Genome analysis

Three minimal sequences found in Ebola virus genomes and absent from human DNA

Raquel M. Silva1,*,†, Diogo Pratas1,2,†, Luisa Castro1, Armando J. Pinho1,2 and Paulo J. S. G. Ferreira1,2

1IEETA and 2DETI, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

Motivation: Ebola virus causes high mortality hemorrhagic fevers, with more than 25 000 cases and 10 000 deaths in the current outbreak. Only experimental therapies are available, thus, novel diagnosis tools and druggable targets are needed.

Results: Analysis of Ebola virus genomes from the current outbreak reveals the presence of short DNA sequences that appear nowhere in the human genome. We identify the shortest such sequences with lengths between 12 and 14. Only three absent sequences of length 12 exist and they consistently appear at the same location on two of the Ebola virus proteins, in all Ebola virus genomes, but nowhere in the human genome. The alignment-free method used is able to identify pathogen-specific signatures for quick and precise action against infectious agents, of which the current Ebola virus outbreak provides a compelling example.

Availability and Implementation: EAGLE is freely available for non-commercial purposes at http://bioinformatics.ua.pt/software/eagle.

Contact: raquelsilva@ua.pt; pratas@ua.pt

Supplementary Information: Supplementary data are available at Bioinformatics online.

1 Introduction

Ebola virus (EBOV) is a negative strand-RNA virus from the Filoviridae family that causes high mortality hemorrhagic fevers, for which no vaccine or treatment currently exist (Sarwar et al., 2014). There are five Ebolavirus species, namely, Zaire ebolavirus, Sudan ebolavirus, Bundibugyo ebolavirus, Tai Forest ebolavirus and Reston ebolavirus, with the first (1976) and major (2014) outbreaks caused by the type species Zaire ebolavirus (Baise et al., 2014).

The numbers of the largest ever EBOV outbreak are worrying and continue escalating, with over 25 000 cases and 10 000 deaths from the virus mainly in Guinea, Liberia and Sierra Leone, according to the World Health Organization. The current outbreak is also the first where transmission has occurred outside Africa, with reported cases in Europe (Spain) and America (USA; Butler and Morello, 2014). Promising vaccine candidate tests are being rushed to face the epidemics and could be available within a few months (Gulland, 2014). These yet experimental therapies include, for example, recombinant viral vectors (Jones et al., 2005) or antibodies that target the viral glycoprotein (GP; Friedrich et al., 2012; Sarwar et al., 2014), but 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 genomic data of a given organism. They have been studied before to describe properties of prokaryotic and eukaryotic genomes and to develop methods for phylogeny construction or PCR primer design (Chairungsee and Crochemore, 2012; Falda et al., 2014; Garcia et al., 2011; Herold et al., 2008; Pinho et al., 2009; Wu et al., 2010). Here, we introduce minimal relative absent words (RAW), a concept which 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.
We use the current EBOV outbreak sequences, which were recently published (Gire et al., 2014), to discover and characterize the minimal RAWs that are present in EBOV genomes but absent from the human genome. Moreover, we show that these words are also absent from the other Ebolavirus species and even from the genomes obtained from previous outbreaks. Thus, the sequences that we identify are species-specific and important for future development of diagnosis or therapeutic strategies for EBOV. The method that we introduce can be applied to other emerging pathogens or to show evidence of evolutionary patterns and signatures across species.

2 Methods

2.1 Relative absent words

Consider a target sequence (e.g. a virus sequence), \( x \), and a reference sequence (e.g. the human genome), \( y \), both drawn from the finite alphabet \( \sum = \{A, C, G, T\} \). We say that \( x \) is a factor of \( y \) if \( x \) can be expressed as \( x = xyu \), with \( uv \) denoting the concatenation between sequences \( u \) and \( v \).

We denote by \( W_k(x) \) the set of all \( k \)-size words (or factors) of \( x \). Also, we represent the set of all \( k \)-size words \( \not\in x \) as \( W_k(x) \). For each word size \( k \), we define the set of all words that exist in \( x \) but do not exist in \( y \) by

\[
R_k(x, y) = W_k(x) \cap W_k(y)
\]

and the subset of words that are minimal, in the sense presented in Pinho et al. (2009), as

\[
M_k(x, y) = \{ x \in R_k(x, y) : W_{k-1}(x) \cap M_{k-1}(x, y) = \emptyset \}
\]

i.e. a minimal absent word of size \( k \) cannot contain any minimal absent word of size less than \( k \). In particular, \( l,r \) is a minimal absent word of sequence \( x \), where \( l \) and \( r \) are single letters from \( \sum \), if \( lr \) is not a word of \( x \) but both \( ls \) and \( ar \) are (Pinho et al., 2009). In this work, we were particularly interested in the non-empty set \( M_k(x, y) \) corresponding to the smallest \( k \). These are referred as RAWs.

