Proceedings of the International Conference on Communication and Computational Intelligence – 2010,
Kongu Engineering College, Perundurai, Erode, T.N.,India.27 – 29 December,2010.pp.28-34.

Analysis of Genome Signature Strength of SARS Coronavirus Using Self-Organizing Map Neural Network

Francis Thamburaj[#1], Gopinath Ganapathy[#2]

[#1]*Computer Science Department, St. Joseph's College,*
Tiruchirappalli, Tamilnadu, India

[#2]*Computer Science Department, Bharathidasan University*
Tiruchirappalli, Tamilnadu, India

[1]francisthamburaj@gmail.com

[2]director@butp.org

***Abstract*— The nucleotide usage patterns vary not only from organism to organism, but also between genes in the same genome. Each genome has its own characteristics. This unique identity, called genome signature, of a genome is multidimensional. One of the ways to probe into this area is to analyze the nucleotide sequence composition of the genome. In this paper, the nucleotide compositional structure of SARS Corona virus which is the cause of the Severe Acute Respiratory Syndrome (SARS) is analyzed. Both the mono, di and tri nucleotides compositions are explored to find out the genomic nucleotide pattern. The Kohonen's self-organizing map neural network model is used as a tool to analyze the strengths of different nucleotide signatures of the genome. The analysis reveals that SARS virus is Thymine dominated, AT-rich and has the dinucleotide signature as qualitatively best signature, although codon and RSCU based SOM results in clearer cluster maps.**

***Keywords*— Genome Signature, Self-Organizing Map, Cluster Analysis, SARS Coronavirus, Unsupervised Neural Network**

I. INTRODUCTION

The comparative genomic studies show that organisms use nucleotides in a non-random, species specific fashion. The nucleotide compositional constraints, due to environmental pressures, affect both coding and noncoding regions and influence not only the structure but also the function of the genome [3]. Hence, the analysis of the nucleotide arrangement becomes crucial. But, predicting how much variation in nucleotide usage exists in a genome can be quite a difficult and challenging task. This article explores the nucleotide patterns that are buried in SARS genome.

As the molecular sequence data continue to grow exponentially due to the various sequencing projects, the nucleotide pattern analysis becomes crucial. The abundance of the genomic data, calls for new tools and techniques to identify biologically relevant features in the sequences. One of the recent tools to analyze the bio-sequence data is the Self-Organizing Map (SOM) invented by Teuvo Kohonen [8], [9]. The SOM neural network was applied successfully to a variety of problems, such as to extract uncommon sequences [2], to categorize interspecies genes [5], to analyze the codon usage [6], to classify prokaryotic and eukaryotic genomes [1], to identify atypical sequence composition [12], to discover motifs [13], to discover new groups in human endogenous retroviral sequences [15]. The list is endless. But none of them analyzed the SARS Corona virus composition using the SOM neural network.

SARS is an atypical form of pneumonia that first appeared in November 2002 in Guangdong Province, China. On April 7, 2003, WHO announced that it was generally agreed that a newly identified corona virus is the major causative agent of SARS. The SARS virus belongs to a class of viruses known as coronaviruses. They are distinguished by the presence of a single-stranded plus-sense RNA genome about 30 kb in length [14]. The genome has all the features characteristic of a coronavirus, but is sufficiently different from all previously known coronaviruses to represent a new coronavirus group. The SARS Corona virus genome contains five major open reading frames (ORFs) that encode the replicase polyprotein (rep); the spike (S), envelope (E), and membrane (M) glycoproteins; and the nucleocapsid protein (N) in the same order and of approximately the same sizes as those of other coronaviruses. [4], [16]. The nucleotide sequence can be obtained from the NCBI Genbank website

In this paper the nucleotide compositional analysis of SARS Corona virus is done with the aim to find out the strengths of different nucleotide signatures with the help of the SOM neural network. The SARS Corona virus genome is taken for this study. The complete genome sequence is got from the NCBI (Refseq: NC_004718, GenBank: AY274119). The length of the genome is 29,751 nucleotides (nt). It has 40% GC content and there are 13 genes with 14 protein coding

segments. The DNA content has linear topology. There are no structural RNAs or pseudo genes. As much as 97% of the genome is used for coding. It has two large overlapping reading frames (ORFs) encompassing 71% of the genome. These are the non-structural polyproteins with 6880 nucleotides and 4189 nucleotides. The third largest gene is the structural spike protein consisting of 1241 nucleotides. There are as many as 8 hypothetical proteins.

