UNIVERSIDADE DE AVEIRO

DEPARTAMENTO DE ELECTRÓNICA, TELECOMUNICAÇÕES E INFORMÁTICA

Algorithmic Information Theory (2024/2025)

Lab Work #2

Deliver: 11 April 2025

1 The challenge

In the context of Algorithmic Information Theory, the compressibility of sequences provides a powerful tool for assessing similarities and identifying patterns in biological datasets. One of the most intriguing challenges in Exobiology is the analysis of exogenous metagenomes, which consist of genetic sequences from multiple organisms, often without direct references.

In this practical assignment, all the groups will receive a sequenced metagenomic sample that may come from the European Space Station. This sample may contain complete or fragmented sequences (genomes) of unknown organisms, which could be of terrestrial or extraterrestrial origin. The goal of this challenge is to develop a program that implements the Normalized Relative Compression (NRC) to estimate which known organisms from a given reference database share the highest similarity with the genomes found in the sample.

The reference database will contain known organisms, allowing each group to use NRC to infer relative similarity, possible contaminations, or even hints of new life forms. The program's performance will be evaluated based on the accuracy of the identification and the computational efficiency of the proposed solution.

2 Methodology

The program (named **MetaClass**) should follow these steps:

1. **Train a finite-context model (Markov model)** using only the **metagenomic sample** y. During this process, the model will learn the frequency of patterns present in the sample.

2. **Freeze the model's counts**, meaning that after training on y, the model counts will no longer be updated.

3. For each sequence x^i in the reference database of known organisms:

   - **Estimate the amount of bits required to compress** x^i using the model trained on y.
   - **Apply the Normalized Relative Compression (NRC)** to measure the relative compression of x^i to y.

4. **Sort by NRS the names of the sequences** and **print a top** 20.

3 Normalized Relative Compression

Formally, the NRC of a reference y and a target x sequence is given by

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

where $C(x||y)$ represents the number of bits needed to lossless compress x given exclusively a model trained with y, $|x|$ the size of sequence x, and $\log_2(\mathcal{A})$ the alphabet of sequence x, which in our case $\mathcal{A} = 4$ and, hence, $\log_2(4) = 2$.

Since in our application we require to compute the NRC for each x^i according to a reference y, then

$$\mathcal{NRC}(x^i||y) = \frac{C(x^i||y)}{2|x^i|}. \tag{2}$$

4 Data

In Moodle, two zipped files are available in the "Project #02" folder:

- **meta.zip** — A metagenomic file containing the sequenced DNA (y).
- **db.zip** — A database file where lines beginning with "@" represent identifiers (names) corresponding to the subsequent DNA reference sequences (x^i).

5 How to Deliver the Work

Send an email to both teachers (an@ua.pt and pratas@ua.pt), with all authors in CC, containing the compressed repository or the project link. If the repository is private on GitHub, include an invitation.

The repository must contain:

- Source code for the `MetaClass` program.
- Necessary files for testing.
- A `README.md` file with the following information:
  - **Installation Instructions**: Describe how to compile the program on a Linux machine, including any dependencies.
  - **Running the Programs**: Provide example commands to run the `MetaClass` programs, such as:
    - `MetaClass -d db.txt -s meta.txt -k 10 -a 0.1 -t 20`
  - **Compilation Instructions**: Include the exact commands to compile the programs, such as: `make` or `gcc`.
  - **Dependencies**: List any required libraries or tools and how should they be installed (the exact commands).
- Provide the **Report** in **PDF** format only.
- Provide a **video** presenting your work (**maximum video time: 5 minutes**).