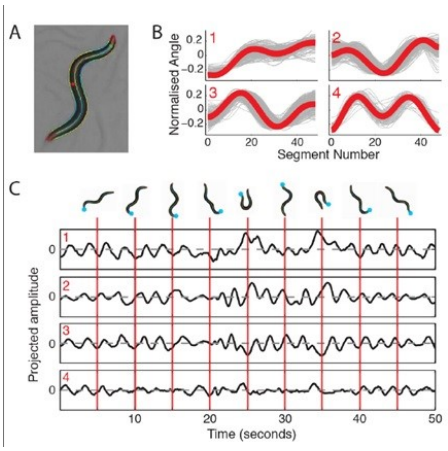
FINAL TASK FOR ANALISI DEI DATI – 2024/25

1. Introduction

Caenorhabditis elegans is a roundworm commonly used as a model organism in the study of genetics. The movement of these worms is known to be a useful indicator for understanding behavioural genetics. Brown *et al.*[1] describe a system for recording the motion of worms on an agar plate and measuring a range of human-defined features[2]. It has been shown that the space of shapes Caenorhabditis elegans adopts on an agar plate can be represented by combinations of six base shapes, or eigenworms. Once the worm outline is extracted, each frame of worm motion can be captured by six scalars representing the amplitudes along each dimension when the shape is projected onto the six eigenworms. Using data collected for the work described in [1], the aim is to address the problem of classifying individual worms as wild-type or mutant based on the time series of their motion.



Date: June 6, 2024.

[1] A. Brown, E. Yemini, L. Grundy, T. Jucikas, and W. Schafer, *A dictionary of behavioral motifs reveals clusters of genes affecting caenorhabditis elegans locomotion*, Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 10, no. 2, pp. 791–796, 2013.

[2] E. Yemini, T. Jucikas, L. Grundy, A. Brown, and W. Schafer, *A database of caenorhabditis elegans behavioral phenotypes*, Nature Methods, vol. 10, pp. 877–879, 2013.

2. Goals

Your first task is to explore possible ways of developing methods for classifying worm types, based on their motion. You are not restricted to using a single classification method (also in view of your second goal), and you can, in principle, develop different methods, taking also into account their computational complexity.

Your second task is to develop methods that allow to understand which characteristics of the motion (values along the six "eigenworms") are more relevant for the outcome of your classification method. Please consider that the results should be as much informative as possible and able to identify not only *which* eigendirections are more relevant, but also, if possible, give a quantitative measure of their influence.

3. Data

The dataset can be downloaded from

https://www.timeseriesclassification.com/description.php?Dataset=EigenWorms,

and contains 259 cases, split into 131 train and 128 test. You can ignore this partition and consider the dataset in its entirety, possibly using the methods discussed in the course of the lectures to assess errors.

Each worm is classified as either wild-type (the N2 reference strain) or one of four mutant types: `goa-1`, `unc-1`, `unc-38` and `unc-63`.

Data is in `.arff` format. You can use, for instance, the library `foreign` in `R`, or the package `SciPy` in `python`, to import the data.

4. Submission

The results of your own analysis and ideas must be summarised in a report (a PDF file) which explains how you have planned to tackle the problem and the possible strategies you have tried to solve the problem. The emphasis is not on the performances of the final method(s) proposed, but on the way you have dealt with the problem.

You are not only allowed but actually encouraged to read up on the subject, and it is recommended to include a list of references in your report. In order to be complete and fair, you are required to cite *all* sources of research material you have used (books, scientific papers, etc.).

4.1. **Fairness.** This final assignment is a personal piece of work and must not be done in groups. Discussions with colleagues or experts, although discouraged, should be reported for fairness.

4.2. **Use of AI.** Use of AI tools, and Large Language Models (LLM) in particular, is in general not allowed. You are welcome to use any tool that is suitable to improve your report, for instance spell checker, grammar suggestions, etc. If you use LLM for editing purposes, please declare it in your report. As the author of your work and of your report, you are responsible for the entire content of the report, which should be correct and *original*. Please avoid, for instance, including reference lists generated by an LLM, but include only those references that are pertinent (and that you have actually read!)

4.3. **Report submission.** You will upload your report on the *e-learning* website. The deadline is by **July 28, 2025** (but early submissions are appreciated). Do not forget to register to the exam on the page `https://esami.unipi.it` (deadline is July 21). This is not mandatory to conclude the exam, but useful for several purposes. If, for some reason, you need to complete the exam before the scheduled date, for instance for degree completion or as an Erasmus student, please contact us as soon as possible.

You should add, at the end of your report, the link to a script (`R` or `python`, or any other programming language of your choice) containing the implementation of the *final* method(s) proposed, based on the analysis developed. The script must be shared via a *notebook* on Google Colab. Obviously the script must not contain any errors. Please add a link to the notebook in your report.

It is not necessary (and in fact useless) for the script to contain the entire analysis. The recommendation is that the output of your scripts will be a detailed account of your conclusions. The numbers, without any explanation about their meaning, are not really helpful.