Modelling of Function and Evolution of Catalytic RNA

DISSERTATION

eingereicht von

Mag. rer. nat. Bärbel Krakhofer

zur Erlangung des akademischen Grades

Doctor rerum naturalium

an der Formal- und Naturwissenschaftlichen Fakultät

der Universität Wien

Dezember 1998

Ich möchte all jenen herzlich danken, die zum Entstehen dieser Arbeit beigetragen haben.

Peter Schuster gab mir die Möglichkeit zur Dissertation. Er war immer ein anregender Diskussionspartner und sorgte für die Bereitstellung aller Ressourcen am Institut.

Walter Fontana verdanke ich viele wichtige Ideen, er hatte stets Zeit für Diskussionen und Ratschläge.

Peter Stadler war ein kompetenter Quell richtiger Antworten auf schwierige Fragen.

Ivo Hofacker stand mir in allen Programmierfragen mit Rat und Tat und viel Geduld zur Seite.

Ich danke Ronke Babajide, Judith Jakubetz, Susanne Rauscher, Thomas Griesmacher, Alex Renner, Stephan Kopp, Herbert Kratky, Robert Happel, Christoph Flamm, Jan Cupal, Stefan Müller, Norbert Tschulenk, Stefan Wuchty, Martin Fekete, Günther Weberndorfer und Christian Haslinger für das angenehme Arbeitsklima und die schöne Zeit am Institut.

Vor allem danke ich meinen Eltern für ihre Unterstützung und die vielen unbeschwerten Jahre.

Abstract

A collision flow reactor model has been developed to simulate evolutionary behaviour in populations of catalytically active RNA molecules. Randomly colliding RNA strings interact by folding into their cofolded secondary structure. This secondary structure is computed using a dynamic programming folding algorithm; if certain predefined structural motifs occur, the sequences are assumed to act as ribozymes catalyzing cleavage or ligation reactions at defined catalytic sites. These reactions lead to the formation of new sequences and are a main source of population diversity.

Influences of various boundary conditions, such as conservation of mass during reactions, preferred replication of reactive sequences, constant organization and different replication error rates and cleavage criteria, on the emergence of quasi-stationary sequence distributions are analyzed. The resilience of such quasi-stationary distributions, where all sequences in the population are produced by a network of catalytic interactions between members of the population, against changes of boundary conditions is investigated.

In simulations with conservation of mass the populations did not evolve to sequence distributions with self-sustaining reaction networks. Under the constraint of constant organization setting - reaction substrates remain in the reactor, the total number of strings is kept constant by an unspecific dilution flux - reactivity quickly increases and the population converges to some self-sustaining sequence distribution.

If reaction substrates are replaced by mutants instead of being left unchanged in the reactor, population development strongly depends on replication accuracy. At low error rates we observe the formation of quasi-stationary string length distributions, with interactions between members of different string length classes resulting in similar structures that fulfill reaction criteria. These subpopulations of same string length give rise to a self-sustaining network of reactions. At higher mutation rates string lengths are randomly

distributed and reactivity stays low.

Self-sustaining populations which resulted from simulations without mutation were perturbed by introducing erroneous replication at error rates that were previously found to allow for the development of reaction networks. The original sets of sequences quickly disappeared and were replaced by quasi-stationary string length distributions that were again self-sustaining.

Deutsche Zusammenfassung

Ein Kollisions-Flußreaktormodell wurde entwickelt, um evolutionäres Verhalten in Populationen katalytisch aktiver RNA-Moleküle zu simulieren. Zufällig kollidierende RNA-Strings interagieren durch Faltung in eine gemeinsame Sekundärstruktur. Diese Sekundärstruktur wird mit Hilfe eines "dynamic programming"-Faltungsalgorithmus berechnet; wenn bestimmte vordefinierte Strukturmotive auftreten, können die Sequenzen als Ribozyme wirken und Schneide- oder Ligationsreaktionen an katalytischen Zentren in diesen Strukturmotiven katalysieren. Diese Reaktionen führen zur Entstehung neuer Sequenzen und sind eine Hauptquelle der Populationsvielfalt.

Einflüsse verschiedener Randbedingungen, wie Massenerhaltung bei Reaktionen, bevorzugte Replikation reaktiver Sequenzen, "constant organization" und verschiedene Replikationsfehlerraten und Schneidekriterien, auf das Entstehen quasi-stationärer Sequenzverteilungen werden analysiert. Die Elastizität solcher quasi-stationärer Verteilungen, in denen alle Sequenzen einer Population durch ein Netzwerk katalytischer Wechselwirkungen zwischen Mitgliedern der Population erzeugt werden, gegenüber Änderungen der Randbedingungen wird untersucht.

In Simulationen mit Massenerhaltung bei Reaktionen entwickelten sich die Populationen nicht zu Sequenzverteilungen mit selbsterhaltenden Reaktionsnetzwerken. Mit der Randbedingung "constant organization" - Substrate einer Reaktion verbleiben im Reaktor, die Gesamtzahl der Strings wird durch einen unspezifischen Verdünnungsfluss konstant gehalten - steigt die Reaktivität rasch an, und die Population konvergiert zu einer selbsterhaltenden Sequenzverteilung.

Wenn Reaktionssubstrate durch Mutanten ersetzt werden, statt unverändert im Reaktor zu verbleiben, hängt die Populationsentwicklung stark von der Replikationsgenauigkeit ab. Bei niedrigen Fehlerraten beobachten wir die Bildung quasi-stationärer Kettenlängenverteilungen, wobei Wechselwirkungen

zwischen Mitgliedern verschiedener Kettenlängenklassen zu ähnlichen, die Reaktionsbedingungen erfüllenden Strukturen führen. Diese Subpopulationen gleicher Kettenlänge bilden ein selbsterhaltendes Reaktionsnetzwerk. Bei höheren Mutationsraten sind die Kettenlängen zufallsverteilt, und die Reaktivität bleibt niedrig.

Selbsterhaltende Populationen, die aus Simulationen ohne Mutation hervorgegangen waren, wurden durch Einführung fehlerhafter Replikation mit Fehlerraten, die zuvor zum Auftreten von Reaktionsnetzwerken geführt hatten, gestört. Die ursprünglichen Sequenzen verschwanden rasch und wurden durch quasi-stationäre Kettenlängenverteilungen, die wiederum selbsterhaltend waren, ersetzt.

Contents

1 Introduction **3**

1.1 The RNA World 6

1.2 Organization of This Work 7

2 RNA Secondary Structure Prediction **9**

2.1 Definitions 11

2.2 Secondary Structure Representation 12

2.3 Structure Analysis by Phylogenetic Comparison 14

2.4 Energy Directed Folding 15

2.5 Folding Algorithms 17

2.5.1 Kinetic Algorithms 17

2.5.2 Dynamic Programming Folding Algorithms 18

3 Ribozymes **23**

3.1 The Hammerhead Ribozyme 25

3.2 The Hairpin Ribozyme 29

3.3 Evolutionary Biotechnology 30

4 Evolution Reactors and Other Reactors **33**

4.1 Serial Transfer Experiments 33

4.2 The Continuously Stirred Tank Reactor 34

4.3 The Recycling System 34

4.4 The Evolution Reactor 35

4.5 The Catalytic RNA Collision Reactor 38

5 Numerical Results **47**
5.1 Conservation of Mass During Reaction 47
5.1.1 No Replication 47
5.1.2 Autocatalytic Replication 48
5.1.3 Removal of Unreactive Strings 55
5.1.4 Preferred Replication of Reactive Strings 55
5.2 Constant Organization 62
5.2.1 Runs Without Additional Replication Events 62
5.2.2 Autocatalytic Replication 67
5.2.3 Preferred Replication of Reactive Strings 67
5.2.4 Mutation During Reaction 71
5.3 Analysis of Self-Sustaining Sequence Distributions 77
5.3.1 Hammerhead Cleavage 77
5.3.2 Cleavage of Hairpin Loops 87
5.4 Stability of Self-Sustaining Sequence Distributions 90
5.4.1 Stability Against Addition of Random Sequences . . . 90
5.4.2 Stability Against Mutation 93

6 Conclusions and Outlook **98**

References **103**

1 Introduction

Almost 140 years after Charles Darwin's mile stone on the way to modern biology, *"The Origin of Species"*, it is common knowledge that the diversity and complexity of present day organisms is the result of biological evolution.

The Darwinian optimization process which proceeds via variation and selection can be thought of as an uphill walk on a fitness landscape that can find the best suited forms within a given class. However, this process does not lead to radical innovation and is not able to drive the major transitions and cause the apparent jumps in complexity, such as the step from prokaryotes to eukaryotes. To model innovation in evolution other approaches, for example modular design and symbiontic mechanisms have to be made [46, 60, 70].

Since the times of Darwin and Mendel we have learned a lot about the sources of variation: we know that DNA - or in some cases RNA - is the molecule that encodes all information necessary to build and run all complex organisms, we know about genes, inheritance, mutation and the benefits of sexual reproduction.

While variation takes place on the genotype level which can at least in principle be fully described by the nucleotide sequences in a cell's genetic material, selection works on the level of the various functional properties of the complex phenotype.

Scarceness of resources in the environment leads to competition between individuals: those who exploit their environment better and have more (fertile) progeny have a higher chance to be selected.
But how does nature evaluate a phenotype's "fitness" in the competition? Variants that have more or more fertile offspring will increase in further generations, less effective competitors will die out. This *a posteriori* definition of fitness leads to the unsatisfactory reduction of the phrase "survival of the fittest" to the tautology "survival of the survivor".

Since we lack understanding of the relation between genotypes and phenotypes, the outcome of the selection process cannot be predicted even for very simple systems: Fitness is a combined and weighted result of all the different functions of a molecule, cell or organism; these functions are properties of the phenotype.

The problem of predicting fitness from genotype can thus be divided into

- prediction of phenotype from genotype
- prediction of functions from a phenotype
- mapping these functions to the fitness of the individual.

In search of a simple model system to study different aspects of evolutionary behaviour we are inevitably attracted by RNA molecules. Currently they offer the only tractable system to study genotype-phenotype relationships. In the RNA case genotype and phenotype are two features of the same molecule: the sequence of the bases adenine, guanine, cytosine and uracil in the nucleotide chain defines the genotype, while the phenotype is given by the spatial structure of the folded polymer. Relating genotype to phenotype in this case reduces to the problem of prediction of structure from sequence.

The determination of the exact three-dimensional structure of RNA molecules is a hard task and has been solved only for a few relatively small molecules such as tRNAs. RNA is a polyelectrolyte and thus solvatation in aqueous environment has to be considered. Fortunately, in the RNA case we do not need a structure to be resolved up to atomic coordinates; the secondary structure, which is described by the pattern of Watson-Crick base pairs, provides us with a very good discrete coarse graining of the 3D-structure. It is much more easily accessible by experiments and can be predicted and compared with the help of computer programs. Sets of energy parameters derived from melting experiments on small oligonucleotides enable us to compute a sequence's minimum free energy secondary structure. Secondary structures are often conserved on evolutionary time scales.

Since the number of sequences of a given chain length is much higher than the number of different secondary structures many sequences have to fold into the same structure. The fraction of common structures is small and decreases with increasing chain length, while the percentage of sequences folding into these common structures increases with chain length [59, 62, 65].

Moreover, populations of RNA molecules are capable of showing evolutionary behaviour [57]. The first experiments on *in vitro* evolution of RNA were performed by Sol Spiegelman [69]; by using a Qβ-RNA replication assay he did not only show that RNA strings can be replicated outside of cells, but also that evolution can take place in a test tube. Variation was created by replication errors of the enzyme Qβ-RNA-replicase. In serial transfer experiments RNA sequences that are better substrates for the replicase and are therefore faster replicated by the enzyme were enriched and the rate of RNA replication was speeded up by more than one order of magnitude.

A theoretical model of molecular evolution based on ordinary differential equations describing correct and erroneous replication was proposed by Manfred Eigen and Peter Schuster [16, 18]. This model shows that not a single fittest sequence is selected in evolutionary competition; stationary states of the population are characterized by sequence distributions, the so called "quasispecies" [17]. Computer simulations of an RNA evolution reactor, in which the predicted secondary structure of a sequence determined its replication and degradation rate constants, have been performed by Walter Fontana et al. [25, 27]. The properties of fitness landscapes that are obtained by folding RNA sequences into their secondary structures have been subject of extensive studies [58, 61, 64, 66].

Besides all these properties that make RNA molecules the perfect toy for evolutionary biologists there is another good motivation to use RNA as a model to study evolution: RNA could have stood at the beginning of all life as we know it.

1.1 The RNA World

In living cells nucleic acids and proteins strongly depend on each other. Nucleic acids store the heritable information needed for metabolism and reproduction; they work as a template for the synthesis of proteins. The latter express this information by specifically catalyzing the different chemical reactions necessary for cell metabolism, including RNA replication and protein synthesis.

If nucleic acids contain the heritable genetic information required for protein synthesis, and if proteins are required for nucleic acids synthesis and replication, the chicken-egg question is raised: Which came first in evolutionary history, nucleic acids or proteins, or did they arise simultaneously and developed their relation by coevolution?

A possible solution to this problem came up when the long held biochemical principle of strict division of labor in the cell between nucleic acids and proteins was overturned by the discovery of catalytic activity of RNA molecules by Thomas Cech [10, 11]. RNA thus can function both as a heritable-information encoding molecule and as an enzyme. Life may have evolved in a prebiotic RNA world preceding our present DNA-RNA-protein world [31] in which self-replicating RNA molecules served both as the chicken and as the egg.

The catalytic abilities of RNA might not be as universal as those of protein catalysts, but seem to be sufficient for processing and replicating RNA molecules under prebiotic conditions. Ribozymes capable of catalyzing cleavage and ligation of the sugar-phospate backbone of RNA have been found in nature; complementary base pairing allows for sequence specificity of these reactions. With the help of evolutionary biotechnology techniques ribozymes that can catalyze reactions of non-polynucleotide substrates such as the isomerisation of biphenyl [54] have been isolated.

Exploring the structure and function of ribozymes has been a major field of scientific interest in recent years; often the catalytically active core of RNA enzymes is found to be a rather small structure motif.

If we can model the basic enzymatic activities known to be performed by ribozymes *in vivo*, we are provided with an ideal system to simulate functional properties of the phenotype. In this case the phenotype is given by the structure of interacting RNA molecules that collide in an evolution reactor.

1.2 Organization of This Work

The idea of this work is to model evolutionary behaviour of populations of catalytically active interacting RNA molecules. A link between genotype, phenotype and function - in this case sequence, secondary structure and enzymatic activity - is established. Randomly colliding RNA strings are folded into their secondary structure using a dynamic programming folding algorithm.

The resulting secondary structures are analyzed: if predefined motifs that resemble structures known to occur *in vivo* in the catalytic core of ribozymes are found, the RNA sequences in the model can perform certain ribozymal activities. Catalysis of phosphodiester cleavage and ligation reactions is simulated; these reactions result in the production of new strings and are a powerful source of population diversity.

The tools for prediction, description and comparison of RNA secondary structures are described in chapter 2. Folding algorithms that predict structures of interacting RNA molecules, circular molecules and alternative structures besides the one with minimal free energy are introduced.

Chapter 3 gives a short overview of catalytically active RNA molecules, especially the hammerhead ribozyme, the catalytic core of which serves as a model for a cleavage criterion in the collision flow reactor, and the hairpin ribozyme.

In chapter 4 the setup for the evolution collision flow reactor, secondary structure criteria for cleavage and ligation and various boundary conditions such as conservation of mass during reaction, constant organization, preferred replication of catalytically active strings and varying mutation rates during replication are described.

The population developments of collision reactor simulations under different conditions are the topic of chapter 5; quasi-stationary sequence distributions are analyzed and their stability against changes in the environment such as a change in replication accuracy or addition of random strings are discussed.

In chapter 6 future aspects of the results are discussed.

2 RNA Secondary Structure Prediction

RNA is a biopolymer consisting of a sugar-phosphate backbone built from ribose units linked by 3',5'-phosphodiester bonds; each sugar carries a purine or a pyrimidine base on its 1' carbon atom. The molecule is a polyelectrolyte and is well solvatized in aqueous environment.

The sequence of the bases adenine, cytosine, guanine and uracil along the ribonucleotide chain is called the primary structure of the RNA molecule; this linear molecule forms a complex three-dimensional structure.

Complementary strands of RNA can form stable double helices quite similar to DNA. Single stranded RNA molecules can partially accomplish double helix formation by folding back onto themselves and forming Watson-Crick base pairs ($G \equiv C$ and $A = U$) or less stable $G - U$ pairs. The main stabilizing energies that are the driving forces for RNA structure formation arise from the coaxial stacking of the base pairs, hydrogen bonding makes a smaller contribution to the free energy, since the bases are solvatized in aquaeous environment.

The folding of an RNA molecule into its spatial structure can be partitioned into two steps: the formation of Watson-Crick-type and G-U wobble base pairs gives rise to the secondary structure; this secondary structure is then folded into a three-dimensional object.

The 3D-structure of RNA is hard to determine; X-ray structure analysis of crystallized RNA has been done only for very few molecules such as tRNA, the group I intron and the hammerhead ribozyme [6, 43, 52]. The calculation of an RNA molecule's tertiary structure with minimum free energy is a hard task, since algorithms will soon be trapped in a local optimum.

RNA secondary structures provide a useful and biologically plausible coarse graining of folded RNA molecules:

- The main part of the free energy of a folded RNA string is covered by

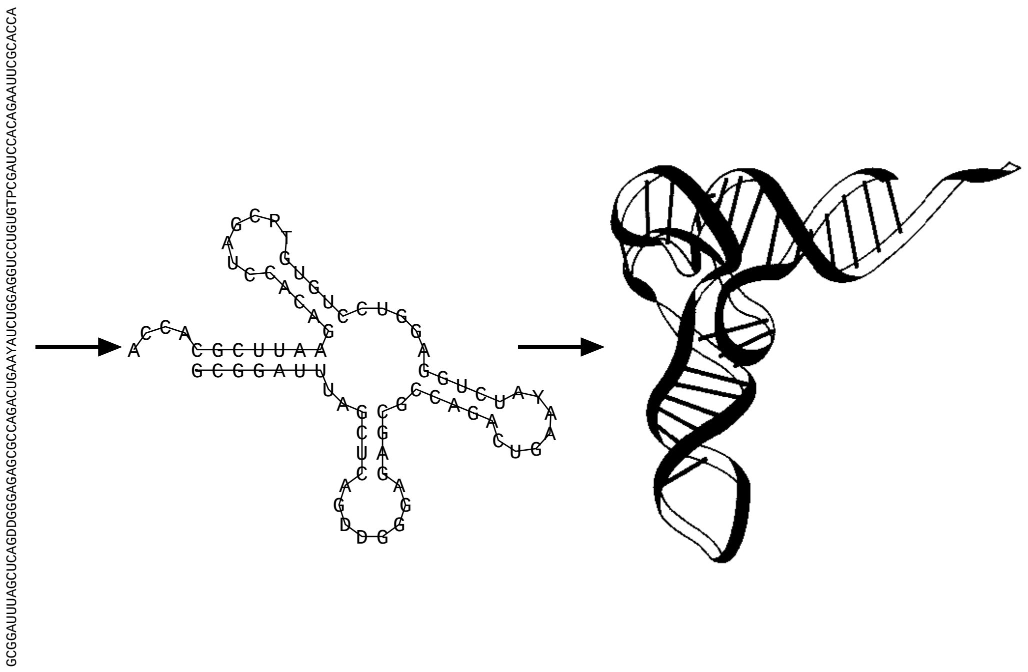


Figure 1: Folding of an RNA sequence into its spatial structure. This folding is a two step process: in the first step only the Watson-Crick-type base pairs are formed; this step contributes the major fraction of the free energy. In a second step the actual spatial structure is built by folding this secondary structure into a three-dimensional object. The example shown here is phenylalanyl-transfer-RNA (tRNA^{phe}); its spatial structure is known from X-ray crystallography.

base pairing and base pair stacking

- Phylogenetic comparison shows that secondary structures have been conserved during evolution
- Secondary structures are discrete and easy to compare
- They can be visualized as planar graphs
- In contrast to tertiary structures, algorithms exist for the computation of the global optimum in the space of secondary structures.

2.1 Definitions

Definition 1. [76] A *secondary structure* is a vertex-labeled graph on n vertices with an adjacency matrix A fulfilling

1. $a_{i,i+1} = 1$ for $1 \leq i < n$;
2. For each i there is at most a single $k \neq i-1, i+1$ such that $a_{ik} = 1$;
3. If $a_{ij} = a_{kl} = 1$ and $i < k < j$ then $i < l < j$.

An edge (i,k), $|i-k| \neq 1$ is called a bond or base pair. A vertex i connected only to $i-1$ and $i+1$ is called unpaired. Condition 3 assures that the structure contains no pseudo-knots. A vertex i is said to be *interior* to the base pair (k,l) if $k < i < l$. If, in addition, there is no base pair (p,q) such that $p < i < q$ we will say that i is *immediately interior* to the base pair (k,l). A base pair (p,q) is said to be (immediately) interior if p and q are (immediately) interior to (k,l).

Definition 2. A secondary structure consists of the following structure elements

1. A *stack* consists of subsequent base pairs $(p-k, q+k)$, $(p-k+1, q+k-1)$, ..., (p, q) such that neither $(p-k-1, q+k+1)$ nor $(p+1, q-1)$ is a base pair. $k+1$ is the *length* of the stack, $(p-k, q+k)$ is the terminal base pair of the stack. Isolated single base pairs are considered as stacks as well.

2. A *loop* consists of all unpaired vertices which are immediately interior to some base pair (p, q), the "closing" pair of the loop.

3. An *external vertex* is an unpaired vertex which does not belong to a loop. A collection of adjacent external vertices is called an external element. If it contains the vertex 1 or n it is a free end, otherwise it is called joint.

If a stack ends in a base pair (p, q) with no unpaired vertices immediately interior to it we speak of a loop with size zero.

Definition 3. The degree of a loop is given by 1 plus the number of terminal base pairs of stacks which are interior to the closing bond of the loop. A loop of degree 1 is called *hairpin (loop)*, a loop of a degree larger than 2 is called *multiloop*. A loop of degree 2 is called *bulge* if the closing pair of the loop and the unique base pair immediately interior to it are adjacent; otherwise a loop of degree 2 is termed *interior loop.*

2.2 Secondary Structure Representation

A secondary structure $\mathcal{S}$ can be represented by a string S by applying the following rules:

- If vertex i is unpaired then $S_i =$ "."
- If (p, q) is a base pair and $p < q$ then $S_p =$ "(" and $S_q =$ ")"

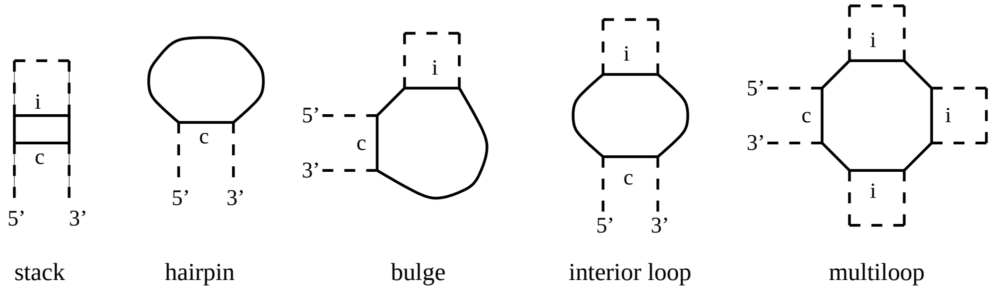


Figure 2: The basic elements of RNA secondary structure. Closing base pairs are denoted by c, interior base pairs by i.

These rules yield a sequence of matching brackets and dots called *bracket notation*.

This bracket notation implies an equivalent representation as a tree. A secondary structure $\mathcal{S}$ can be translated into a rooted ordered tree (linear tree) T by representing a base pair (p, q) by a node x such that the daughters $y_1, \ldots y_k$ of x correspond to the base pairs $(p_1, q_1) \ldots (p_k, q_k)$ immediately interior to (p, q) [25]. For each unpaired vertex z a half-node (leaf) is added to the node representing the closing pair of the loop containing z. An additional node is added as the *virtual root* of the tree that is the mother of all nodes representing external digits and terminal base pairs. This assures that secondary structures with free ends are not represented by a forest. An example is given in figure 3.

A useful method to compare secondary structures are tree edit distances. The application of this method to RNA secondary structure was first proposed by Shapiro and Zhang [68]. The task is to find a sequence of editing steps such as to transform a tree $\mathbf{T}_1$ into a tree $\mathbf{T}_2$ with minimal cost. The allowed edit operations here are deletion and insertion of a node and the total cost of a transformation is given by the sum of the costs of the individual editing operations taken from a cost table. Computation of tree edit distances can be

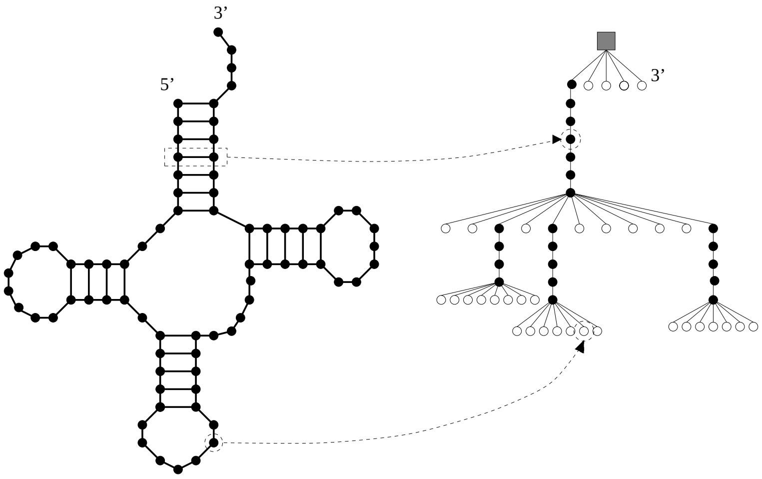


Figure 3: Tree representation of the secondary structure of tRNA^{phe}. Internal tree nodes (black) correspond to base pairs, leaf nodes (white) to unpaired nucleotides. The grey root node is a virtual parent to external elements. Stacks are represented by ropes of internal nodes, loops by bushes of leaves.

done by a dynamic programming algorithm which is included in the Vienna RNA Package [40]. In the tree edit distances computed in this work insertion or deletion of a leaf node adds 1, that of an internal node (corresponding to a base pair) adds 2 to the total costs.

There are several methods to predict RNA structures from a given sequence information. These prediction methods can be divided into two major classes: Folding by *phylogenetic comparison* and *energy directed* folding.

2.3 Structure Analysis by Phylogenetic Comparison

Since the structure of a molecule determines its function, it is assumed that structure is more conserved during evolution than sequence.

Given a large enough number of homologous sequences with identical secondary structure, this structure can be deduced by examining covariances of nucleotides in these sequences [32]. With this method the structure of 16S ribosomal RNAs has been successfully predicted [79]; the clover-leaf structure of tRNAs can easily be found by comparing just a few sequences.

Compensatory mutations such as an A $\rightarrow$ C change in position i of the aligned sequences occurring simultaneously with a change from U to G in position j indicate a base pair (i, j). With this method non-canonical base pairs and tertiary interactions can be detected as well, since an assumption of base pairing rules is not necessary.

Comparative structure analysis currently allows for the most reliable prediction of RNA secondary structure and is therefore frequently used as a comparison for other folding methods. The prediction is accurate as long as the pool of homologous sequences is large enough and shows the proper amount of variation; if the sequences are too similar they do not provide enough covariation data, while for very dissimilar sequences a good alignment is hard to find.

Phylogenetically determined structures usually do not show all actually occurring base pairs, since the method fails for conserved parts of the sequence, where function is sequence dependent, and for non-functional and thus variable parts of sequence and structure.

2.4 Energy Directed Folding

The first attempt to use a crude energy criterion to predict the most stable secondary structure of RNA molecules was made by Nussinov and Jacobson; assuming that base pairing lowers the free energy of a molecule, the "maximum matching" folding algorithm calculates the structure with the highest number of base pairs [49].

Today elaborate models exist for the calculation of the free energy of an RNA secondary structure. The additive model described here is used in the Vienna RNA Package and throughout this work.

The energy of a structure is the sum of independent contributions of the loops of this structure.

$$E(\mathcal{S}) = \sum_{\text{loops } L \text{ in } \mathcal{S}} e(L) + e(L_{ext}), \tag{1}$$

where L_{ext} is the exterior loop containing the free ends. Stacked pairs are treated as minimal loops of degree 2.

