An object-oriented simulation of real occurring molecular biological processes for DNA computing and its experimental verification

Thomas Hinze[1] Uwe Hatnik[2] Monika Sturm[1]
hinze@tcs.inf.tu-dresden.de hatnik@ite.inf.tu-dresden.de sturm@tcs.inf.tu-dresden.de
`http://wwwtcs.inf.tu-dresden.de/dnacomp`

Dresden University of Technology, [1] Institute of Theoretical Computer Science,
[2] Institute of Computer Engineering, Mommsenstr. 13, Dresden, D-01062, Germany

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 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

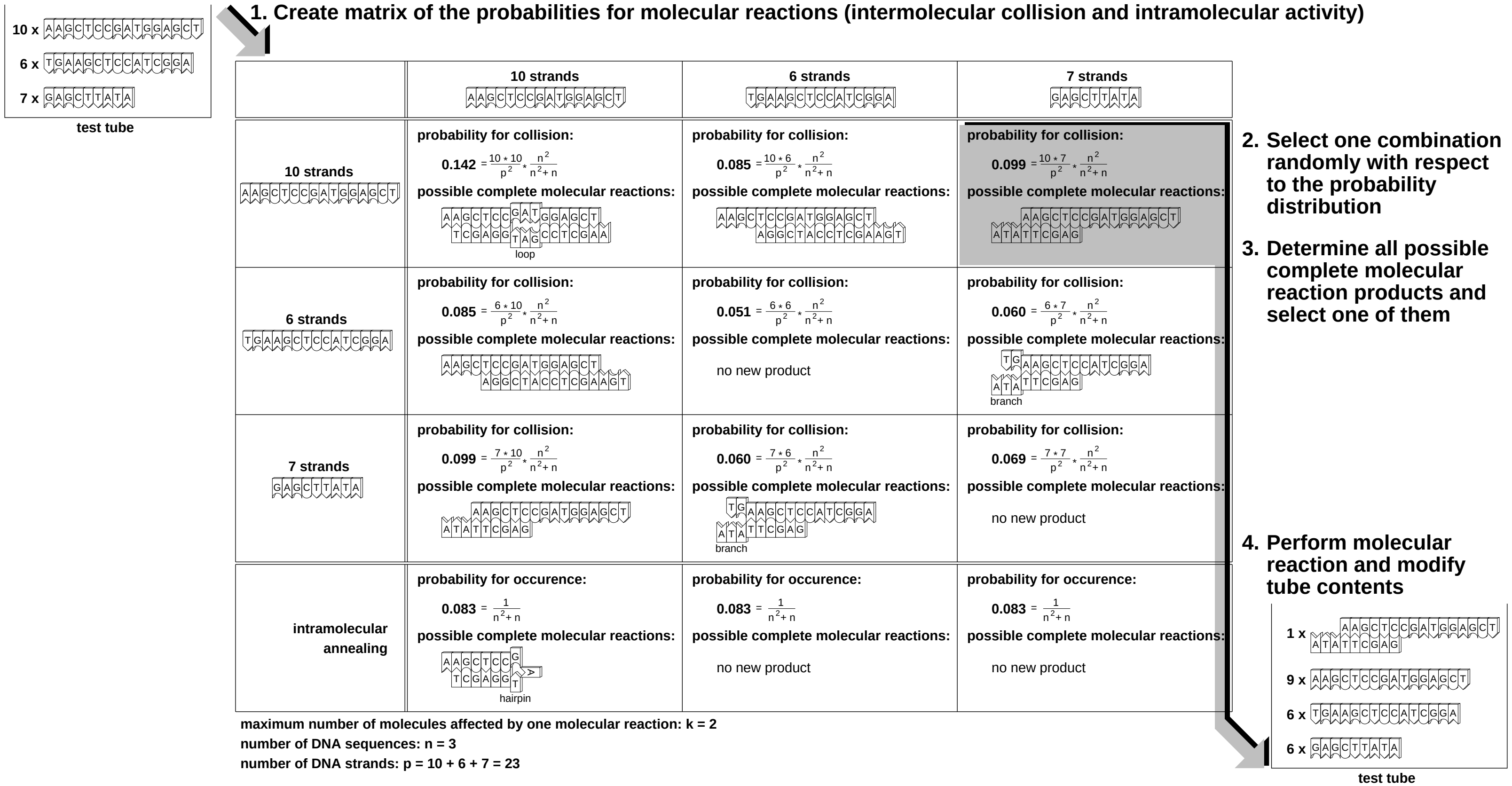

Fig. 1. annealing example of process simulation, one iteration of the process cycle

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.
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.