The rest of the paper is organized as follows. In the second part, the monomer and dimmer nucleotides are explored to find out the compositional pattern of the virus. The third part deals with the codon compositional patterns. The fourth part explains the necessary concept of Relative Synonymous

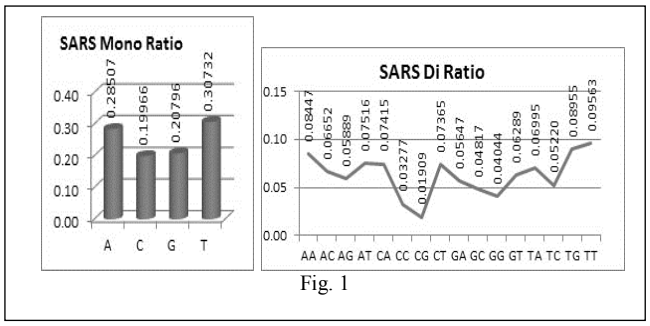

Fig. 1

Codon Usage (RSCU) along with detailed RSCU analysis of the SARS Corona virus and the fifth part gives a short review of the Kohonen Self-Organizing Feature Map. The final part analyses the genomic nucleotide patterns simultaneously with the help of the SOM in order to find out the strengths of various signatures, followed by the conclusion.

II. Mono and Dinucleotide Patterns of SARS

It is well known that both individual genes and entire genomes can vary significantly in nucleotide composition [3]. In fact, one of the most striking features of genomes is the range of nucleotide compositions represented. Some organisms have genomes that are disproportionately rich in guanine and cytosine (G and C), while others have DNA that is rich in adenine and thymine (A and T). For example, *Borrelia burgdorferi* has an overall GC content of only 25.5% in its coding DNA, while *Mycobacterium tuberculosis* has a GC content of 65.9% [17]. So, the nucleotide compositional variations of a genome form the basis of the genome signature.

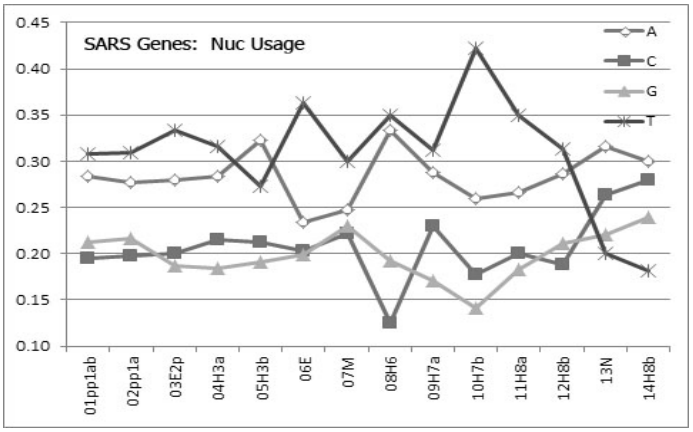

Fig .2

Analyzing the usage of the nucleotides in SARS genome, we see that Adenine and Thymine are used more than the Guanine and Cytosine. The usage priority order goes from Thymine, Adenine, Guanine and Cytosine in a descending order (Fig. 1). Therefore, the SARS virus has A & T rich nucleotide composition. It is evident from the overall GC ratio which is only 40.76% whereas the AT ratio is 59.23%. This AT bias leads to the possibility of high amount of AT-rich codons. Therefore, SARS virus would encode proteins rich in the Phenylalanine(F), Tyrosine(Y), Methionine(M), Isoleucine(I), Asparagine(N), and Lysine(K) amino acids. At the same time we can expect that the Glycine(G), Alanine(A), Arginine(R) and Proline(P) amino acids will be less since these are GC-rich amino acids and the GC percentage of the SARS genome is low.

