ALGORITHMS FOR THE PREDICTION OF 3D RNA MOLECULES

DISSERTATION

zur Erlangung des
akademischen Grades

Doktor rerum naturalium
an der Formal- und Naturwissenschaftlichen
Fakultät der Universität Wien

vorgelegt von

Mag. rer. nat. Alexander RENNER

Wien, im Februar 1998

This work was carried out in the time from April 1995 to February 1998 at the Institute of Theoretical Chemistry of the University of Vienna.
First of all I want to mention Peter Schuster for giving me the chance to join his group and his tremendous support. It always has been a pleasure for me to discuss with him.

Ivo Hofacker the great hacker :⊃ assisted with his computational knowledge and debugged with me endless times.
Peter Stadler supported me with all his tremendous knowledge and wisdom.
Herbert Kratky was of great help during the work and supported me in computational as well in theoretical work. A lot of the data here achieved was a collaboration with him. Christian Haslinger and Günther Weberndorfer assisted by the construction of the genetic algorithm.

Now to more personal wishes:
Thank you Dali for your patience in this last year of my study. I wish to thank my parents for their support and appreciation during my whole study.

Last but not least all the other members of the group: Ronke Babajide, Jan Cupal, Martin Fekete, Walter Fontana, Thomas Griesmacher, Judith Jakubetz, Stephan Kopp, Bärbel Krakhofer, Stefan Müller, Susanne Rauscher, Norbert Tschulenk (our figure-head) and Stefan Wuchty created a superb working (and not only working) atmosphere.

THX Xtof

Contents

Abstract

Investigations of biopolymers by X-ray crystallography or NMR spectroscopy are often time consuming or unfeasable at the current state of the art, as they face considerable technical problems. NMR structure analysis of biopolymers is almost always assisted by computer predictions. A computational approach that reliably extends the experimental data would be very helpful for this task. In this thesis a force field based method for the conformational analysis of RNA structures is presented. Conformational search and energy minimization are combined with an energetic evaluation of solvent solute interaction. The "vacuum energy" of a given confomer contains solvent effects and is defined by the sum of bond and torsion angle deformation energies, the pairwise additive Lennard-Jones and electrostatic contributions. This energy has been calculated using the AMBER and FLEX force fields. The transfer of the solute from vacuum to water results in a polarization of the dielectric medium which interacts with the charges of the solute and is the so-called reaction field (RF) potential which contributes to the total energy of the system. The RF energy has been calculated by the Field Integrated Electro-STatic Approach (FIESTA).
A new technic, based on a Genetic Algorithm (GA), to obtain promising confomers for further optimization, has been developed. It allows to optimize loop-stem structures. It has been successfully used to find "fitter", i.e. lower energy structures for triloops compared to other optimization technics.
A method, based on conformational search, has been applied to four-membered RNA hairpin loops with GNNA sequences, and the Mouse Mammary Tumor Virus (MMTV) pseudoknot. The results obtained show, that structures derived using this new technic conform with spectroscopically predicted structures. The method appears to be a valuable predictive tool for RNA structural motifs.

Zusammenfassung

Röntgenkristallographische oder NMR spektroskopische Untersuchungen von Biopolymeren sind oft zeitintensiv oder beim momentanen Wissensstand überhaupt nicht durchführbar. Für die Strukturaufklärung mittels NMR sind computerunterstützte Methoden unerläßlich. Eine rechnerische Methode, die verläßlich die experimentellen Daten erweitert, wäre extrem hilfreich um dieses Problem zu lösen. Im Rahmen dieser Dissertation, wurde eine auf Kraftfeldern basierende Methode zur Konformationssuche von RNA Strukturen entwickelt. Die Konformationssuche und Energieminimierung wird mit der energetischen Evaluation der Wechselwirkung zwischen Lösungsmittel und Festkörper kombiniert. Die Energie eines gegebenen Konfomers im "Vakuum" enthält Lösungsmitteleffekte und ist sowohl über die Summe von Bindungs- und Torsionswinkeldeformationen, als auch durch paarweise additive Lennard-Jones und elektrostatische Beiträgen, definiert. Diese Energie wurde mit Hilfe der AMBER und FLEX Kraftfelder berechnet. Der Transfer des Festkörpers vom Vakuum in Wasser führt zu einer Polarisation des dielektrischen Mediums, welches mit den Ladungen des Festkörpers interagiert. Dies ist das sogenannte *reaction field* (RF) Potential und trägt zur Gesamtenergie des Systems bei. Die RF Energie ist mittel FIESTA (Field Integrated ElectroSTatic Approach) errechnet worden.
Eine neue Technick, basierend auf einen Genetischen Algorithmus (GA), wurde entwickelt, um vielversprechende Konfomere für weitere Optimierungen, zu erhalten. Mittels dieser Methode ist es möglich *loop-stem* Strukturen zu minimieren. Sie ist erfolgreich an Triloops angewandt worden.
RNA *hairpin loops* mit GNNA Sequenzen und der Pseudoknoten des Mouse Mammary Tumor Virus (MMTV) sind mittels der obgenannten Konformationssuche betrachtet worden. Die auf diese Weise erhaltenen Strukturen zeigen eine gute Übereinstimmung mit den spektroskopischen Daten. Diese Methode scheint ein wertvolles Werkzeug für die Strukturvorhersage von RNA zu sein.

1 Introduction

Understanding the properties and functions of biopolymers is a core issue of biophysics. Biopolymer structures are considered useful intermediates in the prediction of biomolecular functions. The analysis of relations between sequences and structures is difficult but not impossible at the current state of knowledge. Both, RNA molecules and proteins are linear polymers of defined sequences, folding back on themselves to form a lattice of specific interactions between residues. The ways however, how proteins and RNA achieve their compact conformations are rather different. While protein secondary structures are relatively instable on their own and formation of a hydrophobic core provides the driving force of folding, RNA secondary structures are relatively stable, due to strong stacking interactions between basepairs, even as isolated fragments. Close packing of double helices in order to form compact RNA cores requires compensation of the repulsion between negatively charged phosphates in the backbone by means of metal cations. Purine and pyrimidine bases aggregate in planar complexes that have geometries determined by hydrogen bonds. The original set consisting of the Watson-Crick basepair (bp) (G≡C and A=U) was soon complemented by G-U "wobble" pairs, which are also admissible in RNA double helices. Recently other motifs of non-Watson-Crick bp have been detected in RNA structures and new stable conformations are being steadily added to list. The non-Watson-Crick bp deform double helices and thus appear outside regular structures. Examples are U-U and other bp in internal loops [6, 88, 161] as well as A-A, G-A, or G-G closing pairs at the ends of double helical regions or in multiloops [17, 60, 115]. Because of these additional strong interactions between bases internal loops and multiloops seem to be much less flexible than previously thought, predicted by conventional methods. Structures derived by x-ray crystallography of tRNA-molecules and hammerhead ribozymes have revealed that individual double helices may stack coaxially upon each other by forming extended double-helical stretches [115, 139]. Interactions between double helices is mediated by specific motifs, for example, by hairpin loops forming pseudoknots in the neighborhood of stacks. Structures of single stranded nucleic acids call for an intermediate step in structure prediction. Secondary structures (being listings of Watson-Crick and G-U base pairs that can be drawn in two dimensions without knots or pseudoknots) are as much as ever indispensable for the predictions of three-dimensional (3D) structures. Highly efficient

algorithms are available for secondary structure predictions, in particular for minimum free energy (mfe) structures, suboptimal foldings, or partition functions. Comparative sequence analysis leads to "phylogenetic structures" derived from cross-correlations in base substitutions [34, 49, 74]. Remarkable and not yet fully understood differences between mfe and phylogenetic structures are observed.
Modeling and prediction of 3D structures is still kind of an art requiring extensive input of spectroscopic, chemical probing and biochemical degradation data. Following the elimination of methodological artifacts MD simulations of RNA structures are now becoming useful tools for the analysis of structures. But, unfortunately, 3D structural information obtained for RNA is at this state of the art quite limited. X-ray technics were successful, only for tRNAs and small fragments, as well as recently for the hammerhead ribozyme [115, 139] and its group I domain [17]. Small secondary structure motifs such as bulges, internal loops, pseudoknots and hairpin loops have been investigated by NMR spectroscopy [31, 104, 145]. Computational methods refine X-ray chrystallography data, and are of atmost importance for distance-geometry in NMR spectroscopy. Until now the maximum number of nucleotides which are manageable are in the range of 40. Within this work new algorithms to predict the 3D structure for RNA are going to be presented.
The relationships between sequences and functions of biopolymers are not only of current interest in structural biology, but they are also of primary importance for recent developments in biotechnology [137]. The success of rational design and of the planning of efficient search strategies in evolutionary methods depend crucially on the state of the art in understanding RNA structure and function. The interplay between sequence and structure conservation in evolution, subtle and hard to decipher as it may be, becomes an issue of increasing importance to which structural biology is expected to contribute. Phylogenetic structures allow to detect constraints in the sequences of RNA molecules that fulfill multiple functions in nature and thus provide hints on the role of structure in hitherto unknown tasks[62, 96].

2 Biopolymer Structures

In the following chapter a short overview of nucleic acid structure will be given. In addition recent findings in the 3D structure as well as the target structures used in this study will be presented.

2.1 Short Description of Nucleic Acid Structure

Nucleotides are composed out of three molecular fragments:

- pentose
  The pentose is of furanoside-type (β-D-ribose in RNA or β-D-2′-deoxyribose in DNA), and it is phosphorylated in 5′ position and substituted at C1′ by one of the four different heterocycles attached by a β-glycosyl C1′-N linkage. Because of the additional OH at C2' RNA is thermodynamically less stable than DNA.

- heterocyclic bases
  The heterocycles are the purine bases adenine (**A**) and guanine (**G**) and the pyrimidine bases cytosine (**C**) and uracil (**U**, uracil is replaced in DNA by the functionally equivalent thymine-5-methyluracil).

- phosphate
  The phosphates are linking monomers.

Figure 1 shows a short strand of RNA containing the four usual bases adenine (A), guanine (G), cytosine (C) and uracil (U). All four monomers are connected to a single strand, which is directional and starts at the 5′-end (top left of figure 1) and ends at the 3′-end (bottom right of figure 1). Besides these four bases there exists numerous naturally occuring, chemically synthesized and modified nucleotides. Many of these have antibiotic activity, among them the important class of arabinosides, nucleosides with β-D-arabinose instead of β-D-ribose.

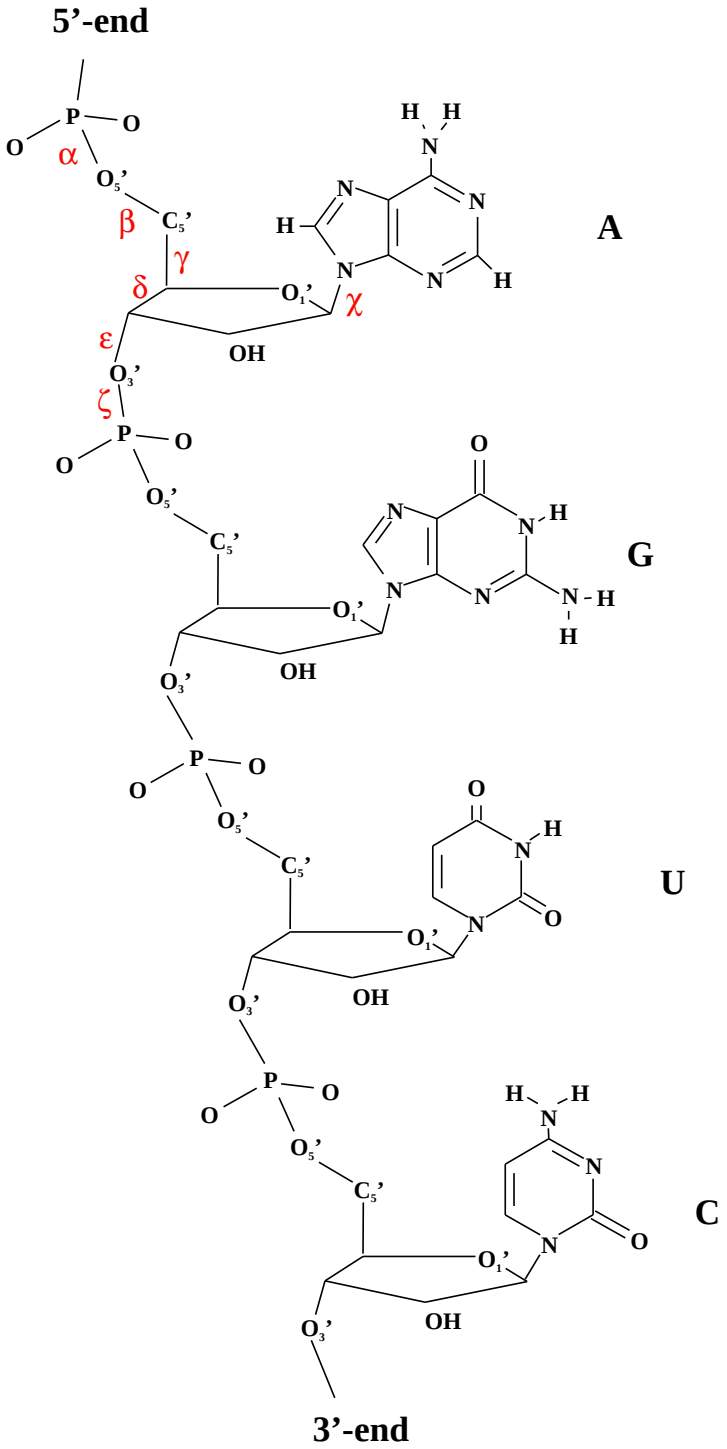


Figure 1: Atomic sample structure of RNA.

A nucleotide sequence, simply gives the order of the nucleotides starting at the 5′-end and ending at the 3′-end. The prediction of RNA structures can be regarded as a two-step process: We first proceed from sequence to secondary structure, then we attempt to construct the three-dimensional geometry of RNA molecules. The first step, the prediction of an RNA secondary structure from a sequence can be solved efficiently using the Vienna RNA Package which has been developed in our group [58]. The so-called secondary structure shows, which bases are paired or unpaired to others. A large variety of base pairs occur in RNAs, starting with the Watson-Crick-types G≡C and A-U in different geometries to G-U pairs and even more uncommon types like G-A, G-G, or A-C^+. RNA secondary structures can be classified in very few types of structural motifs (see figure 2). The most abundant of these motifs are the so-called hairpins consisting of a double-stranded part (the "stem") and a connecting single-stranded part (the loop). Other motifs are the bulge (unpaired bases on one side of the stem), the internal loop (unpaired bases on both sides of the stem), or the multi-loop (several stems connected by short unpaired regions). Unpaired regions at the end of a strand are called "dangling ends".

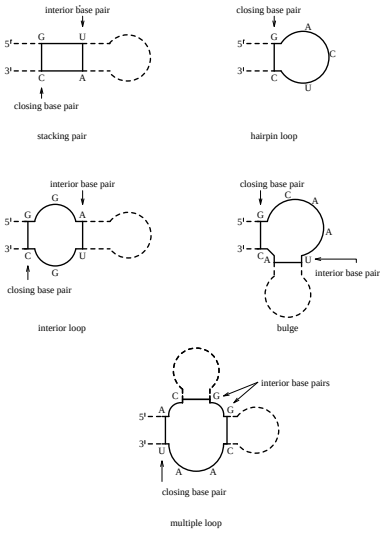


Figure 2: Secondary structure motifs in RNA.

Secondary structure prediction programs usually predict a single mfe structure. However often several suboptimal structures exist withim a few kT of the ground state [176]. Therefor we start from several structural proposals for the next step: the definition of a 3D structure. The whole process can be seen in figure 3. From left to right we start with the sequence of a tRNA, obtain the secondary structure and finally yield the three-dimensional structure.

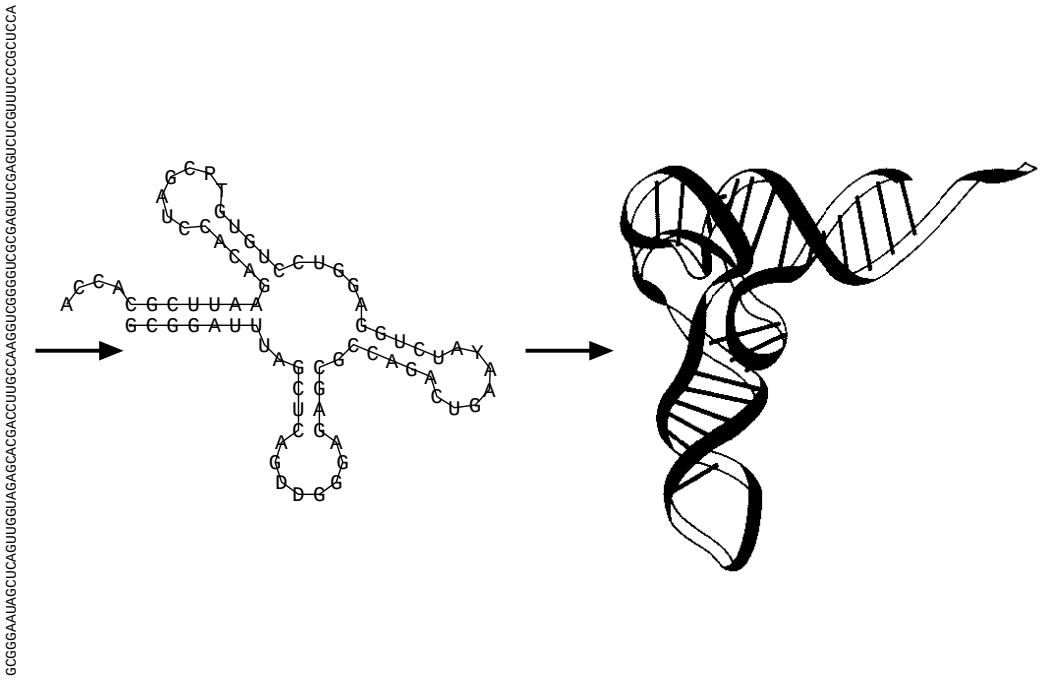


Figure 3: From sequence to structure.

In essence the 3D structure shows the relative position of the secondary structure elements with respect to each other. At the highest resolution the position of each atom is known. In addition several interactions can only be seen if the 3D structure is considered. The most prominent examples are pseudoknots [121, 167], base triples [19, 20, 39, 43, 72, 98], G-quartets [5, 21, 107, 141], Helix loop interactions and helix - helix interactions. The range of possible secondary and tertiary structural elements is rather large,

proving that RNAs are very flexible molecules. Since both the pentoses and - even more - the heterocyclic bases are very rigid structures, most of the conformational flexibility comes from the backbone. In figure 1 seven torsional angles are designated by greek letters. Six of them are along the backbone, and coming from the 5′-end of the molecule their definition is as follows:

- α O3′-P-O5′-C5′
- β P-O5′-C5′-C4′
- γ O5′-C5′-C4′-C3′
- δ C5′-C4′-C3′-O3′
- ϵ C4′-C3′-O3′-P
- ζ C3′-O3′-P-O5′

The last torsional angle of major importance to the 3D structure is angle χ (O1′-C1′-N-C4 in purines and O1′-C1′-N-C2 in pyrimidines). It can be used as a very good assumption that these seven internal degrees of freedom per monomer unit define the whole conformational space of an RNA molecule. Two angles are of special interest as they usually assume only very specific values.
δ: This torsional angle lies within the sugar ring system and is therefore restricted by a ring closure criterion. Since a five-membered ring has no flat geometry, one or two of the atoms are lying above or below the plain defined by the other four or three atoms. If the atom is on the same side of the plain as the C5′ the conformation is called *endo*, if it is on the opposite side it is called *exo*. This behavior is also called "sugar-puckering". Figure 4 shows two of the most frequent sugar-puckers in RNA, C2′-endo (left-hand-side of figure 4) and C3′-endo (right-hand-side of figure 4). Nucleotides in the standard A-RNA-helix are of C3′-endo conformation, C2′-endo conformations occur mostly in small loops, because of their tendency to elongate the backbone. C3′-endo-sugars are also referred to as sugars of N-type, whereas C2′-endo-sugars are of S-type. Apart from the two major types other conformations occur mainly in loop regions. χ: Though this torsional angle is not involved in a ring system, its values are nevertheless restricted to two distinct regions, one around 0 degrees and the other around 180 degrees. This angle determines the

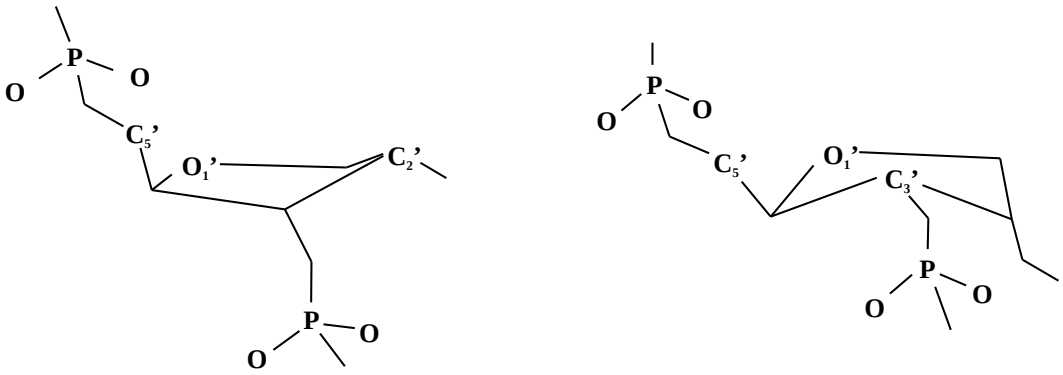


Figure 4: Major puckering modes of sugars in RNA (left-hand-side: C2′-endo, right-hand-side: C3′-endo).

position of the heterocycle with respect to the sugar ring. If the heterocycle is rotated towards the C5′-atom (the torsional angle being 0 degrees) the conformation is called *syn*, if the heterocycle is in the opposite position (away from the C5′-atom, the torsional angle being 180 degrees) the conformation is called *anti*. In standard A-RNA-helixes all bases are of *anti*-conformations, *syn*-conformations can be found in loop regions and in some non-Watson-Crick base pairs. All other torsional angles possess also certain preferred ranges, that are not so well defined as in cases shown above. A comprehensive introduction to nucleic acid structure can be found in Saenger's book [133].

