Systems Analysis
Semester 2024-III
Workshop No. 1 — Entrophy and Divide&Conquer

Eng. Carlos Andrés Sierra, M.Sc.
Computer Engineering
Universidad Distrital Francisco José de Caldas

Welcome to the first workshop of the *Systems Analysis* course.

You have been hired by a *bioinformatics company* to be the new `computer science researcher`; we don't know how, but **you did it**. Your *first task* in one of the most typical, just to give you a proper on-boarding in the company: try to detect **motif** in data.

To simplify, a **motif** is a `pattern of strings` that appears *most frequently* in a set of *genetic sequences*.

So, let's describe the **steps** you need to follow in this challenge:

1. Create an `artificial database` of n *sequences* ($1000 <= n <= 2000000$), of *size* m ($5 <= m <= 100$).
   - Each *sequence* must be composed of `A`, `C`, `G`, `T`, the typical **nucleotide bases**.
   - The *probability* of selection of any **nucleotide base** must be a *parameter*.
   - The **artificial database** must be saved as a `.txt` file.
   - Think of a `divide and conquer` or `distributed computing` **strategy** to achieve this task.
2. Define an *algorithm* to iterate over the `data` and obtain the **motif** of a *size* s ($4 <= s <= 10$), given as a *parameter*.
   - Remember, try all **combinations** of *nucleotide bases* of `size` s.
   - If there are some `patterns` with the *same occurrences*, stick with the one with the *highest consecutive repeated bases*.

Carlos Andrés Sierra, Computer Engineer, M.Sc. on Computer Engineering, Titular Professor at Universidad Distrital Francisco José de Caldas.

Any comment or concern related to this document could be send to Carlos A. Sierra at e-mail: *cavirguezs@udistrital.edu.co*

- Think in an *optimized way* to accomplish this task using `available programming knowledge`.

3. Perhaps, depending on the **randomness**, there will be `many repetitions` in each *sequence*. Also, you want to have a **chaotic system**, so the *most diverse sequences* are `the best`. Remember, **entropy** is a *measure of chaos* in a system, so you could use a measure of entropy (e.g. `Shannon entropy`) to define a filter to remove sequences with *so many repetitions of the same base*.

You want to perform some `experiments` in order to validate the work you did.

1. Generate different `artificial datasets` using different probabilities for each base and different amount of sequences, and search **motifs** with different size. Here I recommended you to create a `table of results` when you could summarize the results as: *Database Size*, *Probability of Bases*, *Motif Size*, *Motif*, *Motif Ocurrences*, *Time to Find Motif*.
2. Using the same `datasets` from the previous item, apply **Shannon Entropy** to *filter sequences*, but you must define the `best filter threshold` value to remove sequences in order to have a better definition of *chaos in the data*. Using the new `filtered datasets`, **repeat** the *same experiments* and create a `new results table`.
3. Write a **report** with the following *sections*: `systemic analysis`, `complexity analysis`, `chaos analysis`, `results`, `discussion of results`, and `conclusions`.

Remember that you want to become a *rock star systems engineer*, so you have to do your best. The report must be in **English**, `PDF format`.

Additionally, you must submit a `URL` to a *GitHub repository* where you will leave both the documentation & the code for all the course workshops. Create a folder for each one. Remember, the repository should have a general `README`, and inside aech folder a `README` with a summary of the workshop development, the contents of the workshop folder, where the `report` and `code` should also be placed.

Deadline: **Sunday, September 15, 2024, 5:00 PM.**