

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]

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

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

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].

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.

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

Table 1: Datasets used in the experiments.

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

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.

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

126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188

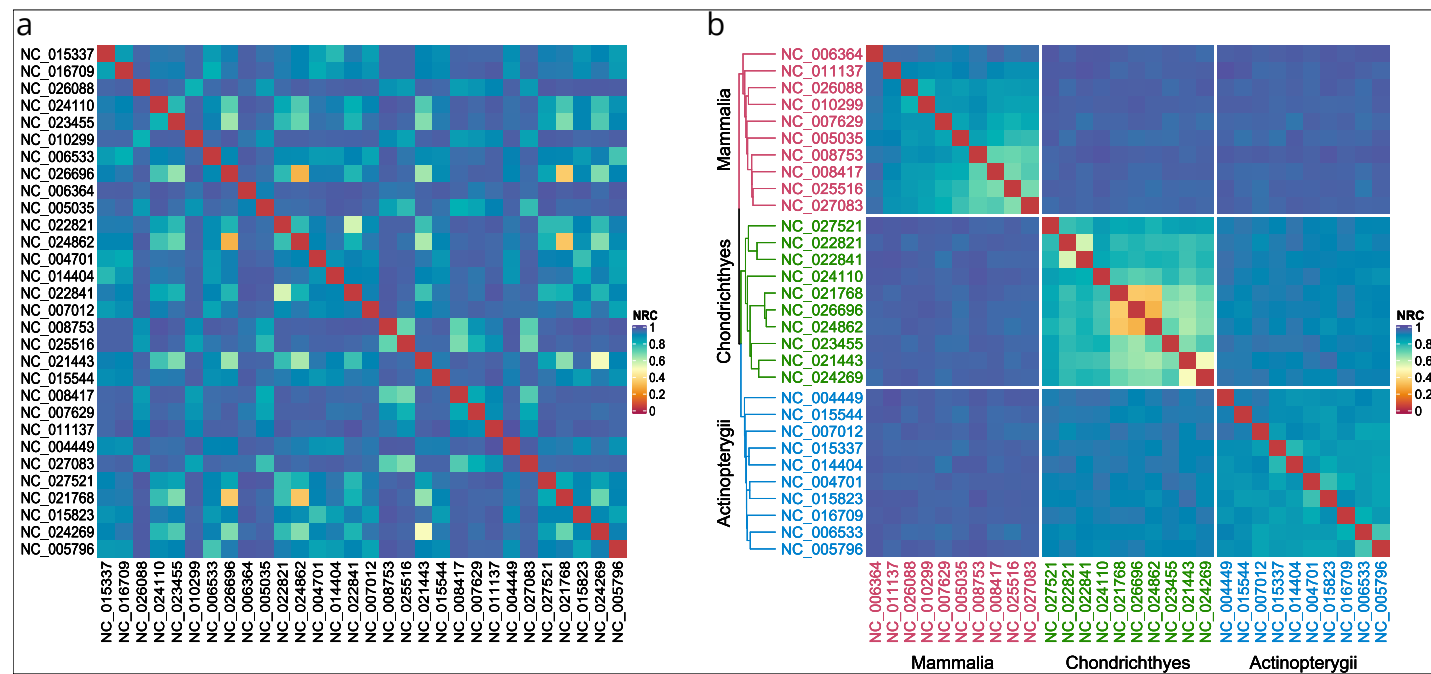


Figure 1: NRC.