Task 1 - Finding Neighbours Using Compression

Domen Lušina
Bioinformatics algorithms

November 9, 2018

1 Introduction

The goal of the assignment was to measure similarity between pairs of organisms. We were provided with a file containing corresponding segments of DNA from 8 organisms. In order to compare two organisms we had to compress their DNA represented as a string. We used two compress methods GenCompress and zlib.

2 Solution description

My general solution can be described in following steps:

1. Reading FAST file and obtaining DNA string of each organism (denoted with letters A-H).
2. Concating DNA strings of two organisms.
3. Compressing strings and calculating pairwise distances (only for distances *D(u,v), u<v*).
4. Computing distance matrix and building a dendrogram using hierarchical clustering.

2.1 Measuring similarity with GenCompress

Our first method of compressing DNA is using a GenCompress[1]. In order to use this program we had to use command prompt and generate files obtaining DNA strings (strings of single organisms and concatenated strings). After that I used a *.bat* file to run the compression and save the output into a text files in *out* directory. After that I used python to get length of the compressed file from text output files. In order to measure distance we had to use the following equation:

$$D(u,v) = 1 - \frac{|Compress(u)| \; - \; |Compress(u|v)|}{|Compress(uv)|}, \tag{1}$$

where $|Compress\ (u)|$ and $|Compress\ (v)|$ are the lengths of compressed DNA strings, $|Compress(uv)|$ is the length of compressed concatenated string of two organisms u and v and $|Compress(u|v)|$ is length of compressed string u knowing v.

[1]GenCompress can be found on: http://www1.spms.ntu.edu.sg/ chenxin/GenCompress/

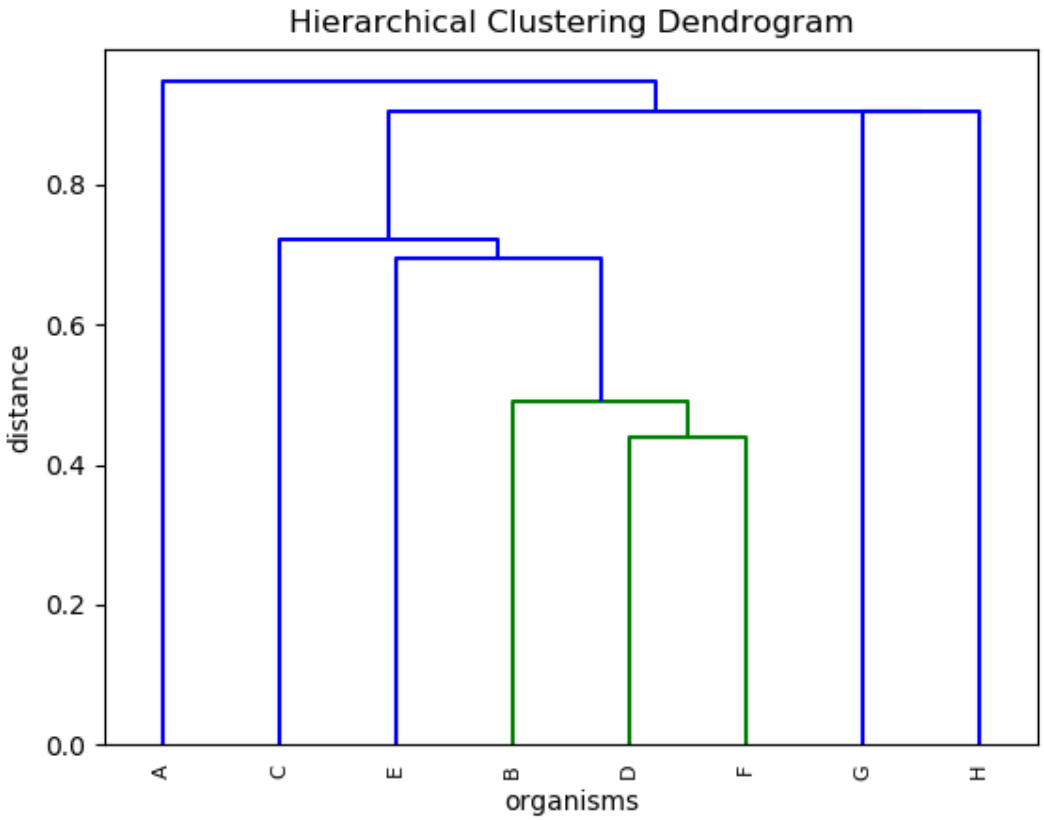

Figure 1: Phylogenetic tree obtained with hierarchical clustering and using GenCompress.

	A	B	C	D	E	F	G	H
A	0.000	0.950	0.956	0.951	0.947	0.951	0.958	0.965
B	0.950	0.000	0.731	0.501	0.695	0.491	0.908	0.930
C	0.956	0.731	0.000	0.757	0.723	0.765	0.930	0.942
D	0.951	0.501	0.757	0.000	0.708	0.439	0.914	0.925
E	0.947	0.695	0.723	0.708	0.000	0.711	0.904	0.919
F	0.951	0.491	0.765	0.439	0.711	0.000	0.913	0.929
G	0.958	0.908	0.930	0.914	0.904	0.913	0.000	0.903
H	0.965	0.930	0.942	0.925	0.919	0.929	0.903	0.000

Table 1: Distance matrix computed using GenCompress compression algorithm.

