DINCON 2011
10ª Conferência Brasileira de Dinâmica, Controle e Aplicações
28 de agosto a 1º de setembro de 2011

PHYLODYNAMICS OF FOOT-AND-MOUTH DISEASE VIRUS: A COMPLEX NETWORK APPROACH

Luiz Max Fagundes de Carvalho[1,3], *Leonardo Bacelar Lima Santos*[2], *Pedro Jeovah Pereira*[3], *Waldemir de Castro Silveira*[3]

[1]Sector of Infectious Diseases Epidemiology – Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, luizepidemiologia@gmail.com
[2]National Institute os Space Research – INPE, São José dos Campos – SP , Brazil, santoslbl@gmail.com
[3]Laboratory of Bioresources – Pan American Foot-and-mouth Disease Center (PANAFTOSA) – Pan American Health Organization (PAHO), Duque de Caxias – RJ, Brazil, silveiraw@paho.org, ppereira@paho.org

Resumo: The foot-and-mouth disease is the most economically important disease of domestic livestock. In order to obtain insights about the risk of distinct antigenic variants arising, we apply an approach based on complex networks. The network's results show a modularity signature, with sequences of same year and region linking different modules.

Palavras-chave: foot-and-mouth disease, applications in biology, medicine e sciences of health, complex networks

1. INTRODUCTION

Foot-and-mouth disease virus (FMDV) is a highly infectious *aphthovirus*, *Picornaviridae*, that causes the most economically important disease of domestic livestock, Foot-and-Mouth Disease (FMD). The virus is present in seven distinct serotypes (O, A, C, SAT-1, SAT-2, SAT-3 and Asia-1) divided in multiple subtypes. In South America, only the circulation of serotypes O, A and C has been detected, and in this study we aim at the first. The virus of genomic RNA (aprox. 8200 bp) encodes 4 structural (VP1-4) proteins [1], among the VP1 is a surface exposed protein with 211 amino-acid residues (639 bp), present in the virus capsid in 60 copies, being the most antigenic and important for vaccine design. FMDV phylodynamics is marked by high rates of mutation (4×10^{-4} – 4×10^{-2} mutations per locus per replication), and the same bias can be extended to VP1 (10^{-2} – 10^{-3} changes per locus per replication) [2]. In this sense, diversity studies are required to assess the risk of loss in vaccine coverage due to mutations in antigenic regions present in novel viral strains.

The mathematical modeling [3, 4] and computational analysis of biological phenomena and data [5, 6] grows at an high speed. A key challenge of contemporary biology is to carry out an integrated theoretical and experimental program to analyze, in quantifiable terms, the topological and dynamic properties of diverse biological networks [7]. Complex networks have been proved useful for understanding FMD dynamics in the sense of migration patterns of livestock [8]. In this paper they are employed in a different way, providing insights about the complex properties of VP1 evolution in Ecuador.

The paper was organized as follows: after the **Introduction**, the section **Material and Methods** brings the datasets construction and pre-processing, and the complex networks indexes used on the paper. After that, the **Results and Discussion** summarizes the more relevant findings of the research, finishing with the investigation's **Conclusions and Perspectives**.

2. MATERIAL AND METHODS

In the next section is shown the databases and softwares applied on this research.

2.1. Data Preparation and Statistical analysis

We created a sequence database in order to comparatively study FMDV phylodynamics in Ecuador. The database was composed by 43 VP1-encoding nucleotide sequences from Ecuador in 2002-2010 (epidemic) period (data obtained from PANAFTOSA sequences bank) combined to 67 sequences from a world-wide at various years recovered from GenBank. Sequences were aligned with MEGA 5 software [9], the final alignment had 540 bp and was used to generate difference matrices. MEGA 5 output either for nucleotide (NT) and (translated) amino-acid (AA) databases were processed with R statistical computing environment [10] in order to create full symmetric distance matrices that were later used in complex network analysis. By this process, we generated 2 matrices of distances: AA and NT.

