Towards novel theranostics for Ebola virus: exploring genomic relative absent words

Raquel M. Silva[1,2], Diogo Pratas[3], Luísa Castro[1], Armando J. Pinho[3], Paulo J. S. G. Ferreira[3]

Ebola virus (EBOV) causes high mortality hemorrhagic fevers, for which no vaccine or treatment currently exist. In the latest and largest ever EBOV outbreak, over 28000 cases and 11000 deaths from the virus occurred mainly in West Africa. Although experimental therapies are being tested, namely, recombinant viral vectors or antibodies for the viral glycoprotein, innovative approaches are still needed for the development of diagnosis tools and identification of druggable targets.

Minimal absent words are the shortest sequence fragments that are not present in the genome of a given organism. In this study, published in Bioinformatics (doi: 10.1093/bioinformatics/btv189), we introduce minimal relative absent words (RAWs), a concept that has not been used so far in the context of personalized medicine, but which is deemed useful for differential identification of sequences that are derived from a pathogen genome but absent from its host. To identify RAWs we have developed the EAGLE tool (freely available from http://bioinformatics.ua.pt/software/eagle/).

Analysis of 165 EBOV genomes allowed the discovery of RAWs that are present in these viral sequences but absent from the human genome. Only three RAWs of length 12 exist and consistently appear within conserved regions of two EBOV proteins, the nucleoprotein (NP) and the viral RNA-polymerase (LP). Moreover, these words can discriminate between the different Ebolavirus species and even between EBOV sequences from different outbreaks (Figure 1). Both NP and LP are critical for the virus replication and constitute good targets for therapeutic intervention.

The alignment-free method used is able to identify species-specific sequences that are important for future therapeutics and diagnosis, or theranostics, strategies for EBOV. It can also be applied to other pathogens of biomedical or economical relevance to detect genomic signatures for quick and precise action against infectious agents, namely in outbreak scenarios.

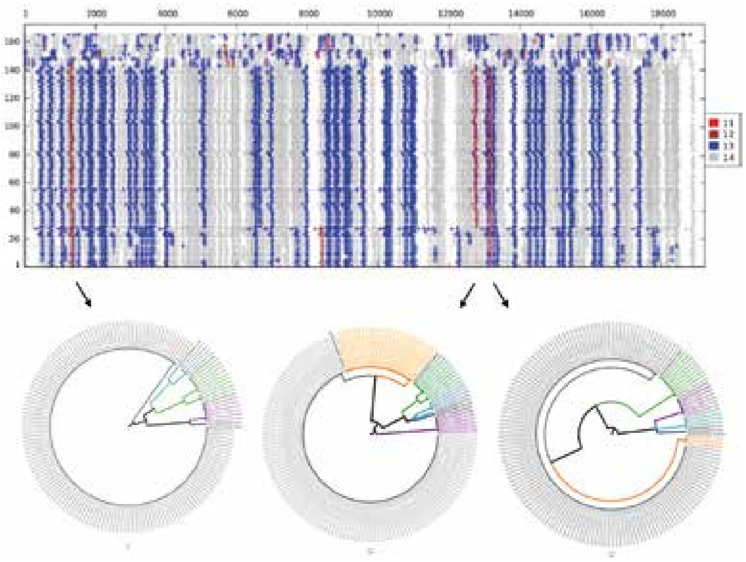

FIGURE 1
Top panel, identification of relative absent words (RAWs) in 165 Ebolavirus genomes with the human genome as reference. RAW sequences are shown in red (k= 11), dark red (k= 12), blue (k = 13) and grey (k = 14). Sequences 1-24, Zaire ebolavirus (EBOV) genomes from previous outbreaks; 25-28, EBOV genomes from the 2014 DRC (Democratic Republic of the Congo) unrelated outbreak; 29-142, EBOV genomes from the West African 2014 outbreak; 143-147, Bundibugyo ebolavirus (BDBV) genomes; 148-154, Reston ebolavirus (RESTV) genomes; 155-164, Sudan ebolavirus (SUDV) genomes; and 165, Tai Forest ebolavirus (TAFV) genome. Bottom panel, phylogeny of 165 Ebolavirus based on RAW1 (left), RAW2 (middle) and RAW3 (right) sequences. EBOV, SUDV, BDBV, TAFV, and RESTV are shown in grey, green, blue, dark blue and purple, respectively. EBOV sequences that diverge from the West African 2014 outbreak are shown in orange (28 and 4 genomes for RAW2 and RAW3, respectively). Reference genomes are displayed in black.

[1] — IEETA, University of Aveiro
[2] — Department of Medical Sciences & iBiMED, University of Aveiro
[3] — Department of Electronics, Telecommunications and Informatics & IEETA, University of Aveiro