A Probabilistic Approach to Description of Molecular Biological Processes on DNA and Their Object Oriented Simulation

U. HATNIK[1] hatnik@eas.iis.fhg.de
T. HINZE[2] hinze@tcs.inf.tu-dresden.de
M. STURM[2] sturm@tcs.inf.tu-dresden.de
[1]Fraunhofer Institute for Integrated Circuits, Division EAS, Zeunerstr. 38, D-01069 Dresden
[2]Dresden University of Technology, Inst. for Theoretical Computer Science, D-01062 Dresden
GERMANY
http://wwwtcs.inf.tu-dresden.de/dnacomp

Abstract: We introduce a probabilistic approach to simulate molecular biological processes controlled by appropriate statistical parameters. Our approach considers molecular biological processes based on strands of deoxyribonucleic acid (DNA): oligonucleotide synthesis, annealing, melting, union, ligation, digestion, strand end labeling, polymerisation, affinity purification, and gel electrophoresis with regard to frequently used laboratory techniques. The simulation on the molecular level allows prediction of experimental results including side effects that can occur in the laboratory. Fields of application consist in experimental setup for genetic engineering as well as particularly in DNA computing. We point out that object oriented simulation is preferably suited for modeling those processes.

Key-Words: object oriented simulation, stochastic processes, molecular biological processes, DNA computing, probabilistic approach, model for computation

1 Introduction

DNA based recombination techniques enable genetic engineering and led to a huge progress in medicine and agriculture. The success of laboratory experiments significantly depends on the ability to predict the outcome of a series of molecular biological reactions and processes. Since those reactions and processes run in a nondeterministic manner using stochastic collisions of molecules, a reliable and efficient preparation of laboratory experiments with precise and reproducible results is seen as one of the main scientific and economic challenges.

Inspired by the potential of DNA based recombination techniques, computer science also profits from the molecular biology. DNA computing, originated by L.M. Adleman [1], emerges as a promising idea how to compute: DNA serves as a data carrier and storage medium for information, and molecular biological reactions and processes are used as operations on DNA. DNA computing features by its massive data parallelism (up to 10^{21} molecular operations per second), because one operation can affect all DNA strands inside a test tube in parallel. Furthermore, DNA owns a persistent high storage capacity and density (up to 10^{20} base pairs per litre), and its processing is very energy efficient, environmentally friendly and without any mechanical wear. DNA algorithms are sequences of DNA operations with appropriate DNA single or double strands encoding the input data. The output is done by determination of the strand lengths using gel electrophoresis or even by determination of the sequence of nucleotides (A, C, G, and T) the final DNA strands consist of. NP problems (combinatorial problems) can be solved in polynomial time using DNA computing. Practical experiments succeeded promising in general, but 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 can be seen as the main challenge in the research field of experimental DNA computing. DNA computing particularly convinces by its practicability of laboratory implementations based on a formal model of computation [4].

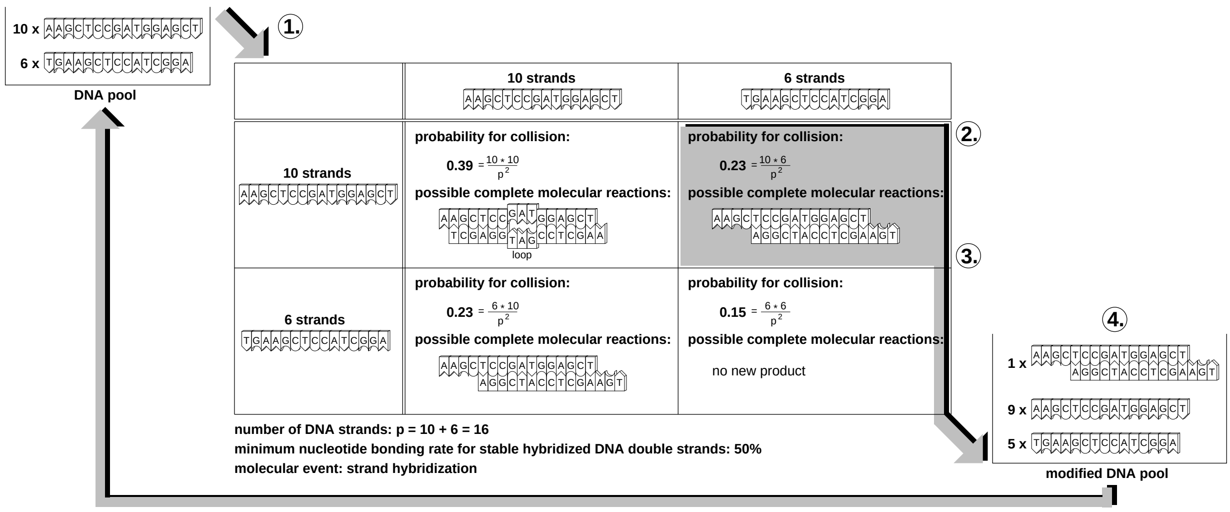
We have analyzed processes used in DNA computing at the molecular level in laboratory studies with the aim to specify these processes as detailed as possible. Based on this knowledge, we have developed a simulation tool of real occurring molecular biological processes considering side effects. Side effects are described by appropriate statistical parameters. The comparison of simulation results with real observations in the laboratory shows a high degree of accordance [3]. 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. The simulation considers the DNA based reactions and processes synthesis, annealing, melting, union, ligation, digestion, labeling, polymerisation, affinity purification, and gel electrophoresis.

2 Molecular Biological Processes

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. 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 DNA operations 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 oriented complementary bases. Let assume that the input strands are arranged in spatial equipartition and that one molecular reaction affects max. 2 DNA molecules at once. Figure 1 shows the first iteration of process simulation. DNA single strands are denoted in 5'-3' direction.

The matrix derived from the test tube contents lists the probabilities for inter- resp. intramole-



1. **Create list (matrix) of molecular events and their probabilities including side effects**
2. **Select one molecular event randomly with respect to the probability distribution**
3. **Determine all possible reaction products from this molecular event and select one of them**
4. **Modify DNA pool**

Figure 1: annealing example of process simulation, one iteration of the process cycle

cular 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. The annealing example should point out the principle how to model molecular biological processes. Other reactions resp. processes can be described in a similar way. Our studies include the DNA operations listed above. They affect as follows:

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.

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 occurrences 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, 5' biotin can be used in a similar way.

polymerisation: conversion of DNA double strand sticky ends into blunt ends; enzym DNA polymerase (e.g. New England Biolabs) catalyzes the replenishment of recessed 3' ends and the removal of protruding 3' ends.

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.

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.

Molecular biological processes annealing and ligation induce interactions between different DNA strands. They are able to produce a variety of strand combinations. Other DNA operations listed above affect the DNA strands inside the test tube independently and autonomously.

3 A Probabilistic Approach to Modelling the Processes

Every DNA operation above is specified by operation parameters and statistical side effect parameters. The operation parameters characterize the process behaviour and give any necessary information to perform the operation itself. The operation parameters correspond to the defaults of the laboratory protocol that can be adjusted or choosen directly. In contrast, additional side effect parameters classify the intensity of the nondeterministic process behaviour in a statistical way. Any DNA operation is usually performed inside a test tube. Following parameters are used with regard to their significant occurrence in the operations:

synthesis: *op. parameters:* tube name, nucleotide sequence (5'-3'), number of strand copies; *side effect parameters:* point mutation rate, deletion rate, maximum deletion length (in % of strand length)

annealing: *op. parameters:* tube name, minimum bonding rate for stable double strands, maximum length of annealed strands; *side effect parameters:* base pairing mismatch rate, rate of unprocessed strands

melting: *op. parameters:* tube name; *side effect parameters:* rate of surviving double strands

union: *op. parameters:* tube name, name of tube whose contents is added; *side effect parameters:* strand loss rate

ligation: *op. parameters:* tube name, maximum length of ligated strands; *side effect parameters:* rate of unprocessed strands

digestion: *op. parameters:* tube name, recognition sequence, restriction site; *side effect parameters:* rate of not executed molecular cuts, rate of star activity (unspecificity), recognition sequence with wildcard base pairs specifying star activity

labeling: *op. 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

polymerisation: *op. parameters:* tube name; *side effect parameters:* point mutation rate

affinity purification: *op. parameters:* tube name, kind of extracted strands (with or without biotin label); *side effect parameters:* rate of false positives (unspecificity), rate of false negatives (unspecificity)

gel electrophoresis: *op. parameters:* tube name, minimum number of strand copies with same length necessary for detection; *side effect parameters:* strand loss rate, rate of strands with forged length, maximum length derivation (forgery)

4 The Simulation Tool

The simulation tool adapts the probabilistic approach to specify molecular biological reactions and processes including significant side effects. The main features focus on:

- Specification of DNA operations is set on the level of single nucleotides and strand end labels using the principle of random probability-controlled consecutive interactions between DNA strands and reactants.
- Number of strand copies is considered to distinguish concentrations of different DNA strands and their influence to the behaviour in the operational process.
- Each DNA operation is processed inside a virtual test tube that collects a set of DNA strands. Specific parameters control the process. The simulation tool is able to manage several test tubes.
- Arbitrary sequences of DNA operations including the propagation of side effects can be visualized and logged.

Since a test tube can be considered as a system containing groups of DNA strands and reactants as (autonomous) subsystems, an object oriented approach for simulation is preferred: object oriented simulation means that a system is split into subsystems which are simulated autonomously [5]. A subsystem in this context is named "object" and may contain other objects forming an object hierarchy. An object embeds its own simulation algorithm that can represent both, a small code fragment and an extensive simulator, see figure 2. All implementation details are encapsulated by the object, only an interface allows data exchange and simulation control. The advantage of this approach lies in its flexibility with respect to object combination and exchange. Furthermore, the simulation algorithm can be optimally adapted to the models [7], [2].

The application of the object oriented simulation technique using a hierarchy of autonomous subsystems models the behaviour of DNA molecules during an operational process in a natural way. On the basic level (sequences), each molecule acts as a subsystem with certain properties to form or break hydrogen or covalent bonds. Collisions between these molecules are controlled by probability distribution using a generator for random numbers. The next upper level (tube) collects all DNA strands belonging to a test tube contents. For example, identical DNA double strands can have different representations. Summarization and unification of identical molecules is done in this level. The top level

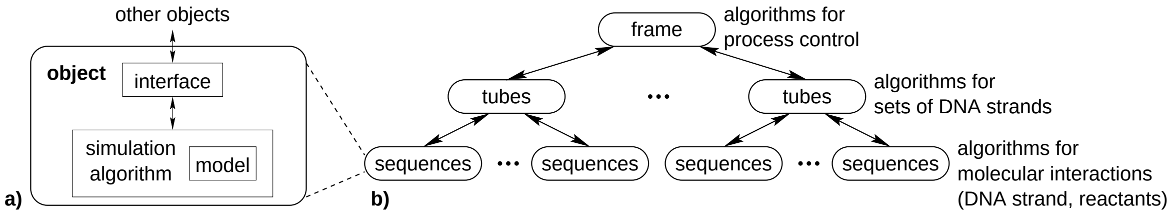


Figure 2: basic object structure **a)** and hierarchical composition of objects **b)**

(frame) manages the collection of test tubes and the user interactions.

The implementation uses Java to ensure a wide interoperability to different platforms because of its object oriented paradigm. The simulation tool requires at least Java Development Kit 2.0.

5 Conclusions

The simulation tool represents a model for molecular biological processes on DNA, directly adapted from their analysis in the laboratory. In contrast to known models for DNA computing, the simulation tool also considers the influence of significant side effects. The intensity of side effects can be controlled by suitable statistical parameters in a range from no influence to absolute dominance. The consistent parameterization of DNA operations as well as side effects entails a high degree of flexibility and ergonomics for the simulation tool. The object oriented simulation approach supports the modelling of interactions between DNA strands and reactants as autonomous subsystems that are combined to test tubes. The implementation in Java guarantees interoperability to different platforms. Recently, the simulation tool features by the DNA operations synthesis, annealing, melting, union, ligation, digestion labeling, polymerisation, affinity purification, and gel electrophoresis. To verify the simulation tool, results of selected laboratory experiments were compared to simulated predictions. A high degree of accordance was obtained. Further studies focus on the extension to additional effects concerning nonlinear DNA structures.

References:

[1] L.M. Adleman. Molecular computation of solutions to combinatorial problems. *Science,* Vol. 266, 1994, pp. 1021–1024.

[2] U. Hatnik, J. Haufe, P. Schwarz. Object Oriented System Simulation of Large Heterogeneous Communication Systems. *Workshop on System Design Automation SDA2000,* Rathen, Germany, 2000, pp. 178–184.

[3] T. Hinze, U. Hatnik, M. Sturm. An object-oriented simulation of real occurring molecular biological processes for DNA computing and its experimental verification. In N. Jonoska, N. Seeman, ed., *PreProc. Seventh International Meeting on DNA Based Computers,* Tampa, USA, 2001, pp. 13–22.

[4] T. Hinze, M. Sturm. Towards an in-vitro Implementation of a Universal Distributed Splicing Model for DNA Computation. In R. Freund, ed., *Proc. Theorietag 2000,* TU Wien, Austria, 2000, pp. 185–189.

[5] J.A. Joines, S.D. Roberts. Fundamentals of object-oriented simulation. In D.J. Medeiros, E.F. Watson, et.al., ed., *Proc. of 1998 Conference on Winter Simulation,* Washington, USA, 1998, pp. 141–150.

[6] P.D. Kaplan, G. Cecchi, A. Libchaber. Molecular computation: Adleman's experiment repeated. *Technical report,* NEC Research Institute, 1995.

[7] G. Zobrist, J.V. Leonard. *Object-Oriented Simulation – Reusability, Adaptability, Maintainability.* IEEE Press, 1997.