Clustering DNA sequences by relative compression

Morteza Hosseini[1]
seyedmorteza@ua.pt
Diogo Pratas[1,2]
pratas@ua.pt
Armando J. Pinho[1]
ap@ua.pt

[1] IEETA/DETI,
University of Aveiro
[2] DV,
University of Helsinki.

Abstract

With advancement of high-throughput sequencing technologies, a huge volume of data is produced every day, which has led to an acceleration of biological and medical research and discovery. We present a novel method for clustering DNA sequences based on relative compression. The method does not use any subject-specific feature or background knowledge. Tested on 30 different sequences, we could successfully classify them into three groups. The results show, evidently, that two groups of Actinopterygii and Chondrichthyes, that are both fishes, are more similar to each other, compared to Mammalia group.

1 Introduction

Ever-increasing growth of high-throughput sequencing technologies has led to the production of a huge volume of genomic data. Some of this data are more alike than others. It is essential to have methods for identifying groups with similar patterns. In this paper, we propose a method for clustering DNA sequences, using a similarity metric based on relative compression. This method does not use any background knowledge about the data. We employ normalized relative compression (NRC) as the similarity metric. For the purpose of clustering, we apply a bottom-up hierarchical clustering approach.

It has been shown in the literature, [1], that clustering by compression is not restricted to the specific area of genomics, and can have application in areas of literature, music, virology, languages, handwritten digits and astronomy.

In the following sections, we describe the proposed method in detail. Then, we demonstrate the results of running the method on a dataset including 30 DNA sequences. Finally, we draw some conclusions.

2 Methods

In order to classify the sequences, we first find similarity between sequences. For this purpose, we use GeCo [5] to compress all sequences, considering them as references as well as targets. The compression is an approximation of Kolmogorov complexity, which is not computable, and can yield the information (complexity) included in a sequence [4]. For measuring the similarity, normalized relative compression is used, that can be calculated as [7]

$$\mathrm{NRC}(x||y) = \frac{C(x||y)}{|x| \log_2 |\Phi|}, \tag{1}$$

in which $C(x||y)$ is the information in the sequence x and is obtained by compressing x relatively to the sequence y, $|x|$ is the size of sequence x, Φ is the alphabet $\{A,C,G,T\}$ used in DNA sequences and $|\Phi|$ is the cardinality of the alphabet, i.e. $\mathrm{size}(\Phi) = 4$. Values of NRC falls within the range $(0,1]$ and the more similar two sequences are, the less is this value. This metric is computationally lightweight, meaning that it only needs to compress a target sequence (using a reference), and does not need to compress the reference sequence or pairwise concatenation of any kind [7].

GeCo works based on a mixture of finite-context models (FCMs) and extended finit-context models (XFCMs), in which mixture weights are frequently updated during the compression process, according to the performance of each probabilistic model [5]. FCMs are probabilistic models that rely on Markov property and consider the k most recent symbols of an information source to estimate the probability of the next symbol [3]. XFCMs are probabilistic-algorithmic models that consider the occurrence probabilities stored in memory and assume that the next symbol is the one with the highest probability. Therefore, they do not consider the actual symbol in the sequence [6].

In the next step, we use weighted pair group method with arithmetic mean (WPGMA), which is a bottom-up hierarchical clustering method, to classify the sequences based on NRC values. The WPGMA algorithm employs a similarity matrix to construct a rooted tree (dendrogram) [2, 8]. At each step, the nearest two clusters a and b are combined into a higher-level cluster $a \cup b$. Then, its distance to another cluster c is the arithmetic mean of the distance of a,c and b,c:

$$d_{(a\cup b),c} = \frac{d_{a,c} + d_{b,c}}{2}, \tag{2}$$

in which d denotes the distance.

3 Results

The proposed method is implemented and publicly available at `github.com/smortezah/Clusico`, under GPLv3 license. The machine used for the tests had an 8-core 3.40 GHz Intel® Core™ i7-6700 CPU with 32 GB RAM.

For the experiments, we have used 30 mitochondrial DNA (mtDNA) sequences from three groups of Actinopterygii (Ray-finned fishes), Chondrichthyes (Cartilaginous fishes) and Mammalia, that can be downloaded from `www.ncbi.nlm.nih.gov/nuccore`. Each groups contains 10 sequences and their sizes varies from 16,189 to 18,431 bases. These sequences are listed in Table 1.

Table 1: Datasets used in the experiments.

