Analysis of Dimer Similarity Features in Viral Genomes with Growing Hierarchical Self Organized Maps

E. Bautista-Thompson
DES-DACI
Universidad Autónoma del Carmen
ebautista@pampano.unacar.mx

B. Tass-Herrera
DES-DACI
Universidad Autónoma del Carmen
btass@pampano.unacar.mx

Abstract

An analysis of similarity between viral genomes at the dimer level was developed by means of a Growing Hierarchical Self Organized Map (GHSOM), a GHSOM network allows the clustering of multidimensional data structures by projection of such structures in a two dimensional space this preserves the spatial structure of similarity between the genomes based on their dimer frequency. The clusters generated with GHSOM show a general correspondence with the standard taxonomical classification that groups the different virus by families based on information about morphological features of the three-dimensional viral structures, but also the analysis allows the identification of specific similarities and differences at the dimer level between some families of virus, then mining the genomic databases at the dimer level (one-dimensional information structures) allows the extraction of useful information at two levels: genomic sequence and morphology (3D-structure).

1. Introduction

Nucleotide sequences studies of DNA are of interest because the insight that they can provide about the evolutionary processes of species [1]. In particular, the study of dimer (dinucleotide) 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 taxonomical 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 gene expression studies, DNA clone classification, analysis of coding and non coding regions in DNA [5], focused on DNA from bacteria, plants and animals, but few studies were found about analysis and search of similarity patterns on viral genomes [5, 6, 7]. Unfortunately, our knowledge about the relations between the information codified inside the genomic sequences, the molecular machinery and the way this machinery controls in detail the processes of the dynamics of the virus at the cellular level is still limited.

In particular, the search for DNA-dimer frequency similarities between a set of different kinds of viruses is of our interest. We believe that identifying basic genomic similarities between viruses and connecting this information with other sources of information (taxonomical, gene functionality, etc.) can give us new knowledge about how these similarities at the level of genomic information are related with the viral structure and its dynamics at the molecular and cellular level. This paper is one of a series of works we are developing in order to gain insight about the aforementioned ideas. In order to search and identified similarity patterns between the viral genomes at the dimer level, we applied a clustering technique: Growing Hierarchical SOM (GHSOM), so we search for information at two levels. First at the level of DNA-dimer frequency similarity patterns between families of virus and second at the level of how the dimers features contributes to the similarity or dissimilarity between the different families of virus.

In section 2, we present the taxonomical information about the collection of viruses under study. In section 3, we describe the experimental methodology and results for the search of similarity patterns. Finally, in section 4 we present the conclusions of this work.

2. Collection of Viruses and their Taxonomical Features

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 [8]. There are a well known taxonomical classification of viruses based on the type of organization of viral genome, the strategy of viral replication and the structure of the virion [8, 9], but the explosion of taxonomical information available in public data bases thanks to the application of Information Technologies and the sequencing of virus genomes [9, 10], has complicated the analysis of the information and the discovery of new knowledge inside these data bases.

Table 1. Families of virus analyzed and some representative examples of them

Family	Virus
Adenoviridae	Human Adenovirus B
	Human Adenovirus F
Bunyaviridae	Uukuniemi Virus
Caliciviridae	Hepatitis E Virus
Coronaviridae	Human Coronavirus HKU1
Filoviridae	Sudan Ebolavirus
	Zaire Ebolavirus
Flaviviridae	Dengue Virus Type 1
	Yellow Fever Virus
Hepadnaviridae	Hepatitis B Virus
Herpesviridae	Human Herpesvirus 1
Paramyxoviridae	Measles Virus
	Mumps Virus
Papovaviridae	Human Papillomavirus 1
Parvoviridae	Human Parvovirus 4
Picornaviridae	Encephalomyocarditis Virus
	Human Enterovirus B
Poxviridae	Variola Virus
Retroviridae	Human Immunodeficiency Virus 1
	Human Immunodeficiency Virus 2
Togaviridae	Rubella Virus

The set of virus under study are representative of different taxonomical families and sources of different common and non common human and non human diseases [8, 10], see Table 1. We select 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 longitude 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 [9].

3. Methodology for Analysis of Dimer Similarity in Viral Genomes

In this section we describe the computational techniques applied in the analysis of the genomic data and the results of such analysis.

3.1. Frequency of Dimers in Genomic Sequences

In order to calculate the dimer frequency pattern of a genome, we need to count the number of times that each dimer appears in the genome. The total number of dimer combinations based on the four bases that form the genomic code are: AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GT, TA, TC, TG, and TT. To be able to compare among different genomes it is necessary to convert the frequencies to percentages, since the genomes are of different longitude. The first step in our analyses of similarity search is to build 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 intensity table is then used as input for the GHSOM technique, the Table 2, shows examples of the intensity data for some viruses.

