Dimer Patterns in Database of Viral Genomes: An Analysis with GHSOM

Ernesto Bautista-Thompson[1], Gustavo Verduzco-Reyes[1], and Luis De la Cruz-De la Cruz[2]

[1] Centro de Tecnologías de la Información, DES-DACI, Universidad Autónoma del Carmen, Avenida 56 Número 4,
C.P. 24180 Ciudad del Carmen, Campeche, México
[2]DACB, Universidad Juaréz Autónoma de Tabasco
C. P. 86690 Cunduacán, Tabasco, México
{ebautista, gverduzco}@pampano.unacar.mx, santanadelacruz@gmail.com

Abstract. An analysis about the dimer patterns in a database of 150 genomes of viruses from sixteen taxonomic families was developed with the technique Growing Hierarchical Self-Organized Map (GHSOM), the GHSOM neural network allows the hierarchical clustering of the viruses by their similarity features in our case dimers frequencies. The clusters generated shows certain degree of correspondence with the taxonomic families but a sharp differentiation was not observed. In the case of the Retroviridae family was observed a strong dispersion of their members between different clusters, reflecting diversity in the frequencies of their dimers. Some families are characterized by specific dimers such as: AA, AT, TA, and TT, as the case of the Poxviridae family.

Keywords: Dimer Patterns, Virus Genome, GHSOM.

1 Introduction

Studies of nucleotide sequences from DNA are of interest because the insight that they can provide about the evolutionary processes of species [1]. In particular, the study of dimer sequences is of interest due to the hypothesis that exists a relation between dimer statistical distribution and the basic conditions for DNA physicochemical stability [2, 3], and also because is possible that dimer distribution is related with a genetic signature useful for phylogenetic and taxonomic classification of species based on a underlying level of information not present in trinucleotide sequences (codons) that are known to carry on the coding information in DNA [4].

Different studies are reported in scientific publications about the application of clustering techniques for the analysis of genomic sequences in

Advances in Artificial Intelligence: Algorithms and Applications
Research in Computing Science 40, 2008, pp. 3-12

gene expression studies, comparison of interspecies characteristics, DNA clone classification, analysis of coding and non coding regions in DNA, focused on DNA from bacteria, plants and animals [5, 6, 7, 8], but few studies were found about analysis and search of similarity patterns on viral genomes [4, 9].

In order to search and identified similarity patterns between the viral genomes at the dimer level, we applied the neural network: Growing Hierarchical Self Organized Maps (GHSOM) for the clustering task. We are interested in exploring different machine learning approaches such as GHSOM for the fusion of data from multiple features and origins in order to extract knowledge in genomic databases, and this work is part of such effort.

In section 2, we briefly present the taxonomic information about the database of viruses under study. In section 3, we describe the experimental methodology and the results of the search of dimer patterns with GSHOM. Finally, in section 4 we present the discussion of this work.

2 Taxonomic Features and Database of Viral Genomes

Viruses are one of the most primitive biological forms on earth, although there is a controversy about if they are living forms or not. They are believed to had been components of cells that became autonomous, in fact some virus are similar to portions of DNA sequences of genes, another hypothesis is that viruses evolved from unicellular organism [10]. There are a well known taxonomic classification of viruses based on the type of organization of viral genome, the strategy of viral replication and the structure of the virion [10, 11], but the explosion of taxonomic information available in public data bases thanks to the application of Information Technologies and the sequencing of virus genomes [11, 12], has complicated the analysis of the information and the discovery of new knowledge inside these databases. The application of techniques such as GHSOM for the dual task of clustering and visualization of the results, allows the identification of patterns of interest from the sets of features contained in genomic databases.

The set of viruses under study are representative of different taxonomic families (sixteen families in this study) and sources of different common and non common human and non human diseases [10, 12], see Table 1. We select randomly 150 virus genomes with different features, virus of DNA and RNA, highly aggressive virus as the Zaire Ebolavirus, virus that produce not very dangerous diseases as the Rhinovirus B and Coronavirus. The size of the genomes is also very variable; there are genomes of less than 5,000 bp like the Parvovirus 4 and genomes around 130,000 bp like the Herpesvirus 1 and Herpesvirus 4. All the genomic sequences were taken from the GenBank through the Entrez Documental Retrieval System [11, 13], and loaded inside a database that was built for the management of the collected data and the dimer frequencies data to be generated.

