Metaheuristics for String Problems in Bio-informatics

This book is dedicated to my parents Maria and Dieter, who currently pass through a difficult period of their lives.
(Christian Blum)

This book is dedicated to my daughters Iara, Mara, and Nara, the most beautiful gift that life could give me.
(Paola Festa)

Metaheuristics Set

coordinated by
Nicolas Monmarché and Patrick Siarry

Volume 6

Metaheuristics for String Problems in Bio-informatics

Christian Blum
Paola Festa

ISTE

WILEY

First published 2016 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Control Number: 2016945029

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-84821-812-3

Contents

Preface

DNA (deoxyribonucleic acid) acts as the information archive of most living beings. Due to the fact that a strand of DNA can be expressed as a set of four-letter character strings, so-called *string problems* have become abundant in bioinformatics and computational biology. Each year, new optimization problems dealing with DNA (or protein) sequences are being formulated that require efficient optimization techniques to arrive at solutions. From this perspective, bioinformatics is a burgeoning field for optimization experts and computer scientists in general. In this book, we will focus on a mixture of well-known and recent string optimization problems in the bioinformatics field. We will focus on problems that are combinatorial in nature.

One of the obstacles for optimization practitioners is the atypical nature of these problems, i.e. although combinatorial in nature, these problems are rather different to the classical traveling salesman problem or the quadratic assignment problem. This implies that a different type of expertise is required to efficiently solve many of these problems. Therefore, one of the main goals of this book is to pass on this kind of expertise and experience to newcomers to this field. The book provides several examples of very successful (hybrid) metaheuristics for solving specific string problems. One such example concerns the use of beam search (an incomplete branch and bound method) in solving longest common subsequence problems. The application of this algorithm in 2009 marked a breakthrough in the solution of this type of problem.

Finally, we would like to address a few words to the interested readers, especially biologists. We apologize for any imprecision in the description of biological processes, which we have tried to keep to a minimum. Keep in mind that, after all, we are only computer scientists and mathematicians.

Christian BLUM
Paola FESTA
June 2016

Acknowledgments

This work was supported by grant TIN2012-37930-C02-02 from the Spanish Government. Support from CSIC (Spanish National Research Council) and IKERBASQUE (Basque Foundation for Science) is also acknowledged. We thank RDlab[1], a BarcelonaTech facility, for allowing us to perform the experiments partly in the high-performance computing environment.

1 http://rdlab.lsi.upc.edu.

List of Acronyms

ACO	Ant Colony Optimization
B&B	Branch & Bound
CMSA	Construct, Merge, Solve & Adapt
CO	Combinatorial Optimization
DNA	Deoxyribonucleic Acid
DP	Dynamic Programming
EA	Evolutionary Algorithm
GA	Genetic Algorithm
ILP	Integer Linear Programming
ILS	Iterated Local Search
IP	Integer Programming
LNS	Large Neighborhood Search
MCSP	Minimum Common String Partition
MSFBC	Most Strings With Few Bad Columns
RNA	Ribonucleic Acid
SA	Simulated Annealing
TS	Tabu Search
TSP	Traveling Salesman Problem
UMCSP	Unbalanced Minimum Common String Partition