3.2. Growing Hierarchical SOM

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 [11]. The GHSOM is a variant from the SOM neural network where a hierarchy of multiple layers of SOM neural networks are generated, 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 [12]. The features of GHSOM mentioned above were the reason for its choice as an analytical tool for the present study.

Table 2. Examples of dimer frequency for some of the viral genomes in percentage values

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 information about the frequency of occurrence (in percentage values) of the dimers for each virus genome is used as input for the GHSOM technique, instead of associate the specific name of the virus with its corresponding set of dimer frequency data, we associate such data with the virus taxonomical 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 and the associated virus family.

3.3. Analysis of the GHSOM Map

In the GHSOM map (see Figure 1), in general the similarity clusters to which the viruses belongs are in correspondence with the associated taxonomical families (grouping of the different virus by its corresponding family), 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, 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 [8, 13]. Then, we observed that at the dimer level new similarities between members of different families can be identified with this analysis. 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
Adenoviridae	1
Hepadnaviridae	2
Herpesviridae	3
Papovaviridae	4
Poxviridae	5
Bunyaviridae	6
Filoviridae	7
Paramyxoviridae	8
Rhabdoviridae	9
Caliciviridae	10
Coronaviridae	11
Flaviviridae	12
Picornaviridae	13
Retroviridae	14
Togaviridae	15
Parvoviridae	16

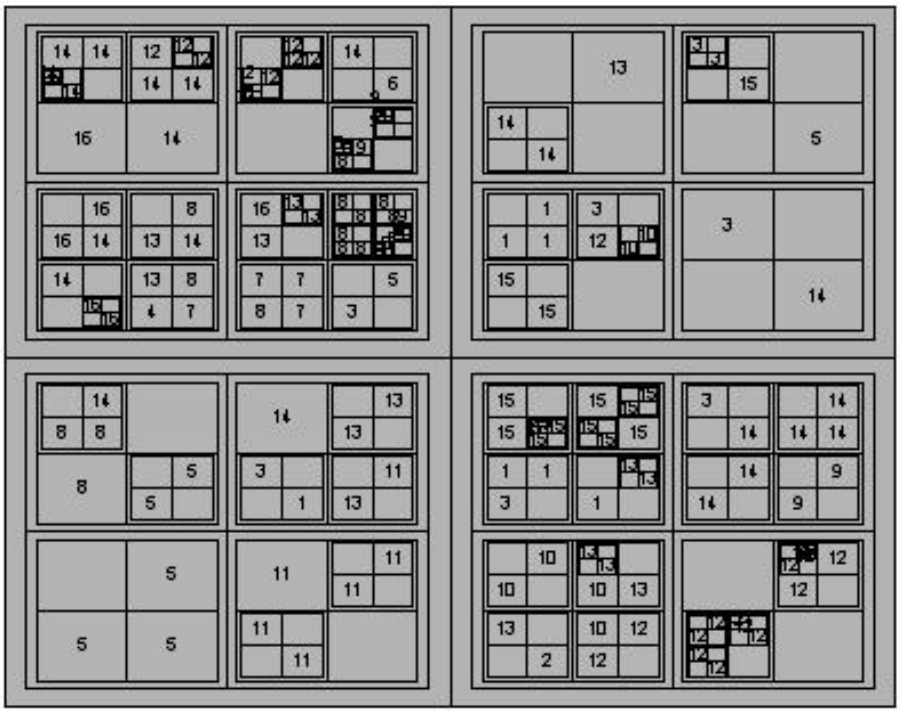

Figure 1. Hierarchical Map showing the similarity at the taxonomical families level between different viral genomes, where such similarity is based on the dimer frequency for each genome

