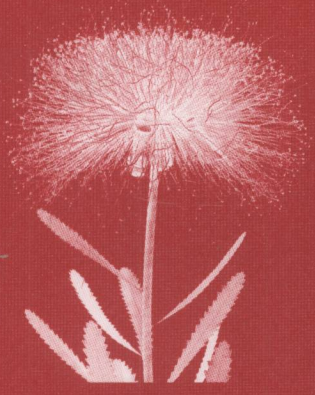
Carlos Cotta
Jano van Hemert (Eds.)

Evolutionary Computation in Combinatorial Optimization

LNCS 4446

7th European Conference, EvoCOP 2007
Valencia, Spain, April 2007
Proceedings



Springer

Carlos Cotta Jano van Hemert (Eds.)

Evolutionary Computation in Combinatorial Optimization

7th European Conference, EvoCOP 2007
Valencia, Spain, April 11-13, 2007
Proceedings

Springer

E2007003201

Volume Editors

Carlos Cotta
Universidad de Málaga, Dept. Lenguajes y Ciencias de la Computación
ETSI Informática, Campus Teatinos, 29071 Málaga, Spain
E-mail: ccottap@lcc.uma.es

Jano van Hemert
University of Edinburgh, National e-Science Institute
15 South College Street, Edinburgh EH8 9AA, UK
E-mail: jano@vanhemert.co.uk

Cover illustration: Morphogenesis series #12 by Jon McCormack, 2006

Library of Congress Control Number: 2007923599

CR Subject Classification (1998): F.1, F.2, G.1.6, G.2.1, G.1

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN 0302-9743
ISBN-10 3-540-71614-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-71614-3 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12041480 06/3180 5 4 3 2 1 0

Lecture Notes in Computer Science 4446

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Preface

Metaheuristics have often been shown to be effective for difficult combinatorial optimization problems appearing in various industrial, economical, and scientific domains. Prominent examples of metaheuristics are evolutionary algorithms, simulated annealing, tabu search, scatter search, memetic algorithms, variable neighborhood search, iterated local search, greedy randomized adaptive search procedures, estimation of distribution algorithms, and ant colony optimization. Successfully solved problems include scheduling, timetabling, network design, transportation and distribution, vehicle routing, the traveling salesman problem, satisfiability, packing and cutting, and general mixed integer programming.

EvoCOP began in 2001 and has been held annually since then. It was the first event specifically dedicated to the application of evolutionary computation and related methods to combinatorial optimization problems. Originally held as a workshop, EvoCOP became a conference in 2004. The events gave researchers an excellent opportunity to present their latest research and to discuss current developments and applications as well as providing for improved interaction between members of this scientific community. Following the general trend of hybrid metaheuristics and diminishing boundaries between the different classes of metaheuristics, EvoCOP has broadened its scope over the last years and invited submissions on any kind of metaheuristic for combinatorial optimization.

This volume contains the proceedings of EvoCOP 2007, the seventh European Conference on Evolutionary Computation in Combinatorial Optimization. It was held in Valencia, Spain, April 11–13, 2007, jointly with EuroGP 2007, the Tenth European Conference on Genetic Programming, EvoBIO 2007, the Fifth European Conference on Evolutionary Computation and Machine Learning in Bioinformatics, and EvoWorkshops 2007, which consisted of the following seven individual workshops: EvoCOMNET, the Fourth European Workshop on the Application of Nature-Inspired Techniques to Telecommunication Networks and Other Connected Systems; EvoFIN, the First European Workshop on Evolutionary Computation in Finance and Economics; EvoIASP, the Ninth European Workshop on Evolutionary Computation in Image Analysis and Signal Processing; EvoInteraction, the Second European Workshop on Interactive Evolution and Humanized Computational Intelligence; EvoMUSART, the Fifth European Workshop on Evolutionary Music and Art; EvoSTOC, the Fourth European Workshop on Evolutionary Algorithms in Stochastic and Dynamic Environments, and EvoTransLog, the First European Workshop on Evolutionary Computation in Transportation and Logistics. Since 2007, all these events are grouped under the collective name EvoStar, and constitute Europe's premier co-located meetings on evolutionary computation.

Accepted papers of previous EvoCOP editions were published by Springer in the series *Lecture Notes in Computer Science* (LNCS – Volumes 2037, 2279, 2611, 3004, 3448, and 3906).

