Protein sequence for clustering DNA based on Artificial Neural Networks

Gamal. F. Elhadi[1], R. M. Farouk[2] and Abdalhakeem. T. Issa[3]

[1] Computer Science Department, Faculty of Computers and Information's,
Menofia University, Menofia, Egypt.

[2] Department of Mathematics, Faculty of Science, Zagazig University,
Zagazig, Egypt.

[3] Department of Computer Engineering, DCC, Shaqra University, KSA

Abstract

DNA is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms and some viruses. Clustering is a process that groups a set of objects into clusters so that the similarity among objects in the same cluster is high, while that among the objects in different clusters is low. In this paper, we proposed an approach for clustering DNA sequences using Self-Organizing Map (SOM) algorithm and Protein Sequence. The main objective is to analyze biological data and to bunch DNA to many clusters more easily and efficiently. We use the proposed approach to analyze both large and small amount of input DNA sequences. The results show that the similarity of the sequences does not depend on the amount of input sequences. Our approach depends on evaluating the degree of the DNA sequences similarity using the hierarchal representation Dendrogram. Representing large amount of data using hierarchal tree gives the ability to compare large sequences efficiently

Keywords: *DNA Sequences, Protein Sequences, ANN, Clustering*

1. Introduction

Most biomacromolecules, such as proteins and nucleic acids, occur in preferred conformations. Examples include n-helices and r-sheets in proteins in nucleic acids. These preferred conformations are the basic keys for structural stability and biological activity of the molecules. Therefore, biological importance to understand the relation between a preferred conformation and the responsible structural parameters. Which structural parameters can be used to study this relation. Possible candidates in the case of nucleic acids are helical parameters, such as, roll, twist and rise [10].

A critical problem in bio-data analysis is to classify bio-sequences or structures based on their critical features and functions. For instance, gene sequences isolated from diseased and healthy tissues can be compared to identify critical differences between the two classes of genes [2]. Such features can be used for classifying bio-data and predicting behaviors. Some approaches have been developed for bio-data classification. For example, one can first retrieve the gene sequences from the two tissue classes and then find and compare the frequently occurring patterns of each class. Usually, sequences occurring more frequently in the diseased samples than in the healthy samples indicate the genetic factors of the disease: those occurring only more frequently in the healthy samples might indicate mechanisms that protect the body from the disease. Similar analysis can be performed on microarray data and protein data to identify similar and dissimilar patterns [7,8].

This work assumes that there is an unknown mapping called clustering structure that assigns a class label to each observation, and the goal of cluster analysis is to estimate this clustering structure, that is, to estimate the number of clusters and cluster assignments. In traditional cluster analysis, it is assumed that such unknown mapping is unique. However, since the observations may cluster in more than one way depending on the variables used, it is natural to permit the existence of more than one clustering structure [12,13,17,23]. This generalized clustering problem of estimating multiple clustering structures is the focus of this paper. The contribution of this paper is to propose a new approach to cluster large DNA data set more efficiently. The proposed system enables researchers to analyze biological data in ease and rapidly using SOM is a specific kind of two-layer Artificial Neural Network

(ANN) can be illustrate by Fig .1, this work proposes an algorithm for finding multiple clustering structures involving clustering variables and observations. The number of clustering structures is determined by the number of variable clusters. The dissimilarity measure for clustering variables is based on nearest-neighbor graphs. The observations are clustered using weighted distances with weights determined by the clusters of the variables.

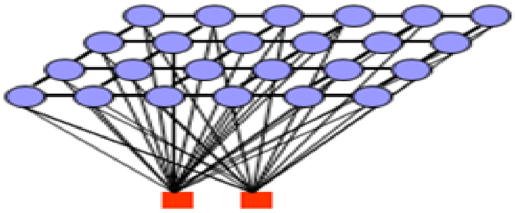

Fig.1 *Input* x_1 & x_2 SOM layers

The rest of this paper is organized as follows: in sec.2; we illustrate the previous work, in sec3, we discuss our problem methodology, sec.4 include SOM and system analysis algorithm, result and conclusion is given and in sec. 5 we present the future work in sec .6.