The energy parameters $e(L)$ were experimentally derived from melting experiments on small oligonucleotides assuming a nearest neighbour model; the energy of a loop depends only on the size and type of the loop (interior loop, bulge, hairpin) and on the two enclosing basepairs. The parameters most widely in use are from Freier et al. [30].

The major stabilizing contribution to the free energy comes from stacked base pairs; the parallel stacking of the bases is more important than the hydrogen bonds between complementary bases. Energy parameters for all possible combinations of valid basepairs have been measured in several oligonucleotides.

Single unpaired bases adjacent to a helix - dangling ends - can also stabilize the structure by stacking onto the last basepair of the helix. Terminal mismatch energies that are assigned to bases next to interior and closing basepairs of loops depend on the two unpaired bases and the basepair on which they can stack.

Loop energies are destabilizing and depend only on the size and type of the loop (hairpin, bulge, interior loop). The loop energy parameters are rather unreliable, since only few experimental data exist, mostly for hairpins. The minimum loop size for hairpins is 3, values for large loops are extrapolated logarithmically.

It is assumed that the base pairs of bulges of size 1 are stacked; asymmetric interior loops are assigned an additional destabilizing energy depending on the difference of unpaired bases on each side of the loop [50]. Certain hairpin loops of size 4 - especially stable tetraloops - are given an additional bonus energy. Since no experimental data exist on the energy of multiloops their contribution is approximated by the linear ansatz

$$\Delta G = a + bu + cm \tag{2}$$

where u is the number of unpaired bases in the multiloop and m is the number of interior basepairs.

Free ends do not contribute to the free energy, since all energies are measured relatively to the open chain.

2.5 Folding Algorithms

Folding algorithms used for the prediction of RNA secondary structure can be divided into dynamic programming folding algorithms, which always find the globally optimal solution to the problem with respect to the underlying model, and other optimization techniques, such as kinetic algorithms.

2.5.1 Kinetic Algorithms

Kinetic algorithms try to mimic the folding process. The underlying assumption is that during the folding process the structure might be trapped in a local optimum and therefore the biologically relevant structure does not necessarily have to be the one that is thermodynamically most stable. In 1984 Martinez developed the first kinetic algorithm [45]; the energies of possible helices are computed, and it is assumed that the most stable helix will form first. Out of the remaining helices that are compatible to this structure again the most stable one is selected and added to the structure, and so on, until there are no more helices are left that could lower the free energy. In this model refolding is not possible, i.e. helices once formed in the folding process cannot open again. Since the folding of RNA already starts during transcription, helices near the 5' end should be formed first.

A kinetic folding algorithm similar to Martinez' has been implemented by Manfred Tacker in our group; comparison of the structures this algorithm predicted for 16S rRNAs with phylogenetic data did not show significant improvement over the minimum free energy algorithm [71, 72].

2.5.2 Dynamic Programming Folding Algorithms

The fastest algorithms for the prediction of the minimum free energy structure or the equilibrium ensemble of RNA molecules are dynamic programming algorithms [49, 82, 83].

The RNA secondary structure prediction algorithms used throughout this work are based on the dynamic programming folding algorithm for the computation of the minimum free energy structure of an RNA sequence. This algorithm described below in section **(i)** is part of the Vienna RNA Package implemented by Ivo Hofacker [38, 39, 40]; tree edit distances between secondary structures [68] are computed by an algorithm provided in the package. The package is available via the web site **http://www.tbi.univie.ac.at**.

(i) Minimum Free Energy Folding

The dynamic programming folding algorithm for the prediction of the secondary structure with lowest free energy calculates optimal structures for all subsequences of the sequence to be folded, starting at small fragments and proceeding to larger ones. The energy C_{ij} of some substructure enclosed by a base pair (i, j) is the sum of the energy of the loop closed by (i, j) and the energy of any loops directly interior to it.

$$C_{ij} = \min_{\substack{\text{loops } L \\ \text{closed by } i,j}} \left\{ E(L) + \sum_{\substack{\text{interior pairs} \\ (p,q)\in L}} C_{pq} \right\} \tag{3}$$

with $C_{ii} = \infty$.

The minimum energy of the subsequence i, j is given by

$$F_{ij} = \min\Big\{C_{ij},\ (d_{i;i+1,j} + C_{i+1,j}),\ (C_{i,j-1} + d_{i,j-1;j}),\\ (d_{i;i+1,j-1} + C_{i+1,j-1} + d_{i+1,j-1;j}),\ \min_{i \le h < j}(F_{ih} + F_{h+1,j})\Big\} \quad (4)$$

with the d terms being contributions from dangling ends and $F_{ii} = 0$.

This equation is only evaluated for loops of degree $<=$ 2; for multiloops the linear ansatz given in equation 2 is used. The minimum energy of the subsequence between i and j given that i and j are part of a multiloop is stored in another array, F^M_{ij}. If (i, j) is the closing basepair of a multiloop

$$C_{ij} = \min_{i<h<j-1} \left\{F^M_{i+1,h} + F^M_{h+1,j-1}\right\} + a. \quad (5)$$

F^M_{ij} is calculated analogously to equation 4.

After having calculated all entries of the arrays C, F and F^M the structure with lowest free energy is calculated by backtracking through these arrays. This procedure is very fast, since only the entries belonging to the minimum free energy structure have to be recalculated.

(ii) CoFold

In order to be able to simulate ribozymal activities of interacting RNA molecules the folding algorithm was altered to predict the minimum free energy secondary structure of two RNA molecules folding together.

To determine the difference between the secondary structure of two sequences folding together and that of a single string composed by concatenation of these two sequences the extent of refolding was measured for stringlengths $n = 20$ to $n = 100$. Random sequences were generated and cut at a randomly chosen position; the average tree edit distance between the secondary structures of the original string and the cleavage products and the distribution of tree edit distance frequencies for $n = 100$ can be seen in figure 4.

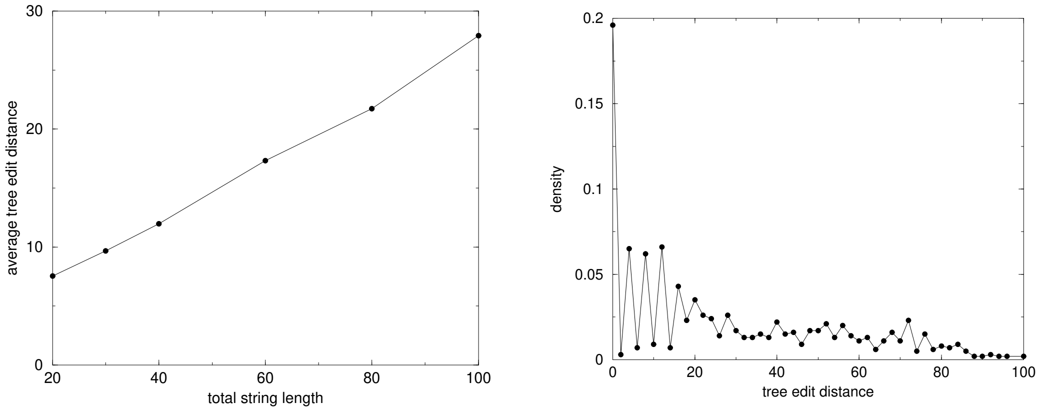


Figure 4: Tree edit distances between the minimum free energy structure of a sequence and the cofolded structure of the two sequences that result from cleavage at a random position (left). Distribution of these tree edit distances for a total string length of $n = 100$ (right). Sample size was 1000 sequences for each string length.

(iii) Circular Folding Algorithm

A dynamic programming algorithm for the folding of circular RNA was implemented following an algorithm proposed by Zuker [81].
In a circular molecule a base pair between bases i and j divides the structure into two parts, the included fragment $\mathcal{S}_{ij}$ and the excluded fragment $\mathcal{S}_{ji}$.

The minimum free energy of the sequence is

$$mfe = \min_{\text{basepairs}(i,j)} \{C_{ij} + C_{ji}\} \tag{6}$$

with C_{ji} being the energy of the best structure of the excluded fragment defined by the basepair (i, j). In circular RNA the choice of origin is arbitrary; to fold a circular sequence I of length n a linear sequence of length $2n$ fulfilling $I(n+i) = I(i)$ is constructed and all matrix entries C_{ij},F_{ij},$F^M_{i,j}$ are calculated for this linear sequence. Then a base pair (i, j) is chosen such that $C_{ij}+C_{j(i+n)}$ is minimal; backtracking through the arrays starting from basepairs (i, j) and $(j, i+n)$ yields the basepairs of the minimum free energy structure.

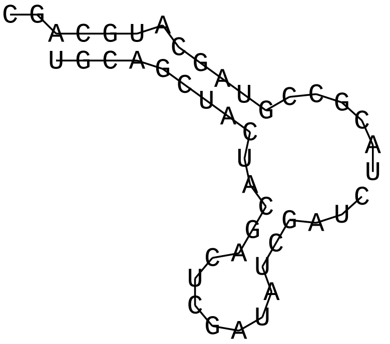
(iiii) Suboptimal Folding Algorithm

Usually numerous foldings exist with energies very close to the minimum free energy; these structures can be completely different from the minimum free energy structure. The idea behind this algorithm to find suboptimal structures is similar to the circular folding algorithm [81]: the minimum energy of a structure $\mathcal{S}(i, j)$ that includes the base pair (i, j) is the sum of the energies of the best foldings of the included fragment $\mathcal{S}_{ij}$ and the excluded fragment $\mathcal{S}_{ji}$,

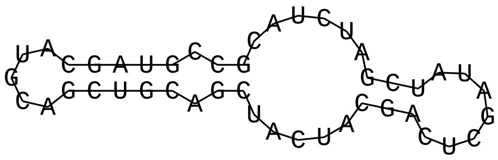
$$E(\mathcal{S}(i, j)) = C_{ij} + C_{ji}. \tag{7}$$

Two sets of arrays C,F and F^M are filled, one for the included and one for the excluded fragment, which contains the 5' (origin) and 3' ends. Then $E(\mathcal{S}(i, j))$ is evaluated for every possible base pair (i, j). If $E(\mathcal{S}(i, j))$ lies within a certain energy range from the mfe, the suboptimal structure is determined by backtracking starting from base pair (i, j) into two directions.

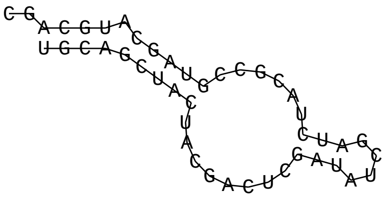
Figure 5 shows different foldings of an example string. The string is cut and the two parts are folded together, the structure of the circularized string, the minimum free energy folding and 3 suboptimal structures lying within a 5 percent energy range from the mfe are shown.



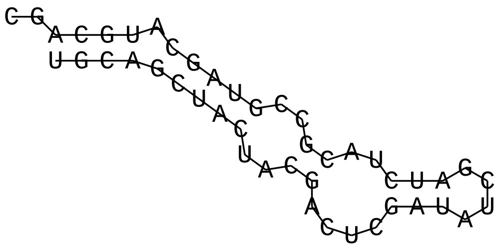
CoFold



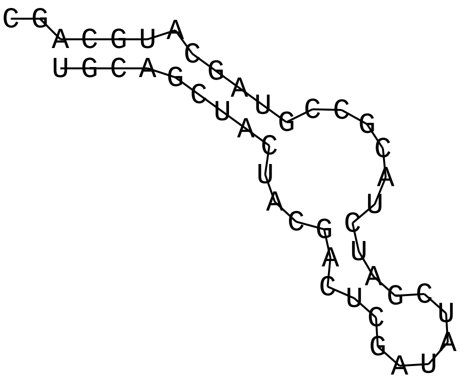
circular folding



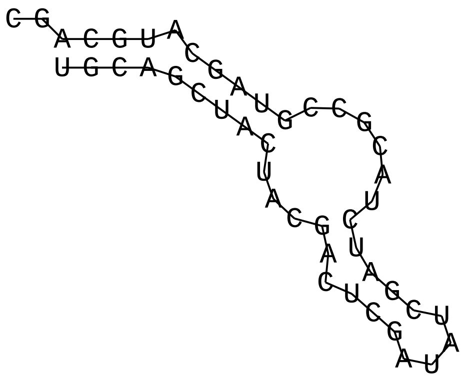
mfe folding



suboptimal structure I



suboptimal structure II



suboptimal structure III

Figure 5: Secondary structures of an RNA sequence as predicted by the algorithms described in in section 2.5.2.

3 Ribozymes

How does a protein enzyme fulfil its catalytic activities? The well defined three-dimensional structure of an enzyme is automatically folded and determined by its primary structure. Enzymes increase the reaction rate in different ways: substrates of the enzyme are bound by chemical groups on the surface, so reactants can be brought into close proximity and have a higher chance to react than they would if they had to collide by chance in solution. During reaction a transition state complex, a structure with high energy, is formed; the higher the activation energy barrier, the slower the reaction rate. Enzymes decrease this activation energy by binding to the reactants and changing their structure. Moreover enzymes are very flexible, by torsion and bending they can guide the reaction. After the reaction has taken place, the products leave the enzyme which can start the next catalytic cycle.

Thomas Cech discovered that an RNA molecule can catalyze its own splicing [8, 10]; the pre-rRNA of the single celled eucaryote *Tetrahymena thermophila* folds back onto itself, exhibiting a complex three-dimensional structure that enables it to excise an intron without the help of protein enzymes. This structure activates two phosphodiester bonds at the 5' and 3' ends of the intron by exposing them to nucleophilic attack. A guanosine is bound by the intron in a position favouring its attack on one of the activated bonds; once the linkage is broken, the newly formed hydroxyl group of one exon attacks and cleaves the other activated bond leading to the excision of the intron and ligation of the two exons. This self splicing closely resembles the action of a protein enzyme: the reactions are highly specific, their rates are accelerated, and the three-dimensional folding of the molecule is essential for its catalytic activity. The term "ribozyme" was coined for this autocatalytically active RNA molecule.

Since the work of Cech many other catalytically active RNA molecules have been discovered [44].

A prerequisite for the existence of an RNA world during early evolution of life is the ability of ribozymes to replicate RNA with high template fidelity; different ribozymes capable of catalyzing reactions necessary for RNA self-replication [7, 9] have been found.

Doudna and Szostak demonstrated the ability of a modified *Tetrahymena* ribozyme to catalyze a ligation reaction that forms an RNA molecule which is complementary to a template strand [13]. In later experiments they used an altered *sunY* intron to ligate oligonucleotides to a primer in a template-dependent manner [14].

An RNA molecule that can add up to 6 mononucleotides to an RNA primer by forming 3',5'-phosphodiester linkages was described by Ekland and Bartel [20]. The ribozyme is a class I ligase derived from a pool of random RNA sequences [19, 21]; it utilizes mononucleoside triphosphates as substrates. The monophosphates are added to the 3'-hydroxyl of the primer retaining Watson-Crick complementarity to a template string, which is covalently bound to the ribozyme; pyrophosphate is displaced.

Complementary base pairing with RNA substrates enables ribozymes to catalyze reactions between RNA molecules in a sequence specific way; all known ribozymes that occur in nature, like the group I and group II introns [56], the hammerhead and the hairpin ribozyme act on the sugar-phosphate backbone.

Ribozymal activities necessary for a transition from an RNA world to modern protein biology include catalyzation of reactions involving carbon centers; Piccirilli et al. engineered the active site of the *Tetrahymena* ribozyme to catalyze the hydrolysis of an aminoacyl ester bond between N-formyl-L-methionine and an oligonucleotide derived from the 3'-end of the corresponding tRNA suggesting that an RNA molecule could have acted as the first aminoacyl tRNA synthetase [51]. For a current overview over the role ribozymes may have played in early RNA replication and protein synthesis, see [33].

3.1 The Hammerhead Ribozyme

The hammerhead ribozyme is one of the few well characterized catalytic RNA motifs. The sequence motif was first recognized in satellite RNAs of plant viroids [53] and was found to catalyze self cleavage which is thought to process replicative intermediates. Since the motif is very small it can easily be synthetized and modified. The structure (Figure 6, numbering according to Hertel [36]) consists of three stems, two non-helical segments and an unpaired nucleotide at the cleavage site [55]. This secondary structure was predicted from the consensus sequence that is necessary for catalytic activity and was supported by mutagenesis experiments [55]. Synthetic oligonucleotides with the hammerhead consensus sequence can enzymatically cleave substrates *in vitro* [34, 75].

The kinetics of the hammerhead cleavage reaction have been explored in great detail [22, 34, 37, 75].

The cleavage rate of the hammerhead ribozyme is typically $\sim 1min^{-1}$. A divalent cation in millimolar concentration (Mg^{++} or Mn^{++}) seems to stabilize the structure necessary for the cleavage reaction. The hammerhead does not cleave DNA strands, since a 2'-hydroxyl has to be present in the cleavage site, but an all-DNA strand with a single ribonucleotide at the cleavage site can act as a substrate, although the turnover rate is substantially lower. The 3D-structure has been solved by Pley et al. [52] by crystallizing a complex of an all-DNA substrate, which acts as an inhibitor for the cleavage reaction, but differs only slightly from a valid substrate, and by Tuschl et al. [74] using fluorescence resonance energy transfer (FRET).

Stems I and II diverge at a small angle at one side of the core, while stem III points in the opposite direction. In the central core of 15 conserved nucleotides five non-Watson-Crick base pairing interactions are observed. The sharp turn of the enzyme strand between stems I and II involves a conserved CUGA sequence. Stem II together with two absolutely conserved reverse-

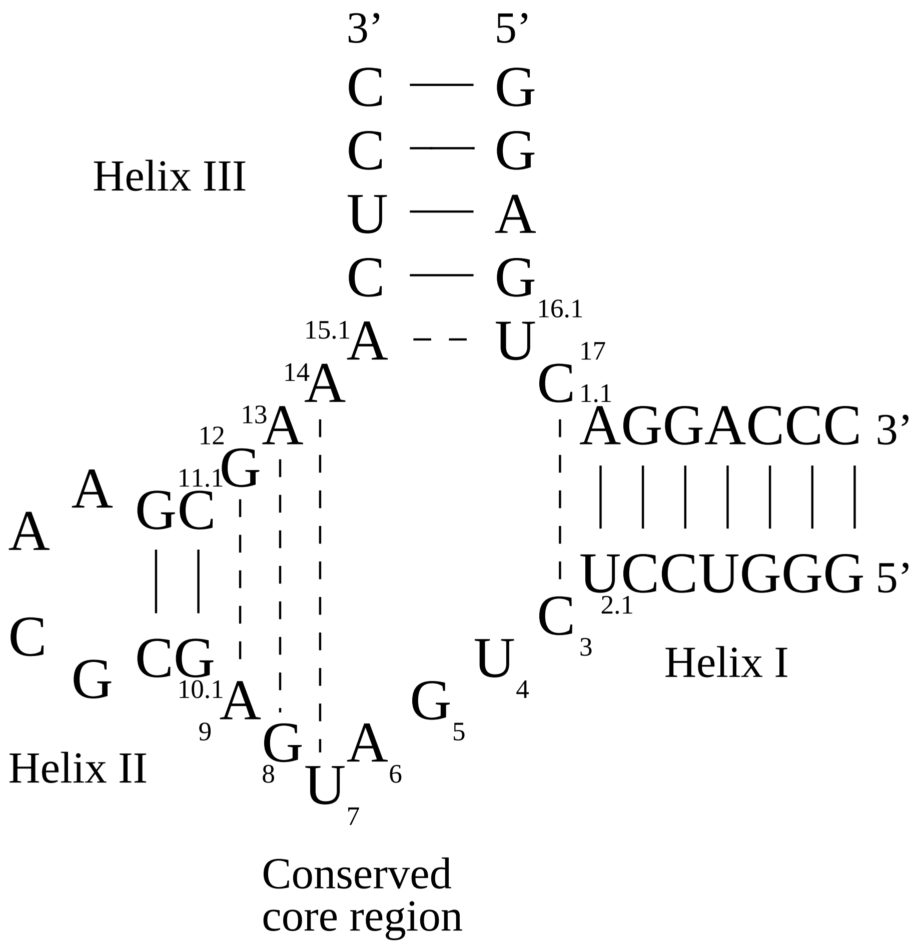


Figure 6: Secondary structure of the hammerhead ribozyme; the sequence is taken from [74], non-standard base pairs (dashed lines) as described in [12]. Stems I and II diverge at a small angle at one side of the central core of 15 conserved nucleotides, stem III is coaxial with the “augmented” stem II helix.

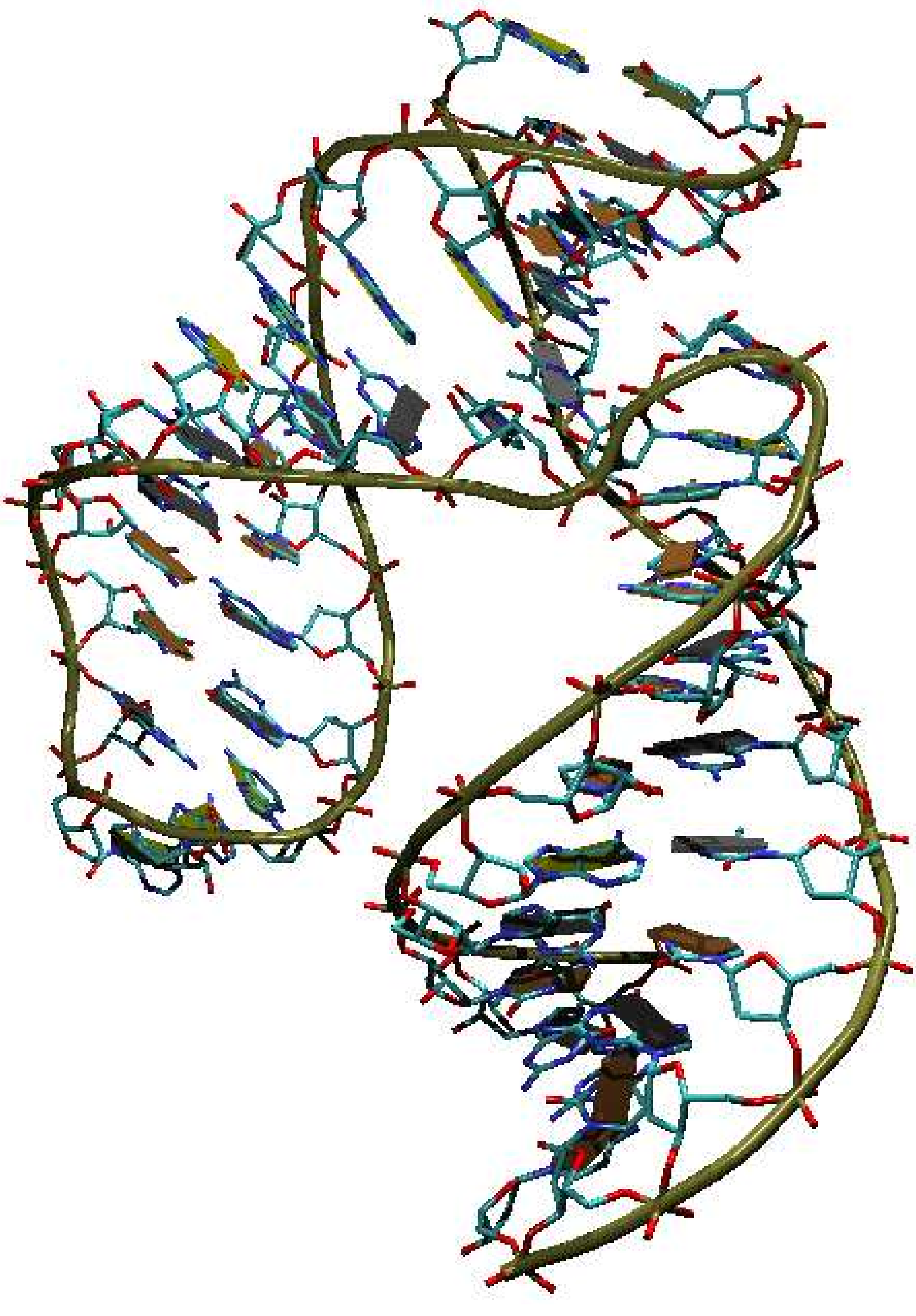


Figure 7: Three dimensional structure of the hammerhead ribozyme. Data from [52]

Hoogsteen GA pairs followed by a singly hydrogen bonded AU base pair form an “augmented” stem II helix which stacks directly upon the absoutely conserved non standard $A_{15.1}$-$U_{16.1}$ base pair of stack III, forcing C_{17} outwards to stack upon the end of stem I [67]. The two conserved GA pairs and a G at the bottom of stem II are the site of metal^{++}- ion coordination. Since this site is not in vicinity of the cleavage site it is assumed that this metal ion is stabilizing the structure and is not involved in the actual cleavage reaction.

Cleavage occurs at a unique site in the motif. The ribozyme cleaves its RNA substrate behind a nucleotide triplet of the general formula NUH, where N is any nucleotide and H is U, C or A but not G; the most efficiently cleaved substrate contains a GUC triplet [15]. The reaction is a transesterification, cutting the 3’,5’ phosphodiester bond between nucleotides 17 and 1.1 producing a cyclic 2’,3’-phosphodiester on nucleotide 17 and a free 5’-hydroxy terminus on nucleotide 1.1 [41] and proceeds with inversion of the configuration at phosphorus. Multiple turnover is possible when the substrate dissociates from the cleaving strand.

Mutagenesis experiments have shown that the cleavage rate is insensitive to base pair exchanges in most helical positions, but only few of the core residues can be altered without reducing turnover rates.

Optimal lengths of helices I and III to obtain higher cleavage rates have been determined by Hendry and McCall [35]; substrates in a complex with shorter helix I and longer helix III are cleaved faster by one to two orders of magnitude than those in complexes with longer helix I and shorter helix III. The influence of variations in helix II on cleavage rates has been investigated [47, 73]. The extent to which ribonucleotides can be replaced by desoxyribonucleotide-analogues while retaining catalytic activity has been determined by Yang et al. [80].

The small size of the hammerhead ribozyme and the fact that it can be easily adapted to cleave specific target sequences [34] makes it a possible candidate for a therapeutic agent that could cleave viral RNA.

3.2 The Hairpin Ribozyme

The hairpin ribozyme was discovered in 1989 [34]; it originates from satellite RNAs associated with certain plant viruses, most commonly from Tobacco Ringspot Virus. It acts as a reversible endoribonuclease and cleaves single stranded RNA via a transesterification reaction creating cleavage products with 5'-hydroxy and 2',3'-cyclic phosphate termini. The reaction is reversible. *In vivo*, the cleavage reaction processes the multimeric concatamers produced by RNA replicase from a (+)-strand template, while ligation cyclizises monomeric (-)-strands which serve as templates for (+)-strand synthesis.

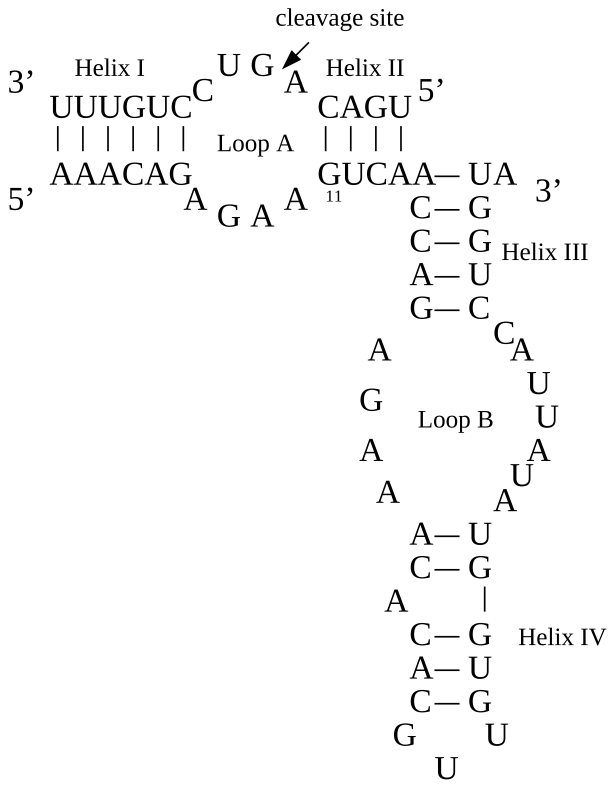


Figure 8: Secondary structure of the complex between the (-)sTRS hairpin ribozyme and its substrate [5].

