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 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.

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 https://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 on all sequences, considering them as references as well as targets. As the result, normalized relative compression (NRC) values were obtained, that can be calculated as

063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125

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.