Insights into primates genomic evolution using a compression distance

Diogo Pratas
pratas@ua.pt
Armando J. Pinho
ap@ua.pt

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

Abstract

With the recently emergence of the high-throughput sequencing technologies, full genomic DNA sequences have been released to the public, allowing for the first time large genomic computational studies between species. One of those studies is chromosomal distances between *Hominidae* primates in order to unveil insights of evolution.

The most successful and promising distance metrics seems to rely on the compressed based type. Therefore, we propose and use an admissible normalized compression distance, based on a specific conditional compression analogue which internally is composed by a mixture of static and dynamic finite-context models.

1 Introduction

The recent full genomic sequencing of primates (*Hominidae*) have brought new challenges, such as the identification of chromosomal distances in order to find insights of large genomic evolution. Although the existence of several distance metrics, such as Hamming and Levenshtein, the compressed-based distances seem to be the most successful ones.

A compression-based distance measure assesses the distance between two objects using the number of bits needed to describe one of them when a description of the other is available. The foundations of compression distances are built upon the Kolomogorov notion of complexity, also known as algorithmic entropy, where $K(x)$ of a string x is the length of the shortest binary program x^* that computes x in an appropriate universal Turing machine. As such, $K(x) = |x^*|$, the length of x^*, denotes the number of bits of information from which x can be computationally retrieved [6]. The conditional Kolomogorov complexity, $K(x|y)$, denotes the length of the shortest binary program, in the universal prefix Turing machine, that on input y outputs x. A special case occurs when y is an empty string, $y = \lambda$, and hence $K(x|\lambda) = K(x)$. Bennett introduced the information distance [1], $E(x,y) = \max\{K(x|y), K(y|x)\}$, defined as the length of the shortest binary program for the reference universal prefix Turing machine that with input x computes y, as well as with y computes x. The normalized version (NID [7]) of $E(x,y)$ is formally known as

$$\mathrm{NID}(x,y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}, \tag{1}$$

up to an additive logarithmic term. The normalized compression distance (NCD) [2] emerged to efficiently compute the NID, formalized as

$$\mathrm{NCD}(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}, \tag{2}$$

up to an additive logarithmic term, where $C(x)$ and $C(y)$ represent, respectively, the number of bits outputted in the compression of x and y, and $C(xy)$ the number of bits outputted by the compression of object x concatenated with y (conjoint compression analogue).

In this paper, we propose an admissible normalized compression distance, based on an analogue conditional compressor, that builds an internal model of the data using a mixture of static and dynamic finite-context models, in order to explore distances between the primates genomic sequences, namely, *H. sapiens*, *P. troglodytes*, *G. gorilla* and *P. abelii*.

2 Method

2.1 Admissible compression distance

Instead of using the conventional conjoint compression to deduce the mutual information content in (2), we have created a specific analogue of a conditional compressor, $C(x|y)$ (see next subsection), in order to compute (1). Therefore, the admissible NCD can be calculated using

$$\mathrm{NCD}(x,y) = \frac{\max\{C(x|y), C(y|x)\}}{\max\{C(x), C(y)\}}, \tag{3}$$

where, asymptotically, $C(x|\lambda) = C(x)$, $C(x|x) = 0$, $C(x|y) \leq C(x)$, $C(x|y) + C(y) = C(y|x) + C(x)$ and $C(x,y) + C(z) \geq C(x,z) + C(y,z)$, up to an additive logarithmic term.

2.2 Conditional compressor

We have conducted a NCD compressor based on a mixture of two classes (static and dynamic) with multiple finite-context models. Accordingly, the compression is performed in two phases. In the first phase, the static class of finite-context models, with variable orders, accumulate the counts regarding the y object. When the entire y object is processed, the models are kept frozen and, hence, the second phase starts. At this point, the x object starts to be compressed using the static models, from the first phase, in cooperation with the new multiple finite-context models, of variable orders, that dynamically accumulate the counts only from the x object.

The per symbol information content average provided by the finite-context model of order-k, after having processed n symbols, is given by

$$H_{k,n} = -\frac{1}{n}\sum_{i=0}^{n-1} \log_2 P(x_{i+1}|x_{i-k+1..i}) \quad \text{bpb}, \tag{4}$$

where "bpb" stands for bits per base. The process of supervision is held by mixture weights which relate each static and dynamic model. Therefore, the probability estimate can be given by a weighted average of the probabilities provided by each model, according to

$$P(x_{n+1}) = \sum_k P(x_{n+1}|x_{n-k+1..n})\, w_{k,n}, \tag{5}$$

where $w_{k,n}$ denotes the weight assigned to model k and $\sum_k w_{k,n} = 1$.