2.2 Recent Findings in 3D Structure

The most recent progress in understanding RNA spatial structure came from high-resolution crystallography of one of the two structural domains of the catalytic core of a group I intron [17]. Every crystal structure of an RNA solved has yielded a surprise, and this one is no exception: "adenosine platforms" consisting of two unpaired As stacking upon the end of two helices were discovered. They seem to mediate different types of long range interactions, for example by providing binding sites for GNRA tetraloops (where N represents A,C,G,U, and R represents A or G). In addition, the compact domain of the group I intron shows remarkably close packing of two helices in

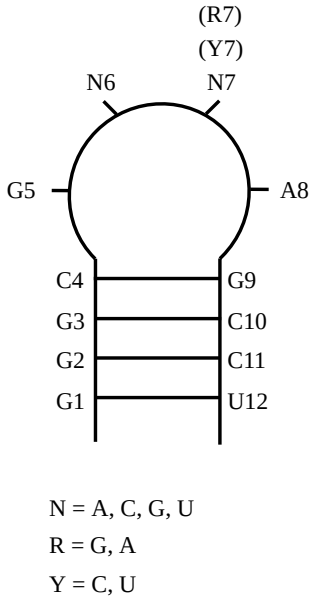


Figure 5: An example of a hairpin.

which Mg^{2+} cations as well as 2'-HO groups of ribose stabilize the negative charges of the backbone phosphates. It is worth mentioning that the early structural model of a group I intron reflects the essential features of the now known structure [99] very well. Adenosine platforms add to the variety of tertiary interactions we mentioned in the introduction: pseudoknots, other non-Watson-Crick bp such as U-U [6, 88, 161] in internal loops, purine-purine base pairs preferentially at the ends of double helical stacks [17, 60, 115], base triplets and other classifiable motifs. Defined tertiary interactions such as the ones we have mentioned can, in principle, be incorporated in structure prediction algorithms. At present the pace of the discovery of new structural motifs, is so fast that it does not allow to accumulate sufficient empirical data. Another example of the successful modeling of a catalytic RNA molecule is RNase P in which protection data from chemical probing were used as constraints for model building [168]. The new data on the structures of catalytic RNAs have provided new insights but in essence, have confirmed the ideas

on the mechanisms of RNA catalysis.
Despite the structural data emerging from crystallographic analysis firm information on RNA structures is still rare and hence all sources, experimental as well as computational, have to be exploited in order to make reliable predictions on molecules for which no crystallographic data are available yet. One approach to molecular modeling of RNA structures is using MC-SYM [156], it is based on the symbolic creation of coarse structures that are refined by means of energy minimization and/or molecular dynamics (MD) calculations. MC-SYM has been successfully applied to derive structural models for the Rev-binding element of HIV-1 from the structures of aptamers, small artificial RNA ligands [85]. A strategy for modeling docking of peptides onto RNA has been developed and applied to complexes formed between peptides from retroviruses and the corresponding RNA counterparts [152]. The modeled structure is in general agreement with an NMR study of the HIV-1 Rev peptide-Rev binding element RNA [8].
The power of modeling 3D structures using experimental constraints became evident when a structural model of the hammerhead ribozyme based on distance data derived from fluorescence resonance energy transfer [157] was found to be in good agreement with simultaneously published x-ray data [115]. Recent modeling studies are dealing with the structural specialties of mRNAs that lead to selenocysteine incorporation in ribosomal protein synthesis [61, 160]. Representative for other investigations using modeling and chemical as well as enzymatic probing is a study on a tRNA-like domain in tobacco mosaic virus RNA [32].
The prediction of RNA 3D structure based on computation of minimal potential energies faces a formidable problem because of the enormously large numbers of local optima. A method based on conformational searches using a genetic algorithm followed by refinement via energy minimization has been conceived and applied to two stem-loop structures of tRNAs, the anticodon and the TY-loop [108]. Energy minimizations are performed using wide-spread empirical potential functions of which AMBER, in various versions, is the most common [110]. The problems of RNA solvation and appropriate positioning of cations that compensate the electric charges of the phosphate groups are not yet solved satisfactorily.

2.3 Target Structures

2.3.1 Loops - Tetraloops

One of the most common structural motifs in RNA is the hairpin, comprising of double stranded stem and a single stranded loop. Stability of hair pins depends on the nature of the closing pair [142], on the length of the loop and on its sequence. Hairpin loops exist in various sizes, ranging from three- and four-membered types in ribosomal RNA up to loop sizes of 7, 8 or 9 nucleotides in tRNAs. A comparison between different force fields and loop size will be given in the result section. In ribosomal RNA the most frequent hairpins consist of four nucleotides [44] and are, therefore, often referred to as tetraloops. A sample structure is given in figure 6. Phylogenetic studies show that in ribosomal RNA nucleotide loops constitute 55% of all hairpins [171] formed in highly conserved region. Therefore it was concluded that they do not only possess a high thermodynamic stability but also an important biochemical functionality. They occur at transcription termination sites, provide sites for interaction with proteins and can be involved in stabilization of RNA 3D structure [17].
The picture of a highly structured stem and a disordered loop region has been questioned by a variety of recent publications [22, 36, 55, 76, 101, 159]. Notable for their abundance are three types of tetraloops: GNRA, UNCG, and CUUG [172] (again: N stands for any base and R for a purine, either G or A). Together the GNRA and UUCG loops make up 70% of all tetraloops in 16S-RNA [171]. UNCG-loops are presumed to be nucleation sites for RNA folding and to act as a protein recognition site, whereas GNRA-loops are thought to function as “anchors” during tertiary folding [99, 105]. It was suggested that the last two bases of a GNRA loop can contact two consecutive purine bases in the minor groove of an A-RNA helix, thus forming a pseudoknot. Examples for this behavior could exist in the conserved core of group I self-splicing introns. A model for the interaction between a GAAA-loop and an RNA helix was published by Pleij [114]. Jaeger et al. [63] pointed out that it is necessary to distinguish between the hairpin structure of a small RNA fragment in solution, and the structure of the same fragment in a larger biomolecule where tertiary interactions (e.g. formation of pseudoknots) are possible.
Most available studies are a combination of NMR-methods and distance ge-

ometry calculations or restrained MD. Examples can be found in [22, 100, 134, 159] for the UUCG and the UUUG loops, respectively, and in [55, 173] for the GNRA-loops. Though the sequences of the tetraloops studied are quite different, they have some structural features in common; the most prominent being the formation of an additional base pair (if permitted by the loop sequence) which is stacked on top of the stem, thereby reducing the loop to length two. This kind of behavior was found for the UNCG as well as the GNRA-loops.
Experimental investigations - mostly 2D-NMR measurements - are often a very time consuming task as they face considerable difficulties in the correct assignment of NMR-signals or in excluding the possibility of dimerization because of the comparatively high concentrations needed for NMR-experiments. An alternative is offered by a pure computational approach, even more so because most tetraloops present a system with severe steric constraints caused by the additional base pair in the loop. Kajava and Rüterjans [65] investigated the conformations of the 16 possible NUUN tetraloops to examine the structure of the "new" pair in the loop without the use of NMR-data. For most of the tetraloops in question the molecular modeling approach yields a few equivalent 3D-structures so that a "family" of conformations is obtained rather than a unique energy minimum geometry.

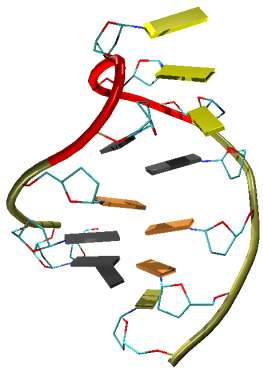


Figure 6: Sample tetraloop.

2.3.2 Pseudoknots

Recent work has indicated that pseudoknots are only marginally more stable than simple secondary structures (although thermodynamic data in this area are still scarce [94, 119]). This observation suggests a role for pseudoknots as conformational switches or control elements in several biological functions [135]. In molecules that lack an overall three-dimensional fold, pseudoknots fold locally and their positions along the sequence reflect their function [93]. For example, pseudoknots that are folded at the 5'-end of mRNAs tend to be involved in translational control whereas those at the 3'-end maintain signals for replication. In molecules with catalytic activities, pseudoknots are located at the core of the tertiary fold and involve nucleotides that are far apart in the sequence (RNAseP). The diversity of molecular biological functions performed by pseudoknots can be subdivided into tree different groups:

- **Translational control:** 5'-end pseudoknots appear to adopt two roles in the control of mRNA translation: either specific recognition of a pseudoknot by some protein is required for control, as described for the 5'-end of mRNAs in some prokaryotic systems [111, 135]; or, the presence of a folded pseudoknot is necessary with no requirements on the nucleotide sequence [13, 18, 158]. In several viruses, the expression of replicase is controlled either by *ribosomal frame shifting* [13, 18, 29, 26, 158] or by *in-frame read-through* of stop codons [169]. In both cases, pseudoknot formation is necessary [13, 29, 158]. The requirements appear, however, more strict for read-through than for frame shifting. Nevertheless, the correct position of the pseudoknot in the 3' direction with respect to the slip site in ribosomal frame shifting, and with respect to the AUG codon in read-through is an absolute requirement [13, 169]. The presence of three pseudoknots in 16S rRNA has been suggested on the basis of comparative sequence analyses [112]. In general these pseudoknots are assumed to show strong interactions with ribosomal proteins. One pseudoknot is known to be important for the binding of tRNA to the ribosomal A site [102, 170], and was shown to be essential for ribosomal function [116]. These observations are particularly interesting in view of the suggested conformational switch that involves the other two pseudoknots.

- **Core pseudoknots:** are necessary to form the reaction center of ribozymes. Most of the enzymatic RNAs with core pseudoknots are involved in cleavage or self-cleavage reactions [14, 35, 50, 99].

- **3'-end pseudoknots:** replication control is the common function of tRNA-like motifs at the 3'-end of several groups of plant viral RNA genomes [93]. This structural similarity is paralleled in biological function as the tRNA-like motifs are recognized by many tRNA-specific enzymes such as aminoacyl-tRNA synthetases, nucleotidyl transferase, or RNAseP [93]. The tRNA-like structure has been shown to be necessary for the initiation of replication [93]. A telomeric function of the tRNA-like structure was also demonstrated [123], in agreement with the genomic tag model associated with such 3'-terminal tRNA-like motifs [163]. Recently, the stretch of three pseudoknots preceding the tRNA-like structure in tobacco mosaic virus was shown to act as the functional equivalent of a poly(A) tail, stabilizing a reporter mRNA and increasing gene expression up to 100-fold [38].

In general a pseudoknot is formed when nucleotides within a hairpin loop interact with nucleotides outside the stem as shown in figure 7. The interaction was first proposed as a viable RNA folding motif by Pleij and co-workers [113] based on a study of the tRNA-like structures at the 3'-termini of certain plant viral RNAs. The stem and loop regions are usually short regions. As in tRNA, the 3D structure is characterized by coaxial stacking of the stems.

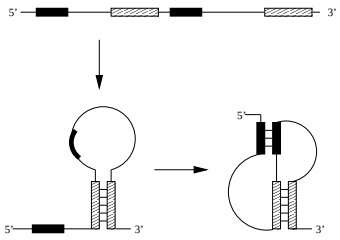


Figure 7: Pseudoknot formation.

3 Methods

3.1 Molecular Mechanics

Although quantum chemical calculations of molecular electronic structure can be highly accurate, they are also costly in computational time. Simulations of large molecules and biological macromolecules can therefore only be performed with classical mechanics. With these empirical methods one calculates the mechanical and electrostatic energy between bonded and non-bonded atoms. Early reports on molecular mechanics date from the seventies [16, 30, 51, 103]. Nowadays, molecular mechanics calculations are being performed on macromolecules taking into account surrounding solvent molecules. It is a method to calculate the structure and energy of molecules based on nuclear motions. Electrons are not considered explicitly, but rather it is assumed that they will find their optimum distribution once the positions of the nuclei are known. This assumption is based on the Born-Oppenheimer approximation of the Schrödinger equation. The Born-Oppenheimer approximation states that nuclei are much heavier and move much more slowly than electrons. Thus, nuclear motions, vibrations and rotations can be studied separately from electrons; the electrons are assumed to move fast enough to adjust to any movement of the nuclei. In a very crude sense molecular modeling treats a molecule as a collection of weights connected with springs, where the weights represent the nuclei and the springs represent the bonds (see figure 8).

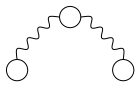


Figure 8: Nuclei and springs ;).

A force field is used to calculate the energy and geometry of a molecule. It is a collection of atom types (to define the atoms in a molecule), parameters (for bond lengths, bond angles, etc.) and equations (to calculate the energy of a molecule). In a force field a given element may have several atom types. The molecular energy is calculated by summing the potentials for bond distance,

bond angle and torsion angle deformation between 2, 3 and 4 bonded atoms and the dispersion and electrostatic potentials between non-bonded atoms. All potentials are based on structural parameters and empirically derived constants, stored in the force field.

3.1.1 A Short Glimpse at Force Fields

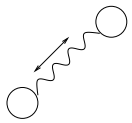
Again: A force field is used to calculate the energy and geometry of a molecule. In a force field a given element may have several atom types. For example, ethylbenzene contains both sp^3-hybridized carbons and aromatic carbons. sp^3-Hybridized carbons have a tetrahedral bonding geometry, while aromatic carbons have a trigonal bonding geometry. The C-C bond in the ethyl group differs from a C-C bond in the phenyl ring, and the C-C bond between the phenyl ring and the ethyl group differs from all other C-C bonds in ethylbenzene. The force field contains parameters for these different types of bonds. The total energy of a molecule is divided into several parts called force potentials, or potential energy equations. Force potentials are calculated independently, and summed to give the total energy of the molecule. They can be divided into **bonded** and **non-bonded** interactions. Examples of force potentials are the equations for the energies associated with bond stretching, bond bending, torsional strain and van der Waals interactions. These equations define the potential energy surface of a molecule. A sample force field is give in equation 1:

$$E_{tot} = \underbrace{E_{bond} + E_{ang} + E_{tors}}_{\text{bonded interactions}} + \underbrace{E_{vdW} + E_{Col} + (E_{HB} + E_{dipol} + \ldots)}_{\text{non-bonded interactions}} \qquad (1)$$

Energy due to bond stretching occurs, whenever a bond is compressed or stretched. The energy potential for bond stretching and compressing is described by an equation similar to Hooke's law for a spring, except a cubic term is added. This cubic term helps to keep the energy from rising too sharply as the bond is stretched.

$$E_s = \frac{k_s}{2}(I - I_0)^2(1 - 2(I - I_0)) \qquad (2)$$

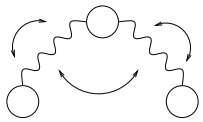
k_s is the force constant;
I_0 is the natural bond length;
I is the actual bond length;



As angles are bent from their norm the energy increases. The potential function below works very well for bends of up to about 10 degrees. To handle special cases, such as cyclobutane, special atom types and parameters are used in the force field.

$$E_\theta = k_\theta(\theta - \theta_0)^2(1 + 10^{-8}(\theta - \theta_0)^4) \tag{3}$$

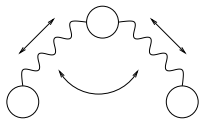
k_s is the force constant;
θ_0 is the natural bond angle;
θ is the actual bond angle;



When a bond angle is reduced the two bonds forming the angle will stretch to alleviate the strain. To handle phenomena such as this, cross term potential functions are introduced. Cross term potential functions take into account at least two terms such as bond stretching and bond bending.

$$E_{s\theta} = k_{s\theta}(\theta - \theta_0)[(I - I_0)_a + (I - I_0)_b] \tag{4}$$

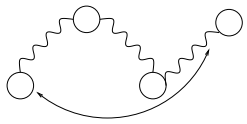
$k_{s\theta}$ is the force constant;
a and b represent bonds to a common atom;
I_0, I, θ_0, θ are as above;



Intramolecular rotations (rotations about torsion or dihedral angles) require energy. For example, it takes energy for cyclohexane to go from the chair conformation to the boat conformation. The torsion potential is a Fourier series that accounts for all 1-4 through-bond relationships.

$$E_{Tor} = \frac{V_1}{2}(1 + cos\omega) + \frac{V_2}{2}(1 + cos2\omega) + \frac{V_3}{2}(1 + cos3\omega) \tag{5}$$

V_1, V_2, V_3 are force constants in the Fourier series;
ω is the torsion angle

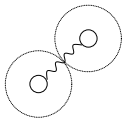


The van der Waals radius of an atom is its effective size. As two non-bonded

atoms are brought together the van der Waals attraction between them increases (a decrease in energy). When the distance between them equals the sum of the van der Waals radii the attraction is at a maximum. If the atoms are brought still closer together there is strong van der Waals repulsion (a sharp increase in energy).

$$E_{vdW} = \epsilon e^{\left(\frac{r_0}{r_v}\right)} - \left(\frac{r_v}{r_0}\right)^6 \tag{6}$$

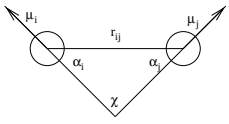
ϵ is the energy parameter which sets the depth of the potential energy well
r_v is the sum of the van der Waals radii of the interacting atoms
r_0 is the distance between the interacting centers



In some force fields electrostatic interactions are accounted for by atomic point charges. In other force fields, such as MM2 and MMX, bond dipole moments are used to represent electrostatic contributions. One can readily see that the equation below stems from Coulomb's law. The energy is calculated by considering all dipole-dipole interactions in a molecule. If the molecule has a net charge (e.g., $NH4^+$), charge-charge and charge-dipole calculations must also be carried out.

$$E_{Col} = \frac{\mu_i \mu_j}{D(r_i j)^3}(cos\xi - 3cos\alpha_i cos\alpha_j) \tag{7}$$

D is the dielectric constant of the solvent;
ξ is the angle between two dipoles μ_i, μ_j;
$\alpha_i \alpha_j$ are the angles between the dipoles and a vector connecting the two dipoles;
$r_i j$ is the distance between the dipoles



An example of a force field is the AMBER [165, 164] force-field (see equation 8) which only differs with the CHARMM (Chemistry at HARvard Macromolecular Mechanics) and GROMOS equations in the use of a hydrogen bond term, which is absent in GROMOS and optional in CHARMM. The optional hydrogen bond term in CHARMM has the advantage of an angular component, accounting for directionality. In the GROMOS force field,

however, an extra harmonic term is used to constrain the improper dihedral angles at their preferred value. The non-bonded Lennard-Jones-Coulomb potential or 1-6-12 potential should be able to account for hydrogen bonding interactions in a well tuned force field. More recently, the AMBER force field was complemented with a polarization potential [110].

$$\begin{aligned} E_{\text{total}} = & \quad \textstyle\sum_{\text{bonds}} K_r(r - r_{eq})^2 \\ & + \quad \textstyle\sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 \\ & + \quad \textstyle\sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \\ & + \quad \textstyle\sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right] \\ & + \quad \textstyle\sum_{\text{H-bonds}} \left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}} \right] \end{aligned} \tag{8}$$

Though the underlying formulas of a standard force field are rather simple its quality depends particularly on the parameterization. In this process the force constants and the equilibrium values in equations 2 - 7 are assigned appropriate values. The quality of a force field depends also very much on the kind of data used for parameterization and the class of molecules these data were taken from. Data useful for calculation of parameters include for example structural data, energy data or vibrational frequencies.
In principle there are two different methods of parameterization: It can be done "by hand", i.e. one looks at where the largest errors in comparison with experimental data are and tries to make adjustments in the parameters to minimize these errors, or it can be done by least square minimization. Whereas the first method is useful for force fields where very little data are available for parameterization it soon becomes difficult to use when the amount of data rises. An example for the second method was implemented by Lifson and coworkers [51, 106, 162] who referred to this method as the "consistent force field". The advantages of this method are obvious since the optimization is done in a precise and mechanical way. Nevertheless there are several disadvantages like the amount of computer time necessary for calculations involving a lot of data and most important the fact that least square optimization depends on all variables being measured in the same units. Therefore to compare for example errors in bond lengths and valence angles it is necessary to estimate how much an error in one case is equal to how much error in the other case. For this purpose weighting schemes were devised (e.g. Wertz et al. [166]) which are used iteratively. Probably the best

way for parameterization is a combination of both methods, using "intuition" to get reasonable starting values and numerical methods when huge amounts of data are involved.

3.1.2 Force Fields Used

Both the standard force-field of JUMNA termed FLEX and the AMBER4.1 force field were used in this thesis. Within JUMNA fixed bond lengths are assumed so that the bond length stretching term is not relevant. Both force-fields calculate internal energies merely under vacuum conditions. A distance dependent dielectric function $\epsilon(R)$ (see equation 9) can be introduced in the electrostatic term in order to account for the dielectric damping induced by the solvent.

$$\epsilon(R) = D - \frac{(D-1)}{2}[(RS)^2 + 2RS + 2]e^{-RS} \tag{9}$$

The formulation of this function allows to vary both the plateau value of the dielectric reached at long distances (D) and the slope of the sigmoidal function (S). Combining the damping with a reduction of all phosphate net charges to $-0.25e$ mimics the effect on counter-ions. This is of course just a crude model. In particular in the case of highly charged nucleic acids it is not surprising that such a simple approach cannot be used for estimating the relative stability of different confomers.
For this purpose the external electrostatic term of the AMBER force field has been replaced by a continuum treatment of solvent-induced interactions. The most important term is the RF (Reaction Field) potential which describes the interactions of the solute charges with the polarize aqueous medium and adds to the Coulomb type potential. The RF contribution has been confined on the Poisson equation:

$$\nabla\epsilon(r)\nabla\Phi(r) = -4\pi\sum_{i=1}^{N} q_i\delta(r - r_i) \tag{10}$$

q_i denotes the atomic charges and $\Phi(r)$ is the potential. $\epsilon(r)$ describes the change of the dielectric permittivity from the solute (ϵ_{solute}) to the solvent (ϵ) at the solvent/solute interface. If $\epsilon_{solute} = 1$, RF is defined as follows,

$$\Phi_R(r) = \Phi(r) - \sum_{i=1}^{N}\frac{q_i}{|r - r_i|} = -\int_{solvent} d\tau' P(r')\nabla'(\frac{1}{|r - r'|}) \tag{11}$$

where $P(r)$ describes the polarization of the solvent by the solute charges. This potential can be redefined in terms of virtual sources, which are located exclusively within the volume defined by the solute surface envelope,

$$\Phi_R(r) = \frac{1}{\epsilon}\left[\int_{solvent} d\tau' P(r')\nabla'(\frac{1}{|r-r'|}) - (\epsilon - 1)\sum_{i=1}^{N}\frac{q_i}{|r-r_i|}\right] \tag{12}$$

where the virtual polarization is defined by $P(r) = (\epsilon - 1)\nabla\Phi(r)/4\pi$. This representation is the starting point of the Field Integrated electrostatic Approach (FIESTA) which introduces a series expansion in terms of spherical harmonics and reasonable approximations in order to calculate the polarization $P(r)$ analytically. The final expression used to calculate molecular energy in solution and to estimate the relative stability of the different confomers is given below:

$$\begin{aligned} E &= \tfrac{1}{2}\sum q_i\Phi_R(r_i) + \sum(\tfrac{q_iq_j}{R_{ij}} - \tfrac{Aij}{R_{ij}^6} + \tfrac{B_{ij}}{R_{ij}^{12}}) \\ &+ \tfrac{1}{2}\sum V_s(1 \pm cos(N_s\tau_s)) + \sum F_a(\sigma_a - \sigma_a^o)^2 \end{aligned} \tag{13}$$

The partial charges q_i the Lennard-Jones parameters A_{ij} and B_{ij}, and the parameters V_s, N_s and K_a defining the distortion energy associated with torsion angle τ_s and valence angle σ_a, respectively, were taken from the AMBER parameterization. R_{ij} is the distance between atoms i and j.

