

Systematic study of inverted repeated sequences in prokaryotic genomes

Open PhD position in data science and computational biology

October 2023

1 PhD subject

DNA sequence motifs are short sequences found many times in a genome, often associated with a biological function. Amongst particular motifs, an inverted repeat (IR) is a nucleotide sequence followed downstream by its reverse complement (e.g., ATACGG followed by CCGTAT), potentially with a gap in the center (e.g., ATACGGnnnCGTAT). They are involved in many biological processes, including gene regulation, replication, translocation, and recombination. Thanks to the advent of high-throughput sequencing, the number of fully sequenced genomes has grown exponentially, paving the way for systematic studies of motifs. Yet, despite two types of IRs found in prokaryotes revolutionizing molecular biology (CRISPR-cas and restriction enzyme cutting sites), IRs have never been systematically studied in prokaryotic genomes.

In this project, we propose to **develop computational tools to systematically detect and characterize IRs in prokaryotic genomes.**

- Develop **computational and statistical tools to detect, analyze, and characterize IRs** in prokaryotic genomes (phylogenetic distributions, position along the genomes, presence in coding/non-coding regions, relationship to 3D structure. . .)
- Systematically apply these tools to the 30,000 complete prokaryotic genomes of the NCBI to **create a catalog of IRs in bacteria and archaea**;
- Create a **statistical learning model to infer putative functions** of IRs, and investigate a possible function of an IR previously found in Neisseriaceae genomes on bacterial cell physiology, gene expression, and genome folding.

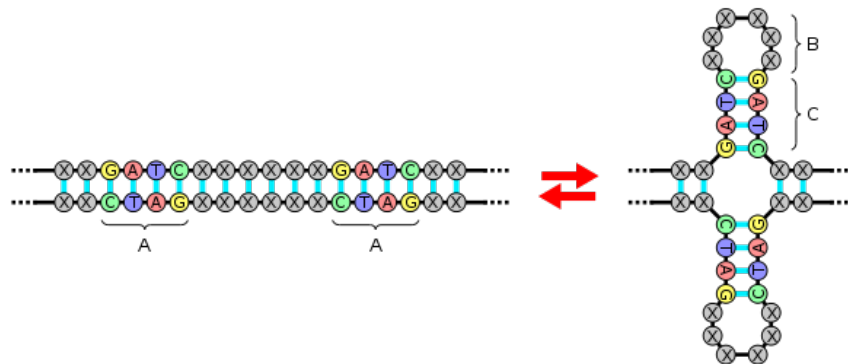


Figure 1: Inverted repeat changing to/from an extruded cruciform. A: Inverted Repeat Sequences; B: Loop; C: Stem with base pairing of the inverted repeat sequences – source : wikipedia

2 Scientific environment

TIMC is a joint CNRS/Grenoble Alpes University research unit that brings together scientists and clinicians through the development and use of computer science and computational biology to understand and control normal and pathological processes in biology and health. The unit is based in Grenoble, one of the largest student and academic cities in France, located near the Alps, La Tronche campus (Tram B).

The candidate will be hosted at the TIMC, in the TrEE team, supervised by Nelle Varoquaux and Ivan Junier. The TrEE team combines experimental and in silico interdisciplinary approaches to understand the evolutionary mechanisms of adaptation of microorganisms to their environment or to the host they inhabit or infect, and to develop biotechnological applications.

TrEE's computational group (compbio@tree) develops methods and tools to analyze, interpret and decipher biological data and processes, with a focus on the evolution of prokaryotes. To this purposes, it brings together expertise in biophysics, comparative genomics, and machine learning.

This project is part of a close collaboration with Virginia Lioy (CNRS, I2BC, Paris Saclay) and Silvia Bulgheresi (University of Vienna).

- The computational biology group website: <https://tree-timc.github.io/compbio>
- The team website: <http://www.timc.fr/en/tree>
- TIMC's website: <https://www-timc.imag.fr/en/>
- Website of the SIRIG project which funds this thesis: <https://tree-timc.github.io/grant-2023-SIRIG/>

This PhD is funded by an 80Prime CNRS grant. More information on <https://tree-timc.github.io/grant-2023-SIRIG/>

3 Profile

The profile sought is that of a graduate student (Master degree or equivalent) in Computer Science (Major in Artificial Intelligence, Data Science, or Bioinformatics), Applied Mathematics (Major in Signal Processing or Statistics), or Biology (with a strong computational biology background) who has a strong interest in interdisciplinary work in biology. They must have programming skills (R or Python) and be fluent in either French or English.

Applicants must apply through the CNRS website

References

- [1] E. D. Ladoukakis and A. Eyre-Walker. The excess of small inverted repeats in prokaryotes. *J Mol Evol*, 67(3):291–300, Sep 2008.
- [2] E. D. Ladoukakis and A. Eyre-Walker. Searching for sequence directed mutagenesis in eukaryotes. *J Mol Evol*, 64(1):1–3, Jan 2007.
- [3] L. Jia, Y. Li, F. Huang, Y. Jiang, H. Li, Z. Wang, T. Chen, J. Li, Z. Zhang, and W. Yao. LIRBase: a comprehensive database of long inverted repeats in eukaryotic genomes. *Nucleic Acids Res*, 50(D1):D174–D182, Jan 2022.
- [4] H. Alamro, M. Alzamel, C. S. Iliopoulos, S. P. Pissis, and S. Watts. IUPACpal: efficient identification of inverted repeats in IUPAC-encoded DNA sequences. *BMC Bioinformatics*, 22(1):51, Feb 2021.