Nataša Jonoska
Nadrian C. Seeman (Eds.)

LNCS 2340

DNA Computing

7th International Workshop on DNA-Based Computers, DNA7
Tampa, FL, USA, June 2001
Revised Papers

Springer

Nataša Jonoska Nadrian C. Seeman (Eds.)

DNA Computing

7th International Workshop on DNA-Based Computers, DNA7
Tampa, FL, USA, June 10-13, 2001
Revised Papers

Springer

Volume Editors

Nataša Jonoska
University of South Florida, Department of Mathematics
4202 E. Fowler A. PHY 114, Tampa FL 33620, USA
E-mail: jonoska@math.usf.edu

Nadrian C. Seeman
New York University, Department of Chemistry
New York, NY 10003, USA
E-mail: ned.seeman@nyu.edu

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

DNA computing : revised papers / 7th International Workshop on DNA Based Computers, DNA 7, Tampa, FL, USA, June 10 - 13, 2001. Nataša Jonoska ; Nadrian C. Seeman (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Tokyo : Springer, 2002
(Lecture notes in computer science ; Vol. 2340)
ISBN 3-540-43775-4

CR Subject Classification (1998): F.1, F.2.2, I.2.9, J.3

ISSN 0302-9743
ISBN 3-540-43775-4 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
Printed on acid-free paper SPIN: 10869927 06/3142 5 4 3 2 1 0

Preface

Biomolecular computing is an interdisciplinary field that draws together molecular biology, chemistry, physics, computer science, and mathematics. DNA nanotechnology and molecular biology are key relevant experimental areas, where knowledge increases with each passing year. The annual international meeting on DNA-based computation has been an exciting forum where scientists of different backgrounds who share a common interest in biomolecular computing meet and discuss their latest results. The central goal of this conference is to bring together experimentalists and theoreticians whose insights can calibrate each other's approaches. DNA7, *The Seventh International Meeting on DNA Based Computers*, was held at The University of South Florida in Tampa, FL, USA, June 10–13, 2001. The organizers sought to attract the most significant recent research, with the highest impact on the development of the discipline. The meeting had 93 registered participants from 14 countries around the world. The program committee received 44 abstracts, from which 26 papers were presented at the meeting, and included in this volume. In addition to these papers, the Program Committee chose 9 additional papers from the poster presentations, and their revised versions have been added to this volume.

As is now a tradition, four tutorials were presented on the first day of the meeting. The morning started with general tutorials by Erik Winfree (Caltech) and Junghuei Chen (University of Delaware), designed to bridge between their respective areas of expertise, computer science and molecular biology. More specialized tutorials on encoding DNA sequences and on non-standard DNA motifs and interactions were given in the afternoon by Anne Condon (University of British Columbia) and Nadrian C. Seeman (New York University), respectively. Four plenary lectures were given during the conference by Nicholas Cozzarelli (University of California at Berkeley) on DNA topology, Richard Lipton (Georgia Technology Institute) on the state of DNA-based computation, John SantaLucia (Wayne State University) on DNA hybridization thermodynamics and Ronald Breaker (Yale University) on DNA catalysis. Those presentations are not included in this volume.

The research presented here contains a diverse spectrum of ideas and topics. The papers under *Experimental Tools* deal with issues such as optimization of biomolecular protocols or a computer program for designing DNA sequences that could be found useful in performing experiments. The papers in *Theoretical Tools* study theoretical properties of DNA sequences and structures that could be used in designing models and subsequently, experiments. Several papers deal with *Probabilistic Theoretical Models* which try to capture the inexact nature of the biomolecular protocols. As the experience of many has shown, sequence design and computer simulations can be very valuable before preparing an actual experiment and several researchers addressed these issues in *Computer Simulation and Sequence Design.* New algorithms for solving difficult problems

such as the knapsack problem and SAT are introduced in *Algorithms*. Several researchers, in fact, reported on successful experimental solutions of instances of computational problems. Their results are included in *Experimental Solutions*. The papers in *Nano-tech Devices* report on the experimental design of DNA nano-mechanical devices. The section on *Biomimetic Tools* contains research on computational tools that primarily use processes found naturally in the cells of living organisms. Several papers deal with the theory of splicing systems and the formal language models of membrane computing. These papers are included in *Splicing and Membrane Systems*.

