PREDICTION AND VISUALIZATION OF STRUCTURAL SWITCHES IN RNA

ROBERT GIEGERICH [a]
University Bielefeld, Postfach 100131, 33501 Bielefeld, Germany

DIRK HAASE
National Research Center for Environment and Health (GSF), Dept. MIPS, Am Klopferspitz 18a, 82152 Martinsried, Germany

MARC REHMSMEIER
Deutsches Krebsforschungszentrum(DKFZ), Abt. Theoretische Bioinformatik, 69120 Heidelberg, Germany

There are various cases where the biological function of an RNA molecule involves a reversible change of conformation. paRNAss is a software approach to the prediction of such structural switching in RNA. It is based on three hypotheses about the secondary structure space of a switching RNA molecule, which can be evaluated by RNA folding and structure comparison. In the positive case, the predicted structures must be verified experimentally. Additionally, we give an animated visualization of an energetically favourable transition between the predicted structures. paRNAss is available via the Bielefeld Bioinformatics Server [1].
This paper explains the underlying model and shows that the approach performs well in a variety of applications.

1 Motivation

1.1 Conformational Switching in RNA

RNA fulfills a broad range of functions in living cells. In messenger RNA, the plain sequence of bases, the primary structure, is sufficient to determine the sequence of amino acids of the encoded protein. In other cases, e.g. in ribosomal RNA or transfer RNA a certain three dimensional conformation is necessary for the correct function. This structure is not rigid, and sometimes even a significant change of shape is required. Such conformational switches have been proven or are suspected to be involved in several important processes: regulation of gene expression in prokaryotes by attenuation[2], translational regulation of *E. coli* ribosomal protein S15 [3], regulation of self-cleavage activity of *Hepatitis Delta Virus* [4], translocation process in protein biosynthesis [5], trans splicing in trypanosomes [6], splicing of pre mRNA by spliceosomes [7].

[a] To whom correspondence should be addressed: robert@techfak.uni-bielefeld.de

To our knowledge, the problem of software support for the detection of switching phenomena has not been addressed before.

1.2 Properties of Current RNA Structure Prediction Programs

Exploring conformational switching requires knowledge of at least two molecular arrangements. In contrast to 3D-conformation (see e.g. [8]), secondary structure can be determined computationally to a sensible degree of correctness, and with acceptable efficiency. Therefore it must serve as an approximation of the 3D shape. Throughout this paper we assume that the switching involves a change of secondary structure.

paRNAss does not provide another folding program - it uses MFOLD by Zuker [9], and RNAfold [10], which is part of the Vienna RNA Package.

As is, pseudoknots (as opposed to unknotted or planar structures) are not recognized by these folding programs. This implies the possibility that a pseudoknot is detected in the guise of two alternative planar structures close to the energetic minimum. These might be suggested as alternative positions of a switch. Hence, paRNAss includes a specific check for this situation.

2 Outline of the paRNAss Approach

This section introduces some terminology and explains the paRNAss approach by means of a successful application, using a known conformational switch. All algorithmic details, method parameters, pitfalls etc. will go unmentioned until their detailed treatment in later sections.

2.1 Some Observations and Hypotheses about the Structure Space of a given RNA sequence

The basic mechanism of RNA structure formation is base pairing. The *combinatorial structure space* of a given RNA sequence x is solely determined by a given set of pairing rules, most commonly the Watson-Crick pairs (A-U, C-G) and the pair (G-U). It comprises all the structures that can be formed according to these rules. The size of the combinatorial structure space is exponential in the length of x. Waterman gives the asymptotic formula $\sqrt{\frac{15+7\sqrt{5}}{8\pi}} n^{-3/2} \left(\frac{3+\sqrt{5}}{2}\right)^n$ in [11]. Two structures are called *neighbours* in the combinatorial structure space if they differ in a single base pair, i.e. two residues that form a pair in one sequence, but not in the other.

The *biophysical structure space* of x is determined by a certain energy model, given by energy parameters associated with base pairing, base pair

stacking, and loop formation. The elements of the biophysical structure space are those which have minimal free energy with respect to all their neighbours. Thus they are local energy minima in the combinatorial structure space, for short lmfe-structures. A structure attaining the global energy minimum is called an mfe-structure. If an RNA folding program is asked to calculate "the" structure, one such mfe-structure (of possibly many) is returned.

