

# Preface

This monograph discusses some of the most well-known string selection problems (SSP), offering a compendium of the current state-of-the-art methods presented in the literature. In particular, this work is intended as a general and comprehensive guide to understanding the very basic notions of mathematical optimization for sequence selection in biology, and a review of current research directions in this area; it aims to help bridge the gap between the computational and biological aspects of this class of problems.

SSP are generally modeled as optimization problems addressing the selection of strings with specific features from a large set of input sequences. In this context, the notion of “specific features” is captured by an objective function and we want to find strings that minimize or maximize it. Many problems in computational biology address the task of determining meaningful properties of biological systems, by comparing sequences, discovering their similarities and/or differences. From probe and primer design for diagnostics to protein structure prediction, from disease modeling to drug design, many steps of such design processes are characterized by discovering similarities or differences among sequences. However, one of the main challenges that arise when studying optimization methods for biological applications is the need of finding meaningful and high quality results, from both computational and biological points of view. Though theoretical results are usually desirable from a mathematical point of view, practical implications are more attractive from a biological perspective; this is a well-known issue in computational biology, where working at the edge of mathematics and biological sciences usually requires to find a trade-off between the biological findings and the *in silico* results.

Preliminarily, in Chap. 1 we provide a brief introduction to some biological concepts to understand the scenario where SSP arise; in Chap. 2 we introduce the basic notations for representing biological sequences, such as the Levenshtein distance and the Hamming distance, and we give an overview of biological applications that require the solution of SSP. Successively, we introduce some basic notions of mathematical optimization in Chap. 3, aiming at providing a minimal background to non-experts, in order to understand the basis of string selection methods, which are covered in Chap. 4. We introduce some of the most studied

SSP in computational biology and discuss some of the existing approaches, with an emphasis on their computational complexity. As such problems are mainly NP-hard, we put the focus on methods with theoretical guarantee on the quality of solutions and on methods with no theoretical proof but fast and effective in practice. In this perspective, several approaches can be considered, ranging from exact to heuristic methods; although exact methods guarantee to find a locally optimal solution, their computational burden makes them impractical when dealing with large datasets. Conversely, approximated methods guarantee a solution with a bounded distance from the optimum, while heuristics simply seek for a good solution without making any assumption on its optimality. Nevertheless, the latter two approaches allow to overcome the high computational cost required by exact methods and usually reach good performance even for large datasets.

At the end of our discussion on the state-of-the-art methods for string selection, and the comparison, when possible, of the approaches proposed in literature, we draw out the conclusions of this monograph. Many questions remain to be answered, both from theoretical and practical points of view, and many topics of interest remain open for further investigation.

## Acknowledgments

Panos M. Pardalos was partially supported by LATNA Laboratory, NRU HSE, RF government grant, ag. 11.G34.31.0057

Baltimore, MD

Elisa Pappalardo

Optimization Approaches for Solving String Selection Problems

Pappalardo, E.; Pardalos, P.; Stracquadanio, G.

2013, VIII, 49 p. 2 illus. in color., Softcover

ISBN: 978-1-4614-9052-4