EvoCOP	Submitted	Accepted	Acceptance ratio
2001	31	23	74.2%
2002	32	18	56.3%
2003	39	19	48.7%
2004	86	23	26.7%
2005	66	24	36.4%
2006	77	24	31.2%
2007	81	21	25.9%

The rigorous, double-blind reviewing process of EvoCOP 2007 resulted in a strong selection among the submitted papers; the acceptance rate was 25.9%. Each paper was reviewed by at least three members of the International Program Committee. All accepted papers were presented orally at the conference and are included in this proceedings volume. We would like to credit the members of our Program Committee, to whom we are very grateful for their quick and thorough work and the valuable advice on how to improve papers for the final publication. EvoCOP 2007 contributions deal with representations, heuristics, analysis of problem structures, and comparisons of algorithms. The list of studied combinatorial optimization problems includes prominent examples like graph coloring, knapsack problems, the traveling salesperson problem, scheduling, as well as specific real-world problems.

We would like to express our sincere gratitude to the internationally renowned invited speakers who gave the keynote talks at the conference: Ricard V. Solé, head of the Complex Systems Lab at the University Pompeu Fabra, Chris Adami, head of the Digital Life Lab at the California Institute of Technology, and Alan Bundy, from the Centre for Intelligent Systems and their Applications, School of Informatics at the University of Edinburgh.

The success of the conference resulted from the input of many people, to whom we would like to express our appreciation. We thank Marc Schoenauer for providing the Web-based conference management system. The local organizers, led by Anna Isabel Esparcia-Alcázar, did an extraordinary job for which we are very grateful. We thank the Universidad Politécnica de Valencia, Spain, for their institutional and financial support and for providing premises and administrative assistance, the Instituto Tecnológico de Informática in Valencia for cooperation and help with local arrangements, and the Spanish Ministerio de Educación y Ciencia for their financial support. Thanks are also due to Jennifer Willies and the Centre for Emergent Computing at Napier University in Edinburgh, Scotland, for administrative support and event coordination. Last, but not least, we would especially like to thank Jens Gottlieb and Günther Raidl for their support and guidance, to whom we owe a lot. From their hard work and dedication, EvoCOP 2007 has now become one of the reference events in evolutionary computation.

April 2007

Carlos Cotta
Jano van Hemert

Organization

EvoCOP 2007 was organized jointly with EuroGP 2007, EvoBIO 2007, and EvoWorkshops 2007.

Organizing Committee

Chairs	Carlos Cotta, Universidad de Málaga, Spain,
	Jano van Hemert, University of Edinburgh, UK
Local Chair	Anna Isabel Esparcia-Alcázar, Universidad Politécnica de Valencia, Spain
Publicity Chair	Leonardo Vanneschi, University of Milano-Bicocca, Italy

EvoCOP Steering Committee

Carlos Cotta, Universidad de Málaga, Spain,

Jens Gottlieb, SAP AG, Germany,

Jano van Hemert, University of Edinburgh, UK,

Günther Raidl, Vienna University of Technology, Austria

Program Committee

Adnan Acan, Eastern Mediterranean University, Turkey
Hernán Aguirre, Shinshu University, Japan
Enrique Alba, Universidad de Málaga, Spain
Mehmet Emin Aydin, London South Bank University, UK
Ruibin Bai, University of Nottingham, UK
Christian Bierwirth, University of Bremen, Germany
Christian Blum, Universitat Politècnica de Catalunya, Spain
Peter Brucker, University of Osnabrück, Germany
Edmund Burke, University of Nottingham, UK
Pedro Castillo, Universidad de Granada, Spain
David W. Corne, Heriot-Watt University, UK
Ernesto Costa, University of Coimbra, Portugal
Carlos Cotta, Universidad de Málaga, Spain
Peter I. Cowling, University of Bradford, UK
Bart Craenen, Napier University, Edinburgh, UK
Keshav Dahal, University of Bradford, UK
David Davis, NuTech Solutions Inc., USA
Karl F. Dörner, University of Vienna, Austria