The secondary structure of the hairpin ribozyme complex consists of four helical elements and two internal loops.

The catalytically active strand is a 50 nucleotide sequence within the (-)-strand of sTRSV RNA. The substrate binds at a 14 nucleotide sequence at the 5' end of the ribozyme, forming two short helices (Helix I and Helix II) on the sides of an internal loop (loop A) that contains the cleavage site. The two helices within the ribozyme (Helix III and Helix IV) enclose a large asymmetric internal loop (loop B). The bases in the helical regions can vary widely, as long as complementarity of pairing bases is maintained; only one G at ribozyme position 11 and a pyrimidine as its partner at the corresponding substrate position 2 nucleotides downstream from the cleavage site are required. By varying bases in the substrate recognition helices I and II of the ribozyme its substrate specifity can be modified.

For an overview of the current status of research on the hairpin ribozyme, see [5].

3.3 Evolutionary Biotechnology

The RNA world model can be extended with the proposal that the scope of RNA catalysis has by no means been restricted to RNA cleavage and ligation reactions. Ribozymes capable of catalyzing a vast variety of reactions could have acted as enzymes in a metabolism before the evolution of encoded protein synthesis.

Since all known naturally occurring ribozymes catalyze only hydrolysis or transesterification of the sugar-phosphate backbone, evolutionary techniques have been developed to generate, isolate and enrich RNA molecules with desired functions.

By iterative rounds of *in vitro* selection and erroneous amplification of RNA sequences exhibiting the properties one is looking for, ribozymes can be "trained" to perform different reactions at faster and faster rates.

A typical *in vitro* evolution experiment to find new ribozymes performing a predefined task starts with a large pool of random RNA sequences generated by linking smaller DNA pools followed by PCR amplification and *in vitro* transcription of the linked pool [1]. RNA molecules that exhibit the desired function are then isolated by affinity chromatography; selected sequences are reverse transcribed, the resulting DNA is PCR-amplified and *in vitro* transcribed, and a new evolution cycle is started (Figure 9). Bartel, Ekland and Szostak successfully used this method to isolate highly active RNA ligase ribozymes [1, 21].

With the help of evolutionary biotechnology ribozymes that can catalyze reactions other than phosphoryl group transfers and bind their substrates by interactions other than Watson-Crick base pairing have been isolated; a ribozyme that can catalyze the isomerization of a biphenyl to its diastereomer was found by screening for the ability to bind a near-planar transition state analogon [54]. An RNA molecule capable of self-alkylation was isolated by Wilson and Szostak in a series of sequential in vitro selections [77].

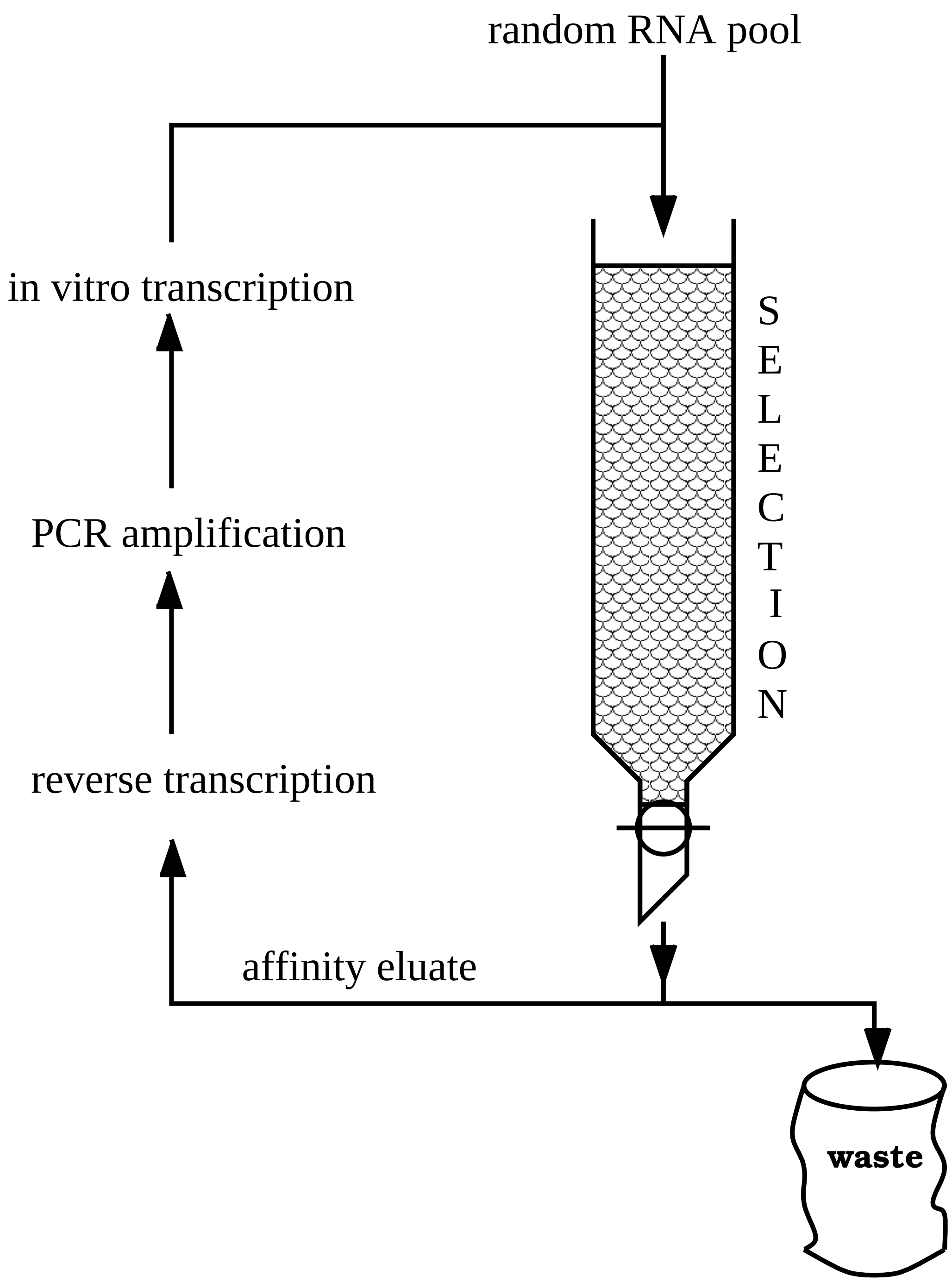


Figure 9: Experimental setting for *in vitro* evolution of RNA using the SELEX method. Catalytic functions or other properties are optimized iteratively in selection cycles. In each cycle affinity chromatography is used to isolate RNA molecules exhibiting desired properties, selected molecules are then amplified and replicated with high error rates.

4 Evolution Reactors and Other Reactors

Various attempts have been made to simulate the evolutionary behaviour of populations of replicating or interacting species or molecules over time. There are different approaches to this task: *in vitro* evolution experiments, theoretical models, and computer simulations of evolving populations.

Open systems that keep polynucleotide replication away from equilibrium can be experimentally realized in different ways. Types of such open systems are serial transfer, the continuously stirred tank reactor (CSTR), the recycling system and the evolution reactor.

4.1 Serial Transfer Experiments

Sol Spiegelman and coworkers did the first optimization experiments on RNA molecules based on the Darwinian variation/selection principle [48, 69]. They made *in vitro* replication assays for RNA molecules using Qβ-RNA replicase; variation is created by replication errors of the enzyme.

In serial transfer experiments RNA templates are added to a solution containing energy rich monomers and Qβ-replicase. After an incubation period a small sample of the reaction mix is transferred to a vial containing fresh solution; in each step this incubation/transfer process is repeated.

Since only single stranded molecules are accepted as templates by the enzyme, the secondary structure has to melt to make replication possible. On the other hand single stranded regions of an RNA molecule can be easily attacked by hydrolytic agents or nucleases.

The rate of replication was speeded up by at least one order of magnitude. The extensive studies of the Qβ-system yielded deep insights into mechanism and kinetic details of Qβ replication and the structural properties RNA molecules have to fulfil in order to be recognized and replicated by the enzyme [2, 3].

These studies on *in vitro* evolution showed that evolutionary phenomena can well be observed with molecules in test tubes and are not constrained to cellular life. The prerequisites to observe test tube evolution are:

- the conditions for replication must be fulfilled.
- variation is created by erroneous replication
- there must be selection pressure by limited resources.

In a finite system with exponential growth rates the latter is always fulfilled.

4.2 The Continuously Stirred Tank Reactor

The kinetics of polynucleotide replication in the continuously stirred tank reactor (CSTR) have been discussed by Peter Schuster and Karl Sigmund [63]. Energy rich material is continuously provided by an influx r that keeps the system away from equilibrium, an evolutionary constraint is imposed on the system by continuous outflux of solution. Replicating template molecules are added to the reactor at time $t = t_0$; depending on the input solution and the flow rate r their concentrations can either increase and reach a stationary value, or they may be diluted out of the reactor.

4.3 The Recycling System

Consumed energy rich monomers are renewed in an irreversible recycling reaction. In this model the evolutionary constraint is provided by means of a degradation process: templates are converted into energy poor material B (in a realistic system B stands for the nucleoside monophosphates AMP, UMP, GMP and CMP). This energy poor material is recycled into starting material A; this recycling reaction might be represented by a photochemical process $B + h\nu \rightarrow A$..

4.4 The Evolution Reactor

The first theoretical model of molecular evolution was proposed by Manfred Eigen in his pioneering work [16]. It deals with the kinetics of replication, mutation and selection in populations of asexually reproducing species. Correct replication and erroneous replication are reactions involving the same template.

This model for polynucleotide replication was based on ordinary differential equations derived from chemical kinetics:

$$\begin{aligned} (A) + I_j &\xrightarrow{Q_{kj}A_j} I_j + I_k \\ I_j &\xrightarrow{D_j} (B) \\ I_i &\xrightarrow{\Phi(t)} 0 \end{aligned}$$

An RNA template sequence I_j is replicated with the rate constant A_j, the probability for the outcome of this process being a sequence I_k is Q_{kj}. The probability for error-free replication of sequence I_j is Q_{jj}. A denotes buffered low molecular weight building materials; degradation of the molecules is taken into account by the degradation rate constant D_j, waste B is constantly removed. An unspecific dilution flux $\Phi(t)$ removes templates from the system.

This reaction network is set up in an evolution reactor (see figure 10). Low molecular weight building material A is added to the reaction mixture by an influx which is regulated such that the concentration of A is kept constant. Degradation products B are steadily removed; outflux is possible through two different channels: one that is impermeable to polymers and one for the well stirred reaction mixture; this dilution flux can be adjusted to keep the sum of numbers of individual polynucleotides and the volume of the reaction mixture constant. This setting is called *constant organization*. The kinetic analysis of this replication-mutation system showed that a sharply defined minimum replication accuracy - the error threshold - exists below which populations become unstable and drift randomly through sequence

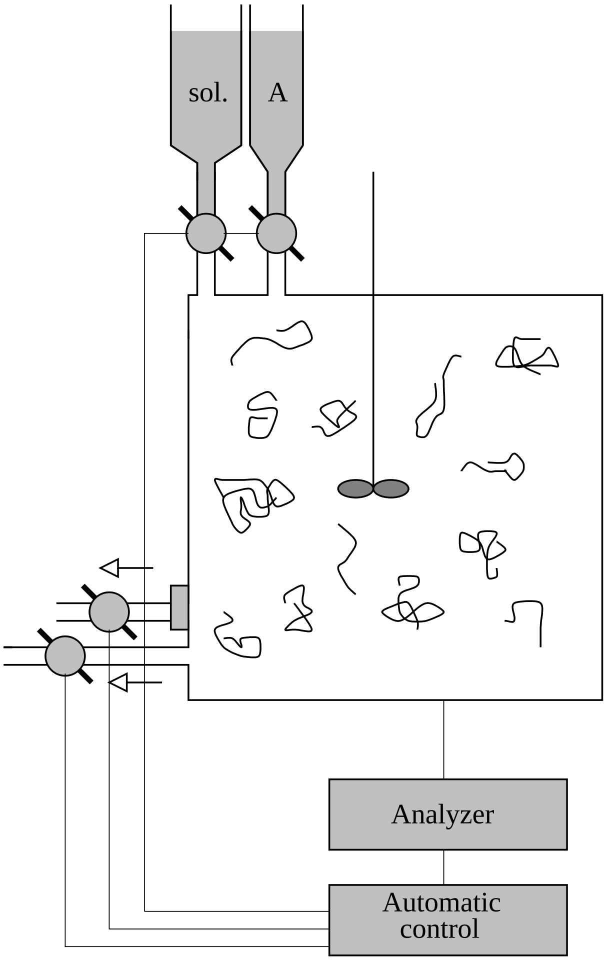


Figure 10: Model of an evolution reactor. The concentration of low molecular weight material (A) is kept constant with the help of a steady influx; replication occurs in the well stirred reaction mixture. Outflux can occur through 2 channels, one of them is impermeable to polymers.

space. Stationary states of a population are characterized by distributions of sequences - so called quasispecies - that cluster around a master sequence. Not a single fittest type, but the quasispecies is selected by the evolutionary process. Such quasistationary states can be destabilized by rare advantageous mutations, and the population moves towards a new quasistationary sequence distribution.

A computer simulation of evolutionary optimization was carried out by Walter Fontana and Peter Schuster [27]. They analyzed a reaction model that simulated correct and erroneous RNA replication together with hydrolytic degradation and a dilution flux in a flow reactor.

Optimization of the population proceeds via mutation and selection. Selective values for replication and degradation were derived from the molecular phenotypes by assigning numerical values to secondary structure elements. Stacked regions are assumed to slow down the replication process, while hydrolytic degradation is eased in unpaired regions. Replication and degradation rate constants are computed for all secondary structures present in the reactor.

In these computer simulation experiments features like replication error thresholds and quasistationary sequence distributions, which are typical for the evolution of populations, could be observed.

In early computer simulations of evolutionary adaptation fitness was computed from kinetic constants derived from RNA structures by a simple model [26]. Recent work on this subject focussed on the distance between secondary structure and a target structure, and on the implications of neutral networks of RNA sequences with identical structure on adaptive evolution [28, 29, 42].

A very interesting approach to model interaction between abstract molecules was made by Walter Fontana and Leo Buss [23, 24]. This model is based on symbolic operators expressed in λ-calculus. λ-expressions can interact by functional application; one expression A acts as an operator on the other expression B which plays the role of an argument. The result C of this interaction, again a valid lambda term, is evaluated by performing a series of reduction steps.

$$(A)B \longrightarrow C \tag{8}$$

Different function-argument combinations can yield the same result; a particular function-argument combination always yields the same result. No chemistry-related assumptions have to be made, no equations have to be set up, yet such an interaction represents in a way a chemical reaction. This elegant model can be used to explore general boundary conditions for evolution. A flow reactor was filled with lambda expressions, after every random collision the resulting term was added to the population, and a randomly chosen expression is removed from the reactor. With this constant organization reaction scheme a motion in object space that frequently converged on self-maintaining systems of λ-expressions could be observed. The λ-expressions of such an ensemble maintain each other by mutual production pathways and share invariant syntactical and algebraic regularities.

4.5 The Catalytic RNA Collision Reactor

In this work computer experiments were designed to simulate evolution in populations of interacting RNA molecules. RNA enzymes act on their substrates in a sequence specific way: by complementary base pairing certain structural motifs are formed; these phenotypic properties lead to well defined reactions such as cleavage and ligation.

In the catalytic RNA reactor collisions induce the formation of a folding which involves both strings, and this phenotype determines the kind of reaction that can take place. Cleavage and ligation reactions give rise to new sequences and thus provide us with an additional source of population diversity besides point mutation.

If the replication frequency of a sequence depends on its reactivity, the fitness value depends on the secondary structures that are formed in collisions with other members of the current population.

Some of the questions that can be addressed in such simulation experiments are:

- How do populations of interacting RNA molecules develop over time?
- Do we find stable sequence distributions?
- Under which boundary conditions can self-sustaining systems evolve?
- Are they stable against changes in the environment such as addition of random strings or higher mutation rates?

In order to simulate interactions between RNA molecules an evolution reactor was designed; collisions between RNA strings are simulated by randomly choosing two strings and folding them together using the dynamic programming algorithm CoFold described in chapter 2. The resulting secondary structure is evaluated; it determines which reaction channels are available to the collision partners. The reactions that can occur during non-elastic collisions are cleavage of a reactant, ligation of the two strings, replication and removal of sequences depending on the secondary structure.

As a secondary structure motif that allows for a molecule to act as a ribozyme and catalyze a cleavage reaction, a multiloop similar to the structure of the hammerhead ribozyme was chosen; figure 12 shows an example.

The ribozyme string has to form a hairpin loop and two adjoining stems that are formed by complementary base pairing with the substrate string; at least one unpaired base has to be present in the loop between the two enzyme-substrate stems, besides this restriction the number of unpaired bases between the stems is arbitrary. Since only the local structure of the catalytic

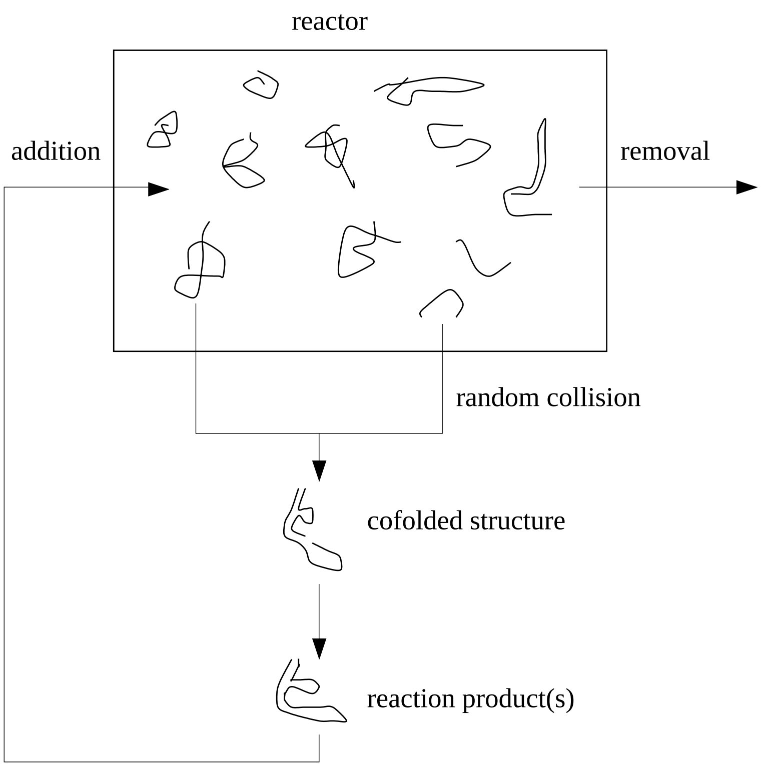


Figure 11: Catalytic RNA flow reactor. Two strings are chosen at random and folded together. Depending on the resulting secondary structure the collision is either elastic or reactive; in the latter case the reaction products are added to the reactor.

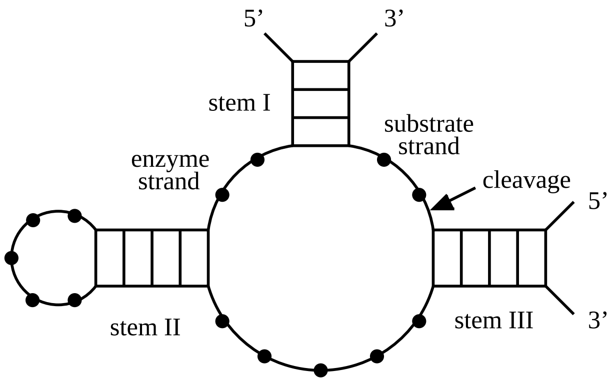


Figure 12: Hammerhead cleavage motif; the substrate is cleaved immediately behind the last paired base of stem III.

core is taken into account the substrate can be part of the ribozyme string as well as of the collision partner. Substrate cleavage occurs after the last base in stem III. In case of multiple cleavage sites in a secondary structure the first hairpin counting from the 5’ end acts as a ribozyme; if both collision partners show hairpins meeting cleavage criteria, the enzyme string is chosen randomly.

In order to enhance cleavage frequency no sequence constraints for the catalytic core are made; the prerequisites for reactions are purely structural.

In contrast to the real hammerhead ribozyme the additional requirement is made that stem II has to end in a hairpin loop.

A different cleavage criterion has been tried in some runs: hairpin loops with at least five unpaired bases are cut in the middle; again the first loop in the sequence counting from the 5’ end is chosen.

The secondary structure motif allowing for ligation has been chosen in analogy to an experiment performed by Ekland, Szostak and Bartel [21]. They isolated highly active RNA ligases from a pool of random RNA sequences by *in vitro* selection methods. RNA molecules that were able to promote the

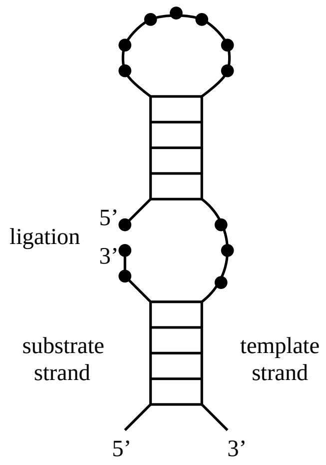


Figure 13: Ligation motif

joining of a substrate oligonucleotide to a complementary template sequence that was covalently linked to the enzyme strand as a 5'-leader sequence were enriched.

Various structural motifs were found to be able to catalyze such a ligation; some of them could also catalyze intermolecular ligations, i.e. formation of a 3',5'-phosphodiester linkage of the substrate to a template that was not part of the ribozyme.

In the evolution reactor runs ligation occurs when parts of the sequence near the 3'-end of one string are complementary to a template region near the 5'-end of the second string; this stem formed by intermolecular base pairing has to "elongate" a stem formed by the second string that ends in a hairpin loop.

For reactivity reasons it is not necessary that the two strings are aligned by Watson-Crick basepairing across the ligation junction; up to 5 unpaired bases at each side of the junction are accepted. The joint connecting the stem regions of the template strand can consist of up to 10 unpaired nucleotides.

A secondary structure motif fulfilling these requirements for ligation is shown in figure 13.

Since folding of the strings into their secondary structure is the time consuming step in collision reactor experiments the maximum string length in the reactor was restricted to 250 nucleotides.

Sometimes two possibilities for ligation reactions arise from a structure; in this case the 3'-end on which the other string is pasted is chosen at random.

A collision may result only in cleavage *or* ligation; in case both reactions are possible, again one is randomly chosen.

Figure 14 shows the frequency at which these secondary structure motifs occur in random sequences of different string length.

A typical evolution reactor experiment starts with a population of 1000 strings. The starting population consists of 100 randomly generated sequences of length n, with 10 copies each. 1 000 000 collisions are performed, the properties of the population are monitored after every 10 000 collisions. To perform a collision 2 strings are randomly chosen and folded together; the secondary structure is scanned for possible reaction sites, if none are found, the collision is elastic, otherwise cleavage or ligation occurs.

Runs were performed with different **boundary conditions:**

- conservation of mass during cleavage or ligation

  The chosen strings are taken out of the reactor, if possible, a reaction is performed, and the products are put back into the reactor; while the number of nucleotides in the reactor is constant, the number of strings is not. Cleavage results in an additional string (two cleavage products, one of the reactants stays unchanged), ligation uses up enzyme as well as substrate, so the total number of strings is reduced by one.

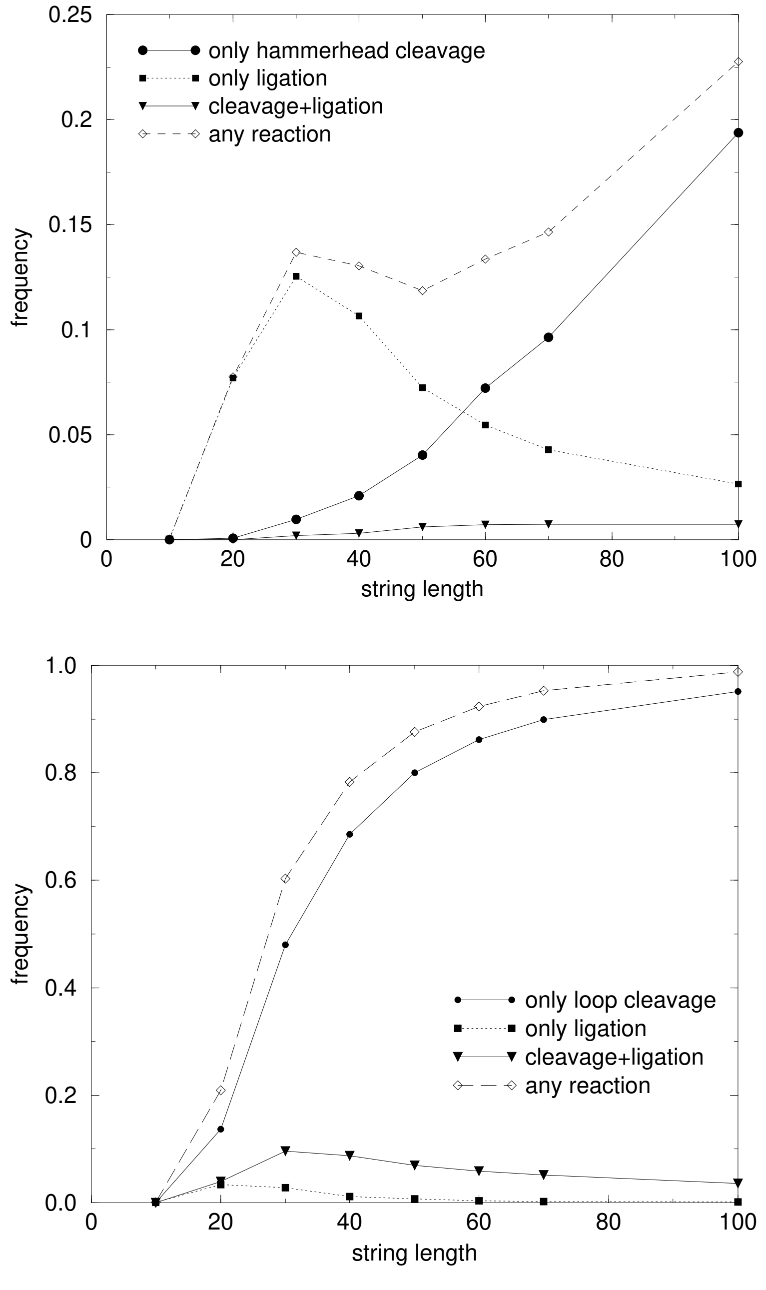


Figure 14: Frequency of structural motifs defined as reactive in the secondary structure of interacting RNA molecules. For each data point 10000 collisions between two random strings of length n (population size 10000) are performed. Reaction criteria as defined above.

- constant organization

  The number of strings in the reactor is constant. The reactants of a collision stay in the reactor together with all the products. An unspecific dilution flow washes one (in case of ligation) or two (in case of cleavage) strings out of the reactor; these strings are randomly chosen from the current population. This means that every string in the reactor has a finite lifetime, although it is not used up during reaction. Since the chance of a sequence to be chosen as a reactant is proportional to its concentration, this reaction scheme favours convergence of the population towards a set of sequences that produce one another.

- autocatalytic replication

  whenever two identical strings collide, an additional copy of this sequence is added; in case of constant organization a randomly chosen string is washed out of the reactor. This replication is independent of an eventual ligation or cleavage reaction.

- favourable replication of enzymatically active strings:

  after every collision a string is chosen and replicated; the more ligations (or cuts) a sequence has previously performed, the more likely it is chosen; a randomly chosen string is washed out of the reactor.

- the partners of unreactive collisions are taken out of the reactor.

- mutation during replication whenever a string is replicated, every nucleotide is copied with a certain replication accuracy p. In case of constant organization both ligation partners (or the cleaved strand) undergo mutation before being put back into the reactor.

A population is described by the current number of copies of every sequence; to monitor individual reactivities the number of ligation and cleavage reactions in which a sequence has taken part in the course of the simulation are tracked as well. In order to be able to identify evolving self-sustaining systems in the population all "parents" of a sequence, i.e. the partners in reactive collisions that produced this sequence, are stored. At every population dump all sequences for which at least one pair of parents still exists in the population are selected; this process is iteratively repeated until a self-sustaining population of sequences is isolated.