We call an lmfe-structure *prominent to the degree* k, if any other lmfe-structure is at least k steps apart, where a step is defined as a transition to a neighbour structure. Clearly, the biophysical structure space is much smaller than the combinatorial one, but to which extent is currently not known.

The *biological structure space* finally is defined to comprise the biologically relevant (functional) structures of x, typically just one structure, or two in the case of a simple conformational switch. (We do not consider switches with more than two states here.) The crux of structure prediction is that the biological structure space is not a subset of the biophysical space under the currently available models. It is also determined by tertiary interactions of the RNA molecule, and by interactions with proteins. Still, the biophysical model can be used to give good approximations; in particular, when these interactions are known, it is possible (and sensible) to fix the residues involved and apply energy minimisation only to the others.

Now let us turn to the phenomenon of structural switching. With respect to the combinatorial structure space, a simple (but maybe surprising) observation has been reported by Grüner [12]: Given two arbitrary structures s_1 and s_2 of equal number of residues, it is always possible to design a sequence x such that both s_1 and s_2 are in the combinatorial structure space of x.

We conclude that the combinatorial structure space is too abstract to provide hints towards potential switches. We need characteristics of the biophysical structure space of conformational switches that can be observed by algorithmic methods. paRNAss is based on the following hypotheses:

1. *The two alternative functional structures of a conformational switch are close to different lmfe-structures, and relatively close to the global energy minimum.* In a case where this hypothesis does not hold, the energy model is not applicable, and our approach will reveal nothing.

2. *These two lmfe-structures are prominent structures of a significant degree, and within a certain energy threshold, there are no other prominent structures.* The justification of this hypothesis is that a switch must have two clearly distinct states, and a molecule in transition must not get caught in other local energy minima.

3. *The two lmfe-structures may reside on different energy levels, but a certain energy barrier must be overcome when switching in either direction.* This hypothesis reflects that the switching should not be spontaneous, but must be triggered by some outside event.

Our approach investigates the biophysical structure space of the target sequence. If it clearly exhibits the two structures as postulated in Hypothesis 2, these will be suggested as the two conformations of a structural switch according to Hypothesis 1.

2.2 A Run through a paRNAss Experiment

In the simplest case, a successful paRNAss experiment takes five steps:

Step 1: Sampling the structure space Using an RNA folding program, we draw a sample set $S = \{s_1, ..., s_p\}$ from the combinatorial structure space of our target RNA. Since current folding programs cannot determine the true biophysical structure space, we permit that there may be some structures in the sample that are not local free-energy minima. If Hypothesis 2 holds, the sample should contain two "families" of structures, since all structures in the sample should be close to either the first or the second of the two prominent structures (which themselves may or may not be contained in the sample).

Step 2: Pairwise distance calculation For all $s_i, s_j \in S$, we calculate their pairwise distance $d_\delta(s_i, s_j)$. We do so for at least two different metrics δ_1, δ_2 on the structure space. We plot the results in a δ_1, δ_2 coordinate system.

If both elements in a pair are from the same structure family, their distance should be small. Conversely, if both are from different families, their distance should be roughly equal to the distance of the two prominent structures. Thus, if Hypothesis 2 holds, the plotted distance diagram should exhibit two clusters of points, one in the lower left, one in the upper right. See Figure 1.

Note that such clusters may also occur by chance. Also, the albeit unlikely case of three equidistant prominent structures would result in a similar plot. Hence, further steps are necessary.

Step 3: Clustering We use a standard clustering algorithm to split S into two disjoint clusters C_1 and C_2, based on the pairwise distances under either δ_1 or δ_2.

Step 4: Consensus structure calculation For each cluster C_i, a consensus structure c_i is derived by first taking all the base pairs present in the majority of the members of C_i, and then reapplying the folding algorithm with these base pairs fixed. Note that the consensi need not be contained in S. Figure 2 shows the consensus structures derived for our example. As the clustering program will always return two clusters, another step is necessary to