The editors would like to acknowledge the help of the conference's Program Committee in reviewing the submitted abstracts. In addition to the editors, the Program Committee consisted of Junghuei Chen, Anne Condon, Masami Hagiya, Tom Head, Lila Kari, George Paun, John Reif, Grzegorz Rozenberg, Erik Winfree, David Wood, and Bernard Yurke. The editors thank Denise L. Marks for helping us with her skillful typesetting abilities. The Organizing Committee (Anne Condon, Grzegorz Rozenberg, and the editors) is grateful for the generous support and sponsorship of the conference by the Center for Integrated Space Microsystems within the Jet Propulsion Laboratory, NASA, and the following branches of The University of South Florida: The College of Arts and Sciences, the Institute for Biomolecular Science, the Department of Mathematics, the Department of Biology, the Department of Chemistry; and the USF Research Foundation.

The meeting was held in cooperation with the ACM Special Interest Group on Algorithms and Computation Theory (ACM SIGACT) and the European Association for Theoretical Computer Science (EATCS).

We note with sadness the passing of Michael Conrad, who participated in several of the earlier conferences. His contributions will be missed by all.

Finally, the editors would like to thank all of the participants in the DNA7 conference for making it a scintillating and fruitful experience. This is a discipline that has not yet found its 'killer ap,' but the excitement that is generated when this group assembles is virtually palpable at the conference. The interactions, collaborations, and advances that result from each of the meetings on DNA Based Computers are the key products of the meeting. We hope that this volume has captured the spirit and exhilaration that we experienced in Tampa.

April 2002

Nataša Jonoska
Nadrian Seeman

Organization

Program Committee

Junghuei Chen, University of Delaware, USA
Anne Condon, University of British Columbia, Canada
Masami Hagiya, University of Tokyo, Japan
Tom Head, Binghamton University, USA
Nataša Jonoska, University of South Florida, USA
Lila Kari, University of western Ontario, Canada
Gheorghe Paŭn, Romanian Academy, Romania
John H. Reif, Duke University, USA
Grzegorz Rozenberg, Leiden University, The Netherlands
Nadrian C. Seeman, New York University, USA (Program Chair)
Erik Winfree, California Institute of technology, USA
David H. Wood, University of Delaware, USA
Bernie Yurke, Bell Laboratories, Lucent Technologies, USA

Organizing Committee

Anne Condon, University of British Columbia, Canada
Nataša Jonoska, University of South Florida, USA
(Organizing Committee Chair)
Grzegorz Rozenberg, Leiden University, The Netherlands
Nadrian C. Seeman, New York University, USA

Table of Contents

Experimental Tools

Theoretical Tools

Probabilistic Computational Models

Computer Simulation and Sequence Design

Algorithms

Experimental Solutions

Nano-tech Devices

Biomimetic Tools

New Computing Models

Splicing Systems and Membranes

An Object Oriented Simulation of Real Occurring Molecular Biological Processes for DNA Computing and Its Experimental Verification

Thomas Hinze[1], Uwe Hatnik[2], and Monika Sturm[1]

[1] Institute of Theoretical Computer Science, Dresden University of Technology, Mommsenstr. 13, i Dresden, D-01062, Germany
{hinze, sturm}@tcs.inf.tu-dresden.de
http://www.tcs.inf.tu-dresden.de/dnacomp

[2] Institute of Computer Engineering, Dresden University of Technology, Mommsenstr. 13, i Dresden, D-01062, Germany
hatnik@ite.inf.tu-dresden.de

Abstract. We present a simulation tool for frequently used DNA operations on the molecular level including side effects based on a probabilistic approach. The specification of the considered operations is directly adapted from detailed observations of molecular biological processes in laboratory studies. Bridging the gap between formal models of DNA computing, we use process description methods from biochemistry and show the closeness of the simulation to the reality.