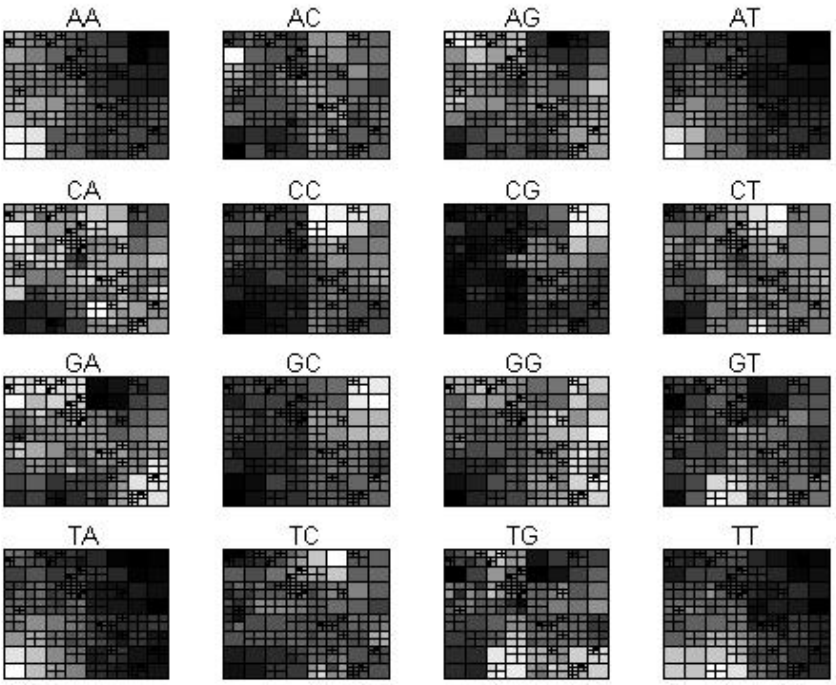

Figure 2. Maps based on dimer weight, each map shows the contribution of dimer frequency to the similarity of the corresponding viral genomes associated to each of the regions inside the GHSOM map, the white color corresponds to a strong contribution of the dimer in a specific region and the black color corresponds to a weak contribution of the dimer in a specific region

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).

The Figure 2, 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.

4. Conclusions

An analysis with GHSOM maps was developed in order to identify at the genome sequence level hierarchies of similarity patterns for viruses from different taxonomical families. At the dimer level certain degree of correspondence with taxonomical families was conserved, but a sharp differentiation by families was not observed. An interesting case was the Retroviridae family, the members of this family showed a strong dispersion between different clusters, the immunodeficiency viruses belongs to this family and they have a strong capability to mutate this implies a diversity in the dimer frequency 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 techniques such as the GHSOM maps, at the dimer level and its connection with biological knowledge at upper levels such as the taxonomical classification is useful for the understanding of relations between multiple levels, in particular: genomic (1D-structure) and morphological (3D-structure). In our case, more research is on the way in order to increment the number and detail of the biological data to be integrated with the genomes under study and the refinement of the methodology presented in this work.

5. References

[1] R. H. R. Stanley, N. V. Dokholyan, S. V. Buldyrev, S. Havlin, and H. E. Stanley, "Clustering of Identical Oligomers in Coding and Noncoding DNA Sequences", *Journal of Biomolecular Structure & Dynamics*, Vol. 17, Num. 1, 1999, pp. 79-87.

[2] K. J. Breslauer, R. Frank, H. Blöcker, and L. A. Marky, "Predicting DNA Duplex Stability from the Base Sequence", *Proc. Natl. Acad. Sci. USA*, Vol. 83, June 1986, 1986, pp. 3746-3750.

[3] P. Miramontes and G. Cocho, "DNA Dimer Correlations Reflect In Vivo Conditions and Discriminate among Nearest-Neighbor Base Pair Free Energy Parameter Measures", *Physica A*, Vol. 321, 2003, pp. 577-586.

[4] A. Quiroz-Gutierrez, "Biophysical Considerations and Evolutionary Aspects of DNA-dimer Frequency in AIDS Retrovirus Genomes", *Topics in Contemporary Physics*, IPN Press, México, 2000, pp. 239-248.

[5] M. C. Frith, M. C. Li, and Z. Weng, "Cluster-Buster: Finding Dense Clusters of Motifs in DNA Sequences", *Nucleid Acids Research*, Vol. 31, Num. 13, 2003, pp. 3666-3668.

[6] A. Figueroa, J. Borneman, and T. Jiang, "Clustering Binary Fingerprint Vectors with Missing Values for DNA Array Data Analysis", *Proceedings of the Computational Systems Bioinformatics (CSB'03)*, IEEE Computer Society Press, U.S.A, 2003.

[7] J. McCallum, and S. Ganesh, “Text Mining of DNA Sequence Homology Searches”, *Applied Bioinformatics*, 2003, pp. S59-S63.

[8] Brooks, G. F., Butel, J. S., Ornston, L. N., Jawitz, E., Melnick, J. L., and Adelberg, E. A., *Jawitz, Melnick, and Adelberg’s Medical Microbiology*, Prentice-Hall, U.S.A., 1991.

[9] C. Büchen-Osmond, “The Universal Virus Database ICTVDB”, *Computing in Science & Engineering*, May/June 2003, 2003, pp. 2-11.

[10] Van Regenmortel, M. H. V., et. al. (Eds.), *Virus Taxonomy. Classification and Nomenclature of Viruses. Seventh Report International Committee on Taxonomy*, Academic Press, U.S.A., 2000.

[11] Kohonen, T., *Self-Organizing Maps*, Springer-Verlag, Berlin, 2001.

[12] M. Dittenbach, D. Merkl, and A. Rauber, “The Growing Hierarchical Self-Organizing Map”, *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2000) Vol. 6.*, IEEE Computer Society Press, U.S.A., 2000, pp. 15-19.

[13] CONASIDA (Ed.), *El Médico Frente al SIDA*, Pangea Editores, México, 1989.