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
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>ix</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>xi</td>
</tr>
<tr>
<td>List of Acronyms</td>
<td>xiii</td>
</tr>
<tr>
<td>Chapter 1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1. Complete methods for combinatorial</td>
<td></td>
</tr>
<tr>
<td>optimization</td>
<td>3</td>
</tr>
<tr>
<td>1.1.1. Linear programming relaxation</td>
<td>6</td>
</tr>
<tr>
<td>1.1.2. Cutting plane techniques</td>
<td>9</td>
</tr>
<tr>
<td>1.1.3. General-purpose ILP solvers</td>
<td>18</td>
</tr>
<tr>
<td>1.1.4. Dynamic programming</td>
<td>19</td>
</tr>
<tr>
<td>1.2. Approximate methods: metaheuristics</td>
<td>20</td>
</tr>
<tr>
<td>1.2.1. Ant colony optimization</td>
<td>22</td>
</tr>
<tr>
<td>1.2.2. Evolutionary algorithms</td>
<td>24</td>
</tr>
<tr>
<td>1.2.3. Greedy randomized adaptive search</td>
<td>25</td>
</tr>
<tr>
<td>procedures</td>
<td></td>
</tr>
<tr>
<td>1.2.4. Iterated local search</td>
<td>26</td>
</tr>
<tr>
<td>1.2.5. Simulated annealing</td>
<td>27</td>
</tr>
<tr>
<td>1.2.6. Other metaheuristics</td>
<td>29</td>
</tr>
<tr>
<td>1.2.7. Hybrid approaches</td>
<td>29</td>
</tr>
<tr>
<td>1.3. Outline of the book</td>
<td>32</td>
</tr>
</tbody>
</table>
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
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\textsuperscript{1}, a BarcelonaTech facility, for allowing us to perform the experiments partly in the high-performance computing environment.

\textsuperscript{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