Anton V. Eremeev, Omsk Branch of Sobolev Institute of Mathematics, Russia
Jeroen Eggermont, University of Leiden, The Netherlands
Antonio J. Fernández, Universidad de Málaga, Spain
Francisco Fernández de Vega, Universidad de Extremadura, Spain
David B. Fogel, Natural Selection, Inc., USA
Bernd Freisleben, University of Marburg, Germany
José E. Gallardo, Universidad de Málaga, Spain
Michel Gendreau, Université de Montréal, Canada
Jens Gottlieb, SAP AG, Germany
Joern Grahl, University of Mannheim, Germany
Walter Gutjahr, University of Vienna, Austria
Jin-Kao Hao, University of Angers, France
Emma Hart, Napier University, Edinburgh, UK
Richard F. Hartl, University of Vienna, Austria
Geir Hasle, SINTEF, Norway
Jano van Hemert, University of Edinburgh, UK
Juhos István, University of Szeged, Hungary
Bryant A. Julstrom, St. Cloud State University, USA
Graham Kendall, University of Nottingham, UK
Joshua D. Knowles, University of Manchester, UK
Gary A. Kochenberger, University of Colorado, USA
Mario Köppen, Fraunhofer IPK, Germany
Jozef J. Kratica, Serbian Academy of Sciences and Arts, Serbia and Montenegro
Rhyd Lewis, Cardiff University, UK
Andrea Lodi, University of Bologna, Italy
José Antonio Lozano, University of the Basque Country, Spain
Arne Løkketangen, Molde University College, Norway
Vittorio Maniezzo, University of Bologna, Italy
Dirk C. Mattfeld, Technical University Braunschweig, Germany
Helmut A. Mayer, University of Salzburg, Austria
Barry McCollum, Queen's University Belfast, UK
Juan Julián Merelo-Guervós, Universidad de Granada, Spain
Daniel Merkle, University of Leipzig, Germany
Peter Merz, University of Kaiserslautern, Germany
Martin Middendorf, University of Leipzig, Germany
Pablo Moscato, The University of Newcastle, Australia
Christine L. Mumford, Cardiff University, UK
Francisco J. B. Pereira, University of Coimbra, Portugal
Adam Prugel-Bennett, University of Southampton, UK
Jakob Puchinger, University of Melbourne, Australia
Günther R. Raidl, Vienna University of Technology, Austria
Marcus C. Randall, Bond University, Australia
Colin Reeves, Coventry University, UK
Marc Reimann, ETH Zurich, Switzerland
Andrea Roli, University of Bologna, Italy

Roberto Santana, University of the Basque Country, Spain
Marc Schoenauer, INRIA, France
Marc Sevaux, University of South Brittany, France
Christine Solnon, University Lyon I, France
Giovanni Squillero, Politecnico di Torino, Italy
Thomas Stützle, Université Libre de Bruxelles, Belgium
El-ghazali Talbi, INRIA Futurs – Lille, France
Jorge Tavares, University of Coimbra, Portugal
Stefan Voß, University of Hamburg, Germany
Ingo Wegener, University of Dortmund, Germany
Fatos Xhafa, Universitat Politècnica de Catalunya, Spain
Takeshi Yamada, NTT Communication Science Laboratories, Japan

Lecture Notes in Computer Science

For information about Vols. 1–4333

please contact your bookseller or Springer

Vol. 4446: C. Cotta, J. van Hemert (Eds.), Evolutionary Computation in Combinatorial Optimization. XII, 241 pages. 2007.

Vol. 4444: T. Reps, M. Sagiv, J. Bauer (Eds.), Program Analysis and Compilation, Theory and Practice. X, 361 pages. 2007.

Vol. 4430: C.C. Yang, D. Zeng, M. Chau, K. Chang, Q. Yang, X. Cheng, J. Wang, F.-Y. Wang, H. Chen (Eds.), Intelligence and Security Informatics. XII, 330 pages. 2007.

Vol. 4429: R. Lu, J.H. Siekmann, C. Ullrich (Eds.), Cognitive Systems. X, 161 pages. 2007. (Sublibrary LNAI).

Vol. 4427: S. Uhlig, K. Papagiannaki, O. Bonaventure (Eds.), Passive and Active Network Measurement. XI, 274 pages. 2007.

Vol. 4425: G. Amati, C. Carpineto, G. Romano (Eds.), Advances in Information Retrieval. XIX, 759 pages. 2007.

Vol. 4424: O. Grumberg, M. Huth (Eds.), Tools and Algorithms for the Construction and Analysis of Systems. XX, 738 pages. 2007.

Vol. 4423: H. Seidl (Ed.), Foundations of Software Science and Computational Structures. XVI, 379 pages. 2007.