As far as the protein coding genes are concerned, there are 14 protein coding genes in the SARS corona virus. The nucleotide patterns of these 14 genes are given in the figure (Fig. 2). The global genomic bias of Thymine and Adenine amino acids is reflected in all the genes. It is interesting to note the signature of each protein coding genes in terms of the number of various nucleotides. In Replicase gene Thymine dominates followed by Adenine guanine and Cytosine. In this way the mono nucleotide signature can be framed as TAGC. Similarly we can form the nucleotide signatures of the other genes. For Spike gene we have the signature as TACG (S), and other signatures are TACG, ATCG, TACG (Eprotein), TAGC, TAGC, TACG, TACG, TACG, TAGC,

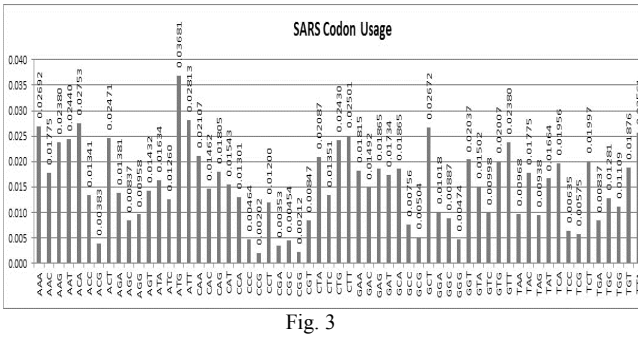

Fig. 3

ACGT(Capsid), ACGT, CGAT respectively. Looking at the signatures, we can see that signatures starting with T are dominating (10 numbers) which is nothing but the reflection of the genomic domination of Thymine. Adding the

percentages of each amino acid can make the signature more unique and not repetitive.

III. Tri-nucleotide Patterns of SARS

Codon usage pattern can reveal many things about a genome as well as the genes of a genome. Codon usage patterns vary not only from organism to organism, but also between genes in the same genome. Predicting how much variation in codon usage exists in a genome can be quite a difficult task. Variation in codon usage bias within a genome suggests many factors like translational selection and mutation-drift etc. [18]. Each and every genome has a particular codon bias. Then a gene from that genome must have the same pattern of codon bias. If it is not the case, then the deviation from the genomic codon bias can be measured and used as an index that can be used for various gene identification or similarity detection purposes. The 64 codon usage pattern is given in Fig.3 with the percentages.

Apparently it looks as if there is no correspondence between the genomic signature and the individual genes. In general the codon pattern analysis needs large number of codons which is possible only in long genes. One can see that the larger genes comparatively kept up the signature trend of the genome. For example the first gene has kept up the trends of the genomic signature.

IV. Relative Synonymous Codon Usage

The variation in nucleotide composition is usually most pronounced at the synonymous codon positions of genes. In a given code, codons that encode the same amino acid are called synonymous codons, while those encoding different amino acids are non-synonymous codons. The specificity of the synonymous codon usage is found not only across the genomes but also across the genes within a genome. The synonymous codon usage depends very much on the nucleotide bias of the particular genome [6]. The Genome Hypothesis postulates that genes in any given genome use the same coding pattern with respect to synonymous codons. Genes in an organism or in related species generally show the same pattern of codon usage [11].