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.

classification of side effects		operations performed with state of the art laboratory techniques	synthesis	annealing	melting	union	ligation	digestion	labeling	polymerisation	PCR	affinity purification	gel electrophoresis
mutations	(differences in DNA sequence)	point mutation (% mutation rate)	■							■	■		
		deletion (% deletion rate, max. length of deletion)	■										
		insertion					■						
artifacts	(diff. from lin. DNA structure)	loss of linear DNA strands by forming hairpins, bulges, loops, junctions, and compositions of them (% loss rate of tube contents)		■			■			■	■		
failures in reaction procedure	(differences from perfect specification of reaction)	incomplete reaction (% unprocessed strands)		■	■		■	■	■	■	■	■	
		unspecificity (% error rate, maximum difference)						■				■	■
		supercoils											■
		strand instabilities caused by temperature or pH		■	■		■	■	■	■	■		■
		impurities by rests of reagences	■				■	■	■	■	■	■	■
		undetectable low DNA concentration (min. # copies)	■	■	■		■	■	■	■	■	■	■
		loss of DNA strands (% loss rate of tube contents)				■						■	■

: considered in simulation tool in brackets: statistical parameters ■ : significant side effect caused by the operation

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

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%
annealing	**operation parameters**		
	• tube name		
	• minimum bonding rate for stable duplexes	$0 \ldots 100\%$	50%
	• maximum length of annealed strands	$1 \ldots 10^6$	
	side effect parameters		
	• base pairing mismatch rate	$0 \ldots 100\%$	$600/L$
	• rate of unprocessed strands	$0 \ldots 100\%$	5%
melting	**operation parameters**		
	• tube name		
	side effect parameters		
	• rate of surviving duplexes	$0 \ldots 100\%$	0.1%
union	**operation parameters**		
	• tube name		
	• name of tube whose contents is added		
	side effect parameters		
	• strand loss rate	$0 \ldots 100\%$	0.5%
ligation	**operation parameters**		
	• tube name		
	• maximum length of ligated strands	$1 \ldots 10^6$	
	side effect parameters		
	• rate of unprocessed strands	$0 \ldots 100\%$	5%
polymerisation	**operation parameters**		
	• tube name		
	side effect parameters		
	• point mutation rate	$0 \ldots 100\%$	0.1%

operation	parameter	range	default
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%
gel electrophoresis	**operation parameters**		
	• tube name		
	• minimum number of strand copies with same length, necessary for detection	$1 \ldots 10^6$	
	• selection of available length (bands)		
	side effect parameters		
	• strand loss rate	0 . . . 100%	1%
	• rate of strands with forged length	0 . . . 100%	1%
	• maximum length derivation (forgery)	0 . . . 100% of L	10%

4 Basic ideas of the simulation tool

A simulation tool based on the molecular biological processes from section 2 including optional side effects from section 3 contributes to the experimental setup in the laboratory and is able to explain unexpected results. Our approach extends the idea from [2]. The main features of the simulation tool focus on:

- Specification of DNA operations is set on the level of single nucleotides and strand end labelings 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 test tube that collects a set of DNA strands. The simulation tool is able to manage several test tubes.
- Each DNA operation is characterized by a set of specific parameters and side effect parameters that can be stored and load together with all test tube contents as a project.
- 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 3. 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 [9]. The object-oriented simulation approach is suitable for a wide range of applications, e.g. [3].

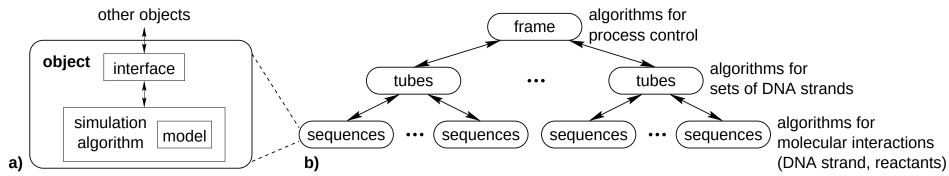

Fig. 3. basic object structure **a)** and hierarchical composition of objects **b)**

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 Comparison of simulation and reality

Two examples (PCR and Cut) were selected to confirm simulation results by laboratory experiments. Both examples compare simulation and laboratory experiment and support an explanation of side effects to be seen in the agarose gel photos.
PCR example: A PCR example should illustrate the consequence of deletions and point mutations in synthesized DNA single strands to subsequent iterated PCR cycles. For laboratory implementation, the PCR template was constructed by oligonucleotide synthesis of two complementary DNA single strands named template1 (`5'-AGGCACTGAGGTGATTGGCAGAAGGCCTAAAGCTCACTTAAGGGCTACGA-3'`) and template2 (`5'-TCGTAGCCCTTAAGTGAGCTTTAGGCCTTCTGCCAATCACCTCAGTGCCT-3'`), both 50 bases (Perkin-Elmer) as well as the primers named primer1 (`5'-AGGCACTGAGGTGATTGGC-3'`) and primer2 (`5'-TCGTAGCCCTTAAGTGAGC-3'`), both 19 bases (Amersham Pharmacia Biotech). The PCR according to standard protocols was done in four samples using 30 cycles including one sample without Taq-Polymerase as negative control. The PCR product was visualized by agarose gel electrophoresis, see figure 4, box below. Lanes 2 until 4 show the amplified band and below a weaker smear of shorter DNA fragments that has to be comprehended by simulation with side effects.

