Detection and visualisation of regions of human DNA not present in other primates

Diogo Pratas[1]
pratas@ua.pt
Raquel M. Silva[2]
raquelsilva@ua.pt
Armando J. Pinho[1]
ap@ua.pt
Paulo J. S. G. Ferreira[1]
pjf@ua.pt

[1] Information Systems and Processing Group, IEETA/DETI,
University of Aveiro,
3810–193 Aveiro, Portugal

[2] IEETA,
University of Aveiro,
3810–193 Aveiro, Portugal

Abstract

Human specific regions are DNA segments that are unique or share high dissimilarity rates relatively to close species, namely primates. Their existence is important to localize evolutionary traits that are often related to novel functionality, besides its obvious discriminative ability.

We propose an unsupervised method, and an associated tool, to detect and visualise these regions. It is based on the detection of relative absent words (RAWs), using a probabilistic high depth model. The experimental results show several regions that are associated with documented human specific regions, namely centromeres and several genes, such as those related with olfact. However, it also shows several undocumented ones, that may express trends in human evolution.

1 Introduction

Relative absent words (RAWs) are sub-sequences that do not occur in a given sequence (the reference), but do occur in another sequence (the target). Consider a target sequence, x, and a reference sequence, y, both from a finite alphabet Θ. We say that β is a factor of x if x can be expressed as $x = u\beta v$, with uv denoting the concatenation between u and v. We denote by $\mathcal{W}_k(x)$ 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 $\overline{\mathcal{W}_k(x)}$. For each k-size word, we denote the set of all words that exist in x but do not exist in y by

$$\mathcal{R}_k(x,\overline{y}) = \mathcal{W}_k(x) \cap \overline{\mathcal{W}_k(y)} \tag{1}$$

and the subset of words that are minimal as

$$\mathcal{M}_k(x,\overline{y}) = \{\beta \in \mathcal{R}_k(x,\overline{y}) : \mathcal{W}_{k-1}(\beta) \cap \mathcal{M}_{k-1}(x,\overline{y}) = \emptyset\}, \tag{2}$$

i.e., a minimal absent word of size k cannot contain any minimal absent word of size less than k. In particular, $l\beta r$ is a minimal absent word of x, where l and r are single letters from Θ, if $l\beta r$ is not a word of x, but both $l\beta$ and βr are ([6]).

Although minimal absent words have been studied before to describe properties of prokaryotic and eukaryotic genomes and to develop methods for phylogeny construction or PCR primer design [3, 9], their pratical usage for differential analysis is relatively new. Recently, we have proposed this approach, exploring the non-empty set $\mathcal{M}_k(x,\overline{y})$ corresponding to the smallest k, referred to as minimal relative absent words (mRAWs), in an application related to the ebola virus [7].

In this paper, we focus on finding large RAWs, with the aim of detecting human regions that are unique (with high probability), relatively to several primates. Hence, we are interested in creating a model of one or more reference sequences, to detect sub-sequences that are present in a target. To achieve this goal, we use a k-mer model with high depth (tipically $k = 30$), that is efficiently implemented using Bloom filters. Thereafter, the unique regions, that are filtered and segmented using a threshold value, are presented in a map. A tool, with the implementation of the unsupervised method, is freely available.

2 Method

If one uses a binary vector to store all the possible entries indicating if a certain k-mer exists or not in the sequence, we would use 4^k bits. For $k = 30$, we would need 131,072 TeraBytes of memory, which is impracticable on current computers. A data structure such as a hash table for implementing such a model would certainly be more reasonable, but the memory becomes dependent on the number of inserted elements. Moreover, for the volumes of data that we usually need to deal with, it still implies high memory requirements.

A third option is a probabilistic data structure, namely a Bloom filter [1], which trades space resources by precision. Notwithstanding, the usage of a very large Bloom filter (with the number of hash functions optimized), can give very high probabilities of becoming very similar to deterministic. Because, for this case, we do not need very large lengths and precise results, since we want to find regions (RAWs) and not mRAWs, this seems the most efficient choice.

