uT=0,uG=0,uC=0;
        double dA=0,dT=0,dC=0,dG=0;
        double [][] Vect = null;
        double [][] Euclidean = null;
        try {
            Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");
            }
        catch(java.lang.ClassNotFoundException e) {
            System.err.print("ClassNotFoundException: ");
            System.err.println(e.getMessage());
        }
        try {
            con = DriverManager.getConnection(url, "myLogin", "myPassword");
            stmt = con.createStatement();
            String query = "SELECT * FROM INFO";
            ResultSet rs = stmt.executeQuery(query);

             while (rs.next()){
```

Figure 4.2: DNA sequences comparison program

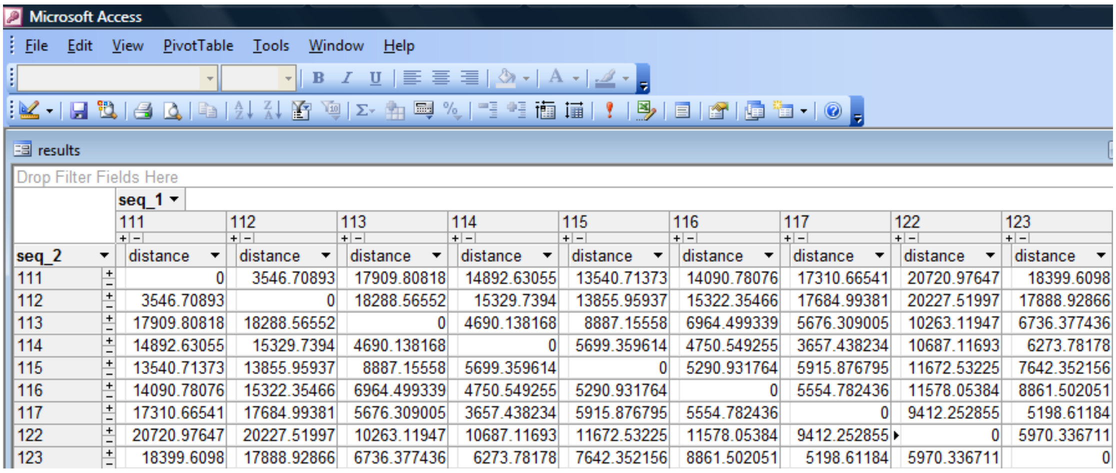
This program retrieves the DNA sequence from the database and then converts it to a 12-dimensional vector, and stores it again in the database to produce a featured vectors database. Fig. 4.3 shows the feature vectors produced from the converter program and how are they stored in the database as a table of featured vectors.



Seq_No	nA	nT	nC	nG	tA	tT	tC	tG	dA	dT	dC	dG
111	138	134	97	131	34800	32867	23894	33689	16719.172652804	25441.498384941	21745.5818896801	19689.6359186528
112	133	133	100	134	33701	34492	24206	32851	18261.9072870145	23280.6599581661	22684.6364	19407.0873802629
113	119	93	133	155	29253	22096	32985	40916	21092.2797825012	21312.9943346052	20034.2480637684	20728.4767533819
114	119	97	127	157	30971	23396	32094	38789	18223.7892804181	20763.3121479435	21907.0568541137	21844.2634589639
115	133	95	122	150	34380	24036	32365	34469	18866.3251738369	20195.3577839335	20916.9914673475	22245.8572888889
116	130	96	127	147	32809	22843	32517	37081	17360.6656213018	23792.5077039931	19485.0456940914	23000.8550141145
117	119	90	136	155	31295	21593	34808	37554	18785.5795494669	20877.9383950617	22802.1288927336	20373.764578564
122	94	106	166	134	25520	25347	39557	34826	20605.3137166139	20758.730242079	20004.9791334011	21393.9144575629
123	106	100	148	146	28666	24037	36070	36477	20811.0380918476	18358.6731	21833.4059532505	21124.5710733721

Figure 4.3: How feature vectors stored in the database

The next step is to compare all the featured vectors that have been stored in the database through the same program and then stores the results (distances) in the database. Fig. 4.4 shows the output for the program in the form of pivot table form designed in MS-Access program.