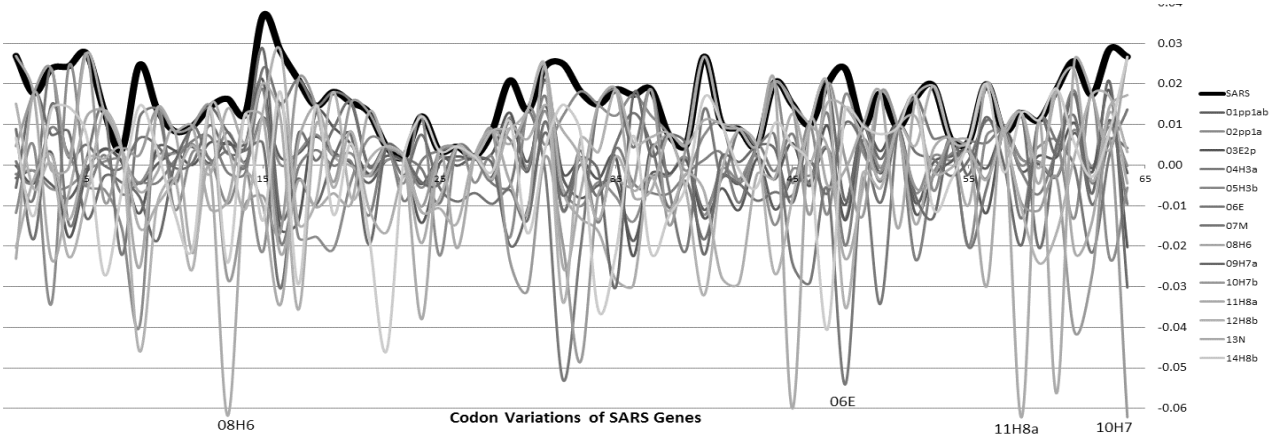

Fig. 4

The genomic bias of SARS protein encoding genes is shown in the following figure (fig 4). The thick line represents the codon usage variations of SARS corona genome. The milder lines show the codon variations of 14 protein coding genes of the genome. One could see clearly that some genes have large deviations from the genomic codon pattern. Especially, the Envelope gene (06E), and the hypothetical genes eight (08H6), ten (10H7), eleven (11H8a) are highly biased (around 60%). These genes are highly contaminated and changed from the global pattern of codons. This fact implies that large amount of translational selection and mutation drift has occurred in these genes. Since these codon biases are due to the necessity, these protein coding genes may be highly expressed compared to the other genes that do not have high genomic bias.

More complex nucleotide signatures can be framed using the codon counts. Such signatures are formed for SARS genome and all of its protein coding genes that are taken from the NCBI. From the figures given in '*SARS Protein Coding Gene Signatures*' (Fig. 5), it is clear that each protein coding gene has got its own unique signature while the genome has got the overall genomic signature.

Most codons have at least one synonymous alternative with the exception of those coding Methionine and Tryptophan. Within the protein-coding regions of most sequenced genomes, the occurrence of synonymous codons does not appear to be random. In other words, genomes seem to display

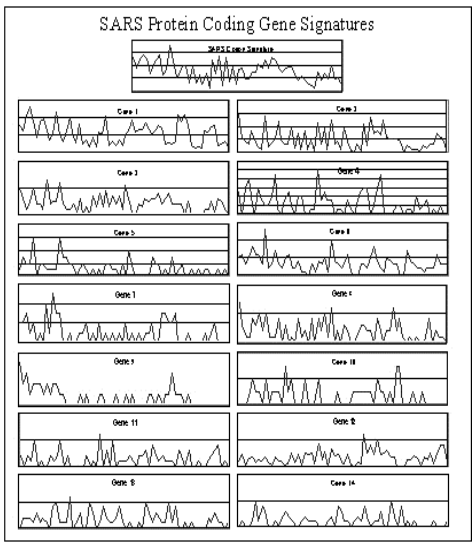

Fig. 5

a clear preference for one codon over a synonymous alternative. This preference is known as the synonymous codon usage pattern of a genome. Relative Synonymous Codon Usage (RSCU) measure for codon 'i' is calculated using the formula $RSCU_i = Obs_i/Exp_i$ where $Observed_i$ is the observed number of occurrences of codon 'i', and $Expected_i$ is the expected number of occurrences of the same codon (based on the number of times the relevant amino acid is present in the genome and the number of synonymous alternatives to 'i', assuming a uniform choice of synonymous codons) [12].

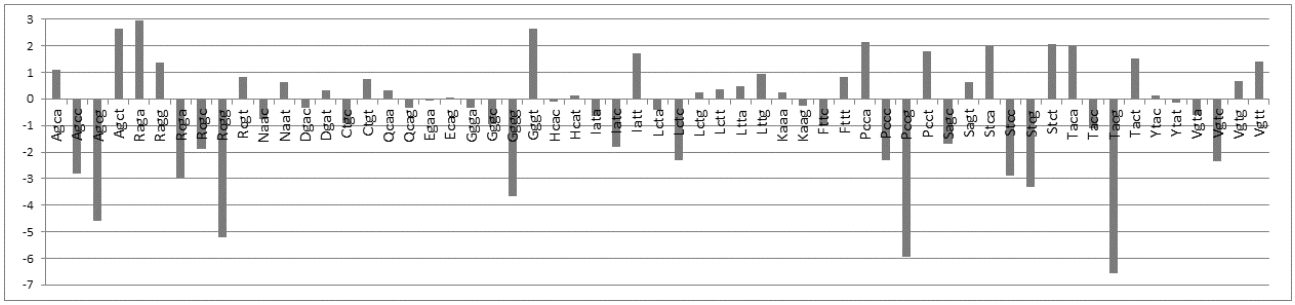

Fig. 6

In order to make the data more compatible with the mathematical methods used, the log10 of each RSCUi value is found in order to center the value around zero so that the resulting value is positive if the codon is used more than expected in that genome and negative if the codon is used less than expected. For each sequence, RSCU values are calculated for each of the 59 codons with synonymous alternatives.

The relative synonymous codon usages of SARS Corona virus is given in a graphical form in Fig. 6. The downward bars from the origin denote the negative measures and the upward bars denote the positive measures of the RSCU values. One can see that many codons are used less than the expected values. The first capital letter of the label denotes the amino acid and the next three letters denote the codon nucleotides. For example, the codons GCA, GCC, GCG, and GCT codes for the Arginine amino acid. Out of the four codons, GCA and GCT are used more than the expected value, while GCC and GCG are used less. This means that SARS Corona virus prefers GCA and GCT than the other two codons for coding the Arginine amino acid.

In order to identify more subtle patterns of nucleotides, SOM neural network is used as a tool. The general algorithm is given in the next section of the article. This method has the ability to automatically cluster the similar genome sequences together. This architecture consists of a two-dimensional output "lattice" of weight vectors. During training of the SOM, the weight vectors end up representing various popular patterns in the dataset and similar patterns are clustered together in neighboring areas of the output lattice. The result is that we can easily visualize variation within a dataset of patterns.

V. Review of SOM Algorithm

The Self-Organizing Map (SOM) is an unsupervised neural network algorithm, used to visualize and cluster high-dimensional data. The SOM is based around the concept of a lattice of interconnected nodes, each of which contains a model and in our case it is a vector of relative frequency ratios of mono, di, codon or RSCU. The models change during training to become similar to common or repeated patterns in the training set. Similar models are clustered together. The SOM used in our analyses of SARS virus is 12x8 nodes in a rectangular lattice configuration. The full SOM training algorithm is explained elsewhere [10]. but we summarize for our use here:

1) A gene's characteristic vector is loaded form the training dataset.
2) The lattice node is found whose gene character model most closely resembles the input vector. This node is denoted as the 'winning node'.
3) The winning node's model, W, as well as a certain number of 'neighborhood bubble' node models, are changed to be more similar to the input vector using
$$\mathbf{W}_{new} = \mathbf{W}_{old} + \eta(\mathbf{X}_i - \mathbf{W}_{old}).$$
4) If all vectors in the training dataset are processed, we say that an epoch has been completed.
5) Repeat the above four steps until number of desired epochs are over or desired error limit is reached.

VI. SOM for Signature Strength Identification

With the help of the SOM neural network the SARS Corona virus genome is studied in order to identify the strengths of various genomic signatures based on monomer, dimmer, codons and RSCU patterns. For the purpose of our study, 250 genomic sequences are artificially generated. Each of these sequences is made up of 29,751 nucleotides. Out of these 250 sequences, 50 are made randomly, and the other 200 sequences are created with the bias weightages of monomers, dimmers, codons, and amino acids of the SARS Corona virus. Along with these 250 artificial sequences, the original sequence is also added to the dataset consisting of 75,67,501 nucleotides. The above process is adopted to study the uniqueness of genomic signatures of closely resembling genomes and the ability of SOM to recognize these subtle differences in these sequences.

The frequency ratios of monomers, dimmers, codons, and

the RSCU values of these genomes are calculated to create various characteristic vectors. The dimensions of these input vectors for the SOM neural network are of 4 and 16, while for the codon vector is 64. The SOM neural network is chosen precisely to deal with such large dimensional data. The RSCU vector contains all the 59 possible codons leaving aside the Methionine and Tryptophan amino acids which do not have synonymous codons and three stop codons. These different characteristic vectors are fed into the SOM in order to train the neural network for clustering them into proper groups.