5 Numerical Results

5.1 Conservation of Mass During Reaction

The condition that the number of nucleotides has to stay constant during reactions is plausible and can model realistic reaction kinetics. For every collision two randomly chosen reactants are taken out of the reactor, after evaluation of the resulting secondary structure the ligated string or the products of a cleavage reaction together with the unaltered collision partner are added to the reactor. If no reactive structure motif is found, the collision is elastic and the reactants are put back into the reactor.

$$I_i + I_j \longrightarrow \mathcal{S}_{i+j} \longrightarrow \begin{cases} I_m & \text{ligation} \\ I_i + I_j & \text{elastic collision} \\ I_i + I_k + I_l & \text{cleavage} \end{cases}$$

5.1.1 No Replication

Figure 15 shows a typical population development in a reactor filled with 100 sequences of length 30 (10 copies each). The cleavage criterion is the hammerhead structure motif described in section 4.5. For 1 000 000 collisions the average stringlength, reactivity and population diversity are monitored.

Since only cleavage and ligation reactions occur and there is no replication, the number of nucleotides in the reactor is constant. In the first 20 000 collisions ligations are significantly more frequent than cleavage reactions; the number of strings decreases, while average string length increases. Reactivity decreases steadily with increasing number of collisions, reaction rates decrease faster for ligation than for cleavage. After an initial short rise the average length of strings decreases to 21 nucleotides, and most sequences exist only as single copies. No self-sustaining system could be identified.

Figure 16 shows the distribution of string lengths after 100 000, 250 000 and 500 000 collisions.

Figure 17 shows the population development in a reactor with the same setup, but with loop cleavage criterion: hairpin loops with at least 5 unpaired nucleotides are cut in the middle.

As with hammerhead cleavage, reactivity decreases steadily with increasing number of collisions. Due to the different secondary structure motif required for cleavage the initial cleavage rate is much higher than the ligation rate, the average string length quickly decreases to 16.8 and almost all collisions become unreactive. For this reason the simulation was terminated after 200 000 collisions.

The distribution of string lengths in the population after 10 000, 20 000 and 100 000 collisions for this simulation can bee seen in figure 18.

5.1.2 Autocatalytic Replication

In addition to the setup of conservation of mass during cleavage and ligation reactions autocatalytic replication was introduced: whenever 2 identical strings collide another copy of this sequence is added to the reactor, and a randomly chosen string is taken out. This replication has no influence on a possible future reaction between the two identical strings.

$$\begin{aligned} 2I_i &\longrightarrow 3I_i \\ I_j &\longrightarrow 0 \end{aligned}$$

The reactor is filled with 100 random sequences of length 30, 10 copies each. The cleavage rate stays practically constant during the first 50 000 collisions, the rate of ligation is initially much higher than the cleavage rate and decreases slowly. The number of different sequences quickly rises from 100 to

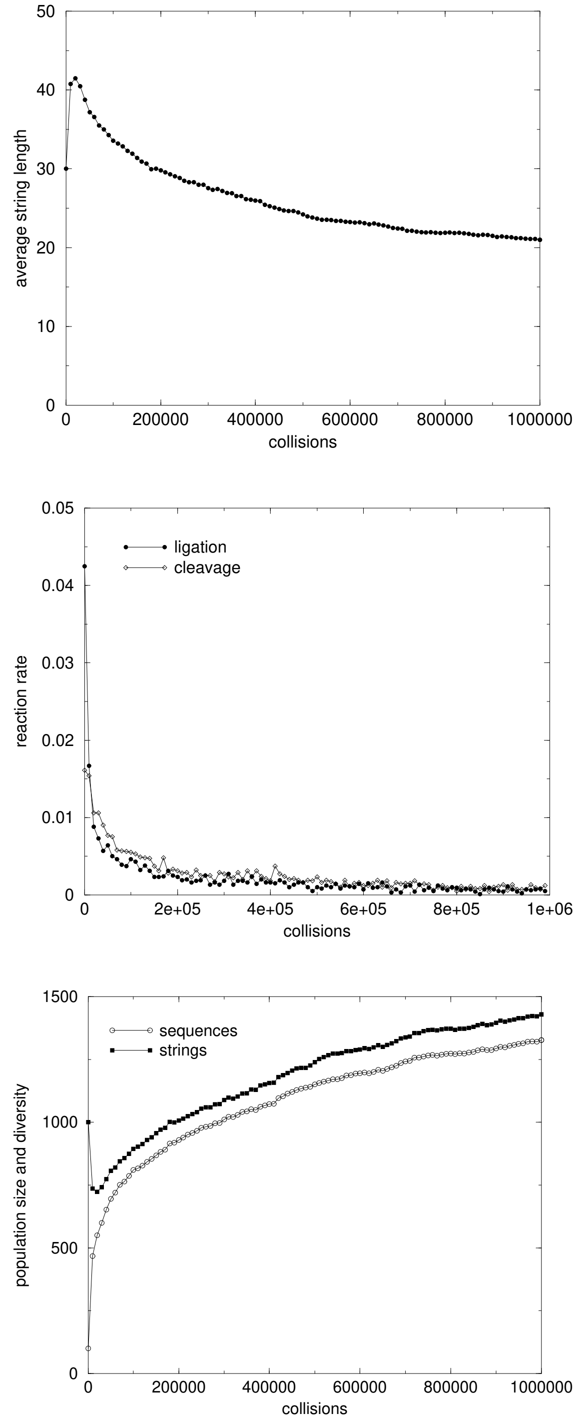


Figure 15: Population development in a reactor with conservation of mass, no replication and hammerhead cleavage. The initial population consists of 100 different random sequences of length $n = 30$, 10 copies each. 1 000 000 collisions were performed, the population was monitored after every 10 000 collisions. Cleavage and ligation rates decrease quickly, the population reaches an unreactive state of equilibrium.

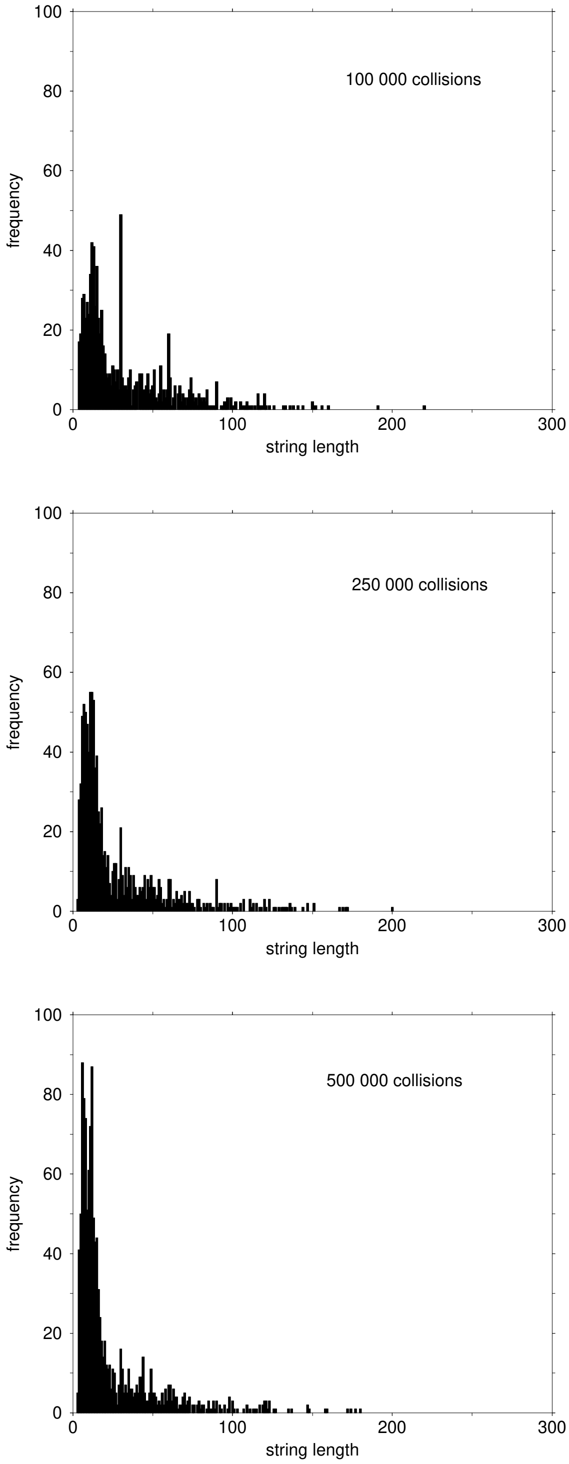


Figure 16: Distribution of string lengths in the population after 100 000, 250 000 and 500 000 collisions in the simulation described in figure 15.

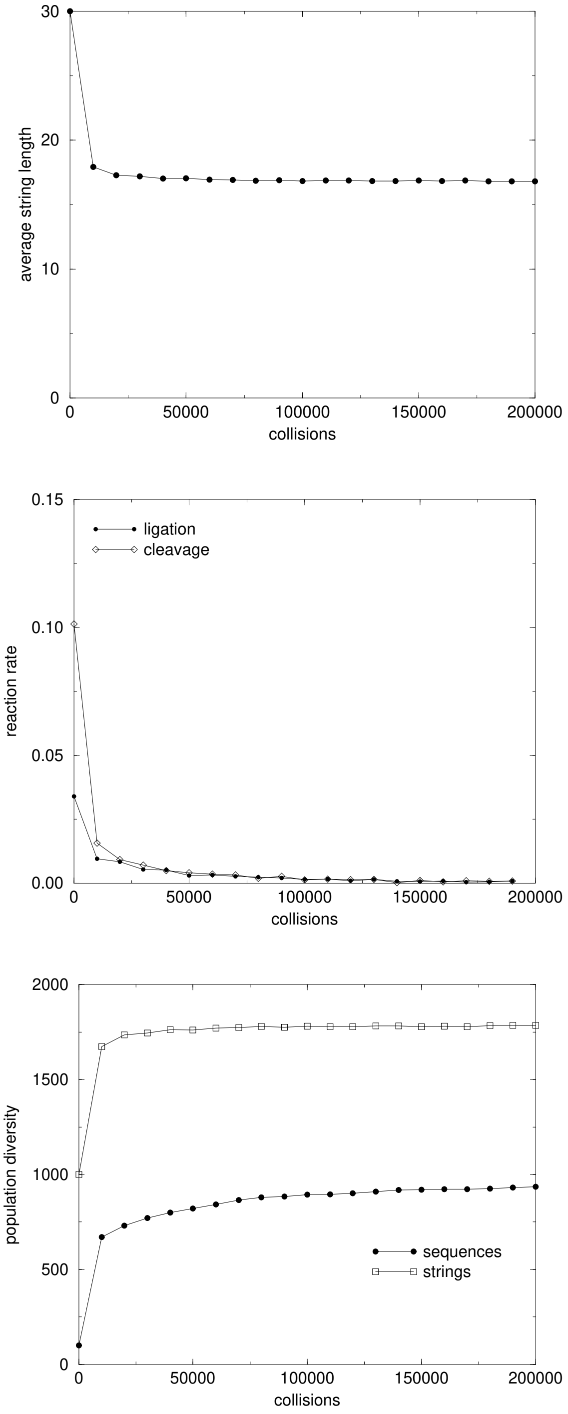


Figure 17: Population development in a reactor with conservation of mass, no replication and loop cleavage. The initial population consists of 100 different random sequences of length $n = 30$, 10 copies each. 200 000 collisions were performed, the population was monitored after every 10 000 collisions. The population quickly becomes unreactive.

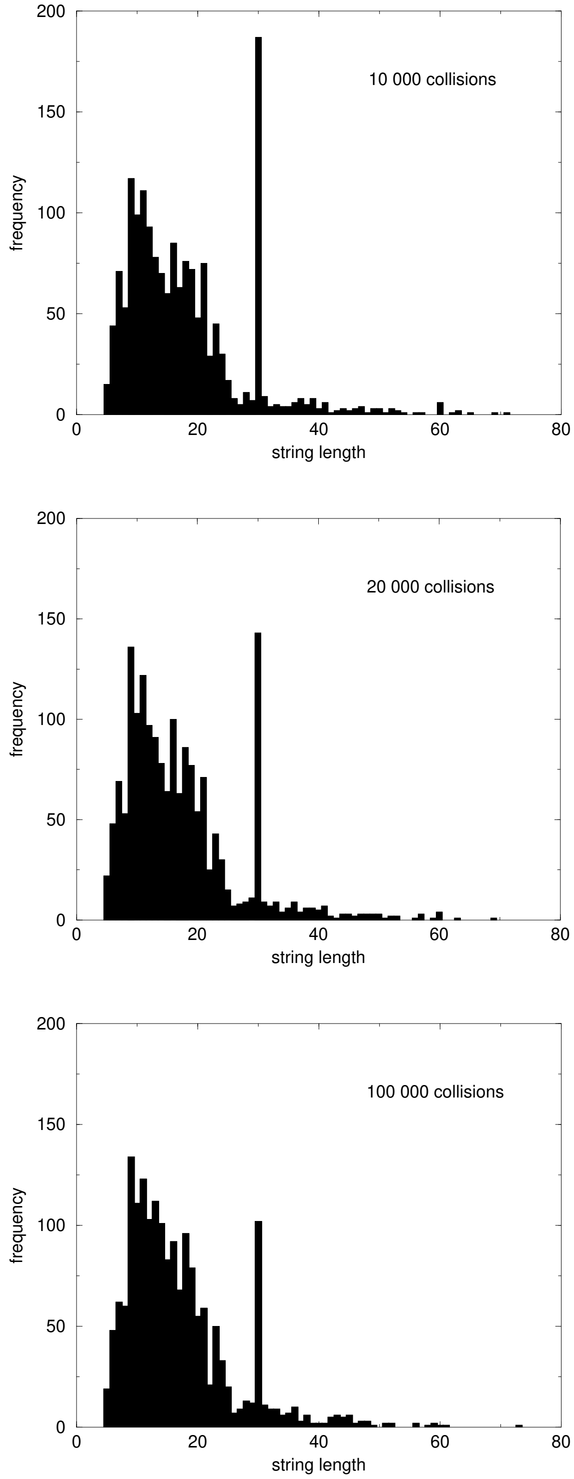


Figure 18: Distribution of string lengths in the population after 10 000, 20 000 and 100 000 collisions in the simulation described in figure 17.

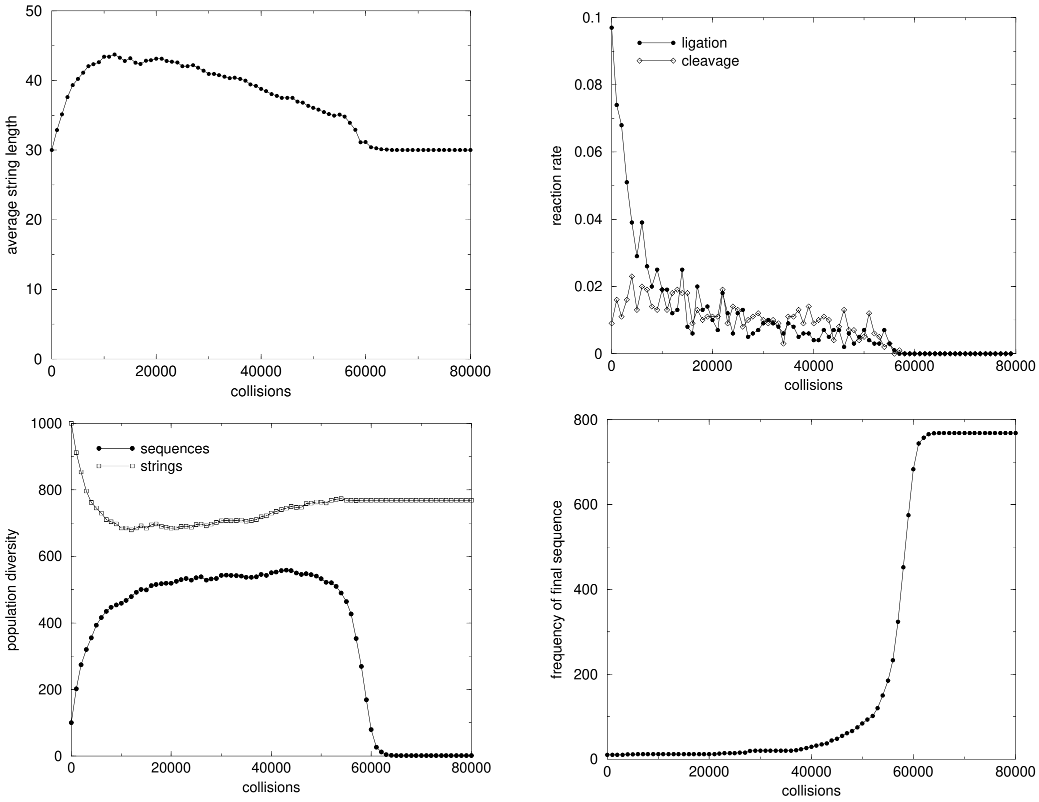


Figure 19: Population development in a reactor with conservation of mass during reactions, autocatalytic replication and hammerhead cleavage. The initial population consists of 100 different random sequences of length $n = 30$, 10 copies each. After 65 000 collisions population diversity collapses to only one sequence, the number of copies of this sequence is plotted against the number of collisions in the lower right box.

500, then stays relatively constant, until it drops sharply after 50 000 collisions. After 65 000 collisions only one sequence remains; this sequence has been part of the initial population and is unreactive. Since it did not take part in any reactive collision, it has not been cleaved or ligated. It exists in 769 copies and cannot perform reactions with itself (figure 19).

An attempt to eliminate unreactive sequences from the reactor was made by taking the partners of elastic collisions out of the system; after every elastic collision two randomly chosen strings are replicated. Autocatalytic replication and conservation of mass during reactions are left unchanged.

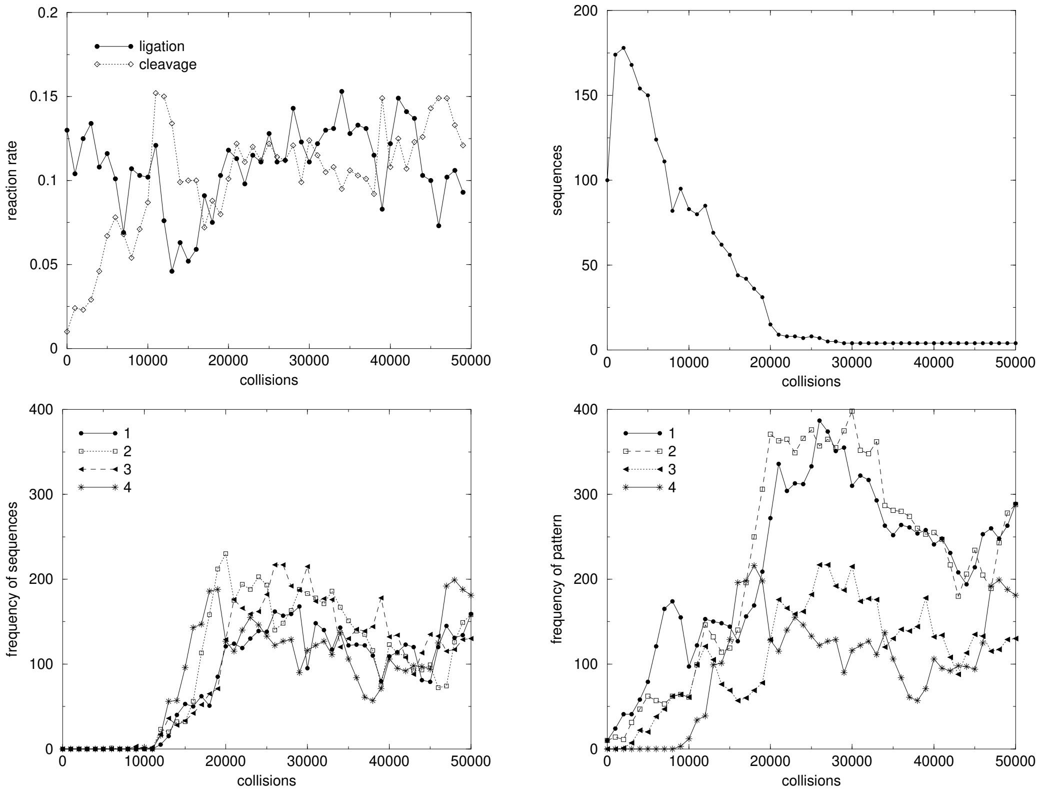


Figure 20: Population development in a reactor with conservation of mass during reactions, autocatalytic replication, hammerhead cleavage and removal of partners of inreactive collisions. The initial population consists of 100 different random sequences of length $n = 30$, 10 copies each. After 29 000 collisions the population diversity is reduced to 4 sequences. The frequency with which these 4 sequences and their pattern occur in the population can be seen in the lower plots.

The population development for this setup can be seen in figure 20.

After 29 000 collisions the population consists of 4 sequences, the matrix of their interaction is shown in Table 1.

Every sequence can perform exactly one reactive collision. Sequence 4 is not produced by any cleavage or ligation reaction, but is not used up in the cleavage reaction with 3, and can be produced by autocatalytic replication events. Since 8 out of 10 possible collisions are elastic and thus lead to random replication of 2 sequences besides removal of the colliding strings, the population

	1	2	3	4
1	-	3	-	-
2		-	-	-
3			-	2,1
4				-

Table 1: Matrix of interaction between the four sequences of the final population of a simulation with autocatalytic replication and removal of unreactive collision partners as described in figure 20.

is self-sustaining under these boundary conditions. Figure 20 monitors the frequencies of these four sequences and their pattern in the population.

5.1.3 Removal of Unreactive Strings

In this setup partners in inreactive collisions are taken out of the reactor, and two randomly chosen strings are replicated. The cleavage criterion again is defined by the hammerhead motif, but there is no autocatalytic replication.

The population quickly reduces to 2 sequences that exist in multiple copies, but do not interact. They are composed of only one sequence pattern occurring 4 respectively 6 times; this pattern is first found after 17 000 collisions and makes up for the whole population after 20 000 collisions. Figure 21 shows the development of population reactivity and diversity and the upcoming of the final pattern.

5.1.4 Preferred Replication of Reactive Strings

A different attempt to favour catalytically active strings was made by introducing replication/dilution events that do not depend on the outcome of a single collision such as the removal of partners in unreactive collisions did.

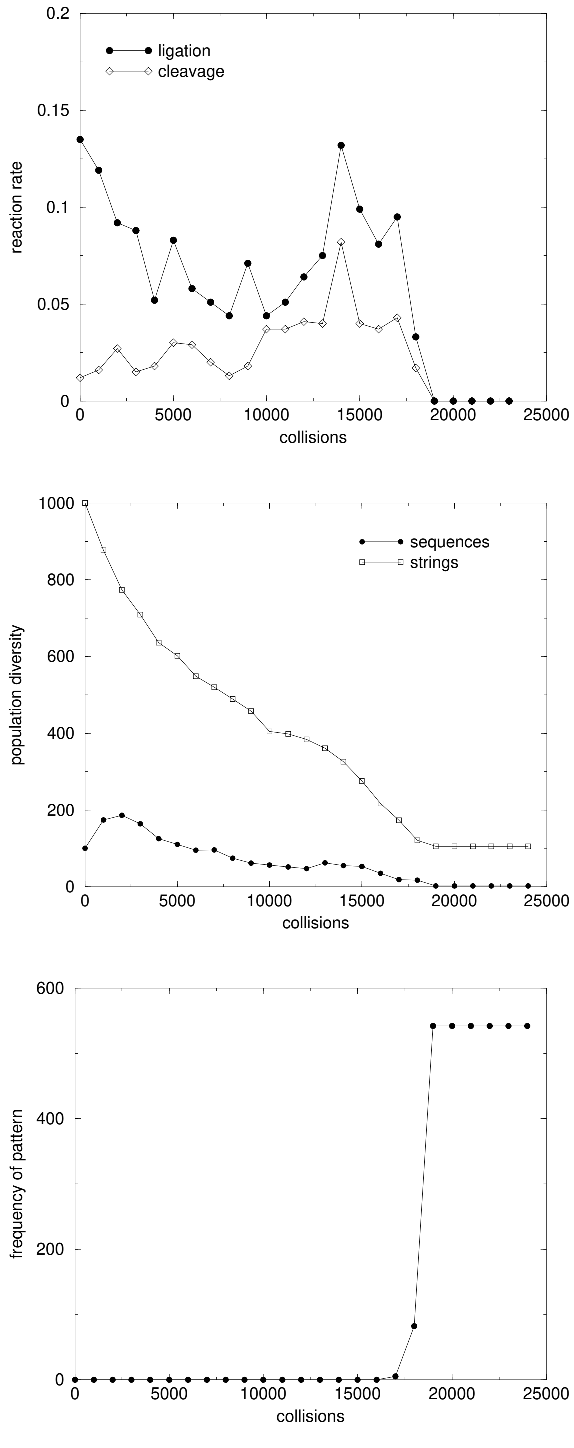


Figure 21: Population development in a reactor with conservation of mass during reactions, hammerhead cleavage and removal of partners of inreactive collisions combined with replication of 2 randomly chosen strings. The initial population consists of 100 different random sequences of length $n = 30$, 10 copies each. After 19 000 collisions the population diversity is reduced to 2 sequences which are composed of only one sequence pattern and do not interact. The frequency of this pattern in the population is monitored in the lower plot.

Instead, the overall reactivity of a sequence determines its chances to be chosen for replication, while every string has an equal chance to be diluted.

After every collision a string is randomly chosen and replicated, and a randomly chosen string is washed out. The probability that a sequence is chosen for replication is higher for catalytically active strings: it is proportional to the number of copies plus the number of reactions (ligations, cleavages or both) this sequence has been involved in so far.

$$\begin{aligned} I_i &\longrightarrow 2I_i \\ I_d &\longrightarrow 0 \end{aligned}$$

(i) Preferred replication of ligating strings

As one can see in figure 22 (top), the population diversity is quickly reduced to only one sequence, which does not react with itself and has been present in the starting population. Since ligations are much more frequent than cleavage reactions, the total population size also shows a fast decrease.

Whenever two strings ligate, they are taken out of the reactor due to the condition of conservation of mass during reactions; however, remaining strings (copies) of the 2 species have a higher probability of being replicated. Newly added sequences that are the product of a reactive collision usually exist in only one copy and may be swept out after a replication event before they can perform a reactive collision. As long as a sequence has not taken part in a ligation reaction, its chances to be replicated are slim; if only one copy exists and this string takes part in a ligation, the species is deleted from the reactor.

Therefore the ligating strings of the primary population, which start with 10 copies, have the highest chance of being replicated, and the population consists almost only of these sequences and their ligation products. Figure 22 shows the distribution of string lengths after 1000, 5000 and 7000 collisions (middle, bottom) and the decrease in diversity of sequences of string length $n = 30$ (bottom).

(ii) Preferred replication of strings that have been involved in cleavage reactions

Initially the ligation rate is higher than the rate for cleavage; the former decreases quickly to a constant value, the latter increases slowly; after 6 000 collisions the cleavage rate is higher, after 12 000 collisions cleavage reactions have outnumbered ligations (figure 23).

The population size grows steadily, since only one reactant is used up during cleavage, but two new strings are added to the reactor. Both cleavage reaction partners get a replication bonus. The number of different sequences stays in a range between 100 and 200.

The 9 most frequently occurring sequences include 2 enzymatically active strings, their substrates and the resulting cleavage products and make up for 64 percent of the total population after 50 000 collisions.

(iii) Preferred replication of all reactive strings

Since the initial ligation rate is significantly higher than the cleavage rate, if all reactive sequences are equally preferred when choosing a string for replication the overall picture is the same as in the case of ligation biased replication. Another simulation run starting with a population of the same size, but of higher diversity (1000 random sequences of length $n = 30$, one copy each) showed basically the same result: after 10 000 collisions only one sequence that cannot react with itself is left.