2. Previous Work

The Basic Local Alignment Search Tool (BLAST) is typically the first Bioinformatics tool that biologists use when examining a new DNA sequence [5,18]. BLAST compares the new sequence to all sequences in the database to find the most similar known sequences [4, 14]. BLAST use Clusters of Orthologous Groups (COGs) [1] tool to compare DNA sequences encoded in 43 complete genomes, representing 30 major phylogenetic lineages. Each COGs consists of groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain. It also uses VAST search to structure similarity search service [4]. It compares 3-D coordinates of a newly determined protein structure to those in the MMDB/PDB database [6].

For the sake of completeness of our investigations, let us offer some introductory notes about the underlying DNA processing. Basically, DNA consists of polymer chains, usually refereed to as DNA strands. This chain is composed of nucleotides, and nucleotides may differ only in their bases. There are four bases which are A (adenine), G (guanine), C (cytosine) and T (thymine).

The familiar double helix of DNA arises by the bonding of two separate strands known as Watson–Crick complementarily, which comes in the formation of such double strands. These four bases are bounded: A always bonds with T and G with C. Fig. 2 shows DNA bases in a double helix form.

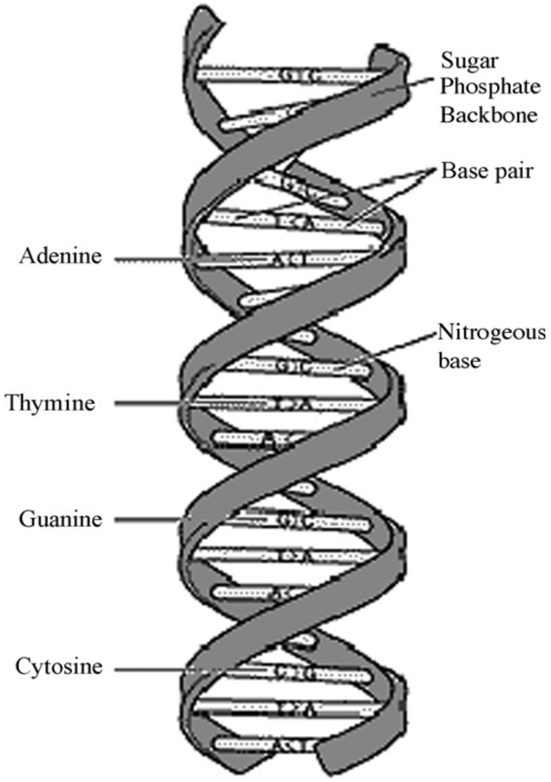

Fig. 2. DNA in a double helix form

Characterizing the DNA-binding specificities of transcription factors is a key problem in computational biology that has been addressed by multiple algorithms. These usually take as input sequences that are putatively bound by the same factor and output one or more DNA motifs [20]. A common practice is to apply several such algorithms simultaneously to improve coverage at the price of redundancy. In interpreting such results, two tasks are crucial: *clustering* of redundant motifs, and attributing the motifs to transcription factors by more DNA motifs. In interpreting such results, two tasks are involving motif comparison. Microarray technology has made it possible to simultaneously measure the expression levels of large numbers of genes in a short time. Gene expression data is information rich; however, extensive data mining is required to identify the patterns that characterize the underlying mechanisms of action. Clustering is an important tool for finding groups of genes with similar expression patterns in microarray data analysis. Hard clustering approaches assign each gene exactly to one cluster, are poorly suited to the analysis of microarray datasets because in such datasets the clusters of genes frequently overlap [10, 11].

3. Problem Methodology

In this paper, we propose an approach depends on microarray analysis approach, and the most suitable approach recently used in bioinformatics. We have used SOM approach for clustering DNA, which is very simple, easy to understand and efficient. Matlab tools are used to implement our proposed approach. Clustering DNA sequence is performed using SOM algorithm by comparing DNA sequences based on bases found in the sequences. Then DNA sequence is converted into protein sequence by using the genetic code and save this protein sequence in a database. In the next step clustering is performed based on protein sequence. After getting DNA sequence from the user or from a database, we have converted it into protein sequence by using the standard genetic code and then performing clustering on protein sequence. In addition, we will also perform similarity recognition between DNA sequences and plot a figure of the dendrogram illustrating this similarity. A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. Dendrograms are often used in computational biology to illustrate the clustering of genes. The proposed approach enables the user to insert to, update or open the database. The framework of our proposed approach is shown in Fig.3.