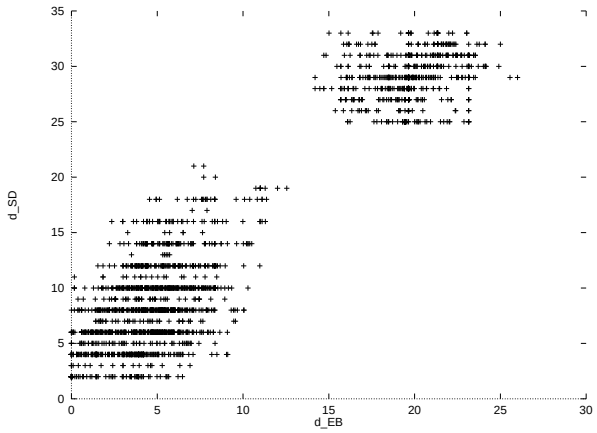

Figure 1: Distance Plot showing a clear separation. Average energy barrier d_EB between the two structure families is 20 kcal/mol; string distance d_SD is about 10 edit operations.

safeguard against the case of several equidistant prominent structures or other artefacts.

Step 5: Consensus structure validation To take care of potential pseudoknots, the pseudoknot distance of c_1 and c_2 is calculated (see section 3.4). Furthermore, for all $s_i \in S$ we calculate the distances $d_\delta(s_i, c_1)$ and $d_\delta(s_i, c_2)$. We plot these distance pairs as points in a coordinate system. If Hypothesis 2 holds, the c_1 family of sample structures will show up as a cloud of points near the x-axis, the other family near the y-axis. See Figure 3.

If the outcome of steps $1 - 5$ is as described above, we say that paRNAss predicts the possibility of conformational switching between structures c_1 and c_2. We then use the tool RNA Movies[13] to visualize an energetically favourable transition path from c_1 to c_2.

2.3 Applicability of the paRNAss Approach

There is no intrinsic obstacle to further automate the paRNAss approach. But at present, human interaction is essential. paRNAss takes great care to produce visualizations of all its intermediate results. These convey various hints to the expert, much more than can be discussed here. paRNAss should be applied in a context where there is some indication for the presence of a conformational switch, i.e. knowledge about autocatalytic behaviour, inconsistent methylation data, or different functions of closely related RNA molecules that cannot be explained by sequence variation. Ultimately, the suggested confor-

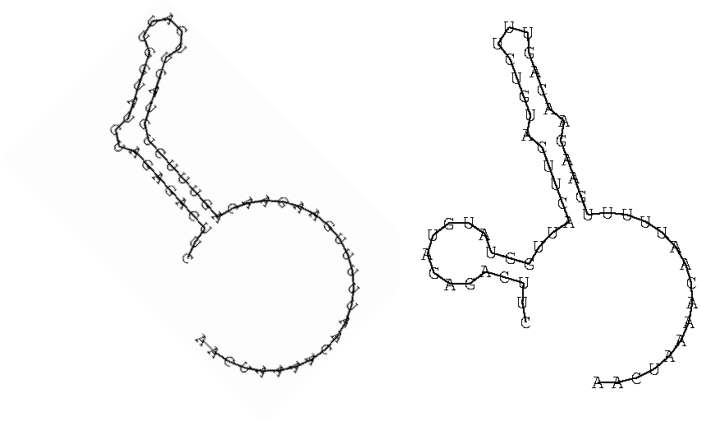

Figure 2: Predicted structures for *L. collosoma*

mations should be verified experimentally.

3 Algorithmic Methods

This section describes the algorithms used in steps 1 - 5, explains their Parameters and discusses problems of interpreting the results.

3.1 Generating the Structure Sample

We use MFOLD or RNAfold to enumerate (sub)optimal structures within an energy threshold of the mfe value. This gives rise to three parameters: the folding temperature T, the suboptimality threshold P (in percent of the mfe-value), a bound N on the number of structures. If more than N structures are generated under the given settings of T and P, N of them are randomly chosen as the sample set for the subsequent steps.

The number of structures generated increases with P and decreases for increasing T. Note that the bound N is applied after structure generation, so it bounds the computational effort only for the subsequent steps. Typical (and default) values are T=37, P=15, N=50.

3.2 Metrics for Pairwise Structure Comparison

In contrast to pairwise sequence comparison, there is no generally accepted model for comparing structures. paRNAss provides three alternative approaches.

