Towards personalized medicine: ebola virus absent words in the human genome

Raquel M. Silva
raquelsilva@ua.pt
Luisa Castro
luisa.castro@ua.pt
Diogo Pratas
pratas@ua.pt
Armando J. Pinho
ap@ua.pt

Signal Processing Lab, IEETA / DETI,
University of Aveiro,
3810–193 Aveiro, Portugal

Abstract

Next generation sequencing technologies are driving a novel revolution in the personalized medicine field. The unprecedented availability of genomes and computational tools enables the development of novel diagnosis and therapeutic strategies. Here, using an alignment-free method based on the relative minimal absent words, we show that the identification of pathogen signatures is possible to allow quick intervention for infectious agents, as exemplified by the current Ebola virus outbreak genome analysis.

Introduction

The $1200 genome milestone has been reached, with sequencing times of one week, and it has been argued that these values will substantially decrease with the incoming of the third generation. The personalized medicine field is now newborn, increasingly dependent on the advances of sequencing technologies, and yielding major contributions for diagnosis or genetic counseling to maximize the probability of welfare. Under this framework, personalized vaccines are a possibility and their development is essential with the emergence of pathogen resistance, namely for fungal, viral and bacterial infections.

New successful applications strongly depend on the ability to detect regions in pathogen genomes that are absent in the healthy host, mainly to eliminate the pathogen without harming the host. Given the task, genomic information sizes create a challenge to the wet lab, and currently this can only be solved with comparative genomics that can be time-consuming, expensive and more susceptible to error than computational approaches.

Although computational methods can be employed to solve these tasks, the success of the detection algorithms are strongly dependent on the time and memory requirements to run on common computers. Several studies indicate that minimal absent words on the human genome exist above size 10. Therefore, efficient data structures, such as suffix arrays and hash-tables, among others, play a key role to minimize memory requirements maintaining affordable processing times.

On the other hand, identifying comparative specific novel regions is also a way to discover new genes and genome structures, and evidence of evolutionary patterns and signatures across species. For example, one of the current research topics is the identification of modern human specific genes, compared with the newly sequenced genomes of Denisovans and Neanderthals. Typically, these genomes have approximately 3 giga bases, establishing the importance of efficient fast-compact data structures.

Minimal absent words have been studied and computed by many researchers [1, 3, 5, 7]. In this paper, we follow an emerging branch, the relative minimal absent words. These are minimal absent words of a certain genome that exist in another one. We propose a method to detect these words and report their position. We have applied the method to 20 Ebola virus (EBOV) genomes and detected several minimal absent words of the human reference genome (GRC) that occur in the virus, as well as their positions.

Method

Consider a reference sequence, X, and n target sequences $Y_1, Y_2, \ldots, Y_n$. All sequences are from a finite alphabet, $\Sigma = \{A, C, G, T\}$, and $|X|$ denotes the size of sequence X.

We compute the k-mers of X and the Y_i sequences using a sliding window of size k. Each k-mer is converted into a numeric index, i, and stored into a binary array if $k < 17$, otherwise in a hash table. Parallel, we perform the following mapping: $A \rightarrow T$, $T \rightarrow A$, $C \rightarrow G$ and $G \rightarrow C$. The mapping is applied for each reversed k-mer, converted into an index, i and stored as described above. Each k-mer from X is loaded, including those from the reverse mapping, and stored in memory. We call this the training phase.

Then we start the matching phase. The intention is to find exact k-mers on each Y_i. Therefore, for each Y_i, a boolean array is created, B_i, with $|B_i| = |Y_i| - k$, containing a true value when a k-mer exists in the memory.

The objective is to detect relative absent words, therefore the interest is on the false elements from B_i. Since the process of matching is sequential, each position of a false element in B_i reports the exact position in the target sequence, Y_i.

Finally, the results are presented in a map, along each Y_i, that depicts the regions or points where the k-mer (absent in X) occurs.

Results

For the results in this section we have used the full GRC-38 human reference genome [2] downloaded from the NCBI[1], including the mitochondrial, unplaced and unlocalized sequences. For the 20 EBOV genomes, the sequences (first 20 genomes) have been also downloaded from NCBI[2] [4].