Vol. 4422: M.B. Dwyer, A. Lopes (Eds.), Fundamental Approaches to Software Engineering. XV, 440 pages. 2007.

Vol. 4421: R. De Nicola (Ed.), Programming Languages and Systems. XVII, 538 pages. 2007.

Vol. 4420: S. Krishnamurthi, M. Odersky (Eds.), Compiler Construction. XIV, 233 pages. 2007.

Vol. 4419: P.C. Diniz, E. Marques, K. Bertels, M.M. Fernandes, J.M.P. Cardoso (Eds.), Reconfigurable Computing: Architectures, Tools and Applications. XIV, 391 pages. 2007.

Vol. 4418: A. Gagalowicz, W. Philips (Eds.), Computer Vision/Computer Graphics Collaboration Techniques. XV, 620 pages. 2007.

Vol. 4416: A. Bemporad, A. Bicchi, G. Buttazzo (Eds.), Hybrid Systems: Computation and Control. XVII, 797 pages. 2007.

Vol. 4415: P. Lukowicz, L. Thiele, G. Tröster (Eds.), Architecture of Computing Systems - ARCS 2007. X, 297 pages. 2007.

Vol. 4414: S. Hochreiter, R. Wagner (Eds.), Bioinformatics Research and Development. XVI, 482 pages. 2007. (Sublibrary LNBI).

Vol. 4410: A. Branco (Ed.), Anaphora: Analysis, Algorithms and Applications. X, 191 pages. 2007. (Sublibrary LNAI).

Vol. 4407: G. Puebla (Ed.), Logic-Based Program Synthesis and Transformation. VIII, 237 pages. 2007.

Vol. 4405: L. Padgham, F. Zambonelli (Eds.), Agent-Oriented Software Engineering VII. XII, 225 pages. 2007.

Vol. 4403: S. Obayashi, K. Deb, C. Poloni, T. Hiroyasu, T. Murata (Eds.), Evolutionary Multi-Criterion Optimization. XIX, 954 pages. 2007.

Vol. 4400: J.F. Peters, A. Skowron, V.W. Marek, E. Orłowska, R. Slowinski, W. Ziarko (Eds.), Transactions on Rough Sets VII, Part II. X, 381 pages. 2007.

Vol. 4399: X. Llorà, T. Kovacs, K. Takadama, P.L. Lanzi, S.W. Wilson, W. Stolzmann (Eds.), Learning Classifier Systems. XII, 345 pages. 2007. (Sublibrary LNAI).

Vol. 4398: S. Marchand-Maillet, E. Bruno, A. Nürnberger, M. Detyniecki (Eds.), Adaptive Multimedia Retrieval: User, Context, and Feedback. XI, 269 pages. 2007.

Vol. 4397: C. Stephanidis, M. Pieper (Eds.), Universal Access in Ambient Intelligence Environments. XV, 467 pages. 2007.

Vol. 4396: J. García-Vidal, L. Cerdà-Alabern (Eds.), Wireless Systems and Mobility in Next Generation Internet. IX, 271 pages. 2007.

Vol. 4395: M. Daydé, J.M.L.M. Palma, Á.L.G.A. Coutinho, E. Pacitti, J.C. Lopes (Eds.), High Performance Computing for Computational Science - VECPAR 2006. XXIV, 721 pages. 2007.

Vol. 4394: A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing. XVI, 648 pages. 2007.

Vol. 4393: W. Thomas, P. Weil (Eds.), STACS 2007. XVIII, 708 pages. 2007.

Vol. 4392: S.P. Vadhan (Ed.), Theory of Cryptography. XI, 595 pages. 2007.

Vol. 4391: Y. Stylianou, M. Faundez-Zanuy, A. Esposito (Eds.), Progress in Nonlinear Speech Processing. XII, 269 pages. 2007.

Vol. 4390: S.O. Kuznetsov, S. Schmidt (Eds.), Formal Concept Analysis. X, 329 pages. 2007. (Sublibrary LNAI).

Vol. 4389: D. Weyns, H.V.D. Parunak, F. Michel (Eds.), Environments for Multi-Agent Systems III. X, 273 pages. 2007. (Sublibrary LNAI).

Vol. 4385: K. Coninx, K. Luyten, K.A. Schneider (Eds.), Task Models and Diagrams for Users Interface Design. XI, 355 pages. 2007.