Simulations with the same boundary conditions, but starting with a population of random strings of length $n = 70$ are shown in figure 24. If ligating sequences are preferred for replication (top), the result is the same as for string length $n = 30$: only one sequence survives that was already present in

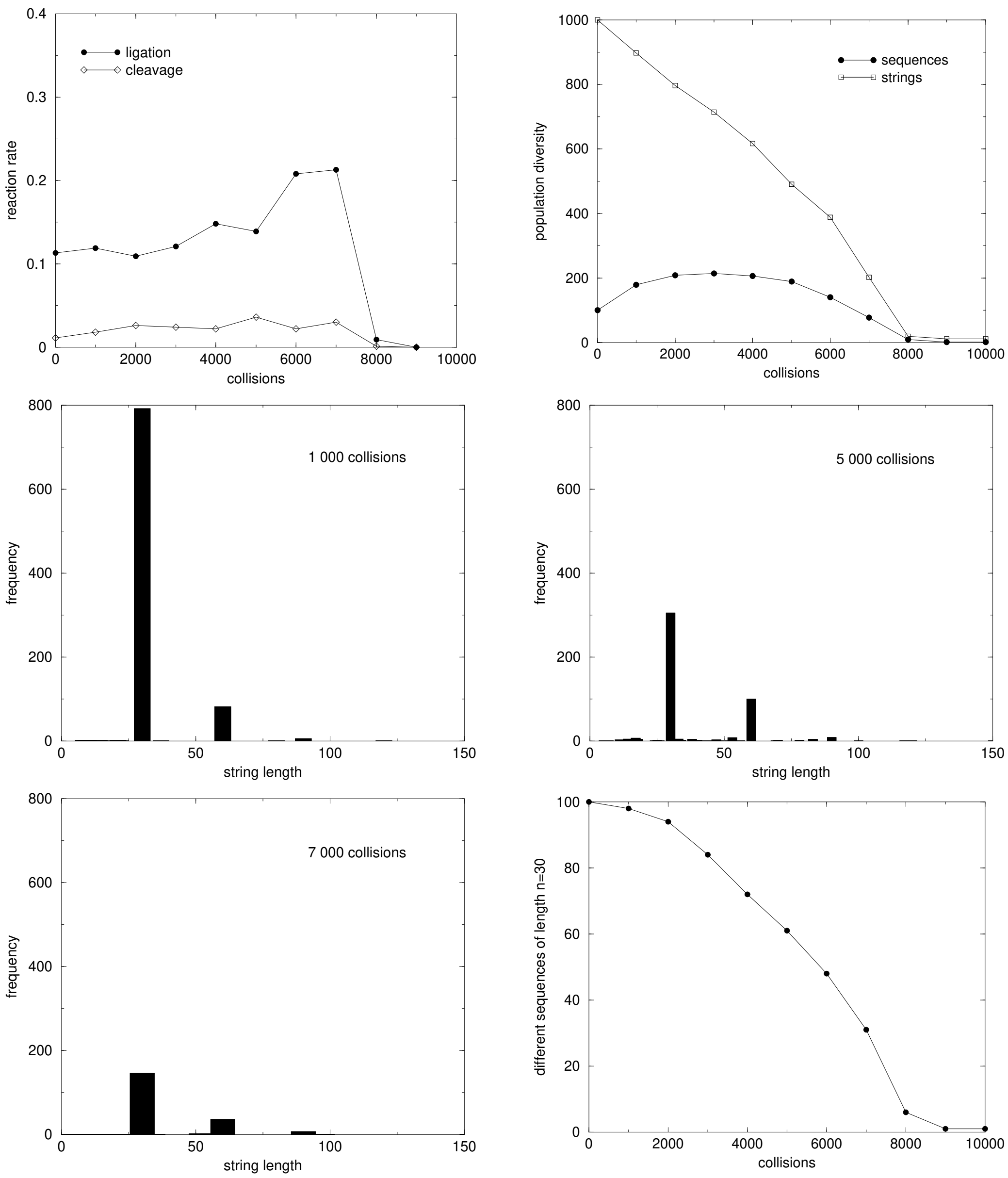


Figure 22: Reactivity, population diversity and distribution of string lengths are monitored for a run starting with 100 sequences of length n = 30, 10 copies each, conservation of mass during reactions and hammerhead cleavage, and a replication/dilution event after every collision. Sequences that have been involved in ligations have a higher chance to be chosen for replication. The population is quickly reduced to one unreactive sequence.

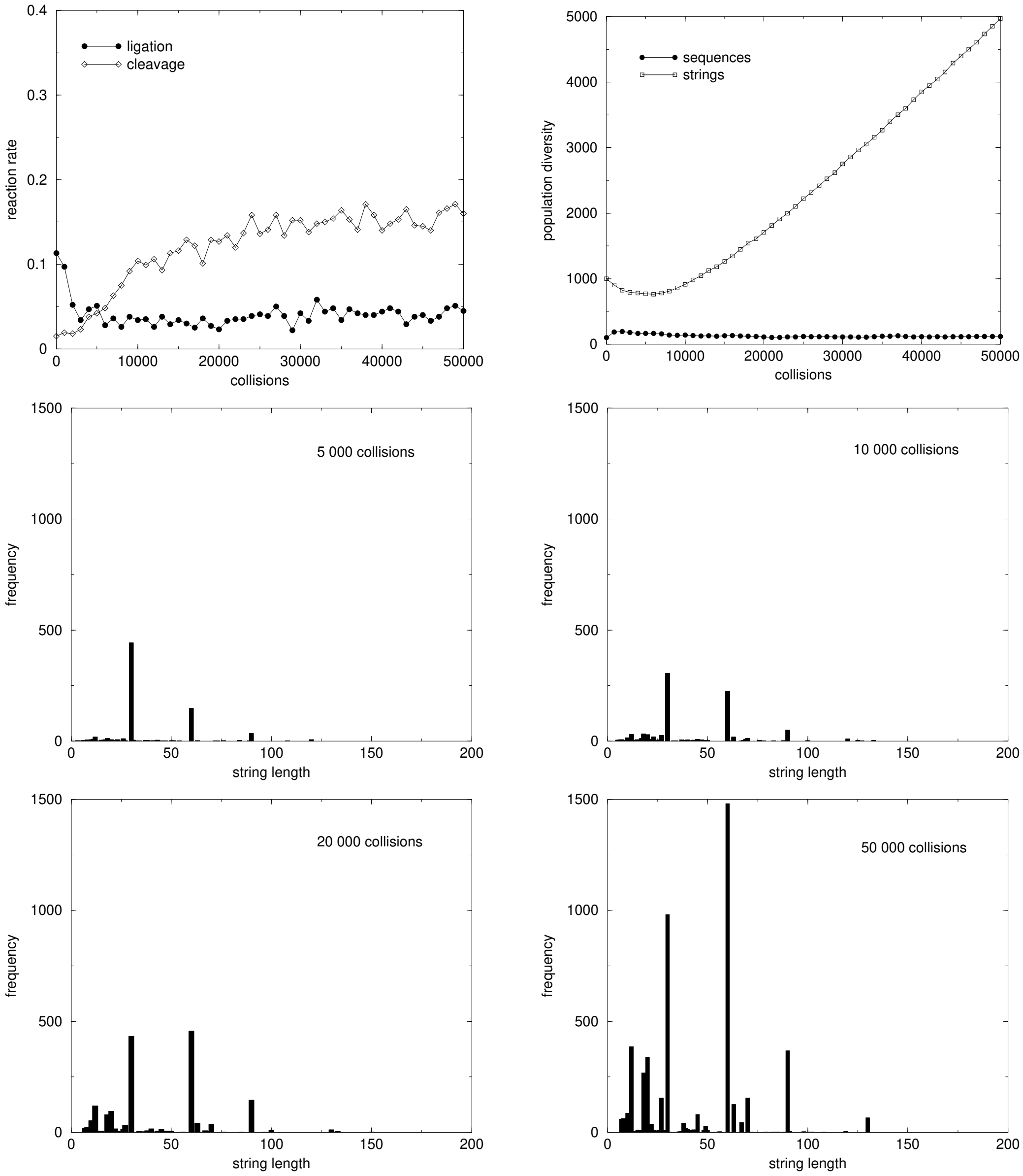


Figure 23: Reactivity and population diversity are monitored for a run starting with 100 sequences of length n = 30, 10 copies each, hammerhead cleavage, with preferred replication of sequences that have been involved in cleavage reactions. The number of different sequences stays relatively constant, the total number of strings increases. After 50 000 collisions 9 reactive strings make up for over 60 percent of the total population size.

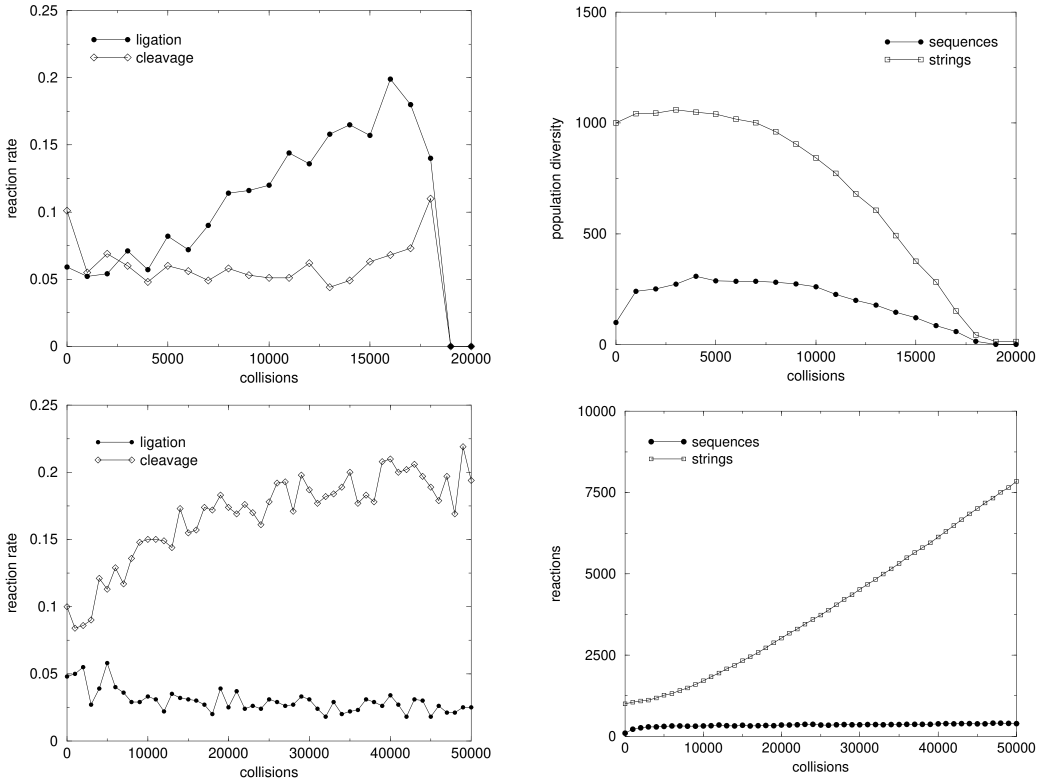


Figure 24: Reactivity and population diversity are monitored for a run starting with 100 sequences of length n = 70, 10 copies each, conservation of mass during reactions, hammerhead cleavage and a replication/dilution event after every collision. Top: preferred replication of ligating sequences; only one sequence remains. Bottom: preferred replication of sequences that have been involved in cleavage reactions. The 18 most frequent sequences compose 65 percent of the total population.

the starting population.

Since a collision between two random strings of length $n = 70$ is more likely to result in a cleavage reaction than in ligation, the picture for preferred replication of reactive sequences is similar to the simulation with a bias towards sequences performing cleavage reactions. The latter can be seen in figure 24 (bottom); after 50 000 collisions the 18 most frequent sequences compose 65 percent of the total population. 6 of them are cleavage substrates and are not produced by any interaction between sequences of this subpopulation.

5.2 Constant Organization

When cleavage or ligation takes place during a collision, the parent strings stay in the reactor together with the reaction products. An unspecific dilution flow keeps the size of the population constant.

This condition is equivalent to specific replication of the cleavage substrate or of both of the ligated strings, respectively, and faciliates emerging of self-sustaining systems.

$$
\begin{array}{rrcl}
\text{cleavage:} & I_i + I_j & \longrightarrow & I_i + I_j + I_k + I_l \\
 & I_d & \longrightarrow & 0 \\
 & I_e & \longrightarrow & 0 \\
\text{ligation:} & I_i + I_j & \longrightarrow & I_i + I_j + I_m \\
 & I_d & \longrightarrow & 0 \\
\text{elastic:} & I_i + I_j & \longrightarrow & I_i + I_j
\end{array}
$$

5.2.1 Runs Without Additional Replication Events

Under these boundary conditions populations reach stable self-sustaining sequence distributions, some of which are described and analyzed in section 5.3. The total population size of 1000 strings does not change during simulations.

Figures 25 and 27 show simulation runs starting with populations of 100 random strings of length $n = 30$, each of them existing in 10 copies, and hammerhead cleavage criterion. Figure 26 shows a run starting with a higher sequence diversity in the starting population: 200 different random sequences exist in 5 copies.

Figure 28 monitors a run starting with the first setup (100 sequences, 10 copies), but loops with at least 5 unpaired bases are cleaved.

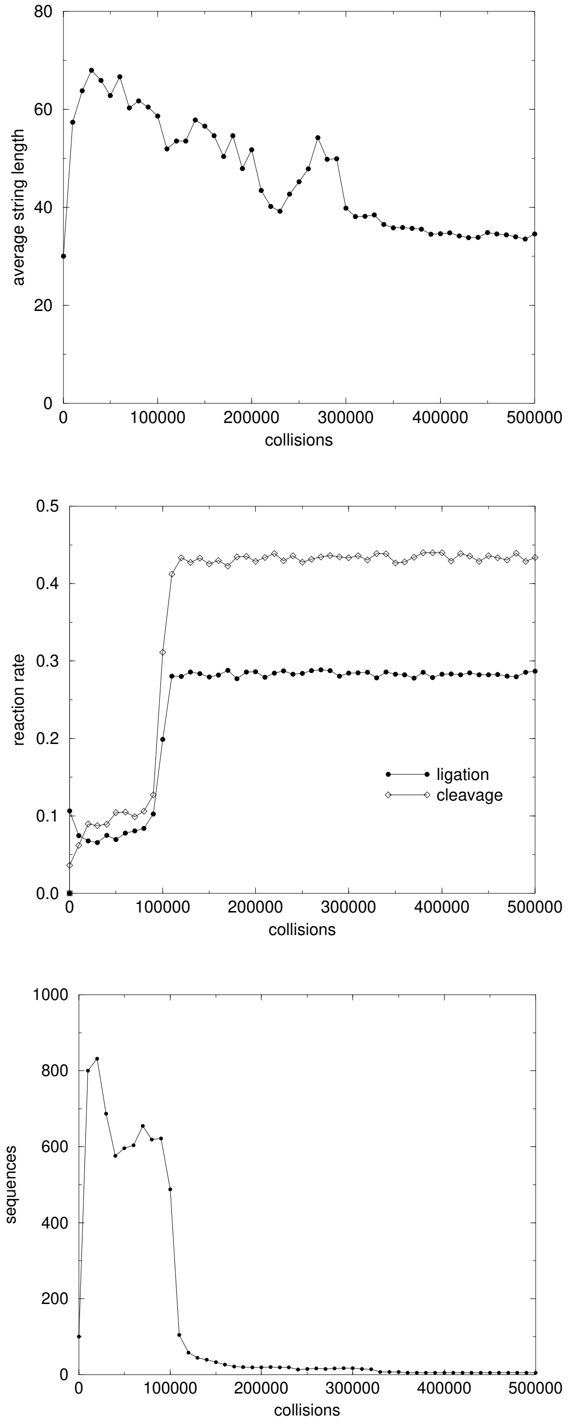


Figure 25: Average string length, reactivity and population diversity are monitored for a run starting with 100 sequences of length n = 30, 10 copies each, constant organization and hammerhead cleavage. After 360 000 collisions the population is reduced to a stable self-sustaining system consisting of 5 sequences. The properties of this system are discussed in section 5.3.1 (i) (System 1).

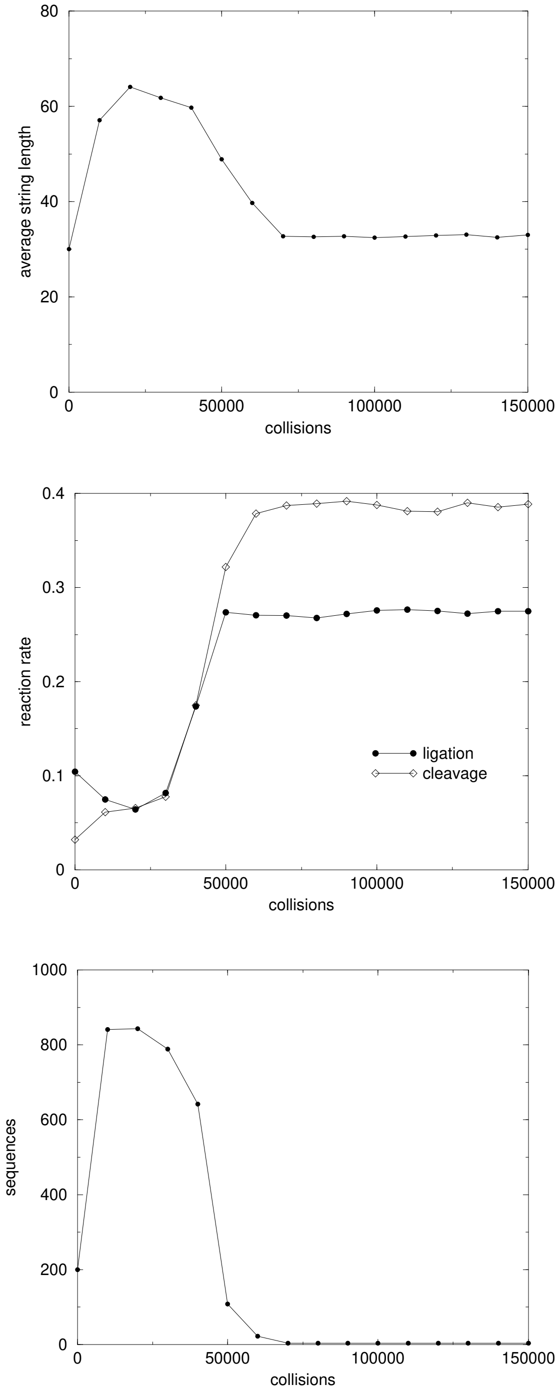


Figure 26: Average string length, reactivity and population diversity are monitored for a run starting with 200 sequences of length n = 30, 5 copies each, constant organization and hammerhead cleavage. After 70 000 collisions the population is reduced to a stable self-sustaining system consisting of 3 sequences. The properties of this system are discussed in section 5.3.1 (ii) (System 2)

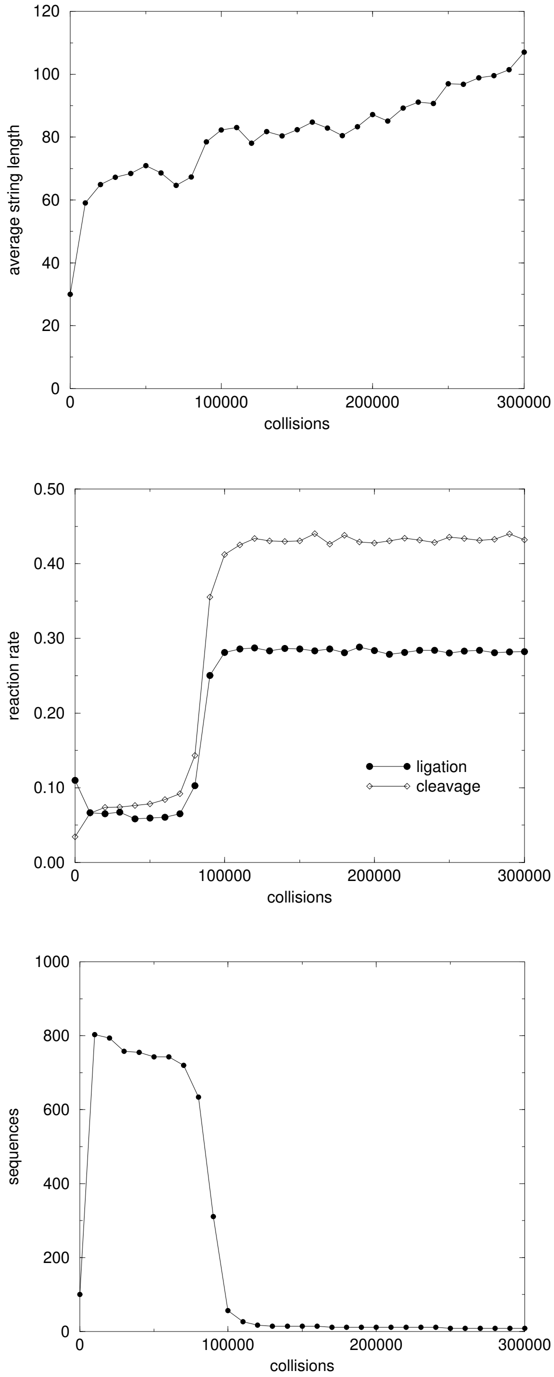


Figure 27: Average string length, reactivity and population diversity are monitored for a run starting with 100 sequences of length n = 30, 10 copies each, constant organization and hammerhead cleavage. After 340 000 collisions the population is reduced to a stable self-sustaining system consisting of 5 sequences. The properties of this system are discussed in section 5.3.1 (iii) (System 3)

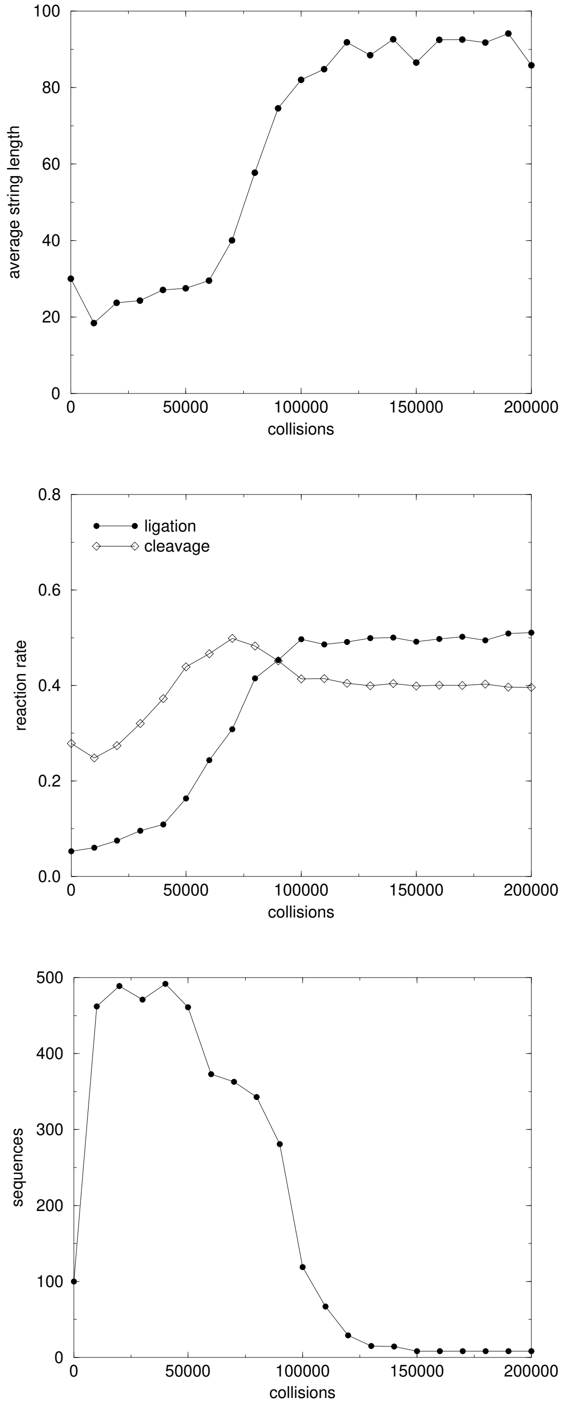


Figure 28: Average string length, reactivity and population diversity are monitored for a run starting with 100 sequences of length n = 30, 10 copies each, constant organization and loop cleavage. After 150 000 collisions the population is reduced to a stable self-sustaining system consisting of 8 sequences, which are composed of only one sequence pattern. The properties of this system are discussed in section 5.3.2. (i) (Loopsystem 1).

5.2.2 Autocatalytic Replication

In addition to the setup described above autocatalytic replication is introduced: whenever 2 identical strings collide another copy of this sequence is added to the reactor, and a randomly chosen string is deleted.

$$\begin{aligned} 2I_i &\longrightarrow 3I_i \\ I_d &\longrightarrow 0 \end{aligned}$$

Figure 29 shows a simulation starting with 100 different random sequences of length $n = 30$, 10 copies each. After 110 000 collisions the population consists of 7 different sequences, which are all built by concatenation of one sequence pattern; this population is self-sustaining and is discussed in section 5.3.1.(iiii) (System 4).

5.2.3 Preferred Replication of Reactive Strings

In addition to the condition that reactants are not used up in cleavage or ligation reactions an attempt to favour catalytically active sequences even more was made: after every collision a replication step as described in section 5.1.4. that favourably increases the number of copies of reactive sequences takes place. Does this boundary condition accelerate the formation of a stable sequence distribution?

$$\begin{aligned} I_i &\longrightarrow 2I_i \\ I_d &\longrightarrow 0 \end{aligned}$$

As we can see in figure 30, which shows the case of preferred replication of sequences that have taken part in ligations, the population diversity stays high and is relatively constant. In the final population after 1 000 000 collisions only 23 sequences exist in more than 10 copies, 12 of which have already been

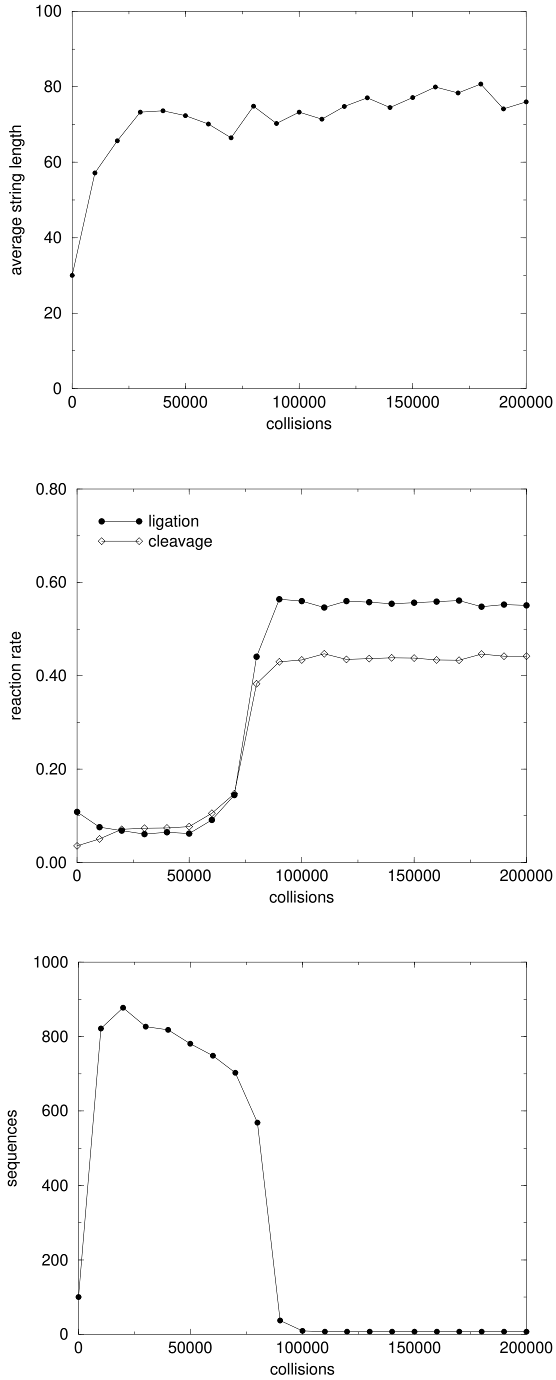


Figure 29: Average string length, reactivity and population diversity are monitored for a run starting with 100 sequences of length n = 30, 10 copies each, constant organization, hammerhead cleavage and autocatalytic replication. After 110 000 collisions the population is reduced to a stable self-sustaining system consisting of 7 sequences, which are composed of only one sequence pattern. The properties of this system are discussed in section 5.3.1. (iiii) (System 4).