2.2 Protein structural models

Protein 3D structural models were built by homology modeling as previously described (Duarte-Pereira et al., 2014). Appropriate templates were selected from PDB (www.rcsb.org; Berman et al., 2000), where several nucleoprotein (NP) structures from viruses within Mononegavirales (negative-sense genome single-stranded RNA viruses) are available, whereas for the region of interest in \( \iota \)-protein only structures from more distant viruses exist.

Structures from the Nipah virus NP (PDB ID:4CO6; Yabukarski et al., 2014) and the BVDV (bovine viral diarrhea virus) RNA polymerase (PDB ID:1S48; Choi et al., 2004) were used as templates in MODELLER (Eswar et al., 2006; Sali and Blundell, 1993), to predict the structure of the N-terminal regions of Ebola virus NP (residues 1–380) and RNA-polymerase (residues 177–805), respectively (Supplementary Figs. S3 and S4). Accuracy of the predicted models (Supplementary Fig. S5) was estimated using ProSA-web (https:// prosa.services.came.sbg.ac.at/prosa.php; Sippl, 1993; Wiederstein and Sippl, 2007) and structures were visualized with PyMOL (Schrodinger, 2010).

3 Results

To identify RAWs, we have developed the EAGLE tool that implements the method described above (Supplementary Data). We have used the full GRC-38 human reference genome (Church et al., 2011) downloaded from the NCBI, including the mitochondrial, unplaced and unlocalized sequences. The sequences of 99 EBOV genomes from the current outbreak in Sierra Leone (Gire et al., 2014) and additional 66 Ebolavirus genomes have been also downloaded from NCBI (Supplementary Table S1). The code used in this analysis is available (Pratas, 2015).

Figure 1 shows the computation for word sizes 12, 13 and 14 (for computer characteristics see Supplementary Section Software and Hardware). As expected, the number of absent words decreases as the \( k \)-mer size decreases. Specifically, for \( k = 11 \) (not represented), there are no EBOV RAW. On the other hand, for \( k = 12 \), three groups of points emerge (RAW1, RAW2 and RAW3) representing the position of a RAW in each of the 99 unaligned viral genomes (Fig. 1a).

Alignments of 124 Ebolavirus sequences (Gire et al., 2014), including additional EBOV genomes from the current outbreak in Guinea (Baize et al., 2014) and from previous outbreaks, show that the identified minimal RAWs fall into conserved protein regions (Fig. 1b). However, several mutations can be found in the genomes that discriminate between the different species of Ebolavirus and even between EBOV sequences from the current and previous outbreaks (Fig. 1c). The identification of these viral genome signatures is important for quick diagnosis in outbreak scenarios. Additional analysis with all 165 Ebolavirus genomes confirmed these results (Supplementary Fig. S1). In particular, RAW1 is conserved within EBOV and can distinguish EBOV from other Ebolavirus species. RAW2 is conserved in all sequences from the West African 2014 outbreak in Guinea, Sierra Leone and Liberia, and only one nucleotide difference exists between these sequences and unrelated outbreak genomes. RAW3 is also conserved at the species level, excluding the four EBOV 1976/777 genomes, and can distinguish between all Ebolavirus species (Supplementary Fig. S2).

From the three EBOV sequence motifs absent in the human genome, the first (RAW1) is included in the virus NP, while the other two (RAW2 and RAW3) fall within the sequence of the viral RNA-polymerase (\( \iota \)-protein; Fig. 1c). Previous studies show that the N-terminal region of EBOV NP participates in both the formation of nucleocapsid-like structures through NP–NP interactions and in the replication of the viral genome (Watanabe et al., 2006), and RAW1 sequence (TTTCGCCCGACT) is part of this N-terminal region. The \( \iota \)-protein (LP) produces the viral transcripts to be translated by host ribosomes and is involved in the replication of the viral genome as well. The LP contains the two additional minimal RAWs, RAW2 (TACCCCTATCG) and RAW3 (CCTACCGCGAAA).

