International Journal of Scientific and Innovative Mathematical Research (IJSIMR)
Volume 7, Issue 3, 2019, PP 25-33
ISSN No. (Print) 2347-307X & ISSN No. (Online) 2347-3142
DOI: http://dx.doi.org/10.20431/2347-3142.0703003
www.arcjournals.org

Trinucleotides Based Species Identification by Genomic Taxonomy Using Self Organizing Feature Map

G. Manimannan[1], S. Jasmine Farzana[2], R. Lakshmi Priya[3], V. Suriya[4]

[1]*Assistant Professor, Department of Mathematics, TMG College of Arts and Science, Chennai*

[2]*Senior Post Graduate Students, Department of Statistics, SDNB Vaishnav College for Women, Chennai*

[3]*Assistant Professor, Department of Statistics, Dr. Ambedkar Government Arts College, Chennai*

[4]*Assistant Professor, Department of Statistics, SDNB Vaishnav College for Women, Chennai*

Corresponding Author:** ***G. Manimannan, *Assistant Professor, Department of Mathematics, TMG Colleae of Arts and Science. Chennai*

Abstract: *On recent times, the collection of biological data has shown rapid increase with the availability of improvised technologies. These massive amounts of biological data are organized and maintained to assist conducting of experiments and research programs at large scale. This study deals with the extraction of genome sequences, identification of species and their similarities by applying data mining techniques for genomic data and using clustering algorithms to group the related species with similar DNA sequences. Each species have their own specific genome sequences and considering these sequences, we represent the DNA string in numerical characterization. Thus, the frequency counts for each three lettered words of nucleotides are framed. Thus it gives 43=64 keywords of three lettered strings which are recognized as tri-nucleotides. Then these species are represented geometrically for the identification of the species using polar plot. Finally, the method of artificial neural network is used to reduce high dimensionality of the data. The use of Self-Organizing Feature Map (SOFM) demonstrates the clustering of similar species in close 2D neighborhood and dispersion among the clusters of dissimilar species.*

Keywords: *Genomic data, DNA sequences, Data mining techniques, artificial neural network, self-organizing feature map.*

1. INTRODUCTION

Biological data mining is an emerging field of research and development. Rapid advances in automated DNA sequencing technology have generated the need for statistical summarization of large volumes of sequence data so that efficient and effective statistical analysis can be carried out. There is a general assumption made for all evolutionary biology that any groups of organisms are related by descent from a common ancestor. It literally means that all organisms are interrelated in one or another way. As time passes the evolutionary changes takes place and this results in characteristic changes in organisms over time. These changes are used to classify the species as there must be a classification of the organisms into groups to carry out many biological researchers. Ideally, the classification should be based on the evolutionary history of life, such that it predicts properties of newly discovered or poorly known organisms. All methods for the identification of species that rely on DNA or protein sequence analysis presuppose the neutral theory of molecular evolution, in which different lineages diverge over evolutionary times by the accumulation of molecular changes (most of them neutral.

In this study, we aim at identifying and classifying the similar species considering their genome sequences. Since the analysis are carried for large DNA sequences, we can make it easier by statistically summarizing the data using clustering techniques, while capturing some of the fundamental structural information contained in the sequence data to help classify different species on the basis of their genomic data alone.

2. BIOINFORMATICS

Deoxyribonucleic acid (DNA) is the basis of genetic material. Essentially, DNA is the coded information contained within the threadlike, tubular structures known as chromosomes confined in the nucleus of a cell and a small quantity is found in the mitochondria and chloroplasts. DNA exists to instruct the cell as to how and when to synthesize specific products, such as proteins, which perform and regulate most basic functions. A DNA molecule consists of two strands of sugar phosphate, tightly wound around each other to form a right-handed double helix structure.

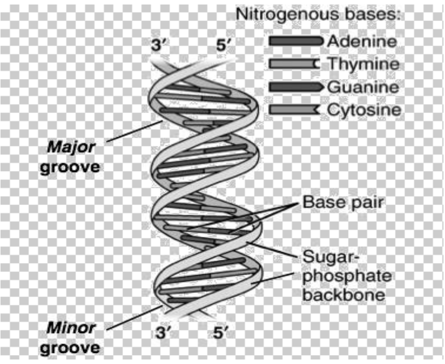


Fig1. *Structure of DNA with nitrogenous bases [Watson and Crick model,1953]*

On each sugar phosphate strand resides a linear arrangement of nitrogenous amine bases (adenine, thymine, guanine or cytosine), with bases on one strand paired with bases on the other strand, resembling rungs on a ladder. The bases are joined through hydrogen bonds in a specific manner, in accordance to the complementary base pairing rules of Watson and Crick. Thus, guanine will always be paired with cytosine; similarly, adenine is paired with thymine. A complementary base is the term used to describe the amine base that forms a bonding pair with another amine base (i.e. A with T and G with C). The particular order of the bases along any one strand of the DNA structure is referred to as its DNA sequence.