For using a Bloom filter, we set a vector of dimension m and the number of hash functions h, obtaining a balance that is also related with the number of elements that are filtered, n. Asymptotically, for a given m and n, the value of the number of hash functions that minimizes the probability of false positives is given by

$$h = \frac{m}{n} \ln 2, \tag{3}$$

that can be re-written as

$$2^{-h} \approx 0.6185^{m/n}. \tag{4}$$

The more elements that are added to the set, the larger the probability of false positives. Given n and a desired false positive probability p (assuming that the optimal value of h is used), we can find the required number of bits m using

$$m = -\frac{n \ln p}{(\ln 2)^2}. \tag{5}$$

This means that, asymptotically, for a given false positive probability p, the length m of a Bloom filter is proportional to the number of elements being filtered, n. For finite values, the false positive probability for a finite Bloom filter with m bits, n elements and h hash functions is, at most (see [2] for more details),

$$\left(1 - e^{-h(n+0.5)/(m-1)}\right)^h. \tag{6}$$

This method allows whole genome analysis using T_n targets and R_n references. To solve this, we write to disk each RAW detected from R_i in relation to each T_i. Next, for each T_i, the RAWs are considered only if they exist in all R_i. A file containing the whole genome RAWs in relation to each T_i is stored (these are the unique regions).

An example of the method, from the sequences to the maps, using three reference sequences and one target is depicted in Fig. 1. For T_n targets, the process is repeated n times. Moreover, when using inverted repeats, the reverse complemented sequence is also loaded into memory (for the same reference model).

For visualising the unique regions, after low-pass filtering of a binary sequence containing the presence/absence of RAWs, a threshold is used to segment the regions and then they are presented in a visual map (see Fig. 1 for an example).

We have created a fully automatic tool (CHESTER), written in C language, with the implementation of the unsupervised method. It is available at `http://github.com/pratas/chester`, under GPL-2, and can be applied to any genomic sequence, in FASTA, FASTQ or SEQ (ACGTN) formats.

In Fig. 2, we show the results of running the tool against several synthetic sequences, used to better illustrate the method. As can be seen,

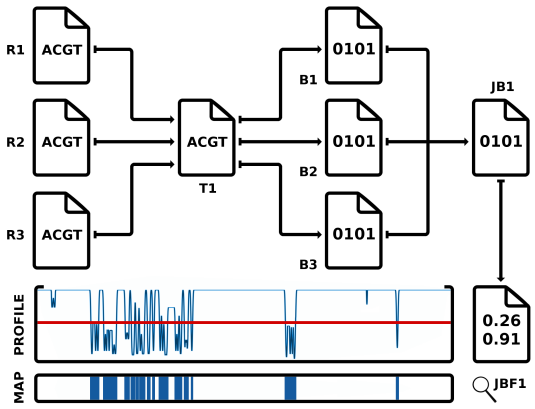

Figure 1: Visual description of the method. The genomic sequences contained in the files $R1$, $R2$ and $R3$ are independently processed against a target, $T1$. From each computation a binary sequence is generated, $B1$, $B2$ and $B3$, describing the presence/absense of a RAW according to the order of $T1$. Next, the binary files are merged using a logic or ($\vee$), $B1 \vee B2 \vee B3$ and the result is $JB1$. The $JB1$ sequence is then low-pass filtered, resulting in the real sequence described as $JBF1$. Finally, a threshold (line in red) is used to segment the information contained in the $JBF1$, where each segmented region is represented in the RAWs map.

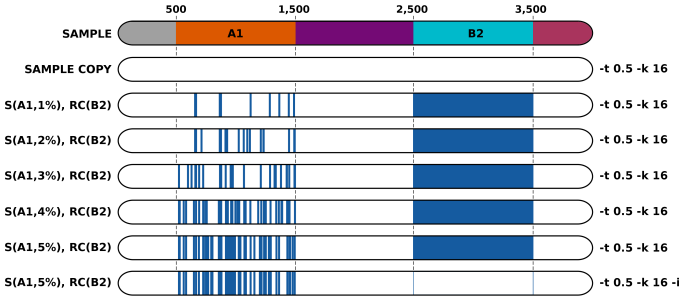

Figure 2: Running CHESTER using several synthetic sequences. Blocks $A1$ and $B2$ have been edited according to the functions refered on the left. Function S stands for a substitution mutation of the input block with the defined percentage. Function RC applies the reverse complement of the input. CHESTER running parameters are defined on the right, using a threshold of 0.5 and a k-mer size of 16, while only the bottom map has been run using inversions. The blue color on the computed maps represents the unique regions according to the RAWs.

the method identifies as novel the regions that are mutated ($A1$) and also the inversions ($B2$). When the tool runs with the "-i" parameter (handle inversions), these where successfuly not reported. When dealing with sequenced data, the sequences might have several inverted regions due to errors of assemblage or sequencing. As we have shown in Fig. 2, this method is prepared to overcome those limitations.