The simulation uses 1000 copies of template1 considering a point mutation rate of 0.06% and a deletion rate of 0.06%, maximum deletion length 12 bases. These side effect parameters were adapted from properties of oligonucleotide synthesis. Template2 was generated in a same way. 8000 copies from each primer (point mutation rate 0.06%, no deletions) were used. All subsequent DNA operations were assumed to be perfect inside the example. Figure 4 shows a screenshot of synthesized strands after union into one common test tube (box above) and of the simulation result after three PCR cycles affirming unwanted shorter bands (box below). The test tube output lists the DNA strands sorted descendingly by number of copies. Screenshots are truncated to the top of the lists.

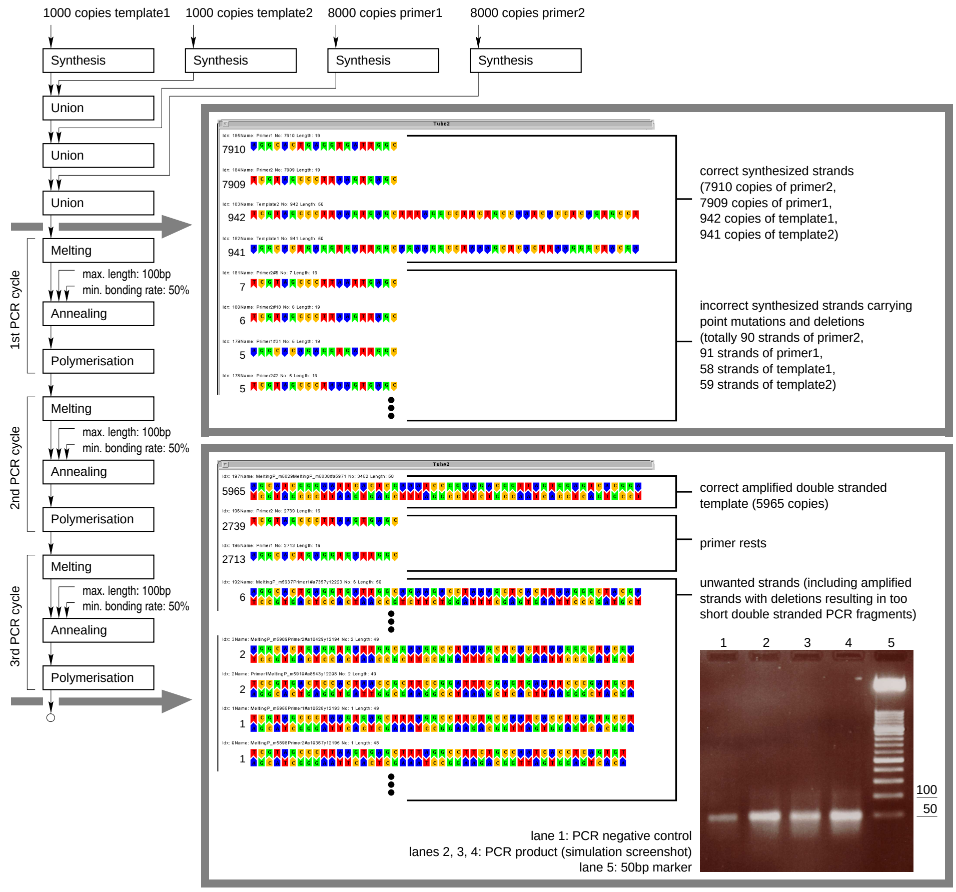

Fig. 4. PCR example: side effect considering simulation result vs. laboratory experiment

The example demonstrates the consequences of point mutations and deletions in synthesized DNA strands to a subsequent PCR decreasing the amount of error free template after three cycles from 8000 expected copies to 5965.

Cut example: A cut example should illustrate incomplete and unspecific reactions by digestion of annealed synthesized oligonucleotides. For laboratory implementation, two complementary DNA single strands named oligo1 (`5'-AGGCACTGAGGTGATTGGCAAGTCCAATCGCGAAAGTCCAAGCTCACTTAAGGGCTACGA-3'`) and oligo2 (`5'-TCGTAGCCCTTAAGTGAGCTTGGACTTTCGCGATTGGACTTGCCAATCACCTCAGTGCCT-3'`), both 60 bases (Perkin-Elmer), were synthesized. Aliquots of each were merged and annealed using standard protocols. The subsequent digestion using NruI, a blunt cutter, should cleave all double stranded fragments in the middle producing only $30bp$ strands. The agarose gel photo shows the result of an incomplete reaction, and base pair mismatching supporting unspecific cleavages, see figure 5.

6 Conclusions

The simulation tool represents a restricted and multiset-based model for DNA computing whose operations were specified and adapted directly from the analysis of molecular biological processes in the laboratory. In contrast to the most 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 assigns to the simulation

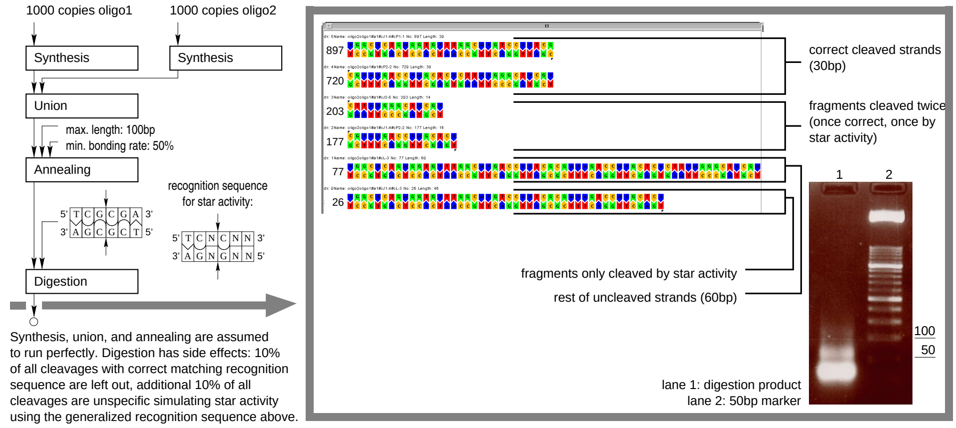

Fig. 5. Cut example: side effect considering simulation result vs. laboratory experiment

tool a high degree of flexibility and ergonomics. The object-oriented simulation approach supports the modelling of interactions between DNA strands and reactants as autonomous subsystems that are combined to test tubes with frame controlled behaviour. 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, PCR, affinity purification, and gel electrophoresis. Further studies focus on the extension to additional effects concerning nonlinear DNA structures.

Acknowledgements. This work is a result of an interdisciplinary collaboration between Institute of Theoretical Computer Science, Institute of Computer Engineering, and Department of Surgical Research, Dresden University of Technology, Dresden, Germany.

References

1. K. Chen, E. Winfree. Error correction in DNA computing: Misclassification and strand loss. In E. Winfree, D.K. Gifford, editors. Proceedings 5th DIMACS Workshop on DNA Based Computers, Cambridge, MA, USA, DIMACS Vol. 54, pp. 49–64, 2000
2. M. Garzon, R.J. Deaton, J.A. Rose, D.R. Franceschetti. Soft molecular computing. In E. Winfree, D.K. Gifford, editors. Proceedings 5th DIMACS Workshop on DNA Based Computers, Cambridge, MA, USA, DIMACS Vol. 54, pp. 91–100, 2000
3. U. Hatnik, J. Haufe, P. Schwarz. Object Oriented System Simulation of Large Heterogeneous Communication Systems. Workshop on System Design Automation SDA2000, Rathen, pp. 178–184, March 13–14, 2000
4. T. Hinze, M. Sturm. A universal functional approach to DNA computing and its experimental practicability. PreProceedings 6th International Meeting on DNA Based Computers, University of Leiden, Leiden, The Netherlands, p. 257, 2000
5. J.A. Joines, S.D. Roberts. Fundamentals of object-oriented simulation. In D.J. Medeiros, E.F. Watson, J.S. Carson, M.S. Manivannan, ed., Proceedings of 1998 conference on Winter Simulation, Washington, USA, pp. 141–150, December 13–16, 1998
6. P.D. Kaplan, G. Cecchi, A. Libchaber. Molecular computation: Adleman's experiment repeated. Technical report, NEC Research Institute, 1995
7. F. Lottspeich, H. Zorbas. Bioanalytik. Spektrum Akad. Verlag Heidelbg., Berlin, 1998
8. M. Sturm, T. Hinze. Distributed Splicing of $\mathcal{RE}$ with 6 Test Tubes. Romanian Journal of Information Science and Technology, Publishing House of the Romanian Academy **4(1–2)**:211–234, 2001
9. G. Zobrist, J.V. Leonard. Object-Oriented Simulation – Reusability, Adaptability, Maintainability. IEEE Press, 1997