The figure (Fig. 7) shows the clustering of the genomic sequences based on the monomer signature. As seen in the picture, that the strength of the signature can only separate the random sequences from non-random sequences that are coding the SARS Corona virus. The monomer signature is not capable of separating groups within the coding genomes. This proves the basic fact that the nucleotide patterns are not random in any given genome sequence.

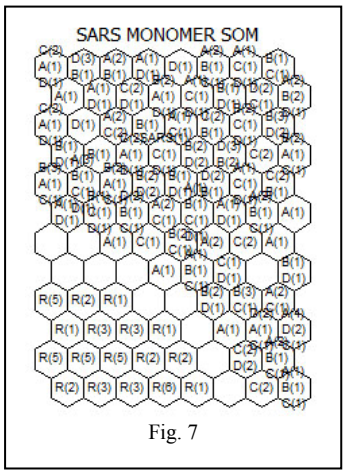

Fig. 7

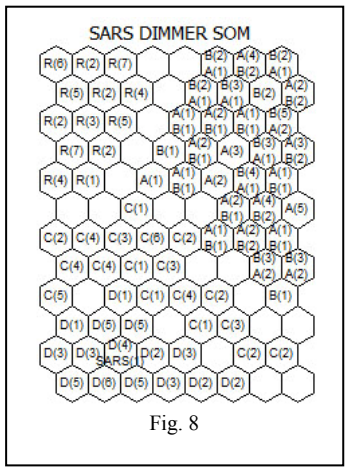

Fig. 8

The clustering based on the dimmer signature is capable of identifying the groups within the coding genomic sequences (Fig. 8). This scenario proves the fact that dinucleotide is unique to each organism and varies from genome to genome. But, the identified clusters are not separated as clearly as the random sequences from the coding sequences. Though the 'C' group and 'D' group seems to be mingling, the other representation like 'hit ratio diagram' (which is not shown here) of SOM clusters shows that it is not so. This shows that this signature is stronger than the monomer signature. The original SARS Corona virus is clearly identified and placed in the middle of SARS like di-nucleotide patterns.

The SOM clustering on the basis of the codon frequency ratio (Fig. 9) shows clearer separation of the different SARS virus groups. The A and B types of sequences have, though appears mixed, many have found unique neuron. These are the sequences generated using the amino acid frequency weights and base frequency weights. The difference between these two types of sequences is very subtle and close. This may be the reason for grouping them together. It can also be interpreted as the limit of codon signature to segregate the different groups. In other words, it is the measure of the codon signature strength.

Looking at the picture of the SOM clusters based on the RSCU values (Fig. 10), again we see a clear cut grouping of the random, dinucleotides, and codons biased sequences. There is a little attempt to classify the mixed groups of monomer and amino acid biased sequences. There is an empty neuron in the SOM map. Above this neuron there are four neurons containing only the 'B' type sequences and there are two neurons below containing only the 'A' type sequences. This attempt is also found in the codon based SOM clustering. Therefore, the RSCU based signature is equally stronger as the codon based signature.

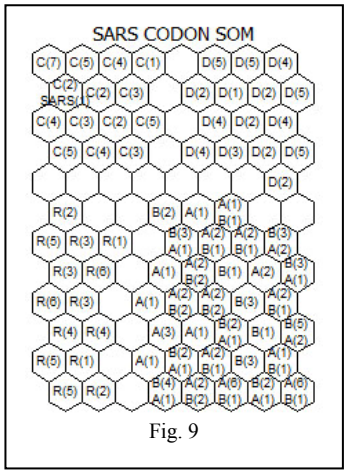

Fig. 9

In all the three (Di, Codon, RSCU) SOM clustering pictures, the original SARS genome is placed well in the middle of the group based on which the SOM map is constructed. This indicates that the original sequence has strong dual nucleotide signatures: dinucleotide signature and trinucleotide or codon signature.

The inability to divide the biased sequences based on the amino acid counts and monomer counts, reveal the fact that the monomer mutations and amino acid mutations do not radically change the strength of the signature. The second fact is that these mutations do not turn the sequences into a random sequence. They stand as a unique group separated from the random sequence group. A relaxation factor does exist in the basic identity of the coding genome sequence. This implies that there may be inevitable mutations due to the environmental pressures and other causes. But they do not affect the basic nature of the whole genome and their di and tri nucleotide signatures. In other words, the basic identity of the genome is unaltered due to the small changes in the genomic sequences.