Vol. 4384: T. Washio, K. Satoh, H. Takeda, A. Inokuchi (Eds.), New Frontiers in Artificial Intelligence. IX, 401 pages. 2007. (Sublibrary LNAI).

Vol. 4383: E. Bin, A. Ziv, S. Ur (Eds.), Hardware and Software, Verification and Testing. XII, 235 pages. 2007.

Vol. 4381: J. Akiyama, W.Y.C. Chen, M. Kano, X. Li, Q. Yu (Eds.), Discrete Geometry, Combinatorics and Graph Theory. XI, 289 pages. 2007.

Vol. 4380: S. Spaccapietra, P. Atzeni, F. Fages, M.-S. Hacid, M. Kifer, J. Mylopoulos, B. Pernici, P. Shvaiko, J. Trujillo, I. Zaihrayeu (Eds.), Journal on Data Semantics VIII. XV, 219 pages. 2007.

Vol. 4378: I. Virbitskaite, A. Voronkov (Eds.), Perspectives of Systems Informatics. XIV, 496 pages. 2007.

Vol. 4377: M. Abe (Ed.), Topics in Cryptology – CT-RSA 2007. XI, 403 pages. 2006.

Vol. 4376: E. Frachtenberg, U. Schwiegelshohn (Eds.), Job Scheduling Strategies for Parallel Processing. VII, 257 pages. 2007.

Vol. 4374: J.F. Peters, A. Skowron, I. Düntsch, J. Grzymała-Busse, E. Orłowska, L. Polkowski (Eds.), Transactions on Rough Sets VI, Part I. XII, 499 pages. 2007.

Vol. 4373: K. Langendoen, T. Voigt (Eds.), Wireless Sensor Networks. XIII, 358 pages. 2007.

Vol. 4372: M. Kaufmann, D. Wagner (Eds.), Graph Drawing. XIV, 454 pages. 2007.

Vol. 4371: K. Inoue, K. Satoh, F. Toni (Eds.), Computational Logic in Multi-Agent Systems. X, 315 pages. 2007. (Sublibrary LNAI).

Vol. 4370: P.P Lévy, B. Le Grand, F. Poulet, M. Soto, L. Darago, L. Toubiana, J.-F. Vibert (Eds.), Pixelization Paradigm. XV, 279 pages. 2007.

Vol. 4369: M. Umeda, A. Wolf, O. Bartenstein, U. Geske, D. Seipel, O. Takata (Eds.), Declarative Programming for Knowledge Management. X, 229 pages. 2006. (Sublibrary LNAI).

Vol. 4368: T. Erlebach, C. Kaklamanis (Eds.), Approximation and Online Algorithms. X, 345 pages. 2007.

Vol. 4367: K. De Bosschere, D. Kaeli, P. Stenström, D. Whalley, T. Ungerer (Eds.), High Performance Embedded Architectures and Compilers. XI, 307 pages. 2007.

Vol. 4366: K. Tuyls, R. Westra, Y. Saeys, A. Nowé (Eds.), Knowledge Discovery and Emergent Complexity in Bioinformatics. IX, 183 pages. 2007. (Sublibrary LNBI).

Vol. 4364: T. Kühne (Ed.), Models in Software Engineering. XI, 332 pages. 2007.

Vol. 4362: J. van Leeuwen, G.F. Italiano, W. van der Hoek, C. Meinel, H. Sack, F. Plášil (Eds.), SOFSEM 2007: Theory and Practice of Computer Science. XXI, 937 pages. 2007.

Vol. 4361: H.J. Hoogeboom, G. Păun, G. Rozenberg, A. Salomaa (Eds.), Membrane Computing. IX, 555 pages. 2006.

Vol. 4360: W. Dubitzky, A. Schuster, P.M.A. Sloot, M. Schroeder, M. Romberg (Eds.), Distributed, High-Performance and Grid Computing in Computational Biology. X, 192 pages. 2007. (Sublibrary LNBI).

Vol. 4358: R. Vidal, A. Heyden, Y. Ma (Eds.), Dynamical Vision. IX, 329 pages. 2007.

Vol. 4357: L. Buttyán, V. Gligor, D. Westhoff (Eds.), Security and Privacy in Ad-Hoc and Sensor Networks. X, 193 pages. 2006.

Vol. 4355: J. Julliand, O. Kouchnarenko (Eds.), B 2007: Formal Specification and Development in B. XIII, 293 pages. 2006.