3.1.3 Structure Optimization

Calculating the energy with respect to a given geometry is only one part of optimizing the structure of a molecule. To improve the structure it is necessary to change the geometry in such a way, that the total energy is lowered. This process is repeated iteratively so that an energy minimization corresponds to a geometry optimization. The potential function is a function of a large number of variables which specify the molecule's geometry in either internal or Cartesian coordinates. The ideal solution for geometry optimization would be the **global** minimum of this function corresponding to the molecule in a state of minimal free energy. Since there is no known method to determine the global minimum of a function of many variables, one usually is trapped in a **local** minimum, a behavior often called the "global minimum problem". One consequence of ending the optimization in an local minimum is the fact that the "optimized" structure will depend on the starting geometry so that

it is usually necessary to use different starting geometries and compare the resulting structures to get lower energies.
The global minimum problem is known for a long time since it occurs in many fields of science. Consequently general optimization procedures are of great interest and there is a wide selection of available algorithms, some of which are described in this paragraph. Probably the simplest of all optimization algorithms is the method of **steepest descent**, in which only the first derivative of the energy with respect to the atomic coordinates is calculated so that the geometry can be changed in direction of the largest energy gradient. This method leads directly into the next local minimum and is therefore only used at the very beginning of an optimization to get rid of the largest energy contributions. Convergence of this method is best when one is far from the minimum and thus the gradient is largest. A more sophisticated method is for example the **Newton Raphson method** which uses first and second derivatives which can be calculated numerically or analytically (see for example. The advantages of the Newton Raphson method lie in the faster convergence (even in the vicinity of a minimum) and the smaller computational effort to reach a minimum (i.e. a smaller number of steps). In most programs a combination of steepest descent and Newton Raphson method is used.
So far all optimization methods considered were purely analytical calculations where no random elements were involved. Another approach to optimization processes is the use of stochastic techniques as it is done with method of **simulated annealing**. Simulated annealing is a widely used optimization procedure that originally came from the field of statistical physics (e.g. [69]). In effect it tries to simulate the cooling and the crystallization process occuring in a heated solid. Starting point is a configuration space E and a so-called energy function U which is defined in the following way $U : E \rightarrow R$. In the case of molecular mechanics U corresponds to the potential function whereas E is the conformational space constructed from all possible conformations of the molecule. In addition a temperature is defined. Beginning from a starting geometry the energy of the molecule is calculated giving the energy E_0. The next step is a random step in conformational space which in this case equals a random change of the molecular geometry. Again the energy is calculated resulting in energy E_1. Now there are two possibilities: if $E_1 < E_0$ the random

step is accepted in any case, if $E_1 > E_0$ it is only accepted if:

$$p = \begin{cases} 1 & : \quad E_0 \leq E_1 \\ e^{-\frac{E_1 - E_0}{kT}} & : \quad E_1 > E_0 \end{cases} \tag{14}$$

p is a random number between 0 and 1, and k and T are the Boltzmann constant and the above defined temperature, respectively. This criteria is also known as the Metropolis algorithm [97]. It ensures that the optimization cannot be trapped in a local minimum since higher energies are accepted with a certain probability so that energetical barriers can be overcome. If n is the number of steps that are calculated the global minimum is always found for $n \to \infty$. The optimization is continued making a given number of steps at a given temperature, then the temperature is lowered by a certain value (the so-called cooling schedule). Simulated annealing is most useful for systems that are not too restricted and usually gives good results when a high computational effort is used.

Another principal possibility for an optimization algorithm is the combination of the two principles mentioned above, namely a combination of analytical and heuristical methods as it is done in the so-called **Bremermann method**. This technique was originally devised for the use in biomathematics by Hans Bremermann [12], but it can be adopted to geometry optimization of molecules as it was done by Eberhard von Kitzing for the AMBER force field [70, 71]. The first step in a Bremermann optimization is the definition of a certain number n ($n = 10 - 20$) of axes in a molecule around which atoms or groups of atoms are allowed to rotate. These axes may be "conventional" axes along bonds between atoms, but they can also be defined to enable rotations of larger parts of the molecule as can be seen in figure 9. Figure 9 shows two rotational axes where Φ is defined by two adjacent phosphorus atoms and allows the rotation of the nucleoside together with the sugar whereas Ψ is defined by the glycosidic bond between nucleoside and sugar thus allowing for a variation of the χ angle. By defining rotational axes in this way more "global" changes in the molecules geometry are made possible since larger parts of the molecule become flexible.

The configurations made accessible by rotation around these axes form the conformational space which is to be sampled by the Bremermann method, each axis representing a coordinate in this space. Again a starting point has to be given (e.g. geometry x_k) then a random direction R_k within the restricted conformational space is chosen by taking n random numbers from a

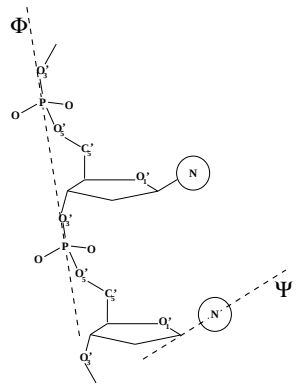


Figure 9: Definition of rotational axes for the Bremermann method.

Gaussian distribution. Along this "search direction" the energy is calculated at five different points: $U(x_k + \lambda_m R_k)\ with\ \lambda_m \in \{-2, -1, 0, 1, 2\}$ where the size of λ_m is a parameter of the method. The five energy values are interpolated by a fourth order polynomial, the global minimum of which is calculated using Cardan's formula. If the energy of the configuration corresponding to this minimum is lower than the energy of the starting point the minimum becomes the starting geometry for the next iteration x_{k+1}.
Since the Bremermann procedure involves the use of heuristical elements it is obvious that two Bremermann "runs" starting from the same geometry will not necessarily end up in the same minimum, so that the best way to use this algorithm is to make several runs from the same starting geometry and then choose the "best" configuration as the begin for the next set of Bremermann optimizations. The Bremermann method works best for molecules that are already coarsely optimized and it requires some experience in choosing the right magnitude of λ_m and in the definition of the rotational axes. Best results are obtained for not too constrained systems (e.g. four- or higher-membered loops) where the energy can be lowered by an improvement of stacking and by increasing the number of base pairs in the loop region.

3.1.4 An Introduction to Molecular Dynamics

The molecular energy can be mathematically minimized by alteration of the structural parameters (the internal coordinates). The inherent shortcoming of finding sometimes only a local minimum in conformational space, led to the more realistic simulation of atomic and molecular movement.
MD was initiated by Alder and Wainwright 1957 [1]. Integration of Newton's laws of motion, using Verlet's or a leapfrog algorithm, leads to atomic trajectories in space and time

$$F_i(t) = m_i a_i(t) = m_i \frac{\partial^2 r_i(t)}{\partial t^2} \tag{15}$$

The forces on the atoms are the negative gradient of the potential energy function mentioned in Equation 1

$$F_i = -\frac{\partial}{\partial r_i} E_{tot} \tag{16}$$

MD simulations enable the calculation of properties within their time scale. The prediction of tertiary biopolymer structure out of the primary sequence could increase the amount of biopolymeres with known structure from approx. 6000 (available in the Protein Data Bank of the Brookhaven National Laboratory) to 200,000 (available in several sequence databases). The process of folding, however, takes place in the millisecond scale. Since integration of Newton's laws only gives reliable results at time steps of 1 femtosecond, the simulation of folding is not within reach of present day computers. Faster sampling of conformational space is another advantage of MD. Simulated annealing is the heating of structures up to 1000 or 2000 K, before cooling down into an energy minimum. Another application of MD is calculation of the Gibbs free energy using free energy perturbation theory [7, 90], where atoms are slowly grown into other atoms. Reviews on MD and its use in biochemistry have been numerous [46, 48, 66].
The procedure for a dynamics simulation is subject to a lot of user-defined variables. The evaluation of the atomic positions is not performed on a continuous basis, but at intervals of a femtosecond. Since this is the time-scale of stretches of the bonds with hydrogen atoms, these stretches should be constrained in order to permit time steps of 2 fs. The algorithm that permits this increase in simulation speed is called SHAKE [45, 132]. Application of

constraints on all bond lengths and bond angles would increase the permitted time step even more, without too much loss of information. In solvated dynamics the solute is immersed in a cubic box of solvent molecules. In order to prevent solvent molecules from evaporating, periodic boundary conditions are imposed. The box is surrounded by its own image 26 times. In GROMOS a truncated octahedron permits the same advantages at less computer time. Since most of the computational time is spent on evaluating the non-bonded interactions between atoms, the evaluation time step for non-bonded atom pairs can be increased to a small extent. The amount of non-bonded atom pairs can be significantly reduced by using a cut-off distance beyond which atoms are no longer considered to interact. Shifting and switching functions make the gradient over this distance smooth. Treating atom groups with non-polar hydrogen atoms as an ensemble (united atom approach) and defining charge groups of atoms with net zero charge, are also time-saving simplifications. To impose experimental heat bath conditions [9], a rescaling of velocities is periodically carried out, to ensure calibration around the simulation temperature, according to a Langevin equation:

$$\frac{m_i \partial^2 r_i(t)}{\partial t^2} = F_i + m_i \beta \left[\frac{T_0}{T(t)} - 1 \right] \frac{\partial T_i(t)}{\partial t} \tag{17}$$

A small value of β is related to a long temperature relaxation time. Short relaxation times may give rise to non-Newtonian behaviour and can only be used in an equilibration stage.

Since 1983 MD simulations of nucleotides have been reported. Nucleic acids have been considered a challenge for simulations because of the negative backbone charge and the polyelectrolyte behaviour. The first simulation over 90 ps in vacuum was performed with neglect of electrostatics [87]. Tidor et al. [155] did not neglect electrostatics but reduced the phosphate charge to -0.2. Vacuum simulations over 250 ps with X-Plor [11] and 100 ps with CHARMM [77] have been reported. The GROMOS force field was used for vacuum simulations over 6.6 ps [117] using an additional hydrogen bond potential and for a 30 ps simulation by Ravishanker et al. [126]. In vacuum simulations over 50 ps [37] and 20 ps [67] have been performed with the AMBER force field. Song [151] used the AMBER force field and the Discover potentials for a 100 ps vacuum simulation. In 1985 Singh et al. introduced the use of hydrated counterions to neutralize the negative charges of the phosphate backbone [148].

The same approach was used in a 50 and 84 ps simulation [124, 125] and a 100 ps simulation [152], based on the AMBER force field. The explicit inclusion of water molecules and counterions, however, has only been described in some reports, starting with a 106 ps simulation [140]. Other simulations in solution with the AMBER force field were over 50 ps [53], 40 ps [54] and 48 and 20 ps [56, 57]. With the GROMOS force field solvated dynamics of oligonucleotides were reported over 80 ps [47], 60 ps [174, 175] and 140 ps [153]. In the latter investigation explicit constraints were put on the Watson-Crick hydrogen bonding of the base pairs. Simulations of solvated dynamics with other force fields have also been reported [23, 154].
In oligonucleotide dynamics simulations particular attention should be paid to the atomic charges. The negatively charged phosphate groups may very well influence the trajectory. Scaling down these charges, as in the CHARMM force field, is one solution. Explicitly using sodium counterions at a fixed distance is another solution [140]. Here, more than anywhere else, reliable charges are of crucial importance [131]. Much discussion has been focused to calculation of partial atomic charges, based on empirical procedures, population analysis and molecular electrostatic potential derived (PD) fits [130]. In the GROMOS force field empirical charges are used to best reflect experimental results. For the CHARMM force field, atomic charges are initially calculated with a 6-31G* basis-set and afterwards corrected to fit experimental results. The AMBER force field uses atomic charges fitted to reflect a MEP from STO-3G basis sets [148]. The MEP was sampled at 1.4, 1.6, 1.8 and 2.0 times the solvent accessible molecular surface as determined by Connoly [24]. The idea of PD charges was proposed by Cox et al. [25]. The possibility of obtaining equally reliable MEP fitted charges based on a semi-empirical calculation was reported by Besler et al. [10] and others [129]. These PD charges are equally well reflected by charges from a Distributed Multipole Analysis (DMA) [33, 118]. A straightforward charge calculation is of great benefit to force field users, who have to add new residues (for example drugs) or covalently altered residues to the force field. The charges should be in tune with the existing force field charges to avoid compromising the integrity of the force fields. This may well be the major bottleneck in performing simulations, since often two methods of charge calculation give opposite results.
In contrast with molecular mechanics where a structural minimum is the end result, MD offers so much information that it is hard to quantify this. Most programs offer animation of the structures written out at specific intervals.

Monitoring certain geometric features, including hydrogen bonds is another means of analysis. Also, the energies can be plotted and monitored during the animation. Since MD does not permit the making and breaking of bonds, the ligands have to be bonded covalently to the bases at forehand. The most nucleophilic regions in DNA are found in the negative wells of the MEP of DNA. According to Pullman et al. [122] the N7, O6 and N2 atoms of guanine as well as the N6 and N7 atoms of adenine are to be considered. For the nitrogen mustards, which resemble these BABQs, N7 alkylation is preferred [73, 95] and the same alkylation position has been reported by several authors for aziridines [68] and BABQ compounds [52, 86]. After manually binding the ligand, only the lack of distortion of the oligonucleotide can be used as a measure of likelihood of binding. To quantify the movements of the original B-type oligonucleotide structure a program called Curves, Dials-and-Windows (CDW) 1991 [81, 149], offers help. Here, not only the torsion angles of the nucleoside backbone, but also the movement of the bases compared to a central helicoidal axis can be monitored.

3.1.5 Molecular Dynamics RNA Remarks

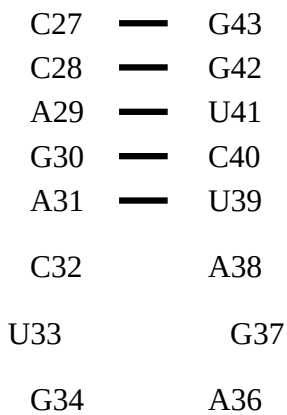
As we have seen MD calculations provide information on thermal motion in biopolymers that is otherwise hard to obtain. They are extremely costly as far as computer time is concerned and thus can be extended only over short periods commonly in the range of nanoseconds. They suffer also from various artifacts, for example finite size effects as the volumes considered in simulations have to be kept to a minimum and cut-off problems with long-range electrostatic forces. For an example of accurate nanosecond dynamics we recommend a simulation of the motions in dinucleotides in an crystal environment [110].
MD calculations of anticodon hairpin in tRNAs has been studied in a series of large scale computations [2, 3, 4, 89]. The calculations were performed on the hairpin fragments of $tRNA^{asp}$ with different truncation radii for solvent interactions (8 and 16). Auffinger and Westhof [4] presents a computation by means of the particle mesh Ewald algorithm that explicitly allows the handling of all long-range electrostatic forces. Six 500ps-long trajectories yield a total observation time of 3ns that is sufficiently long for an identification of several structural features: first, differences in the dynamics of the Watson-Crick bp, the Y-C pseudo-bp, and the noncanonical G-U "wobble"

bp; second, the existence of two C-H$\cdots$O hydrogen bonds which contribute to the overdvips-all stability of the fragment; and third, local heterogeneity attributed to an ensemble of accessible structural microstates between which the RNA molecule drifts in random manner. Another hairpin loop was studied using MD [147] in order to understand the differences in stability between an especially stable tetraloop and its less stable mutant.

3.2 Programs Used

3.2.1 MC-SYM

MC-SYM stands for **M**acromolecular **C**onformations by **SYM**bolic programming and is not a force field program but a tool to obtain 3D nucleic acid structures which are in accordance to a list of input constraints. The program was written and tested by the group of Cedergren and Gautheret (see references below). The backtracking algorithm in MC-SYM searches the conformational space of an RNA molecule and all geometries that fulfill the constraints are returned in PDB-format to be optimized by a force field program. The conformational space explored is determined by the choice of pre-computed nucleotide conformations and transformations. MC-SYM has been successfully used for RNA hairpins [41, 92], for tRNAs [91], or for the Rev-binding site of HIV-1 [85].
The program input for MC-SYM consists of a simple ASCII-file divided into two sections. The first section, the so-called "sequence-section" defines the sequence and secondary structural information of a macromolecule. It lists all the nucleotides and fragments that compose the RNA and information on how these parts are connected or related to others. The second section, the "constraints-section" consists of additional constraints which might be local (i.e. they are valid for just one base or a base pair) or global (i.e. they are valid for all nucleotides). The following example shows the description of a simple stem-loop structure (RNA hairpin) and was taken from the MC-SYM manual (see figure 10). The molecule modeled is the anticodon stem-loop of a tRNA.




```
SEQUENCE
; 5' helical strand
  A   rC   27  reference          type_A
  A   rC   28  connect      27    type_A
  A   rA   29  connect      28    type_A
  A   rG   30  connect      29    type_A
  A   rA   31  connect      30    type_A
; 3' helical strand
  A   rU   39  wc           31    stk_AA
  A   rC   40  connect      39    type_A
  A   rU   41  connect      40    type_A
  A   rG   42  connect      41    type_A
  A   rG   43  connect      42    type_A
; 3' loop strand
  A   rA   38  connect      39    stk_AA
  A   rG   37  connect      38    stk_AA
  A   rA   36  connect      37    stk_AA
  A   rA   35  connect      36    stk_AA
  A   rG   34  connect      35    stk_AA
; 5' loop strand
  A   rC   32  connect      31    stk_AA
  A   rU   33  connect      32    stk_AA
; Constraints section
ADJACENCY
    1      4
CONSTRAINT
   33      34     distance  O3'  P   1   3
GLOBAL
   P       P      3.5
   C1'     C1'    3.5
```

Figure 10: Input file for MC-SYM for a simple stem-loop structure.

The secondary structure shown on the left-hand side of figure 10 indicates that bases C27 to A31 form base pairs with G43 to U39. It is assumed that bases A38 to G34 are stacked and as a first attempt C32 over A31 and U33 over C32 are stacked as well (following a quite common strategy in RNA modeling that tries to maximize stacking). These assumptions lead to the input file shown in figure 10. In the first section of the input file a typical line consists of several entries of the following format:

• *chain-identifier*: a letter indicating the strand which is important only for molecules with more than one strand.

• *nucleotide-type*: gives the sequence of the molecule and can be one of rA, rC, rG, or rU.

• *nucleotide-identifier*: a unique number identifying a certain nucleotide.

• *connection-function*: a keyword that specifies the position of the current nucleotide relative to another. Keywords can be chosen from a wide range of possibilities such as all kinds of base-pairs (Watson-Crick, Hoogsteen, reverse Hoogsteen, Wobble, unusual base pairs like G-A, base pairs with different numbers of hydrogen bonds, ...), standard RNA or DNA helix forms, stacking, or simple connections between two adjacent bases.

• *reference-nucleotide*: the number of an already defined nucleotide which the connection-function refers to.

• *conformational-set*: a set of pre-computed conformations and transformations which is taken from a database. This set comprises the "allowed" movements for the given nucleotides. The "allowed" movements range from a simple "type_A" which stands for a base in C3′-endo conformation taken from an A-RNA helix to the keyword "sample+" which represents a total of 59 different conformations and transformations. The total number of conformations in the example is 6561 ($= 3^8$). This stems from the combination of 9 A-type nucleotides ("type_A", 1 conformation) and 8 A-type nucleotides stacked over other A-type bases ("stk_AA", 3 conformations).

Whereas the first part of the input file specifies the largest possible search tree for the MC-SYM run, the following section (starting with keyword "ADJACENCY") reduces the number of possible conformations significantly by introducing a number of constraints. The "ADJACENCY" keyword refers to the O3′-P bonds in the molecule and is used when MC-SYM detects a loop-construction (i.e. when unpaired bases are not at the end of a stem, but between paired regions). In the given example this distance may vary between 1 and 4 Ångstrøms. Adding the "ADJACENCY"-section to the

input file reduces the number of conformations to 645. In the following "CONSTRAINT"-section an example for a local constraint can be seen. It is specified that the distance between atoms O3′ of U33 and P of G34 must be larger than 1 Å and must not be greater than 3 Å , thus reducing the number of possible conformations to 56. The last section, labeled "GLOBAL" is for definition of global constraints that are valid for all nucleotides in the molecule. In the example from figure 10 this means that only conformations in which P and C1′ atoms are at least 3.5 Å apart are acceptable, which reduces the total number to 52 different geometries.
MC-SYM is a very handy tool which is useful for finding possible molecular geometries when only the secondary structure and some additional data are available. For small molecules it can also be used to generate a "pool" of starting geometries when only the secondary structure is known. These starting geometries can than be minimized by a force field program and the "best" geometries (in terms of energy) can then be selected for further optimization.