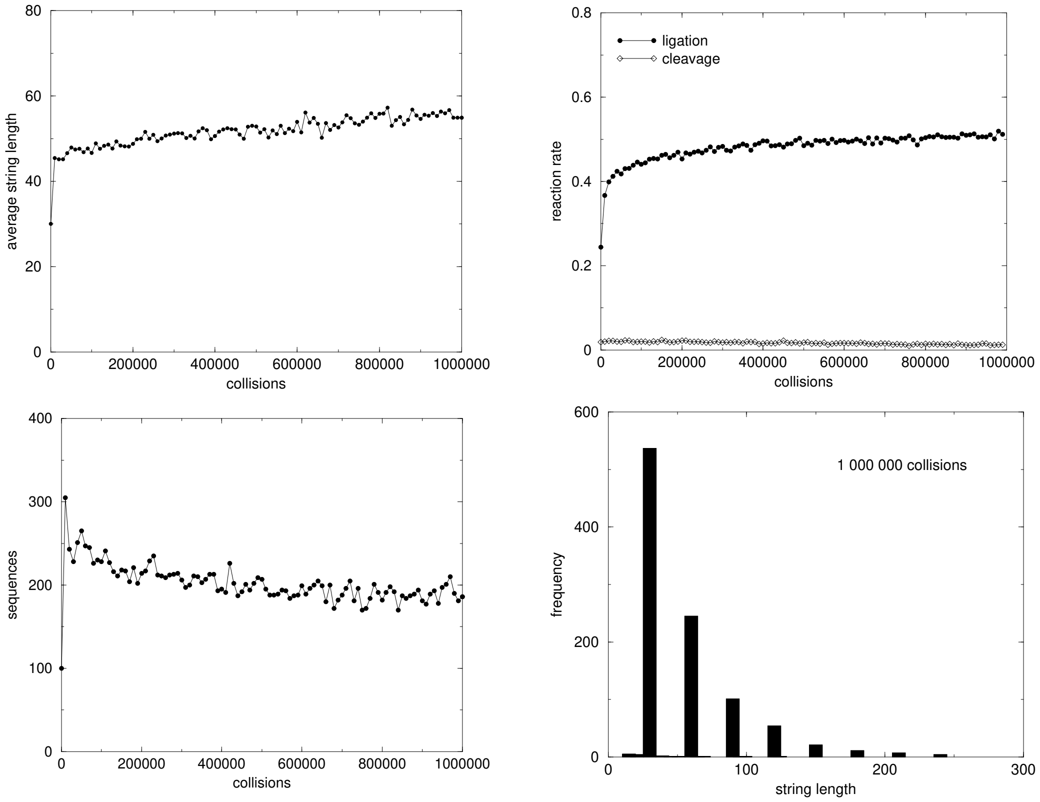


Figure 30: Average string length, reactivity, population diversity and string length distribution after 1 000 000 collisions are displayed for a run starting with 100 sequences of length n = 30, 10 copies each, constant organization, hammerhead cleavage, and preferred replication of ligating sequences. All sequences of length 30 in the final population have been present at the beginning, no self-sustaining subpopulation is found.

part of the starting population; the distribution of string lengths in the population changes only slightly during the simulation and looks practically the same for the population after 10 000 and 1 000 000 collisions. All sequences of length $n = 30$ in the final population have been present at the start. No self-sustaining sequence distribution or subpopulation can be found.

In the case of preferred replication of sequences that have taken part in cleavage reactions the picture is a little different (figure 31). The population diversity is quickly reduced to 14 to 17 different sequences: of the 5 cleavage

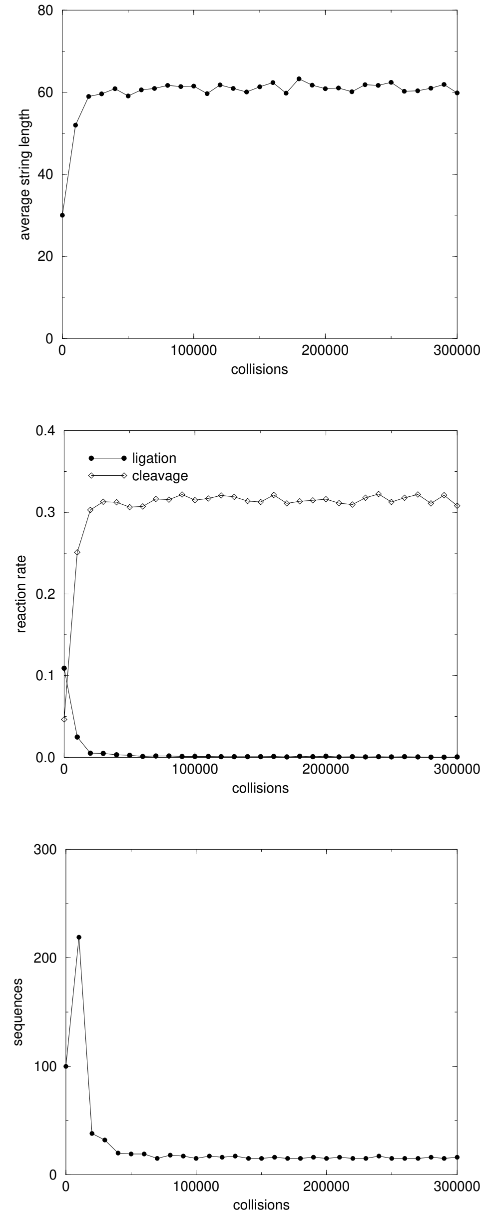


Figure 31: Average string length, reactivity and population diversity are displayed for a run starting with 100 sequences of length n = 30, 10 copies each, constant organization, hammerhead cleavage, and preferred replication of sequences that have taken part in cleavage reactions. The population soon is composed of about 15 different sequences not all of which can be produced by interactions of other members of the population.

substrates only one can be produced by another cleavage reaction, none by ligation; the set of sequences is not self-sustaining. If strings that have been involved in any reaction are preferentially replicated, the result is the same as for ligation biased replication.

5.2.4 Mutation During Reaction

In the constant organization setting described so far strings were not used up in reactive collisions; to keep population size constant, one or two randomly chosen strings were deleted after ligations or cleavage reactions, respectively.

To introduce point mutations as an additional source of variation strings I_j that are cleaved or ligated are not left unchanged in the reactor; instead, they are replaced by an erroneous copy I_j'. This copy is the result of a process that introduces point mutations at a predefined rate uniformly along the chain. For every nucleotide the probability to be faithfully copied is given by the replication accuracy p, the probability of being replaced by one of the 3 other nucleotides is $1 - p$.

$$\begin{array}{lrcl}
\text{cleavage:} & I_i + I_j & \longrightarrow & I_i + I_j' + I_k + I_l \\
 & I_d & \longrightarrow & 0 \\
 & I_e & \longrightarrow & 0 \\
\text{ligation:} & I_i + I_j & \longrightarrow & I_i' + I_j' + I_m \\
 & I_d & \longrightarrow & 0 \\
\text{replication:} & I_r & \longrightarrow & I_r + I_r' \\
 & I_d & \longrightarrow & 0
\end{array}$$

Simulations with different mutation rates have been performed. Figure 32 shows population developments for replication accuracy 0.990 and 0.999.
A comparison of the distribution of string lengths in the populations after

500 000, 750 000 and 1 000 000 collisions is plotted in figure 33. At replication accuracy 0.990 almost all sequences exist only as single copies, and no regularity or pattern was found in the population. The picture is the same if replication accuracy is lowered to 0.950.

For a simulation with replication accuracy 0.999 the picture looks different. The population does not collapse to a self-sustaining system consisting of only a few strings as in the simulations without mutation; instead, the number of different sequences in the reactor moves around 500 for a total population of 1000 strings. However, string lengths are by no means equally distributed. After 300 000 collisions a string length distribution comes up that stays stable until the end of the simulation (1 000 000 collisions). Figure 33 (right) shows string length frequencies in the population after 500 000, 750 000 and 1 000 000 collisions. The 10 most frequent string lengths make up 80 to 86 percent of the total population.

Moreover, the sequences of equal length are quite similar. Figure 34 (left) shows the average hamming distance of strings of the same length to the most frequent sequence of that string length in the population after 1 000 000 collisions.

The most frequent sequences change slightly with time. Figure 34 (right) shows the hamming distance of the most frequent sequence of the string length class 71 to the most frequent one after 300 000 collisions and to the most frequent one 100 000 collisions ago.

If we group all sequences of equal length together and study the interactions between different string length classes, we observe a self-sustaining reaction network due to structural neutrality of some point mutations. Table 2 shows the string lengths of the most common interaction products between string length classes at the end of the simulation. Figure 35 shows some of the most frequently occurring secondary structures resulting from collisions between members of the most densely populated string length classes.

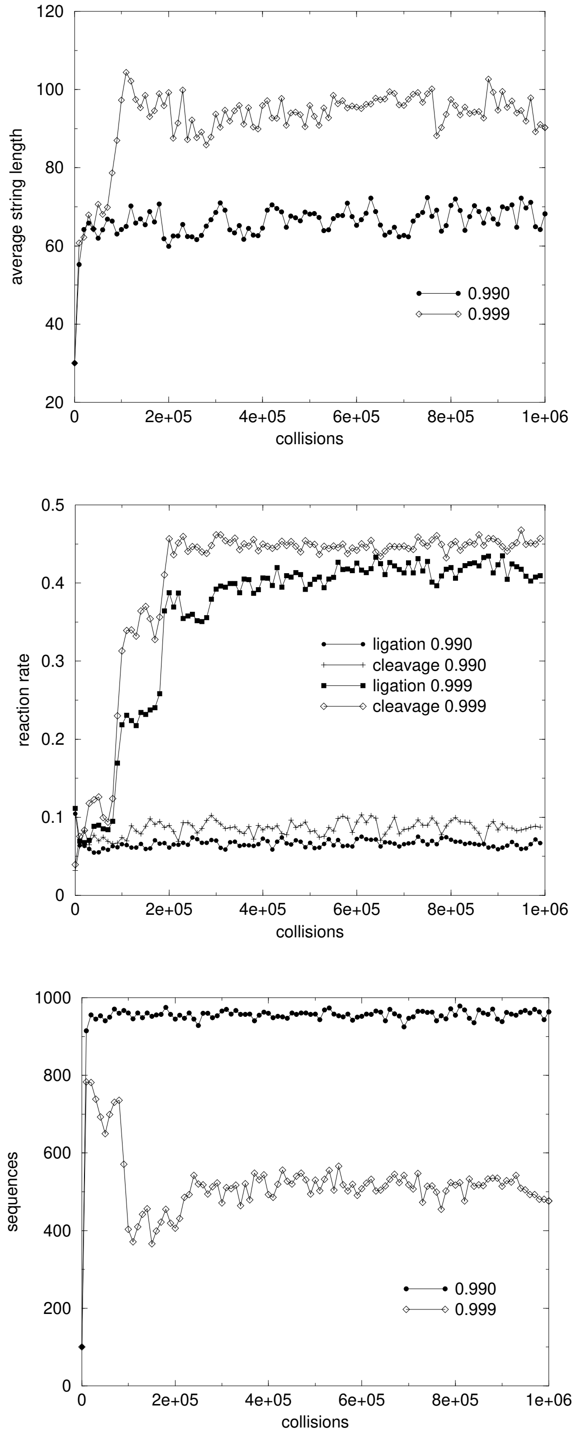


Figure 32: Average string length, reaction rates and population diversity are displayed for runs starting with 100 sequences of length n = 30, 10 copies each, constant organization, hammerhead cleavage, and mutation of cleaved or ligated strings. Values from simulations with replication accuracy 0.990 and 0.999 are compared. At the higher mutation rate almost all sequences exist as single copies. At the lower mutation rate reactivity is much higher.

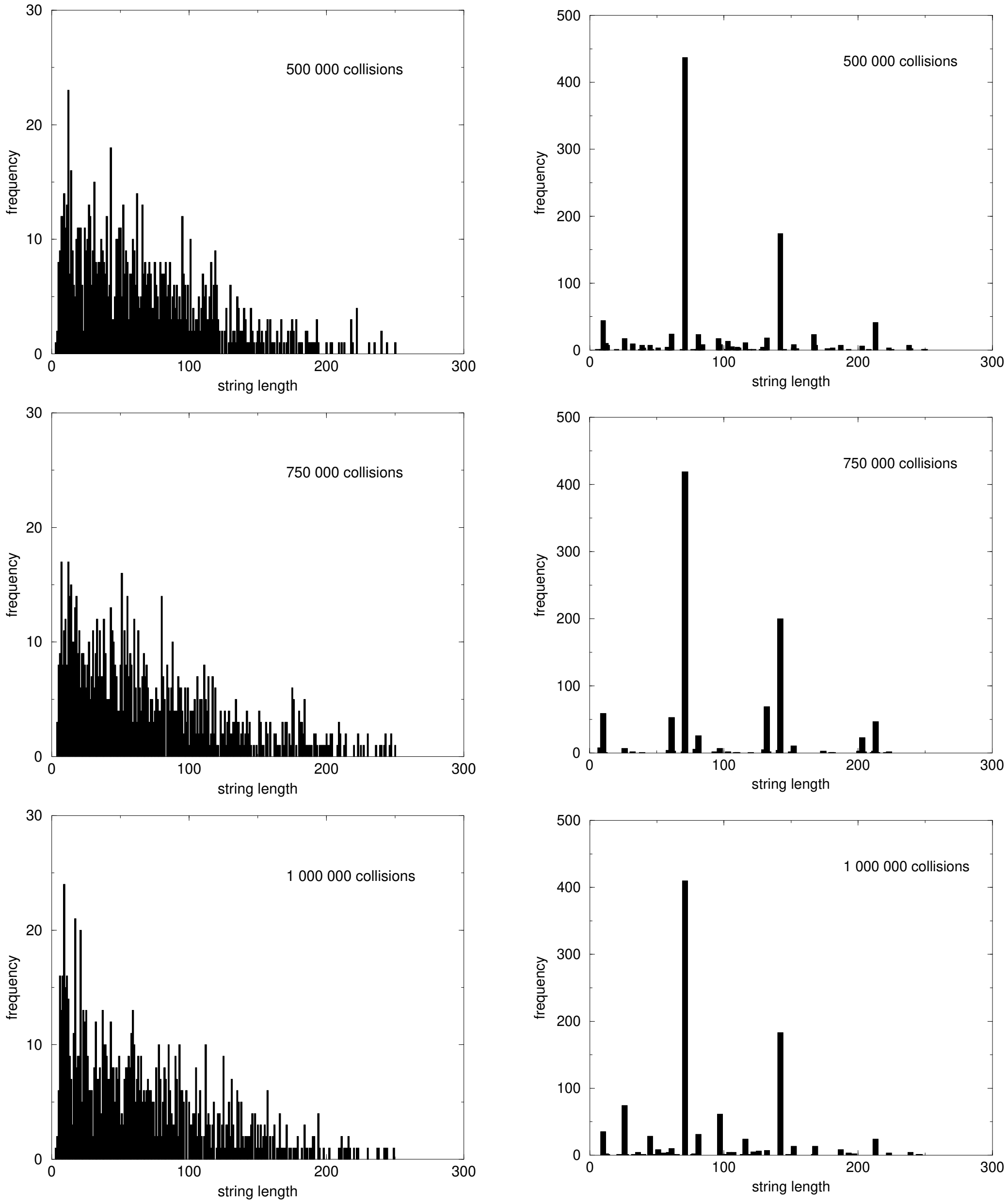


Figure 33: Distribution of string lengths in the populations from simulations described in figure 32 after 500 000, 750 000 and 1 000 000 collisions. Replication accuracy 0.990 (left): diversity of string lengths is high; no regularities or pattern were found in the population. Replication accuracy 0.999 (right): the 10 most frequent string lengths make up 80 to 86 percent of the total population.

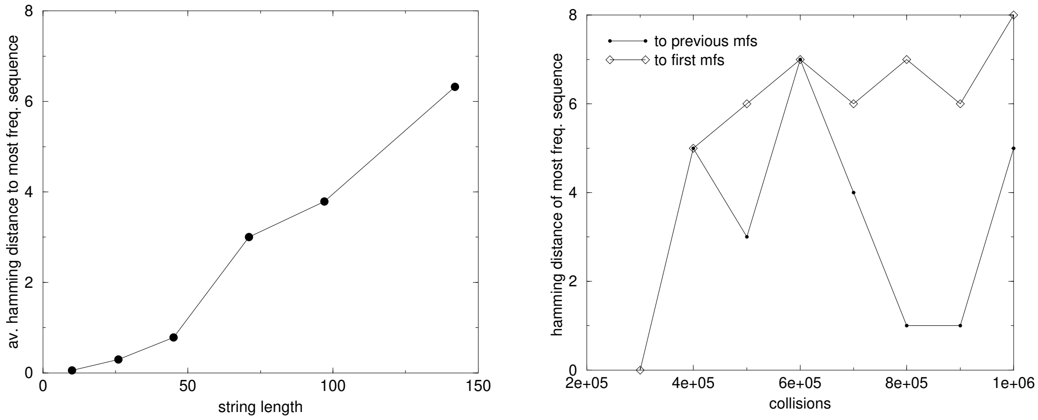


Figure 34: Hamming distances to most frequent sequences for the simulation with replication accuracy 0.999 described in figures 32 and 33. In the left figure we see the average hamming distance of all sequences of a string length to the most frequent sequence of this length for the most frequent string length classes in the population after 1 000 000 collisions. The figure on the right monitors the hamming distances between most frequent sequences of length $n = 71$; distances to the most frequent sequence after 300 000 collisions and to that of the population 100 000 collisions earlier are plotted against collision number.

	45	71	97	142
10		81	26,71; 107	71,71;152
26	71	97;26,45	26,71	71,71;168;45,97
45		116;26,45	26,71;142	71,71
71		142; 26,45	26,71	71,71; 213
97			26,71	
142				

Table 2: Interaction matrix between sequences of the most densely populated string length classes in the final population of the simulation described in figures 32, 33 and 34, replication accuracy 0.999. Collisions between all sequences of any two classes were performed; entries in the table are the string lengths of the most common interaction products.

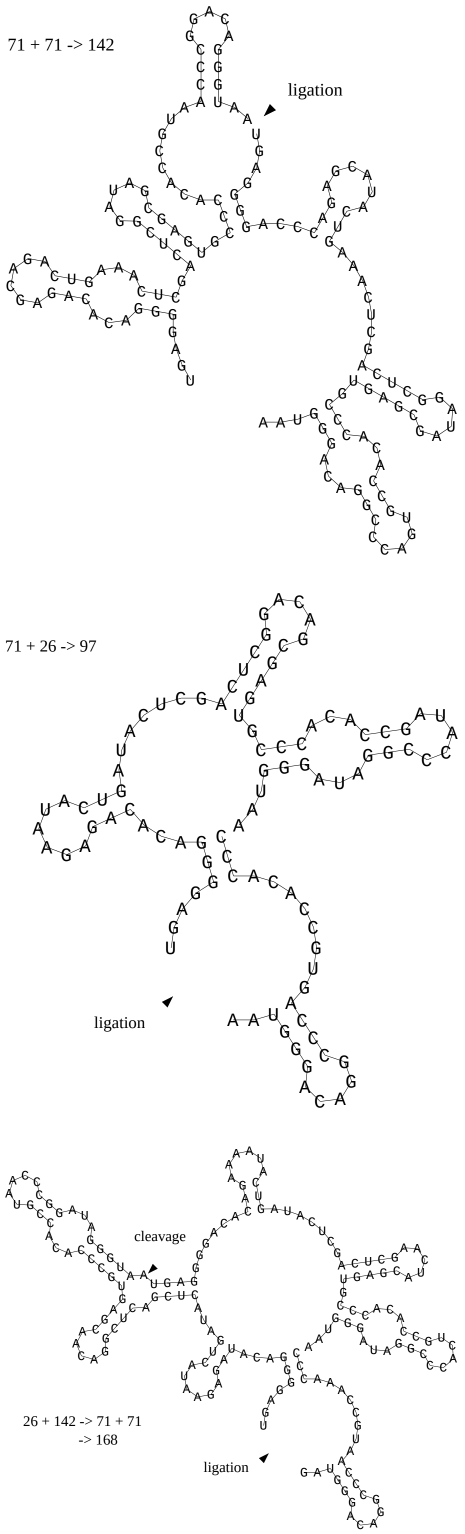


Figure 35: Most frequent secondary structures resulting from interactions between members of the most densely populated string length classes in the final population of the simulation (replic. acc. 0.999) discussed in figures 32 and 33.

5.3 Analysis of Self-Sustaining Sequence Distributions

In this section we take a closer look on the properties of sequence distributions that were reached by populations evolving under constant organization without mutation events. The sets of sequences described here are self-sustaining in the sense that all sequences of the population can be produced by cleavage or ligation reactions occurring during interactions between members of the population.

5.3.1 Hammerhead Cleavage

(i) System 1

The population of the collision reactor simulation monitored in figure 25 (section 5.2.1) evolved into a stable self-sustaining system consisting of 5 different sequences after 360 000 collisions were performed.

1 = AGGGUCAAUUGAUGAAUGUGGCCUUCGAC*AUAA***UCUCGA GCGC**
2 = *AUAA***UCUCGAGCGC**
3 = *AUAA***UCUCGAGCGC**AGGGUCAAUUGAUGAAUGUGGCCUU CGAC*AUAA***UCUCGAGCGC**
4 = **UCUCGAGCGC**
5 = **UCUCGAGCGC**AGGGUCAAUUGAUGAAUGUGGCCUUCGAC *AUAA***UCUCGAGCGC**

These 5 sequences can be decomposed into 3 patterns B, D and E.

$$\begin{aligned} \mathrm{B} &= \mathbf{UCUCGAGCGC} \\ \mathrm{D} &= \mathit{AUAA} \\ \mathrm{E} &= \mathrm{AGGGUCAAUUGAUGAAUGUGGCCUUCGAC} \end{aligned}$$

$$\begin{aligned} 1 &= \quad\; \text{E-D-B} \\ 2 &= \text{D-B} \\ 3 &= \text{D-B-E-D-B} = 2 - 1 \\ 4 &= \;\; \text{B} \\ 5 &= \;\; \text{B-E-D-B} = 4 - 1 \end{aligned}$$

The matrix of interactions between these sequences is shown in table 3. There are no additional reaction products that do not belong to the system. The subsystems (1,2,3) and (1,4,5) are also self-sustaining.

The secondary structures resulting from these interactions can be seen in figure 36.

	1	2	3	4	5
1	-	3	2,1	5	4,1
2		-	2,1	-	4,1
3			2,1	2,1	2,1;4,1
4				-	4,1
5					4,1

Table 3: Matrix of interactions between the sequences of system 1

The frequency of these 5 sequences, their sequence pattern and of pattern B, D and E during the simulation can be seen in figure 37.

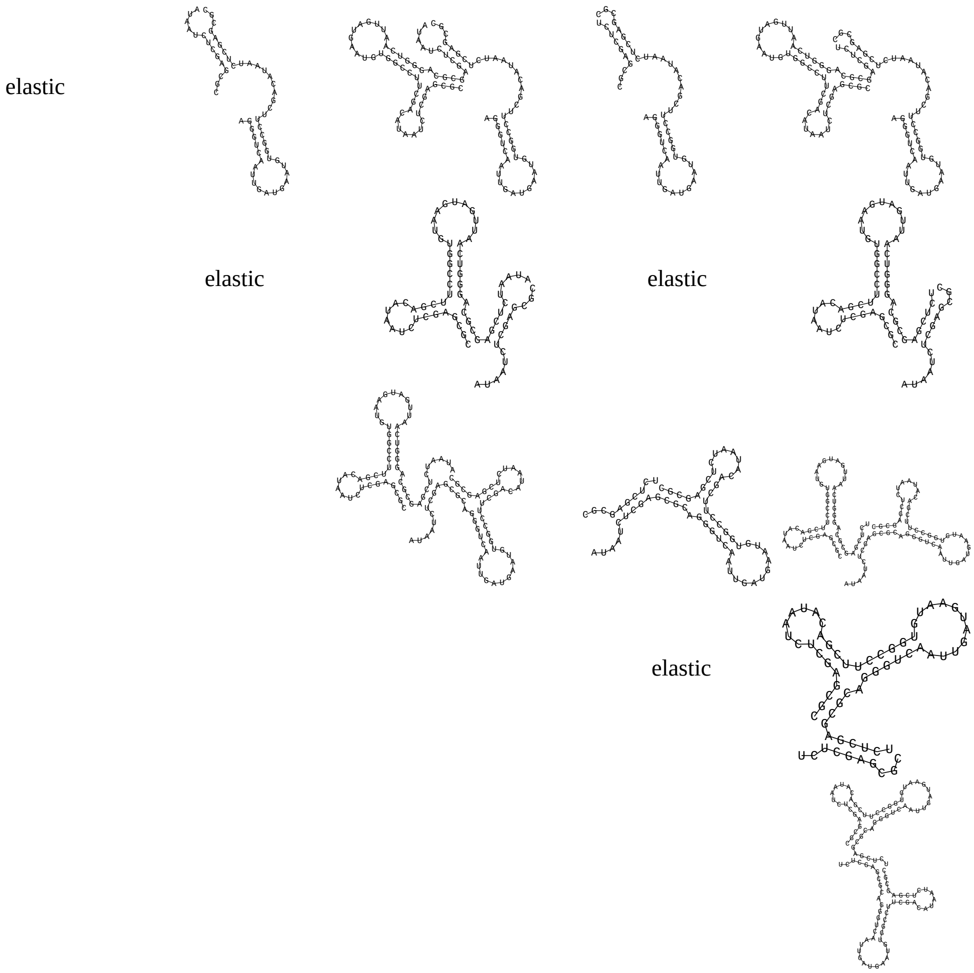
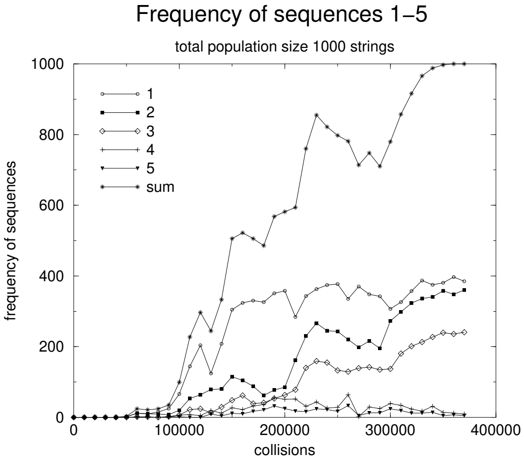
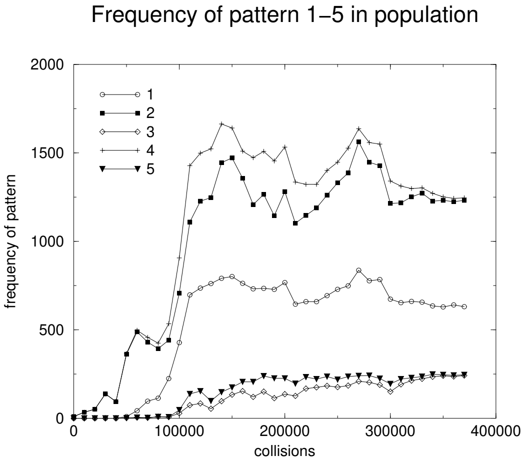


Figure 36: Secondary structures resulting from interactions between the 5 sequences of system 1





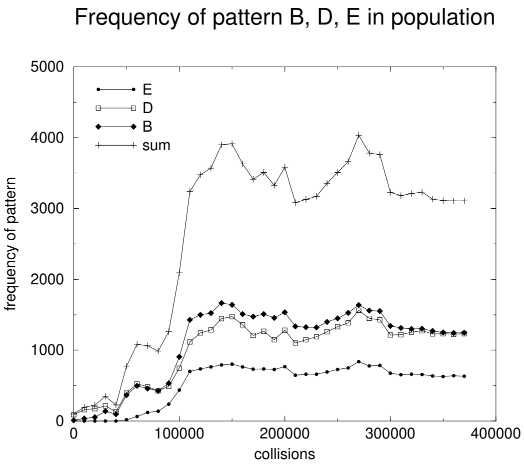


Figure 37: Frequency of the 5 sequences of system 1, their sequence pattern and of pattern B, D and E of which these sequences are composed, in the population of the simulation described in figure 25.

(ii) System 2

The simulation described in figure 26 resulted in a population of three different sequences: when interacting with sequences 1 or 3 sequence 2 is cleaved into these sequences 1 and 3, which in turn can interact to form sequence 2.

1	=	GAGCCAUCGGAACGGAAACUGAGCGGUAUGCGCGGU
2	=	GAGCCAUCGGAACGGAAACUGAGCGGUAUGCGCGGU
		GGGCUCCGGCCAGCGC
3	=	GGGCUCCGGCCAGCGC

2	=	1-3

The matrix of interactions and the corresponding secondary structures are displayed in Table 4 and figure 38.