Vol. 4354: M. Hanus (Ed.), Practical Aspects of Declarative Languages. X, 335 pages. 2006.

Vol. 4353: T. Schwentick, D. Suciu (Eds.), Database Theory – ICDT 2007. XI, 419 pages. 2006.

Vol. 4352: T.-J. Cham, J. Cai, C. Dorai, D. Rajan, T.-S. Chua, L.-T. Chia (Eds.), Advances in Multimedia Modeling, Part II. XVIII, 743 pages. 2006.

Vol. 4351: T.-J. Cham, J. Cai, C. Dorai, D. Rajan, T.-S. Chua, L.-T. Chia (Eds.), Advances in Multimedia Modeling, Part I. XIX, 797 pages. 2006.

Vol. 4349: B. Cook, A. Podelski (Eds.), Verification, Model Checking, and Abstract Interpretation. XI, 395 pages. 2007.

Vol. 4348: S.T. Taft, R.A. Duff, R.L. Brukardt, E. Ploedereder, P. Leroy (Eds.), Ada 2005 Reference Manual. XXII, 765 pages. 2006.

Vol. 4347: J. Lopez (Ed.), Critical Information Infrastructures Security. X, 286 pages. 2006.

Vol. 4346: L. Brim, B. Haverkort, M. Leucker, J. van de Pol (Eds.), Formal Methods: Applications and Technology. X, 363 pages. 2007.

Vol. 4345: N. Maglaveras, I. Chouvarda, V. Koutkias, R. Brause (Eds.), Biological and Medical Data Analysis. XIII, 496 pages. 2006. (Sublibrary LNBI).

Vol. 4344: V. Gruhn, F. Oquendo (Eds.), Software Architecture. X, 245 pages. 2006.

Vol. 4342: H. de Swart, E. Orłowska, G. Schmidt, M. Roubens (Eds.), Theory and Applications of Relational Structures as Knowledge Instruments II. X, 373 pages. 2006. (Sublibrary LNAI).

Vol. 4341: P.Q. Nguyen (Ed.), Progress in Cryptology - VIETCRYPT 2006. XI, 385 pages. 2006.

Vol. 4340: R. Prodan, T. Fahringer, Grid Computing. XXIII, 317 pages. 2007.

Vol. 4339: E. Ayguadé, G. Baumgartner, J. Ramanujam, P. Sadayappan (Eds.), Languages and Compilers for Parallel Computing. XI, 476 pages. 2006.

Vol. 4338: P. Kalra, S. Peleg (Eds.), Computer Vision, Graphics and Image Processing. XV, 965 pages. 2006.

Vol. 4337: S. Arun-Kumar, N. Garg (Eds.), FSTTCS 2006: Foundations of Software Technology and Theoretical Computer Science. XIII, 430 pages. 2006.

Vol. 4336: V.R. Basili, D. Rombach, K. Schneider, B. Kitchenham, D. Pfahl, R.W. Selby, Empirical Software Engineering Issues. XVII, 193 pages. 2007.

Vol. 4335: S.A. Brueckner, S. Hassas, M. Jelasity, D. Yamins (Eds.), Engineering Self-Organising Systems. XII, 212 pages. 2007. (Sublibrary LNAI).

Vol. 4334: B. Beckert, R. Hähnle, P.H. Schmitt (Eds.), Verification of Object-Oriented Software. XXIX, 658 pages. 2007. (Sublibrary LNAI).

Table of Contents

A New Local Search Algorithm for the DNA Fragment Assembly Problem

Enrique Alba and Gabriel Luque

Grupo GISUM, Departamento de LCC
E.T.S.I. Informática
Campus Teatinos, 29071 Málaga (Spain)
{eat,gabriel}@lcc.uma.es

Abstract. In this paper we propose and study the behavior of a new heuristic algorithm for the DNA fragment assembly problem: PALS. The DNA fragment assembly is a problem to be solved in the early phases of the genome project and thus is very important since the other steps depend on its accuracy. This is an NP-hard combinatorial optimization problem which is growing in importance and complexity as more research centers become involved on sequencing new genomes. Various heuristics, including genetic algorithms, have been designed for solving the fragment assembly problem, but since this problem is a crucial part of any sequencing project, better assemblers are needed. Our proposal is a very efficient assembler that allows to find optimal solutions for large instances of this problem, considerably faster than its competitors and with high accuracy.

