

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.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licenses issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned address:

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

© ISTE Ltd 2016

The rights of Christian Blum and Paola Festa to be identified as the author of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

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	ix
Acknowledgments	xi
List of Acronyms	xiii
Chapter 1. Introduction	1
1.1. Complete methods for combinatorial optimization	3
1.1.1. Linear programming relaxation	6
1.1.2. Cutting plane techniques	9
1.1.3. General-purpose ILP solvers	18
1.1.4. Dynamic programming	19
1.2. Approximate methods: metaheuristics	20
1.2.1. Ant colony optimization	22
1.2.2. Evolutionary algorithms	24
1.2.3. Greedy randomized adaptive search procedures	25
1.2.4. Iterated local search	26
1.2.5. Simulated annealing	27
1.2.6. Other metaheuristics	29
1.2.7. Hybrid approaches	29
1.3. Outline of the book	32

Chapter 2. Minimum Common String Partition Problem	37
2.1. The MCSP problem	38
2.1.1. Technical description of the UMCSP problem	38
2.1.2. Literature review	39
2.1.3. Organization of this chapter	40
2.2. An ILP model for the UMCSP problem	40
2.3. Greedy approach	42
2.4. Construct, merge, solve and adapt	42
2.5. Experimental evaluation	45
2.5.1. Benchmarks	46
2.5.2. Tuning CMSA	46
2.5.3. Results	47
2.6. Future work	54
Chapter 3. Longest Common Subsequence Problems	55
3.1. Introduction	56
3.1.1. LCS problems	56
3.1.2. ILP models for LCS and RFLCS problems	59
3.1.3. Organization of this chapter	61
3.2. Algorithms for the LCS problem	61
3.2.1. Beam search	61
3.2.2. Upper bound	64
3.2.3. Beam search framework	64
3.2.4. Beam-ACO	67
3.2.5. Experimental evaluation	71
3.3. Algorithms for the RFLCS problem	75
3.3.1. CMSA	77
3.3.2. Experimental evaluation	80
3.4. Future work	85

Chapter 4. The Most Strings With Few Bad Columns Problem	87
4.1. The MSFBC problem	88
4.1.1. Literature review	88
4.2. An ILP model for the MSFBC problem	89
4.3. Heuristic approaches	90
4.3.1. Frequency-based greedy	91
4.3.2. Truncated pilot method	91
4.4. ILP-based large neighborhood search	92
4.5. Experimental evaluation	94
4.5.1. Benchmarks	94
4.5.2. Tuning of LNS	96
4.5.3. Results	97
4.6. Future work	104
Chapter 5. Consensus String Problems	107
5.1. Introduction	107
5.1.1. Creating diagnostic probes for bacterial infections	108
5.1.2. Primer design	108
5.1.3. Discovering potential drug targets	108
5.1.4. Motif search	109
5.2. Organization of this chapter	110
5.3. The closest string problem and the close to most string problem	110
5.3.1. ILP models for the CSP and the CTMSP	111
5.3.2. Literature review	112
5.3.3. Exact approaches for the CSP	113
5.3.4. Approximation algorithms for the CSP	113
5.3.5. Heuristics and metaheuristics for the CSP	114
5.4. The farthest string problem and the far from most string problem	117
5.4.1. ILP models for the FSP and the FFMSP	117
5.4.2. Literature review	118
5.4.3. Heuristics and metaheuristics for the FFMSP	119
5.5. An ILP-based heuristic	141
5.6. Future work	146

Chapter 6. Alignment Problems	149
6.1. Introduction	149
6.1.1. Organization of this chapter	150
6.2. The pairwise alignment problem	151
6.2.1. Smith and Waterman’s algorithm	154
6.3. The multiple alignment problem	157
6.3.1. Heuristics for the multiple alignment problem	161
6.3.2. Metaheuristics for the multiple alignment problem	162
6.4. Conclusion and future work	173
Chapter 7. Conclusions	175
7.1. DNA sequencing	175
7.1.1. DNA fragment assembly	176
7.1.2. DNA sequencing by hybridization	177
7.2. Founder sequence reconstruction	180
7.2.1. The FSRP problem	181
7.2.2. Existing heuristics and metaheuristics	182
7.3. Final remarks	184
Bibliography	187
Index	205

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 RDlab¹, 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