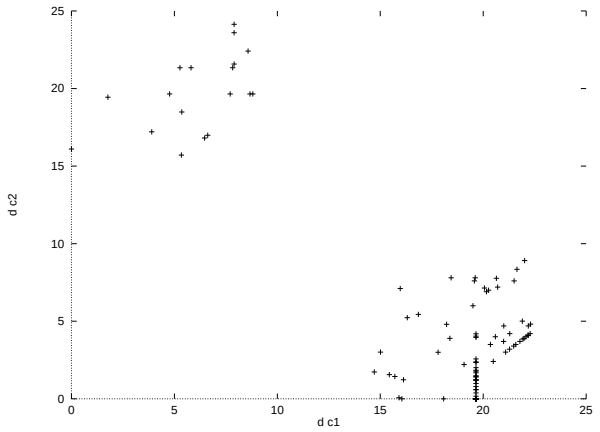

Figure 3: The validation plot shows the distances of all sample structures from the two consensus structures c_1 and c_2. This particular plot is based on the energy barrier distance.

The *morphological distance* d_{MD} is a slightly modified version of a formula suggested by Zuker [14]. Here structures are represented as sets of base pairs. $(i,j) \in s$ means that residues i and j form a base pair in s. For two sequences s_1, s_2 we define

$$d_{MD}(s_1,s_2) = \max\{d'_{MD}(s_1,s_2), d'_{MD}(s_2,s_1)\}, \text{ where} \quad (1)$$

$$d'_{MD}(s_1,s_2) = \sum_{(i_1,j_1)\in s_1} \min_{(i_2,j_2)\in s_2} \max\{|i_1-i_2|, |j_1-j_2|\} \quad (2)$$

d_{MD} is strictly positive and symmetric, but does not satisfy the triangle inequality. Although it is not a metric in the mathematical sense, it behaves quite reasonably as a distance measure.

The *string edit distance* d_{SD} of two structures employs their string representation with dots and parantheses, e. g. (((.(...)))...), as used with the Vienna RNA package. We define

$$d_{SD}(s_1,s_2) = d_w(y_1,y_2) \quad (3)$$

where y_i is the string representation of s_i, and d_w is an edit distance (i.e. the score of an optimal alignment) on strings. Being defined via the edit distance model, d_{SD} is a metric. This distance measure is provided with the Vienna RNA package.

The *energy barrier distance* d_{EB} is designed to capture the minimal amount of energy necessary for the molecule to switch between two structures. A transition path from s_1 to s_2 is given by a sequence of intermediate structures

(where each is a neighbour of its predecessor according to 2.1). Let $e(s)$ denote the free energy of s.

$$
\begin{aligned}
d_{EB}(s_1, s_2) &= \min\{d'_{EB}(s_1, s_2), d'_{EB}(s_2, s_1)\}, \text{ where} && (4)\\
d'_{EB}(s_1, s_2) &= \min\{e(p) | p \text{ is transition path from } s_1 \text{ to } s_2\} && (5)\\
e(p) &= \max\{e(s) - e(s_1) | s \text{ is intermediate structure in } p\} && (6)
\end{aligned}
$$

The energy barrier distance uses a discrete model of transition and only approximates reality, since an RNA molecule need not go through well defined intermediate structures. Furthermore, as the number of paths is excessively large, our implementation uses a greedy heuristic to approximate d_{EB}.

d_{EB} satisfies the axioms of a metric. In the plots we show the two values $d'_{EB}(s_1, s_2)$ and $d'_{EB}(s_2, s_1)$ instead of $d_{EB}(s_1, s_2)$. This gives some extra information about the different energy levels of s_1 and s_2, but does not affect the general appearance of the distance plot.

Sometimes a distance plot is hard to interpret, as it shows a rather weak separation, possibly only in one dimension. In such a case, two further experiments should be done. The first is to relax the parameters to include some more structures in the sample. This may make the signal go away or come out more clearly. If a weak signal persists, this may be an indication of a possible switch which involves only a small part of the overall structure, while the rest remains stable. A relatively low energy barrier between all structures in the example is also a hint in this direction. In that case the sequence should be cut into shorter parts which are then analysed separately.

3.3 Clustering and Structure Prediction

The clustering step [15] takes two parameters: D names the distance measure (d_{MD}, d_{SD}, d_{EB}) upon which the clustering is to be based. Clusters for different distance measures should be obtained and compared — they often are consistent even in the case where one of the two measures yields a poor separation in the distance plot in step 2. C specifies the number of clusters to be generated. Normally, C = 2. Other values can be used when the clustering appears to be artificial.