3 Experimental results

In the experiments with real sequences we used: reference human genome (GRC-38) [4]; reference chimpanzee genome (2.1.4); reference gorilla genome (3.1); reference orangutan genome (2.0.2). The sequences were downloaded from the NCBI. The Y chromosomes of gorilla and orangutan have not been yet sequenced and therefore they are not present. On the other hand, we have included the unlocalized, unplaced and mitochondrial sequences, in order to bypass most assembly challenges.

We ran CHESTER on those sequences, obtaining the map displayed in Fig. 3. The larger blue areas identify the centromeres, corresponding to very repetitive DNA. The smaller areas contain several genes and pseudogenes (genes that are not expressed [8]) associated, for instance, to immunology, blood, smell and brain. Besides, there are several identified motifs in these regions (on the NCBI and Ensemble) that, although considered of importance, their nature has not yet been understood.

Of those sub-sequences which are less understood, we highlight HCP5 HLA complex P5, MAFK, GALNT9, OR11H12, OR11H11, SHOX short stature homebox. For example, on human chromosome 14 the OR11H12

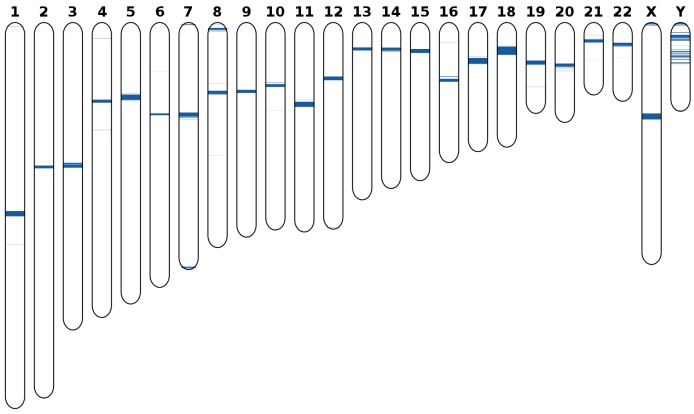

Figure 3: Human specific region (chromosomal) maps relative to the chimpanzee, gorilla and orangutan using CHESTER with $t = 0.6$ and $k = 30$. The blue strips represent the relative unique regions.

olfactory receptor is a gene associated with olfactory receptors that interact with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell. These findings are confirmed by other recent studies that show the loss of olfatory function only in the hominid evolution and therefore the consequent genomic sequence alteration [5].

4 Conclusions

RAWs are unique words that appear in a sequence and nowhere in other sequence. Their fundaments have been used recently in personalized medicine scenarios using minimal absent words. In this paper, we followed a different line, exploring high orders and whole genome analysis, proposing a method and an associated tool (CHESTER) to detect and visualise RAWs. These regions are associated with relative whole genome uniqueness, namely with centromeres and recent evolutionary traits that, relatively to several primates (chimpanzee, gorilla and orangutan), are specific to humans.

Acknowledgment

This work was partially funded 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 by National Funds through FCT - Foundation for Science and Technology, in the context of the project UID/CEC/00127/2013.

References

[1] B. H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM*, 13(7):422–426, July 1970.

[2] A. Broder and M. Mitzenmacher. Network applications of bloom filters: A survey. *Internet Math.*, 1(4):485–509, 2004.

[3] S. Chairungsee and M. Crochemore. Using minimal absent words to build phylogeny. *Theo. Comp. Sci*, 450:109–116, 2012.

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

[5] G. M. Hughes, E. C. Teeling, and D. G. Higgins. Loss of olfactory receptor function in hominin evolution. *PloS ONE*, 9(1):e84714, 2014.

[6] 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.

[7] R. M. Silva, D. Pratas, L. Castro, A. J. Pinho, and P. J. S. G. Ferreira. Three minimal sequences found in ebola virus genomes and absent from human DNA. *Bioinformatics*, page btv189, 2015.

[8] E. F. Vanin. Processed pseudogenes: characteristics and evolution. *Annual Rev. Gen.*, 19(1):253–272, 1985.

[9] Z. Wu, T. Jiang, and W. Su. Efficient computation of shortest absent words in a genomic sequence. *Inf. Proc. Let.*, 110(14):596–601, 2010.