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

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 1: Results. Ours is better.

Abstract

This document demonstrates the format requirements for papers submitted to the Portuguese Conference on Pattern Recognition. The format is designed for easy on-screen reading, and to print well at one or two pages per sheet. Additional features include: pop-up annotations for citations [**? ?**]; a margin ruler for reviewing; and a greatly simplified way of entering multiple authors and institutions.

1 Introduction

2 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`. The size of these sequences varies from 16,189 to 18,431 bases.

In order to classify the sequences, we first ran GeCo [2] on all sequences, considering them as references as well as targets, to find similarity of sequences to each other. For measuring the similarity, normalized relative compression (NRC) were used, that can be calculated as [3]

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

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.

In the next step, we used weighted pair group method with arithmetic mean (WPGMA), which is a bottom-up hierarchical clustring method, to classify the sequences based on NRC values. The WPGMA algorithm employs a similarity matrix to construct a rooted tree (dendrogram) [1, 4].

Figure 2a 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.

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

[3] 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.

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

Figure 1: It is often a good idea for the first figure to attempt to encapsulate the article, complementing the abstract. This figure illustrates the various print and on-screen layouts for which this paper format has been optimized: (a) traditional print format; (b) on-screen single-column format, or large-print paper; (c) full-screen two column, or 2-up printing.

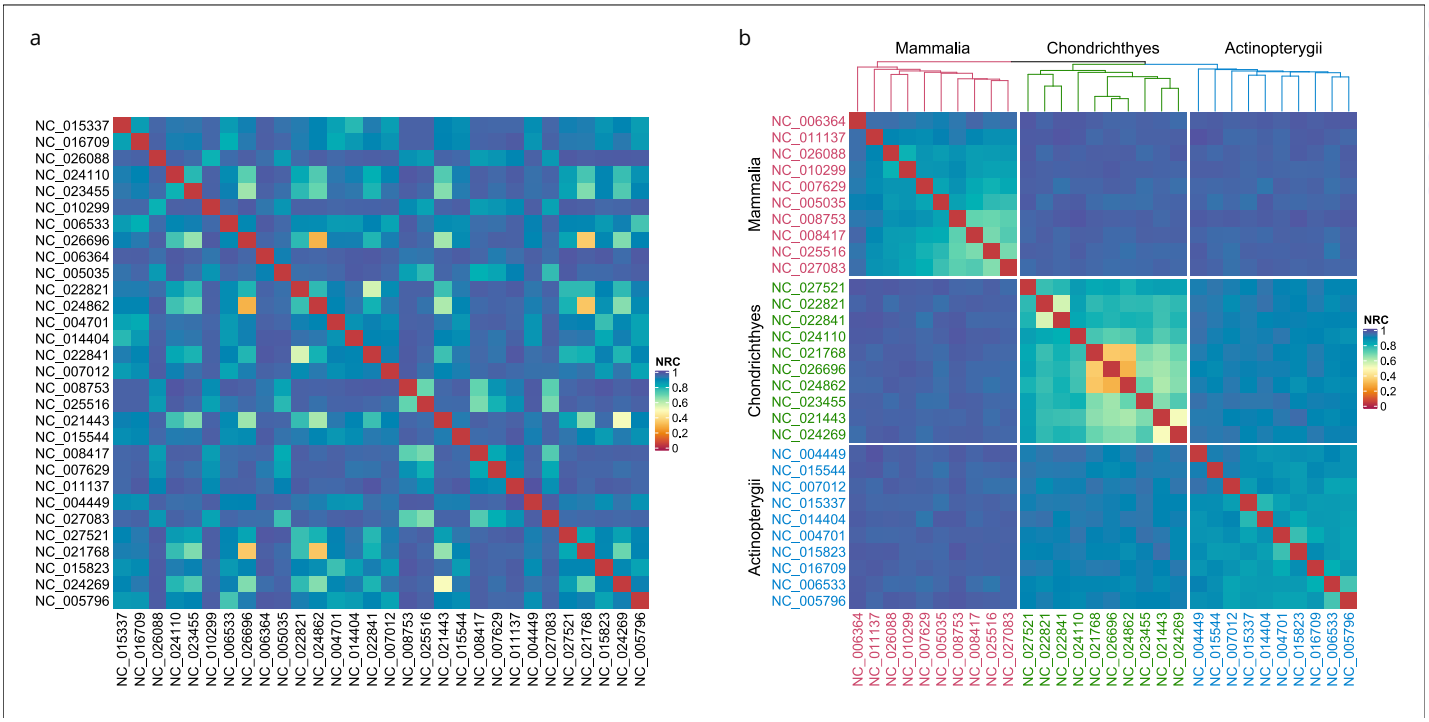

Figure 2: NRC.