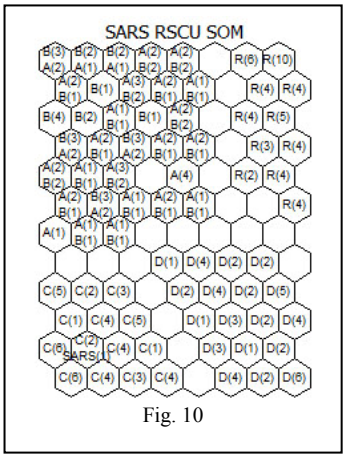

Fig. 10

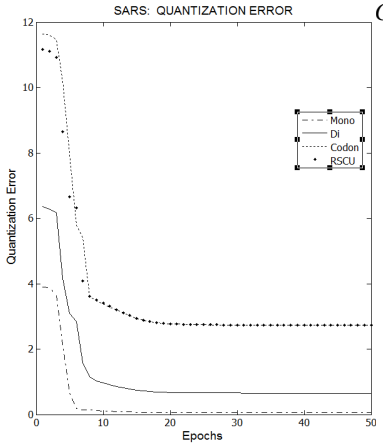

Fig. 11

The quality of the above SOM maps may reveal more information about the strengths of the genomic signature. Two quality measures of the four SOM maps are calculated and plotted for the comparative analysis. The first measure is the quantization error (Qe) that calculates the average distance between each input vector and its 'best matching unit' (BMU) [10]. Therefore, the small value of Qe is more desirable. Hence, the accuracy is minimum for the RSCU and codon value based SOM and less than one, which are acceptable, for the monomer and dimmer value based SOM. The best quality is achieved by monomer SOM map followed by the dimmer SOM map.

Another quality measure of the SOM map is the topographic error (Te). It describes how well the SOM preserves the topology of the studied data set [7]. It calculates the average ratio of the number of nodes for which the first and the second winners are not neighbours to each other. Therefore small value of Te is more desirable. Unlike the quantization error it considers the structure of the map. The dimmer based SOM map has achieved the best topographic quality followed by the monomer and then the codon and RSCU based SOM maps.

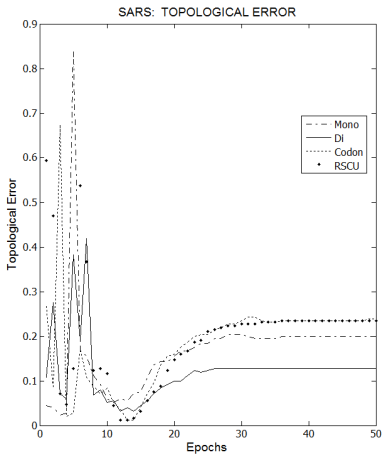

Fig. 12

Comparing both the qualities, the dimmer based SOM has achieved the overall best quality, by having 0.5 quantization error and 0.12 topographic error. So, the dinucleotide signature is the strongest of all signatures as it enables to cluster the closely similar genomic sequences and separate them as individual groups.

VII. CONCLUSION

From the above analysis of the mono nucleotide composition of the SARS virus, shows that the majority of the protein coding genes have the signature with T start and this is due to the genomic domination of Thymine nucleotide. The di nucleotide analysis shows that the genome is AT-rich genome. The tri nucleotide analysis shows that the protein coding genes do have unique codon signatures. The SOM based cluster analysis reveals that the monomer signature is more than enough to separate the coding genomes from the random genome sequences. But to find the clusters among the coding genomes, we need higher order signatures based on the relative frequency of the dinucleotides, codons, and RSCU. The experiments prove that the dinucleotide signature is qualitatively best signature, while the codon and RSCU signatures give clearer cluster separation. Also, it has to be noted that none of these signatures got the strength to cluster and separate the amino acid frequency wise biased genomic sequences from the monomer frequency wise biased genomic sequences. Some new nucleotide signatures may be studied to explore the possibility of clustering them also.

REFERENCES

[1] Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. and Ikemura, T. A Novel Bioinformatic Strategy for Unveiling Hidden Genome Signatures of Eukaryotes: Self-Organizing Map of Oligonucleotide Frequency, *Genome Informatics*, vol.13, pp.13–20, 2002.