3.2.2 AMBER

One of the two force field programs used to produce the results presented in this thesis is the widely used AMBER (**A**ssited **M**odel **B**uilding with **E**nergy **R**efinement) force field [110] in the versions 4.0 and 4.1 (mainly), written by Kollman, Weiner et al. AMBER4.0 is a widely used program that is suitable for the calculation of two of the most important types of macromolecules in biochemistry, i.e. peptides and nucleic acids. Molecules can be treated in a quasi-vacuum as well as in solution and it is also possible to do not only minimization but also molecular dynamics. AMBER4.0 is comprised of several modules that fulfill specific tasks; figure 11 shows the flow of information between the different AMBER modules. Modules represented by a circle stand for data that has to be supplied by the user, whereas modules drawn as a box stand for the actual programs. There are four major types of input data to AMBER modules:
• The actual commands for each module: these are read in from an input file and have a specific format for each module.
• Cartesian coordinates: these are read in via PDB-files and result usually from X-ray-crystallography, NMR, or from model-building.
• Topology: this input comes from the database which is part of the AMBER package. The unit of information within the database is a "residue",

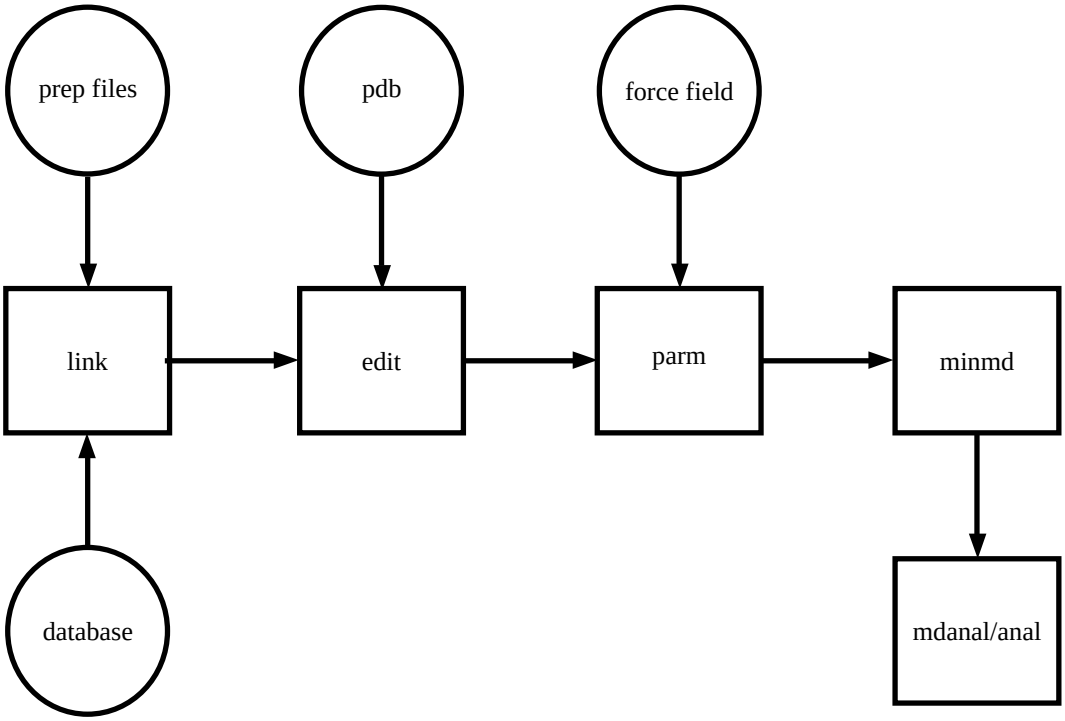


Figure 11: Basic information flow in AMBER4.0.

which can be as small as a single hydrogen atom and as large as complete nucleic acid. The database contains information about the way atoms within a residue are connected as well as a standard topology (i.e. a complete set of bond lengths, valence angles and torsional angles for each residue).
• Force field parameters: these are sets of parameters of each combination of atom types occuring in the database. Both the database and the force field parameters can be changed and expanded by the user, in case of the topology database a special program, **Prep**, is needed to do so.
The functions of the modules shown in figure 11 are as follows:
• **Link**: Link deals only with topology. Its main user input is a list of residues that correspond to the sequence of the molecule. Link reads the information for these residues from the topology and creates a (binary) topology file which is specific to this molecule. AMBER knows two possibilities of representing

molecules: In the so-called "all-atom-model" each atom in a molecule is considered with its Cartesian coordinates, whereas in the "united-atom-model" non-active hydrogens are combined with carbons to a united atom (so that a CH_2-group is treated only as one instead of three atoms).
• **Edit**: Edit deals mainly with coordinates and their conversion. After reading the topology file created by Link its main purpose is to read in PDB-files and apply the contained atomic coordinates to the system designed by Link. Should the set of Cartesian coordinates not be complete Edit is able to create data for the missing atoms from the database file. Edit is also responsible for solvatation of a molecule in water, for the addition of counter ions, for changes to specific coordinates, or for a conversion between Cartesian and internal coordinates. Edit writes out a binary file that contains both topology and Cartesian coordinates.
• **Parm**: Parm will determine which bonds, angles, dihedrals, and atom types exist in the system and extract the appropriate parameters for them from the force field file. Parm writes out another topology file containing the sequence of atoms and the corresponding parameters and a coordinate file containing only the Cartesian coordinates. This method has the advantage that for a given molecule with a given sequence the topology file has to be created only once, even when the geometry is varied as long as no bonds are broken or newly formed. This fact was used for example in the program *randstruct* described below.
• **Minmd**: Minmd is the energy minimizer and the MD program. This module relaxes the structure by iteratively moving the atoms down the energy gradient until a sufficient low average gradient is obtained. Its output consists of several files including a listing file, a summary file and a coordinate file containing the optimized geometry.
• **Mdanal/Anal**: these programs deal with analysis of structure and molecular mechanical energy of a single configuration of a system (Anal) and with trajectory averaging, correlation analysis, and general analysis of MD trajectories (Mdanal). Anal can also be used to generate PDB-files from a minimized structure or to compare two geometries and calculate root mean square distances.
Apart from the quality of the force field itself the clear separation between topological and positional information makes AMBER ideal for experiment with new optimization algorithms as has been done in the Bremermann method, conformational sampling and genetic algorithms.

3.2.3 JUMNA 10

JUMNA stands for **Ju**ction **M**inimization of **N**ucleic **A**cids and is a molecular mechanics program that was designed by Richard Lavery and Heinz Sklenar [83, 79] especially for dealing with nucleic acid structures. JUMNA differs from AMBER not only in the specialization to nucleic acids but also in a different force field (JUMNA uses the FLEX force field [80, 83, 84]) and in a different description of molecular structure.
The basic approach of the JUMNA methodology is to split nucleic acid fragments into a collection of 3′-mono-phosphates (with the exception of the 3′-termini which are simple nucleosides). This division is achieved by cutting the O5′-C5′ bonds of the phosphodiester backbone. These nucleotides are positioned with respect to a local helical axis with a set of 6 helicoidal parameters (according to the Cambridge convention [28]). These helicoidal variables consist of three translations (xdisp, ydisp, and zdisp) and three rotations (inclination, tip, and twist). The structure of the fragment can then be energy optimized in terms of helicoidal parameters plus variables describing the internal conformation of each nucleotide (glycosidic angle, sugar torsions and valence angles and two backbone torsions ϵ and ζ). The remaining backbone torsions are treated as dependent variables. During energy minimization energy penalties ensure that the sugar rings and the phosphodiester junctions between successive nucleotides close properly. One distance constraint, O5′-C5′, and two angle constraints P-O5′-C5′ and O5′-C5′-C4′, are used per nucleotide junction. This approach leads to an important reduction in the number of variables required compared to classical molecular mechanical algorithms and also gives more control over the conformations which are generated. Dielectric conditions can be varied through the use of a sigmoidal distance dependent dielectric function of variable slope and plateau, the use of a chosen fixed dielectric constant or the function $\epsilon = nr$. The net charge on each phosphate group can also be varied to mimic counter-ion screening. Explicit mobile counter-ions or water molecules can also be included through a ligand option. JUMNA can build, manipulate and energy optimize fragments of DNA or RNA having up to 4 strands. Many structural features can be blocked during minimization and certain global or local features can be constrained such as base pair opening angle, average twist or rise per base step, radius of curvature, sugar phase and amplitude, atom pair distances, and torsion and valence angles. This makes for an easy use of experimen-

tal data like atom-atom distances determined by NMR. The simple use of constraints and the representation of the molecule in terms of helicoidal and backbone parameters are the most powerful features of JUMNA, since the description of molecular geometry is thus sequence independent, so that the effects of sequence changes can be tested very easily.

3.2.4 Randstruct

Yet another possibility to optimize molecular geometry is conformational sampling. Here a simple "greedy" algorithm (i.e. only conformations with a lower energy than the previous are accepted) is applied to random changes of the geometry. The program *randstruct* (from **rand**om **struct**ure) which applies these principles to the optimization of rather rigid loop structures was written by Herbert Kratky and shall be described in the following paragraph.
The program *randstruct* actually consists of two main modules: the evaluation module and the geometry randomizer. As the evaluation module the minimizer ("minmd") of the AMBER4.0 force field was chosen and also the geometry format used in the program corresponds to the AMBER data structure. The purpose of *randstruct* is to further optimize molecules that were already treated with standard optimization techniques or for example with the Bremermann method. *Randstruct* assumes that the molecule consists of a rigid part and a certain number of flexible bases. It tries to optimize the overall energy by changing the conformation of the flexible part. Information about the respective size of the parts and other optimization parameters are supplied on the command line.
At the begin of an optimization process the PDB-file, the file containing the Cartesian coordinates, and the topology file of the molecule are read in. From the PDB-file information concerning the sequence and the size of the molecule are taken whereas the actual Cartesian coordinates are read from a separate file as the values in PDB-files proved to have too little accuracy. The geometry of the flexible part is then reduced to the bare backbone connecting the ends of the flexible region, which can be treated as a loop region (see figure 12). The conformation of this backbone can be described simply be using q torsional angles going from the $5'$-end of the loop region to its $3'$-end. In another command line option the number p of torsional angles that are to be changed randomly can be specified. Then p of the q angles are chosen

randomly and assigned random values, after which the loop is closed again by an iterative procedure. This four-step process is shown in figure 12. A stands for the beginning of the flexible region on the 5′-side of the stem, B stands for the 3′-end of the flexible region, and C stands for the beginning of the rigid region on the 3′-side of the stem. Step a stands for the starting geometry including all atoms in the molecule. In step b all atoms except those along the backbone in the flexible region have been removed and the conformation of the backbone has been changed randomly. In step c the flexible region is again connected to the rigid regions by the following procedure: starting from point A the rest of the flexible region is rotated around the bond between atom A and the following atom in such a way that the distance between points B and C is minimized; then this process is repeated for the next bond along the backbone until the distance between points B and C is lower than a previously defined value. Step d finally shows the new, optimized structure; in the ideal case this structure has lower energy than the starting geometry, usually this results in a more compact structure. Figure 13 shows a flowchart of the program *randstruct*. At the beginning a starting geometry and various command line options are read in and the energy of the starting geometry is calculated. In the next step the geometry is randomized following the procedure described above and an optimization is started using a very small number of iterations (in the range of 100 to 500).

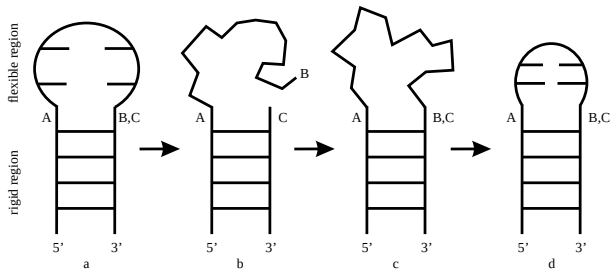


Figure 12: Schematic representation of an optimization using Randstruct.

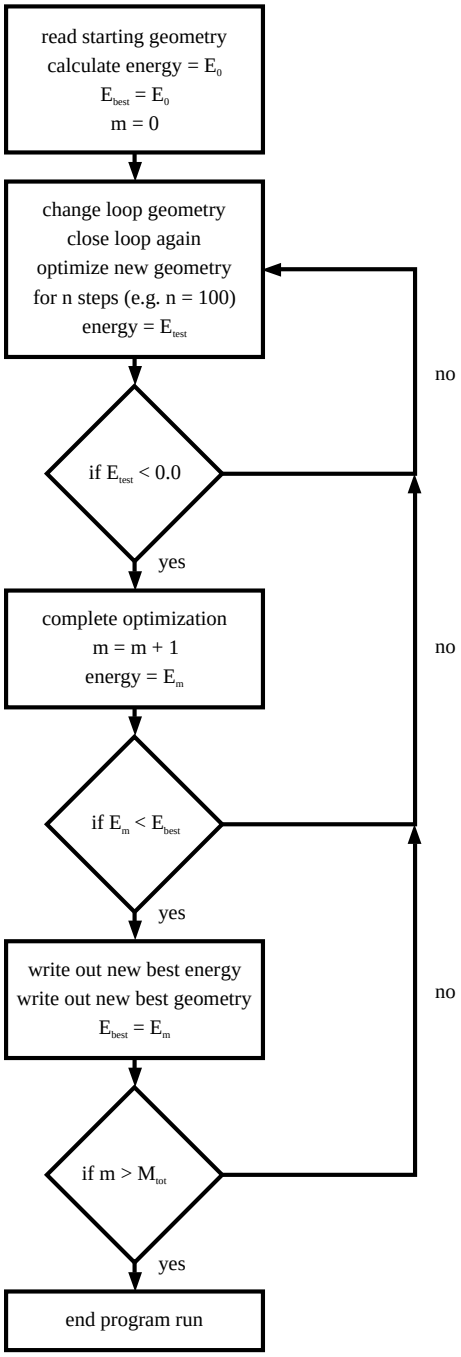


Figure 13: Flowchart of an optimization using Randstruct

Only if the energy after this short minimization is lower than a given value (also supplied via command line; in figure 13 this value is 0.0) optimization is continued until convergence. The energy after these few hundreds of iterations is used as a crude estimate for the minimum energy; experience has shown that if the total energy after a few hundred steps of optimization is not at least lower than 0.0 kcal/mol the underlying structure is usually not a "realistic" model for RNA molecules (in most cases these calculations won't converge at all) so that this estimate is a convenient way to save computational effort. When the optimization has converged the resulting minimum energy is compared to the energy of the starting conformation; the geometry corresponding to the lower energy is taken as starting point for the next randomizing run. The program ends after a given number of runs, writing out the best energy and the optimized geometry. Experience has shown that the best use of this program lies in the final optimization starting from already pre-optimized structures. The improvement in energy for small RNA molecules is usually in the range of 5 - 10% of the total energy, most of which is gained by formation of more compact structures.

3.2.5 GEN-3D

Last but not least another optimization technic a **G**enetic **A**lgorithm (GA) has been used, to obtain "sets of structures" for further minimization. The program has been written by the author and will be described in further detail.
In the 1950s and the 1960s several computer scientists independently studied evolutionary systems with the idea that evolution could be used as an optimization tool for engineering problems. The idea in all these systems was to evolve a population of candidate solutions to a given problem, using operators inspired by natural genetic variation and natural selection. In the 1960s, Rechenberg [127, 128] introduced "evolution strategies", a method he used to optimize real-valued parameters for devices such as airfoils. This idea was further developed by Schwefel [138]. The field of evolution strategies has remained an active area of research, mostly developing independently from the field of GAs. GAs were invented by John Holland in the 1960s [59]. In contrast with evolution strategies and evolutionary programming, Holland's original goal was not to design algorithms to solve specific problems, but rather to formally study the phenomenon of adaptation as it occurs in na-

ture and to develop ways in which the mechanisms of natural adaptation might be imported into computer systems. Given a clearly defined problem to be solved and a symbol string representation for candidate solutions, **a simple GA works as follows** (see figure 14):

1. Start with a randomly generated population of n 1-bit chromosomes (candidate solutions to a problem).
2. Calculate the fitness $f(x)$ of each chromosome x in the population.
3. Repeat the following steps until n offspring have been created:
   (a) Select a pair of parent chromosomes from the current population, the probability of selection being an increasing function of fitness. Selection is done "with replacement," meaning that the same chromosome can be selected more than once to become a parent (Gillespie-wheel [42]).
   (b) With probability pc (the "crossover probability" or "crossover rate"), cross over the pair at a randomly chosen point (chosen with uniform probability) to form two offspring. If no crossover takes place, form two offspring that are exact copies of their respective parents. (Note that here the crossover rate is defined to be the probability that two parents will cross over in a single point. There are also "multi-point crossover" versions of the GA in which the crossover rate for a pair of parents is the number of points at which a crossover takes place.)
   (c) Mutate the two offspring at each locus with probability pm (the mutation probability or mutation rate), and place the resulting chromosomes in the new population. If n is odd, one new population member can be discarded at random.
4. Replace the current population with the new population.
5. Go to step 2.

Each iteration of this process is called a generation. A GA is typically iterated for anywhere from 50 to 500 or more generations. The entire set of generations is called a run. At the end of a run there are often one or more

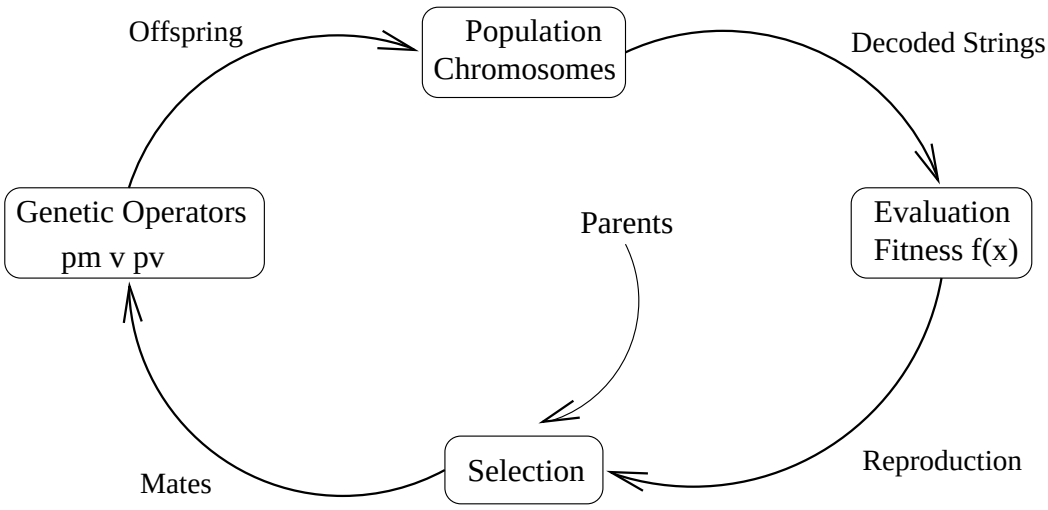


Figure 14: A simple GA.

highly fit chromosomes in the population. Since randomness plays a large role in each run, two runs with different random-number seeds will generally produce different detailed behaviors. GA researchers often report statistics (such as the best fitness found in a run and the generation at which the individual with that best fitness was discovered) averaged over many different runs of the GA on the same problem.
The simple procedure just described is the basis for most applications of GAs. There are a number of details to fill in, such as the size of the population and the probabilities of crossover and mutation, and the success of the algorithm often depends greatly on these details. There are also more complicated versions of GAs (e.g., GAs that work on representations other than strings or GAs that have different types of crossover and mutation operators).
The simplest form of GA involves three types of operators:

- Selection:
  This operator selects chromosomes in the population for reproduction. The fitter the chromosome, the more times it is likely to be selected to reproduce.

- Crossover:
  This operator randomly chooses a locus and exchanges the subsequences before and after that locus between two chromosomes to create two offspring. For example, the strings 10000100 and 11111111 could be crossed over after the third locus in each to produce the two offspring 10011111 and 11100100. The crossover operator roughly mimics biological recombination between two single-chromosome (haploid) organisms.

- Mutation
  This operator randomly flips some of the bits in a chromosome. For example, the string 00000100 might be mutated in its second position to yield 01000100. Mutation can occur at each bit position in a string with some probability, usually very small (e.g., 0.001).

In order for GAs to surpass their more traditional cousins in the quest for robustness, GAs must differ in some very fundamental ways. GAs are different from more common optimization and search procedures in four ways:

1. GAs work with a coding of the parameter set, not the parameters themselves.

2. GAs search from a population of points, not a single point, therefor the search is highly parallel.

3. GAs use payoff (objective function) information, not derivatives or other auxiliary knowledge.

4. GAs use probabilistic transition rules, not deterministic rules.

The evolutionary program created in this thesis has been inspired by Ogata's work [108]. It's a conformational search program based on a simple GA containing two genetic Operations (crossover and mutation). It is suitable for loop stem structures and has been used to obtain confomers for triloops and pentaloops. The most straightforward way to describe the 3D structure of RNA is obviously to list the three dimensional coordinates of each nucleotide, or even each atom. In principle, a GA could use such a representation, evolving vectors or coordinates to find a plausible structure. But, because of a number of difficulties involved (e.g., usual crossover operators would be too

likely to get physically impossible structures) another representation has been chosen, introduced by Schulze-Kremer [136] for Proteins. The conformation of a nucleotide is defined by seven variables (six angles for the backbone and one for the base). Each chromosome, representing a candidate structure with N nucleotides is a set of these variables (see. figure 15).

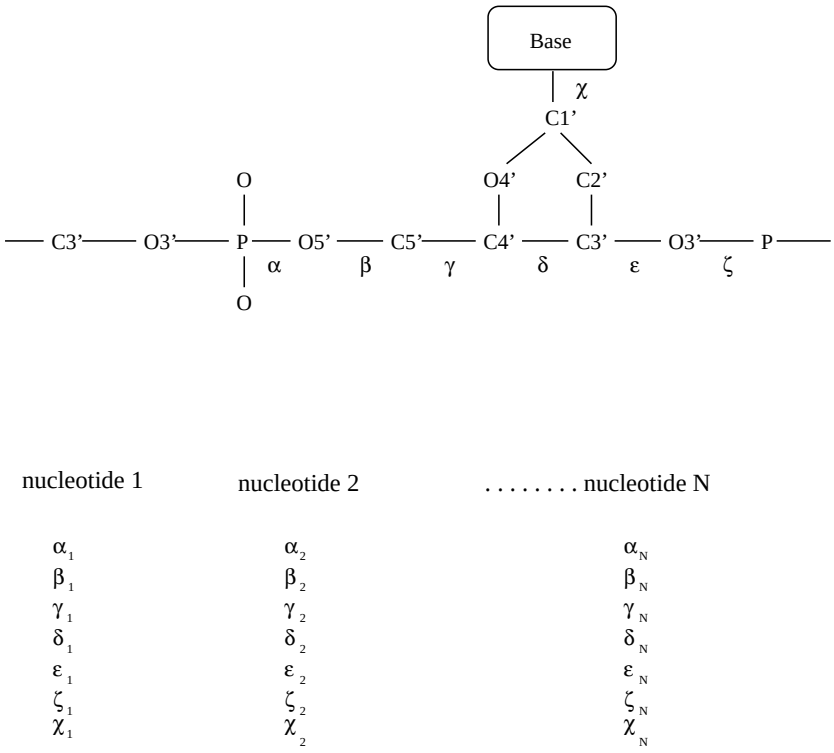


Figure 15: Schematic representation of encoding in GEN-3D.

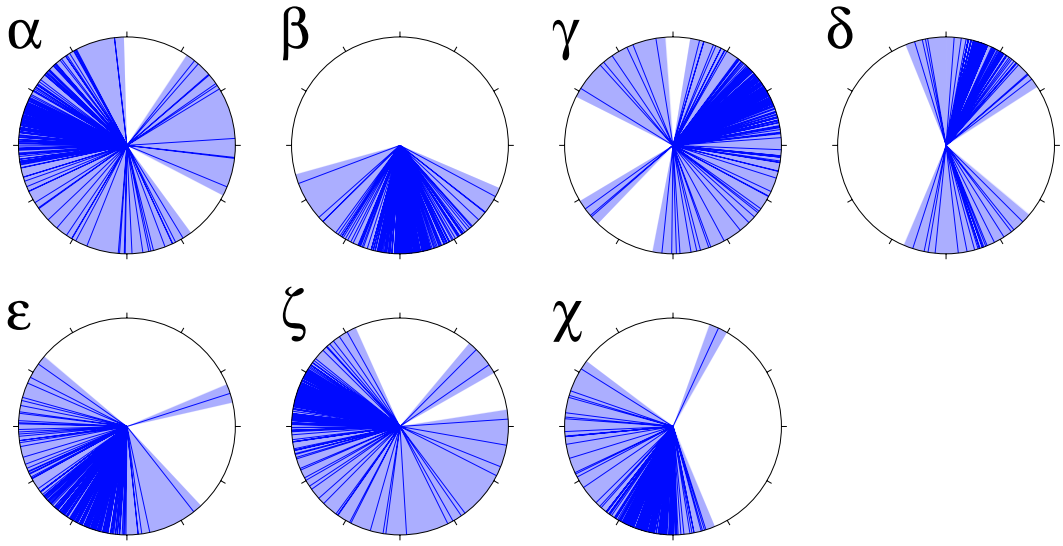


Figure 16: Used values for the seven variables in GEN-3D.

The variables can only take discrete values, obtained from known RNA structures (see figure 16), they are in good agreement to previously arrived data [133]. For the fitness function the AMBER-force field has been implemented. A flowchart-diagram of GEN-3D can be seen in figure 17. GEN-3D is used to obtain a set of starting conformations for further analyzation. Using this technic it was possible to find confomers with lower energies (approx 2-10%) than with usual methods. Details will be presented in the following chapters.