Briefly, the sequence is broken into triplet subsets of bases (called codons), which each acts as a code for the production of a particular amino acid. A chain of amino acids from consecutive cordon constitutes an individual protein molecule.

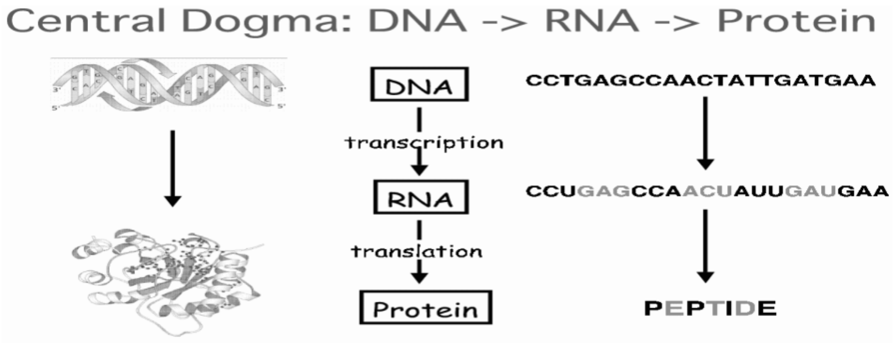


Fig2. *Central dogma of molecular biology*

DNA in the human genome comprises of 23 pairs of chromosomes containing about 3.5 billion pairs of nucleotides, within which exists the coding for many different genes, which themselves can range quite widely in length from a few hundred base pairs to several thousand. Each gene occupies a specific location (or locus) on a particular chromosome, which allows it to be specifically identified. This provides scope for mapping genes and allows the impact of chromosomal location to be considered.

Bioinformatics is the science of integrating, managing, mining, and interpreting information from biological data sets. Biological databases continue to grow rapidly. A huge volume of data is available for the extraction of high level information including the development of new concepts, concept interrelationships and interesting patterns hidden in the databases.

Data analysis is seen as the largest and possibly the most important area of microarray bioinformatics [11]. More complex data may require analysis via ANOVA or general linear models. Principal

Component Analysis (PCA) and Multidimensional Scaling (MDS) provide a good way to visualize data without imposing any hierarchy on them. Hierarchical clustering can be used to identify related genes or samples and portray the usage of dendrogram. There are several methods for classifying samples, each with advantages and disadvantages, including: K-nearest neighbour, centroid classification, linear discriminant analysis, and neural network, support vector machines (Stekel, 2003).

3. DATA MINING

Data mining is the extraction of hidden predictive information from large databases. Data mining techniques and applications have been recently developed and used for genomic data analysis. Mining bioinformatics data is an emerging area of intersection between bioinformatics and data mining. There has been a great explosion of genomic data in recent years. Data mining is the application of specific tools for pattern discovery and extraction. Knowledge discovery is concerned with the theoretical and practical issues of extracting high level information (knowledge) from volumes of low level data. It combines techniques from databases, statistics and artificial intelligence. It comprises of several data pre-processing steps as well as data mining and knowledge interpretation steps.

4. REVIEW OF LITERATURE

This secttion deals with few literature related to species identification and application of Self Organizing Feature Map (SOM). It covers a general reviews on factors associated with genomic cluster and data mining application. It further includes various studies that are related directly and indirectly to the topic this study.

In 2007, the work by Shreyas Sen, Seetharam Narasimhan, and Amit Konar- “Biological Data Mining for Genomic Clustering Using Unsupervised Neural Learning”, (Engineering Letters, 14:2, EL_14_2_8.), aims at designing a scheme for automatic identification of a species from its genome sequence. A set of 64 three-tuple keywords is first generated using the four types of bases: A, T, C and G. These keywords are searched on N randomly sampled genome sequences, each of a given length (10,000 elements) and the frequency count for each of the 4 = 64 keywords is performed to obtain a DNA-descriptor for each sample. This paper presents a novel method for identifying a species from its genome sequence with the help of a two dimensional map of neuronal clusters, where each cluster represents a particular species. PCA is a proven technique for reducing data dimensionality without losing data accuracy. The use of PCA demonstrated better classification of different species without sacrificing similarity in biologically close species. The work entails huge amount of genomic data, and so the learning ability of the neural network is used to reduce high dimensionality of the data. The self organizing feature map is shown to provide an easier technique for recognition and classification of a species based on its genomic data.

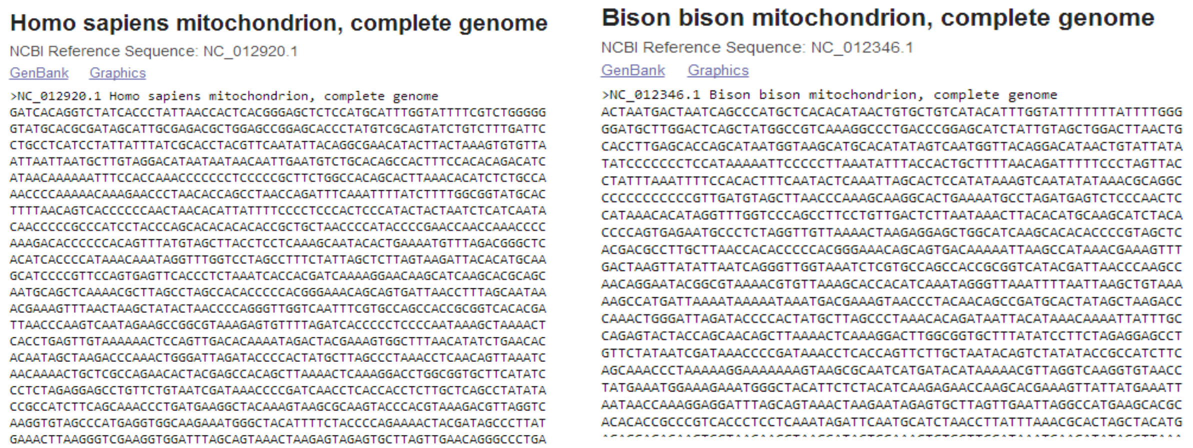
In 1999, the work of Juha Vesanto “SOM−Based Data Visualization Methods In Intelligent Data Analysis”, Volume 3, Number 2, Elsevier Science, pp. 111−126**,** gives an overview and categorization of both old and new methods for the visualization of SOM is presented. The study gives an idea of what kind of information can be acquired from different presentations and how the SOM can be best be utilized in exploratory data visualization. Most of the presented methods can also be applied in the more general case of first making a vector quantization (e.g. k-means) and then a vector projection**.** The major point of this paper is to bring together the many visualization methods for the self-organizing map and thus bring forth the power of the SOM, and VQ-P methods in general, in visualization of multidimensional numerical data. The SOM combines vector quantization and projection which together provide a map of the data giving a visual insight to the properties of the data: the shape of the data cloud, clusters, their properties and correlations between vector components. Comparing new data with the map helps in classifying the data and gives an indication of whether the new data belongs to the same data distribution as the map helps in classifying as the map was trained with. In the general framework of knowledge discovery, the methods presented in this paper are especially useful in the exploratory phase of knowledge discovery task.

In 2013, the work by Hongjie Yu and Deshuang Huang,-“Descriptors for DNA Sequences Based on Joint Diagonalization of Their Feature Matrices from Dinucleotide Physicochemical Properties”, (Tsinghua science and technology, ISSNll1007-0214ll03/11llpp446-453,Volume 18, Number 5), aims to visualize and compare different DNA sequences in less space by a novel descriptors extraction approach proposed for numerical characterizations and similarity analysis of sequences. Initially, a transformation method was introduced, then, based on the approximate joint diagonalization theory, an eigenvalue vector was extracted from each DNA sequence, which could be considered as descriptor of the DNA sequence. Moreover, similarity analyses were performed by calculating the pair-wise distances among the obtained eigenvalue vectors. The proposed approach (AJD-PCM) allows the fusion of the physicochemical properties with the sequential property of the biological sequence with consideration of the sequential property at the stage of mapping each dinucleotide into a 12-tuple vector. Moreover, at the second stage, AJD could extract the features among multiple sequences jointly rather than separately, facilitating the simultaneous discovery of sub-groups of organisms having a common structure at the molecular level. The clustering results were consistent with the evolutionary facts demonstrating the rationality of the proposed numerical characterization of DNA sequence.

The study by Huai-Chun Wang, Jonathan Badger, Paul Kearney, Ming Li -"Analysis of Codon Usage Patterns of Bacterial Genomes Using the Self-Organizing Map ", (Mol Biol Evol (2001) 18 (5): 792-800), in 2001, presents the Kohonen's self-organizing map applied to analysis of the codon usage pattern of the genomes of 7 organisms for evidence of highly expressed genes and horizontally transferred genes. All of the analyzed genomes had a clear category of horizontally transferred genes, and their apparent percentages ranged from 7.7% to 21.4%. The apparent percentage of highly expressed genes ranges from 0% to 11.8%. An SOM of the average codon usage for the major gene categories of 7 organisms are framed. It can be seen that two proteobacteria, E. coli and H. influenza, form a codon usage cluster, as do the thermophilic organisms (with exception of M.jannaschii). the gene categories from A. aecolicus, A. fulgidus, M. thermoautrophicum, and P. horikoshiiare intermingled on the map, suggesting that horizontal transfer events may have occurred between these organisms. Further classification of the three gene categories in E. coli and H. influenza according to gene function revealed that genes involved in communication (such as regulation and cell process) and structure (cell structure and structural proteins) are more likely to be horizontally transferred than are genes involved in information (transcription, translation, and related processes) and in some groups of energy (such as energy metabolism and carbon compound catabolism).

5. Data Base

Databases of nucleic acid and protein sequences maintain facilities for a very wide variety of information retrieval and analysis operations such as retrieval of sequences from the database, sequence comparison, translation of DNA sequences to protein sequences, simple types of structure analysis and prediction, pattern recognition and molecular graphics. In this study, the genome sequences of different species [Homo Sapiens (Humans), Bison Bison (American Bison), Pan Troglodytes (Chimpanzee), Loxodonta Africana (African Elephant), Chlamydomonas reinhardtii (unicellular green algae) , Plasmodium Falciparum (causal agents of human malaria), Pongo Abelii (Sumatran orangutan)] are retrieved from GenBank, in FASTA format. GenBank can be accessed through the website (www.ncbi.nlm.gov/GenBank). The data is the complete genome sequences of DNA in mitochondria. A portion of two species are shown:



Homo sapiens mitochondrion, complete genome

NCBI Reference Sequence: NC_012920.1

GenBank Graphics

```
>NC_012920.1 Homo sapiens mitochondrion, complete genome
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGG
GTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTC
CTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTA
ATTAATTAATGCTTGTAGGACATAATAATAACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATC
ATAACAAAAAATTTCCACCAAACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCA
AACCCCAAAAACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAATCTCATCAATA
CAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATACCCCGAACCAACCAAACCCC
AAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAAGCAATACACTGAAAATGTTTAGACGGGCTC
ACATCACCCCATAAACAAATAGGTTTGGTCCTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAA
GCATCCCCGTTCCAGTGAGTTCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCAGC
AATGCAGCTCAAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGCGGTCACACGA
TTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCACCCCCTCCCCAATAAAGCTAAAACT
CACCTGAGTTGTAAAAAACTCCAGTTGACACAAAATAGACTACGAAAGTGGCTTTAACATATCTGAACAC
ACAATAGCTAAGACCCAAACTGGGATTAGATACCCCACTATGCTTAGCCCTAAACCTCAACAGTTAAATC
AACAAAACTGCTCGCCAGAACACTACGAGCCACAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATC
CCTCTAGAGGAGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATA
CCGCCATCTTCAGCAAACCCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCACGTAAAGACGTTAGGTC
AAGGTGTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTTCTACCCCAGAAAACTACGATAGCCCTTAT
GAAACTTAAGGGTCGAAGGTGGATTTAGCAGTAAACTAAGAGTAGAGTGCTTAGTTGAACAGGGCCCTGA
```

Bison bison mitochondrion, complete genome

NCBI Reference Sequence: NC_012346.1

GenBank Graphics

```
>NC_012346.1 Bison bison mitochondrion, complete genome
ACTAATGACTAATCAGCCCATGCTCACACATAACTGTGCTGTCATACATTTGGTATTTTTTTATTTTGGG
GGATGCTTGGACTCAGCTATGGCCGTCAAAGGCCCTGACCCGGAGCATCTATTGTAGCTGGACTTAACTG
CACCTTGAGCACCAGCATAATGGTAAGCATGCACATATAGTCAATGGTTACAGGACATAACTGTATTATA
TATCCCCCCCTCCATAAAAATTCCCCCTTAAATATTTACCACTGCTTTTAACAGATTTTTCCCTAGTTAC
CTATTTAAATTTTCCACACTTTCAATACTCAAATTAGCACTCCATATAAAGTCAATATATAAACGCAGGC
CCCCCCCCCCGTTGATGTAGCTTAACCCAAAGCAAGGCACTGAAAATGCCTAGATGAGTCTCCCAACTC
CATAAACACATAGGTTTGGTCCCAGCCTTCCTGTTGACTCTTAATAAACTTACACATGCAAGCATCTACA
CCCCAGTGAGAATGCCCTCTAGGTTGTTAAAACTAAGAGGAGCTGGCATCAAGCACACACCCCGTAGCTC
ACGACGCCTTGCTTAACCACACCCCCACGGGAAACAGCAGTGACAAAAATTAAGCCATAAACGAAAGTTT
GACTAAGTTATATTAATCAGGGTTGGTAAATCTCGTGCCAGCCACCGCGGTCATACGATTAACCCAAGCC
AACAGGAATACGGCGTAAAACGTGTTAAAGCACCACATCAAATAGGGTTAAATTTTAATTAAGCTGTAAA
AAGCCATGATTAAAATAAAAATAAATGACGAAAGTAACCCTACAACAGCCGATGCACTATAGCTAAGACC
CAAACTGGGATTAGATACCCCACTATGCTTAGCCCTAAACACAGATAATTACATAAACAAAATTATTTGC
CAGAGTACTACCAGCAACAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTATATCCTTCTAGAGGAGCCT
GTTCTATAATCGATAAACCCCGATAAACCTCACCAGTTCTTGCTAATACAGTCTATATACCGCCATCTTC
AGCAAACCCTAAAAAGGAAAAAAAGTAAGCGCAATCATGATACATAAAAACGTTAGGTCAAGGTGTAACC
TATGAAATGGAAAGAAATGGGCTACATTCTCTACATCAAGAGAACCAAGCACGAAAGTTATTATGAAATT
AATAACCAAAGGAGGATTTAGCAGTAAACTAAGAATAGAGTGCTTAGTTGAATTAGGCCATGAAGCACGC
ACACACCGCCCGTCACCCTCCTCAAATAGATTCAATGCATCTAACCTTATTTAAACGCACTAGCTACATG
```

Fig3. *A section of the FASTA format of complete mitochondrion genome sequence of Homo Sapiens and Bison Bison (American Bison)*

6. Self Organizing Map

Self-organizing map is an unsupervised learning type Neural Network that can be used for clustering the input data and find features inherent to the problem. The Self-Organizing Map was developed by professor Kohonen. The SOM has been proven useful in many applications. SOM is useful for clustering data without knowing the class memberships of the input data. The SOM can be used to detect features inherent to the problem and thus has also been called SOFM, the Self-Organizing Feature Map.

Map units, or neurons, usually form a two-dimensional lattice and thus the mapping is a mapping from high dimensional space onto a plane. The property of topology preserving means that the mapping preserves the relative distance between the points. Points that are near each other in the input space are mapped to nearby map units in the SOM. The SOM can thus serve as a cluster analyzing tool of high-dimensional data. Also, the SOM has the capability to generalize.

6.1. Network Architecture

The architecture of SOMs is shown in figure 5.1. SOMs consist of two types of units; the input units and the output units. The input to the network is features of input patterns and the number of input units is equal to the number of features. The competing neurons are the cluster units. The number of cluster units (i.e. neurons) varies depending on the problem domain and the choice of the users. SOMs assume a topological neighbourhood structure among the cluster units.

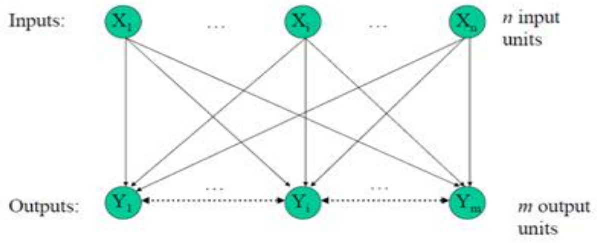


Fig4. *NETWORK ARCHITECTURE*

Each competing neuron has its specific position (an x, y coordinate) in a two dimensional space. Each neuron has a vector of weights of the same dimension as the input vectors. If the input data consists of vectors X of N dimensions $(x_1, x_2, \ldots, x_N)$ then each node contains a corresponding weight vector W W of N dimensions $(w_1, w_2, \ldots, w_N)$.

The weight vector for a cluster unit serves as an exemplar of the input patterns associated with that cluster. During the self-organization process, the cluster unit whose weight vector matches the input pattern most closely (according to the distance function used) is chosen as the winner. The weights of the winner and its neighbours are updated according to the topology and radius (Laurene, 1993).

6.2. Algorithm

The basic SOM algorithm is iterative. Each neuron i has a d-dimensional prototype vector $[,\ldots,]w_i =, w_{i1}, w_{id}$ or weight of i^{th} neuron. At each training step, a sample data vector x is randomly chosen from the training set. Distances between x and all the prototype vectors are computed. The best-matching unit (BMU) or the winner unit, denoted here by x_i^*, is the map unit with prototype closest to x (Kaski (1997))

The algorithm is shown below (Laurene, 1993)

Step 0: Initialize weights.

-Set topological neighborhood parameters.

-Set learning rate (α) parameters.

Step1: While stopping condition is false, do Steps 2-8

Step 2: For each input vector x, do step 3-5.

Step 3: For each j, compute :(i.e. compute all distances between input data and each neuron by using the selected distance function) $D(j) = \sum (w_{ij} - x_i)^2$

Step 4: Find index j such that $D(j)$ is a minimum. (i.e. find the winner neuron)

Step 5: For all units j within a specified neighborhood of j, and for all i;
$w_{ij}(new) = w_{ij}(old) + \alpha \lfloor x_i - w_{ij}(old) \rfloor$.

Step 6: Update learning rate.

Step 7: Reduce radius of topological neighborhood at specified times.

Step 8: Test stopping condition.

The learning rate α is a slowly decreasing function of time or training epochs. It was indicated that a linearly decreasing function is satisfactory for practical computations; for example a geometric decrease would produce similar results (Laurene, 1993). The radius of the neighborhood for a winner neuron also decreases as the clustering process progresses (Laurene, 1993).

6.3. Data Preprocessing

Data preprocessing in general can be just about anything: simple transformations or normalizations performed on single variables, filters, calculation of new variables from existing ones. The data inputted in excel are the standardized values with mean zero and a unit variance of the frequency counts of tri-nucleotides (variables) which are previously standardized in R programming using the scale() function.

6.4. Initialization and Training

Initially, the genome sequence of a species is considered in the study and frequency counts are performed on it, for each species. Then, the weights of the neurons are calculated using the (1 × 64) vector as an input, the distance between this input vector and the weights of the neurons are calculated. The winner for this input is found to be a neuron from the map. Whenever a new sequence is obtained, its DNA-descriptor is computed and the distance between the new input and existing neurons is calculated. The winning neuron will declare to which species it belongs or if it is of a new species.

The batch training algorithm is generally much faster than the incremental algorithm, and it is the default algorithm for SOFM training. Training of the data automatically stops when the full numbers of epochs have occurred. Training multiple times will generate different results due to different initial conditions.

7. Result and Discussion

The self organizing feature map is carried out in Matlab version 12 using nctool functions (Neural Network Clustering tool) to select data, train and evaluate using the visual representation. The neurons are arranged in a 2D topology and the network is trained using batch algorithm, by default. The standardized values for the variables (frequency count of tri-nucleotides) are selected as an input data and the Matlab code is generated. Thus, we have inputted the (7 x 64) vector, representing 7 samples. The numbers of neurons are, by default, selected as 10, which proceeds to training of the data. This will produce the following visual representation. In training window, different aspects of SOM can be visualized. The SOM Weight Positions plots the input data and SOM's nodes with lines between the neighboring nodes. The figure indicates, after only 200 iterations of the batch algorithm, the map is well distributed through the input space.

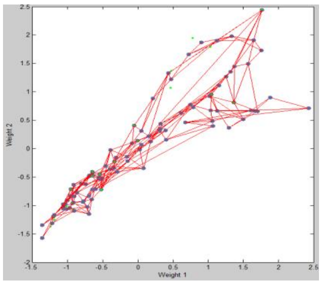


Fig6.2. *Weight Positions*

When the input space is high dimensional, all the weights can be visualized at the same time. SOM Neighbor Distances visualizes the distances between neighboring neurons using a color coding.

The colors coding are:

1. The blue hexagons represent the neurons.
2. The red lines connect neighboring neurons.
3. The colors in the regions containing the red lines indicate the distances between neurons.
4. The darker colors represent larger distances.
5. The lighter colors represent smaller distances.

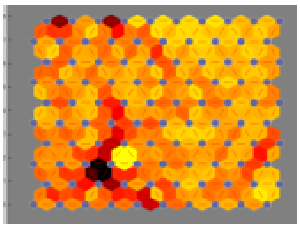


Fig6.3. *SOM Neighbor Weight Distance*

A group of light segments appear in the upper-right region, bounded by some darker segments. This grouping indicates that the network has clustered the data into two groups. These two groups can be seen in the previous weight position figure. The lower-left region of that figure contains a small group of tightly clustered data points. The corresponding weights are closer together in this region, which is indicated by the lighter colors in the neighbor distance figure. Where weights in this small region connect to the larger region, the distances are larger, as indicated by the darker band in the neighbor distance figure. A small group in the lower-left region of the neighbor distance figure is darker than those in the upper right. This color difference indicates that data points in this region are farther apart. This distance is confirmed in the weight positions figure. Another useful figure to show how many data points are associated with each neuron is SOM Sample Hits. It is best if the data are fairly evenly distributed across the neurons.

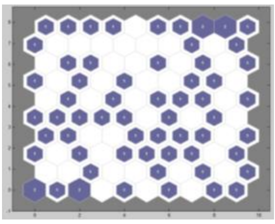


Fig6.4. *SOM sample Hits*

The weights can also be visualized using the weight plane figure by SOM Weight Planes in the training window. There is a weight plane for each element of the input vector (two, in this case). They are visualizations of the weights that connect each input to each of the neurons. (Lighter and darker colors represent larger and smaller weights, respectively.) If the connection patterns of two inputs are very similar, you can assume that the inputs were highly correlated.

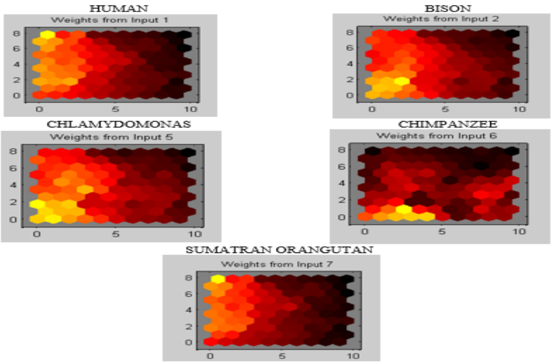


Fig6.5. *SOM WEIGHT PLANES for each samples*

Finally the M-file i.e. the script file, is generated. It gives a text file with MATLAB commands. When the file is run, MATLAB reads the commands and executes them exactly as it would if you had typed each command sequentially at the MATLAB prompt. All m-filenames must end with the extension '.m'.

8. Conclusion

Development of novel data mining methods will play a fundamental role in understanding these rapidly expanding sources of biological data. A type of neural network, SOMs provides an elegant solution to many arduous problems with large or difficult to interpret the biological data. Through their intrinsic properties, such as preserving topological relationships between input data, they allow the visualization of complex data. Comparative genomics is the study of the relationship of genome structure and function across different biological species. The SOM method is useful in identifying the species and the similar species thus enabling the ease of comparative genomics.

From the study, we conclude that the genome sequence can be expressed in numeric by three-lettered DNA bases which are the tri-nucleotides. In some cases, the studies based these tri-nucleotides are useful for research works on finding tri-nucleotide repeated disorders. Then, the geometric representation based on the tri-nucleotides using polar plot gives the unique structures of the organisms. The organisms with similar structure and those belong to same family and classes in the taxonomy can also be visually identified.

Self Organizing Feature Map (SOM) applied to the data gives a visual representation of the organisms. The genomic data of size 7 x 64 neurons with different training parameters are performed and a final map of 10 x 10 is obtained. For the inputs which have a similar connection pattern indicates they were highly correlated. From the SOM weight planes, it can be concluded that Human and Chimpanzee share a similar pattern indicating that they are highly related. Orangutan also has fairly a similar pattern to Human and Chimpanzee. Whereas it shows that the patterns of Plamodium Falciparum and Chlamydomonas have entirely different structure indicating that there is no similarity among them and the other organisms.

References

[1] Edder, John F and Abbott, Dean-“A Comparison of Leading Data Mining Tools.” Fourth International Conference on Knowledge discovery and Data Mining, Friday, Aug 28, 1998, NY, pp 19-25.

[2] Zhang , Dongsang and Zhou, Lina. “Data Mining Techniques in Financial Application”. IEEE Transactions on Systems, Man and Cybernetics – Part C : Applications and Reviews, Vol – 34, No- 4, Nov-2004, pp. 513 – 522.

[3] George Tzanis, Christos Berberidis, and Ioannis Vlahavas, “Biological Data Mining”, Aristotle University of Thessaloniki, Greece.

[4] Shreyas Sen, Seetharam Narasimhan, and Amit Konar- “Biological Data Mining for Genomic Clustering Using Unsupervised Neural Learning”, Engineering Letters, 14:2, EL_14_2_8 (Advance online publication: 16 May 2007)

[5] Jia-Feng Yua, b, , Xiao Suna, Ji-Hua Wangb “TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications”, Journal of Theoretical Biology, Volume 261, Issue 3, 7 December 2009.

[6] Yakovchuk P, Protozanova E, Frank-Kamenetskii MD (2006). "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix". Nucleic Acids Res. 34 (2):564–74.doi: 10.1093/nar/gkj454. PMC 1360284 . PMID 16449200.

[7] Definition of GC-ratio on Northwestern University, IL, USA.

[8] Furey TS, Haussler D (May 2003). "Integration of the cytogenetic map with the draft human genome sequence". Hum. Mol. Genet. 12 (9): 1037–44. doi:10.1093/hmg/ddg113. PMID 12700172

[9] Birdsell JA (1 July 2002). "Integrating genomics, bioinformatics, and classical genetics to study the effectsof recombination on genome evolution". Mol. Biol. Evol. 19 (7): 1181 97.doi:10.1093/oxford journals. molbev.a004176. PMID 12082137

[10] Redford GI1, Clegg RM., “Polar plot representation for frequency-domain analysis of fluorescence lifetimes”. J Fluoresc. 2005 Sep;15(5):805-15.

[11] Irina Arhipova, “The role of statistical methods in computer science and bioinformatics”, Latvia University of Agriculture, Latvia, ICOTS-7, 2006: Arhipova (Refereed).

[12] Juha Vesanto, Johan Himberg, Esa Alhoniemi and Juha Parhankangas-"Self-organizing map in Matlab: the SOM Toolbox", Laboratory of Computer and Information Science, Helsinki University of Technology, Finland

[13] Junbai Wang, Jan Delabie, Hans Christian Aasheim, Erlend Smeland and Ola Myklebost,-" Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study", http://www.biomedcentral.com/1471-2105/3/36.

[14] Alper Özdogan,-"Clustering Genes by Using Different Types of Genomic Data and Self-Organizing Maps", University of Skövde, SWEDEN,June, 2008.

[15] Petri To«ro«nena, Mikko Kolehmainenb, Garry Wonga, Eero Castre¨na;c;-" Analysis of gene expression data using self-organizing maps", FEBS Letters 451 (1999) 142^146.

[16] Shaun Mahony, James O McInerney, Terry J Smith and Aaron Golden,-" Gene prediction using the Self-Organizing Map: automatic generation of multiple gene models", BMC Bioinformatics 2004, http://www.biomedcentral.com/1471-2105/5/23.

[17] Marghny Mohamed, Abeer A. Al-Mehdhar, Mohamed Bamatraf, Moheb R. Girgis,-"Enhanced Self-Organizing Map Neural Network for DNA Sequence Classification", Intelligent Information Management, 2013, 5, 25-33.

Citation: *Manimannan. G ,et.al., (2019). Trinucleotides Based Species Identification by Genomic Taxonomy Using Self Organizing Feature Map. International Journal of Scientific and Innovative Mathematical Research (IJSIMR), 7(3), pp.25-33. http://dx.doi.org/10.20431/2347-3142.0703003*