1 Introduction

It is well-known that DNA operations can cause side effects in a way that the results of algorithms do not fit to the expectation. Any molecular biological operation used for DNA computing seems to be closely connected with certain unwanted effects on the molecular level. Typical side effects are for instance unwanted additional DNA strands, loss of wanted DNA strands, artifacts, mutations, malformed DNA structures or sequences, impurities, incomplete or unspecific reactions, and unbalanced DNA concentrations. Unfortunately, side effects can sum up in sequences of DNA operations leading to unprecise, unreproducible or even unusable final results [6]. Coping with side effects is to be seen as the main challenge in the research field of experimental DNA computing. We have analyzed processes used in DNA computing at the molecular level in laboratory studies with the aim to specify these processes as detailed as possible. The analysis led to a classification and to a statistical parametric logging of side effects. Based on this knowledge, we have developed a simulation tool of real occurring molecular biological processes considering side effects. The comparison of simulation results with real observations in the laboratory shows a high degree of accordance. Our main objective is to construct error reduced and side effect compensating algorithms. Furthermore, the gap between formal models of DNA

N. Jonoska and N.C. Seeman (Eds.): DNA7, LNCS 2340, pp. 1–13, 2002.

computing and implementations in the laboratory should be bridged. A clue to handle side effects in DNA computing can consist in the idea to include them into the definition of DNA operations as far as possible. DNA computing as hardware architecture particularly convinces by its practicability of laboratory implementations based on a formal model of computation.

The simulation tool and continued laboratory studies extend our results presented at DNA6 [4]. Our work focuses a reliable implementation of an optimized distributed splicing system TT6 in the laboratory [8]. Using the simulation tool, prognoses about resulting DNA strands and influences of side effects to subsequent DNA operations can be obtained. The number of strand duplicates reflecting DNA concentrations is considered as an important factor for a detailed description of the DNA computing operations on the molecular level in the simulation. This property allows to evaluate the quantitative balance of DNA concentrations in a test tube. Here, we show the abilities of the simulation using the operations synthesis, annealing, melting, union, ligation, digestion, labeling, polymerisation, PCR, affinity purification, and gel electrophoresis by means of selected examples with comparison to laboratory results.

2 Modelling Molecular Biological Processes

The knowledge about underlying molecular biological processes grows up more and more rapidly. In the meantime, the principles of biochemical reactions are understood very well. Precise descriptions can be found in recent handbooks of genetic techniques like [7]. This pool of knowledge mostly aims at applications in medicine, agriculture, and genetic engineering. Our intention is to use this knowledge and to apply it for approaches in DNA computing.

Biochemical reactions on DNA are generally caused by collisions of the reactants with enough energy to transform covalent or hydrogen bonds. This energy is usually supplied by heating or by addition of instable molecules with a large energy potential. Thus the vis viva of the molecules inside the test tubes increases and they become more moveable. One test tube can contain up to 10^{20} molecules including water dipoles. Which reactive molecules of them will interact indeed? The answer to this question requires to abstract from a macroscopic view. A microscopic approach has to estimate the probability of an inter- or intra-molecular reaction for all combinations of molecules inside the test tube. This can be done by generating a probability matrix whose elements identify all possible combinations how molecules can hit to react together. The probabilities for a reaction between the molecules forming a combination depend on many parameters e.g. chemical properties of the molecules, their closeness and orientation to each other and the neighbourship of other reactive molecules. After creating the matrix of molecular reaction probabilities, a certain combination with acceptable probability > 0 is selected randomly according to the given probability distribution. The molecular reaction is performed and produces a modified contents of the test tube. Using this contents, the subsequent matrix of molecular reaction probabilities is generated and so on. The whole reaction

can be understood as a consecutive iterated process of matrix generation, selection of a molecular reaction and its performance. The process stops if all further probabilities for molecular reactions are very low or an equilibrium of the test tube contents occurs. This strategy to model molecular biological processes implies side effects and a nondeterministic behaviour in a natural way. The simulation tool adapts this basic idea to model processes of DNA computing on the molecular level controlled by suitable parameters. A simple annealing example should illustrate the idea how to simulate biochemical reactions closed to the laboratory. Annealing (hybridization) is a process that pairs antiparallel and complementary DNA single strands to DNA double strands by forming hydrogen bonds between opposite orientated complementary bases. Let assume for simplicity that a (very small) test tube contains three different DNA sequences in solution: 10 copies of the DNA single strand `5'-AAGCTCCGATGGAGCT-3'`, 6 copies of `5'-TGAAGCTCCATCGGA-3'`, and 7 copies of `5'-GAGCTTATA-3'`. Further let assume that these strands are spatially distributed in equipartition and that one molecular reaction affects max. $k = 2$ DNA molecules at once. Figure 1 shows the first iteration of process simulation.

