Gene 528 (2013) 277–281

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Triplet entropy analysis of hemagglutinin and neuraminidase sequences measures influenza virus phylodynamics

Günther J.L. Gerhardt [a,*], Agnes A.S. Takeda [d], Tahila Andrighetti [d], Ivaine T.S. Sartor [b], Sergio L. Echeverrigaray [b], Scheila de Avila e Silva [b], Laurita dos Santos [c], José L. Rybarczyk-Filho [d]

[a] *Departamento de Física e Química da Universidade de Caxias do Sul, Rua Francisco Getulio Vargas 1130, 95001-970 Caxias do Sul, RS, Brazil*
[b] *Instituto de Biotecnologia da Universidade de Caxias do Sul, Rua Francisco Getulio Vargas 1130, 95001-970 Caxias do Sul, RS, Brazil*
[c] *LAC, Instituto Nacional de Pesquisas Espaciais, 12227-010 Sao Jose dos Campos, SP, Brazil*
[d] *Department of Physics and Biophysics, Institute of Biosciences, Univ Estadual Paulista "Júlio de Mesquita Filho", Distrito de Rubião Junior S/N, 18618-970 Botucatu, SP, Brazil*

ARTICLE INFO

Article history:
Accepted 17 June 2013
Available online 11 July 2013

Keywords:
Sequence analysis
H1N1
Entropy
Influenza

ABSTRACT

The influenza virus has been a challenge to science due to its ability to withstand new environmental conditions. Taking into account the development of virus sequence databases, computational approaches can be helpful to understand virus behavior over time. Furthermore, they can suggest new directions to deal with influenza. This work presents triplet entropy analysis as a potential phylodynamic tool to quantify nucleotide organization of viral sequences. The application of this measure to segments of hemagglutinin (HA) and neuraminidase (NA) of H1N1 and H3N2 virus subtypes has shown some variability effects along timeline, inferring about virus evolution. Sequences were divided by year and compared for virus subtype (H1N1 and H3N2). The nonparametric Mann–Whitney test was used for comparison between groups. Results show that differentiation in entropy precedes differentiation in GC content for both groups. Considering the HA fragment, both triplet entropy as well as GC concentration show intersection in 2009, year of the recent pandemic. Some conclusions about possible flu evolutionary lines were drawn.

1. Introduction

Throughout the last century the flu was (and indeed still remains) one of the greatest challenges in modern medicine. Together with the recent return of H1N1 to worldwide media attention (during the 2009 pandemic), a simple question also returned: How is humanity prepared to face this challenge: namely the erradication or living with the flu? According to the (World Health Organization) the flu is responsible for an average of a fourth to half a million deaths per year worldwide, and countless loss in productivity since it can derail each worker for several weeks. However, a novelty that came up during the last outbreak was the possibility of organizing virus sequence databases in order to draw new computational approaches shedding some light on the organization and mutation of these viruses (Greenbaum et al., 2008; Kobayashi and Suzuki, 2012; Pompei et al., 2012; Santos et al., 2011; Trifonov, 1999). Such approaches have shown promising results in the study of sequences which are small in size but large in quantity (Benson et al., 2011). The sequences carry with them some phylogenetic information, the biggest challenge is how to decode it (Frenkel and Trifonov, 2008; Trifonov, 1999) and many efforts have been done to link influenza phylogenetics and sequence analysis (Xu et al., 2012).

H1N1 and H3N2 are both subtypes of influenza that infect humans and animals. H1N1 is known as the swine flu pandemic and by the famous 2009 outbreak. The H3N2 flu became famous as the "Hong Kong flu" and has an endemic cycle in birds and pigs. With each new outbreak, new viral mutations appear resulting from selection pressures that create new resilient virus subtypes (Garten et al., 2009; Girard et al., 2010; Watanabe et al., 2010; Webster et al., 1992). Understanding the influenza virus' physical and bio-compositional structure is essential to study its spread. The Influenza genome comprises eight segments of viral RNA and two of them encode surface glycoproteins that are essential in the infection process: Hemagglutinin (HA) and Neuraminidase (NA) (Fig. 1). The influenza nomenclature itself takes its abbreviation letters from these two proteins (H and N). The infection is initiated by HA, present on the outer surface of the virus, that recognizes the sialic acid of the host cell membrane, promoting a receptor mediated endocytosis (Weis et al., 1988; Wilson et al., 1981). On the other hand NA is important in the end of the infection process, when it removes the sialic acids from infected cell membrane surfaces and new-formed virions, allowing the virus progeny to be released from the cell and promoting infection of other host cells (Xu et al., 2008). Summarizing: HA is responsible for binding the virus to the host and NA is responsible for rapid contamination and spread.

Abbreviations: AT, Adenine and Thymine; GC, Guanine and Cytosine fraction; CpG, C-phosphate-G-; HA, Hemagglutinin; NA, Neuraminidase; pH, Power of hydrogen; RNA, Ribonucleic Acid.

[*] Corresponding author. Tel.: +55 32312137.
E-mail address: gunther_lew@yahoo.com.br (G.J.L. Gerhardt).