3 GHSOM and Dimer Patterns

Our feature space was the dimers frequencies for each viral genome. The sixteen dimer combinations based on the four bases that form the genomic code are: AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, and TT. In order to be able to compare among different genomes it was necessary to convert the frequencies to percentages, since the genomes are of different size, so each dimer frequency was normalized by the total number of dimers of the corresponding genome. With the frequency data we built an intensity table, in which each row corresponds to a certain virus genome and the columns to the percentage of frequency of each dimer in the genome. This table is then used as input for the GHSOM technique; we apply the GHSOM Toolbox for MatLab® [14] for the generation of the maps. The Table 2 shows examples of the normalized frequency data for some viruses.

Table 1. Examples of different taxonomic features of the viruses under study.

Virus	Family	Molecule	Topology	Capsid	Envelope
Human adenovirus A	Adenoviridae	dsDNA	Linear	Icosahedral	No
Uukuniemi virus	Bunyaviridae	nssRNA	Linear	Helical	Yes
Sapporo virus	Caliciviridae	pssRNA	Linear	Icosahedral	No
SARS coronavirus	Coronaviridae	pssRNA	Linear	Helical	Yes
Sudan ebolavirus	Filoviridae	nssRNA	Linear	Helical	Yes
Dengue virus type 1	Flaviviridae	pssRNA	Linear	Icosahedral	Yes
Hepatitis B virus	Hepadnaviridae	dsDNA	Circular	Icosahedral	Yes
Human herpesvirus 1	Herpesviridae	dsDNA	Linear	Icosahedral	Yes
Measles virus	Paramyxoviridae	nssRNA	Linear	Helical	Yes
Human papillomavirus - 1	Papovaviridae	dsDNA	Circular	Icosahedral	No
Parvovirus H1	Parvoviridae	ssDNA	Linear	Icosahedral	No
Foot-and-mouth disease virus A	Picornaviridae	pssRNA	Linear	Icosahedral	No
Variola virus	Poxviridae	dsDNA	Linear	Icosahedral	Yes
Human immunodeficiency virus 1	Retroviridae	pssRNA	Linear	Helical	Yes
Rubella virus	Togaviridae	pssRNA	Linear	Icosahedral	Yes

Table 2. Examples of normalized frequency values used to build the intensity table.

AA	AC	AG	AT	VIRUS
5.23	6.11	4.79	5.36	HEPATITISB
2.54	6.15	3.99	2.59	HERPES1
4.08	5.89	6.55	3.72	HERPES4
4.54	7.29	4.71	3.81	HERPES5
2.19	5.92	3.91	2.29	HERPES2
9.91	6.15	5.59	7.66	HERPES6
13.24	5.81	5.45	9.47	HERPES7
8.05	6.8	4.7	7.6	HERPES3
5.68	6.88	6.31	4.89	HERPES8
7.71	6.38	6.46	6.02	ADENOB
7.74	6.61	5.66	4.73	ADENOF
5.11	6.5	6.33	4.13	ADENOE
5.88	6.46	6.46	4.55	ADENOD
11.98	5.45	6.3	9.47	HEMORRHAGICENT

The Growing Hierarchical Self Organizing Map (GHSOM) is an unsupervised clustering technique that allows the generation of hierarchies of clusters based on the similarity of the input data, the basis of this map is the SOM neural network that exploits the non supervised competitive learning, the algorithm generates a mapping that preserves the space topology of greater dimension in the space of the neuron units. The neuron units form a two-dimensional grid then a mapping from n-dimension to 2-dimension is generated. The property of topological preservation means that a SOM groups sets of vectors with similar information in neighbor neural units. A SOM network is able to generalize, in this way new information can be added and integrated to the map, also it is able to work with incomplete data inside the vectors [15]. The GHSOM is a variant from the SOM neural network where a hierarchy of multiple layers of SOM neural networks are generated (see Figure 1), each unit of the SOM can generated a new SOM network based on a dissimilarity threshold (quantization error), in this way a hierarchy of similarity clusters is created, the deepness of the hierarchy shows the non uniformity that can be expected from real world data sets [16]. There are other clustering techniques such as K-Nearest Neighbors, Multidimensional Scaling Analysis, Principal Component Analysis, etc. [17], but they are not able to generate a hierarchy of clusters, generalize (preserve the clusters when new information is added), and to work with incomplete input datasets; these features of the GHSOM technique were the factors for its choice as an analytical and visualization tool for the present study.