Accession	Group	Organism
NC_005796	Actinopterygii	*Pterothrissus gissu*
NC_015337	Actinopterygii	*Canthigaster valentini*
NC_014404	Actinopterygii	*Hapalogenys nigripinnis*
NC_015823	Actinopterygii	*Diplomystes nahuelbutaensis*
NC_004449	Actinopterygii	*Gadus chalcogrammus*
NC_004701	Actinopterygii	*Eigenmannia sp. CBM-ZF-10620*
NC_016709	Actinopterygii	*Clupeoides borneensis*
NC_015544	Actinopterygii	*Horadandia atukorali*
NC_006533	Actinopterygii	*Anguilla australis schmidti*
NC_007012	Actinopterygii	*Scleropages formosus*
NC_024269	Chondrichthyes	*Lamna ditropis*
NC_021768	Chondrichthyes	*Glyphis glyphis*
NC_024862	Chondrichthyes	*Carcharhinus macloti*
NC_026696	Chondrichthyes	*Carcharhinus amboinensis*
NC_021443	Chondrichthyes	*Alopias superciliosus*
NC_022841	Chondrichthyes	*Rhinobatos hynnicephalus*
NC_022821	Chondrichthyes	*Pristis clavata*
NC_023455	Chondrichthyes	*Rhincodon typus*
NC_024110	Chondrichthyes	*Pristiophorus japonicus*
NC_027521	Chondrichthyes	*Dipturus trachyderma*
NC_027083	Mammalia	*Lynx lynx*
NC_006364	Mammalia	*Zaglossus bruijni*
NC_007629	Mammalia	*Lipotes vexillifer*
NC_025516	Mammalia	*Mustela erminea*
NC_011137	Mammalia	*Homo sapiens neanderthalensis*
NC_010299	Mammalia	*Daubentonia madagascariensis*
NC_005035	Mammalia	*Mogera wogura*
NC_008753	Mammalia	*Ursus thibetanus mupinensis*
NC_026088	Mammalia	*Megaladapis edwardsi*
NC_008417	Mammalia	*Arctocephalus pusillus*

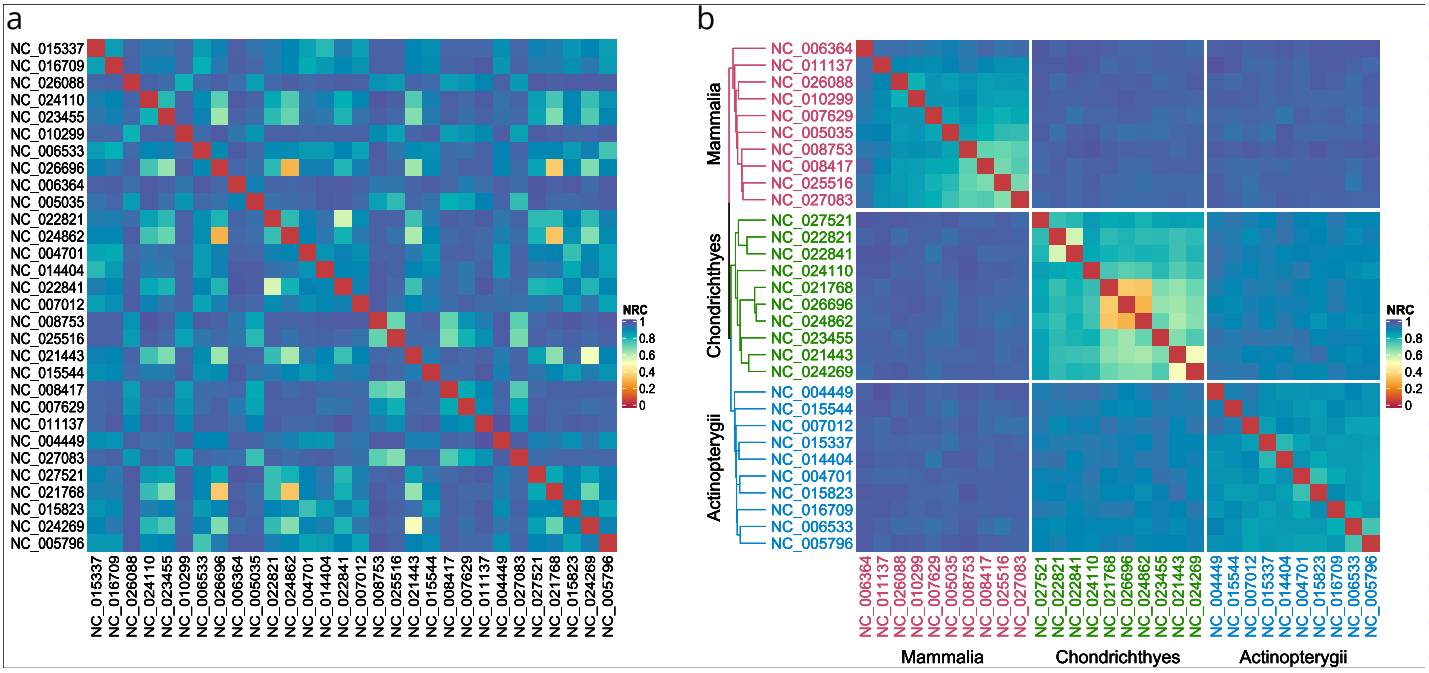

Figure 1: (a) Similarity between different DNA sequences. The more the NRC value is, the less the similarity of two sequences will be; (b) classification result of the sequences. The sequences in Mammalia, Chondrichthyes and Actinopterygii groups are shown in red, green and blue colors, respectively.