1 Introduction

With the advance of computational science, bioinformatics has become more and more attractive to researchers in the field of computational biology. Genomic data analysis using computational approaches is very popular as well. The primary goal of a genomic project is to determine the complete sequence of the genome and its genetic contents. Thus, a genome project is accomplished in two steps, the first one is the genome sequencing and the second one is the genome annotation (i.e., the process of identifying the boundaries between genes and other features in raw DNA sequence).

In this paper, we focus on the genome sequencing, which is also known as the DNA fragment assembly problem. The fragment assembly occurs in the very beginning of the process and therefor other steps depend on its accuracy. At present, DNA sequences that are longer than 600 base-pairs (bps) cannot routinely be sequenced accurately. For example, human DNA is about 3.2 billion nucleotides in length and cannot be read at once. Hence, large strands of DNA need to be broken into small fragments for sequencing in a process called *shotgun sequencing*. In this approach, several copies of a portion of DNA are each broken into many segments short enough to be sequenced automatically by machine. But this process does not keep neither the ordering of the fragments nor the

C. Cotta and J. van Hemert (Eds.): EvoCOP 2007, LNCS 4446, pp. 1–12, 2007.

portion from which a particular fragment came. This leads to the DNA fragment assembly problem [1] in which these short sequences have to be reassembled to their (supposed) original form. The automation allows shotgun sequencing to proceed far faster than traditional methods. But comparing all the tiny pieces and matching up the overlaps requires massive computation.

The assembly problem is therefore a combinatorial optimization problem that, even in the absence of noise, is NP-hard: given k fragments, there are $2^k k!$ possible combinations. Over the past decade a number of fragment assembly packages have been developed and used to sequence different organisms. The most popular packages are PHRAP [2], TIGR assembler [3], STROLL [4], CAP3 [5], Celera assembler [6], and EULER [7]. These packages deal with the previously described challenges to different extents, but none of them solves all of them. Each package automates fragment assembly using a variety of algorithms. The most popular techniques are greedy-based while other approaches have tackled the problem with metaheuristics [8]. This work reports on the design and implementation of a new problem aware local search algorithm to find fast and accurate solutions for large instances of the DNA fragment assembly problem. We additionally study the behavior of several variants of the basic method. Finally, we also compare the results of our approach with the ones of classical (real world) assemblers in order to test the actual interest of our method.

The remainder of this paper is organized as follows. In the next section, we present background information about the DNA fragment assembly problem. In Section 3, the details of our proposed heuristic are presented. We analyze the results of our experiments in Section 4. Finally, we end this paper by giving our final thoughts and conclusions in Section 5.

2 The DNA Fragment Assembly Problem

In order to determine the function of specific genes, scientists have learned to read the sequence of nucleotides comprising a DNA sequence in a process called DNA sequencing. To do that, multiple exact copies of the original DNA sequence are made. Each copy is then cut into short fragments at random positions. These are the first three steps depicted in Fig. 1 and they take place in the laboratory. After the fragment set is obtained, a traditional assemble approach is followed in this order: overlap, layout, and then consensus. To ensure that enough fragments overlap, the reading of fragments continues until a coverage is satisfied. These steps are the last three ones in Fig. 1. In what follows, we give a brief description of each of the three phases, namely overlap, layout, and consensus.

Overlap Phase - Finding the overlapping fragments. This phase consists of finding the best or longest match between the suffix of one sequence and the prefix of another. In this step, we compare all possible pairs of fragments to determine their similarity. Usually, a dynamic programming algorithm applied to semiglobal alignment is used in this step. The intuition behind finding the pairwise overlap is that fragments with a significant overlap score are very likely next to each other in the target sequence.