For stationary sources, we could compute weights such that $w_{k,n} = P(k|x_{1..n})$, i.e., according to the probability that model k has generated the sequence until that point. In that case, we would get

$$w_{k,n} = P(k|x_{1..n}) \propto P(x_{1..n}|k)P(k), \tag{6}$$

where $P(x_{1..n}|k)$ denotes the likelihood of sequence $x_{1..n}$ being generated by model k and $P(k)$ denotes the prior probability of model k. Assuming $P(k) = \frac{1}{K}$, where K denotes the number of models, we obtain $w_{k,n} \propto P(x_{1..n}|k)$. Calculating the logarithm we get

$$\log_2 P(x_{1..n}|k) = \log_2 \prod_{i=1}^{n} P(x_i|k, x_{1..i-1}) = \sum_{i=1}^{n} \log_2 P(x_i|k, x_{1..i-1}), \tag{7}$$

which corresponds to the code length that would be required by model k for representing the sequence $x_{1..n}$. It is, therefore, the accumulated measure of the performance of model k until instant n. However, since the DNA sequences are not stationary, a good performance of a model in a certain region of the sequence might not be attained in other regions. Hence, the performance of the models have to be measured in the recent past of the sequence, for example over a window of appropriate size, or be equipped with a mechanism of progressive forgetting of past measures. We opted for the latter possibility, using the recursive relation

$$\sum_{i=1}^{n} \log_2 P(x_i|k, x_{1..i-1}) = \tag{8a}$$

$$= \gamma \sum_{i=1}^{n-1} \log_2 P(x_i|k, x_{1..i-1}) + \log_2 P(x_n|k, x_{1..n-1}). \tag{8b}$$

As can be verified, this relation corresponds to a first-order recursive filter that, for $\gamma \in [0,1)$, has a low-pass characteristic and an exponentially decaying impulse response. For more information on finite-context modelling and mixtures see [8, 9].

Table 1: Data set table. The number of expected chromosome pairs for each species is represented by 'Exp', while 'Missing' is a nonexistence sequence and Mb represents the approximated size in Mega bases.

Organism	Build	Exp	Missing	Mb
Homo sapiens	37.p10	23	-	2,861
Pan troglodytes	2.1.4	24	-	2,756
Gorilla gorilla	r100	24	Y	2,719
Pongo abelii	1.3	24	Y	3,028

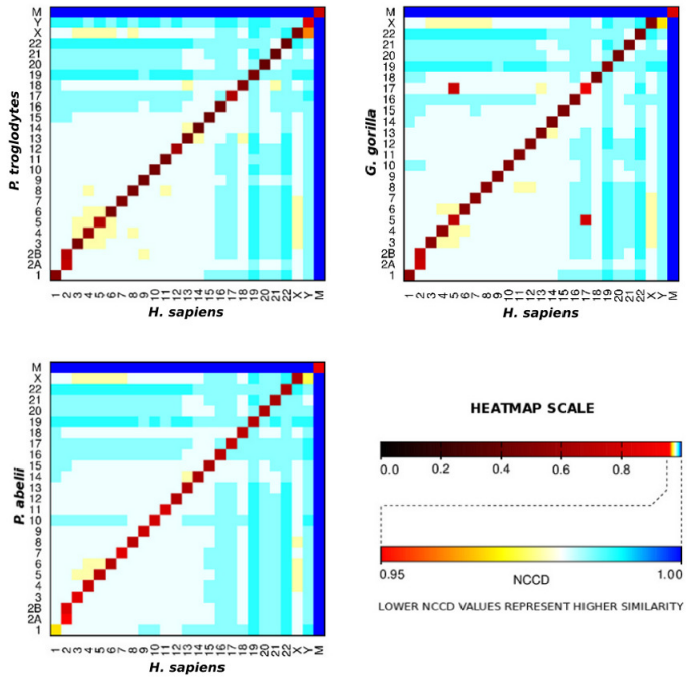

Figure 1: *P. troglodytes*, *G. gorilla* and *P. abelii* inter-genomics chromosomal NCD heatmaps in relation to *H. sapiens*.

3 Results

The data set is composed by 4 genomes (Table 1), downloaded from the NCBI (`ftp://ftp.ncbi.nlm.nih.gov/genomes`).

In Fig. 1 the inter-chromosomal NCD distance heatmaps for the three species relatively to *H. sapiens* have been plotted in an approach *all with all*. As it can be seen, for all species there is a direct correlation with the respective chromosomal number, with the exception of chromosome 2 (related to 2A and 2B). This is justified by a presumed chromosomal fusion in humans from previous ancestors [5].