Figure 1a shows similarity between different sequences (NRC values), obtained by GeCo. As is show, when a sequence is compressed relatively to itself, the NRC value will be approximately 0. These cases are shown with red squares.

Figure 1b demonstrates the result of classification of the sequences, which is obtained by WPGMA algorithm. The sequences in Mammalia, Chondrichthyes and Actinopterygii groups are shown with red, green and blue colors, respectively. On top of this figure, the dendrogram is plotted, which shows similarity of different sequences within each group and also, similarity between different groups. To have a better view on the dendrogram, we have shown the clusters in Figure 2. As it is shown, the two groups of Chondrichthyes and Actinopterygii, that are fishes, are more similar to each other, in comparison to Mammalia.

4 Conclusions

We presented a novel method for clustering DNA sequences, using a similarity metric obtained by relative compression. The method is unsupervised in that it does not use any background knowledge about the DNA sequences. We tested the proposed method on 30 different sequences from three groups of Actinopterygii, Chondrichthyes and Mammalia, with 10 sequences in each group. The results showed that we could successfully cluster these sequences into each group. Also, the results showed that the two groups of Actinopterygii and Chondrichthyes, that are both fishes, are more alike each other, in comparison to Mammalia group.

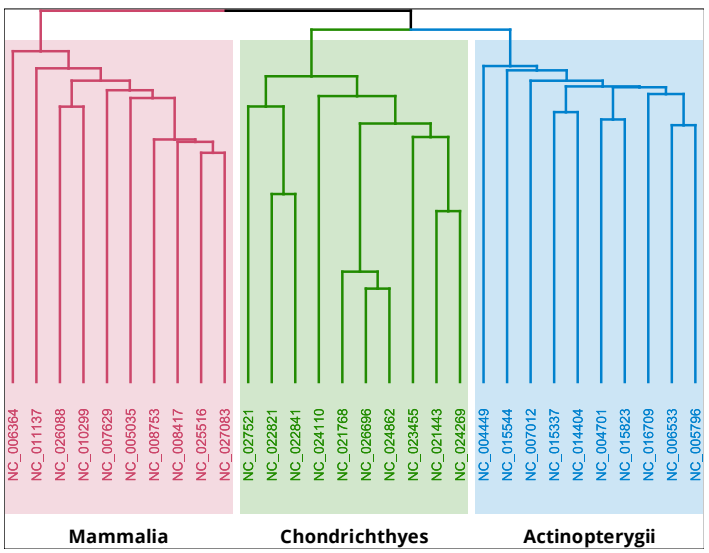

Figure 2: Dendrogram for the DNA sequences.

Acknowledgements

This work was supported by FEDER (Programa Operacional Factores de Competitividade — COMPETE), and by national funds through the FCT, in the context of the projects UID/CEC/00127/2019 and PTCD/EEI-SII/6608/2014 and the grant PD/BD/113969/2015.

References

[1] Rudi Cilibrasi and Paul MB Vitányi. Clustering by compression. *IEEE Transactions on Information theory*, 51(4):1523–1545, 2005.

[2] Harry Clifford, Frank Wessely, Satish Pendurthi, and Richard D Emes. Comparison of clustering methods for investigation of genome-wide methylation array data. *Frontiers in genetics*, 2:88, 2011.

[3] Morteza Hosseini, Diogo Pratas, and Armando J Pinho. AC: A compression tool for amino acid sequences. *Interdisciplinary Sciences: Computational Life Sciences*, 11(1):68–76, 2019.

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

[5] Diogo Pratas, Armando J Pinho, and Paulo JSG Ferreira. Efficient compression of genomic sequences. In *Data Compression Conference (DCC)*, pages 231–240. IEEE, 2016.

[6] Diogo Pratas, Morteza Hosseini, and Armando J Pinho. Substitutional tolerant markov models for relative compression of dna sequences. In *International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 265–272. Springer, 2017.

[7] Diogo Pratas, Raquel Silva, and Armando J Pinho. Comparison of compression-based measures with application to the evolution of primate genomes. *Entropy*, 20(6):393, 2018.

[8] Robert R Sokal and Charles D Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.