# COMPUTER AND CONTROL ENGINEERING

## CRT/DAUIN/DISAT - Statistical learning of genomic data

| | |
|---|---|
| **Funded By** | Dipartimento DAUIN<br>Dipartimento DISAT<br>FONDAZIONE CRT CASSA DI RISPARMIO DI TORINO<br>[P.iva/CF:06655250014] |
| **Supervisor** | FERRERO RENATO - renato.ferrero@polito.it |
| **Contact** | CARBONE ANNA FILOMENA - anna.carbone@polito.it |
| **Context of the research activity** | The information content of biological data can be quantified using various information-theoretic measures, such as entropy, mutual information, and compression complexity. The research activity concerns the study of different information-theoretic measures and clustering methods capable of providing insights into the functional and evolutionary properties of the genome. The research will be extended to the pangenome, i.e., the complete set of genes within a species, encompassing genetic variation found across different individuals. |
| **Objectives** | The structure and properties of DNA molecules can be conveniently represented by translating the sequence of nucleotides (A, T, C, G), for example into a one-dimensional numerical sequence (known as DNA walk). Then, different theoretic measures can be adopted to evaluate the information content of such generated sequence. For example, entropy measures the uncertainty or randomness of the sequence, with higher entropy indicating more randomness and lower entropy indicating more predictability. Mutual information can be used to quantify the dependence between different regions of the genome. Compression complexity quantifies the shortest possible description length of the sequence.<br><br>The research activity of the PhD candidate will provide valuable insights into various structural and functional aspects of DNA sequences. The complexity of DNA sequences will be assessed by examining the variability and unpredictability in the walk. The randomness (or order) of the sequence will be quantified by measuring the Shannon entropy. The analysis can detect periodic patterns that may have functional or regulatory roles, as well as identify large-scale structural variations such as inversions, translocations, or duplications that alter the walk's trajectory. Finally, the identified DNA walk patterns may be related to disease-causing variants or mutations.<br>The work plan will ensure progressive development from foundational studies to comprehensive analyses and practical applications. Foundation and preliminary studies will be carried out during the first year. An extensive review will survey methodologies, applications, and key findings on DNA |

| | |
|---|---|
| | walks. A comprehensive set of DNA sequences will be gathered from the Human Pangenome Reference Consortium. Existing software for sequence alignment, visualization, and statistical analysis will be evaluated.<br>The second year will focus on comprehensive data analysis and pattern identification. Statistical and computational methods will be applied to identify recurring patterns, motifs, and structural variations in the DNA walks. Complexity and entropy analyses will be performed to quantify the randomness and order within the DNA sequences. Efficient data management practices will be implemented to handle the large volume of generated data, ensuring data integrity and accessibility.<br>The third year finalized the research activities as it will concentrate on its applications. The candidate will attempt to correlate patterns in DNA walks with known functional elements such as genes, promoters, and enhancers. Findings will be validated using experimental data and existing annotations from genomic databases.<br><br>In order to engage with the scientific community for feedback and collaboration, findings will be published on peer-reviewed journals that focus on bioinformatics, genomics, and computational biology, such as:<br>- IEEE/ACM Transactions on Computational Biology and Bioinformatics<br>- Bioinformatics (Oxford University Press)<br>- Pattern Recognition (Elsevier) |

| | |
|---|---|
| **Skills and competencies for the development of the activity** | The candidate should have strong programming skills in languages commonly used in bioinformatics (Python, R, and MATLAB), and experience with software development for creating custom analysis tools and scripts. A genomics and molecular biology background is preferable, with familiarity with genomic concepts, including DNA structure, gene function, and regulatory elements. Some data analysis and statistics knowledge are required, including handling large datasets and performing statistical tests. |