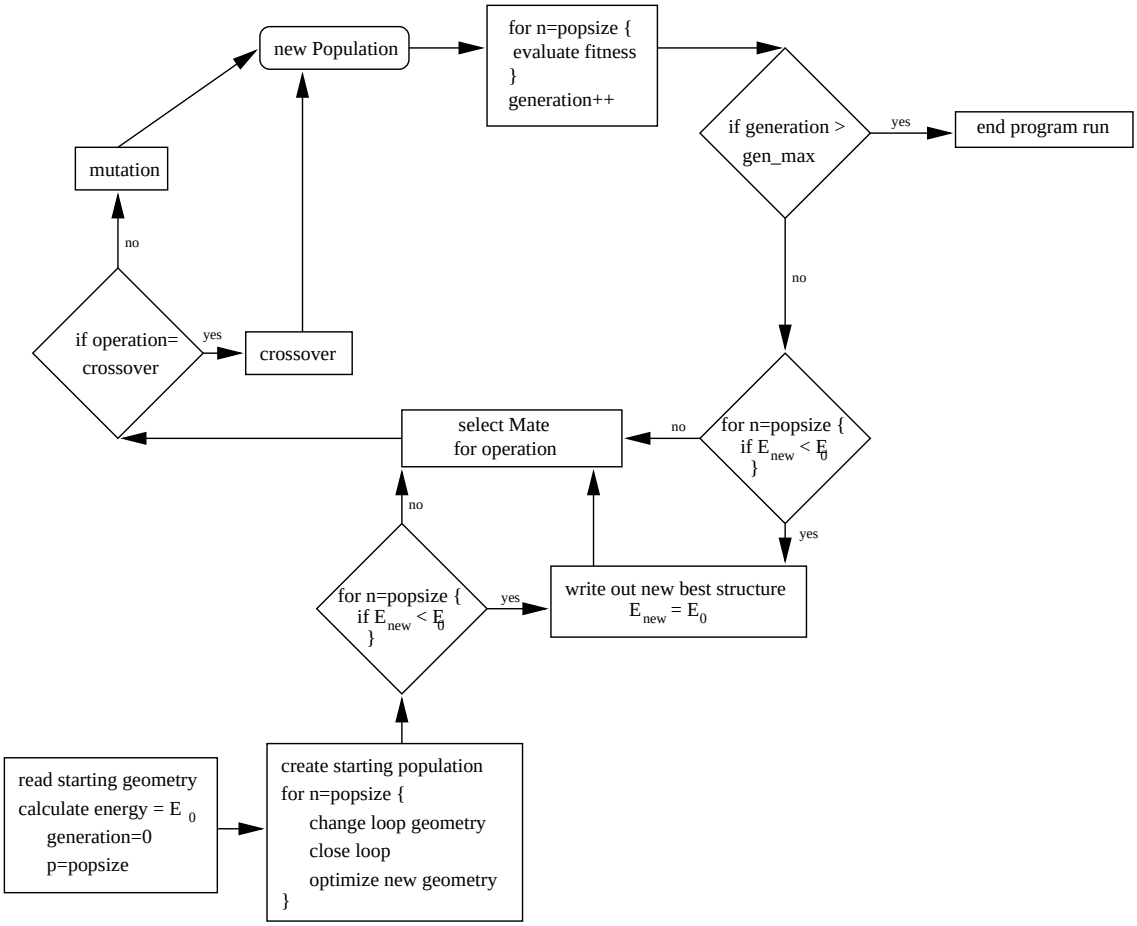


Figure 17: Schematic representation of GEN-3D.

3.3 Editing and Visualization of Molecules

The regular microscope stops where biology stops and chemistry starts. Although, eventually electron microscopes might be able to visualize molecules, currently computer graphics has to do the job. A picture is worth more than a thousand words. The pictures presented in this thesis has mainly been created using VMD:
VMD is designed for the visualization and analysis of biological systems such as proteins, nucleic acids, lipid bilayer assemblies, etc. It may be used to view more general molecules, as VMD can read standard Protein Data Bank (PDB) files and display the contained structure. VMD provides a wide variety of methods for rendering and coloring a molecule: simple points and lines, CPK spheres and cylinders, licorice bonds, backbone tubes and ribbons, cartoon drawings, and others. It can be obtained at the WEB ("http://www.ks.uiuc.edu/Research/vmd/").

4 Results

4.1 Performance of GEN-3D and it's Application

For structure prediction it is of crucial importance to know the CPU-time expected for calculations. A simple sample loop (see figure 18) has been chosen to measure the performance and the best set of genetic parameters. The loop and its properties has been investigated by Herbert Kratky [75]. It

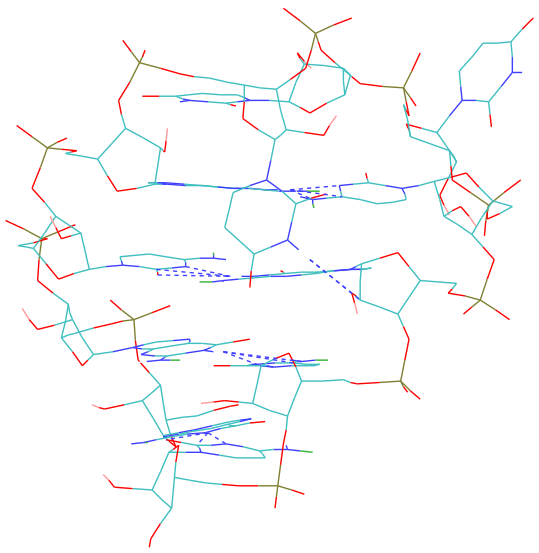


Figure 18: Starting triloop for GEN-3D.

is a triloop with the following sequence and structural features:

- 5'-GGCC*UUU*GGCC-3'
- the stem structure corresponds to the standard RNA-A-helix
- the first base in the loop lies parallel to the axis of the stem, where it is stabilized by additional hydrogen bonds.
- the second nucleotide in the loop is stacking on the closing pair on the 5'-side of the stem.

- the position of the third base is given by steric constraints of the two other nucleotides which are in energetically favorable positions.

The biggest caveat of GEN-3D algorithm is the fitness function, it's an implementation of the AMBER force field. The most time consuming part is the execution of the AMBER-force-field. In figure 19 we notice: The CPU time is strongly correlated to the number of AMBER-iterations per optimization I of each individual-structure. At a certain point ($\approx$ at 200 AMBER iterations) we loose a linear relationship. This could be due to the fact, that all the essential arrays in the force field are build up and ready for further iterations.

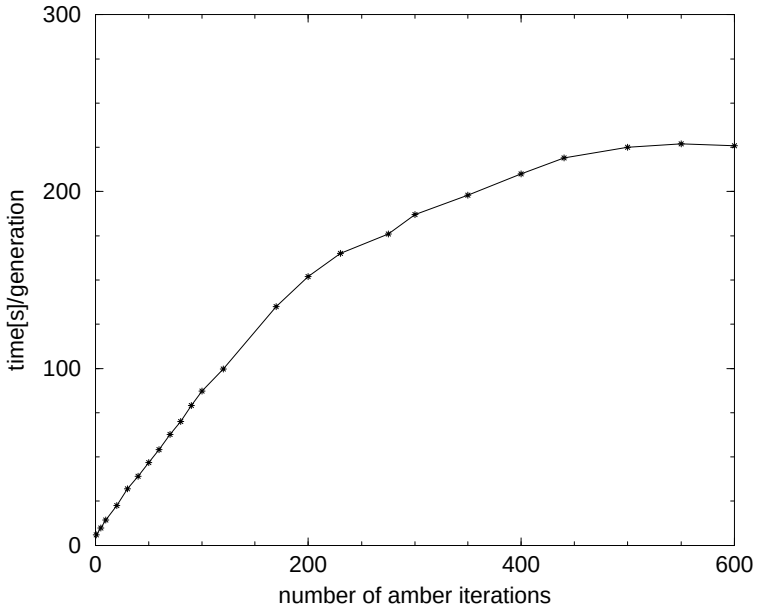


Figure 19: Performance data of GEN-3D.

Another essential task is the optimization of genetic parameters (e.g., Population size P, Mutation rate p_m, Crossover rate p_c and number of generations G). For this purpose a set of runs have been investigated. The primary goal has been to achieve a "fitter" structure (e.g., a structure with less energy E_G according to the AMBER force field) within a reasonable amount of time. To be able to compare the resulting energies, the resulting populations of the GA have been optimized completely in the AMBER-force field. The final energy is denotated with E_F. Note the geometry of the stem is **not** touched. In tables 2-4 the results are presented to yield the best parameters. The target structure had an energy of -72.38 kcal/mol before the GA. Each table represents a set of runs for a different number of iterations. The best results are obtained using 30 iterations and a high mutation rate p_m. A lower crossover rate p_c is also of advantage. This can be observed in nearly all runs. We also note that 30 AMBER iterations are enough to obtain satisfactory results. Using 100 iterations or big populations will result in low energy structures as well, however the computational time needed will exhaust any human. So the best results are achieved using the following set of parameters: All of

G	=	1000
I	=	30
P	=	30
p_m	=	0.5
p_c	=	0.01

Table 1: Best set of genetic parameters.

this calculations have been done on an SGI-workstation with an RS-10000 processor and 1 Gigabyte of Memory. The memory requirement is not essential, for the optimization of a triloop structure with the best suited genetic parameters the program needs about 19 Mbyte of RAM.

N. of I	N. of P	p_m	p_c	E_G [kcal/mol]	E_F [kcal/mol]
5	10	0.01	0.01	10.3	-55.4
5	10	0.1	0.01	8.3	-54.2
5	10	0.5	0.01	7.3	-31.4
5	10	0.01	0.1	21.4	-44.2
5	10	0.1	0.1	12.2	-43.1
5	10	0.5	0.1	15.7	-23.4
5	10	0.01	0.5	22.0	-48.9
5	10	0.1	0.5	15.7	-33.2
5	10	0.5	0.5	12.6	-21.3
5	30	0.01	0.01	7.5	-54.0
5	30	0.1	0.01	4.5	-51.9
5	30	0.5	0.01	3.6	-21.6
5	30	0.01	0.1	12.5	-44.1
5	30	0.1	0.1	10.5	-33.6
5	30	0.5	0.1	12.3	-33.5
5	30	0.01	0.5	17.3	-47.9
5	30	0.1	0.5	14.2	-36.1
5	30	0.5	0.5	12.7	-25.7
5	70	0.01	0.01	8.2	-51.3
5	70	0.1	0.01	5.3	-34.7
5	70	0.5	0.01	4.5	-41.2
5	70	0.01	0.1	13.0	-49.3
5	70	0.1	0.1	9.5	-45.2
5	70	0.5	0.1	7.5	-33.1
5	70	0.01	0.5	16.2	-50.1
5	70	0.1	0.5	12.2	-48.7
5	70	0.5	0.5	15.3	-32.1

Table 2: Genetic parameters (all values for G=1000).

N. of I	N. of P	p_m	p_c	E_G [kcal/mol]	E_F [kcal/mol]
30	10	0.01	0.01	-45.5	-69.4
30	10	0.1	0.01	-43.3	-64.3
30	10	0.5	0.01	-46.2	-61.7
30	10	0.01	0.1	-32.4	-65.3
30	10	0.1	0.1	-31.5	-63.7
30	10	0.5	0.1	-30.4	-62.1
30	10	0.01	0.5	-24.5	-58.1
30	10	0.1	0.5	-21.4	-49.1
30	10	0.5	0.5	-22.5	-50.7
30	30	0.01	0.01	-46.5	-72.9
30	30	0.1	0.01	-42.2	-73.1
30	30	0.5	0.01	-44.1	**-74.9**
30	30	0.01	0.1	-39.3	-66.5
30	30	0.1	0.1	-35.5	-67.2
30	30	0.5	0.1	-34.5	-64.7
30	30	0.01	0.5	-29.6	-67.9
30	30	0.1	0.5	-28.4	-56.8
30	30	0.5	0.5	-24.7	-59.8
30	70	0.01	0.01	-47.1	-70.4
30	70	0.1	0.01	-41.3	-70.2
30	70	0.5	0.01	-44.2	-68.8
30	70	0.01	0.1	-38.4	-67.8
30	70	0.1	0.1	-34.5	-65.1
30	70	0.5	0.1	-33.5	-65.0
30	70	0.01	0.5	-22.5	-66.3
30	70	0.1	0.5	-23.4	-64.2
30	70	0.5	0.5	-21.3	-62.1

Table 3: Genetic parameters (all values for G=1000).

N. of I	N. of P	p_m	p_c	E_G [kcal/mol]	E_F [kcal/mol]
100	10	0.01	0.01	-41.4	-70.4
100	10	0.1	0.01	-40.2	-64.8
100	10	0.5	0.01	-44.2	-65.1
100	10	0.01	0.1	-31.2	-67.6
100	10	0.1	0.1	-33.5	-63.2
100	10	0.5	0.1	-34.7	-61.4
100	10	0.01	0.5	-22.3	-54.7
100	10	0.1	0.5	-24.5	-59.3
100	10	0.5	0.5	-21.7	-52.0
100	30	0.01	0.01	-44.3	-71.8
100	30	0.1	0.01	-43.5	-72.4
100	30	0.5	0.01	-42.4	-72.6
100	30	0.01	0.1	-34.3	-69.5
100	30	0.1	0.1	-36.5	-68.1
100	30	0.5	0.1	-32.1	-62.8
100	30	0.01	0.5	-27.3	-66.4
100	30	0.1	0.5	-27.2	-53.7
100	30	0.5	0.5	-25.9	-58.7
100	70	0.01	0.01	-49.1	-71.3
100	70	0.1	0.01	-45.4	-70.8
100	70	0.5	0.01	-43.5	-67.3
100	70	0.01	0.1	-39.2	-65.7
100	70	0.1	0.1	-35.1	-65.8
100	70	0.5	0.1	-36.2	-63.2
100	70	0.01	0.5	-21.6	-67.1
100	70	0.1	0.5	-22.7	-63.0
100	70	0.5	0.5	-23.1	-61.8

Table 4: Genetic parameters (all values for G=1000).

4.1.1 Optimizing Structures with GEN-3D

RNA secondary structure can be classified by the use of a small number of motifs, as mentioned above. The most common among these motifs is the so-called hairpin structure consisting of a single-stranded loop and a double-stranded stem. Hairpin loops occur in many sizes from 3 up to 9 or 10 nucleotides in the unpaired region. Triloops (three-membered loops) are therefore the smallest loop size that exist in nature and they are good candidates for a structural investigation for a variety of reasons. Though they are not as common as tetraloops, tri-nucleotide loops occur in bacterial as well as eukaryotic 16S-RNA [171, 172] where they replace their more abundant four-membered relatives [171]. The energy difference between the three- and the four-membered hairpins is of the order of 1 kcal/mol in favor of the tetraloops [44]. Their small size and the high rigidity makes them an ideal starting point for experimental as well as computational approaches. So far very few studies have been done on triloops, among them the rCGC(UUU)GCG- [27] and the rGCGAUU(UCU)GACCGCC-hairpin [120]. Both investigations presented solution structures of the respective molecules obtained by multi-dimensional Proton-NMR and they agreed on a stem structure that was close to the RNA-A-helix while the structure of the loop regions could not be clarified satisfyingly.
As described in the previous chapter GEN-3D allows to optimize loop-stem structures. In this section an example will be given. The whole procedure of optimization can bee seen in figure 30. During this optimization the genetic parameters of table 1 have been used.
In figure 20 we can see the energies of the best structures of each population, in which a fitter individual has arised. *E before* denotes the best energy before GEN-3D has been used. In the small picture the energy of the "fittest" structure during the whole simulation is depicted. Note, here we only have 30 AMBER-iterations, so that the energies are not in the range of *E before*. Figure 21 depicts the best structures of the last 8 generations. As we can see the stem has not been modified, but a variety of loop conformations have arised.

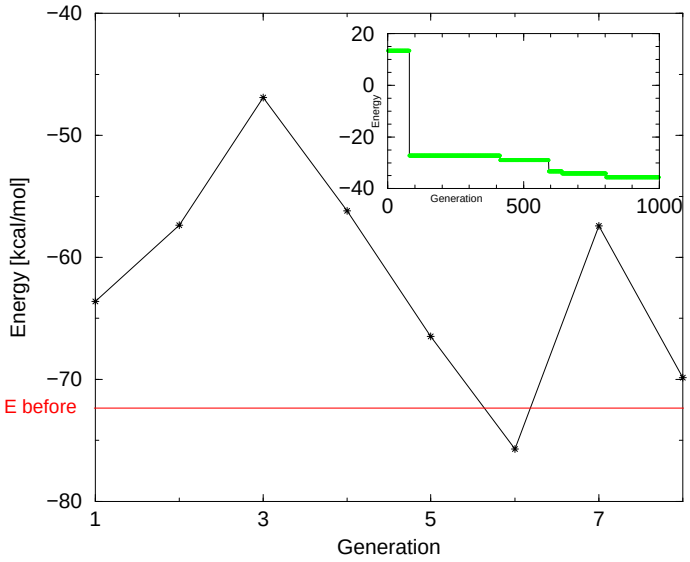


Figure 20: Best structures of the last 8 generations, in which a fitter individual has arised.

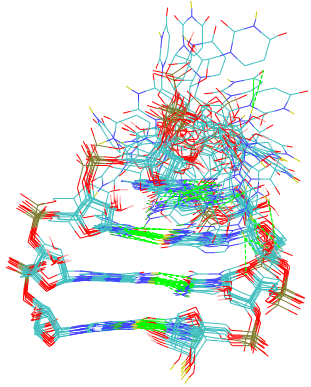


Figure 21: All best structures of the last 8 generations.

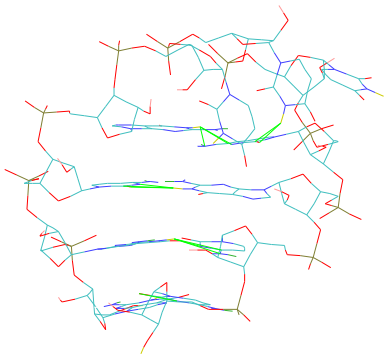


Figure 22: Generation 1.

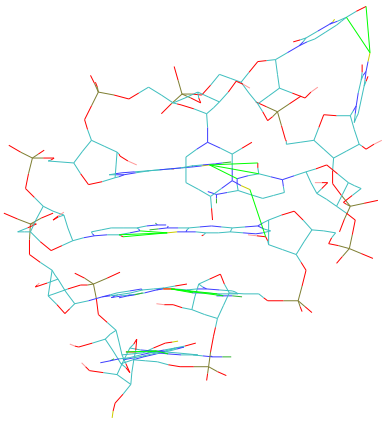


Figure 23: Generation 2.

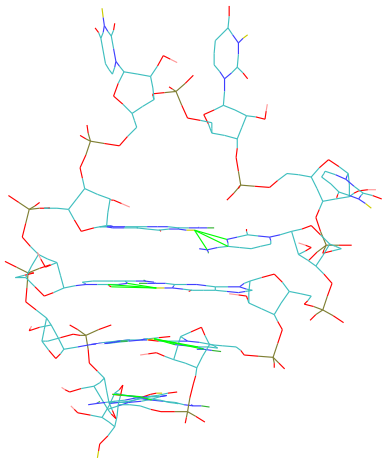


Figure 24: Generation 3.

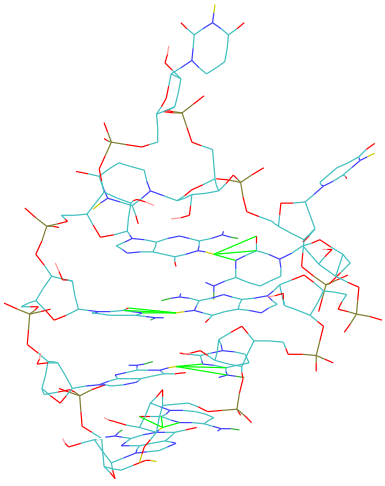


Figure 25: Generation 4.

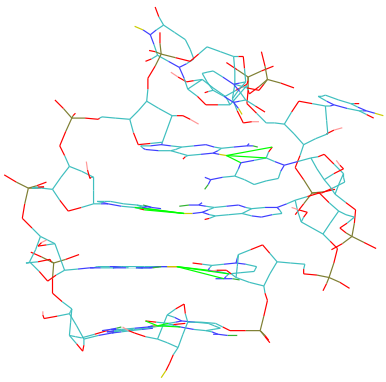


Figure 26: Generation 5.

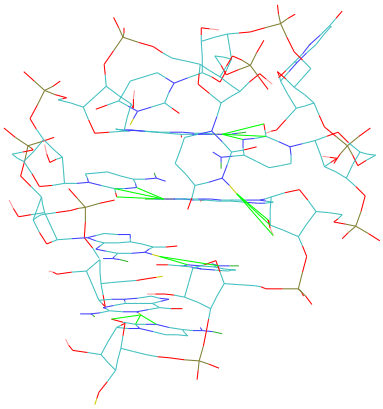


Figure 27: Generation 6.

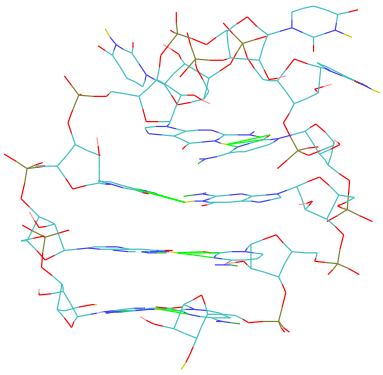


Figure 28: Generation 7.

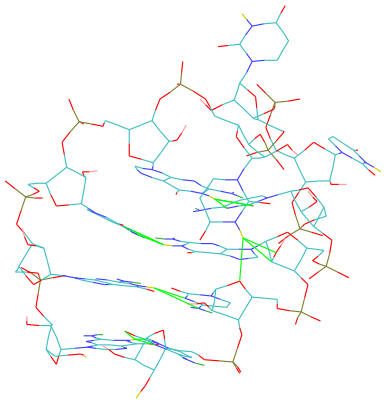


Figure 29: Generation 8.

In figures 22-29 the "fittest" structures of the last 8 generations are shown of a GEN-3D run. In Generation 6 we yield a structure with lower energy, than found in previous investigations. Possible H-bonds are represented with green lines. They will now be described in further detail.

- The fittest structure of generation 1 has an RMSD of 2.65 Å compared to the "start structure". It has quite a high energy and is not a suitable candidate for a possible structure. Neither of the two criteria, such as stacking of the loop base U6 or possible H-bond interaction between base U8 and the stem is fulfilled.

- The fittest structure of generation 2 has an RMSD of 2.47 Å compared to the "start structure". It has a very distorted geometry and possible H-bond interaction within the loop.

- The fittest structure of generation 3 has an RMSD' of 3.57 Å compared to the "start structure". It is the "worst" structure in this series and can be seen as an intermediate structure to gain to more interesting candidates.

- The fittest structure of generation 4 has an RMSD of 2.59 Å compared to the "start structure". Still the loop bases have whether stacking nor hydrogen-bonding.

- The fittest structure of generation 5 has an RMSD of 2.04 Å compared to the "start structure". This structure is more compact and the second nucleotide is stacked on the closing pair of the 5'-side of the stem.

- The fittest structure of generation 6 has an RMSD of 1.22 Å compared to the "start structure". Here we got a new "fittest" structure with approx. 5 % lower energy, than found in previous investigations [75]. The structure is no surprise, it has both the stacking and the H-bond. In addition it is quite compact.

- The fittest structure of generation 7 has an RMSD of 2.56 Å compared to the "start structure". Again we drift away from structures containing promising features.

- The fittest structure of generation 8 has an RMSD of 2.45 Å compared to the "start structure". In this last structure the population has totally drifted away.