The matrix derived from the test tube contents lists the probabilities for inter- resp. intramolecular collisions that can result in molecular reactions for all combinations of molecules. Subsequently, one combination is selected randomly with respect to the probability distribution. The example uses the collision marked by a grey background. For this selected combination, all possible molecular hybridization products have to be determined.

Two DNA strands can stable anneal to each other if at least approximately 50% of the bases of one participating strand form hydrogen bonds with their complementary counterparts of the other one. A lower bonding rate mostly produces not survivable DNA double strands that melt again. The minimum bonding rate describes the process parameter of annealing. Based on the bonding rate parameter, all possible stable molecular hybridization products from the selected combination are generated. One of these products is selected randomly as performed molecular reaction. The test tube contents is modified accordingly completing one iteration of the process cycle. The modified test tube contents serves as input for the next iteration and so on until no new products can appear.

The annealing example should point out the principle how to model molecular biological processes. Other reactions resp. processes can be considered in a similar way. Our studies include the DNA operations synthesis, annealing, melting, union, ligation, digestion, labeling, polymerisation, PCR, affinity purification, and gel electrophoresis. They affect as follows:

operation	**effect**
synthesis	generation of DNA single strands (oligonucleotides) up to maximum approximately 100 nucleotides; there are no limitations to the sequence. Most methods use the principle of a growing chain: Fixed on a surface, the DNA single strands are constructed by adding one nucleotide after the other using a special coupling chemistry. Finally, the DNA single strands are removed from the surface and purified.

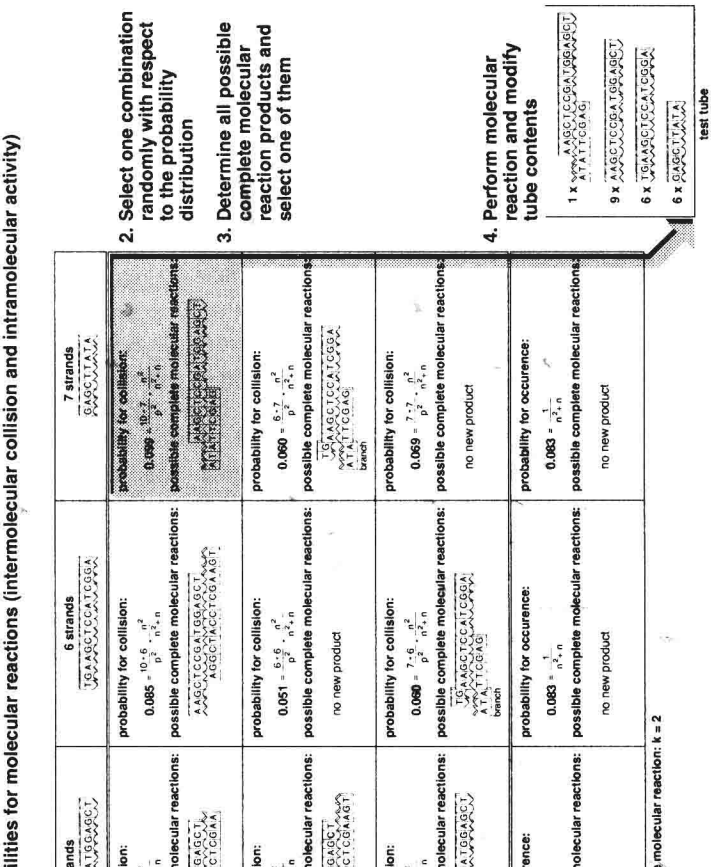

Fig. 1. Annealing example of process simulation, one iteration of the process cycle.