The frequencies of these 3 sequences and their respective pattern during the simulation are shown in figure 39.

	1	2	3
1	-	1,3	2
2		-	1,3
3			-

Table 4: Matrix of interactions between the sequences of system 2

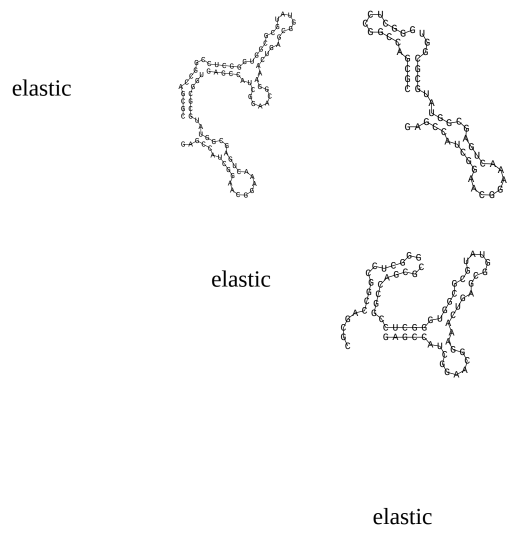


Figure 38: Secondary structures resulting from interactions between sequences of system 2

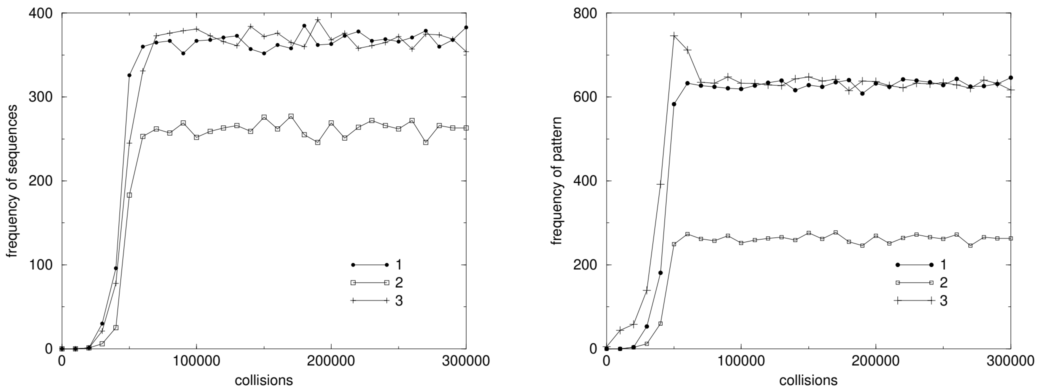


Figure 39: Frequency of the 3 sequences of system 2 and their sequence pattern in the population of the simulation described in figure 26.

(iii) System 3

The simulation described in figure 27 resulted in a population of 5 different sequences which are composed from 3 sequence pattern D, E and B.

D = ACACUACGACUAGCGUUGGGUACCAGCCGACAAUGGUAG UGAAUUCCGUC
E = AUUGGACCCUUUU
B = UUCCAUUGGGUGGCCGCCGAUGGUAUUUUUCUUGGAAGU AGACGUGACGACGCCCAUCGCUAAUGAGGAGGGUAACAC ACGGGCGGUGGAAGCUAUCCAUCCAGAAGGUG
1 = D
2 = D-B-E
3 = E-D
4 = E-D-B-E
5 = B-E

The interaction matrix can be seen in Table 5, the frequencies of the 5 sequences and the 3 pattern in figure 40.

	1	2	3	4	5
1	-	1,5	-	3,5	2
2		1,5	1,5	1,5;3,5	1,5
3			-	3,5	4
4				3,5	3,5
5					-

Table 5: Matrix of interactions between the sequences of system 3

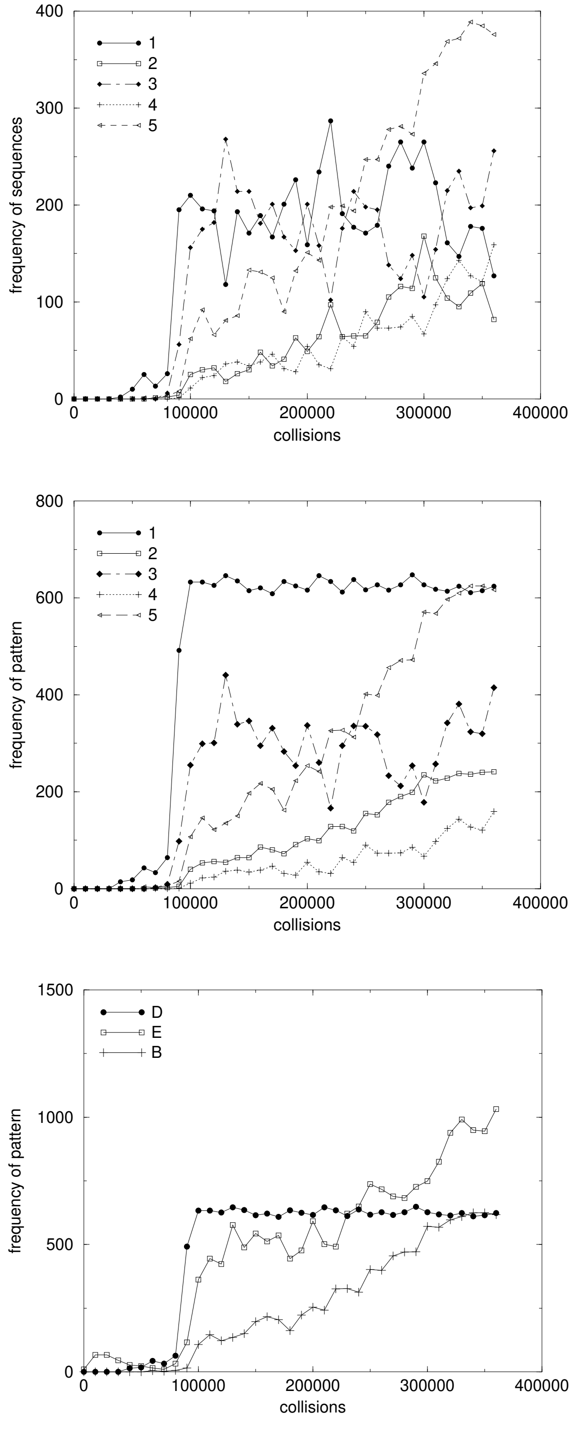


Figure 40: Frequency of the 5 sequences of system 3, their sequence pattern and of pattern D, E and B of which these sequences are composed, in the population of the simulation described in figure 27.

(iiii) System 4

This 7 sequence system evolved in a simulation with autocatalytic replication (figure 29). All seven sequences consist of only one pattern (sequence 1) that can ligate with itself by forming a hairpin; this leads to concatamers which refold and meet the criteria for cleavage, producing only the pattern or multimers of it. Since the first hammerhead in a structure is chosen for cleavage reactions, sequence 1 is the main cleavage product.

$$
\begin{array}{lcl}
1 & = & \text{ACUGUCGAUCGGAUAGCUAAUGCUAGGCCUCGCC} \\
2 & = & \text{1-1} \\
3 & = & \text{1-1-1} = \text{2-1} \\
\vdots & & \vdots
\end{array}
$$

Figure 41 shows the interaction structures up to sequence 6, the interaction matrix is shown in Table 6.

	1	2	3	4	5	6	7
1	2	1,1;3	4	2,2;5	1,4;6	1,5;7	4,3
2		1,1	1,2;5	1,1;2,2;6	1,1;2,3;7	1,1;1,5	1,1;2,5
3			1,2	1,2;1,3	1,2;1,4	1,2;1,5	1,2;1,6
4				1,3	1,3;1,4	1,3;1,5	1,3;1,6
5					1,4	1,4;1,5	1,4;1,6
6						1,5	1,5;1,6
7							1,6

Table 6: Matrix of interactions between the sequences of system 4

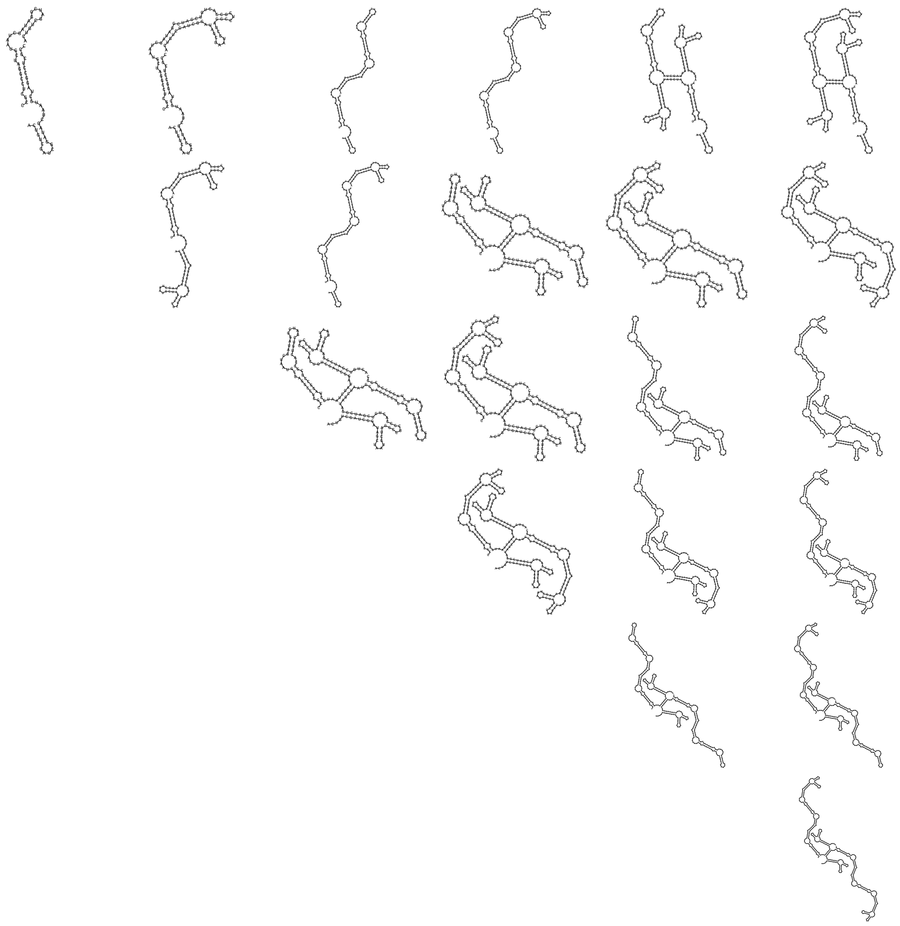


Figure 41: Secondary structures resulting from interactions between sequences of system 4

5.3.2 Cleavage of Hairpin Loops

(i) Loopsystem 1

The stable sequence distribution that was reached after 150 000 collisions in a simulation with boundary conditions constant organization and cleavage of hairpin loops with at least 5 unpaired bases consists of 8 sequences that are concatamers of only one pattern.

1	=	AGAGGCAAUGUGCGCGGUGGGGCCAGCCGAC
2	=	1-1
3	=	1-1-1
⋮		⋮

The interaction matrix is shown in Table 7, the secondary structures resulting from interactions between sequences 1 to 4 are displayed in figure 42. The frequency of sequences 1 to 8 and of the only pattern in the population can be seen in figure 43.

	1	2	3	4	5	6	7	8
1	2	1,1;3	4	1,3;5	1,4;6	1,5;7	1,6;8	1,7;9
2		1,1;4	1,2;5	1,1;1,3;6	1,1;1,4;7	1,1;1,5;8	1,1;1,6;9	1,1;1,7;10
3			1,2;6	1,2;1,3;7	1,2;2,3;8	1,2;1,5;9	1,2;1,6;10	1,2;1,7;11
4				1,3;8	2,2;1,4;9	1,3;3,3;10	1,3;1,6;11	1,3;1,7;12
5					1,4;10	2,3;1,5;11	1,4;3,4;12	1,4;1,7;13
6						1,5;12	2,4;1,6;13	1,5;3,5;14
7							1,6;14	2,5;1,7;15
8								1,7;16

Table 7: Matrix of interactions between the sequences of loopsystem 1

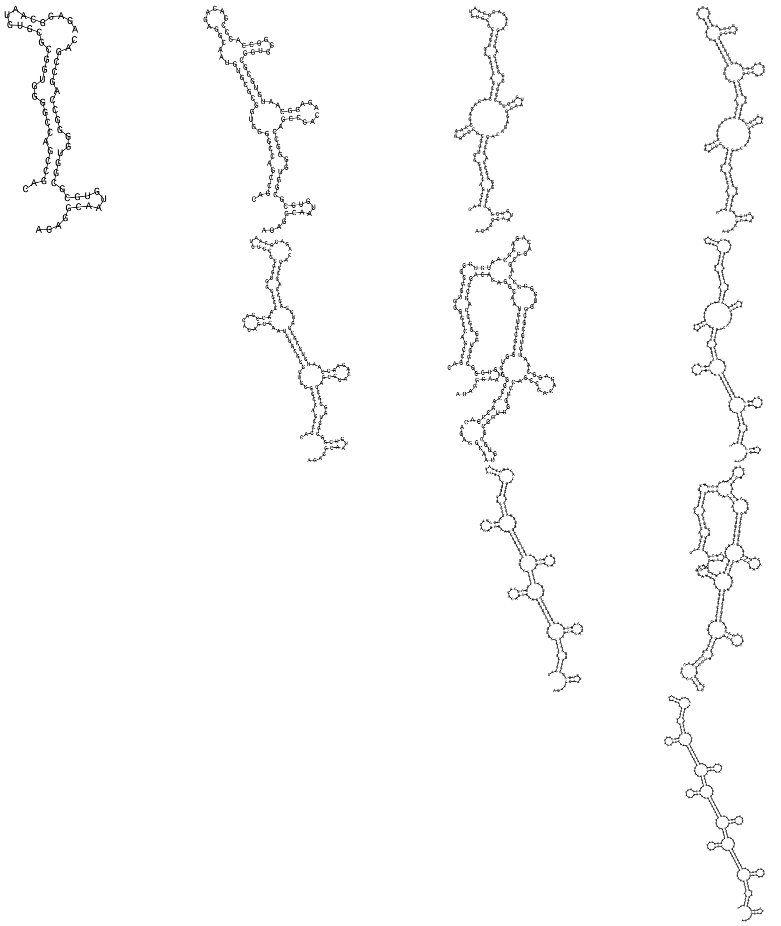


Figure 42: Secondary structures resulting from interactions between sequences 1 to 4 of loopsystem 1

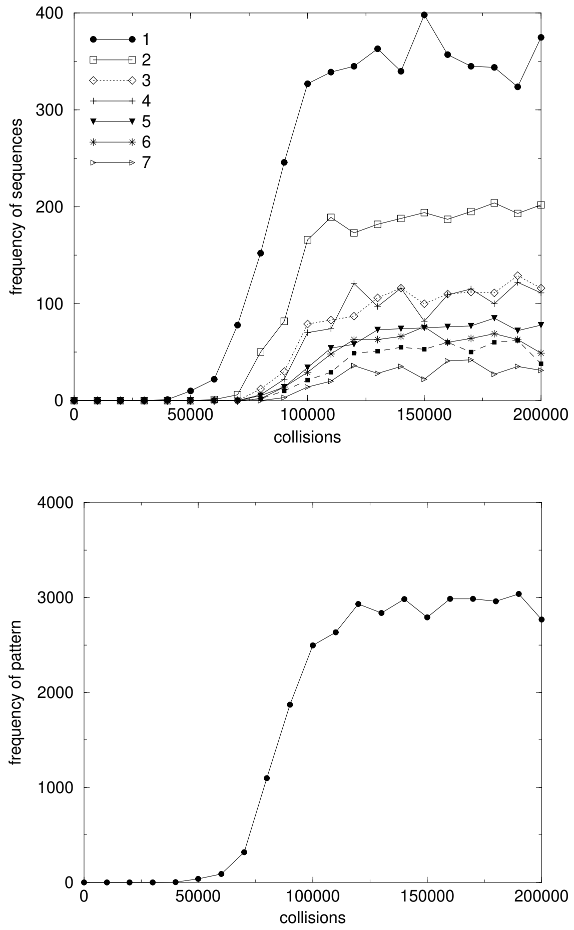


Figure 43: Frequency of sequences and the only pattern of loopsystem 1

5.4 Stability of Self-Sustaining Sequence Distributions

5.4.1 Stability Against Addition of Random Sequences

(i) System 1

An attempt to check for stability of system 1 was made by adding random sequences: the starting population (population size 1000) consisted of the 5 sequences of system 1 and different percentages of random sequences of length $n = 30$ (10 copies each). Simulations were performed with constant organisation setting and with conservation of mass during reactions.

(a) Conservation of Mass

The stable sequence distribution was reached in a simulation with constant organization. Relative concentrations of these 5 sequences with conservation of mass, but no perturbation by random sequences starting with 200 copies each can be seen in figure 44 (top). Average string length drops slightly from 35.4 to around 30, cleavage and ligation rates are both around 0.323.

Figure 44 (middle) and (bottom) shows the relative concentrations of these 5 sequences and reaction rates if 10 or 50 percent of the total population is composed of random strings, respectively.

In simulations with conservation of mass and no replication events the self-sustaining system can only survive in absence of other strings that can interact with sequences of the system, since strings from the system are used up in reactive collisions, but are not likely to be produced by such interactions.

(b) Constant Organization

The system is perturbed by addition of 50 or 80 percent random sequences (figure 45 left and right, respectively) with constant organization setting.

The subsystems (1,2,3) and (1,4,5) are self-sustaining.

In the simulation with 50 percent random sequences between collisions 14 000 and 35 000 two additional sequences (6 and 7) are added to the system;

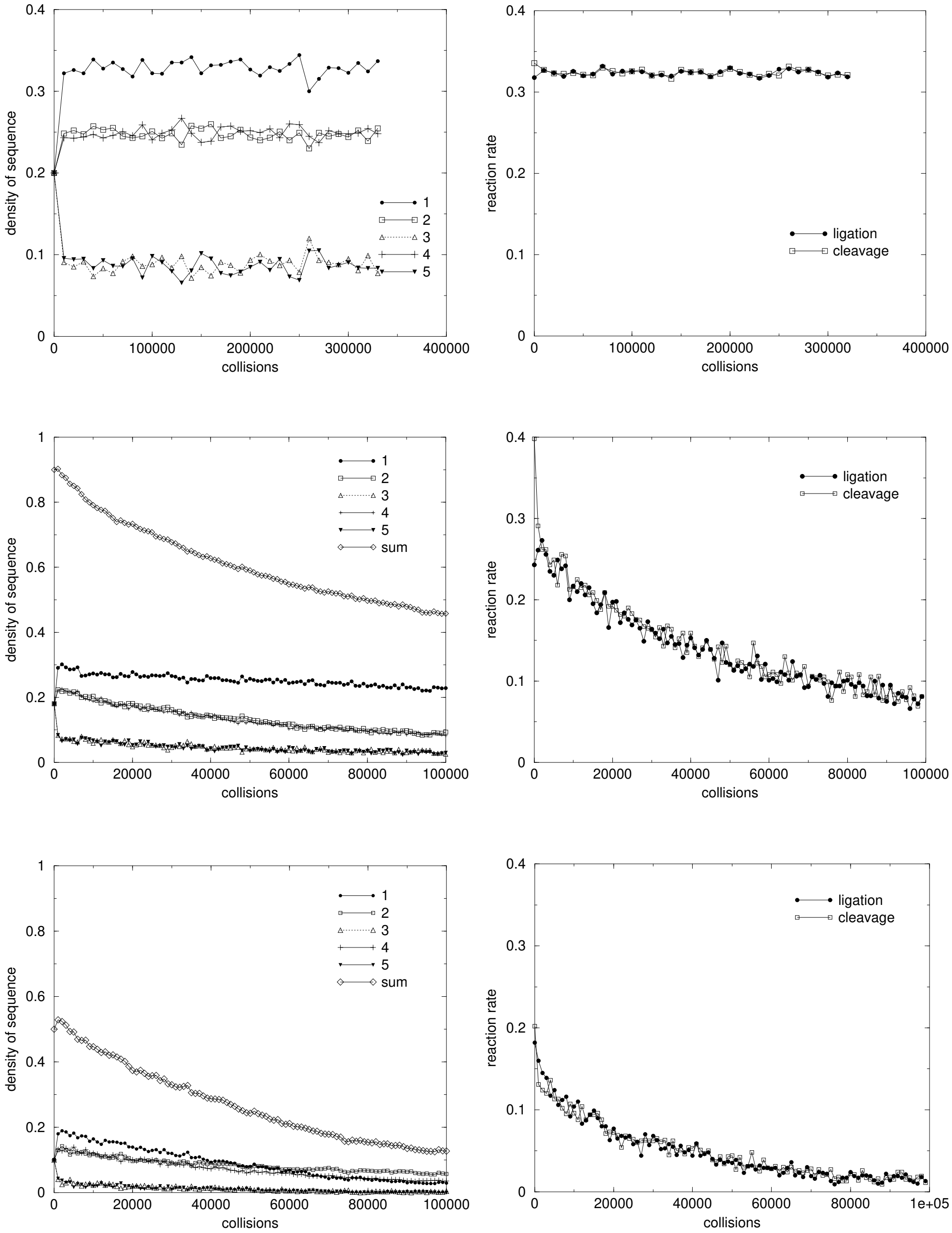


Figure 44: Frequency of sequences of system1 and reaction rates; conservation of mass during reactions, no replication, perturbation by addition of 0 (top), 10 (middle) and 50 (bottom) percent random strings of length $n = 30$, 10 copies each. Total population size 1000 strings, the sequences of the system start with equal number of copies.

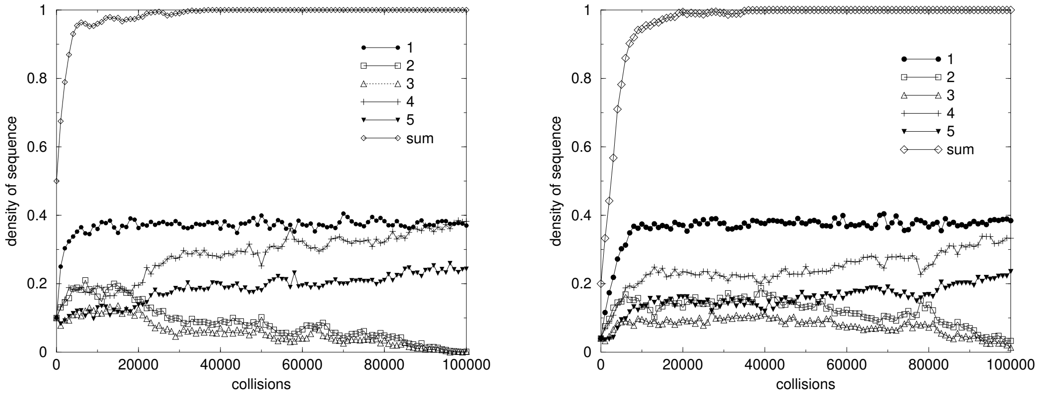


Figure 45: Frequency of sequences of system1; constant organization, perturbation by addition of 50 (left) and 80 (right) percent random strings of length $n = 30$, 10 copies each. Total population size 1000 strings, the sequences of the system start with an equal number of copies.

(1,6,7) again is a self-sustaining subsystem. Due to the fact that when the system first accounts for the whole population (after 37 000 collisions) sequences 4 and 5 outnumber sequences 3 and 2, and to random fluctuations the population converges to subsystem (1,4,5).

In the simulation with 80 percent random sequences 3 additional sequences (6, 7, 8) are incorporated into the self-sustaining system between collisions 5 000 and 34 000 by addition of the same subsequence to 1, 3 and 5. The subsystem (6, 7, 8) is not self-sustaining; interactions between these three sequences produce (6, 2, 4), resulting in a self-sustaining subsystem (2, 4, 6, 7, 8); but since the number of copies of the new sequences is too low, they are lost again. The population again converges to subsystem (1,4,5).

5.4.2 Stability Against Mutation

(i) System 1, replication accuracy 0.995

The reactor was filled with the five sequences of system 1, 200 copies each. Boundary conditions of this simulation are constant organization, the replication accuracy is 0.995 per digit, uniformly along the chain.

The original network of reactions is replaced by a set of reactions that consist of ligation of strings of length 43, resulting in concatamers that can be cleaved producing strings of length 43 and multiples thereof.

Figure 46 shows the string length distribution in the population from start to 340 000 collisions. After 340 000 collisions the 4 string length classes (43, 86, 129 and 172) participating in the self-sustaining reaction network account for 76 percent of total population size. Of these 4 classes only string length 43 has been present in the initial population (sequence 1).

Figure 47 shows the hamming distances of the most frequent sequences of length 43 to sequence 1 and the distances to most frequent sequences 50 000 collisions before. While a collision between 2 sequence1 molecules does not result in ligation, after 30 000 collisions one of the most frequent sequences of string length class 43, which has hamming distance 3 to sequence1, already shows this property.

In a simulation with replication accuracy 0.999 a development towards the same string length distribution is observed.

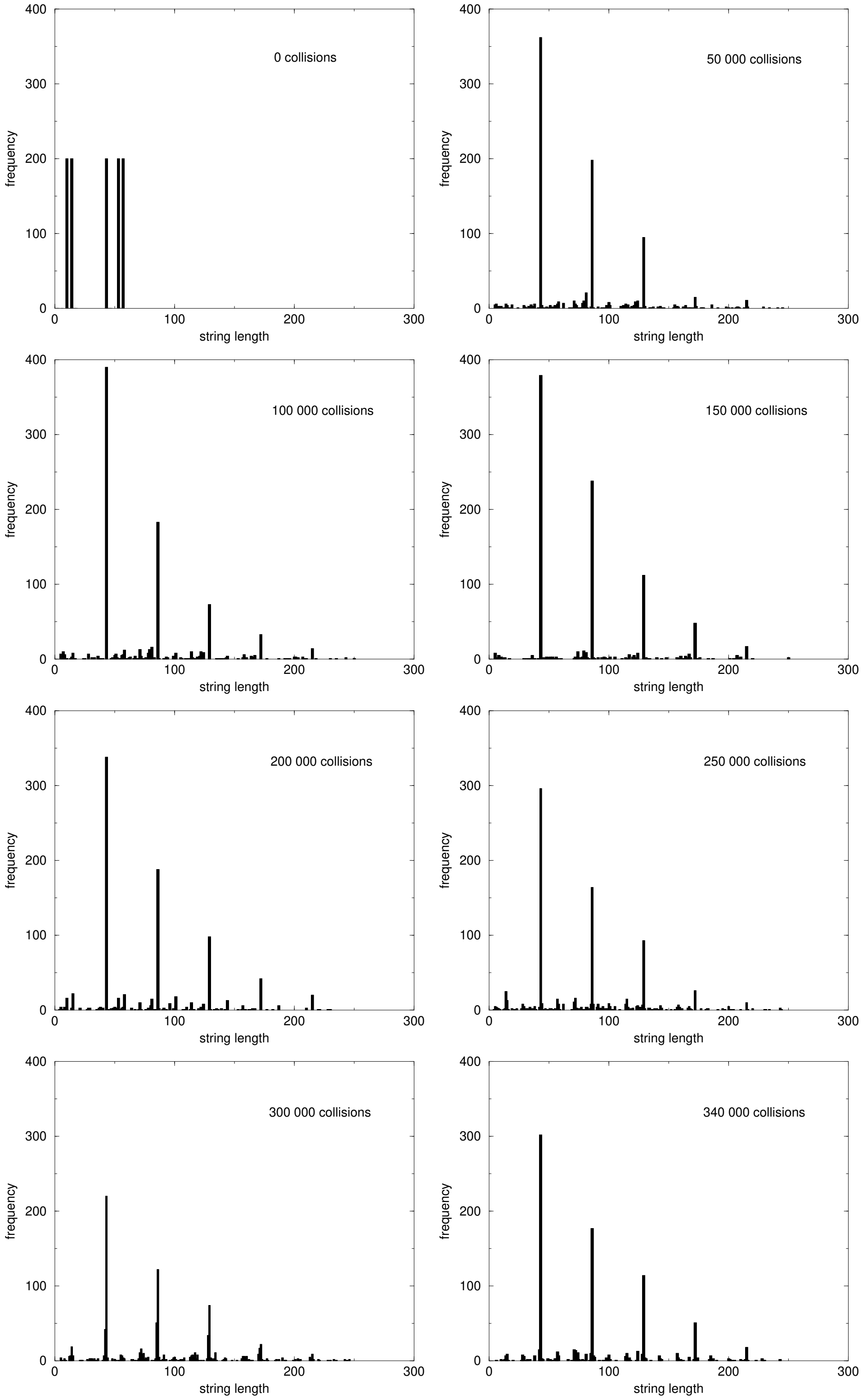


Figure 46: Perturbation of system 1 by mutation; constant organization, replication accuracy 0.995, total population size 1000 strings. The distribution of string lengths in the population after every 50 000 collisions is displayed. The original network of reactions between the 5 sequences of system 1 is replaced by a different one between members of string length classes, which is again self-sustaining.