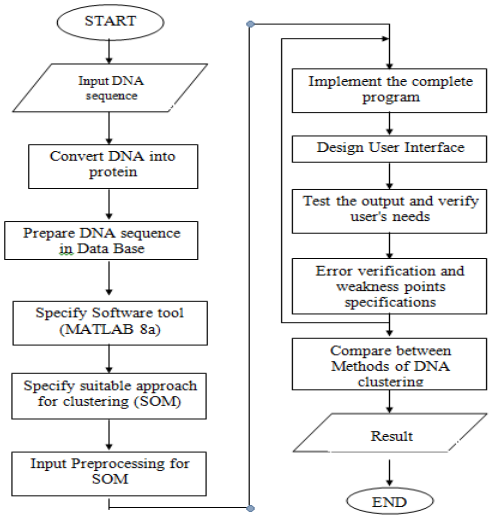
START
Input DNA sequence
Convert DNA into protein
Prepare DNA sequence in Data Base
Specify Software tool (MATLAB 8a)
Specify suitable approach for clustering (SOM)
Input Preprocessing for SOM
Implement the complete program
Design User Interface
Test the output and verify user's needs
Error verification and weakness points specifications
Compare between Methods of DNA clustering
Result
END

Fig.3. framework of proposed approach

3.1 Self-Organizing Maps

A self-organizing map (SOM) as an explicit brand of Artificial Neural Network (ANN), that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map. Self-organizing maps are different than other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space. This makes SOM useful for visualizing low-dimensional views of high-dimensional data, akin to multidimensional scaling. The model was first described as an artificial neural network by Kohonen map [15]. Like most artificial neural networks, SOMs operate in two modes: training and mapping. Training builds the map using input examples. It is a competitive process, also called vector quantization. Mapping automatically classifies a new input vector. A self-organizing map consists of components called nodes or neurons. Associated with each node is a weight vector of the same dimension as the input data vectors and a position in the map space, the usual arrangement of nodes is a regular spacing in a hexagonal or rectangular grid see Fig.4.

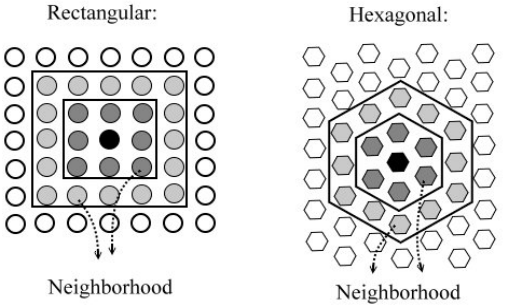

Fig.4. The neurons of the map can be arranged either on a rectangular or a hexagonal map

Self-organizing feature maps (SOFM) learn to classify input vectors according to how they are grouped in the input space. They differ from competitive layers in that neighboring neurons in the self-organizing map learn to recognize neighboring sections of the input space. Thus, self-organizing maps learn both the distribution (as do competitive layers) and topology of the input vectors they are trained on [16].

3.2Conversion DNA Sequence into Protein Sequence

This section includes the technique used to analysis the proposed system and SOM clustering algorithm. Our proposed approach divides the input into a number of clusters, the value of the clusters must be less than the number of the input sequences, as it is not logical to divide inputs into a number of clusters larger than the number of the inputs. So, by logic if the number of the inputs and the number of clusters are the same, each input will be in a separate cluster and this is of course meaningless [19, 21, 22]. In our proposed approach we accept data from database for DNA sequence [9] and converting each DNA sequence to protein sequence by searching for the starting code (ATG) from DNA sequence then converting each triple of DNA sequence base (A, C, G and T) to amino acids (using genetic code stored in amino acid database in our proposed approach) and we repeat that for each triple of the sequence bases until we reach the one of the ending codes (TAA, TGA, TAG), these amino acids form a protein sequence [15]. One major problem of SOM is that it requires a value for each dimension of each member of samples in order to generate a map. As we use two-dimensional map, we have to convert each protein sequence into two numbers. These two numbers must be chosen to distinguish each sequence from any others; we proposed a formula for this conversion as follows:

$$x_j = \sum_{i=1}^{M} A_i(X_i, Y_i) \Big/ \sum_{i=1}^{M} L_i \qquad (1)$$

Where A_i denotes to the ASCII code of each character (A, C, G and T), and (X_i, Y_i) the position of each character in the sequence, L_i the length sequence of protein sequence.