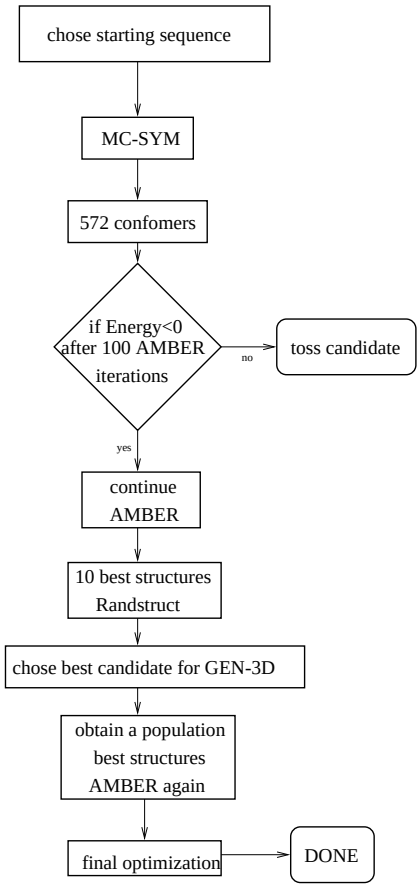


Figure 30: Flowchart of triloop optimization.

4.2 Tetraloops

In this section another optmization technic is introduced. A method of conformational search has been applied. It is suitable even in the absence of experimental constraints. Such a merely theoretical approach to predicting molecular structures of biological interest, is presently hampered by several problems and therefore, the results obtained by such techniques have been met with some scepticism. The difficulties in treating large molecular systems are twofold:
The first is due to the tremendous number of conformational states which must be included in a complete conformational analysis. The second problem is closely related to global minimization in a high-dimensional multiminimum landscape. The progress achieved in this field is documented, in particular, by successful applications to searching the conformational space of oligopeptides. The problem of the large search space can be surmounted by using the experimental knowledge on conformational preferences in similar systems, however one still ends up with a large set of different structures where only a few of them are relevant for the real system. Thus, the choice of the force field model is critically important for using conformational search as a predictive tool. The predicted relative stabilities of different nucleic acid conformers in solution depend, in particular, on the description of solvent electrostatic effects.
In the search for the origin of the unusual stability of GNRA hairpins the analogous RNA molecules with GNYA loop sequences were included in the study, using a different approach. A comparison of the optimized geometries of all GNNA loop sequences shows surprisingly little variation of the overall structures of G–A base pair geometry [40]. Releasing the strain introduced by chain closure in NR and NY dinucleotides and re-optimizing the open structures, however, makes evident that GNRA loops have higher internal stability than GNYA loops.

4.2.1 Conformational Search

To carry out an extensive conformational search the program MC-SYM [92] was used to generate starting structures for the tetraloop sequences GAAA, GCAA and GCUA. This restriction was necessary because of the tremendous amount of computation time needed for the conformational search. For

these three sequences MC-SYM tested 810.000 different conformations respectively and yielded more than 4000 geometries that were optimized using the JUMNA program [79] with the FLEX force field [83].
JUMNA was designed specifically to build, manipulate and optimize nucleic acid structures. The main differences between JUMNA and "conventional" force field programs lies in the internal representation of the molecule. Here the basic idea is to split the nucleic acid fragment into a collection of 3′-monophosphate nucleotides. These nucleotides are then positioned with respect to a helical axis using the six helicoidal parameters (x displacement, y displacement, rise, inclination, tip, and twist) [82, 150] obeying the Cambridge conventions [28]. A reduced set of parameters (glycosidic angle, sugar pucker, and the two backbone torsion angles epsilon and zeta) describe the internal conformation of the nucleotides. During structural optimization constraints for the O5′-C5′ bond distance and associated valence angles are used to fulfill the chain closure condition.
By modeling nucleic acid structures in this way the number of variables for an optimization is strongly reduced (in total there are only 14 variables per nucleotide unit including the dihedral angles of the backbone, the sugar pucker, and bond angles in the sugar ring) and internal angle variations are applied locally, acting on neighboring units only via the constraints. The representation of the structure is completely independent from the underlying sequence, so that a systematic variation of sequences is fairly simple. Furthermore, the JUMNA concept allows control of the molecular geometry in a versatile manner. This feature has been used for the construction of loop structures with different sugar puckers, and the possibility to release the chain closure conditions in different loop positions has enabled us to quantify the sequence dependent conformational stress associated with loop formation.
The 74 best (i.e. lowest in energy) conformations were then put into JUMNA, optimized again using the AMBER force field [110, 165] and sorted according to their energy. From the structures obtained, a subset of altogether 74 confomers with lowest energy was selected which has been used for the final round of conformational analysis. This subset contains 26 confomers with a G-A base pair closing the loop. The conformational energies calculated without taking into account the electrostatic RF (Reaction Field) at this stage of the procedure, however do not show a clear preference for confomers which are close to the structures derived from NMR experiments. Since representation of nucleic acid structures in JUMNA is independent from the

underlying sequences, it is fairly simple to use the whole subset of the low-energy confomers as starting structures for all GNRA sequences. In this step, the AMBER force field was used for energy minimization and the RF contributions were calculated and added to the final energies for all structures obtained for each of the sequences.

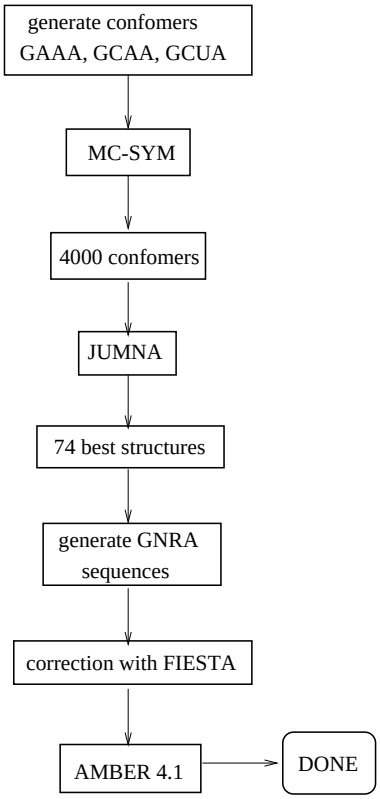


Figure 31: Flowchart of tetraloop optimization.

Eleven conformers lie within an energy range of two kcal/mole compared to the lowest energy structure for each sequence and were taken for structural analysis. In table 5 the energies of the nine confomers with common stacking pattern (as described by Jucker and Pardi [64]) are shown. The remaining two structures (613 and 576) show differences in the position of A8 which lies in the major groove and so didn't fulfill the results of NMR-studies about

ID	GAAA	GGAA	GCAA	GUAA	GAGA	GGGA	GCGA	GUGA
508	0.7	0.9	1.5	2.4	**0.0**	**0.0**	**0.0**	**0.0**
418	0.1	1.2	2.6	1.7	2.2	2.2	2.6	2.3
419	0.1	**0.0**	2.0	2.2	2.0	1.1	2.8	2.8
445	0.8	2.1	1.8	2.5	2.9	3.0	3.2	2.9
404	**0.0**	0.7	0.7	1.0	2.0	1.7	1.7	1.2
799	2.7	2.0	1.9	2.3	1.4	0.8	2.5	1.6
857	0.3	0.2	**0.0**	0.7	3.0	1.7	2.2	1.6
838	1.2	1.5	1.7	1.9	3.3	2.6	2.8	2.1
576	2.15	0.8	**0.0**	**0.0**	7.2	5.3	2.0	2.7

Table 5: Energies of the GNRA confomers.

GNRA tetraloops [55, 109]. Those nine structures have differences in the backbone structure. The first and the last residues of the tetraloop form a G5-A8 base pair on top of the four Watson-Crick base pairs in the stem. In figure 32 the G-A bp of confomer 404 with sequence GAAA (i.e, the structure with the minimum energy) is depicted. In conformer 576 A8 forms hydrogen bonds with residues within the stem region (i.e. Gl and C10) and the sugar pucker of the residue A8 is of the S type for all eight tetraloop sequences (see figure 33. In the following part of this work the ten structures containing the G5-A8 base pair will be described in more detail.
There are some structural features in all the received energy minimum structures independently of their sequence. One of the most interesting feature of the GNRA loops is the presence of an unusual G5-A8 base pair. A hydrogen bond is formed between one of the G5 amino protons and N7 of the residue A8. Beside this hydrogen bond conformer 445 with the tetraloop sequences GAGA and GCGA has a second hydrogen bond between one of the A8 amino protons and N3 on G5, though this hydrogen bond isn't a very strong one with a length of 2.67 Å (for GAGA) and 2.75 Å (for GCGA). For all other conformers the distances between the A8 amino protons and N3 on G5 are greater than 3 Å and so this second hydrogen bond cannot be formed. For further stabilization there are additional hydrogen bonds between one of the G5 amino protons and the A7pA8 phosphate oxygen in all conformations and one between the H02' on G5 and N7 on residue R7 in some conformers (508

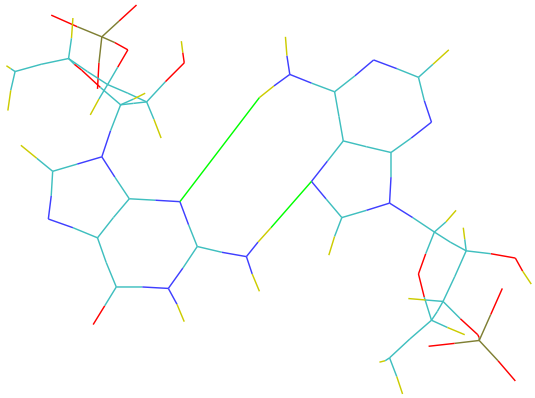


Figure 32: G-A base pair of GAAA 404.

for all GNRA sequences, 857 for the GNGA tetraloops). If all structures are compared it can be seen that the stacking pattern is common for all GNRA sequences. The stacking between the stem and the loop is very good. The G5A8 pair stacks on the closing pair of the stem and A8 lies directly over the Cl' proton on G9. The nucleotides N6 and R7 are positioned in such a way that stacking is continued from the 3'-end of the stem. There is a sharp bend between the first and the second nucleotide in the GNRA tetraloops. The Xdisplacement, Ydisplacement and inclination values for, the bases N6, R7 and A8 within the loop region differ from these values for the stem region. Furthermore in most confomers the tilt and/or roll angles concerning the A8/G9 base step are significant different compared to these values for the other steps. The three bases after the sharp bend are positioned to form tertiary interactions (involving their Watson-Crick faces). The most noticeable torsion angle-pattern is the $\alpha\gamma$-flip where the α torsion angle is +ac or +ap instead of -sc within a residue. The analysis of the torsion angles shows that for all conformers the G5 α torsion angle is +ac or +ap instead of the normal -sc but there is no $\alpha\gamma$-flip observed for the residue G5. The conformer 508 has a change of the G9-torsion angle from trans to gauche as it was found in the NMR study by Heus and Pardi [55] and so 508 meets most of the structural features derived with NMR techniques. Concerning the sugar conformation one can see that all the nucleotides adopt the Ntype (C3'-endo and

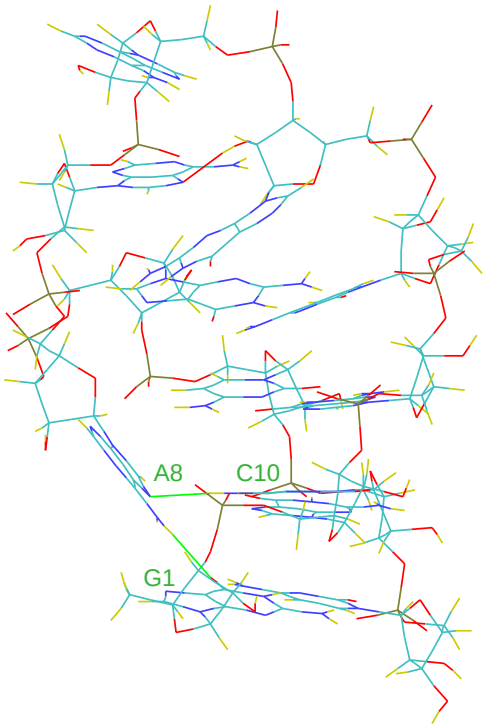


Figure 33: H-bonds of GAAA 576.

C2'-exo) pucker found in A-form helices. The 60% S-type (C2'endo) sugar pucker for the residues N6 and R7 of the tetraloops GCAA and GAAA, mentioned in [55], cannot be found in the conformers of this study. The sugar puckers within the GAGA structures are in good agreement with the results of the NMR studies [109], though a mixture of N- and S-type sugar puckers for G7 isn't found in the GAGA conformers described here. The glycosidic angle conformation for all nucleotides is anti. The overall structure of the tetraloops isn't effected on the variation of the middle bases. It seems to be that the formation of the G5-A8 pair is the important step for the stabilization of the molecule whereas the nature of the middle bases (N6 and R7) doesn't play such an important role due to the fact that they didn't interact with other parts of the molecule. In table 6 the RMSD values between the

confomers are shown, as reference the confomer with the lowest energy have been taken for each sequence. It is not surprising, that confomer 576 has the largest RMS deviation compared to the other structures, the reasons have been given. To calculate one or more measures for the structure dissimilarity,

ID	GAAA	GGAA	GCAA	GUAA	GAGA	GGGA	GCGA	GUGA
508	0.887	0.946	1.166	3.234	0.000	0.000	0.000	0.000
418	0.325	0.665	1.122	3.262	0.517	0.692	0.517	0.692
419	0.775	0.000	0.746	3.073	0.959	0.892	0.996	0.510
445	0.343	0.681	1.123	3.258	0.526	0.693	0.527	0.525
404	0.000	0.751	1.175	3.300	0.655	0.792	0.527	0.692
799	1.367	0.880	0.405	2.767	1.055	0.950	1.378	1.224
857	1.191	0.780	0.000	2.857	1.334	1.183	1.362	1.286
838	0.663	0.361	0.648	3.053	0.937	0.958	0.958	0.918
576	3.133	3.086	2.850	0.000	3.200	3.066	3.408	3.127

Table 6: RMSD values of the GNRA confomers.

based on RMSD values shown in table 6 tree editing or string alignment algorithms have beens used (e.g., Shapiros- and Ward- method [144, 143]). As both trees have similar branches and roots, we can assume that the resulting families are significant. We can determine 2 big families:

- Family one consists out of confomers 419, 799, 838 and 857
- Family two consists out of confomers 404, 418, 445 and 508

Confomer 576 is as expected on a totally different branch. The same results are obtained using all GNRA sequences investigated.

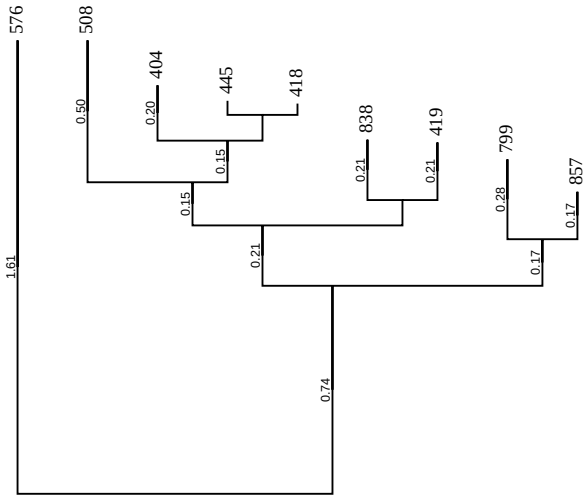


Figure 34: GAAA Shapiros method.

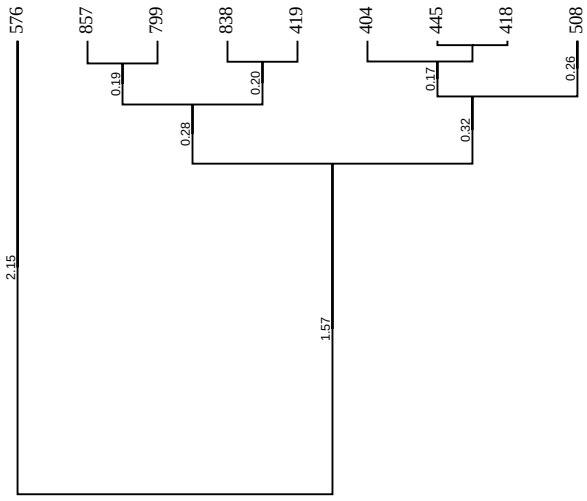


Figure 35: GAAA Wards method.

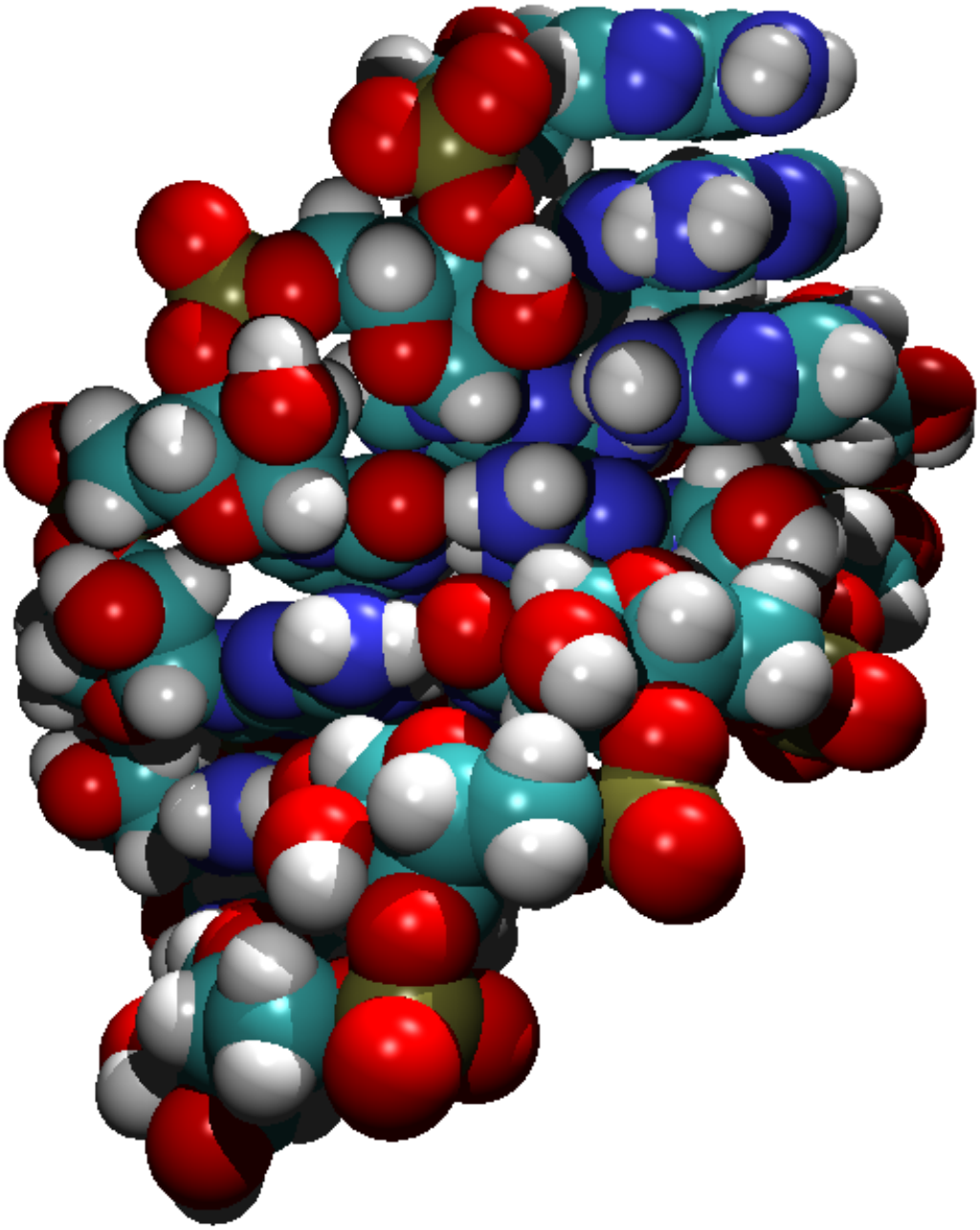


Figure 36: CPK-model of the optimized GAAA structure. The stacking between the bases A6, A7 and A8 is easy to recognize.

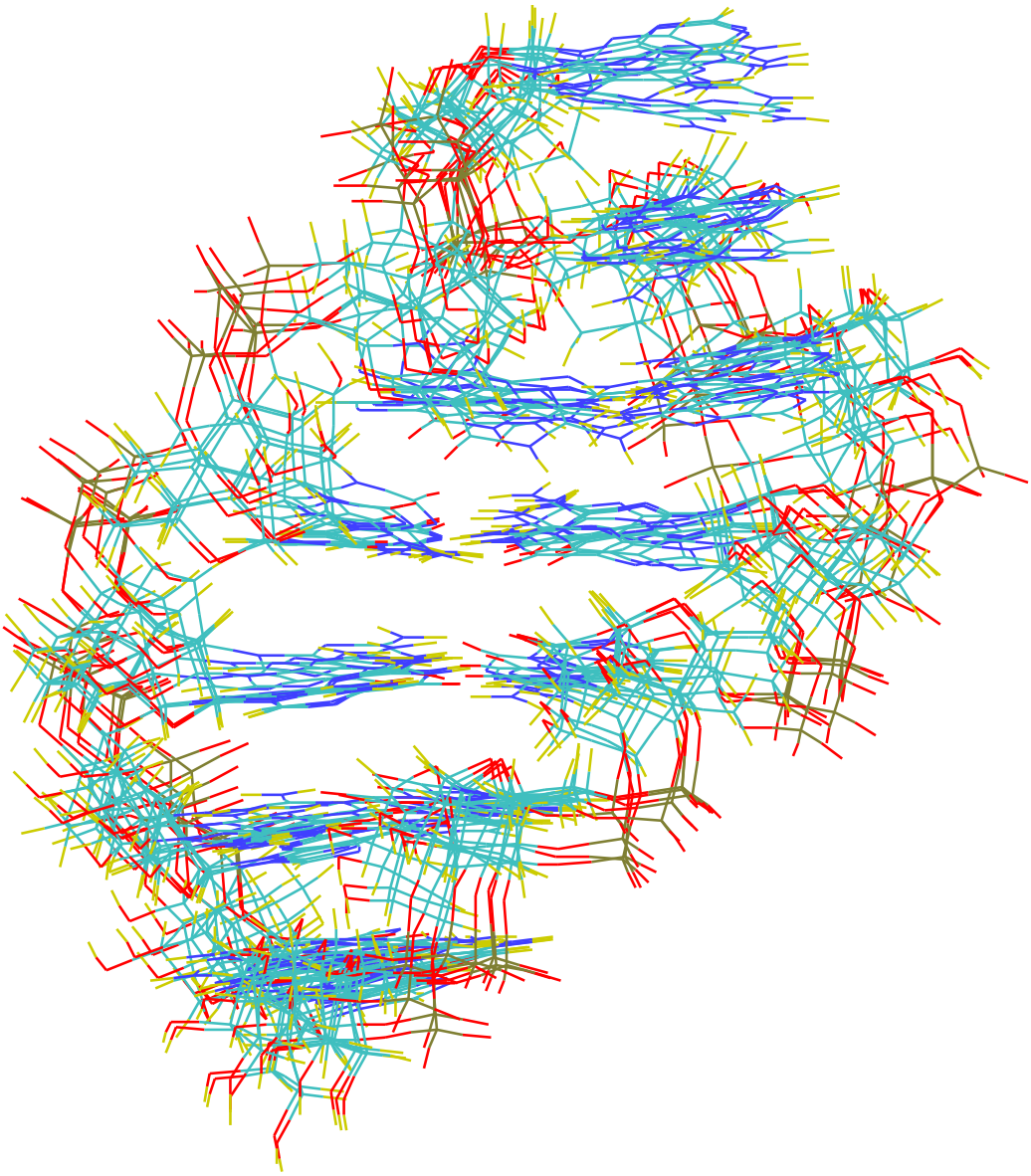


Figure 37: All GAAA-confomers except confomer 576. The data is described in the text and in good agreement with the literature.

