

# 000 Clustering DNA sequences by relative compression

001

002 Morteza Hosseini  
003 seyedmorteza@ua.pt

004 Diogo Pratas  
005 pratas@ua.pt

006 Armando J. Pinho  
007 ap@ua.pt

IEETA/DETI,  
University of Aveiro

008

009

## 010 Abstract

011

012 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.

016

017

## 018 1 Introduction

019

## 020 2 Results

021

022 The proposed method is implemented and publicly available at [github.com/smortezah/Clusico](https://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.

024

025 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](http://www.ncbi.nlm.nih.gov/nuccore). The size of these sequences varies from 16,189 to 18,431 bases.

029

030 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]

032

$$033 \text{NRC}(x|y) = \frac{C(x|y)}{|x| \log_2 |\Phi|}, \quad (1)$$

034

035 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.  $\text{size}(\{A, C, G, T\}) = 4$ . Values of NRC falls within the range  $(0, 1]$  and the more similar two sequences are, the less is this value.

040

041 In the next step, we used 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, 4].

044

045 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.

046

047 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. This dendrogram shows that Chondrichthyes and Actinopterygii groups, that are both fishes, are more similar to each other, rather than to

054

## 055 References

056

057 [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.

060

061 [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.

062

- [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.