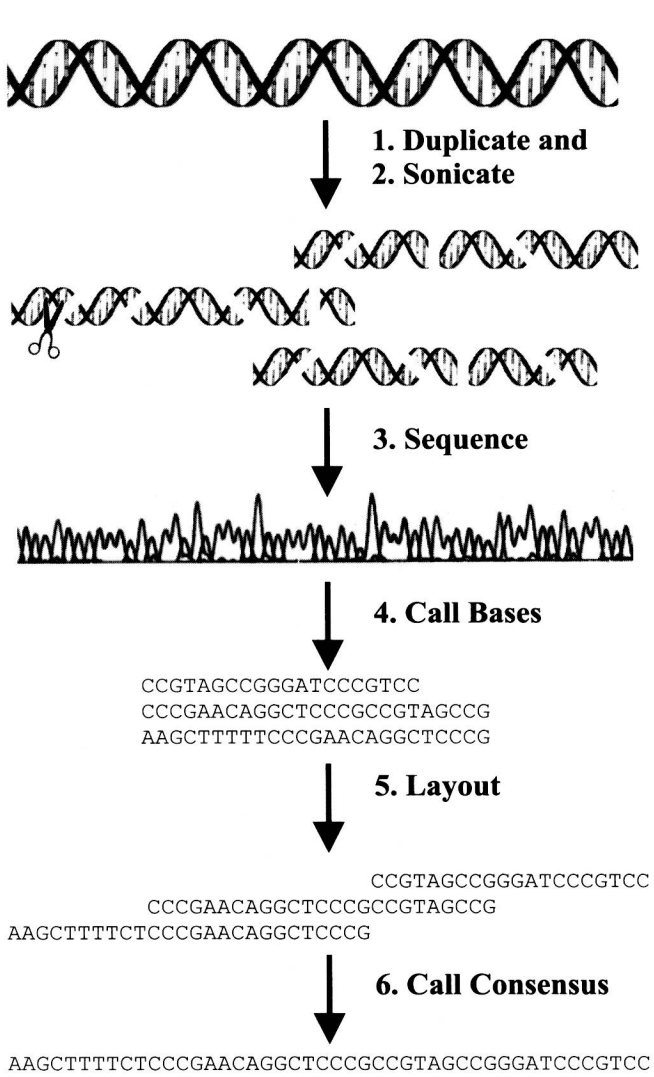


Fig. 1. Graphical representation of DNA sequencing and assembly

Layout Phase - Finding the order of fragments based on the computed similarity score. This is the most difficult step because it is hard to tell the true overlap due to the following challenges:

1. Unknown orientation: After the original sequence is cut into many fragments, the orientation is lost. One does not know which strand should be selected. If one fragment does not have any overlap with another, it is still possible that its reverse complement might have such an overlap.

2. Base call errors: There are three types of base call errors: substitution, insertion, and deletion errors. They occur due to experimental errors in the electrophoresis procedure (the method used in the laboratories to read the ADN sequences). Errors affect the detection of fragment overlaps. Hence, the consensus determination requires multiple alignments in highly coverage regions.
3. Incomplete coverage: It happens when the algorithm is not able to assemble a given set of fragments into a single contig. A contig is a sequence in which the overlap between adjacent fragments is greater or equal to a predefined threshold (cutoff parameter).
4. Repeated regions: "Repeats" are sequences that appear two or more times in the target DNA. Repeated regions have caused problems in many genome-sequencing projects, and none of the current assembly programs can handle them perfectly.
5. Chimeras and contamination: Chimeras arise when two fragments that are not adjacent or overlapping on the target molecule join together into one fragment. Contamination occurs due to the incomplete purification of the fragment from the vector DNA.

After the order is determined, the progressive alignment algorithm is applied to combine all the pairwise alignments obtained in the overlap phase.

Consensus Phase - Deriving the DNA sequence from the layout. The most common technique used in this phase is to apply the majority rule in building the consensus.

To measure the quality of a consensus, we can look at the distribution of the coverage. Coverage at a base position is defined as the number of fragments at that position. It is a measure of the redundancy of the fragment data, and it denotes the number of fragments, on average, in which a given nucleotide in the target DNA is expected to appear. It is computed as the number of bases read from fragments over the length of the target DNA [1].

$$Coverage = \frac{\sum_{i=1}^{n} length\ of\ the\ fragment\ i}{target\ sequence\ length} \tag{1}$$

where n is the number of fragments. The higher the coverage, the fewer number of the gaps, and the better the result.

3 Our Proposal: Problem Aware Local Search (PALS)

Classical assemblers use fitness functions that favor solutions in which strong overlap occurs between adjacent fragments in the layouts, using equations like 2 [9] (where $w_{i,j}$ is the overlap between fragments i and j). But the actual objective is to obtain an order of the fragments that minimizes the number of contigs, with the goal of reaching one single contig, i.e., a complete DNA sequence composed of all the overlapping fragments. Therefore, the number of contigs is used as a