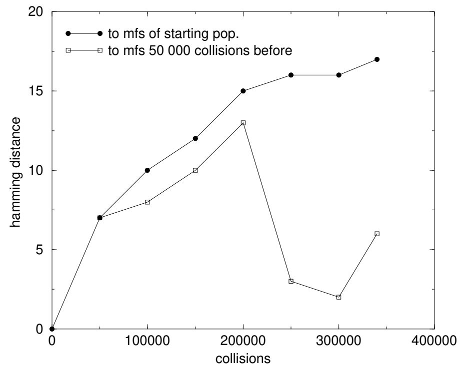


Figure 47: Hamming distances of the most frequent sequences of length 43 to sequence 1 and to the most frequent sequence 50 000 collisions before. Data from the simulation described in figure 46.

(ii) System 3, replication accuracy 0.999

The reactor was filled with the five sequences of system 3, 200 copies each. Boundary conditions in this simulation are constant organization and hammerhead cleavage, the replication accuracy is 0.999 per digit, uniformly along the chain.

The original network of reactions is not stable against mutation. After 400 000 collisions a system of 3 different interacting string length classes (19, 50, 69) turns up; this system is stable for 500 000 collision, then disappears. The most common sequence of length 50 in this network changes only slightly from 200 000 to 900 000 collisions, but has only 20 percent in common with the sequence of length 50 from the starting population.

In figure 48 average string length, reactivity and population diversity are monitored for this simulation. The sudden change in these values after 920 000 collisions marks the disappearance of the reaction network.

Figure 49 shows the distribution of string lengths during the simulation. This simulation was continued; after 1 100 000 collisions a new system of 3 different string length classes came up (32, 104, 136).

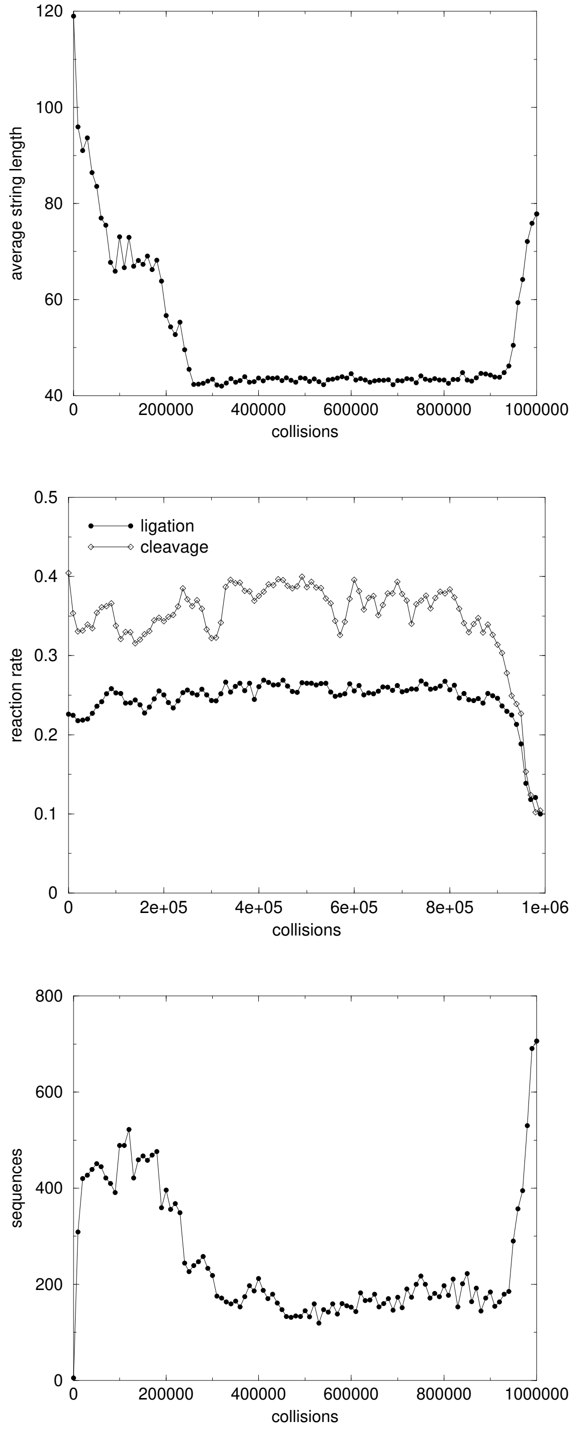


Figure 48: Average string length, reactivity and population diversity of the perturbation of system 3 by introducing mutation (replication accuracy 0.999). A quasi-stationary distribution of sequences between collisions 400 000 and 900 000 could be observed. The sharp change of these values marks the disappearance of this self-sustaining network.

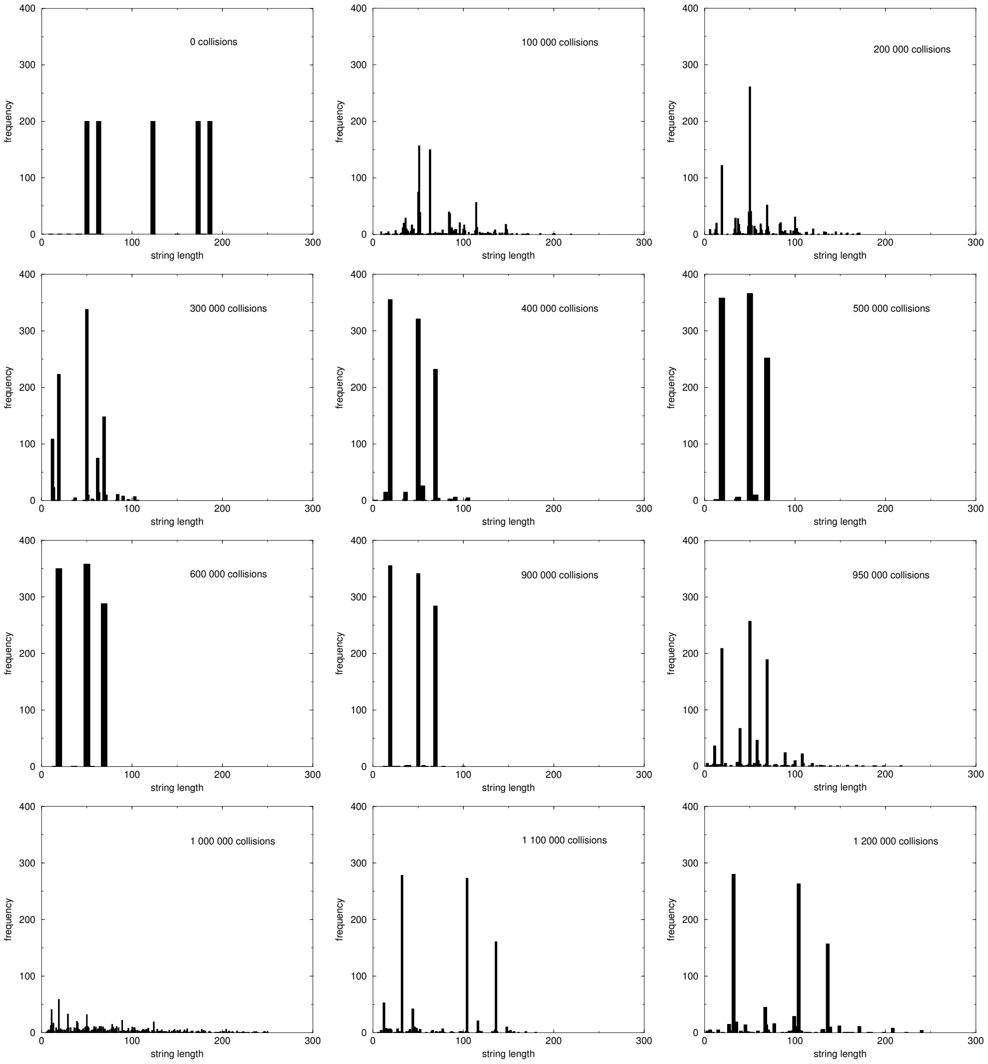


Figure 49: Perturbation of system 3 by mutation; constant organization, replication accuracy 0.999, total population size 1000 strings. The development of the distribution of string lengths in the population is displayed. After 400 000 collisions the original network of reactions between the 5 sequences of system 3 is replaced by a different system of 3 string length classes, which disappears after 900 000 collisions. 200 000 collisions later a new network is established.

6 Conclusions and Outlook

Populations of RNA molecules are known to exhibit all properties necessary for evolutionary adaptation: they provide a template for their own replication, variation is created by erroneous copying and recombination, whereas selection operates on the phenotype, which is the three-dimensional RNA structure that is determined by the sequence of bases in the polynucleotide chain.

RNA is particularly well suited to study evolution and functional self-organization both experimentally and in computer simulation. In order to model evolution in an RNA scenario the 3D-structure is replaced by the secondary structure which represents a coarse grained version focussing on base pairs. The secondary structure represents a compromise between being as realistic as possible and as simple as necessary. In particular, for RNA secondary structure predictions highly efficient algorithms are available which allow to handle up to billions of sequences in reasonable times.

Accordingly, we can design computer experiments to model different aspects of evolutionary behaviour in a world populated by RNA sequences.

In previous experiments computer simulations were carried out to model the development of populations of replicating and mutating RNA molecules in an evolution reactor, with fitness values of sequences determined by replication and degradation rate constants derived from their minimum free energy secondary structure. Here a different approach is made: inspired by the finding that naturally occurring RNA can catalyze reactions involving other RNA molecules, catalytic functions are assigned to the phenotype. This phenotype is not a property of an individual molecule, but is determined by its interaction with an other RNA sequence in the cofolded complex.

Reactions known to be catalyzed by ribozymes involve the sugar-phosphate backbone; ligation and cleavage of polynucleotide strands take place in the catalytic core of a ribozyme-substrate-complex, which often consists of a rather small structural motif.

The development of populations of catalytically active RNA molecules is simulated in an evolution reactor where randomly colliding RNA molecules interact by cofolding into a common secondary structure. If certain predefined structural motifs occur, cleavage or ligation reactions take place at these catalytic sites.

The question to be answered was: "Under which conditions can populations evolve to a quasi-stationary sequence distribution where all sequences of this population are produced by a network of catalytic interactions between members of that same population, in other words, are there sets of RNA species with functional closure?".

The problem was addressed in two steps. First simulations without replication errors have been performed; in this setup all sequence diversity comes from rearrangement of sequences resulting from cleavage or ligation processes. Second, experiments with erroneous replication were carried out.

Results for the error-free model system

(i) Mass is conserved during reactions - substrates are used up:

- Without any replication events the population can change only by reassembly of subsequences and quickly approaches an unreactive state of equilibrium.
- In a setup where partners in unreactive collisions are taken out of the reactor and two randomly chosen strings are replicated, population diversity was reduced to 2 sequences that did not interact. When in addition to this condition collisions between identical strings lead to autocatalytic replication of this sequence, a system of 4 reactive sequences was found, only three of which could be produced by an interaction.

- An attempt to favour reactive strings was made by introducing an additional replication/dilution event after every collision. When ligating strings were preferentially selected for replication, the population quickly collapsed to one sequence, favouring sequences that have been involved in cleavage reactions led to a growing population composed mainly of a few interacting sequences and their cleavage products.

(ii) Constant organization - substrates remain in the reactor, products are added, population size is kept constant by an unspecific dilution flow:

- This setup without additional replication events favours emerging of self-sustaining reaction networks. Reactivity quickly increases, and after about 100 000 collisions, populations consist of a few sequences all of which are produced in reactive interactions between members of the population.

- Additional autocatalytic replication events lead to a similar self-sustaining system which is composed of concatamers of only one sequence pattern.

- Introducing an additional replication/dilution event after every collision favouring reactive sequences results in high ligation or cleavage rates, but does not lead to self-sustaining sequence distributions.

After identification of suitable boundary conditions, which resulted in quasi-stationary self-sustaining sequence distributions, simulations with different error rates were performed. In the constant organization setting, substrates were not left unchanged in the reactor as before, but were replaced by mutants.

Results for the model system with different error rates

- If replication accuracy is high, reaction rates quickly rise and we observe the formation of quasi-stationary string length distributions. Hamming distances between members of the same string length class are low, with interactions between different string length classes resulting in the same or similar structures that fulfill the criteria for reaction. These interactions between subpopulations consisting of sequences of the same length give rise to a self-sustaining network of reactions.

- At higher mutation rates string lengths are randomly distributed and reactivity stays low.

Two self-sustaining populations that had resulted from simulations without mutation events were perturbed by introducing erroneous replication. The stability of their interactions against error rates that were previously found to allow for the development of reaction networks was investigated. Both sets of sequences quickly disappeared and were replaced by different quasi-stationary string length distributions that were again found to be self-sustaining.

"Molecular ecology experiments" with functionally coupled amplifying systems based on nucleic acids have been carried out in John McCaskill's group [78]. These model systems simulate *in vitro* coevolution in a molecular predator-prey system. Computer simulations in configurable hardware have been used to study artificial DNA/RNA chemistries similar to the *in vitro* biochemistry of these systems [4].

An interesting field for further investigations is the effect of different mutation rates on the rise and fall of quasi-stationary string length distributions in larger populations of interacting RNA molecules. Can we find an error threshold for reaction networks?

The effect of the interplay between cleavage and ligation reactions in a population of RNA strings can be compared with that of recombination events.

Recombination occurs at random positions in the genome; two sequence parts of equal length are exchanged. Since the sites of these sequence exchanges are not motif-dependent such recombination events usually lead to loss of biological function. Cleavage and ligation sites in reactions catalyzed by ribozymes strongly depend on RNA structure and are therefore more likely to lead to the exchange of functional units that have variable length.

To determine whether a set of sequences of equal length stays clustered around a point in sequence space in a quasispecies-like manner, or if it slowly drifts through sequence space together with its interaction partners, longer simulations have to be performed in larger populations that reduce the probability that the observed system disappears due to stochastic effects.

Further insight into the properties of reaction networks could be gained by studying the role of neutrality in this context.

References

[1] D. P. Bartel and J. W. Szostak. Isolation of new ribozymes from a large pool of random sequences. *Science*, 261:1411–1418, 1993.

[2] C. K. Biebricher. Darwinian selection of self-replicating RNA molecules. *Evolutionary Biology*, 16:1–52, 1983.

[3] C. K. Biebricher, M. Eigen, and W. C. Gardiner Jr. Kinetics of RNA replication. *Biochemistry*, 22:2544–2559, 1983.

[4] J. Breyer, J. Ackermann, and J. McCaskill. Evolving reaction-diffusion ecosystems with self-assembling structures in thin films. *Artificial Life*, 4(1):25–40, 1998.

[5] J. M. Burke. Hairpin ribozyme: current status and future prospects. *Biochem. Soc. Trans.*, 24:608–615, 1996.

[6] J. H. Cate, A. R. Gooding, E. Podell, K. Zhou, B. L. Golden, C. E. Kundrot, T. R. Cech, and J. A. Doudna. Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science*, 273:1678–1685, 1996.

[7] T. R. Cech. A model for the RNA-catalyzed replication of RNA. *Proc Natl Acad Sci U S A*, 83:4360–3, 1986.

[8] T. R. Cech. RNA as an enzyme. *Scientific American*, 11:76–84, November 1986.

[9] T. R. Cech. RNA chemistry. Ribozyme self-replication? *Nature*, 339:507–8, 1989.

[10] T. R. Cech. Self-splicing of group I introns. *Ann. Rev. Biochem.*, 59:543, 1990.

[11] T. R. Cech. The efficiency and versatility of catalytic RNA: implications for an RNA world. *Gene*, 135:33–6, 1993.

[12] T. R. Cech and O. C. Uhlenbeck. Ribozymes. Hammerhead nailed down. *Nature*, 372:39–40, 1994.

[13] J. A. Doudna and J. W. Szostak. RNA-catalysed synthesis of complementary-strand RNA. *Nature*, 339:519–22, 1989.

[14] J. A. Doudna, N. Usman, and J. W. Szostak. Ribozyme-catalyzed primer extension by trinucleotides: a model for the RNA-catalyzed replication of RNA. *Biochemistry*, 32:2111–5, 1993.

[15] F. Eckstein. RNA interactions: Ribozymes and antisense. *Biochem. Soc. Trans.*, 24:601–604, 1996.

[16] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 10:465–523, 1971.

[17] M. Eigen, J. McCaskill, and P. Schuster. The molecular Quasispecies. *Adv. Chem. Phys.*, 75:149 – 263, 1989.

[18] M. Eigen and P. Schuster. The hypercycle A: A principle of natural self-organization: Emergence of the hypercycle. *Naturwissenschaften*, 64:541–565, 1977.

[19] E. H. Ekland and D. P. Bartel. The secondary structure and sequence optimization of an RNA ligase ribozyme. *Nucleic Acids Res*, 23:3231–8, 1995.

[20] E. H. Ekland and D. P. Bartel. RNA-catalysed RNA polymerization using nucleoside triphosphates. *Nature*, 382:373–376, 1996.

[21] E. H. Ekland, J. W. Szostak, and D. P. Bartel. Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science*, 269:364–370, 1995.

[22] M. J. Fedor and O. C. Uhlenbeck. Kinetics of intermolecular cleavage by hammerhead ribozymes. *Biochemistry*, 31:12042–54, 1992.

[23] W. Fontana and L. W. Buss. The arrival of the fittest: toward a theory of biological organizazion. *Bull. Math. Biol*, 56:1 – 64, 1994.

[24] W. Fontana and L. W. Buss. What would be conserved if 'the tape were played twice'? *Proc. Natl. Acad. Sci. USA*, 1994.

[25] W. Fontana, T. Griesmacher, W. Schnabl, P. F. Stadler, and P. Schuster. Statistics of landscapes based on free energies, replication and degradation rate constants of RNA secondary structures. *Monatsh. Chem.*, 122:795–819, 1991.

[26] W. Fontana, W. Schnabl, and P. Schuster. Physical aspects of evolutionary optimization and adaptation. *Physical Review A*, 40(6):3301–3321, Sep. 1989.

[27] W. Fontana and P. Schuster. A computer model of evolutionary optimization. *Biophysical Chemistry*, 26:123–147, 1987.

[28] W. Fontana and P. Schuster. Continuity in evolution: On the nature of transitions. *Science*, 280:1451–1455, 1998.

[29] W. Fontana and P. Schuster. Shaping space. The possible and the attainable in RNA genotype-phenotype mapping. *J.Theor.Biol.*, 194:491–515, 1998.

[30] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for prediction of RNA duplex stability. *Proc.Natl.Acad.Sci.USA*, 83:9373–9377, 1986.

[31] W. Gilbert. The RNA world. *Nature*, 319:618, 1986.

[32] R. Gutell. Comparative studies of RNA: Inferring higher order structure from patterns of sequence variation. *Current Opinion in Structural Biology*, 3:313, 1993.

[33] A. J. Hager, J. D. Pollard, and J. W. Szostak. Ribozymes: aiming at RNA replication and protein synthesis. *Chemistry & Biology*, 3:717–725, 1996.

[34] J. Haseloff and W. L. Gerlach. Simple RNA enzymes with new and highly specific endoribonuclease activities. *Nature*, 334:585–591, 1988.

[35] P. Hendry and M. McCall. Unexpected anisotropy in substrate cleavage rates by asymmetric hammerhead ribozymes. *Nucleic Acids Research*, 24(14):2679–2684, 1996.

[36] K. J. Hertel. Numbering system for the hammerhead. *Nucleic Acids Research*, 20:3252, 1992.

[37] K. J. Hertel, D. Herschlag, and O. C. Uhlenbeck. A kinetic and thermodynamic framework for the hammerhead ribozyme reaction. *Biochemistry*, 33:3374–85, 1994.

[38] I. L. Hofacker. *The rules of the evolutionary game for RNA: A statistical characterization of the sequence to structure mapping in RNA*. PhD thesis, University of Vienna, 1994.

[39] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh.Chem.*, 125(2):167–188, 1994.

[40] I. L. Hofacker, W. Fontana, P. F. Stadler, and P. Schuster. `Vienna RNA Package`. `http://www.tbi.univie.ac.at/ ivo/RNA/`, 1994. (Free Software).

[41] C. J. Hutchins, P. D. Rathjen, A. C. Forster, and R. H. Symons. Self-cleavage of plus and minus RNA transcripts of avocado sunblotch viroid. *Nucleic Acids Research*, 14:3627–3640, 1986.

[42] M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. (USA)*, 93:397–401, 1996.

[43] S. Kim, F. Suddath, G. Quigley, A. McPherson, J. Sussman, A. Wang, N. Seeman, and A. Rich. Three dimensional tertiary structure of yeast phenylalanine tRNA. *Science*, 185:435, 1974.

[44] D. M. Long and O. C. Uhlenbeck. Self-cleaving catalytic RNA. *FASEB*, 7:25–30, 1993.

[45] H. M. Martinez. An RNA folding rule. *Nucl.Acid.Res.*, 12:323–335, 1984.

[46] J. Maynard Smith and E. Szathmary. *The Major Transitions in Evolution.* W. H. Freeman, Oxford, UK, 1995.

[47] M. J. McCall, P. Hendry, and P. A. Jennings. Minimal sequence requirements for ribozyme activity. *Proc Natl Acad Sci U S A*, 89:5710–4, 1992.

[48] D. R. Mills, R. L. Peterson, and S. Spiegelman. An extracellular darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Nat. Acad. Sci., USA*, 58:217–224, 1967.

[49] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. USA*, 77(11):6309–6313, 1980.

[50] C. Papanicolau, M. Gouy, and J. Ninio. An energy model that predicts the correct folding of the tRNA and the 5S RNA molecules. *Nucl.Acid.Res.*, 12:31–44, 1984.

[51] J. A. Piccirilli, T. S. McConnell, A. J. Zaug, H. F. Noller, and T. R. Cech. Aminoacyl-esterase activity of the *tetrahymena* ribozyme. *Science*, 256:1420–1424, 1992.

[52] H. W. Pley, K. M. Flaherty, and D. B. McKay. Three-dimensional structure of a hammerhead ribozyme. *Nature*, 372:68–74, 1994.

[53] G. A. Prody, J. T. Bakos, J. M. Buzayan, I. R. Schneider, and G. Bruening. Autolytic processing of dimeric plant virus satellite RNA. *Science*, 231:1577–1580, 1986.

[54] J. R. Prudent, T. Uno, and P. G. Schultz. Expanding the scope of RNA catalysis. *Science*, 264:1924–1927, 1994.

[55] D. E. Ruffner, G. D. Stormo, and O. C. Uhlenbeck. Sequence requirements of the hammerhead RNA self-cleavage reaction. *Biochemistry*, 29:10695–10702, 1990.

[56] R. Saldanha, G. Mohr, M. Belfort, and A. M. Lambowitz. Group I and group II introns. *FASEB J.*, 7:15–24, 1993.

[57] P. Schuster. Mechanisms of molecular evolution. In S. W. Fox, editor, *Selforganization*, pages 57 – 91. 1986.

[58] P. Schuster. Complex optimization in an artificial RNA world. In D. Farmer, C. Langton, S. Rasmussen, and C. Taylor, editors, *Artificial Life II*, volume XII. Addison-Wesley, SFI Studies in the Science of Complexity, 1991.

[59] P. Schuster. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *J. Biotechnology*, 41:239–257, 1995.

[60] P. Schuster. How does complexity arise in evolution? *Complexity*, 2:22–30, 1996.

[61] P. Schuster. Genotypes with phenotypes: Adventures in an RNA toy world. *Biophys. Chem.*, 66:75–110, 1997.

[62] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc.Roy.Soc.(London)B*, 255:279–284, 1994.

[63] P. Schuster and K. Sigmund. Dynamics of evolutionary optimization. *Berichte der Bunsen-Gesellschaft für physikalische Chemie*, 89:668–682, 1985.

[64] P. Schuster and P. F. Stadler. Landscapes: Complex optimization problems and biopolymer structures. *Computers Chem.*, 18:295–314, 1994.

[65] P. Schuster and P. F. Stadler. Sequence redundancy in biopolymers: A study on RNA and protein structures. In G. Myers, editor, *Viral Regulatory Structures*, volume XXVIII of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 163–186. Addison-Wesley, Reading MA, 1998. Santa Fe Institute Preprint 97-07-67.

[66] P. Schuster, P. F. Stadler, and A. Renner. RNA Structure and folding. From conventional to new issues in structure predictions. *Curr. Opinion Struct. Biol.*, 7, 1997. 229-235.

[67] W. G. Scott. Molecular palaeontology: understanding catalytic mechanisms in the RNA world by excavating clues from a ribozyme three-dimensional structure. *Biochem. Soc. Trans.*, 24:604–608, 1996.

[68] B. A. Shapiro and K. Zhang. Comparing multiple RNA secondary structures using tree comparisons. *CABIOS*, 6:309–318, 1990.

[69] S. Spiegelman. An approach to the experimental analysis of precellular evolution. *Quart. Rev. Biophys.*, 17:213, 1971.

[70] E. Szathmary and J. Maynard Smith. The major evolutionary transitions. *Nature*, 374:227–232, 1995.

[71] M. Tacker. *Robust Properties of RNA Secondary Structure Folding Algorithms*. PhD thesis, University of Vienna, 1993.

[72] M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. Algorithm independent properties of RNA secondary structure prediction. *Eur. Biophys. J.*, 25:115–130, 1996.

[73] T. Tuschl and F. Eckstein. Hammerhead ribozymes: importance of stem-loop II for activity. *Proc Natl Acad Sci U S A*, 90:6991–4, 1993.

[74] T. Tuschl, C. Gohlke, T. M. Jovin, E. Westhof, and F. Eckstein. A three-dimensional model for the hammerhead ribozyme based on fluorescence measurements. *Science*, 266:785–789, 1994.

[75] O. C. Uhlenbeck. A small catalytic oligoribonucleotide. *Nature*, 328:596–600, 1987.

[76] M. S. Waterman. Secondary structure of single - stranded nucleic acids. *Studies on foundations and combinatorics, Advances in mathematics supplementary studies, Academic Press N.Y.*, 1:167 – 212, 1978.

[77] C. Wilson and J. W. Szostak. In vitro evolution of a self-alkylating ribozyme. *Nature*, 374:777–82, 1995.

[78] B. Wlotzka and J. S. McCaskill. A molecular predator and its prey: coupled isothermal amplification of nucleic acids. *Chemistry & Biology*, 4(1):25–33, 1997.

[79] C. Woese, L. Magrum, R. Gupta, R. Siegel, D. Stahl, J. Kop, N. Crawford, J. Brosius, R. Gutell, J. Hogan, and H. Noller. Secondary structure model of bacterial 16s ribosomal RNA: Phylogenetic, enzymatic and chemical evidence. *Nucl.Acid.Res*, 8:2275–2293, 1980.

[80] J. H. Yang, N. Usman, P. Chartrand, and R. Cedergren. Minimum ribonucleotide requirement for catalysis by the RNA hammerhead domain. *Biochemistry*, 31:5005–9, 1992.

[81] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52, 1989.

[82] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull.Math.Biol.*, 46(4):591–621, 1984.

[83] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.

Curriculum vitae

Bärbel Krakhofer

∗ 1966-12-17, Wien

1973-1977	:	Volksschule, Bruck/L. und Eisenstadt
1977-1985	:	Bundesgymnasium Eisenstadt
1985	:	Abschluß u. Matura mit Auszeichnung
10/1985 - 03/1995	:	Studium der Biochemie, Universität Wien
10/1993 - 01/1995	:	Diplomarbeit am Institut für Theoretische Chemie und Strahlenchemie der Universität Wien bei Prof. Peter Schuster
02/03/1995	:	2. Diplomprüfung mit Auszeichnung
20/03/1995	:	Sponsion zum Magister der Naturwissenschaften
07/1995 - 08/1998	:	Dissertation am Institut für Theoretische Chemie und Strahlenchemie der Universität Wien bei Prof. Peter Schuster