Clustering DNA sequences by relative compression

Morteza Hosseini
seyedmorteza@ua.pt
Diogo Pratas
pratas@ua.pt
Armando J. Pinho
ap@ua.pt

IEETA/DETI,
University of Aveiro

Abstract

In this paper, we present a method for clustering DNA sequences, using relative compression. Tested on 30 different sequences, we could successfully classify them into three groups. The results show that two groups of Actinopterygii and Chondrichthyes, that are both fishes, are more similar to each other, compared to Mammalia group.

1 Introduction

2 Methods

In order to classify the sequences, we first find similarity of sequences to each other. For this purpose, we use GeCo [3] to compress all sequences, considering them as references as well as targets. For measuring the similarity, normalized relative compression (NRC) is used, that can be calculated as [5]

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

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 and $|\Phi|$ is the cardinality of input DNA sequences, i.e. $\mathrm{size}(\{A,C,G,T\}) = 4$. Values of NRC falls within the range $(0,1]$ and the more similar two sequences are, the less is this value.

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 [3]. 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 [2]. 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 [4].

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) [1, 6]. 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 1a shows similarity of 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 of different groups. As it is show, the two groups of Chondrichthyes and Actinopterygii, that are fishes, are more similar to each other, in comparison with Mammalia.

References

[1] 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.

[2] 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.

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

[4] 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.

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

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

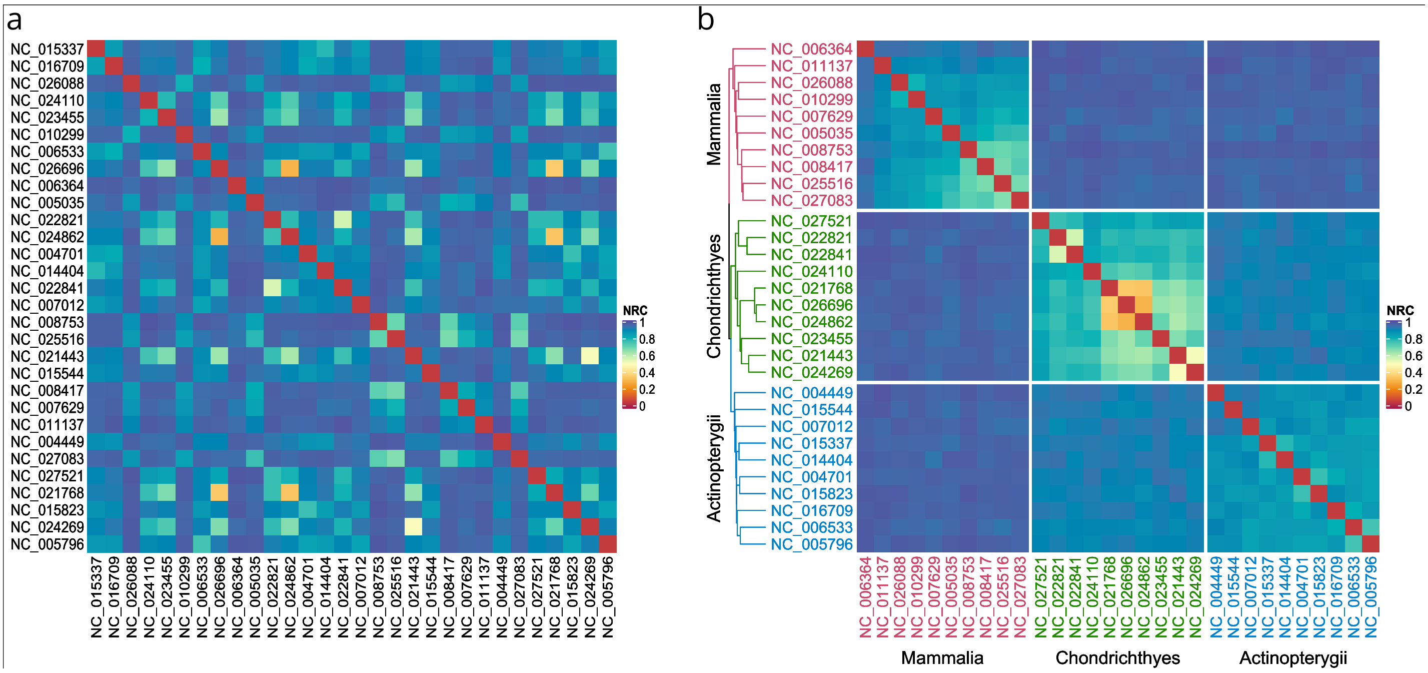
a
NC_015337
NC_016709
NC_026088
NC_024110
NC_023455
NC_010299
NC_006533
NC_026696
NC_006364
NC_005035
NC_022821
NC_024862
NC_004701
NC_014404
NC_022841
NC_007012
NC_008753
NC_025516
NC_021443
NC_015544
NC_008417
NC_007629
NC_011137
NC_004449
NC_027083
NC_027521
NC_021768
NC_015823
NC_024269
NC_005796
NRC
1
0.8
0.6
0.4
0.2
0
b
Mammalia
NC_006364
NC_011137
NC_026088
NC_010299
NC_007629
NC_005035
NC_008753
NC_008417
NC_025516
NC_027083
Chondrichthyes
NC_027521
NC_022821
NC_022841
NC_024110
NC_021768
NC_026696
NC_024862
NC_023455
NC_021443
NC_024269
Actinopterygii
NC_004449
NC_015544
NC_007012
NC_015337
NC_014404
NC_004701
NC_015823
NC_016709
NC_006533
NC_005796
NRC
1
0.8
0.6
0.4
0.2
0
Mammalia
Chondrichthyes
Actinopterygii

Figure 1: NRC.