Reg. No: 2017/27/B/ST6/02720; Principal Investigator: dr hab. Paweł Piotr Górecki

Consistent Models and Efficient Algorithms for Genomic Duplications

Abstract

The primary scientific objective of this project is to develop new methods for practitioners to study genomic duplication events. Our developed methods will be implemented as software tools in tight collaboration with biologists, computer scientists, and mathematicians. These tools will be made publicly available to the scientific community through our website. This project consists of three tasks whose objectives are addressing the development of efficient algorithms, large-scale computations, and biological applications. The following list presents the main scientific objectives for our proposal.

1. Develop biologically consistent models for genomic duplications.
2. Specify in practice effective computational problems based on the developed models.
3. Analyse the computational complexity of the problems, and design efficient algorithms addressing these problems.
4. Efficiently implement the developed algorithms with an easy-to-use interface for biologists.
5. Evaluate applicability and scalability of the developed tools.
6. Use the new tools to identify and study genomic duplication events for large-scale empirical datasets.

Topics proposed in this proposal are not only in the field of interest of computational biology and bioinformatics researchers, but also address serious biological and practical applications. In particular, genomic and whole-genome duplications have occurred for numerous species and, for example, had crucially impacted the evolution of crop plants especially significant for agriculture and industry.

The algorithmic developments will be partially based on our recently published methods. For the first time, these methods allowed solving complex instances of genomic duplications under various well-established models of genomic duplications. However, we observed, that these models are severely limited in practice. Our analysis of these limitations is a starting point to identify biologically more consistent models and corresponding efficient algorithms. Techniques to be employed will include standard mathematical proof techniques, graph theory, rule-based, hill-climbing, genetic algorithms, standard algorithmic design paradigms, run-time complexity analysis, fixed-parameter tractability, computational approximation and amortized analysis.

Our solutions will be implemented C++ and Python programming languages. We will develop pipelines for the data processing and validation of results. In these uniform environments, we will prepare a collection of empirical and simulated benchmark datasets for testing of our tools. Validation of results will be conducted in a tight collaboration with biological experts, by providing critical feedback the algorithmic design tasks.