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.06.060

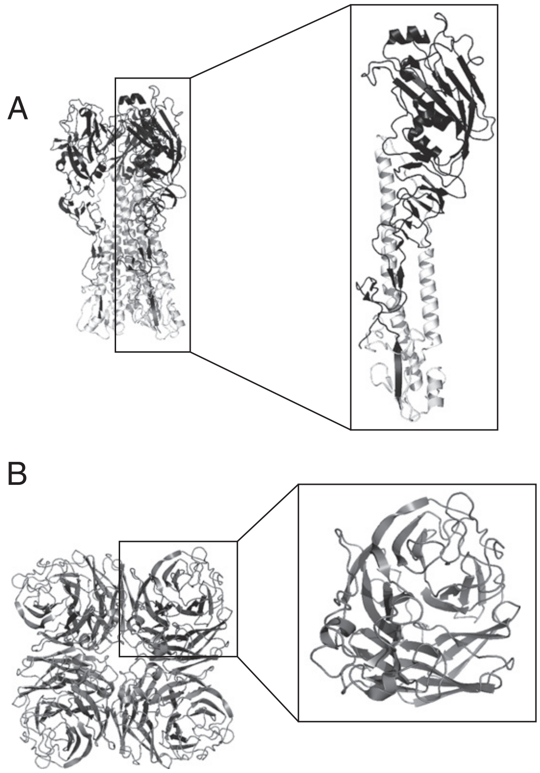

Fig. 1. Crystallographic structures of the surface glycoproteins hemagglutinin (HA) and neuraminidase (NA) shown in ribbon representation. (A) Trimeric biological assembly of HA (PDB ID: 1RUZ (Gamblin et al., 2004)) featuring a monomer. The HA1 subunit (black) corresponds to the first interaction between virus and host cell surface; and HA2 subunit (light gray) acts to fuse the viral envelope to the cellular membrane of the host cell, allowing the infection. (B) Tetrameric biological assembly of NA (PDB ID: 1NN2 (Varghese and Colman, 1991)) featuring one monomer (gray). NA ensures the virus release by the cleavage of the ends of polysaccharide chains. Illustrations generated by the program PyMOL v. 1.0. (Schrödinger, 2010).

The main goal of this paper is to use H1N1 and H3N2 virus data available at *Influenza Virus Resource* (Bao et al., 2008) to track how both main segments (HA and NA) are evolving over time in each virus subtype. Although evolutionary flu lines intersect, it is unclear how the organization of sequences correlates with that evolution. The sequence organization carries together evolutionary information at molecular level. If organization measurements coincide, this could represent a common evolutionary path taken by both virus subtypes.

Aiming to follow the sequence organization over time, we will use GC content and triplet entropy as organization measures (Gerhardt et al., 2006; Santos et al., 2011; Trifonov, 1999). The GC content gives a hint of how the global genomic composition is changing and triplet entropy can inform about codon organization.

2. Methods

The methodology used here has been described previously in detail and only the basic outlines are presented here (Gerhardt et al., 2006; Santos et al., 2011). The main goal of this approach is to measure triplet prevalence entropy, described below, and it is based on the concept of amino acid organization and frequency prevalence (Poland, 2005; Trifonov, 1999).

2.1. Sequences

In this work we analyze the HA and NA segments of H1N1 and H3N2 obtained from Influenza Virus Resource (Bao et al., 2008). Table 1 shows the number of segments used for each year. Complete NA segments were used for this study. The HA segments used in this work have 800–1200 bp long, which comprise the region that codifies to HA1 subunit (Fig. 1A) (Weis et al., 1988). The emphasis in this region was due to the localization of the sialic acid binding site, responsible for the first interaction between virus and host cell surface. Moreover, these HA segments contain antigenic sites recognized by the host immune system that can accumulate amino acid substitutions (Skehel and Wiley, 2000; Sriwilaijaroen and Suzuki, 2012).

Table 1
Number of segments downloaded from (Bao et al., 2008) and used in this work.

Year	HA/H1N1	NA/H1N1	HA/H3N2	NA/H3N2
1995	10	29	93	70
1996	18	31	118	69
1997	12	9	100	83
1998	14	4	112	122
1999	27	10	123	190
2000	67	82	92	234
2001	88	132	44	88
2002	53	50	345	280
2003	46	26	595	417
2004	15	13	610	418
2005	75	79	687	376
2006	199	220	730	217
2007	279	559	524	312
2008	470	698	823	418
2009	558	5193	228	460
2010	63	468	36	154
2011	221	268	75	91

2.2. Triplet entropy

Triplet prevalence entropy is defined following the classic Shannon information theory definition as

$$S = -\sum_{n=1}^{64} P_n log P_n, \tag{1}$$

where P_n is the probability of n^{th} triplet in a given sequence of size W. In our case W is the size of sequence. If window size W is not a multiple of three, we use a periodic contour condition.

Eq. (1) in this form is naturally biased by the GC content itself. So we can define a dispersion coefficient for entropy as

$$D_n = S - S_{GC}(rand), \tag{2}$$

where $S_{GC}(rand)$ is the entropy calculated for a random sequence with same GC content.

This procedure assures that (2) will measure information over triplet organization uncorrelated with GC content (Poland, 2005; Rodríguez-Trelles et al., 2000) and can be considered as a measure that correlates itself with complex AT and GC skewness present in sequence avoiding to take into account the aminoacid natural degeneracy. As D_n becomes closer to zero, the sequence of triplet nucleotides can be considered more random. It is easier to correlate and understand (1) when comparing with GC content variations.

2.3. Statistical analysis

Sequences were divided by year and GC and D_n distributions were compared for each virus type (H1N1 and H3N2). Once the distributions are non-normal but unimodal, the non-parametric Mann–Whitney test was used for comparison between the different groups. Categorical data were analyzed in frequency tables that fulfilled the guidelines for “a large sample” approximation and χ^2 test was used to test the null hypothesis. Since sequence number exponentially increases by year, we used a maximum of 100 randomly chosen sequences from each year in order to compare data among groups. Differences between the