2.2 Measuring similarity with zlib library

In our second approach we used zlib[2] library to compress out DNA strings. In this approach we didn't have to generate string files because zlib is available as a python library. However we could not compute "conditional" compression, so we had to use the following distance equation:

$$D(u,v) = 1 - \frac{|Compress(u)| - (\,|Compress(uv)| - |Compress(v)|)}{|Compress(uv)|}. \tag{2}$$

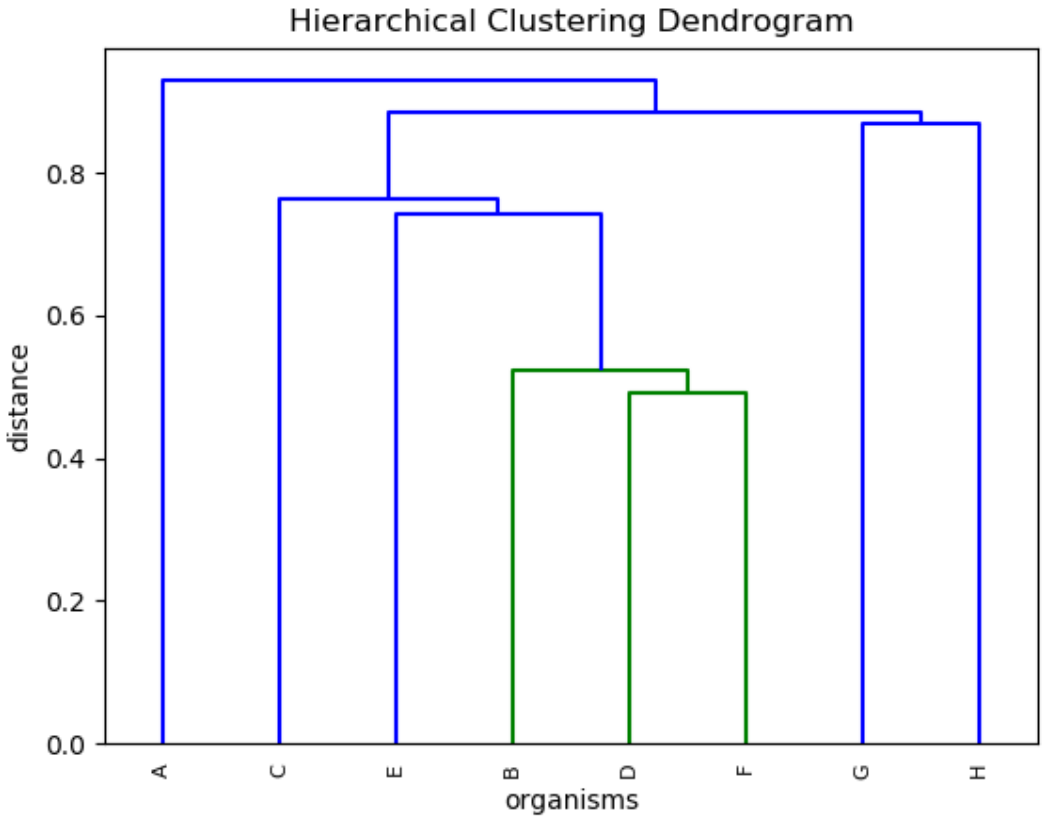

Figure 2: Phylogenetic tree obtained with hierarchical clustering and using zlib library.

[2]More about zlib library: http://www.zlib.net/manual.html

	A	B	C	D	E	F	G	H
A	0.000	0.932	0.946	0.938	0.930	0.934	0.939	0.941
B	0.932	0.000	0.772	0.570	0.743	0.522	0.896	0.899
C	0.946	0.772	0.000	0.799	0.765	0.775	0.914	0.921
D	0.938	0.570	0.799	0.000	0.759	0.491	0.900	0.897
E	0.930	0.743	0.765	0.759	0.000	0.750	0.886	0.888
F	0.934	0.522	0.775	0.491	0.750	0.000	0.894	0.894
G	0.939	0.896	0.914	0.900	0.886	0.894	0.000	0.870
H	0.941	0.899	0.921	0.897	0.888	0.894	0.870	0.000

Table 2: Distance matrix computed using zlib compression algorithm.

3 Conclusion

Both methods showed nice results and displayed that compression can be used to compare similarity between two organisms. There was a slight difference in computed distances and therefor build dendrograms were slightly different. However if we take a closer look at both dendrograms we see that grouping of organisms was practically the same. It would be hard to say which compression method was better, but since we can compute conditional compression with GenCompress we can say that GenCompress is better in that way. However for better evaluation we would need and expert to evaluate our results and we should have used a bigger set of organisms.