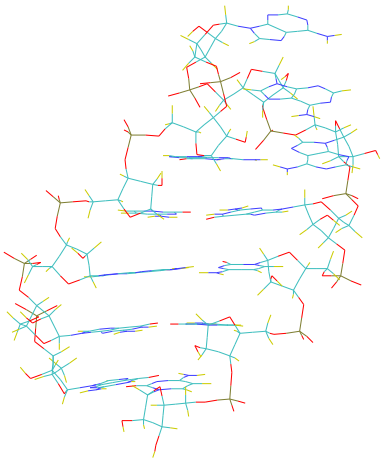


Figure 38: GAAA confomer 404.

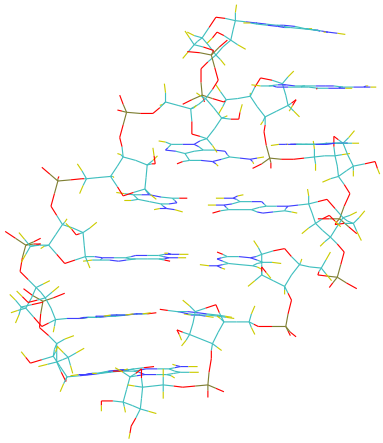


Figure 39: GGGA confomer 419.

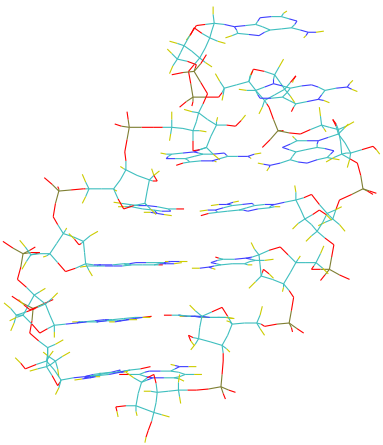


Figure 40: GAGA confomer 508.

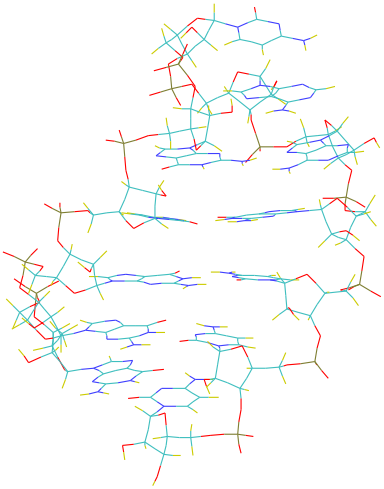


Figure 41: GCAA confomer 857.

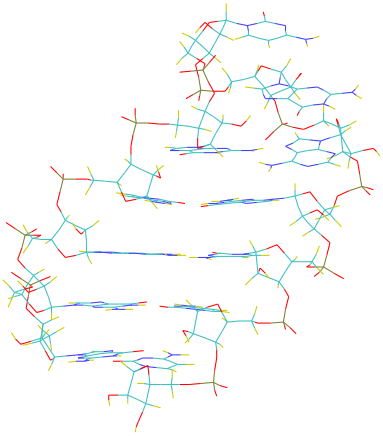


Figure 42: GCGA confomer 508.

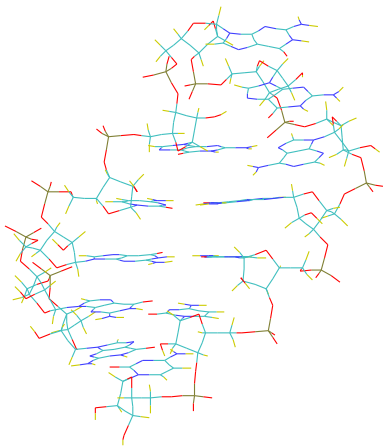


Figure 43: GGGA confomer 508.

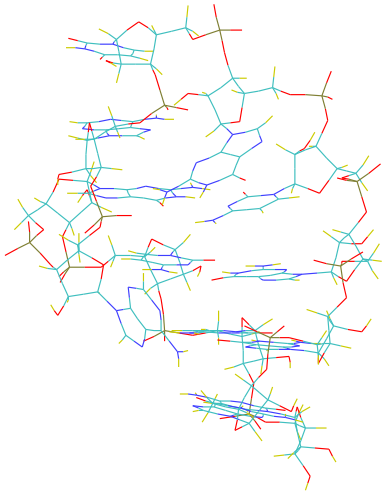


Figure 44: GUAA confomer 576.

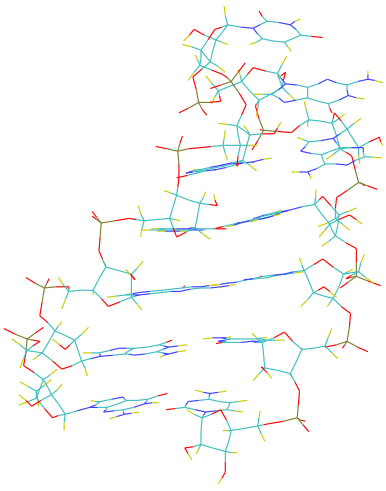


Figure 45: GUGA confomer 508.

All conformers used for conformational search were obtained **without** using any constraints (with exception of using an A-form helix as starting structure for the stem). These results are in strong agreement with the data presented in [55]. The overall dependence of the structure on a variation of the middle bases is surprisingly low, even in cases were pyrimidine bases are replaced with purine. This of course is probably due to the fact that the formation of the G–A pair has the most prominent stabilizing effect on the molecule and that the middle bases are forced in a position where there is little chance of interaction with other parts of the molecule, and thus steric interactions do not seem to be of significant magnitude.

4.2.2 GNRA Stability

A different approach has been used to determine if GNRA loops have higher internal stability than the GNYA loops: Starting from data derived from Heus and Pardis' [55] the molecular graphics program QUANTA (Molecular Simulations Inc.) was used to construct the molecule in such a way that the stem adopted standard A-RNA-helix conformation and the steric requirements to enable base pairing between G5 and A8 were roughly met. This structure was optimized to relax any close contacts that might have occurred during the modeling process. Subsequently the sugars belonging to the four bases in the loop were forced to adopt C3′-endo, O1′-endo, and C2′-endo-conformation by using appropriate constraints for the amplitude and pseudo-rotation angle. The systematic construction of all possible combinations yields $3^4 = 81$ different starting conformations which were all energy minimized. This choice does not mean a full conformational search, but is guided by the experience, that most substates are reachable by energy minimization of structures with different sugar puckers. The five conformations with the best (i.e. lowest) energy were chosen for further investigation. They are denoted in the following with A, B, C, D, and E, in order of increasing energy. The sequences used for these calculations were GGGC(GNNA)GCCU, studied in [55] with just the dangling end on the 3′-end omitted. To study sequence effects on the stability of the loop we have released the chain closure conditions in different positions of the loop and re-optimized these open structures. The energy differences obtained can be considered as a measure of the conformational stress due to loop formation.
Figure 46 shows the stabilization energies, it is the energy difference with and

without loop closure, for each structure type (A – E) and each sequence. Since

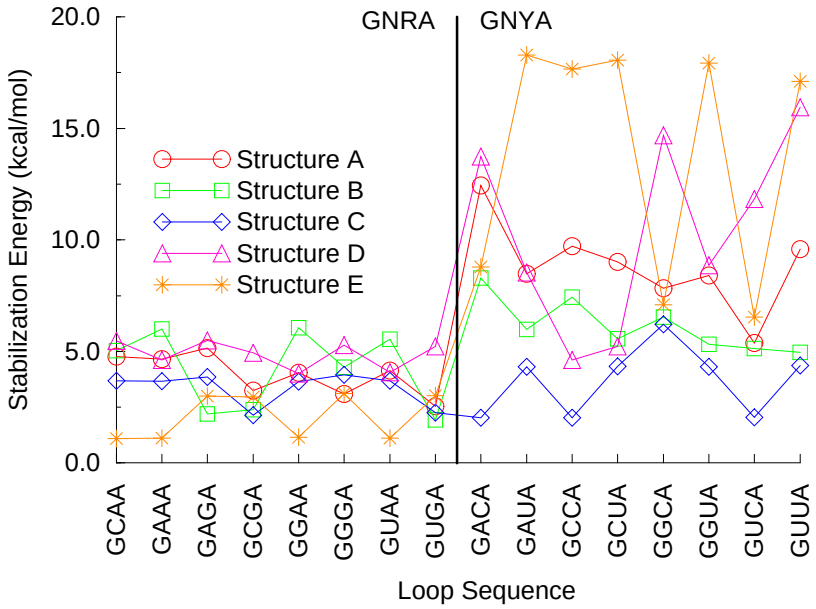


Figure 46: GCAA stabilization energies.

the release of the closing condition allows a relaxation of the loop structure so that steric constraints are decreased, all energy differences are of positive sign. The magnitude of the stabilization energies shows a distinct correlation with the loop sequence. For GNRA loops the stabilization energies have a mean value of 3.7 kcal/mol whereas the mean stabilization energy for GNYA loops is 8.6 kcal/mol.

Using another approach the results about loop-stabilization energies have been confirmed. Single stranded A-form helix used as a starting point for the conformational analysis and energy minimization. The energy differences (between open and closed loop structure) have been calculated relative to the corresponding loop confomers with lowest energies (including RF contri-

butions) and are predominately due to the formation of four G-C base pairs in the stem region of the hairpins. The average values obtained for GNRA and GNYA loop sequences are 39 [kcal/mol] and 37 [kcal/mol], respectively, and indicate a slightly higher stability of GNRA loops compared to GNYA. So even **without** using any constraints we yield results in agreement with experimental data on thermodynamic stability.

4.3 Triloops to Nonaloops

Sequence forming triloops were originally investigated by Herbert Kratky [75] in cooperation with experiments done in Larry Browns group [15] (see chapter 4.1.1.). Experimental investigations has been shown that triloops tend to form dimers in solution rather than remain monomers. This problem was been bypassed by using different solvents and a longer stem for the NMR. By using conformational search technics first hints on new structural features have arised, such as the closing G-C pair is opened and by that means forming a pentaloop. These results were undermined by the NMR studies and confirmed in a recent publication [146]. To get a general feeling about the reliability of the different force-fields a large scale investigation has been started with a triloop sequence *GGCGUUUCGCC* and the results were not encouraging. Following method has been applied: The triloop with the best energy has been taken as a reference structure in table 7 denotated with "Tri A". Using MC-SYM new starting structures have been created for all possible loop-sizes. Yielding 592 starting structures for the pentaloop, 3913 for the heptaloop and 2859 for the nonaloop. Now they where minimized using three different technics:

1. with AMBER
2. with Randstruct
3. with JUMNA + FIESTA

The optimized JUMNA structures where iterated one step in the AMBER force-field, so that resulting energies would be comparable. In figure 47 a bar-chart of the best resulting energies is depicted. The results are not easy to explain. One one hand we have the lowest energy structure obtained by Randstruct forming a heptaloop, on the other, the lowest energy structure

obtained by JUMNA is a pentaloop as proposed with NMR technics. Looking at the structures more closely, see figures 48-51 we can try to explain this behaviour.

Class	label	AMBER	RAND	ΔE	JUMNA
Tri	A	-67.54	-91.44	8.89	-81.72
Penta	77	-80.69	-80.73	0.04	-79.93
	184	-80.69	-87.83	7.14	-65.11
	591	-72.24	-79.13	5.99	-86.74
	673	-69.36	-77.94	8.58	-67.31
Hepta	3132	-64.01	-77.55	13.54	-58.38
	3672	-66.20	-80.81	14.61	-56.85
	3218	-64.34	-101.05	36.71	-13.32
Nona	701	-76.03	-94.28	18.25	-55.60
	700	-75.65	-85.79	10.14	-53.66
	77	-71.41	-91.58	20.17	-61.96
	684	-70.99	-89.93	18.97	-57.43
	79	-69.01	-77.51	7.50	-29.41
	2114	-67.63	-79.87	12.24	-80.46

Table 7: Energy $[kcal/mol]$ of optimized triloops - nonaloops.

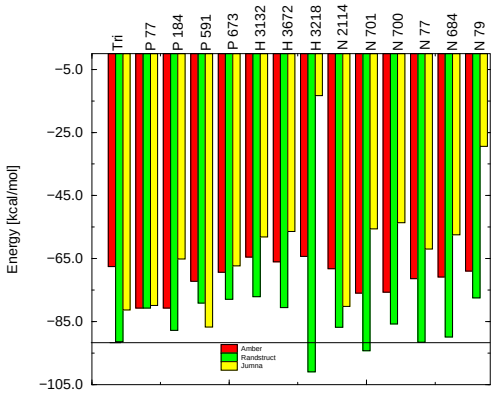


Figure 47: Triloops - nonaloops: Force field results.

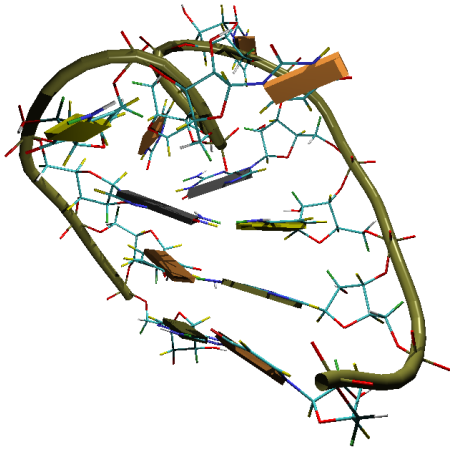


Figure 48: NMR.

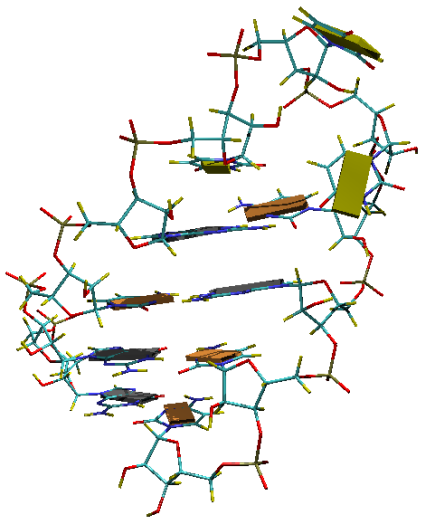


Figure 49: P-519 JUMNA.

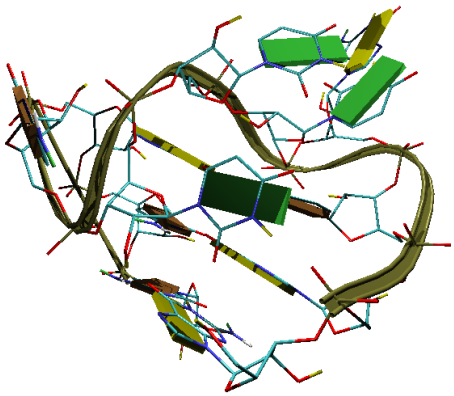


Figure 50: P-184 AMBER.

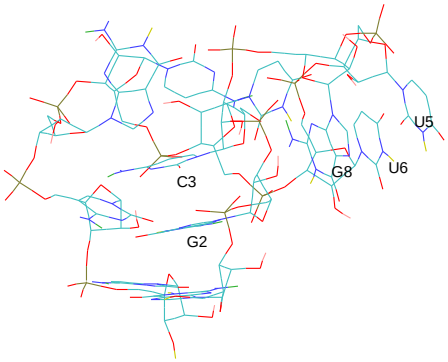


Figure 51: H-3218 Randstruct.

The lowest energy structure according to JUMNA is a pentaloop (see figure 49), which is conform to the NMR findings. However the RMSD between the JUMNA and NMR structure is 5.7 Å. The closing G4 and C8 are not able to construct a H-bond. Stacking is possible in the loop-region between U6 and U7 and thus explaining the low energy. In figure 50 the lowest AMBER structure is depicted. The structure is very compact in the loop region. As an additional structural feature in the loop H-bond between U7 and G2 occurs. The compactness of this structure could explain the low energy. In figure 51 the Randstruct structure is shown, it is a heptaloop here we have an additional H-bond between C3 and G2 and stacking between the bases G8,U6 and U5 is possible. So using each of the three conformational search methods different results are obtained. The forming of a heptaloop using randstruct could be explained, by taking the "stressed" conformation of loop structures into account. The loop bases have more "freedom" and by that means have the possibility to form "relaxed" structures. The best results, i.e. structure closest to the NMR structure, is obtained with JUMNA (and FIESTA). This is conform to the findings of chapter 4.2.1..

4.4 Pseudoknots

Another structure motif has been investigated using conformational search technics: A pseudoknot, in particular the MMTV (Mouse Mammary Tumor Virus) pseudoknot. It is an remarkable structure, forming loops with different sizes: loop1 is formed by A6 and G7 and loop2 is formed by nucleotides 20 to 27. A secondary structure can be seen in figure 52.
The frameshifter pseudoknot possesses structural features not observed in previously reported model pseudoknots. It has a compact structure with a pronounced bend at the junction of its GC-rich stems. A single adenylate residue-nr. 14 is intercalated between the two stems so that direct coaxial stacking of the stems is not possible. The lack of an opposing nucleotide for the stacked, intervening adenylate creates a hinge in the pseudoknot. Most of the loop nucleotides are restrained by base stacking interactions which keep the loops from adopting extended conformations. The sterically constrained loops direct the bending of the pseudoknot at the stem-stem junction. The roles of the intercalated adenylate and loop lengths in causing bending can explain their requirement for efficient frame-shifting [145]. The following method have been applied: It was not possible to gain any

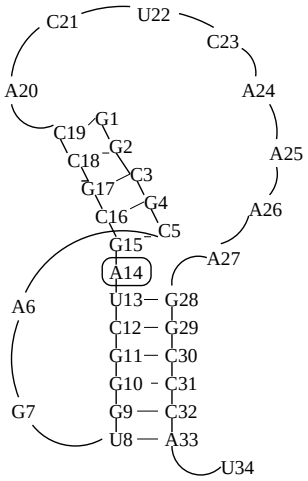


Figure 52: Secondary structure of MMTV pseudoknot.

satisfactory starting structures using MC-SYM for the complete pseudoknot, cause the conformational space was too large and thus this would exceed our computational power. It turned out to be useful to start modeling the 2 stem regions seperatly with Macromodel, and afterwards add the loop structures. As the two loop structures don't interact within the three-dimensional model this is reasonable. For loop1 119 confomers were found and minimized (using AMBER and JUMNA). The resulting loop is in excellent agreement with the NMR-structure. For loop2 2318 confomers were obtained, and optimized. The best confomers are also in quite good agreement with the NMR structure (see table 8). Now confomer 65 of loop1 and confomer 1323 of loop2 were merged and optimized again. The resulting structure had a good agreement, and contained all crucial structural features, with the NMR structure proposed of Tinoco. The RMSD value is 3.2 Å compared to the NMR structure and the energy -241.33 $[kcal/mol]$. Again this was an example of conformational search using tools described in this thesis. It is of importance to notice, that the structure closest to the NMR is *not* the lowest energy structure, but is about 3% higher.

Loop 1			Loop 2			
ID	$E[kcal/mol]$	RMSD	ID	$E[kcal/mol]$	RMSD	RMSD/base
65	-239.21	0.044	1323	-241.98	1.73	0.22
74	-241.43	0.096	1453	-242.33	2.05	0.25
5	-244.32	0.159	200	-243.14	2.44	0.30
19	-243.54	0.182	867	-241.44	2.50	0.31
33	-238.21	0.221	2202	-237.65	2.94	0.98

Table 8: Energies $[kcal/mol]$ and RMSDs of loops.

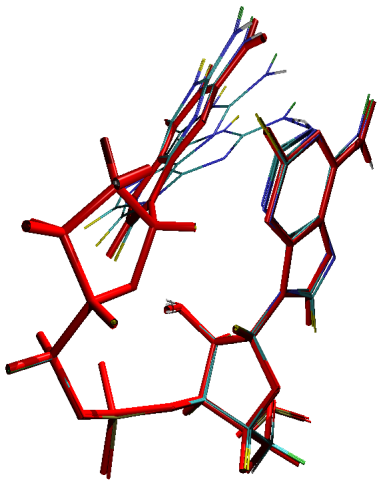


Figure 53: Loop1 of MMTV pseudoknot. The NMR structure is depicted in red color.

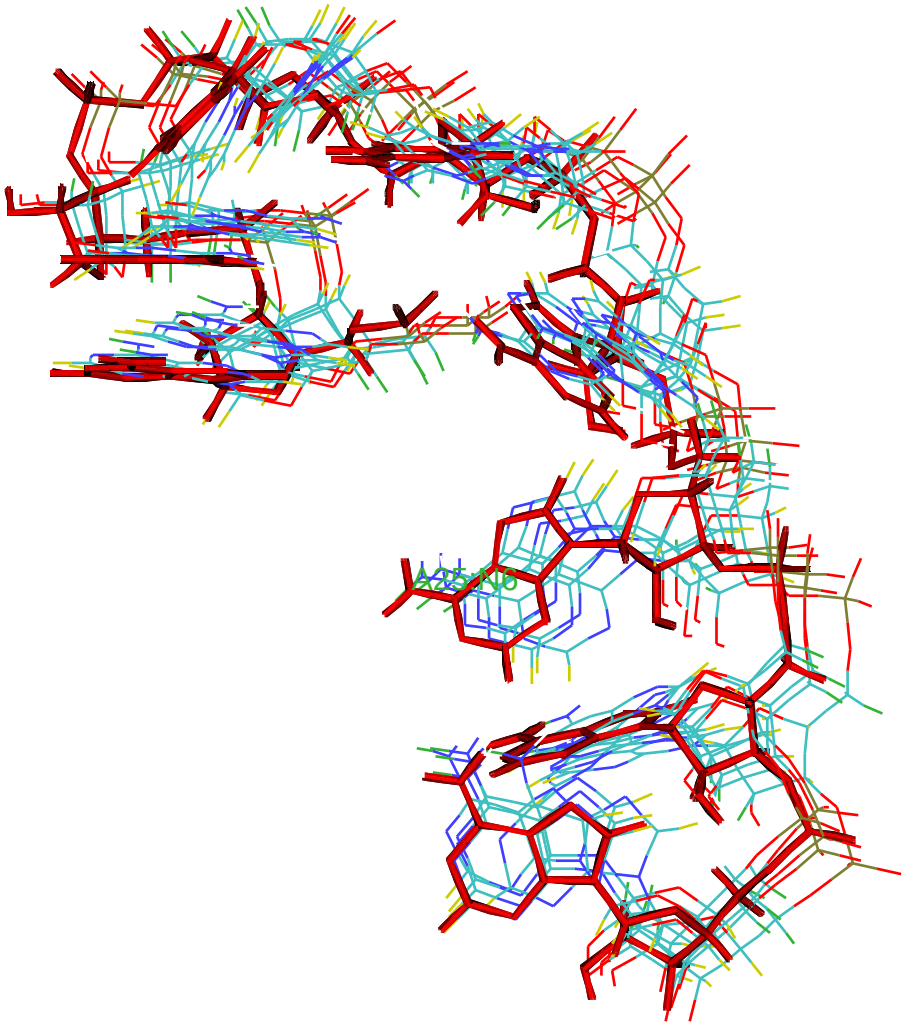


Figure 54: Loop2 of MMTV pseudoknot. The NMR structure is depicted in red color.