2.2. *Softwares for networks analysis*

Over all this paper is considered a non-weighted, undirected complex network R with N nodes and E edges. There are a lot of measurements for complex networks analysis, here three of them will be calculated: node degree $k(i)$, clustering coefficient $c(i)$ and the mean minimal distance $l(i,j)$, with i and $j \in N$. The degree k of a node counts the number of edges connected to it, while $< k >$ is the average number of edges per node over the network. The clustering coefficient c of node i, $c(i)$, is defined as the ratio between the number of edges among the immediate neighbors of i and $k(i)(k(i)-1)/2$, which is the maximum number of edges between the set of neighbors of i. The average of $c(i)$ over i leads to the network clustering coefficient $< c >$. The $l(i,j)$ index measures the number of edges between the nodes i and j, and your average over the network is $< l >$. The softwares employed were written in C++ language and used in UNIX Operational System (OS). For networks visualization was used the software PAJEK [11] in Windows OS.

2.3. *Networks generation and analysis*

After the step of preparing the NT and AA distance's matrixes, one similarity matrix (S) is construct for each database as follows:

1. Get the greatest value of the distance $d(i,j)$ between the sequences i and j, with i and $j \in N : max$;
2. Define the similarity between the sequences i and j, $s(i,j)$, as: $s(i,j) = 100 - \frac{d(i,j)}{max}$

The networks's nodes correspond to the protein sequences, and the presence of edges between two nodes depends on how similar the related sequences are. The same rule of connection was applied in [12, 13] using the software BLAST for the similarity evaluation, on the context of phylogenetic analysis. Each network can be defined by its adjacency matrix (M), for which any matrix element i,j is set to 1, if the nodes i and j are connected, or to 0, if not. The network's edges depends of a threshold value s_{min}, where the elements of its adjacency matrix M is set to 1, if $s(i,j) \geq s_{min}$, or to 0, if not. Were constructed for each database, AA and NT, 101 networks, each of them for one s_{min} value, from 0 to 100. For each network was evaluated: $< k > /N$, $< c >$ and $< l >$.

3. RESULTS AND DISCUSSION

The Exploratory Statistical Analysis (ESA) of these datasets is summarized in table 1. In order to assess differences in VP1 conservation – i.e., the relative proportion of differences between protein and RNA sequences – chi-square test was performed.

As expected, ESA showed that AA matrices are less variable than NT ones. Chi-square test results showed no differences in the degree of protein conservation in the isolates from Ecuador when compared to world-wide isolates. The FMDV high variability can account for this finding in the sense that the virus genome evolves in error catastrophe threshold, being shaped mostly by negative selection [14].

Table 1 – Exploratory Statistical Analysis under the distance datasets

Parameter/Subset	subset NT	subset AA
Identitity (%)	79.4	90.6
Mean	101.16	33.25
SD	12.39	3.14
Min	82.16	27.43
Max	134.63	46.27

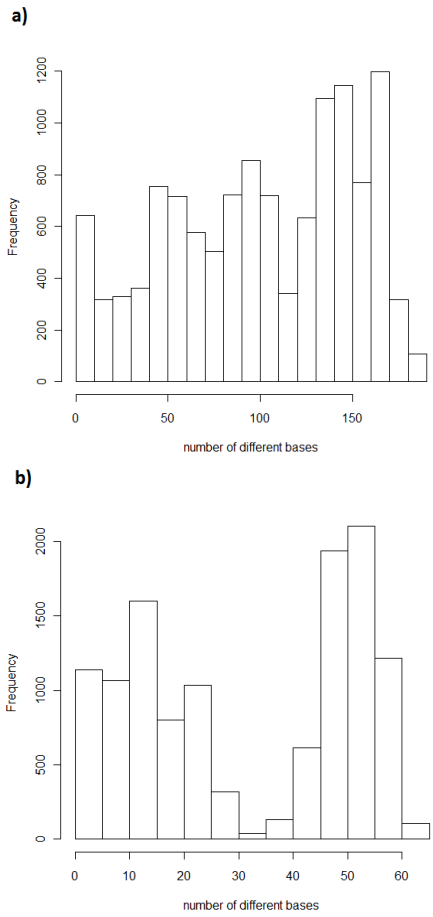

Figure 1 – Histograms of sequence's distances values – $d(i,j)$ – for (a) NT and (b) AA subsets.