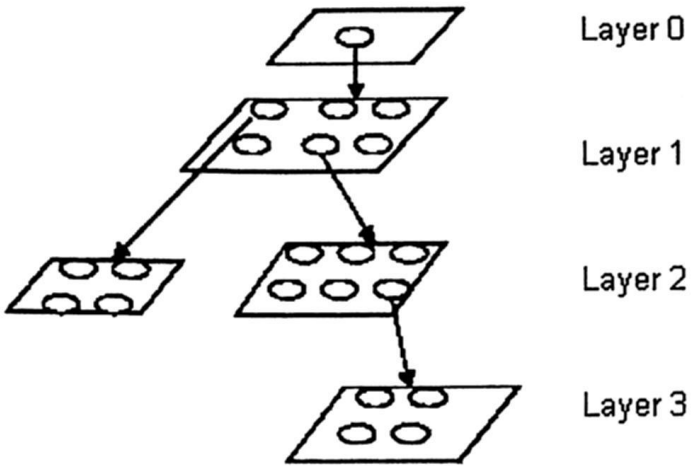

Fig. 1. Hierarchical generation of SOM layers inside a GHSOM neural network, when a new object represented by a node is dissimilar (quantization error) to its neighborhood a new layer is generated through this node, so a new cluster is created.

Instead of associate the specific name of the virus with its corresponding set of dimer frequency data, we associate such data with its corresponding taxonomic family. The generated GHSOM map was analyzed in order to identify global similarities between families of virus; the map shows a similarity hierarchy based on the contributions of the frequency values for the different dimers. Complementary maps were generated that shows in a grey scale the weight of each dimer for different regions inside the GHSOM map. The Table 3 shows the correspondence between the tags in the map, the associated virus family, and the number of virus for each family.

In the GHSOM map (see Figure 2), in general the similarity clusters to which the viruses belongs are in correspondence with the associated taxonomic families (grouping of the different viruses by their corresponding families), but some families presents a strong dispersion of its members: Picornaviridae (tag 13) and Retroviridae (tag 14) families, this shows that differences at the dimer feature level are greater for members of these families, in particular the immunodeficiency viruses belong to the Retroviridae family (see Table 1) and they are known to have a high rate of mutation so their genomic sequences are very variable [10, 18]. Some families have a strong localization of its members: Paramixoviridae (tag 8), Flaviviridae (tag 12) and Togaviridae (tag 15). This is indicative of a strong similarity between the genome of its members at the dimer level.

Table 3. Associated tags for the interpretation of the GHSOM map.

Family	Tag	Number of virus
Adenoviridae	1	7
Hepadnaviridae	2	1
Herpesviridae	3	8
Papovaviridae	4	1
Poxviridae	5	7
Bunyaviridae	6	1
Filoviridae	7	4
Paramyxoviridae	8	18
Rhabdoviridae	9	7
Caliciviridae	10	6
Coronaviridae	11	6
Flaviviridae	12	25
Picornaviridae	13	15
Retroviridae	14	23
Togaviridae	15	15
Parvoviridae	16	6

At the first level of the hierarchy that corresponds to the main four similarity cluster of the GHSOM map, the union of different families can be observed, some examples are in the lower right region of the map: Rhabdoviridae (tag 9), Caliciviridae (tag 10), Flaviviridae (tag 12), Togaviridae (tag 15), and some elements from the Retroviridae family (tag 14). In the lower left region of the map: Poxviridae (tag 5), some elements from the Paramyxoviridae family (tag 8), Coronaviridae (tag 11), and some elements from the Picornaviridae family (tag 13). In the upper left region of the map: Filoviridae (tag 7), Paramyxoviridae (tag 8), and the Parvoviridae family (tag 16). Then, we observed that at the dimer level new similarities between members of different families can be identified with this analysis.

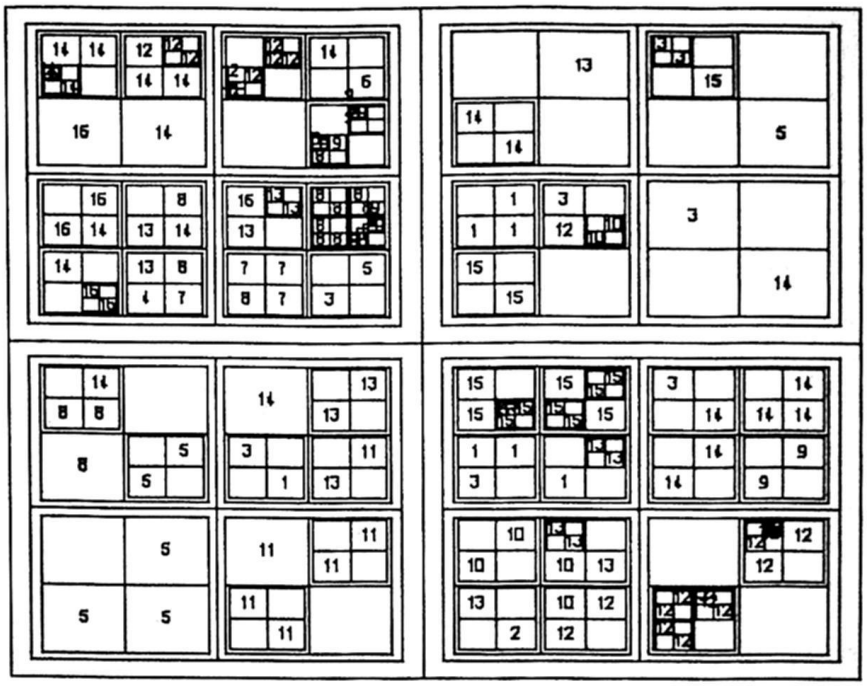

Fig. 2. Hierarchical Map showing the similarity at the taxonomic families level between different viral genomes, where such similarity is based on the dimers frequencies for each genome.

The Figure 3, shows a series of maps that corresponds to the different dimers, each map visualize the weight of a dimer for specific regions inside the GHSOM map, white regions means a strong contribution of the dimer for the elements inside such regions and black regions means a weak contribution of the dimer for the elements inside these regions. The map for the dimer AA shows a white region that corresponds to viruses that belongs to the Poxviridae family, this family is highly localized in the GHSOM map, then the high frequency of occurrence of the dimer characterize to this family. The same case occurs for the dimers AT, TA and TT, so these four dimers have a strong presence in this family of virus.

Another interesting case corresponds to the maps for the dimers: CC, CG, and GC; they show a strong contribution for the viruses grouped on the upper right region inside the GHSOM map.

In the cases of the maps for the dimers: AT, TA, and TT; they have a weak and uniform contribution for the virus grouped in the right side of the GHSOM map, in contrast the maps of the dimers: CC, CG, and GC; show that they have a strong contribution for the same virus grouped in the right side of the GHSOM map.

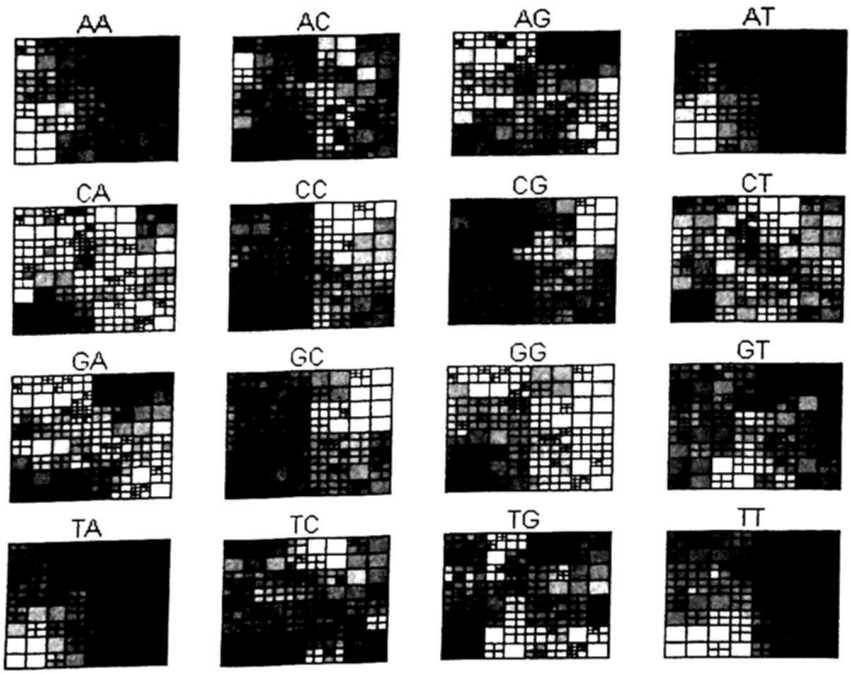

Fig. 3. Maps based on dimer weight, each map shows the contribution of each dimer frequency to the similarity on the corresponding viral genomes associated to each of the regions inside the GHSOM map, the white color corresponds to a strong contribution of the dimer in a specific region and the black color corresponds to a weak contribution of the dimer in a specific region.

4 Discussion

An analysis with the GHSOM technique was developed in order to identify at the dimer frequency level hierarchies of similarity patterns for viruses from different taxonomic families. At the dimer level certain degree of correspondence with taxonomic families was conserved, but a sharp differentiation by families was not observed. An interesting case was the Retroviridae family (which includes the immunodeficiency virus: HIV, SIV, FIV) the members of this family showed a strong dispersion between different clusters reflecting a diversity in the dimer frequency distribution for their genomes. With this analysis was possible to identify similarities between viruses from different families, for example: Filoviridae (tag 7), Paramyxoviridae (tag 8), and the Parvoviridae family (tag 16), where the Filoviridae family has some of the most lethal virus for the human being

(Ebola virus). From the analysis of the dimer contribution to the similarity between the viruses, it was observed that some dimers characterize strongly some families; this was the case of the dimers AA, AT, TA, and TT with the members of the Poxviridae family. The analysis of biological information with clustering techniques such as the GHSOM among others, at the dimer level and its connection with biological knowledge at upper levels such as the taxonomic classification can be a useful tool for the understanding of relations between multiple levels, for example: genomic (1D-structure) and morphological (3D-structure) by combining features from both levels and using techniques such as GHSOM for the identification of patterns of interest. In our case, more research is on the way in order to increment the quantity and detail of the biological data to be integrated from the genomes under study and the exploration and development of techniques for the analysis and visualization of data from genomic databases.

Acknowledgments. The first author thanks the financial support from PROMEP-SEP under the project of generation and application of knowledge UNACAR-PTC-085.

References

1. Stanley, R. H. R., Dokholyan, N. V., Buldyrev, S. V., Havlin, S., Stanley, H. E.: Clustering of Identical Oligomers in Coding and Noncoding DNA Sequences. Journal of Biomolecular Structure & Dynamics 17, 79-87 (1999)
2. Breslauer, K. J., Frank, R., Blöcker, H., Marky, L. A.: Predicting DNA Duplex Stability from the Base Sequence. Proc. Natl. Acad. Sci. USA 83, 3746-3750 (1986)
3. Miramontes, P., Cocho, G.: DNA Dimer Correlations Reflect In Vivo Conditions and Discriminate among Nearest-Neighbor Base Pair Free Energy Parameter Measures. Physica A 321, 577-586 (2003)
4. Quiroz-Gutierrez, A.: Biophysical Considerations and Evolutionary Aspects of DNA-dimer Frequency in AIDS Retrovirus Genomes. In: Topics in Contemporary Physics, pp. 239-248. IPN Press, México (2000)
5. Frith, M. C., Li, M. C., Weng, Z.: Cluster-Buster: Finding Dense Clusters of Motifs in DNA Sequences. Nucleid Acids Research 31, 3666-3668 (2003)
6. Abe, T., Sugawara, H., Kanaya, S., Kinouchi, M., Ikemura, T.: Self Organizing Map (SOM) Unveils and Visualizes Hidden Sequence Characteristics of a Wide Range of Eukariotic Genomes. Gene 365, 27-34 (2006)
7. Figueroa, A., Borneman, J., Jiang, T.: Clustering Binary Fingerprint Vectors with Missing Values for DNA Array Data Analysis. In: Proceedings of the Computational Systems Bioinformatics (CSB'03), pp. 38. IEEE Computer Society Press, U.S.A (2003)
8. McCallum, J., Ganesh, S.: Text Mining of DNA Sequence Homology Searches. Applied Bioinformatics 2(3 Suppl), S59-S63 (2003)
9. Gatherer, D.: Genome Signatures, Self-Organizing Maps and Higher Order Phylogenies: A Parametric Analysis. Evolutionary Bioinformatics Num. 3, 211-236 (2007)

10. Brooks, G. F., Butel, J. S., Ornston, L. N., Jawitz, E., Melnick, J. L., Adelberg, E. A.: Jawitz, Melnick, and Adelberg's Medical Microbiology. Prentice-Hall, U.S.A. (1991)
11. Büchen-Osmond, C.: The Universal Virus Database ICTVDB. Computing in Science & Engineering May/June 2003, 2-11 (2003)
12. Van Regenmortel, M. H. V. et. al. (eds.): Virus Taxonomy. Classification and Nomenclature of Viruses. Academic Press, U.S.A. (2000)
13. The Universal Virus Database ICTVDB, http://www.ncbi.nlm.nih.gov/ICTVdb/
14. GHSOM Toolbox, http://www.ofai.at/~elias.pampalk/ghsom/index.html
15. Kohonen, T.: Self-Organizing Maps. Springer-Verlag, Berlin (2001)
16. Dittenbach, M., Merkl, D., Rauber, A.: The Growing Hierarchical Self-Organizing Map. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN 2000) Vol. 6, pp. 15-19. IEEE Computer Society Press, U.S.A. (2000)
17. Duda, R. O., Hart, P. E., Stork, D. G. : Pattern Classification. John Wiley & Sons, New York (2001)
18. CONASIDA (ed.): El Médico Frente al SIDA. Pangea Editores, México (1989)