5 Conclusion and Outlook

5.1 GEN-3D

During this thesis a new algorithm for conformational sampling based on a GA has been developed (GEN-3D). It allows to optimize loop-stem structures. It has been successfully used to find "fitter", i.e. lower energy structures for triloops compared to other optimization technics. The best results are obtained using 30 iterations of the AMBER force-field and a high mutation rate p_m. A lower crossover rate p_c is also of advantage. This can be observed in nearly all runs. Unfortunately it failed to find new optimal structures for larger loops, responsible for this behaviour is mainly the CPU usage. So, the biggest caveat of GEN-3D algorithm is the fitness function, it is an implementation of the AMBER force field and the most time consuming part. So further investigations to gain a faster algorithm using a coarse grained force-field are in development. The force-field is called *Vieforce* and based on a coarse-grained structure representation. Every nucleotide is represented by a tile. Using *Vieforce*, one might be able to use the GA for larger loop-sizes and by that mean obtain structure proposals for further investigations.

5.2 Tetraloops

The structures of a family of very common and unusually stable RNA hairpins has been determined by a molecular modeling approach. GNRA hairpins are known for their extraordinary thermodynamic stability, that is presumably caused by several specific interactions in the loop. First, all bases but N6 stack on other bases, a G–A pair is formed which stacks on the closing pair of the loop and base N7 stacks on A8. Second, because of the G–A pair the tetraloop is in fact only a "diloop"; the resulting constraints force base N6 in position on top of the loop. Further stabilization comes from additional hydrogen bonds between the base in the loop and the backbone. These features, however, are not restricted to the GNRA-loops, but appear also in the GNYA-loops. Two distinct types of the G–A pair have been found in all of the sequences; the energy differences between both types vary strongly with the nature of the middle bases. Though the loop *geometries* show little dependence on a variation of the middle bases the relative stabilization energies calculated between "open" and "closed" loop structures show a distinct

difference between GNRA and GNYA loops in the order of 5 kcal/mol, that might be a reason for the preference of GNRA over GNYA hairpin loops in nature. The predicted 3D-structures of tetraloop hairpins demonstrate a clear tendency to reduce the interface area between the bases and the aqueous solvent as much as possible. This trend is known to be the major driving force of stack formation, particularly in double helical geometries. It is continued in 3D-structure formation: non-Watson-Crick closing pairs of the stacks are found that, in essence, reduce the size of the loop from four to two bases and, in addition, the remaining two bases try to stack on top of the prolonged stack whenever this is stereo-chemically feasible. In this a new protocol for conformational analyses of ribonucleotid oligomers which will serve in future as a new tool to narrow down 3D structures of RNAs has been presented.

5.3 Force Fields

Sequence forming triloops were originally investigated by Herbert Kratky [75] in cooperation with experiments done in Larry Browns group [15]. Experimental investigations has been shown, that triloops tend to form dimers in solution rather than remain monomers. This problem was been bypassed by using different solvents for the NMR. By using conformational search technics first hints on new structural features have arised, such as the closing G-C pair is opened and by that means forming a pentaloop. These results were supported by the NMR studies and confirmed in a recent publication [146]. To get a general feeling about the reliability of the different force-fields a large scale investigation has been done with the triloop sequence *GGCGUUUCGCC*. Three different conformational search technics have been applied. The best results are obtained using JUMNA (and FIESTA). The lowest energy confomer found using this procedure is a pentaloop (see figure 49), which is conform to the NMR findings. However the RMSD between the JUMNA and NMR structure is 5.7 Å.

5.4 Pseudoknots

Modeling a pseudoknot has also been part of this thesis. Again using conformational sampling methods a surprisingly good result has been obtained. As it would be too time demanding to model the complete pseudoknot the modeling has been focused on the two loops. Both loops were modeled seperatly

and then merged together. As the two loop structures don't interact within the three-dimensional model this is reasonable. For loop1, 119 confomers were found and minimized (using AMBER and JUMNA). The resulting loop is in excellent agreement with the NMR-structure. For loop2, 2318 confomers were obtained, and optimized. The best confomers are also in quite good agreement with the NMR structure (see table 8). The merged structure has been optimized again, it has good agreement with the NMR structure proposed of Tinoco, and containes all its crucial structural features. The RMSD value is 3.2 Å compared to the NMR structure and the energy -241.33 $[kcal/mol]$. Again this was an example of conformational search using tools described in this thesis. It is of importance to notice, that the structure closest to the NMR is *not* the lowest energy structure, but is about 3% higher.

5.5 Concluding Remarks

Some general results found during this thesis are of importance for future calculations of 3D structures of small RNA hairpins. Presenting a new technic, combining different force-fields and conformational search methods, one might have found a way to determine the 3D structure of small RNA molecules. But it is of extreme importance to notice, that in all results presented here the target structures for example NMR structures, where *not* the lowest energy structures, so a "de novo" 3D structure prediction is not possible. Only an *iteratively* process between theoretical and experimental studies, such as chemical enzymatical probing or spectroscopy, can obtain a satisfactory result to achieve a reliable 3D structure prediction. A boom in the investigation of 3D RNA structure prediction is currently under way and great progress is expected for the next years.

List of Figures

References

[1] B. J. Alder and T. Wainwright. Phase transition for a hard sphere system. *J. Chem. Phys.*, 27:1208, 1957.

[2] P. Auffinger, S. Louise-May, and E. Westhof. Multiple molecular dynamics simulations of the anticodon loop of tRNAAsp in aqueous solution with counterions. *J.Am.Chem.Soc.*, 117:6720–6726, 1995.

[3] P. Auffinger, S. Louise-May, and E. Westhof. Molecular dynamics simulations of the anticodon hairpin of tRNAAsp: Structuring effects of C-H...O hydrogen bonds and of long-range hydration forces. *J.Am.Chem.Soc.*, 118:1181–1189, 1996.

[4] P. Auffinger and E. Westhof. H-bond stability in the tRNAAsp anticodon hairpin: 3 ns of multiple molecular dynamics simulations. *Biophys.J.*, 71:940–954, 1996.

[5] G. Awang and D. Sen. Mode of dimerization of HIV-1 genomic RNA. *Biochemistry*, 32:11453–11457, 1993.

[6] K. J. Baeyens, H. L. DeBondt, and S. R. Holbrook. Structure of an RNA double helix including uracil-uracil base pairs in an internal loop. *Nature Struct. Biol.*, 2:6–62, 1995.

[7] P. A. Bash, U. C. Singh, R. Langridge, and P. Kollman. Free energy calculations by computer simulation. *Science*, 236:564–568, 1987.

[8] J. L. Battiste, H. Mao, N. S. Rao, R. Tan, D. R. Muhandiram, L. E. Kay, A. Frankel, and J. R. Williamson. A helix-rna major groove recognition in an HIV-1 Rev peptide-RRE RNA complex. *Science*, 273:1547–1551, 1996.

[9] H. J. C. Berendsen, J. P. Postma, W. F. V. Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[10] B. H. Besler, K. M. Merz, and P. Kollman. Atomic charges derived from semiempirical methods. *J. Comput. Chem.*, 11:431–439, 1990.

[11] K. Boehncke, M. Nonella, K. Schulten, and A. H. J. Wang. Molecular dynamics investigation of the interaction between DNA and distamycin. *Biochemistry*, 30:5465–5475, 1991.

[12] H. Bremermann. A method of unconstrained global optimization. *Mathematical Biosciences*, 9:1–15, 1970.

[13] I. Brierley, N. J. Rolley, A. J. Jenner, and S. C. Inglis. Mutational analysis of the RNA pseudoknot component of a coronavirus ribosomal frameshifting signal. *J. Mol. Biol*, 229:889–902, 1991.

[14] J. W. Brown. Structure and evolution of ribonuclease P RNA. *Biochemie*, 73:689–697, 1991.

[15] L. Brown. *private communications*, 1996.

[16] U. Burkert and N. L. Allinger. *Molecular Mechanics, ACS Monograph 177*. American Chemical Society, 1982.

[17] 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.

[18] M. Chamorro, N. Parkin, and H. E. Varmus. An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA. *Proc. Natl. Acad. Sci. USA*, 89:713–717, 1992.

[19] M. Chastain and I. Tinoco. A base triple structural domain in RNA. *Biochemistry*, 31:12733–12741, 1992.

[20] M. Chastain and I. Tinoco. Nucleoside triples from the group I intron. *Biochemistry*, 32:14220–14228, 1993.

[21] C. Cheong and P. B. Moore. Solution structure of an unusally stable RNA tetraplex containing G- and U- quartet structures. *Biochemistry*, 31:8406–8414, 1992.

[22] C. Cheong, G. Varani, and I. Tinoco. Solution structure of a unusually stable RNA hairpin 5'GGAC(UUCG)GUCC. *Nature*, 346:680–682, 1990.

[23] V. P. Chuprina, U. Heinemann, A. A. Nurislamov, P. Zielenkiewicz, R. E. Dickerson, and W. Saenger. Molecular dynamics simulation of the hydration shell of a B-DNA decamer reveals two main types of minor-groove hydration depending on groove width. *Proc. Natl. Acad. Sci. USA*, 88:593–597, 1991.

[24] M. L. Connolly. Solvent-accesible surfaces of proteins and nucleic acids. *Science*, 221:709–713, 1983.

[25] S. R. Cox and D. E. Williams. Representation of the molecular electrostatic potential by a net atomic charge model. *J. Comput. Chem.*, 2:304–323, 1981.

[26] E. B. T. Dam, C. W. A. Pleij, and L. Bosch. RNA pseudoknots and translational frameshifting on retroviral, coronaviral and luteoviral RNAs. *Virus Genes*, 4:121–136, 1990.

[27] P. W. Davis, W. Thurmes, and I. Tinoco. Structure of a small RNA hairpin. *NAR*, 21(3):537–545, 1993.

[28] R. E. Dickerson, M. Bansal, C. R. Calladine, S. Diekmann, W. N. Hunter, O. Kennard, R. Lavery, H. C. M. Nelson, W. K. Olson, W. Saenger, Z. Shaked, H. Sklenar, D. M. Soumpasis, C. S. Tung, E. von Kitzing, A. H. J. Wang, and V. B. Zhurkin. Definitions and nomenclature of nucleic acid structure parameters. *J.Mol.Biol.*, 205:787–791, 1989.

[29] J. D. Dinman, T. Icho, and R. B. Wickner. A -1 ribosomal frameshifting in a double-stranded RNA virus of yeast forms a gag-pol fusion protein. *Proc. Natl. Acad. Sci. USA*, 88:174–178, 1991.

[30] E. M. Engler, J. D. Andose, and P. R. V. Schleyer. Critical evaluation of molecularmechanics. *J. Am. Chem. Soc.*, 95:8005–8025, 1973.

[31] J. Feigon, T. Diekmann, and F. W. Smith. Aptamer structures from A to zeta. *Chemistry & Biology*, 3:611–617, 1996.

[32] B. Felden, C. Florentz, R. Gieg, and E. Westhof. A central pseudoknotted three-way junction imposes trna-like mimicry and the orientation of three 5' upstream pseudoknots in the 3' terminus of tobacco mosaic virus RNA. *RNA*, 2:201–212, 1996.

[33] G. G. Ferenczy. Charges derived from distributed multipole series. *J. Comput. Chem.*, 12:913–917, 1991.

[34] D. S. Fields and R. R. Gutell. An analysis of large rRNA sequences folded by a thermodynamic method. *Folding & Design*, 1:419–430, 1996.

[35] A. C. Forster and S. Altman. Similar cage-shaped structures for the RNA component of all ribonuclease P and ribonuclease MRP enzymes. *Cell*, 62:407–409, 1990.

[36] M. A. Fountain, T. R. Serra, T. R. Krugh, and T. Turner. Structural features of a six-nucleotide RNA hairpin loop found in ribosomal RNA. *Biochemistry*, 35:6539–6548, 1996.

[37] V. Fritsch and E. Westhof. Molecular dynamics simulations of DNA oligomers under various electrostatic parameters. *J. Chim. Phys. Phys.-Chim. Biol.*, 88:2543–2550, 1991.

[38] D. R. Gallie, J. N. Feder, R. T. Schmike, and V. Walbot. Functional analysis of the tobacco mosaic virus tRNA-like structure in cytoplasmic gene regulation. *Nucleic Acids*, 19:5031–5036, 1991.

[39] D. Gautheret, S. H. Damberger, and R. R. Gutell. Identification of base-triples in RNA using comparative sequence analysis. *J.Mol.Biol.*, 248:27–43, 1995.

[40] D. Gautheret, D. Konnings, and R. R. Gutell. A major family of motifs involving G · A mismatches in ribosomal RNA. *J.Mol.Biol.*, 242:1–8, 1994.

[41] D. Gautheret, F. Major, and R. Cedergren. Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. *J.Mol.Biol.*, 229(4):1049–1064, 1993.

[42] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403, 1976.

[43] R. Green and J. W. Szostak. In vitro genetic analysis of the hinge region between helical elements P5-P4-P6 and P7-P3-P8 in the sunY group I self-splicing intron. *J.Mol.Biol.*, 235:140–155, 1994.

[44] D. R. Groebe and O. C. Uhlenbeck. Characterization of RNA hairpin loop stability. *NAR*, 16:11725–11735, 1988.

[45] W. F. V. Gunsteren and H. J. C. Berendsen. Algorithms for macromolecular dynamics and constraint dynamics. *Mol. Phys.*, 34:1311–1327, 1977.

[46] W. F. V. Gunsteren and H. J. C. Berendsen. Molküldynamik-Computersimulationen: Methodik, Anwendungen und Perspektiven in der Chemie. *Angewandte Chemie*, 102:1020–1055, 1990.

[47] W. F. V. Gunsteren, H. J. C. Berendsen, R. G. Geurtsen, and H. R. J. Zwinderman. A molecular dynamics computer simulation of an eight-base-pair DNA fragment in aqueous solution: Comparison with experimental two-dimensional NMR data. *Ann. New York Acad. Sci.*, 482:287–303, 1986.

[48] W. F. V. Gunsteren and A. E. Mark. On the interpretation of biochemical data by molecular dynamics computer simulation. *Eur. J. Biochem.*, 204:947–961, 1992.

[49] R. R. Gutell, N. Larsen, and C. R. Woese. Lessons from an evolving RNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiological Reviews*, 58:10–26, 1994.

[50] E. S. Haas, D. P. Morse, J. W. Brown, J. F. Schmidt, and N. R. Pace. Long-range structure in ribonuclease P RNA. *Science*, 254:853–856, 1991.

[51] A. T. Hagler, E. Huler, and S. Lifson. Energy functions for peptides and proteins. 1. derivation of a consistent force field including the hydrogen bond from amide crystals. *J. Am. Chem. Soc.*, 96:5319–5327, 1974.

[52] J. A. Hartley, M. Berardini, M. Ponti, N. W. Gibson, A. S. Thompson, D. E. Thurston, B. M. Hoey, and J. Butler. DNA cross-linking and sequence selectivity of Aziridinylbenzoquinones - a unique reaction at 5'-GC-3' sequences with 2,5-Diaziridinyl-1,4-Benzoquinone upon reduction. *Biochemistry*, 30:11719–11724, 1991.

[53] F. H. Hausheer, U. C. Singh, J. D. Saxe, and O. M. Colvin. Identification of local determinants of dna interstrand crosslink formation by cyclophosphamide metabolites. *Anti-Cancer Drug Des*, 4:281–294, 1989.

[54] P. Herzyk, J. M. Goodfellow, and S. Neidle. Molecular dynamics simulations of dinucleoside and dinucleoside-drug crystal hydrates. *J. Biomol. Struct. Dyn.*, 9:363–386, 1991.

[55] H. A. Heus and A. Pardi. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science*, 253:191–194, 1991.

[56] G. C. Hill and W. A. Remers. Computer simulation of the binding of saframycin A to d(GATGCATC)2. *J. Med. Chem.*, 34:1990–1998, 1991.

[57] G. C. Hill, T. P. Wunz, N. E. MacKenzie, P. R. Gooley, and W. Remers. Computer simulation of the binding of naphthyridinomycin and cyanocycline A to DNA. *J. Med. Chem*, 34:2079–2088, 1991.

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

[59] J. Holland. *Adaption in Natural and Artificial Systems*. The University of Michigan Press, Ann. Arbor, 1975.

[60] S. Huang, Y.-X. Wang, and D. E. Draper. Structure of a hexanucleotide RNA hairpin loop conserved in ribosomal RNAs. *J.Mol.Biol.*, 258:308–321, 1996.

[61] A. Hüttendorfer, E. Westhof, and A. Böck. Solution structure of mRNA hairpins promoting selenocysteine incorporation in Escherichia coli and

their base-specific interaction with special elongation factor SELB. *RNA*, 2:354–366, 1996.

[62] M. A. Huynen, D. M. A. Konings, and P. Hogeweg. Multiple coding and the evolutionary properties of RNA secondary structure. *J.theor.Biol.*, 165:251–267, 1993.

[63] L. Jaeger, F. Michel, and E. Westhof. Involvement of a GNRA tetraloop in long-range RNA tertiary interactions. *J.Mol.Biol.*, 236:1271–1276, 1994.

[64] F. M. Jucker and A. Pardi. GNRA tetraloops make a U-turn. *RNA*, 1:219–222, 1995.

[65] A. Kajava and H. Rüterjans. Molecular modeling of the 3d-structure of RNA tetraloops with different nucleotide sequences. *Nucleic Acids Research*, 21(19):4556–4562, 1993.

[66] M. Karplus and G. A. Petsko. Molecular dynamics simulations in biology. *Nature*, 347:631–639, 1990.

[67] Y. Kawakami and A. J. Hopfinger. Molecular dynamics simulations of the intercalation of benzothiopyranoindazole anticancer analogs with DNA models. *Anti-Cancer Drug Des.*, 7:181–201, 1992.

[68] O. Kikuchi, A. J. Hopfinger, and G. Klopman. Chemical reactivity of protonated aziridine with nucleophilic centers of DNA bases. *Biopolymers*, 19:325–340, 1980.

[69] S. Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[70] E. v. Kitzing. Modelling DNA structure. *Progress in Nucleic Acid Research and Molecular Biology*, 30:87–108, 1992.

[71] E. v. Kitzing. Modelling DNA structures: molecular mechanics and molecular dynamics. *Methods in Enzymology*, 211:449–467, 1992.

[72] R. Klinck, J. Liquier, E. Taillandier, C. Gouyette, T. Huynh-Dinh, and E. Guittet. Structural characterization of an intramolecular RNA

triple helix by NMR spectroscopy. *European Journal of Biochemistry*, 233:544–553, 1995.

[73] K. W. Kohn, J. A. Hartley, and W. B. Mattes. Mechanisms of DNA sequence selective alkylation of guanine-n7 positions by nitrogen mustards. *NAR*, 15:10531–10549, 1987.

[74] D. A. M. Konings and R. R. Gutell. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA*, 1:558–574, 1995.

[75] H. Kratky. *Investigations on the three dimensional structure of RNA molecules using force field calculations*. PhD thesis, University of Vienna, 1996.

[76] L. G. Laing and K. B. Hall. A model of the iron responsive element RNA hairpin loop structure determined from NMR and thermodynamic data. *Biochemistry*, 35:13586–13596, 1996.

[77] D. R. Langley, T. W. Doyle, and D. Beveridge. The dynemicin-DNA intercalation complex: a model based on DNA affinity cleavage and molecular dynamics simulation. *J. Am. Chem. Soc*, 113:4395–4403, 1991.

[78] R. Lavery. *Structure & Expression Volume 3 : DNA Bending and Curvature*, pages 191–211. Adenine, Schenectady, New York, 1987.

[79] R. Lavery. Jumna (version 7). *Laboratorie de Biochimie Theorique CNRS, Institute de Biologie Physico-Chimique, Paris*, 1992.

[80] R. Lavery, I. Parker, and J. Kendrick. A general approach to the optimization of the conformation of ring molecules with an application to valinomycin. *J.Struct.Dyn.*, 4:443–461, 1986.

[81] R. Lavery and H. Sklenar. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.*, 6:63–91, 1988.

[82] R. Lavery and H. Sklenar. Defining the structure of irregular nucleic acids: conventions and principles. *J.Struct.Dyn.*, 6:655–667, 1989.

[83] R. Lavery, H. Sklenar, K. Zakrzewska, and B. Pullman. The flexibility of nucleic acids (II): The calculation of internal energy and applications to mononucleotide repeat DNA. *J.Struct.Dyn.*, 3:989–1014, 1986.

[84] R. Lavery, K. Zakrzewska, and A. Pullman. Binding of non-intercalating antibiotics to B-DNA: a theoretical study taking into account nucleic acid flexibility. *J.Struct.Dyn.*, 4:443–461, 1986.

[85] F. Leclerc, R. Cedergren, and A. E. Ellington. A three-dimensional model of the Rev-binding element of HIV-1 derived from analyses of aptamers. *Nature: Structural Biology*, 1:293–300, 1994.

[86] C. S. Lee, J. A. Hartley, M. D. Berardini, J. Butler, D. Siegel, D. Ross, and N. W. Gibson. Alteration in DNA cross-linking and sequence selectivity of a series of aziridinylbenzoquinones after enzymatic reduction by dt-diaphorase. *Biochemistry*, 31, 1992. 3019-3025.

[87] M. Levitt. Computer simulation of DNA double-helix dynamics. *Cold Spring Harbor Symp. Quant. Biol. Vol.*, 47:251–262, 1983.

[88] S. E. Lietzke, C. L. Barnes, J. A. Berglund, and C. E. Kundrot. The structure of an RNA dodecamer shows how tandem U-U base pairs increase the range of stable RNA structures and the diversity of recognition sites. *Structure*, 4:917–929, 1996.

[89] S. Louise-May, P. Auffinger, and E. Westhof. RNA structure from molecular dynamics simulations. *Biological Structure and Dynamics*, pages 1–18, 1995. Adenine Press. Schenectady (NY).

[90] T. P. Lybrand. Computer simulations of biomolecular systems using molecular dynamics and free energy perturbation methods. *Reviews in Computational Chemistry*, 1:295–320, 1990.

[91] F. Major, D. Gautheret, and R. Cedergren. Reproducing the three-dimensional structure of a tRNA molecule from structural constraints. *Proc. Natl. Acad. Sci. USA*, 90:9408–9412, 1993.

[92] F. Major, M. Turcotte, D. Gautheret, G. