Biological Data Mining for Genomic Clustering Using Unsupervised Neural Learning

Shreyas Sen, Seetharam Narasimhan, and Amit Konar

Abstract— **The paper 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 $4^3 = 64$ keywords is performed to obtain a DNA-descriptor for each sample. Principal Component analysis is then employed on the DNA-descriptors for N sampled instances. The principal component analysis yields a unique feature descriptor for identifying the species from its genome sequence. The variance of the descriptors for a given genome sequence being negligible, the proposed scheme finds extensive applications in automatic species identification. An alternative approach to automatic species classification and identification of species using Self-Organizing Feature Map is also discussed in the paper. The computational map is trained by using the DNA-descriptors from different species as the training inputs. The maps for different dimensions are constructed and analyzed for optimum performance. The scheme 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. The map is shown to provide an easier technique for recognition and classification of a species based on its genomic data.**

Index Terms—**DNA-descriptors, Feature Descriptors, Principal Component Analysis (PCA), Self-Organizing Feature Map (SOFM)**.

I. INTRODUCTION

Genomic data mining and knowledge extraction is an important problem in bioinformatics. Identification of a species from its genomic database is a challenging task. The paper explores a new approach to extract genomic features of a species from its genome sequence. Biological data mining is an emerging field of research and development for further progress in this direction [1]. Significant progress on DNA-string matching has been reported in the current literature on Bio-informatics. Among the well-known techniques of DNA-string matching are the Smith-Waterman algorithm [2], [3] for local alignment, the Needleman-Wunsch algorithm [4] for global alignment, Hidden Markov's model, matrix model, evolutionary algorithms for multiple sequence alignment [5] etc. These works, though extremely valuable, have their limitations. The demerits include the use of complicated matrix algebra and dynamic programming, and the results of sequence matching are not free from pre-calculated threshold values. It is to be noted that none of the above-mentioned methods can be directly employed to identify the species from the structural signature of the genomes.

Rapid advances in automated DNA sequencing technology [6] have generated the need for statistical summarization of large volumes of sequence data so that efficient and effective statistical analysis can be carried out. The popular sequence alignment algorithms and techniques for estimating homologies [7] and mismatches among DNA sequences that are used for comparing sequences of relatively small sizes are not applicable to sequences with sizes varying between a few thousand base pairs to a few hundred thousand base pairs. Even for comparison of small sequences, the standard alignment and matching algorithms are known to be time consuming. There is a dearth of rapid and parsimonious procedures that may be somewhat approximate in nature yet useful in producing quick and significant results.

The present paper is an attempt to fill this void. The idea is to make the analysis of large DNA sequences easier by statistically summarizing the data using dimensional reduction and 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.

Since the work entails processing huge amounts of incomplete or ambiguous data, the learning ability of artificial neural networks (ANNs) is utilized in this direction. The learning capabilities of ANNs, typically in data-rich environments, come in handy when discovering regularities from large datasets. This can be unsupervised as in clustering, or supervised as in classification. The connection weights and topology of a trained ANN are often analyzed to generate a mine of meaningful (comprehensible) information about the learned problem in the form of rules. There exist different ANN-based learning and rule mining strategies, with applications to the biological domain [8].

Manuscript received April 30, 2006.

Shreyas Sen is with the Electronics & Telecommunication Engineering Department, Jadavpur University, Calcutta, India – 700032. Phone: 913324147471, (e-mail: shreyas.sen@gmail.com).

Seetharam Narasimhan is with the Electronics & Telecommunication Engineering Department, Jadavpur University, Calcutta, India – 700032. Phone: 913324353079, (e-mail: mail.seetharam@gmail.com).

Amit Konar is with the Electronics & Telecommunication Engineering Department, Jadavpur University, Calcutta, India – 700032 (e-mail: babu25@hotmail.com).

Feature extraction refers to a process whereby a data space is transformed into a feature space that has exactly the same dimension as the original data space. However the transformation is designed in such a way that the data set may be represented by a dimensionally reduced number of effective features and yet retains most of the intrinsic information content of the data; in other words the data set undergoes a dimensionality reduction [9]. The transformation must have a low variance for at least some of its components. The right choice is Principal Components Analysis (PCA) since it maximizes the rate of decrease of variance. Since the main issue is to achieve good data compression, while preserving as much information about the inputs as possible, the use of principal components analysis offers a useful self-learning procedure.

A related issue is the representation of a data set made up of an aggregate of several clusters. Cluster validation is essential, from both the biological and statistical perspectives, in order to biologically validate and objectively compare the results generated by different clustering algorithms. In this context we take the assistance of a very well-known ANN model, the self-organizing feature map (SOFM) for clustering of the extracted features from genomic data.

The self organizing feature maps are a special class of artificial neural networks, based on competitive learning. The neurons become selectively tuned to various input patterns or classes of input patterns in the course of a competitive learning process. A self-organizing map is characterized by the formation of a topographic map of the input patterns in which the spatial locations of the neurons in the lattice are indicative of intrinsic statistical features contained in the input patterns [10]. The SOFM is an established technique for classification of events. Kohonen's SOFM has been used for the analysis of protein sequences [11], involving identification of protein families, aligned sequences and segments of similar secondary structure, in a highly visual manner. Other applications of SOFM include prediction of cleavage sites in proteins [12], prediction of beta-turns [13], classification of structural motifs [14] and feature extraction [15].

To the best of the authors' knowledge, identifying a species from its genomic data is an open problem. The novelty of the work reported in this paper is as follows.

First, the paper takes into account frequency counts of 64 three-lettered primitive DNA attributes in randomly selected samples of the genome sequences of different species (e.g., the bacterium Escherichia coli (E. coli) [16], Drosophila melanogaster [17], Saccharomyces cerevisiae (yeast) [18], Mus musculus (mouse), and Homo sapiens (human beings)).

Second, to reduce the data dimension of extracted features (here, frequency count), principal component analysis (PCA) is employed on the randomly selected samples of genome sequence. The variance of the extracted feature vectors being extremely small for any randomly selected input sequence, the accuracy of the results in identifying the species is very high.

Third, clustering techniques are adopted on the frequency count data of three species: E. coli, Yeast and Mouse. The SOFM algorithm is adopted for this purpose. Maps of different dimensions are constructed and analyzed on the basis of their efficiency in clustering the extracted features from genomic data of different species.

II. DNA

A. Structure

The nucleus of a cell contains chromosomes that are made up of the double helical DNA molecules. The DNA consists of two strands, consisting of a string of four nitrogenous bases, viz., adenine (A), cytosine (C), guanine (G), and thymine (T). DNA in the human genome is arranged into 24 distinct chromosomes. Each chromosome contains many genes, the basic physical and functional units of heredity. However, genes comprise only about 2% of the human genome; the remainder consists of non-coding regions, whose functions may include providing chromosomal structural integrity and regulating where, when, and in what quantity proteins are made. The DNA is transcribed to produce messenger (m)-RNA, which is then translated to produce protein. The m-RNA is single-stranded and has a ribose sugar molecule. Here exist 'Promoter' and 'Termination' sites in a gene, responsible for the initiation and termination of transcription. Translation consists of mapping from triplets (codons) of four bases to the 20 amino acids, which are the building block of proteins.

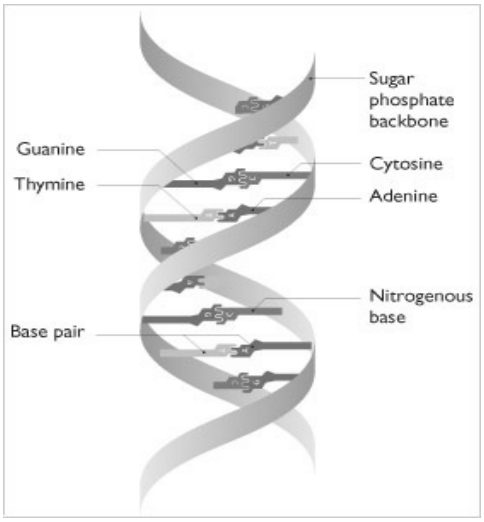


Fig. 1: Double-helical Structure of DNA showing the nucleotide bases.

B. DNA Sequences

Portions of the DNA sequences of the two species Drosophila and E. coli, as available on the web, are shown below. It is quite clear from them that the species, whose DNA-sequences they represent cannot be distinguished on the basis of these sequences alone.

```
>gi|1786181|gb|AE000111|ECAE000111 Escherichia coli , thrL, thrA, thrB,
 thrC, yaaA, yaaJ, talB, mog, yaaH genes from bases 1 to 10596
(section 1 of 400) of the complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
CCCGCACCTGACAGTGCGGGCTTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAA
GTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGAAAGCAATGCC
AGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTG
AAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTT
GACGGGACTCGCCGCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTT

>gi|7289301|gb|AE002567.1|AE002567 Drosophila melanogaster genomic scaffold 14200(
AAATAAAAATGGCCAAAAATCGGATATTTACAAACGGGGTTTTCTCTATAACCTGACCAAAAAAGATATAACCGTTTAAA
TGAAGACTGTTTTGTACTGCCAATAAACATAGCGAACTATCCGTATCACTATTTTTATTATCATCAAAAATAAAAATTTT
AACACATTTTCACATTTTTCGATAAGGGTTACCATCATAAAAATTTGCGAAAAATGGCAAAAAAATTGAATTTCCATTTC
TGAATACAGGTCGATGGGGAATTTTGTTACGAATTCAACGAGGTATGCCATTTCGCTCTTGGACAGTTATTTTTAATGCT
ATTGAGAAAGTACAGATTCTTTCTAGGCAAAAATAACTCTCCCTATATGTGTTTTTATGATTGCCAAAAGTAAATAAGTA
AACAGTAAAAGAAAGTAATTAAACATATTCGCATTTTCGTTAAGGAGTACTTTAATTTAAATTTGGATAGCGGGCATAAA
AAGAATAAGCAGACGACAGGACTTATCGGTGTGCTGGGCCGCGTATAAGCCCCGTTTTCTCATTAAAGCTCCGCCACGAC
TGAAAATAAATAAGCCATATTGGCTGGCGCTTATGGCCGGTGGCGAAGCCAGTTTTCAGGATAGTTATGTTATATGTTGA
GTGAACTATTACTACAAACAATCACTGTAAAAAAAATACTCACCAATTATGCCAAAAATGCCCCTAAATACGAATGACCAA
TGCACACAGAGCACTTTTCTTTCTCTTCTATCACTTTCTCATCACTAAAAGAAAAGGAACAAAAAACACAAAAAGCAACG
AAATGCTTCTTCTTCCTTTTCACTGACGCAAGGGCATGCAAAACCAACCGGCCTTGGTCACGTTTTAGCGCACGGTTTAT
TTCACTATCTGTAAACAGATCACCTAAAAAAAAACACACTTCTCTCTCTCTATTTCATTTATTACGCACTTTCTTACCAC
```

Fig. 2: DNA sequences of Drosophila and E. coli

III. Selection of DNA Descriptors

There are only 4 letters in a DNA-string; naturally the substrings could be one lettered, two lettered, three lettered or four lettered. So the number of possible combinations in each case is 4, 42, 43, or 44. Consequently the total number of substrings would be 4+42+43+44, which indeed is very large. To keep the search time optimum and moderately large search keys, we considered 3-lettered search keys only. Thus, we have 43 =64 search keys. Typical three-lettered search keys are AAA, AAC, AAG, AAT, ACA..... TTT. These 64 search keys thus generate a (1 × 64) frequency count vector, whose each component denotes population of one of the 64 sub-strings or keys in a sample of the genomic data of a species.

To illustrate what we mean by frequency count, let us take the help of some examples. Consider a small portion of the sequence like …AATCG…. It contributes a count of 1 each to the frequencies of occurrence of each of the 3 keywords AAT, ATC and TCG. Similarly for the substring …TTTTT…, we get a count of 3 for the frequency of the keyword TTT. Proceeding similarly for a large sample sequence of 10,000 bases we get frequencies of all the 64 keywords in the form of a frequency count vector of dimension (1 × 64). This (1 × 64) vector is called a DNA-descriptor of a given species.

Experiments undertaken on DNA-string matching reveals that some typical sub-strings have a high population in the DNA-sequence of a given species. Naturally, this result can be used as a basic test criterion to determine a species from its DNA-sequence. Now, we have plotted the DNA-descriptor obtained from a sample of the species Mouse. The plot is in the bar-diagram format. Corresponding to each of the 64 sub-strings, the value from the DNA-descriptor vector is plotted. It is to be noted that the values are normalized.

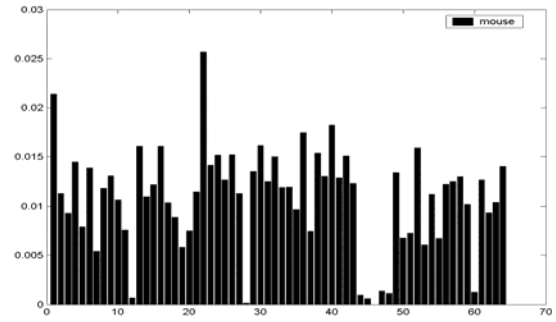


Fig. 3: DNA-descriptor for a sample of the DNA sequence of Mouse in the bar-diagram form

It is important to note that the frequency counts of 64 three-element keywords in a 10,000 element string of genome sequence are more or less invariant with respect to the random sampling of the genome sequence. Naturally, our main emphasis of study was to determine whether the small difference in the counts of a given keyword in N samples is statistically significant. PCA provides a solution to this problem. First, the dimension of (N × 64) is reduced by PCA to (1 × 64). Second, the (minor) disparity in the feature gets eliminated by PCA. Since PCA is a well-known tool for data reduction without loss of accuracy, we claim that our results on feature extraction from the genome database are also free from loss of accuracy.

IV. Application of PCA in the Present Context

The methodology of employing PCA [19] to the given problem is outlined below:

INPUT: A set of N DNA-descriptor vectors (1 × 64) representing the frequency counts of 64 three-tuple keywords.

OUTPUT: A minimal feature descriptor vector sufficient to describe the problem without any significant loss in data.

1. *Normalization*: Let the i^{th} (1 × 64) input vector be denoted by

$$\boldsymbol{a_i} = [a_{i1} \;\; a_{i2} \;\; \ldots\ldots \;\; a_{i64}] \tag{1}$$

To get the vector normalized we use the following transformation:

$$a_{ik} \leftarrow \frac{a_{ik}}{\sum_{j=1}^{64} a_{ij}} \tag{2}$$

2. *Mean adjusted data*: To get the data adjusted around zero mean, we use the formula:

$$a_{ik} \leftarrow a_{ik} - \overline{a_i} \qquad \forall i,k \tag{3}$$

where $\overline{a_i}$ = mean of the i^{th} vector

$$= \frac{1}{64} \sum_{j=1}^{64} a_{ij}$$

The matrix (N × 64) so

$$= \begin{pmatrix} a_{11} & \cdots & a_{164} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{N64} \end{pmatrix}$$

obtained is called the Data Adjust:

$$\text{Data Adjust} \tag{4}$$

3. *Evaluation of the covariance matrix:* The covariance between any two vectors a_i and a_j is obtained by the following formula:

$$\text{cov}(a_i, a_j) = c_{ij} = \frac{\sum_{k=1}^{64}\left(a_{ik} - \overline{a}_i\right)\left(a_{jk} - \overline{a}_j\right)}{(n-1)} \tag{5}$$

Covariance matrix C for the N different (1 × 64) vectors is represented as follows:

$$C = \begin{pmatrix} c_{11} & \cdots & c_{1N} \\ \vdots & \ddots & \vdots \\ c_{N1} & \cdots & c_{NN} \end{pmatrix} \tag{6}$$

where C is an N × N matrix.

4. *Eigenvalue Evaluation*: From the roots of the equation | C - λI | = 0, the eigenvalues of the covariance matrix C are obtained. There would be N eigenvalues of matrix C, and corresponding to each eigenvalue there would be eigenvectors each of dimension N × 1.

5. *Principal Component Evaluation*: The eigenvalues are not the same. In fact, it turns out that the eigenvector λ_{large} corresponding to the highest eigenvalue is the Principal Component (N × 1) of the data set. Therefore

$$\textit{Principal Component} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{bmatrix} \tag{7}$$

where $\lambda_{large} > \lambda_i$ for $1 \le i \le N$

6. *Projection of Data Adjust along the Principal Component:* Now, to get the feature descriptor, the following formula is applied:

Feature Descriptor = Principal Component^T × Data Adjust

where Principal Component^T (1 × N) is the transpose of the Principal Component vector. Thus we get a Feature Descriptor vector of dimension 1 × 64 corresponding to N samples of the genome sequence database of the particular species.

7. *Computing the Mean Feature Descriptor*: We calculate M such feature descriptors from different random samples and then calculate the mean of these vectors and also the variance vector (both 1 × 64).

V. Geometric Representation of Feature Descriptor

The Feature Descriptor Diagrams for different species are described here. We could represent the feature descriptors using bar diagrams, pie-charts or any other standard representation. However, using the polar plot we get figures that are compact yet distinct representations of the mean feature descriptor.

As mentioned earlier, the mean feature descriptor is a 1 × 64 vector. So to construct these diagrams 360° is divided into 64 equal parts, corresponding to 64 keywords. Plotting it in polar (r,ɵ) co-ordinates with r as the values of the mean feature descriptor vector and ɵ as these angles we get the feature descriptor diagrams. The feature descriptor diagrams are distinctly different from species to species. So we can readily detect new species and identify known species by comparing their feature descriptor diagrams.

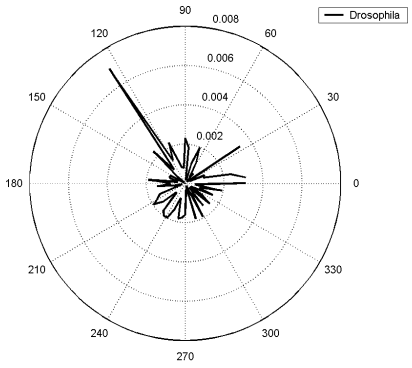


Fig. 4.1: Feature Descriptor Diagram for Drosophila

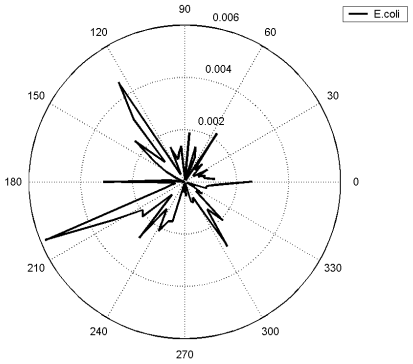


Fig. 4.2: Feature Descriptor Diagram for E. coli

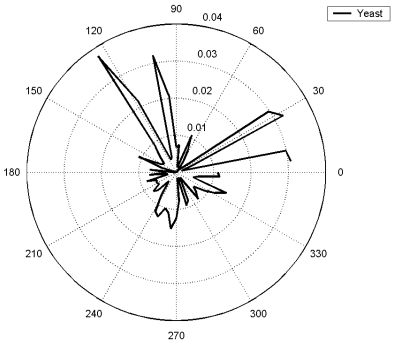


Fig. 4.3: Feature Descriptor Diagram for Yeast

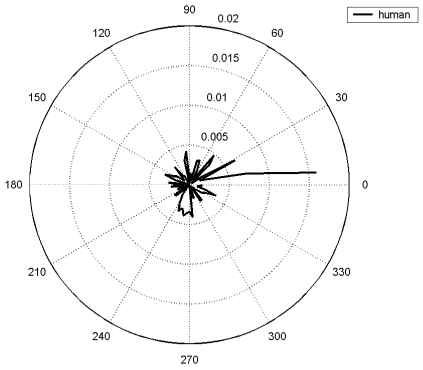


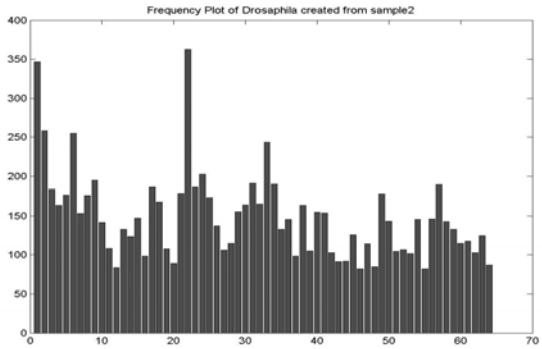
Fig. 4.4: Feature Descriptor Diagram for Human chromosome rp11-433k2

If we closely observe Fig 4.1 which contains the diagrammatic representation of the mean feature descriptor vector for Drosophila we can see some distinct peaks with a prominent one at around 125˚. In contrast, Fig 4.2 drawn for E. coli has its highest peaks near 125˚ and 200˚ and smaller ones around 140˚, 180˚ and 300˚. Similar distinctions are clearly visible from the other diagrams as well.

VI. Comparison of DNA-descriptors and feature descriptors

The DNA-descriptors have been used to generate the feature descriptors for a species. Now, we present the advantages gained herewith. In figure 5, we have plotted the DNA-descriptors obtained from different samples of the same species Drosophila. The plots are in the bar-diagram format.

In figure 6, we have drawn the feature descriptor vectors obtained by applying PCA (as described in section 4) for different sets of samples for the same species. These diagrams are in the polar plot format as described in the previous section.



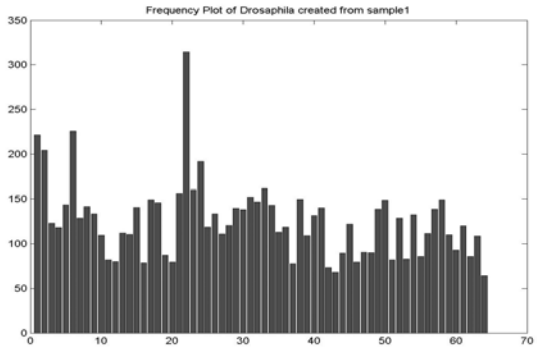


Fig. 5: Frequency count data plotted for two samples of the species Drosophila

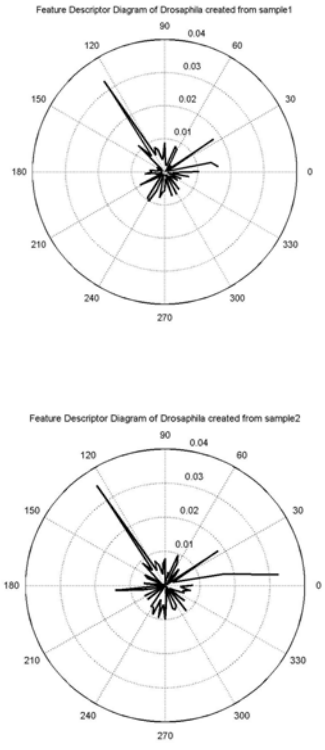


Fig. 6: Feature Descriptor vectors drawn in the form of Feature Descriptor diagrams for two different sets of DNA-descriptors from the DNA sequence of the species Drosophila

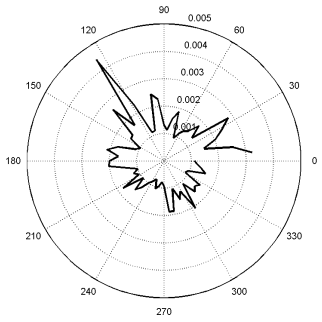


Fig. 7: Polar plot of DNA-descriptor

In figure 7, we have plotted a sample DNA-descriptor in the polar form. It is quite evident from the figures that the feature descriptors provide a more unique identifier for the species from its genomic data. Thus we have certainly gained an advantage by incorporating the data reduction tool PCA into our search for an effective identifier for a species. From the above figures it is clear that there is a significant increase in accuracy after applying PCA to the frequency count data. It has been found out that the DNA-descriptors obtained from different samples of the same species contain wide disparities. Hence their diagrammatic representations alone cannot represent the species.

But the *Feature Descriptors* obtained after processing a different set of DNA-descriptors are unique and present absolutely no significant disparities. Hence the *Feature Descriptor Diagrams* can be used as the unique representation of the genomic characteristics of the different species.

As a further justification of the uniqueness of the *Feature Descriptor Diagrams*, we have plotted in figure 8, the mean and the variance vectors (both 1 × 64) obtained from the different *Feature Descriptor vectors.*

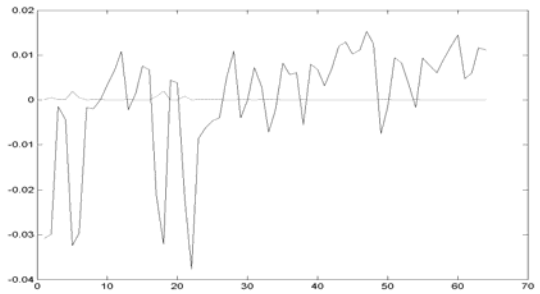


Fig. 8: Mean and Variance of Feature Descriptors

The variance (almost zero) of the descriptors for a given genome sequence being negligible, the proposed scheme finds extensive applications in automatic species identification.

VII. Some More Feature Descriptors Based on Mitochondrial Genomes

Now, that we have proved that the Feature Descriptors are unique, irrespective of the samples taken from the entire genome sequence, we shall henceforth work with only the small portion of the whole genome sequence of different species which correspond to their mitochondria. The mitochondria are those tiny modules present in every living cell, which act as the energy centre of the cell.

Due to their smaller size, the Feature Descriptors from the mitochondrial genomes are obtained by more rapid computational procedures. We have plotted below the feature Descriptors of different species obtained from their mitochondrial genomes.

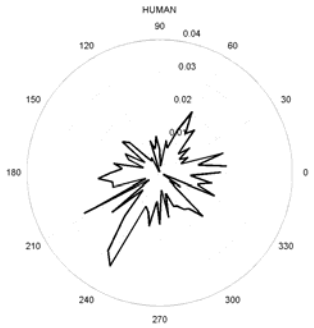


Fig. 9: Feature Descriptor Diagram for Human

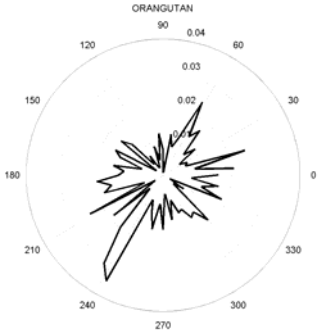


Fig. 10: Feature Descriptor Diagram for Orang-utan

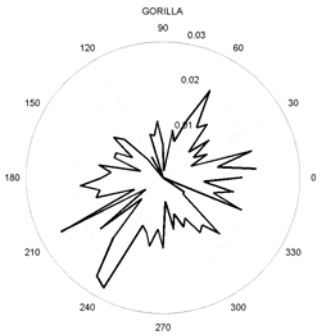


Fig. 11: Feature Descriptor Diagram for Gorilla

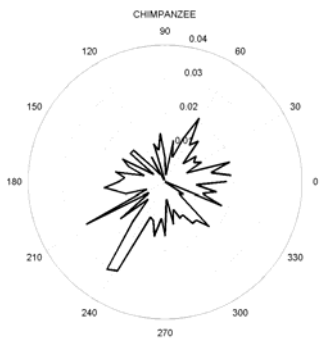


Fig. 12: Feature Descriptor Diagram for Chimpanzee

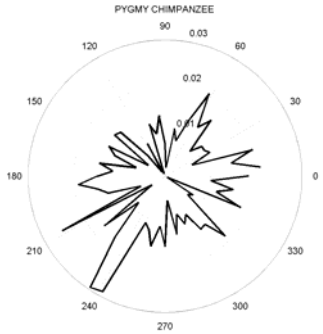


Fig. 13: Feature Descriptor Diagram for Pygmy Chimpanzee

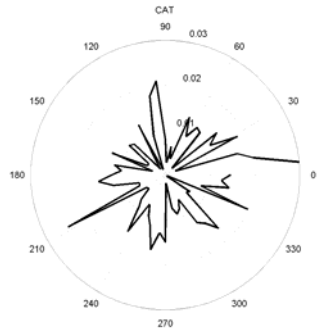


Fig. 14: Feature Descriptor Diagram for Cat

On seeing these diagrams, we can correctly conclude that the species Human, Orang-utan, Gorilla, Chimpanzee and Pygmy Chimpanzee have many similarities in their genome characteristics which can be translated to a similarity in their biological characteristics. However it is also quite clear from these diagrams that these species have some distinctions in their genomic characteristics. Also, a species like Cat which has different characteristics has a vividly distinct *Feature Descriptor diagram.*

VIII. Topological Clustering of DNA-Descriptors by SOFM

However there remains the cumbersome task of applying the abovementioned process to the genomic data of all the species to get figures corresponding to all the species. As an easier approach to automatic species identification, we present in this section a topological clustering method which will give us a single feature map whose different portions contain mappings from the extracted features of different species.

In this section, we make an attempt to map the DNA-descriptors onto a 2D array of neurons by the well-known Self-organizing Feature Map algorithm. Our main interest is to note whether DNA-descriptors of the same species occupy neighborhood neuronal positions and species having close resemblances in their DNA structure form neighborhood clusters.

To verify the above, we considered 36 vectors of each of the following 3 species: Mouse, Yeast and E. coli. Naturally, we have 36 × 3 = 108 vectors to be mapped onto the 2D array of size (k × k). To perform the experiment, we considered (6 × 6) dimensional space for the 2D array of neurons. Later the maps were created for different dimensions ranging from 4 to 11. Training principles and the algorithm are outlined below.

A. Training the map

INPUT: A set of 108 DNA-descriptor vectors each of size (1×64) obtained from 36 samples from each of the abovementioned 3 species. The vectors are in normalized form.

OUTPUT: Clustering of the DNA-Descriptors over a 2D array of neurons.

Normalization: Let the i^{th} (1 × 64) input vector be denoted by

$$\boldsymbol{a_i} = \begin{bmatrix} a_{i1} & a_{i2} & \ldots.. & a_{i64} \end{bmatrix}$$

To get the vector normalized we use the following transformation:

$$a_{ik} \leftarrow \frac{a_{ik}}{\sum_{j=1}^{64} a_{ij}}$$

Creation of Neuron Field: The neuron field of dimension (k × k) is constructed. Each neuron has a weight vector of (1 × 64) dimension. Initially, k is chosen as 6 and later the process was repeated for different values ranging form 4 to 11.

Initialization: All the k^2 neurons are initialized with random values ranging between 0 and 1. While initializing special care should be exercised to ensure that two neurons should not have identical weight vectors.

Choosing the value of learning rate constant η: In the beginning, we keep the value of η (eta), the learning constant, high (0.9) and gradually decrease it with each epoch until it reaches a very small value and thereafter η was kept constant at 0.005. The equation which governs the decay of η is given as:

$$\eta = 0.9 \times (1 - \frac{epoch}{\tau}) \quad \text{for} \quad epoch < \tau$$

where τ is a constant less than maximum value of epoch.

Choosing the size of neighborhood: Initially the neighborhood includes the entire neuronal space and is gradually decreased until it finally contains only the nearest neighbors of the winning neuron. Here the SOFM algorithm uses a neighborhood function which is convex, so as to avoid the occurrence of metastable states [20] which represent topological defects in the configuration of the feature map.

B. Phases of Training Process

We may decompose the adaptation of the synaptic weights in the network into two phases [21], [22]:

a) *Self-organizing or Ordering Phase:* It is during this first phase of the adaptive process that the topological ordering of the weight vectors takes place. In this phase, η is provided with a high value (0.9) and the neighborhood is defined large enough so that all the neurons are trained initially when any neuron wins for a particular data input. The neighborhood size also decreases in subsequent epochs.

b) *Convergence Phase:* The second phase of the adaptive process is needed to fine tune the feature map and therefore provide an accurate statistical quantification of the input space. This phase starts when ordering of similar types of neurons is complete. Then tuning is done to let the best neuron be trained most. In this phase, both eta and the neighborhood size are kept at a constant minimum value.

C. Algorithm

The algorithm for creation of the SOFM is as follows:

```
Begin

Initialize maxepoch

For epoch = 1 to maxepoch

  For each input data

    Compare the input vector with each neuronal weight vector by determining the Euclidean distance between them. The Euclidean distance
```

d_{ij} between the i[th] input data vector x_i and the j[th] neuron's weight vector w_j is computed using the following formula:

$$d_{ij} = \sqrt{\sum_{k=1}^{64} (x_{ik} - w_{jk})^2} \qquad (8)$$

```
    The neuron with the least distance is termed as the winner for that input. Then the winning neuron and the neighboring neurons (the size of the neighborhood depends on the epoch number according to the neighborhood function) are trained according to the following formula, where η is the learning constant:
```

$$w_{jk} = w_{jk} + \eta \times (x_{ik} - w_{jk}) \qquad (9)$$

```
  End For

End For

End.
```

D. Representation of the Map

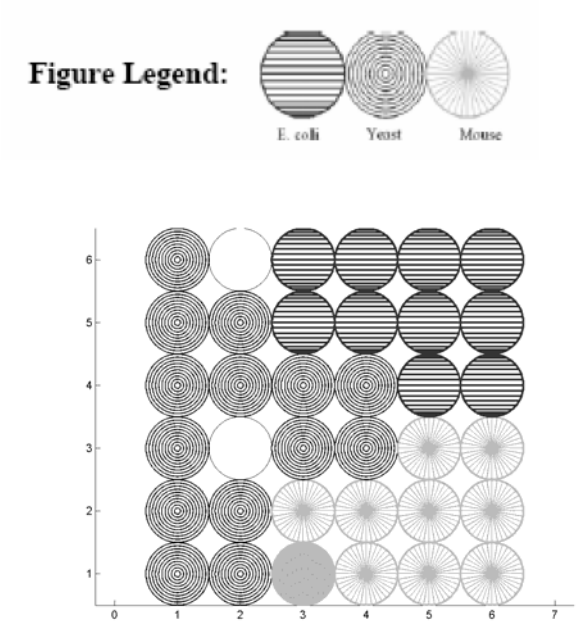
After the whole training process is complete, the SOFM is prepared. It is represented diagrammatically by a 2 D array of circles, each circle representing a neuron. The circles are shaded differently according to whether they were the winner for E. coli, Yeast or Mouse.

E. Cluster Centre

After mapping all the 108 input data vectors onto the 2 D array of neurons, it is noted that the inputs from the same species are mapped onto neurons occupying neighboring positions, thus forming different clusters for different species. Now we define the cluster center as the neuron belonging to a particular cluster which has emerged as the winner the maximum number of times, for that particular species.

IX. Simulation Results

We plot the trained SOFM for k = 6 below. The neurons which have won for Mouse, Yeast and E. coli are shaded differently. The blank circles have not won a single time for any species. We can clearly see that a distinct cluster is formed for each species.



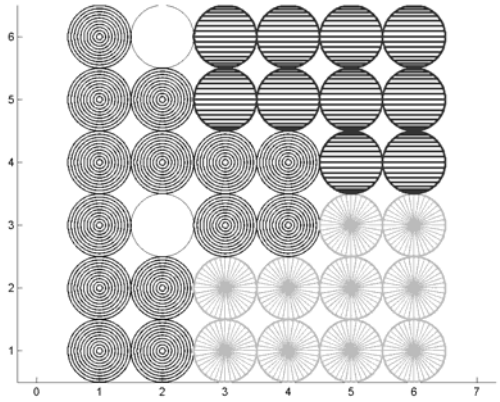
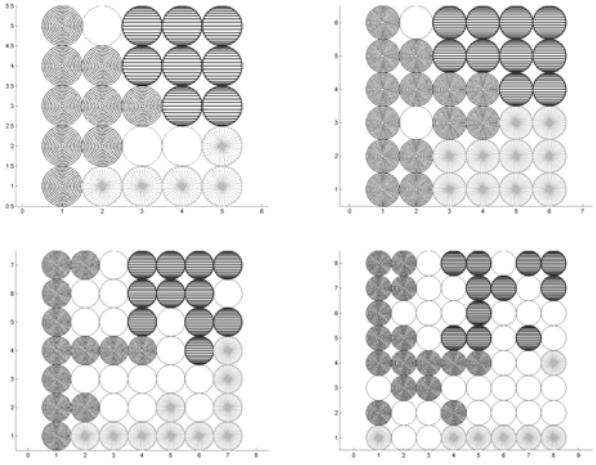


Fig. 15: Trained SOFM showing different clusters and the neuron winning for Mouse

Now we take a random sequence from the genome sequence of the species Mouse and perform frequency count on it. Using this vector as an input, the distance between this input vector and the weights of the neurons are calculated. The winner for this input, indicated by the filled circle in the above diagram, is found to be a neuron from the map belonging to the cluster for the species Mouse.

The above method offers a scheme for using the SOFM for automatic species identification. 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, then to which phylum the species belongs.

The following figures depict the maps obtained for different values of the map dimension from 5 to 10. If we chose the map dimension to be less than 4, the map becomes too small to distinguish between clusters of the 3 species. If it is greater than 11, we have to increase the number of inputs proportionately, either by increasing the number of species for which the map is constructed or the number of samples per species. They are plotted and compared for optimization of the map dimension.



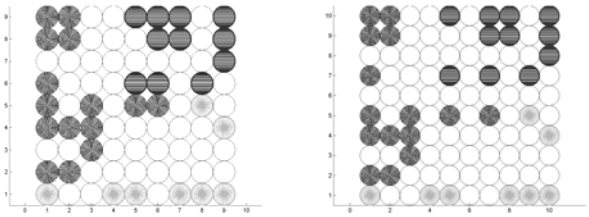


Fig. 16: The SOFMs for map dimension varying from 5 to 10

X. Interpretation of the Results and Performance Evaluation

We can see that, as the size of the map increases, the cluster corresponding to each species becomes more localized and concentrated. The topological property of a self-organizing feature map may be assessed quantitatively in different ways. One such quantitative measure, called the topographic product [23], may be used to compare the faithful behavior of different feature maps pertaining to different dimensionalities. However, the measure is quantitative only when the dimension of the lattice matches that of the input space [9]. Hence, there arose the need to define a new performance index.

To estimate how efficient the map is, we first find out the cluster center for each species in each map and then find the Euclidian distance of the other neurons belonging to that cluster from their cluster center. Now the mean and variance of the distance corresponding to each cluster are computed. The following figure contains a plot of the mean distance along with a tolerance margin (depicted by mean ± variance) for different values of the map dimension for the same species. This parameter i.e. the mean distance of the cluster members from the cluster center is defined as the error and is used as a figure of merit for the SOFM.

As is clearly visible from the figure 17, the error decreases as map dimension increases. This signifies that the cluster becomes more concentrated in a smaller region and the neurons which are a part of the cluster emerge as the winner more number of times as the map dimension is increased. This is also validated by the visual representations of the maps shown in Figure 16.

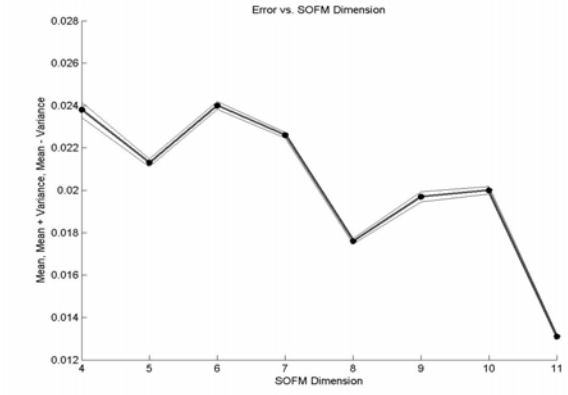


Fig. 17: Error vs. SOFM dimension

XI. CONCLUSIONS

Bioinformatics is a new area of science where a combination of statistics, molecular biology, and computational methods is used for analyzing and processing biological information like gene, DNA, RNA, and proteins. However no significant work has been done towards exploiting the fact that the genomic data of a species holds the structural signature of the species, hence can be used for identification and classification of the species. This paper aims to fill this gap.

To our knowledge, this is the first work of its kind to extract information from complete genome sequences and to distinguish between species by feature descriptor diagrams. Since the work entails processing huge amounts of genomic data, the learning ability of neural networks is utilized in this direction.

Since codons in a DNA consist of triplets of the 4 bases, hence the choice of 3-lettered sub-strings conveys some significance, as the codons are translated into amino acids, which are the building blocks of the proteins with which a species' body is constructed.

The micro array of DNA sequences for a species follows a repetitive pattern of nucleotide bases. This paper emphasizes the above statement by representing the micro array sequence by a specialized *feature descriptor vector*. The mapping of the DNA arrays to feature descriptors need not be unique. In fact; any type of nonlinear mapping that compresses the large DNA array to a vector of very small dimension could be employed to correlate the structural topology of *Feature Descriptor* with a given species.

In this process we have used PCA to reduce the large dimensions of genome sequence data without loss of accuracy. If only the frequency count is plotted then we do get some difference from species to species but it is not enough to distinguish between them. This is where PCA comes in. When PCA is applied to the original data we get enough differences between the *feature descriptor diagrams* of different species that enable us to tell one species from another with the help of these diagrams. Moreover when *feature descriptor vectors* for similar species are calculated, they are effective in bringing out the similarities in the species though they still retain their individual distinguishing features. Thus we claim that by constructing feature descriptor diagrams for each species we get an effective identifier for the species. However, we still need a quicker approach for automatic species identification. Here, we utilize the leaning and clustering abilities of computational maps.

After mapping all the 108 input data vectors onto the 2 D array of neurons, it is noted that data from the same species is mapped onto neurons occupying neighboring positions. Hence, it can be inferred that different vectors computed from different samples of the genomic data of a species are close in many respects and hence are mapped onto neighboring spaces on the map, thus forming separate clusters for different species. It can also be claimed that species which are close in characteristics will have similar DNA-descriptors and hence the clusters corresponding to similar species will lie in neighboring positions. Hence, if clustering techniques are applied to DNA-descriptors of a large number of species we will see that species which are similar in many respects e.g. Human and Gorilla will be forming sub-clusters within a super-cluster belonging to their families.

Also the SOFM can help us demonstrate homology between new sequences and existing phyla. 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, then to which phylum the species belongs.

Currently, works in Bioinformatics and biological data mining are aimed at discovering the parts of the DNA sequence which translate into proteins which lead to the development of different parts of the body i.e. to identify the genes and their functionalities. Another trend is to predict the structures of the proteins and their various conformations. No work has been directed towards utilizing the uniqueness of and similarities between DNA sequences of different species to identify and distinguish species. Hence the work described in this paper is a pioneer in this regard and carries possibilities for further enhancement in the direction of automatic species identification from genomic data.

REFERENCES

[1] "Special Issue on Bioinformatics, Part I: Advances and Challenges," Proceedings of the IEEE, vol. 90, November 2002.

[2] Smith, T. F., and Waterman, M. S., "Identification of common molecular subsequences," *J. Mol. Biol.* pp.147, 195-197, 1981.

[3] Waterman, M. S., and Eggert, M., "A new algorithm for best subsequence alignments with applications to tRNA-rRNA," *J.Mol.Biol.*, pp. 197, 723-728, 1987.

[4] Needleman, S. B., and Wunsch, C. "A general method applicable to the search for similarities in the amino sequence of two proteins," *J. Mol. Biol.* pp. 48,443-453, 1970.

[5] Cai, L., Juedes, D., and Liakhovitch, E., "Evolutionary Computation Techniques for multiple sequence alignment," *Congress on evolutionary computation 2000*, pp. 829-835, 2000.

[6] Galperin, M. Y., and Koonin, E. V., Comparative Genome Analysis, In Bioinformatics- A Practical Guide to the Analysis of Genes and Proteins, Baxevins, A. D., and Oullette, B. F. F.(Eds.), Wiley-Interscience, New York, 2nd ed., p. 387, 2001..

[7] States, D. J., and Boguski, M. s., Similarity and Homology, In Sequence analysis primer Gribskov, M., and Devereux, J. (Eds.), Stockton Press, New York, pp. 92-124, 1991.

[8] Mitra, S., and Acharya, T., Data Mining: Multimedia, Soft Computing, and Bioinformatics. New York: John Wiley, 2003.

[9] Haykin, S., Neural Networks: A Comprehensive Foundation, 2nd edition, India: Pearson Education, Inc., 1991.

[10] Kohonen, T., "The self-organizing feature map," *Proceedings of the Institute of Electrical and Electronics Engineers*, vol. 78, pp. 1464-1480, 1990.

[11] Hanke, J., and Reich, J. G., "Kohonen map as a visualization tool for the analysis of protein sequences: Multiple alignments, domains and segments of secondary structures," *Comput Applic Biosci*, vol. 6, pp. 447–454, 1996.

[12] Cai, Y. D., Yu, H., and Chou, K. C., "Artificial neural network method for predicting HIV protease cleavage sites in protein," *J. Protein Chem.*, vol. 17, pp. 607–615, 1998.
[13] Cai, Y. D., Yu, H., and Chou, K. C., "Prediction of beta-turns," *J. Protein Chem.*, vol. 17, pp. 363–376, 1998.
[14] Schuchhardt, J., Schneider, G., Reichelt, J., Schomberg, D., and Wrede, P., "Local structural motifs of protein backbones are classified by self-organizing neural networks," *Protein Eng*, vol. 9, pp. 833–842, 1996.
[15] Arrigo, P., Giuliano, F., Scalia, F., Rapallo, A., and Damiani, G., "Identification of a new motif on nucleic acid sequence data using Kohonen's self organizing map," *Comput Appl Biosci*, vol. 7, pp. 353–357, 1991.
[16] Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., et al., "The complete genome sequence of Escherichia coli," Vol. K-12, *Science*, pp. 277, 1453-1462.
[17] Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., et al, "The genome sequence of Drosophila melanogaster," *Science* pp. 287, 2185-2195, 2000.
[18] Cherry, J. M., Ball, C., Weng, S., Juvik, G., Schimidt, R., Alder, C., Dunn, B., Dwight, S., Riles, L. et al., "Genetic and Physical maps of Saccharomyces cerevisiae," *Nature* (suppl. 6632) pp. 387, 67-73, 1997.
[19] Smith, L. I., "A tutorial on Principal Components Analysis," 2002.
[20] Erwin, E., Obermayer, K., and Sculten, K., "II: Self-organizing maps: Ordering, convergence properties and energy functions," *Biological Cybernetics*, vol.67, pp. 47-55, 1992.
[21] Kohonen, T., "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol.43, pp. 59-69, 1982.
[22] Kohonen, T., "Exploration of very large databases by self-organizing feature maps,"1997 *International Conference on Neural Networks*, vol. I, pp. PL1-PL6, Houston, 1997.
[23] Bauer, H. U., and Pawezik, K.R., "Quantifying the neighborhood preservation of self-organizing feature maps," *IEEE Transactions on Neural Networks*, vol. 3, pp. 570-579, 1992.