The histograms's shapes represented in figure 1 suggest that VP1 is subjected to a non-random pattern of selection since AA subset presents a marked bi-modality. The finding of more than one population can be explained by FMDV's high rates of mutation and by vicariance (local) factors that can create conditions for different evolutionary dynamics to take place. As shown in figure 2, sequences from 1 to 43 have a markedly lower mean distance when compared to other sequences in the database, suggesting that the structure of viral population is somewhat peculiar in Ecuador since 2002. Further investigation concerning changes in livestock trade pattern – which seems to be determinant for viral evolution and spread – is needed in order to better understand these findings.

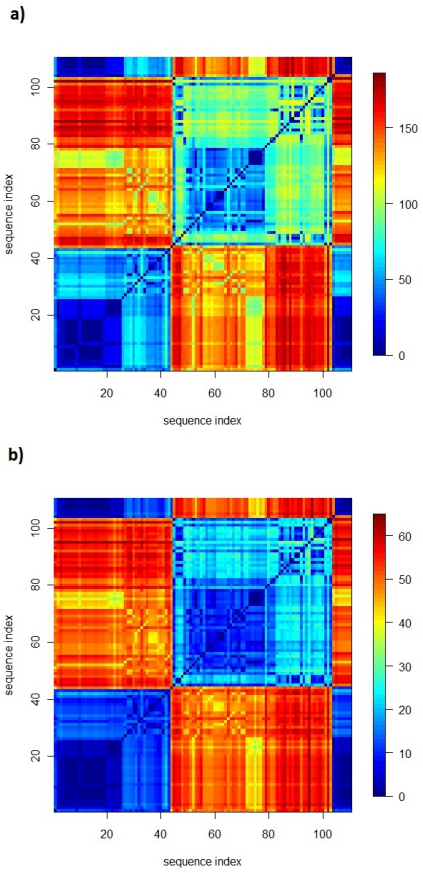

Figure 2 – Color plots for sequence's distances values – $d(i,j)$ – for (a) NT and (b) AA subsets.

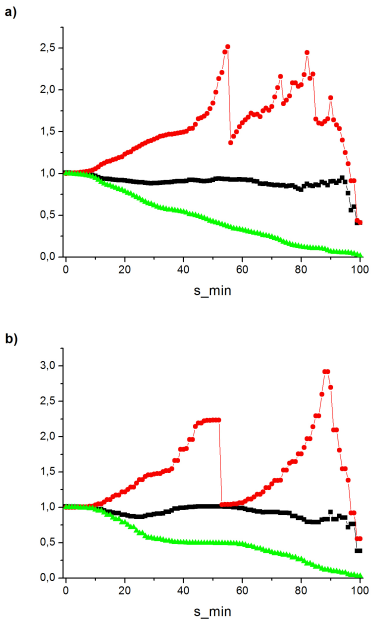

Figure 3 – Dependence of the network's measures against the s_{min} values for (a) NT and (b) AA networks.

After ESA, networks were generated for each database. The dependence of the network's measures against the s_{min} values is shown in the figure 3, for both NT and AA networks.

For both family of networks, AA and NT, for s_{min}=0 until s_{min}=100, we can define three regions in your graphics:

- First – Until s_{min}=55 for NT (s_{min}=52 for AA) as greater is s_{min}, the l index is greater, due some edges disconnecting, and some bypasses loosing. Beside that, the $< k > /N$ is smaller and <c> is practically constant. These three behaviors together can address a modularity signature: node's subsets with a high number of edges only between its own nodes.