operation	effect
annealing	pairing of minimum two antiparallel and complementary DNA single strands or single stranded overhangs to DNA double strands by forming thermic instable hydrogen bonds; the process is performed by heating above the melting temperature and subsequently slowly cooling down to room temperature. Annealing product molecules can survive if at least 50% of the bases of one participated strand bind to their complementary counterpart.
melting	breaking hydrogen bonds by heating above the melting temperature or by using alkaline environments
union	merging the contents of several test tubes into one common test tube without changes of chemical bonds
ligation	concatenation of compatible antiparallel complementary sticky or blunt DNA double strand ends with 5' phosphorylation; enzym DNA ligase catalyzes the formation of covalent phosphodiester bonds between juxtaposed 5' phosphate and 3' hydroxyl termini of double stranded DNA.
digestion	cleavage of DNA double strands on occurences of specific recognition sites defined by the enzym; all arising strand ends are 5' phosphorylated. Enzym type II restriction endonuclease catalyzes the break of covalent phosphodiester bonds at the cutting position.
labeling	set or removal of molecules or chemical groups called labels at DNA strand ends; enzym alkaline phosphatase catalyzes the removal of 5' phosphates (5' dephosphorylation). Enzym Polynucleotide Kinase catalyzes the transfer and exchange of phosphate to 5' hydroxyl termini (5' phosphorylation). Beyond phosphate, other labels like 5' biotin, fluorescent or radioactive labels can be used in a similar way.
polymerisation	conversion of DNA double strand sticky ends into blunt ends; enzym like vent DNA polymerase (New England Biolabs) catalyzes the completion of recessed 3' ends and the removal of protruding 3' ends.
gel electrophoresis	physic technique for separation of DNA strands by length using the negative electric charge of DNA; DNA is able to move through the pores of a gel, if a DC voltage (usually $\approx 80V$) is applied and causes an electrolysis. The motion speed of the DNA strands depends on their molecular weight that means on their length. After switching off the DC voltage, the DNA is separated by length inside the gel. Denaturing gels (like polyacrylamide) with small pores process DNA single strands and allow to distinguish length differences of 1 base. Non-denaturing gels (like agarose) with bigger pores process DNA double strands with precision of measurement $\approx \pm 10\%$ of the strand length.

operation	effect
polymerase chain reaction (PCR)	cyclic process composed by iterated application of melting, annealing, and polymerisation used for exponential amplification of double stranded DNA segments defined by short ($\approx$ 20 bases long), both-way limiting DNA sequences; these sequences denoted as DNA single strands are called primers. Each cycle starts with melting of the double stranded DNA template into single strands. Subsequently the primers are annealed to the single strands and completed to double strands by polymerisation. Each cycle doubles the number of strand copies. PCR can produce approximately up to 2^{40} strand copies using 40 cycles. Higher numbers of cycles stop the exponential amplification leading to a saturation.
affinity purification	separation technique that allows to isolate 5' biotinylated DNA strands from others; biotin binds very easily to a streptavidin surface fixing according labelled DNA strands. Unfixed DNA strands are washed out and transferred to another tube.

Molecular biological processes annealing and ligation induce interactions between different DNA strands. They are able to produce a variety of strand combinations. The potential and power of DNA computing to accelerate computations rapidly is based on annealing and ligation. Other DNA operations listed above affect the DNA strands inside the test tube independently and autonomously. In this case, interactions are limited to DNA with other reactants or influences from the environment. Union, electrophoresis, and sequencing require modelling as physic processes without reactive collisions between molecules.

3 A Probabilistic Approach to Model DNA Operations with Side Effects

The effect of DNA operations on the molecular level depends on random (nondeterministic) interactions (events) with certain probability. The variety of possible events is specified by biochemical rules and experimental experiences. Only a part of them – but not all – forms the description of formal models of DNA computing. Remaining unconsidered events are subsumed by the term "side effect". Formal models of DNA computing include many significant properties but others are ignored (abstraction). The most commonly used assumptions for abstraction are:

- Linear DNA single or double strands are used as data carrier.
- Information is encoded by DNA sequence (words of formal languages).
- unrestricted approach; arbitrary (also infinite) number of strand copies allowed
- Unique result DNA strands can be detected absolutely reliable.
- All DNA operations are performed completely.
- All DNA operations are absolute reproducible.