[2] Arrigo et al., *Identification of a New Motif on Nucleic Acid Sequence Data Using Kohonen's Self-Organizing* Map, *Bioinformatics*, vol. 7, no. 3, pp. 353-357, 1991.

[3] Bernardi G., *Compositional Constraints and Genome Evolution*, *Journal of Molecular Evolution*, vol. 24, pp. 1-11, 1986.

[4] Holmes Kathryn V., and Enjuanes Luis, "The SARS Coronavirus: A Postgenomic Era, *Science, Virology: Perspectives*, vol.300, pp. 1377-1378, May 30, 2003.

[5] Kanaya, S., Kudo, Y., Abe, T., Okazaki, T., Carlos, D.C., and Ikemura, T., Gene *Classification by Self-Organization Mapping of Codon Usage in Bacteria with Completely Sequenced Genome*, *Genome Informatics*, vol. 9, pp. 369-371, 1998.

[6] Kanaya Shigehiko, Kinouchia Makoto, Abe Takashi, Kudoe Yoshihiro, Yamadae Yuko, Nishid Tatsuya, Morib Hirotada, Ikemuraf Toshimichi, *Analysis of Codon Usage Diversity of Bacterial Genes with a Self-Organizing Map (SOM): Characterization of Horizontally Transferred Genes with Emphasis on the E. coli O157 Genome*, *Gene*, vol. 276, pp. 89-99, 2001.

[7] Kiviluoto, Kimmo, *Topology Preservation in Self-Organizing Maps, The 1996 IEEE International Conference on Neural Networks*, ICNN. Part 1 (of 4); Washington, DC; USA; 03-06 June 1996. pp. 294-299.

[8] Kohonen Teuvo, *Self-Organizing Formation of Topologically Correct Feature Maps, Biol. Cyb.*, vol. 43, no. 1, pp. 59-69, 1982.

[9] Kohonen Teuvo,*The Self-Organizing Map*, *Proceedings of the IEEE*, vol.78, no.9, pp.1464 – 1480, Sep 1990.

[10] Kohonen Teuvo, *The Self-Organizing Maps*, 3rd Ed., Springer-Verlog, Germany, 2001.

[11] Li W.-H., *Molecular Evolution*, 2nd edn. Sunderland, MA: Sinauer Associates, 1997.

[12] Mahony Shaun, McInerney O. James, Smith J. Terry, and Golden Aaron , *Gene Prediction using the Self-Organizing Map: Automatic Generation of Multiple Gene Models*, *BMC Bioinformatics*, vol. 5, no. 23, pp. 9, 2004.

[13] Mahony Shaun Aengus, *Self-Organizing Neural Networks for biological Sequence Analysis*, Doctoral Thesis, National University of Ireland, Galway, December 2005.

[14] Marra Marco A., Jones SJM, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS, et al., *The Genome Sequence of the SARS-Associated Coronavirus, Science,* vol.300, pp. 1399-1404, May 30, 2003.

[15] Oja Merja, *Self-Organizing Map based Discovery and Visualization of Human Endogenous Retroviral Sequences Groups*, *Int. Jr. Neural Systems*, vol. 15, no. 3, pp.163-179, 2005.

[16] Rota Paul A., Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, et al., *Characterization of a novel Coronavirus Associated*

with Severe Acute Respiratory Syndrome, vol.300, pp. 1394-1399, May 30, 2003.

[17] Singer A. C. Gregory and Hickey A. Donal, *Nucleotide Bias Causes a Genomewide Bias in the Amino Acid Composition of Proteins*, *Molecular Biology and Evolution*, vol.17, no.11, pp. 1581-1588, 2000.

[18] Wang Huai-Chun, Badger Jonathan, Kearney Paul, and Li Ming, *Analysis of Codon Usage Patterns of Bacterial Genomes Using the Self-Organizing Map*, *Molecular Biology Evolution*, vol.18, no.5, pp. 792-800, 2001.

AUTHORS BIOGRAPHY

First author, Francis Thamburaj, is a professor and an active research scholar in Computer Science Department, specializing in Artificial Neural Networks.