	seq_1								
	111	112	113	114	115	116	117	122	123
seq_2	distance	distance	distance	distance	distance	distance	distance	distance	distance
111	0	3546.70893	17909.80818	14892.63055	13540.71373	14090.78076	17310.66541	20720.97647	18399.6098
112	3546.70893	0	18288.56552	15329.7394	13855.95937	15322.35466	17684.99381	20227.51997	17888.92866
113	17909.80818	18288.56552	0	4690.138168	8887.15558	6964.499339	5676.309005	10263.11947	6736.377436
114	14892.63055	15329.7394	4690.138168	0	5699.359614	4750.549255	3657.438234	10687.11693	6273.78178
115	13540.71373	13855.95937	8887.15558	5699.359614	0	5290.931764	5915.876795	11672.53225	7642.352156
116	14090.78076	15322.35466	6964.499339	4750.549255	5290.931764	0	5554.782436	11578.05384	8861.502051
117	17310.66541	17684.99381	5676.309005	3657.438234	5915.876795	5554.782436	0	9412.252855	5198.61184
122	20720.97647	20227.51997	10263.11947	10687.11693	11672.53225	11578.05384	9412.252855	0	5970.336711
123	18399.6098	17888.92866	6736.377436	6273.78178	7642.352156	8861.502051	5198.61184	5970.336711	0

Figure 4.4: Pivot table form for the output

In this experiment two main approaches have been carried out to produce the characterization vectors. The first one is to produce the feature vectors from the complete sequence nucleotides, and the second one is produce the feature vectors by taking a fixed size of the sequences. The next sections show how the results produced in these two approaches.

4.2.1 Whole Size Sequence Experiment

In this experiment, the whole sequences for the three families (Lysozyme, Myoglobin and Rhodopsin) have been included and the results are shown in Table 4.2, taking in consideration the various lengths for each sequence, which varies from 505(bp) up to 1620(bp). Table 4.1 shows the number of nucleotides in each sequence.

Table 4.2: Whole sequence comparison results

	Lysozyme		Myoglobin					Rhodopsin	
	Bos taurus	Homo sapiens	Danio rerio	Homo sapiens	Mus musculus	Rattus norvegicus	Sus scrofa	Homo sapiens	Rattus norvegicus
	I	II	III	IV	V	VI	VII	VIII	IX
I	0.00E+00	4.05E+05	2.95E+05	1.98E+05	1.76E+05	9.80E+04	1.72E+05	5.60E+05	4.88E+05
II		0.00E+00	1.19E+05	2.57E+05	5.78E+05	3.60E+05	3.37E+05	3.09E+05	2.28E+05
III			0.00E+00	1.52E+05	4.67E+05	2.49E+05	2.29E+05	3.33E+05	2.41E+05
IV				0.00E+00	3.51E+05	1.22E+05	1.49E+05	4.09E+05	3.25E+05
V					0.00E+00	2.32E+05	2.92E+05	6.97E+05	6.36E+05
VI						0.00E+00	1.12E+05	4.88E+05	4.17E+05
VII							0.00E+00	4.06E+05	3.48E+05
VIII								0.00E+00	1.15E+05
IX									0.00E+00

The results shown in Table 4.2 shows that the protein families did not clustered in a proper way, and that can be identified through the ranges that each family fall in. For example in the Lysozyme family the distance between the two sequences is 4.05E+05 but the distance between it and the two families (Myoglobin and Rhodopsin) is not significant. Consequently, there is a need to examine the DNA sequences in such way that have the same length.

4.2.2 Fixed Size Sequence Experiment

In this section, five different experiments have been implemented on the protein families with its DNA sequences. In each experiment the same number of nucleotides has been chosen for the purpose of comparison.

First 100, 200, 300, 400 Nucleotides

The first 100 nucleotides have been chosen to produce the feature vectors for each sequence. Table 4.3 shows that there is improvement for the clustering in each family, but this improvement not yet significant to cluster each protein family and distinguish each family from the other.

Table 4.3: Results for the first 100 nucleotides

	Lysozyme		Myoglobin					Rhodopsin	
	Bos taurus	Homo sapiens	Danio rerio	Homo sapiens	Mus musculus	Rattus norvegicus	Sus scrofa	Homo sapiens	Rattus norvegicus
	I	II	III	IV	V	VI	VII	VIII	IX
I	0.00E+00	3.81E+02	1.06E+03	8.94E+02	1.25E+03	1.39E+03	9.17E+02	1.09E+03	1.56E+03
II		0.00E+00	8.01E+02	5.91E+02	9.37E+02	1.14E+03	5.80E+02	1.04E+03	1.28E+03
III			0.00E+00	7.30E+02	4.32E+02	6.53E+02	6.86E+02	9.91E+02	7.84E+02
IV				0.00E+00	8.13E+02	1.04E+03	2.35E+02	8.57E+02	8.64E+02
V					0.00E+00	5.55E+02	7.14E+02	1.30E+03	9.56E+02
VI						0.00E+00	9.30E+02	1.35E+03	1.10E+03
VII							0.00E+00	9.41E+02	8.69E+02
VIII								0.00E+00	9.53E+02
IX									0.00E+00

By increasing the number of participated nucleotides in the comparison process for the DNA sequences from different families, each family has been clustered in the proper way, and the difference between families become more significant. Table 4.4, Table 4.5 and Table 4.6 shows the results for the first 200, 300 and 400 nucleotides respectively.

Table 4.4: Results for the first 200 nucleotides

	Lysozyme		Myoglobin					Rhodopsin	
	Bos taurus	Homo sapiens	Danio rerio	Homo sapiens	Mus musculus	Rattus norvegicus	Sus scrofa	Homo sapiens	Rattus norvegicus
	I	II	III	IV	V	VI	VII	VIII	IX
I	0.00E+00	1.15E+03	2.60E+03	3.03E+03	2.26E+03	1.92E+03	2.58E+03	4.34E+03	2.76E+03
II		0.00E+00	2.34E+03	2.26E+03	1.64E+03	1.76E+03	1.86E+03	3.93E+03	2.26E+03
III			0.00E+00	2.14E+03	1.80E+03	3.01E+03	1.97E+03	2.16E+03	1.30E+03
IV				0.00E+00	8.89E+02	2.79E+03	8.00E+02	2.33E+03	1.30E+03
V					0.00E+00	2.32E+03	4.09E+02	2.62E+03	1.23E+03
VI						0.00E+00	2.59E+03	4.53E+03	2.80E+03
VII							0.00E+00	2.58E+03	1.40E+03
VIII								0.00E+00	1.91E+03
IX									0.00E+00

Table 4.5: Results for the first 300 nucleotides

	Lysozyme		Myoglobin					Rhodopsin	
	Bos taurus	Homo sapiens	Danio rerio	Homo sapiens	Mus musculus	Rattus norvegicus	Sus scrofa	Homo sapiens	Rattus norvegicus
	I	II	III	IV	V	VI	VII	VIII	IX
I	0.00E+00	1.78E+03	7.51E+03	5.69E+03	4.31E+03	5.17E+03	6.00E+03	9.61E+03	7.06E+03
II		0.00E+00	6.25E+03	4.29E+03	3.06E+03	4.38E+03	4.49E+03	9.03E+03	6.26E+03
III			0.00E+00	2.64E+03	3.48E+03	4.87E+03	2.97E+03	4.58E+03	2.65E+03
IV				0.00E+00	2.10E+03	3.42E+03	1.60E+03	6.39E+03	3.60E+03
V					0.00E+00	3.50E+03	2.05E+03	6.65E+03	4.21E+03
VI						0.00E+00	3.76E+03	6.40E+03	4.56E+03
VII							0.00E+00	6.85E+03	4.65E+03
VIII								0.00E+00	3.56E+03
IX									0.00E+00

Table 4.6: Results for the first 400 nucleotides

	Lysozyme		Myoglobin					Rhodopsin	
	Bos taurus	Homo sapiens	Danio rerio	Homo sapiens	Mus musculus	Rattus norvegicus	Sus scrofa	Homo sapiens	Rattus norvegicus
	I	II	III	IV	V	VI	VII	VIII	IX
I	0.00E+00	3.54E+03	9.64E+03	7.34E+03	6.76E+03	7.65E+03	9.07E+03	1.52E+04	1.23E+04
II		0.00E+00	9.73E+03	7.12E+03	7.25E+03	8.11E+03	9.02E+03	1.35E+04	1.06E+04
III			0.00E+00	4.04E+03	4.41E+03	6.65E+03	3.77E+03	9.33E+03	6.50E+03
IV				0.00E+00	3.89E+03	5.11E+03	2.37E+03	9.66E+03	6.19E+03
V					0.00E+00	4.11E+03	4.08E+03	1.04E+04	7.96E+03
VI						0.00E+00	5.29E+03	1.01E+04	8.77E+03
VII							0.00E+00	9.31E+03	6.29E+03
VIII								0.00E+00	4.64E+03
IX									0.00E+00

The first 400 nucleotides experiment shows significant clustering results Table 4.6. The distances between Lysozyme family is 3.54E+03 and the distance between it and the other families range from 6.76E+03 to 1.52E+04. Myoglobin family sequences range from 2.37E+03 - 6.65E+03, the distances between the Myoglobin family and the other families range from 6.19E+03 - 9.66E+03. For the third family (Rhodopsin), the distance inside this family is 4.64E+03, and the distances between it and the other families range from 6.19 - 1.52E+04. Form the previous results it is easy to distinguish the family cluster and the differences between the families.

First 500 Nucleotides

This experiment has included the first 500 nucleotides of each sequence, and this number of nucleotides has been chosen because the shortest sequence length is 505.

Table 4.7 shows significant results for clustering each family, the vector distance for the Myoglobin family are clustered together ranging from 3.66E+03 - 8.89E+03, also for all families the distance become significantly large between different families Table 4.8.

Table 4.7: Results for the first 500 nucleotides

	Lysozyme		Myoglobin					Rhodopsin	
	Bos taurus	Homo sapiens	Danio rerio	Homo sapiens	Mus musculus	Rattus norvegicus	Sus scrofa	Homo sapiens	Rattus norvegicus
	I	II	III	IV	V	VI	VII	VIII	IX
I	0.00E+00	3.55E+03	1.79E+04	1.49E+04	1.35E+04	1.41E+04	1.73E+04	2.07E+04	1.84E+04
II		0.00E+00	1.83E+04	1.53E+04	1.39E+04	1.53E+04	1.77E+04	2.02E+04	1.79E+04
III			0.00E+00	4.69E+03	8.89E+03	6.96E+03	5.68E+03	1.03E+04	6.74E+03
IV				0.00E+00	5.70E+03	4.75E+03	3.66E+03	1.07E+04	6.27E+03
V					0.00E+00	5.29E+03	5.92E+03	1.17E+04	7.64E+03
VI						0.00E+00	5.55E+03	1.16E+04	8.86E+03
VII							0.00E+00	9.41E+03	5.20E+03
VIII								0.00E+00	5.97E+03
IX									0.00E+00

A complete set of the corresponding characterization vectors in each of the result table can be found in the appendix.

4.3 Discussion

As preceded in the previous section, the experiments that have been implemented on three protein families Lysozyme, Myoglobin and Rhodopsin showed significant results for the first 400 and 500 nucleotides. The results for the first 400 nucleotides are nearly close to the first 500, but it is distinguishable that the first 500 nucleotides have shown more significant discrimination among the three families.

As a summary for all the experiments that have been applied, Table 4.8 show the distances inside and outside the families for all the members of the families.

Table 4.8: Distances inside and outside families in different experiments

Experiment	Protein Family	Distance Inside Family		Distance with Other Families	
		From	To	From	To
First 100 Nucleotides	Lysozyme	3.81E+02		5.80E+02	1.56E+03
	Myoglobin	2.35E+02	1.04E+03	5.80E+02	1.39E+03
	Rhodopsin	9.53E+02		7.84E+02	1.56E+03
First 200 Nucleotides	Lysozyme	1.15E+03		1.64E+03	4.34E+03
	Myoglobin	4.09E+02	2,79E+03	1.23E+03	4.53E+03
	Rhodopsin	1.91E+03		1.23E+03	4.53E+03
First 300 Nucleotides	Lysozyme	1.78E+03		3.06E+03	9.61E+03
	Myoglobin	1.60E+03	4.87E+03	2.65E+03	7.51E+03
	Rhodopsin	3.56E+03		2.65E+03	9.61E+03
First 400 Nucleotides	Lysozyme	3.54E+03		6.76E+03	1.52E+04
	Myoglobin	2.37E+03	6.65E+03	6.19E+03	1.04E+04
	Rhodopsin	4.64E+03		6.19E+03	1.52E+04
First 500 Nucleotides	Lysozyme	3.55E+03		1.35E+04	2.07E+04
	Myoglobin	3.66E+03	8.89E+03	5.20E+03	1.83E+04
	Rhodopsin	5.97E+03		5.20E+03	2.07E+04
Whole Sequence	Lysozyme	4.05E+05		9.80E+04	5.78E+05
	Myoglobin	1.12E+05	4.67E+05	9.80E+04	6.97E+05
	Rhodopsin	1.15E+05		2.28E+05	6.97E+05

For good clustering, Distance Inside Family is minimized whereas the Distance with Other Families is maximized. From Table 4.8, Distance Inside Family for Lysozyme family is minimal for First 500 Nucleotides and maximal for Distance with Other

Families. For the Myoglobin family the minimal Distance Inside Family and maximal in Distance with Other Families shown by the Whole Sequence. For the Rhodopsin family its clearly seen that is has interchangeable values that it has the maximal distance in Distance with Other Families in the First 100 Nucleotides.

In all of the previous experiments in section 4.2.2, the number of the nucleotides that have been associated in the global comparison between the DNA sequences did not exceed the maximum length of the shortest DNA sequence which is referred to the "Mus musculus" sequence from the Myoglobin family with a length of 505(bp). These indications will give this work a novelty in distinguishing the protein families that have DNA sequences nearly the same length. In the same family the length of nucleotides vary, so if there is a big difference in the length of the DNA sequences this will lead to a big difference in distance in the same family and this will reflect on the distances between that family and other families.

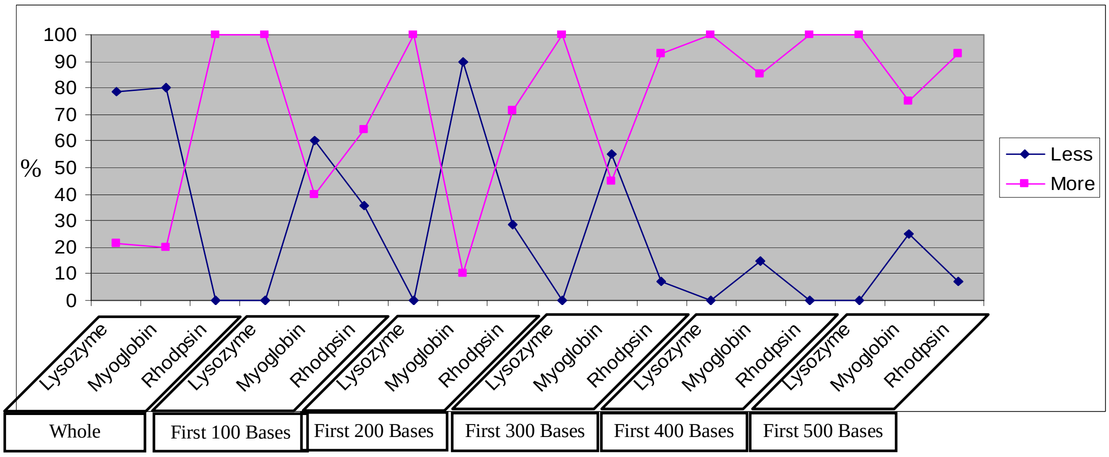


Figure 4.5: All experiment results

All the experiments can be illustrated graphically in Fig 4.5, the lines indicates the differences that is less and more than the maximum distance inside each family, the results for the whole DNA sequence, first 100 bases, first 200 bases, and first 300 bases shows interchangeable values between the percentage of the distances less and more than the maximum distance inside the same protein family. For the first 400 bases and first 500 bases the results shows significant distinguish between the three protein families and this due to the information that is extracted from the DNA sequence to produce the characterization vector that is used to compute the distance become more significant to the DNA sequence. In all cases the numbers of nucleotides that have to be included in the comparison have not to exceed the maximum length of the shortest DNA sequence. This will reduce the number of nucleotides that are associated in the experiments and will reduce the overall calculations for determining the cluster of the new DNA sequence.

As a conclusion, there is no need to compare the new DNA sequences with all the whole DNA sequences, number of nucleotides included in the comparison has not exceed the length of the new DNA sequence. For more details for conclusion and future work will be found in the next chapter.

CHAPTER FIVE

CONCLUSION AND FUTURE WORK

This chapter will conclude what have been done in this study with the results that have been obtained, in addition to the future work and further research.

5.1 Conclusion

The results of DNA sequences comparison among homologous sequences give close distances between their characterization vectors which are easily distinguishable from non-homologous in experiment it with a fixed DNA sequence size that does not exceed the maximum length of the shortest DNA sequence.

Global comparison for multiple DNA sequences simultaneously presented in the genomic space is the main advantage of this work by applying direct comparison of the corresponding characteristic vectors distances. The novelty of this work is that for the new DNA sequence, there is no need to compare the new DNA sequence with the whole

DNA sequences length, just the comparison focused on a fixed number of all the sequences in a way that does not exceed the maximum length of the new DNA sequence. In other words, parts of the DNA sequence can identify the functionality of the DNA sequence, and make it clustered with its family members.

5.2 Future Work

Future work will emphasize on determining the ranges of DNA sequences lengths that have to be included in the comparison (the number of nucleotides). In order not to include all the DNA sequences that vary in the length, just those DNA sequences that fall in this range will be included to reduce the overhead calculations.

Another future work part is to extend this study to be implemented on other protein families that have other amino acids.

REFERENCES

Abonyi, J., & Feil, B. (2005). Computational Intelligence in Data Mining. *Informatica, 29,* 3-12.

Aksoy, S., & Haralick, R. M. (1999). Graph–Theoretic Clustering for Image Grouping and Retrieval. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'99), 1,* 1063.

Anastassiou, D. (2000). Frequency-domain analysis of biomolecular sequences. *Bioinformatics, 16*(4), 1073-1081.

Ansari, A., & Viswanathan, R. (1992). Application of Expectation-Maximization Algorithm to the Detection of Direct-Sequence Signal in pulsed Noise Jamming. *IEEE Military Communications Conference, 3*, 811-815.

Apon, A., Mache, J., Buyya, R., & Jin, H. (2004). Cluster Computing in the Classroom and Integration with Computing Curricula 2001. *IEEE Transactions on Education, 47*(2), 188-195.

Arasa, N., Oommenb, B. J., & Altınelc, I. K. (1999). The Kohonen network incorporating explicit statistics and its application to the travelling salesman problem. *Neural Networks, 12*(9), 1273-1284.

Ayre, L. B. (2006). *Data Mining for Information Professionals*.

Bach, F. R., & Jordan, M. I. (2003). Learning Spectral Clustering. *Learning graphical models with Mercer kernels in Advances Neural Inform, 1*, 1009-1016.

Bolshoy, A., & Volkovich, Z. (2008). Whole-genome prokaryotic clustering based on gene lengths. *Discrete Applied Mathematics, 157*(10), 2370-2377.

Borman, S. (2009). *The Expectation Maximization Algorithm A short tutorial*.

Carvalho, F. A. T. (2006). *Fuzzy clustering algorithms for symbolic interval data based on adaptive and non-adaptive Euclidean distances*.

Draghici S., Graziano, F., Kettoola, S., Sethi, I., & Towfic, G. (2003). Mining HIV dynamics using independent component analysis. *Bioinformatics, 19*(8), 981-986.

Erban, G., & Moldovan, G. S. (2006). A Comparison of Clustering Techniques in Aspect Mining. *Informatica, 1*, 69-78.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases*.

FitzGerald, P. C., Shlyakhtenko, A., Mir, A. A., & Vinson, C. (2004). Clustering of DNA Sequences in Human Promoters. *Genome Res, 14*, 1562-1574.

Gates, M. A. (1985). Simpler DNA sequence representations. *Nature, 31*, 219.

Ghanem M., Chortaras, A., Guo, Y., Rowe, A., & Ratcliffe, J. (2005). *A Grid Infrastructure for Mixed Bioinformatics Data and Text Mining.*

Graham, J., Page, C. D., & Kamal, A. (2003). *Accelerating the Drug Design Process through Parallel Inductive Logic Programming Data Mining*.

Grammalidis, N., Bleris, L., & Strintzis, M. G. (2002). *Using the Expectation-Maximization Algorithm for Depth Estimation and Segmentation of Multi-view Images*.

Guinepain, S., & Gruenwald, L. (2006). *Automatic Database Clustering Using Data Mining*.

Guo, X., & Nandy, A. (2002). *Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy.*

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). *Biological identifications through DNA barcodes*.

Hu, X. O., & Pan, Y. (Eds.). (2007). *Knowledge Discovery in Bioinformatics Techniques, Methods, and Applications.* Hoboken: Wiley.

Huang, G., Liao, B., Li, Y., & Yu, Y. (2009). *Similarity studies of DNA sequences based on a new 2D graphical representation.*

Irene, M. M. (1999). *Hierarchical Clustering*. Retrieved September 29, 2009, from http://www.cse.iitb.ac.in/dbms/Data/Courses/CS632/1999/clustering/node3.html

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Upper Saddle River: Prentice-Hall.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. *ACM Computing Surveys, 31*(3).

Jenssen, R., Hild, K. E., Erdogmus, D., Principe, J. C., & Eltoft, T. (n.d.). *Clustering using Renyi's Entropy*.

Kauer, G., & Blocker, H. (2003). Applying signal theory to the analysis of biomolecules. *Bioinformatics, 19*(16), 2016-2021.

Kozobay-Avrahama, L., Hosid, S., Volkovich, Z., & Bolshoy, A. (2008). *Prokaryote clustering based on DNA curvature distributions.*

Liu, L., Ho, Y., & Yau, S. (2006). *Clustering DNA sequences by feature vectors*.

Lv, T., Huang, S., Zhang, X., & Wang, Z. (2006). *A Robust Hierarchical Clustering Algorithm and its Application in 3D Model Retrieval*.

Myller, N., Suhonen, J., & Sutinen, E. (2002). *Using Data Mining for Improving Web-Based Course Design*.

Ng, H. P., Ong, S. H., Foong, K. W. C., Goh, P. S., & Nowinski, W. L. (2006). *Medical Image Segmentation Using K-Means Clustering and Improved Watershed Algorithm*.

Paccanaro, A., Casbon, J. A., & Saqi, M. A. S. (2006). *Spectral clustering of protein sequences*.

Palace, B. (1996). *Data Mining.* Retrieved September 29, 2009, from http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/index.htm

Qi, Z., & Qi, X., (2009). *Numerical characterization of DNA sequences based on digital signal method*.

Randi, M., Vracko, M., Ler, N., & Plavsi, D. (2002). *Novel 2-D graphical representation of DNA sequences and their numerical characterization.*

Schenker, A. (2003). *Graph-Theoretic Techniques for Web Content Mining*.

Silverman, B. D., & Linsker, R. (1986). *A measure of DNA periodicity*.

Silverman, J. F., & Cooper, D. B. (1988). *Bayesian Clustering for Unsupervised Estimation of Surface and Texture Models.*

Song, J., & Tang, H. (2005). *A new 2-D graphical representation of DNA sequences and their numerical characterization.*

Stoeckle, M. (2003). Taxonomy, DNA, and the Bar Code of Life. *BioScience, 3*(9), 796-797.

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Boston: Pearson Education.

Valgren, C., Duckett, T., & Lilienthal, A. (2007). Incremental Spectral Clustering and Its Application To Topological Mapping. *IEEE International Conference on Robotics and Automation*.

Vinod, V. V., Chaudhury, S., Mukherjee, J., & Ghose, S. (1994). *A Connectionist Approach for Clustering with Applications in Image Analysis.*

Visnick, L. (2003). *Clustering Techniques*.

Voss, R. (1992). Evolution of long-range fractal correlation and 1/f noise in DNA base sequences. *Physical Review Letters*, 68, 3805-3808.

Wang, W., & Johnson, D. H. (2002). Computing linear transforms of symbolic signals Signal Processing. *IEEE Trans. Sig. Proc., 50*(3), 628-634.

Weiming, H. X. L., & Zhang, Z. (2007). *Corner Detection of Contour Images Using Spectral Clustering*.

XL Miner (n.d.). *Hierarchical Clustering*. Retrieved September 29, 2009, from http://www.resample.com/xlminer/help/HClst/HClst_intro.htm

Zhang, H., Ho, T., & Linz, M. (2004). *An Evolutionary K-Means Algorithm for Clustering Time Series Data*.

Zhang, Q., Peng, Q., & Xu, T. (2008). *DNA splice site sequences clustering method for conservativeness analysis.*

Zien, A. , Ratsch, G., Mika, S., Scholkopf, B., Lemmen, C., Smola, A., Lengauer, T., & Muller, K. R. (n.d.). *Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites.*

Appendix

This appendix contains the characterization vectors that have been obtained from conducting the experiments. The Latin numbers in the following table are used in the following tables to indicate the DNA sequence and the family belonging to.

	Lysozyme		Myoglobin					Rhodopsin	
	Bos taurus	Homo sapiens	Danio rerio	Homo sapiens	Mus musculus	Rattus norvegicus	Sus scrofa	Homo sapiens	Rattus norvegicus
	I	II	III	IV	V	VI	VII	VIII	IX

Whole DNA sequence characterization vectors

	nA	nT	nC	nG	tA	tT	tC	tG	dA	dT	dC	dG
I	276	277	155	198	132608	133071	64150	81042	67588.44	70538.11	65868.82	63326.28
II	435	438	306	308	344405	339226	218925	203772	196744.23	183621.25	157667.92	181834.44
III	402	340	288	328	291887	261556	170330	199424	151300.93	155066.07	143049.54	143072.70
VI	356	242	278	314	264119	139662	149610	159430	149983.09	99333.04	95233.57	92678.64
V	135	96	123	151	35385	24539	32870	34971	19455.83	20629.20	21210.31	22585.99
VI	303	197	251	264	176247	96720	122451	120202	103618.72	80336.52	73433.89	71774.74
VII	234	212	310	355	128573	119903	176528	192712	109720.01	104229.91	104802.61	95409.95
VIII	328	314	518	460	267116	238401	412714	394779	201810.10	213958.35	225327.79	222221.95
IX	372	302	416	470	308336	229825	311717	367702	193672.25	201757.14	211208.41	200292.94

First 100 nucleotides characterization vectors

	nA	nT	nC	nG	tA	tT	tC	tG	dA	dT	dC	dG
I	18	37	21	24	811	1873	955	1411	1080.274691	742.1271	668.820862	819.248264
II	19	30	24	27	902	1565	984	1599	1095.091413	669.738889	655.666667	822.691358
III	25	20	27	28	1080	1068	1180	1722	622.88	955.94	720.504801	823.035714
VI	18	22	28	32	855	1193	1325	1677	984.6944444	398.993802	676.432398	1156.74121
V	22	17	22	39	969	937	1060	2084	659.8615702	888.927336	648.694215	966.451019
VI	18	24	22	36	944	969	1025	2112	531.5802469	939.984375	1182.33264	553.333333
VII	18	21	30	31	844	1137	1278	1791	1092.098765	537.931973	753.173333	830.626431
VIII	15	25	37	23	692	1370	1794	1194	738.5155556	746	883.276844	868.340265
IX	19	15	36	30	1139	622	1728	1561	748.6814404	703.448889	800.055556	884.365556

First 200 nucleotides characterization vectors

	nA	nT	nC	nG	tA	tT	tC	tG	dA	dT	dC	dG
I	52	58	38	52	6098	5028	3435	5539	3803.773669	2985.6968	3273.1863	2689.4804
II	50	51	41	58	5592	4688	3597	6223	3495.7344	2988.386	3882.7329	2771.0348
III	48	47	52	53	4419	5392	5048	5241	3224.766927	3665.8171	3818.9556	2402.0627
VI	39	42	52	67	4073	4192	5075	6760	3637.835634	2952.3447	3654.7023	3129.2279
V	45	40	45	70	4530	4487	4614	6469	3875.288889	3263.9444	3446.8711	2805.4141
VI	53	42	44	61	6390	3534	4193	5983	3152.547526	3378.0272	3216.0263	2984.5343
VII	40	42	52	66	4317	4446	4643	6694	3999.519375	3392.3605	3637.4745	2500.5776
VIII	29	51	66	54	2758	5403	6115	5824	3448.989298	3307.3495	3330.6816	3125.7558
IX	38	44	58	60	4091	4895	4991	6123	2960.067175	3307.233	3214.2904	3382.2808

First 300 nucleotides characterization vectors

	nA	nT	nC	nG	tA	tT	tC	tG	dA	dT	dC	dG
I	87	80	59	74	14816	10482	8687	11165	6797.312987	7534.7244	8352.6556	6728.9176
II	84	72	61	83	14012	9940	8647	12551	6881.130385	7451.7469	8907.4969	6673.1819
III	69	60	79	92	9806	8616	11983	14745	8191.667717	6082.44	8541.1277	6831.0022
VI	68	61	73	98	11026	8951	10524	14649	6808.537197	7190.0296	8204.4387	7513.8618
V	77	58	66	99	12399	8965	10103	13683	7773.635689	6526.2452	8035.464	7249.1368
VI	79	58	74	89	12829	7563	11787	12971	5919.453613	8324.3427	8263.2573	7255.1355
VII	70	56	74	100	11727	7975	10166	15282	7438.563469	6751.992	8225.1812	7052.5876
VIII	55	61	105	79	9359	7894	15638	12259	8461.991405	5720.0451	7799.567	7168.399
IX	60	63	90	87	9716	9607	13054	12773	7294.962222	6517.3928	8627.0869	7049.3225

First 400 nucleotides characterization vectors

	nA	nT	nC	nG	tA	tT	tC	tG	dA	dT	dC	dG
I	116	102	78	104	25051	18113	15299	21737	11523.66194	13978	13979	13366.51
II	108	102	82	108	22537	20261	15956	21446	11690.92275	14306	14946	12731.93
III	101	76	108	115	20982	14260	21859	23099	15167.6169	12216	13430	12210.4
VI	97	79	101	123	21076	15353	20433	23338	12142.13902	13489	14939	12489.09
V	108	76	97	119	23265	15527	20869	20539	13147.22454	13149	13849	12034.8
VI	108	79	102	111	22963	15131	21283	20823	11398.14292	16604	12662	13076.76
VII	97	73	108	122	21280	14052	22097	22771	12570.5246	13631	15711	11127.65
VIII	71	89	137	103	15120	17783	26563	20734	13036.04047	14954	12769	12722.66
IX	79	85	121	115	16466	17333	23819	22582	12509.23249	12625	14456	13158.01

First 500 nucleotides characterization vectors

	nA	nT	nC	nG	tA	tT	tC	tG	dA	dT	dC	dG
I	138	134	97	131	34800	32867	23894	33689	16719.17265	25441.498	21745.582	19689.636
II	133	133	100	134	33701	34492	24206	32851	18261.90729	23280.66	22684.636	19407.087
III	119	93	133	155	29253	22096	32985	40916	21092.27978	21312.994	20034.248	20728.477
VI	119	97	127	157	30971	23396	32094	38789	18223.78928	20763.312	21907.057	21844.263
V	133	95	122	150	34380	24036	32365	34469	18866.32517	20195.358	20916.991	22245.857
VI	130	96	127	147	32809	22843	32517	37081	17360.66562	23792.508	19485.046	23000.855
VII	119	90	136	155	31295	21593	34808	37554	18785.57955	20877.938	22802.129	20373.765
VIII	94	106	166	134	25520	25347	39557	34826	20605.31372	20758.73	20004.979	21393.914
IX	106	100	148	146	28666	24037	36070	36477	20811.03809	18358.673	21833.406	21124.571