Differences from these abstractions are considered as side effects. They can be classified into certain groups with specific common properties. The properties are chosen in a way that the side effect can either be defined by statistical parameters with respect to defaults from the reactants (e.g. mutation error rate of DNA polymerase) or the side effect directly results from the process description. Figure 2 shows a proposal for a classification extending the idea from [1] to the set of frequently used DNA basic operations.

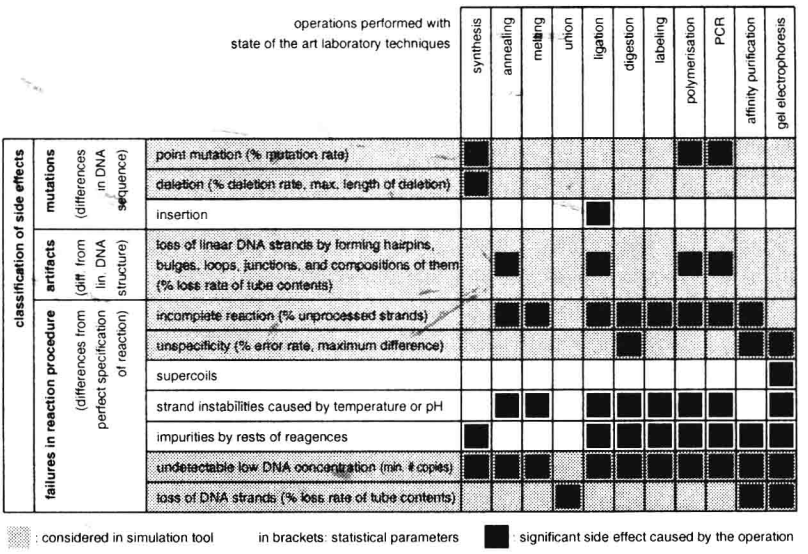

Fig. 2. Significant side effects of frequently used DNA operations

The following table lists the operation parameters and side effect parameters of the considered DNA basic operations. The default values are adapted from laboratory studies. The abbreviation L stands for strand length.

operation	parameter	range	default
synthesis	**operation parameters**		
	• tube name		
	• nucleotide sequence (5'-3')		
	• number of strand copies	$1 \ldots 10^6$	
	side effect parameters		
	• point mutation rate	$0 \ldots 100\%$	5%
	• deletion rate	$0 \ldots 100\%$	1%
	• maximum deletion length	$0 \ldots 100\%$ of L	5%

operation	parameter	range	default
annealing	**operation parameters**		
	• tube name		
	• minimum bonding rate for stable duplexes	0...100%	50%
	• maximum length of annealed strands	$1...10^6$	
	side effect parameters		
	• base pairing mismatch rate	0...100%	$600/L$
	• rate of unprocessed strands	0...100%	5%
melting	**operation parameters**		
	• tube name		
	side effect parameters		
	• rate of surviving duplexes	0...100%	0.1%
union	**operation parameters**		
	• tube name		
	• name of tube whose contents is added		
	side effect parameters		
	• strand loss rate	0...100%	0.5%
ligation	**operation parameters**		
	• tube name		
	• maximum length of ligated strands	$1...10^6$	
	side effect parameters		
	• rate of unprocessed strands	0...100%	5%
polymerisation	**operation parameters**		
	• tube name		
	side effect parameters		
	• point mutation rate	0...100%	0.1%
digestion	**operation parameters**		
	• tube name		
	• recognition sequence		
	• restriction site		
	side effect parameters		
	• rate of not executed molecular cuts	0...100%	5%
	• rate of star activity (unspecificity)	0...100%	5%
	• recognition sequence with wildcard base pairs specifying star activity		
labeling	**operation parameters**		
	• tube name		
	• kind of label (biotin or phosphate)		
	• kind of strand end (3' or 5')		
	• action (set or removal of label)		
	side effect parameters		
	• rate of unprocessed strands	0...100%	5%
affinity purification	**operation parameters**		
	• tube name		
	• kind of extracted strands (with or without biotin label)		
	side effect parameters		
	• rate of false positives (unspecificity)	0...100%	8%
	• rate of false negatives (unspecificity)	0...100%	8%