Both NP and LP are critical for the virus life cycle and constitute good targets for therapeutic intervention. Screening for new antiviral compounds could benefit from knowledge of their protein structures. For EBOV, most protein structures are unknown except for the C-terminal domain of NP, GP, VP24 and VP35 (Shurtleff et al., 2012), thus, we have predicted the structure of the N-terminal regions of the EBOV NP and LP by homology modeling (Supplementary Figs. S3–S5). These structural models show that the amino acids corresponding to the RAW1 motif are enclosed within the structure, while RAW2 and RAW3 are exposed at the protein surface, which can justify its higher degree of conservation.

4 Discussion

The personalized medicine field is now closer to clinical practice with the advances of next-generation sequencing technologies. Personalized therapeutics are a possibility and their development is essential with the emergence of resistance to current available drugs. Additionally, quick diagnosis is required for emerging pathogens and in epidemics
such as the current Ebola outbreak. Here, we have detected minimal RAWs in the human genome that are present in EBOV genomes, and identified nucleotide differences in some of these sequences that can distinguish between \textit{Ebolavirus} species and outbreaks. Also, we show that the corresponding amino acid sequences are conserved within EBOV. These results can now be further explored for diagnosis and therapeutics, sometimes mentioned as theranostics (Picard and Bergeron, 2002). Namely, RAW nucleotide sequences can be used in diagnosis to design primers that identify \textit{Ebolavirus} infections or distinguish between \textit{Ebolavirus} species. For PCR-based methods, longer sequences and multiplex reactions can be developed to avoid primer binding bias. Additional nucleotide or protein-based strategies for therapeutics can be envisaged, as discussed below.

One problem in developing efficient EBOV treatments is the virus ability to evade the immune system. The viral GP is a major target because it mediates attachment and entry into the host cells. However, in addition to the surface envelope protein, the GP gene also produces fragment, soluble GPs that are secreted and direct the
immune system to produce antibodies for variable and non-essential regions of the virus (Cook and Lee, 2013; Mohan et al., 2012). As current efforts based on the viral GP might prove ineffective, additional targets should be sought. Our results show that the viral NP and polymerase (LP) can be attractive targets. As the amino acid sequences of all three 12-mer RAWs are conserved within EBOV, these regions can be used to screen for small molecule inhibitors. In particular, RAW1 is conserved in all Ebola virus NP proteins, which can indicate a functional or structural role. And, considering that the protein model predicts that RAW2 and RAW3 are relatively close in the 3D structure and in exposed domains, these regions can be used to develop novel antibodies. Also, a recently described mechanism shows that the polymerase (LP) from Ebola and Marburg viruses is capable of editing transcripts, resulting in increased variability in the produced proteins, and that the most edited mRNAs are the Ebola GP and Marburg NP and LP itself (Shabman et al., 2014). Thus, the use of combined therapies towards multiple proteins can be more effective, as suggested by studies to develop vaccines for Lassa virus that target both NP and GP (Fisher-Hoch et al., 2000; Lukashевич, 2012).

RNA-based strategies such as RNA interference (RNAi) or antisense therapies are also promising approaches to silence target-specific gene expression. The RAW sequences that we have identified can be used to develop RNAi or antisense probes that bind viral transcripts and prevent their translation, thus, inhibiting viral replication without blocking the host mRNAs. Translation of these technologies into clinical applications have been slowed by challenges in the delivery of small RNAs into cells, but recent developments in delivery systems are bridging the bench to bedside gap (Hayden, 2014; Yin et al., 2014). Among these, gold or lipid nanoparticles (Conde et al., 2014; Draz et al., 2014) were shown to be effective against cancer and viral infections, including EBOV (Geisbert et al., 2010). Gold-nanobeacons can be applied as a combined diagnosis and therapy tool for effective testing, including in low-cost settings (Costa et al., 2014) and, with this purpose, advances in peptide nucleic acid probes for viral detection are also taking place (Joshi et al., 2013; Zhang et al., 2010).

Whichever the technology, the identification of genome signatures for rapid evolving species such as Ebola viruses will be useful for the development of both diagnosis and therapeutics.

Funding
This work was supported by the European Fund for Regional Development (FEDER) through the Operational Program Competitiveness Factors (COMPETE) and by the Portuguese Foundation for Science and Technology (FCT), in the context of projects Pest-OE/EE/UI0127/2014 and Incentivo/EE/UI0127/2014, by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 305444 ‘RD-Connect: An integrated platform connecting registries, biobanks and clinical bioinformatics for rare disease research’, and the project Neuropath (CENTRO-07-ST24-FEDER-002034), co-funded by QREN “Mais Centro” program and the EU.

Conflict of Interest: none declared.

References
Three minimal sequences found in Ebola virus genomes