- Second – To s_{min}=55 to s_{min}=56 for NT (s_{min}=52 to s_{min}=53 for AA) there is an rupture on the network: a great decrease of <c>, due the breakout of the edge whose linked the sequences 23 and 75 (figures 4a – NT, and 4c – AA). To s_{min}=56 until s_{min}=84 for NT (s_{min}=53 until s_{min}=90 for AA) the "noise" on network – the weak edges inside each module – is filtered. At the s_{min}=84 for NT (s_{min}=90 for AA) we have the more significant network: the network with a maximum level of information and the minimum of "noise" (figures 4b – NT, and 4d – AA). This greater value ($90 > 84$) can be explained by the fact the the AA network shows a smaller heterogeneity then the presented by the NT network – been necessary a greater value of s_{min} to "clean" the network.

- Third – After s_{min}=84 for NT (s_{min}=90 for AA) we have more and more isolated nodes, and at s_{min}=100 the networks are composed only by nodes whose genetic distance is null.

Networks topological results are consistent with the findings described above and suggest that FMDV evolution is dependent on spatio-temporal proximity, since the two edges (23 and 75) that connect the great modules are for sequences from the same year (2009) and region (Highland).This can be explained by the fact that the major constraints for FMDV population sizes and genetic diversity are the number and size of the herds affected.

4. CONCLUSION AND PERSPECTIVES

In this paper, a complex network approach was used for obtaining insights about underlying patterns in the phylodynamics of VP1 protein of FMDV. The network's results shown a modularity signature, at three phases of behavior limited for some specific similarity thresholds. The nodes 23 and 75, corresponding to sequences of same year (2009) and region (Highland), linked different modules until a critical value of similarity: a spatial–temporal result. Further investigations will be needed in order to assess the risk of antigenic coverage loss, but this framework has been demonstrated to be very powerful in making clear many patterns in the data.