Moreover, the human Y chromosome is highly related with the X chromosome of all addressed species, namely for the *P. troglodytes*, because the Y chromosome exchanged genetic information with X in the recombination process [3]. Furthermore, there is a low distance between chromosomes 5 and 17 of *G. gorilla* and *H. sapiens*, justified by a chromosomal translocation [10].

On the other hand, in Fig. 2 are presented the chromosomal distances of *P. troglodytes*, *G. gorilla* and *P. abelii* (chromosomes 2A and 2B have been concatenated) according to *H. sapiens* chromosomes order. At glance, *P. troglodytes* has got the lowest distance relatively to *H. sapiens*, and after *G. gorilla* and *P. abelii*, respectively. Specifically, *G. gorilla* chromosomes 5 and 17 have large distances because of the previous mentioned translocation, while *P. abelii* seems to have a very different chromosome 1, besides other relevant dissimilarities.

According to [4], besides the high divergence of Y chromosome, there are several breakpoints in chromosomes 4, 5 and 12, which were tested by fluorscense *in situ* hybridization (FISH), in *P. troglodytes* using *H. sapiens* as reference. Fig. 2 reports the same dissimilarities, surprisingly adding chromosome 17.

Finally, we have found that chromosomes 4, 12 and 18 of *G. gorilla* have lower distances to *H. sapiens* than to the respective *P. troglodytes* chromosomes, while chromosomes 5 and 17 of *G. gorilla* have higher distances than those of *P. abelii*. Mitochondrial sequences, as expected,

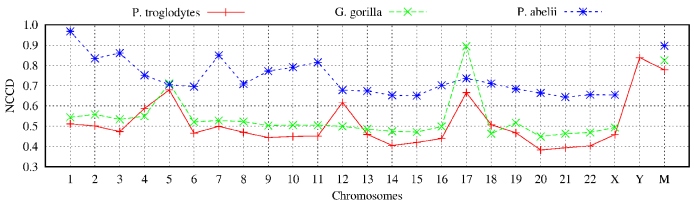

Figure 2: *P. troglodytes*, *G. gorilla* and *P. abelii* related chromosomal NCD values using *H. sapiens* as reference.

show that *P. troglodytes* is the nearest *H. sapiens* species, followed by the *G. gorilla* and, lastly, by *P. abelii*.

4 Conclusions

An admissible normalized compression distance has been proposed, based on a specific conditional compression analogue. The compressor is constituted by a set of multiple static and dynamic finite-context models that are supervised by a mixture model.

We have addressed a study on chromosomal distances between *Hominidae* primates, agreeing with several already documented results (using other approaches), but also unveiling undocumented ones.

The biggest advantage of this method is the process automation for any kinds of genomic DNA sequences, while using biological techniques can be a rather lengthy process.

Acknowledgements

This work was supported in part by FEDER through the Operational Program Competitiveness Factors - COMPETE and by National Funds through FCT - Foundation for Science and Technology, in the context of the projects FCOMP-01-0124-FEDER-022682 (FCT reference PEst-C/EEI/UI0127/2011) and Incentivo/EEI/UI0127/2013

References

[1] C. H. Bennett, P. Gács, M. Li P. M. B. Vitányi, and W. H. Zurek. Information distance. *IEEE Trans. on Information Theory*, 44(4): 1407–1423, July 1998.

[2] R. Cilibrasi and P. M. B. Vitányi. Clustering by compression. *IEEE Trans. on Information Theory*, 51(4):1523–1545, April 2005.

[3] J. Hughes et al. Chimpanzee and human y chromosomes are remarkably divergent in structure and gene content. *Nature*, 463(7280): 536–539, 2010.

[4] T. Mikkelsen et al. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 2005.

[5] J. Ijdo, A. Baldini, D. Ward, S. Reeders, and R. Wells. Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proceedings of the National Academy of Sciences*, 88(20):9051–9055, 1991.

[6] M. Li and P. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 3rd edition, 2008.

[7] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi. The similarity metric. *IEEE Trans. on Information Theory*, 50(12):3250–3264, December 2004.

[8] A. J. Pinho, D. Pratas, and P. J. S. G. Ferreira. Bacteria DNA sequence compression using a mixture of finite-context models. In *Proc. of the IEEE Workshop on Statistical Signal Processing*, Nice, France, June 2011.

[9] D. Pratas and A. J. Pinho. Compressing the human genome using exclusively Markov models. In *Advances in Intelligent and Soft Computing, Proc. of the 5th Int. Conf. on Practical Applications of Computational Biology & Bioinformatics, PACBB 2011*, volume 93, pages 213–220, April 2011.

[10] R. Samonte and E. Eichler. Segmental duplications and the evolution of the primate genome. *Nature Reviews Genetics*, 3(1):65–72, 2002.