We have implemented a version of the method that required 7 minutes and 51 seconds to compute all the k-mers presented in Fig. 1, from the 20 EBOV genomes, including the training phase over a sequence of approximately 3 GB. The maximum RAM memory used was 1 GB. The computation was performed on Linux Ubuntu 12.04 LTS with the following hardware characteristics: 4 Intel Core i7-3520M CPU at 2.90GHz, 8 GB of RAM and a SSD of 243.6 GB.

Figure 1 depicts the computation for word sizes 12, 13, 14 and 15. As expected, the number of absent words decreases as the k-mer size decreases. Specifically, for $k = 11$ (not in Fig. 1), there are no relative absent words. On the other hand, for $k = 12$, three groups of points emerge (P1, P2 and P3). Each position of the corresponding point, according to each genome, is shown in Table 1. In each group, the positions are very close among the different strains of the virus.

In fact, they all degenerate from the same words, namely:

- TTTCGCCCGACT (P1),
- TACGCCCTATCG (P2),
- CCTACGCGCAAA (P3).

From the three EBOV sequence motifs absent in the human genome, the first (P1) is included in the virus nucleoprotein (NP), while the other two (P2 and P3) fall within the sequence of the viral RNA-polymerase (L protein). 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 [8], and P1 sequence is part of this N-terminal region (Fig. 2). The L-protein produces the viral transcripts to be translated by host ribosomes and is involved in the replication of the viral genome as well. Both proteins are critical for the virus life cycle, thus, constitute good targets for therapeutic intervention. The identification of these viral genome signatures is also important for quick diagnosis in outbreak scenarios.

[1] ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/Assembled_chromosomes/seq/

[2] http://www.ncbi.nlm.nih.gov/bioproject/257197

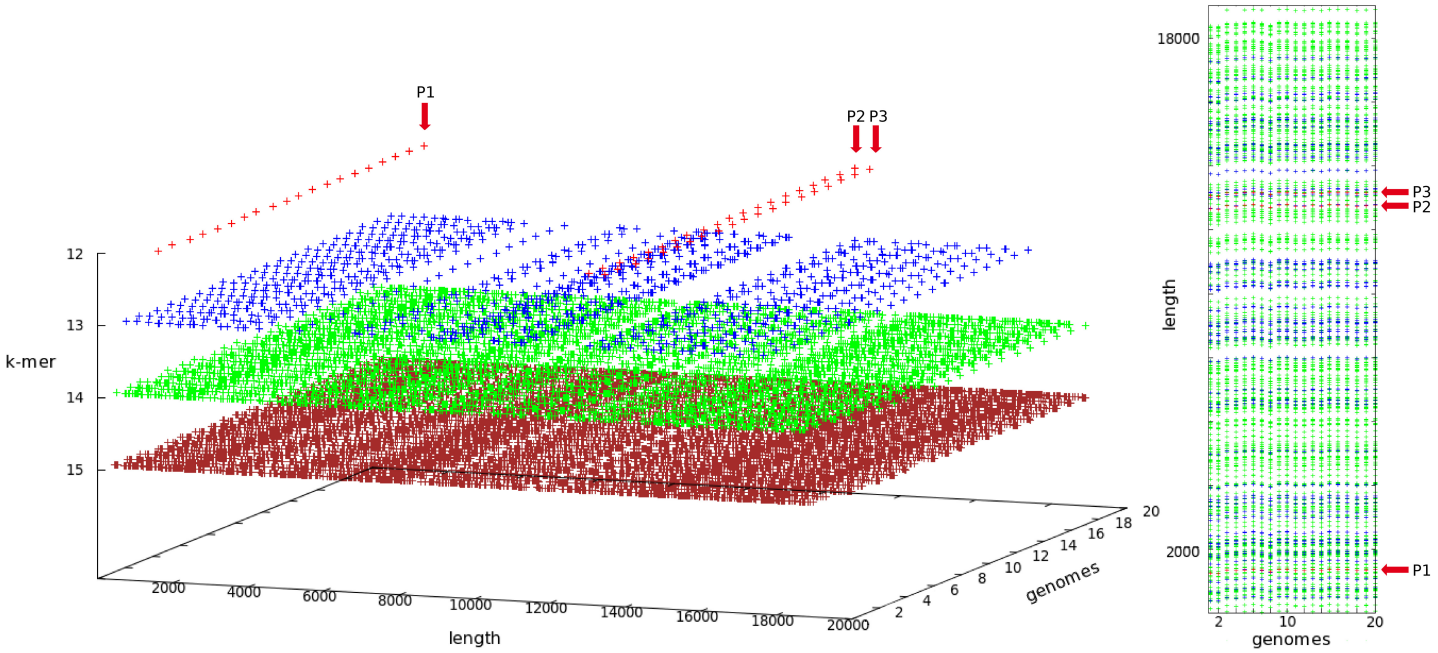

Figure 1: Ebola virus absent words relatively to the GRC reference human genome. Left plot depicts a 3D model with several k-mers for the 20 EBOV genomes along their length, while the right plot contains a (vertical) projection of the genomes and length. Points $P1$, $P2$ and $P3$ represent the relative minimal absent words.