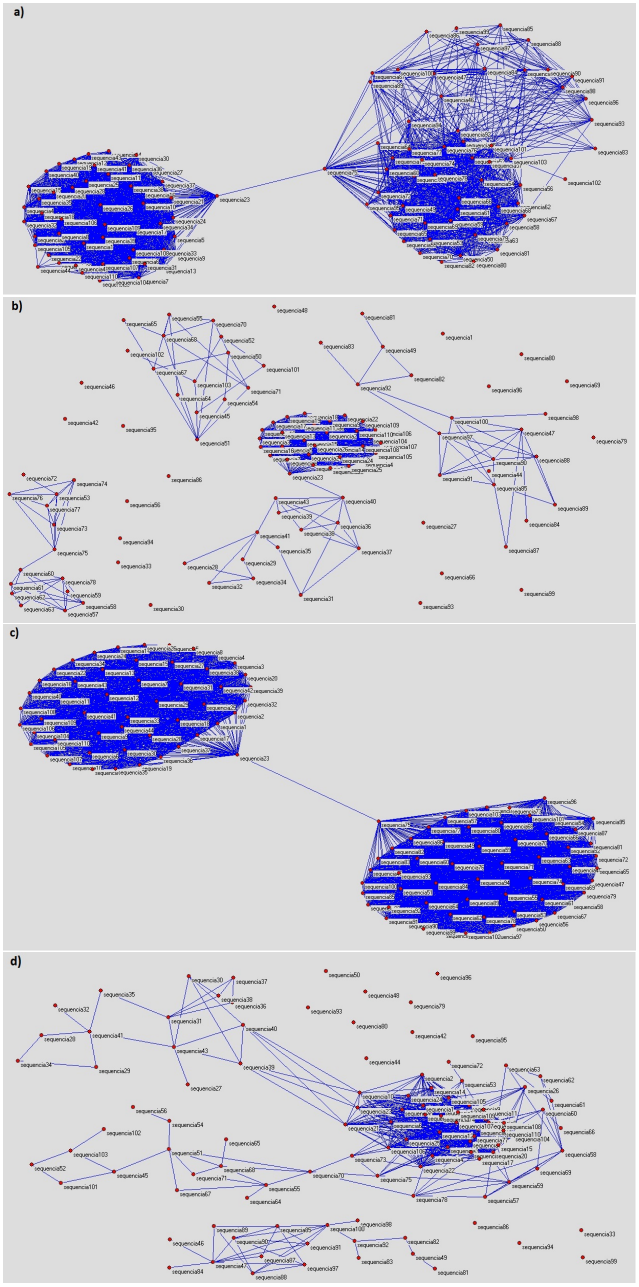

Figure 4 – Networks at some s_{min} values for conection: a) NT, s_{min}=55, b) AA, s_{min}=52, c) NT, s_{min}=84, d) AA, s_{min}=90.

5. ACKNOWLEDGMENTS

The authors thank Jonas da Silva, Dr. José Naranjo, Antônio Mendes and Dr. Júlio Pompei for their insightful contributions and administrative support, and are grateful to Alex Santana, Charles Santana and José Miranda for their contributions to the softwares for complex networks analyzes.

REFERENCES

[1] V. Malirat et al. Phylogenetic analysis of foot-and-mouth disease virus type O re-emerging in free areas of South America. Virus Res., 124, 1-2:22-8, 2007.

[2] A. M. Perez et al. Variation in the VP1 gene of foot-and-mouth disease virus serotype A associated with epidemiological characteristics of outbreaks in the 2001 epizootic in Argentina. J. Vet. Diagn. Invest., 20, 4:433-9, 2008.

[3] E. A. Reis, L. B. L. Santos, S. T. R. Pinho. A cellular automata model for avascular solid tumor growth under the effect of therapy. Physica A: Statistical Mechanics and its Applications, v. 388, n. 7, p. 1303-1314, 2009.

[4] L. B. L. Santos et al. Periodic forcing in a three-level cellular automata model for a vector-transmitted disease. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics (Print), v. 80, p. 016102, 2009.

[5] A. C. Gavin et al. Proteome survey reveals modularity of the yeast cell machinery. Nature 440, 631636, 2004.

[6] C. P. Pirovani et al. Knowledge discovery in genome database: the chitin metabolic pathway in Crinipellis perniciosa. In: Proceedings of IV Brazilian Symposium on Mathematical and Computational Biology/I InternationalSymposium on Mathematical and Computational Biology, vol. 1, E-Papers Servic os Editoriais LTDA, Rio de Janeiro, pp. 122139, 2005.

[7] A. L. Barabasi, Z. N. Oltvai. Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101113, 2004.

[8] C. Dub, C. Ribble, D. Kelton, B. McNab. A review of network analysis terminology and its application to foot-and-mouth disease modelling and policy development. Transbound Emerg Dis, 56, 3:73-85, 2009.

[9] K. Tamura et al. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol, 2011.

[10] R Development Core Team: R: A Language and Environment for Statistical Computing Vienna, Austria: R Foundation for Statistical Computing, 2010.

[11] V. Batagelj, A. Mrvar. Pajekanalysis and visualization of large networks. In: Jnger, M., Mutzel, P. (Eds.), Graph Drawing Software. Springer, Berlin, pp. 77103, 2003.

[12] Góes Neto, A; et al. (2010). Comparative protein analysis of the chitin metabolic pathway in extant organisms: A complex network approach. BioSystems, v. 101, p. 5966.

[13] R. F. S. Andrade et al. Detecting Network Communities: An Application to Phylogenetic Analysis. PLoS Computational Biology, v. 7, p. e1001131, 2011.

[14] S. N. Balinda et al. Diversity and transboundary mobility of serotype O foot-and-mouth disease virus in East Africa: implications for vaccination policies. Infect. Genet. Evol., 10, 7:1058-65, 2010.