For each cluster, the consensus is derived using RNAfold as explained in section 2.2. Again, parameter T indicates the folding temperature.

These two steps generate graphic outputs for the dendrograms and for squiggle plots of the predicted structures, as well as a string representation of structures in the Vienna style.

3.4 Structure Validation

Structure validation takes the string representations of the predicted structures. It uses d_{EB} or d_{SD} to calculate the distances of each sample structure to each of the two predicted structures c_1 and c_2. These are plotted in a c_1, c_2 coordinate system. In the positive case, the two structure families show up as clouds of entries near the d_1 and far from the d_2 axis, and vice versa. This proves that all structures in the sample are actually close to one of the predicted structures (Hypothesis 2), while the structures themselves are sufficiently different (Hypothesis 3).

Let $s_1 \cup s_2$ denote the union of two structures (i.e. base pair sets). Define

$$\begin{aligned} pkDist(s_1, s_2) &= -1, \text{ if } s_1 \cup s_2 \text{ is planar,} && (7)\\ &= 0, \text{ if } s_1 \cup s_2 \text{ contains a pseudoknot,} && (8)\\ &= k, \text{ if } k \text{ is the number of bases with conflicting pairings in } s_1 \cup s_2. && (9) \end{aligned}$$

paRNAss reports $pkDist(c_1, c_2)$ in addition to the above visualization.

3.5 Visualization of Transitions

The calculation of $d_{EB}(c_1, c_2)$ determines an energetically favourable model of the transition from c_1 to c_2. The sequence of intermediate conformations is passed to the RNA Movies visualization tool [13]. The tool offers the functionality of a video player, presenting an animated graphics representation of the transition. This serves as an additional means to check the plausibility of the suggested switch by human expertise.

4 Applications

Space only allows a cursory discussion of results here. A summary of the results obtained in [16] has been made available on the WWW via the paRNAss URL [1]. These include a case of a switching mRNA as well as a case where a pseudoknot is involved.

4.1 Switches

The spliced leader RNA of *Leptomonas collosoma* is part of the RNA section that is added to each mRNA of this species in a process called trans splicing. Two separately transcribed RNA molecules are linked together in a way similar to the connection of neighbouring exons. The structural transition of this

sequence is analysed by LeCuyer [6]. paRNAss clearly predicts the switching structures shown in Figure 2. These are in good correspondence with the published structures.

A well known mechanism in which alternative RNA structures are essential is the regulation of gene expression by attenuation. For this process a leader sequence called attenuator is required upstream of the coding region of an RNA. This leader can be translated to a short peptide. Depending on whether the leader peptide can be built completely or not, the regulated region will fold into different secondary structures. Full translation of the attenuator sequence leads to formation of a terminator structure, which prevents further transcription of the concerned DNA section. If the leader peptide stays unfinished, an anti-terminator is formed and transcription can continue.

As one example of attenuation we examined the leader sequence of the pheS-pheT operon of *E. coli*. The secondary structures of this RNA are analysed in depth in [2]. Application of paRNAss on the (shortened) leader sequence gives a strong hint on the ability to switch. The prediction phase proposes two foldings which have considerable similarity to those published by Fayat [2]. Especially the terminator is almost identical. Finally, the validation plot supports these predictions.

4.2 Non-Switches

We evaluated paRNAss on several mRNAs as well on 20 random sequences generated by ROSE [17]. In general, the distance plots for mRNA are comparable to plots produced for random sequences. No switching is indicated in these plots. As a typical example, we include the distance plot for an mRNA of *Zea mais* in the online documentation.

4.3 Method Reliability

While evaluating paRNAss on a suite of about 40 test sequences, we ran into one false negative and three false positives. The false negative was a sequence from a virusoid, where a known case of a switch was not detected by paRNAss. This was easily explained by Hypothesis 1 – the experimentally determined structure is far from the energy minimum and is not detected by the RNA folding program.

The (possibly) false positives are more interesting: The examination of a coding sequence from *Schistosoma mansoni* led to an ambiguous plot which made us explore a shortened version of the same string. A switch was clearly indicated. The validation phase supported the proposed structures (see [1]). This indicates the possibility that the *S. mansoni* gene carries some function

encoded in structure. While such is known for some viral genes, we are not aware of any biological evidence in this respect with *S. mansoni*. The same applies for the finding of a possible switch in a 5s rRNA from *Neurospora crassa*. We also ran into one case of a random sequence where a switch was strongly indicated. This observation may indicate that switches are not as exceptional as assumed.