Figure 2: Structure of the Ebola virus genome. The negative-stranded RNA genome has about 19 kb in size and encodes for seven proteins: nucleoprotein (N), glycoprotein (GP), polymerase (L) and four additional viral proteins (VP).

Genome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
P1	1,232	1,216	1,323	1,308	1,336	1,353	1,318	1,260	1,351	1,353	1,309	1,308	1,340	1,309	1,353	1,345	1,340	1,308	1,329	1,347
P2	12,654	12,638	12,745	12,730	12,758	12,775	12,740	12,682	12,773	12,775	12,731	12,730	12,762	12,731	12,775	12,767	12,762	12,730	12,751	12,769
P3	13,056	13,040	13,147	13,132	13,160	13,177	13,142	13,084	13,175	13,177	13,133	13,132	13,164	13,133	13,177	13,169	13,164	13,132	13,153	13,171

Table 1: Starting positions for each absent word, for $k = 12$, contained in each EBOV genome relatively to the GRC reference human genome.

Conclusions

The field of personalized diagnosis and terapeutics is largely led by the virology branch. The identification of the regions that are present in a virus genome but are absent in a human genome, will drive the development of innovative terapeutics.

It is well-known that two different human genomes share a high degree of homology, and we explore this characteristic to detect minimal absent words in the human genome that are present in the Ebola virus. For each of the 20 EBOV genomes, we have detected 3 words with size 12, namely, *TTTCGCCCGACT*, *TACGCCCTATCG* and *CCTACGCGCAAA*, using the method that we propose here.

These results can now be further explored from a biological point of view, in order to build a vaccine that has a high probability to damage the virus without harming the human being.

Acknowledgements

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/EEI/UI0127/2014 and Incentivo/EEI/-UI0127/2014. RMS and LC are supported by the project Neuropath (ref: CENTRO-07-ST24-FEDER-002034), co-funded by QREN "Mais Centro" program and the EU. DP is supported by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 305444. Info:"RD-Connect: An integrated platform connecting registries, biobanks and clinical bioinformatics for rare disease research".

References

[1] Supaporn Chairungsee and Maxime Crochemore. Using minimal absent words to build phylogeny. *Theoretical Computer Science*, 450: 109–116, 2012.

[2] D. Church, M. Deanna, V. Schneider, et al. Modernizing reference genome assemblies. *PLoS Biology*, 9(7):e1001091, 2011.

[3] S. P. Garcia, A. J. Pinho, J. M. O. S. Rodrigues, C. A. C. Bastos, and P. J. S. G. Ferreira. Minimal absent words in prokaryotic and eukaryotic genomes. *PLoS ONE*, 6(1):e16065, January 2011. doi: doi:10.1371/journal.pone.0016065.

[4] Stephen K Gire, Augustine Goba, Kristian G Andersen, Rachel SG Sealfon, Daniel J Park, Lansana Kanneh, Simbirie Jalloh, Mambu Momoh, Mohamed Fullah, Gytis Dudas, et al. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science*, page 1259657, 2014.

[5] J. Herold, S. Kurtz, and R. Giegerich. Efficient computation of absent words in genomic sequences. *BMC Bioinformatics*, 9(1):167, 2008. doi: 10.1186/1471-2105-9-167.

[6] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.

[7] A. J. Pinho, P. J. S. G. Ferreira, S. P. Garcia, and J. M. O. S. Rodrigues. On finding minimal absent words. *BMC Bioinformatics*, 10 (137), May 2009. doi: 10.1186/1471-2105-10-137.

[8] Shinji Watanabe, Takeshi Noda, and Yoshihiro Kawaoka. Functional mapping of the nucleoprotein of ebola virus. *Journal of virology*, 80 (8):3743–3751, 2006.