4. SOM Clustering Algorithm

After completing preprocessing function and getting the two input numbers, let the map of size *M* by *M* and the weight vector of neuron *i* is m_i , then we can apply the SOM clustering algorithm as the following steps:

Step 1: Initialize all weight vectors $m_i(0)$ randomly or systematically.

Step 2: A vector x_j is randomly chosen from the training data, then, compute the Euclidean distance *di* between x_j and neuron *i*

$$d_i = \| x - m_i(t) \|, \quad 1 \le i \le M^2 \qquad (2)$$

Step 3: Find the best matching neuron (winning node) *c*.

$$d_C = \| x - m_C(t) \| = \min\{ \| x - m_C(t) \| \}, \forall i \qquad (3)$$

Step 4: Update the weight vectors of the winning node *c* and its neighborhood as follows.

$$m_i(t+1) = m_t(t) + +\alpha(t)\, h_{c,i}(t)[x - mi(t)] \qquad (4)$$

Where $0 \le \alpha(t) \le 1$ is an adaptive function which decreases with time, the $h_{c,i}(t)$ is a neighborhood kernel centered at the winning node *c*, which decreases with time and the distance between neurons *c* and *i* in the topology map.

$$h_{c,i}(t) = \exp\left(- \|r_c - r_i\|^2 \big/ \sigma^2(t) \right) \qquad (5)$$

Where r_c and r_i are the coordinates of neurons *c* and $i, \sigma(t)$ is a suitable decreasing function of time,

Step 5: iterate the Step 2-4 until the sufficiently accurate map is acquired.

Our program allows the user to enter any sequence and perform similarity to all other inputs according to this sequence. Also the program plot a dendrogram graph to show the similarity tree according to specified sequence.

The distance Euclidian is the step to find the similarity between every pair of objects in the data set**.** To find the similarity we calculate the Euclidean distance between objects using the *(pdist)* function. Given an *m-by-n* data matrix *X*, which is treated as *(1-by-n)* row vectors x_1, x_2, x_3 ....., x_n the Euclidean distance between the vector x_r and x_s are defined from equation (6),

$$(pdist)^2 = (x_r - x_s) \qquad (6)$$

After we do similarity according to the processing stage will generate a dendrogram graph which shows the similarity between the specified input and all other inputs. The linkage function takes the distance information generated by *(pdist)* and links pairs of objects that are close together into binary clusters (clusters made up of two objects). The linkage function then links these newly formed clusters to other objects to create bigger clusters until all the objects in the original data set are linked together in a hierarchical tree. The dendrogram is a graphical representation of the results of hierarchical cluster analysis. This is a tree-like plot where each step of hierarchical clustering is represented as a fusion of two branches of the tree into a single one [15,16]. The branches represent clusters obtained on each step of hierarchical clustering.

5. Results and Conclusions

The dendrogram function plots this hierarchical tree information as a graph; the numbers along the horizontal axis represent the indices of the objects in the original data set. The links between objects are represented as upside down U-shaped lines. The height of the U indicates the distance between the objects.

Our approach is implemented using a MATLAB program, the input sequences of DNA sequence [3,11] to the program are clustered into groups with similar sequences. Fig.5. show the comparison of performance for clustering DNA sequences by K-Means, SOM and Linkage Algorithm, where exist similarity between this methods especial SOM Algorithm to clustering DNA after converting into Protein Sequence and K-Means Algorithm.

Fig.6. represents five clusters of two input DNA sequences and its graphical of dendrogram representation the hierarchical clustering of the result is shown in figure 5. We can remark that the subs clusters (1 and 4) are similar in inter cluster distance and also groups (2 and 5) have the same distance. Figure.6 show that two inputs represents ten DNA sequences and its dendrogram is shown in figure.7, in this figure it is clear there are some clusters have the same distance such as (1 and 14), (7 and 9). Figure.8, represents 50 two input sequences and its graphical representation is shown in figure.9, where number of clusters has the same distance such as (4 and 29) and (8 and 12) . Figure .10 represents one hundred two input sequences and its graphical representation is shown in figure.11. The same inter cluster distances refer to the similarity of these clusters. Categorizing the similarity is an efficient and fast process to discriminate the DNA sequences. The results show that our proposed approach is effective in small inputs and also in large inputs. Hierarchal representation of data using dendrogram enables to monitor large number of data easily. The results show that the inter distance between the subs clusters does not relate the amount of data. The inter distance may be long for small two input sequence and vice versa.

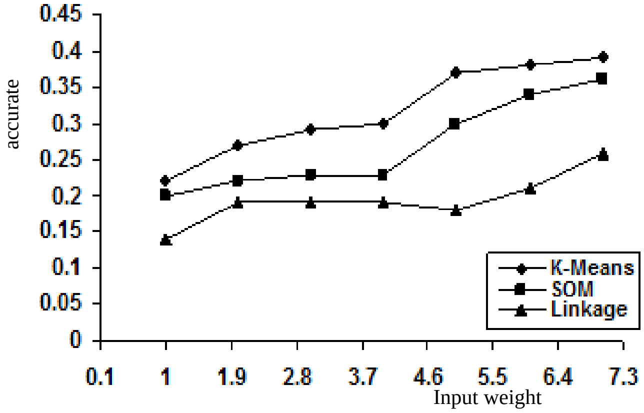

Fig.5. Comparison between K-Means, SOM, and Linkage clustering DNA sequence

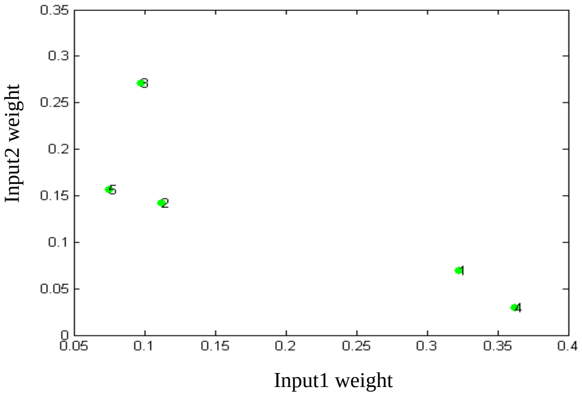

Fig.6. two inputs representation of 5 DNA sequences

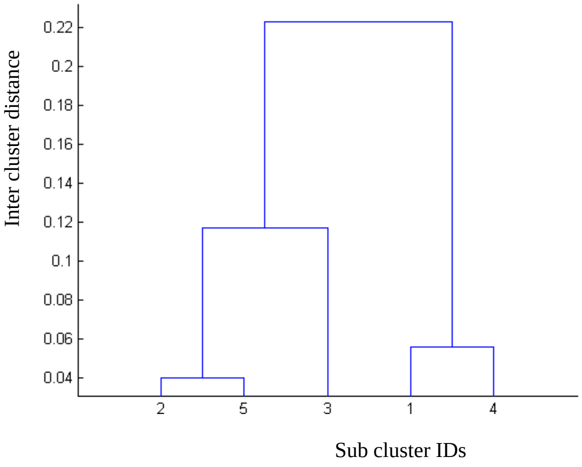

Fig.7. dendrogram of 5 sub clusters

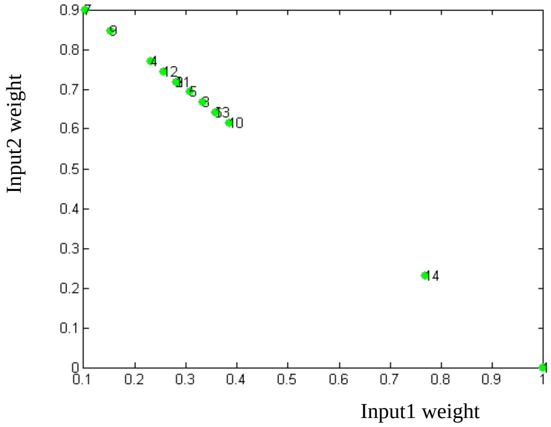

Fig.8. two inputs representation of 10 DNA sequences

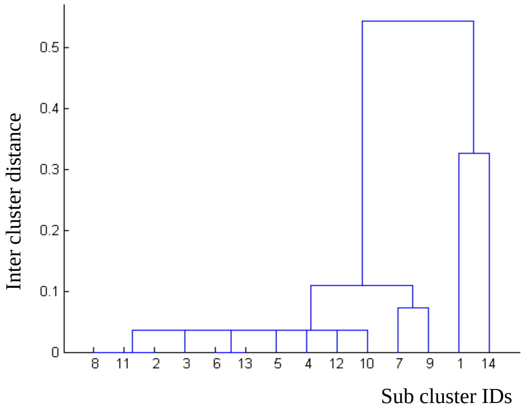

Fig.9. dendrogram of 10 sub clusters

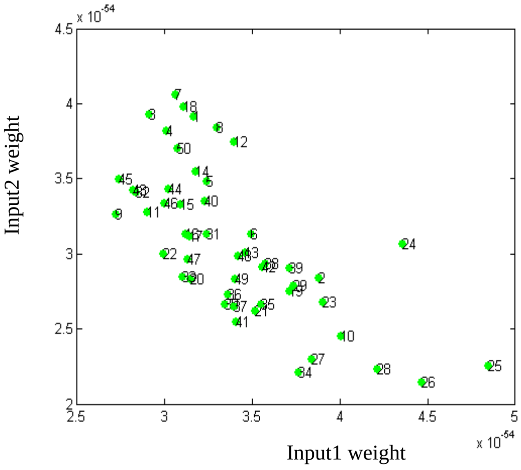

Fig.10. two inputs representation of 50 DNA sequences

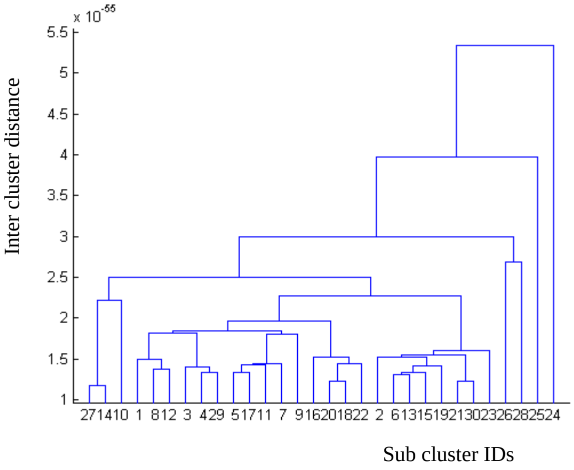

Fig.11. dendrogram of 50 sub clusters

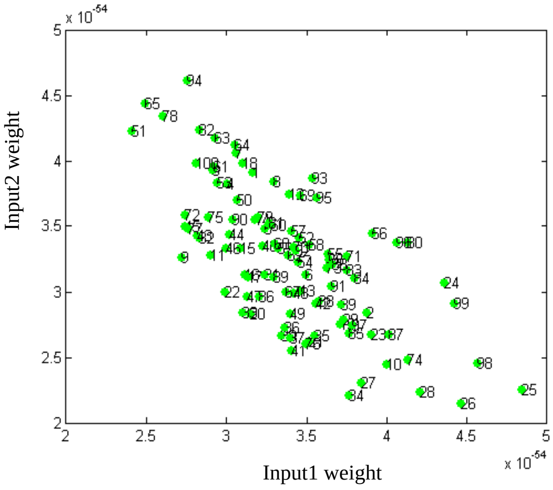

Fig.12. two inputs representation of 100 DNA sequences

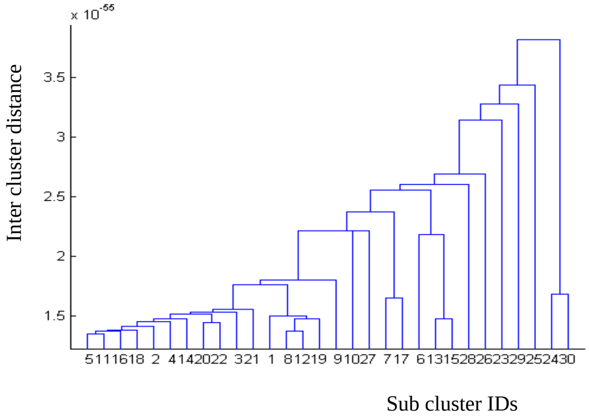

Fig.13. dendrogram of 100 sub clusters

This paper proposes an approach uses an efficient SOM way for clustering in DNA sequences, it takes input from database of DNA sequence [3,11]. This helps researches to save notes about the results in a database, or to compare it with previous one to get a new result for discovering diseases or diseases treatment. We divide DNA sequences into a number of cluster groups where the similarity among the sequences in the same cluster is high, while that among the sequences in different clusters is low. And this is very important for many genetic scientists to recognize the similarity or dissimilarity between sequences. The clusters are shown using a graph to be easy to understand. The proposed approach is also able to find similarity between DNA sequences. It finds the similarity between each pair of the inputs. Then, a dendrogram graph is displayed which clearly shows the similarity between the DNA sequences inputs. Moreover, the similarity can be performed according to a particular sequence. In addition, a conversion from DNA sequences into protein sequences can be performed. This operation is important for biologists. After the conversion, the protein sequences are stored in a database. Clustering operation can be performed on these protein sequences.

References

[1] Alberts, Bruce; Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walters (2002). Molecular Biology of the Cell; Fourth Editio.. New York and London: Garland Science. ISBN 0-8153-3218-1.

[2] Arthur M. Lesk , Introduction to Bioinformatics , ISBN (Pbk)0 19 925196 7,United States by Oxford University Press Inc, 2002.

[3] Anil K. Jain "Data clustering: 50 years beyond K-means" Pattern Recognition Letters 31 (2010) 651–666.

[4] Bock, C., S. Reither, T. Mikeska, M. Paulsen, J. Walter, and T. Lengauer. 2005. BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. Bioinformatics 21:4067-4068.

[5] Beck, S., and V.K. Rakyan. 2008. The methylome: approaches for global DNA methylation profiling. Trends Genet. 24:231-237.

[6] Butler, John M. Forensic DNA Typing "Elsevier". pp. 14–15. ISBN 978-0-12-147951-0, 2001.

[7] Christian Rohde , "New clustering module in BDPC bisulfite sequencing data presentation and compilation web application for DNA methylation analyses" , BioTechniques, Vol. 47, No. 3, September 2009, pp. 781–783.

[8] C. Yu, Q. Liang, C. Yin, R.L. He, S.S.-T. Yau, A novel construction of genome space with biological geometry, DNA Research 17 (2010) 155–168.

[9] Chenglong Yu, Mo Deng, Stephen S.-T. Yau, "DNA sequence comparison by a novel probabilistic method" Information Sciences 181 (2011) 1484–1492

[10] E1 Hassan, M. A. and Calladine, C. R. (1996) Propellertwisting of base-pairs and the conformational mobility of dinucleotide steps in DNA Journal of Molecular Biology 259, 95

[11] Ehrich, M., J. Turner, P. Gibbs, L. Lipton, M. Giovanneti, C. Cantor, and D. van den Boom. 2008. Cytosine methylation profiling of cancer cell lines. Proc. Natl. Acad. Sci. USA 105:4844-4849.

[12] Farthing, C.R., G. Ficz, R.K. Ng, C.F. Chan, S. Andrews, W. Dean, M. Hemberger, and W. Reik. 2008. Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes. PLoS Genet. 4:e1000116.

[13] Jason T.L. Wang, Mohammed J. Zaki, Hannu T.T. Toivonen and Dennis Shasha (Eds). Data Mining in Bioinformatics , ISBN 1852336714, Springer-Verlag London Limited 2005

[14] Kaufman, L., and P.J. Rousseeuw. 2005. Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, Inc., Hoboken, NJ.

[15] Kohonen, T. (1989) *Self-organization and associative memory,* 3rd edn. Springer-Verlag, Berlin-Heidelberg.

[16] Kohonen, T. (1995) *Self-organizing maps.* Springer-Verlag, Heidelberg.

[17] Ladd-Acosta, C., J. Pevsner, S. Sabunciyan, R.H. Yolken, M.J. Webster, T. Dinkins, P.A. Callinan, J.B. Fan. 2007. DNA methylation signatures within the human brain. Am. J. Hum. Genet. 81:1304-1315.

[18] Meissner, A., T.S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B.E. Bernstein. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454:766-770.

[19] Suzuki, M.M., and A. Bird. DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev, 2008, Genet. 9:465-476,

[20] S. L. Salzberg, D. B. Searls, and S. Kasif, eds., Computational approaches in Molecular Biology. Amsterdam: Elsevier Sciences B. V., 1998.

[21] G. Mecca et al., "A new algorithm for clustering search results", Data & Knowledge Engineering (2007) 504–522

[22] Zhang, Y., et al., "DNA methylation analysis by bisulfite conversion, cloning, and sequencing of individual clones.", Methods Mol. Biol, 2009. 507:177-187.

[23] Z. Francis, S. Incerti, R. Capra, B. Mascialino, G. Montarou, V. Stepan, C. Villagrasa, Molecular scale track structure simulations in liquid water using the Geant4-DNA Monte-Carlo processes, Appl. Radiat. Isot. 69 (1) (2011) 220–226.