5 The Software

paRNAss was implemented by two successive Master's Theses: Rehmsmeier [18] provided a prototype proving the viability of the approach, while Haase[16] completed the tool and worked out the applications. D. Evers integrated the RNA Movies tool, and A. Sczyrba built the WWW interface, which was integrated into BiBiServ, the Bielefeld Bioinformatics Server [1].

paRNAss is computationally expensive; the cost mostly comes from folding the sample set of structures and from evaluating d_{EB}. Folding algorithms have a computational complexity of $O(n^3)$, where n is the sequence length. While the heuristics for calculating a single energy barrier is in $O(n^2)$, the overall effort for this phase depends on the sample size p and amounts to $O(p^2 \cdot n^2)$. So both sample size and sequence length must be seen as limiting factors. Typically you can draw a sample of 50 structures of about 150 residues and process it in about 10 minutes of real time on an UltraSparc 1. However, we see possibilities for significant speed-up (cf. below).

6 Conclusion and Future Work

Our immediate goal is to apply paRNAss in situations where conformational switching is suspected, but has not been proved yet. Once more experience has been gained with paRNAss experiments, we may consider to further automate the procedure.

One current limitation of the overall approach is computation time. It should be possible to cut down the complexity of the structure comparison phase. As both RNA folding and the calculation of energy barriers are based on the same physical model, merging the two phases might lead to considerable speedup. This is not trivial, as it requires a redesign of the folding program.

A second limitation today is that paRNAss generates only a rather weak signal in the case where only a small part of the RNA molecule actually changes shape (See the case of *S. mansoni* described above). To better detect such cases, pairs of structures must be analysed with respect to both global and

local similarity and dissimilarity. Methods used in sequence comparison can be generalized to this mode of structure comparison.

7 Acknowledgement

Gerhard Steger contributed numerous hints on known structural switches. Dirk Evers gave valuable advice at all stages of the development of the paRNAss approach.

1. BiBiServ. http://BiBiServ.TechFak.Uni-Bielefeld.DE/parnass/.
2. G. Fayat, J. F. Mayaux, C. Sacerdot, M. Fromant, M. Springer, M. Grunberg-Manago, and S. Blanquet. *Journal of Molecular Biology*, 171:239–261, 1983.
3. C. Philippe, L. Bénard, C. Portier, E. Westhof, B. Ehresmann, and C. Ehresmann. *Nucleic Acids Research*, 23(1):18–28, 1995.
4. David W. Lazinski and John M. Taylor. *RNA*, 1:225–233, 1995.
5. Ira G. Wool, Anton Glück, and Yaeta Endo. *Trends In Biochemical Sciences*, 17:266–269, 1992.
6. Karen A. LeCuyer and Donald M. Crothers. *Proc. Natl. Acad. Sci. USA*, 91:3373–3377, 1994.
7. Hiten D. Madhani and Christine Guthrie. *Cell*, 71:803–817, 1992.
8. Marcel Turcotte, Guy Lapalme, and François Major. *Journal of Functional Programming*, 5(3):443–469, 1995.
9. M. Zuker. In M. S. Waterman, editor, *Mathematical Methods for DNA Sequences*, chapter 7, pages 159–184. CRC Press, Boca Raton, FL, USA, 1989.
10. I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. *Chemical Monthly*, 125:167–188, 1994.
11. M. S. Waterman. Chapman & Hall, London, UK, 1995.
12. Walter Grüner. PhD thesis, University Vienna, 1994.
13. Dirk Evers and Robert Giegerich. 1999. To appear.
14. Michael Zuker. *Science*, 244:48–52, 1989.
15. G. W. Milligan. In P. Arabie, L. J. Hubert, and G. DeSoete, editors, *Clustering And Classification*, pages 341–375. World Scientific Publ., 1996.
16. Dirk Haase. Diploma thesis, Universität Bielefeld, Technische Fakultät, 1997.
17. Jens Stoye, Dirk Evers, and Folker Meyer. 1998.
18. Marc Rehmsmeier. Diploma thesis, Universität Bielefeld, Technische Fakultät, 1996.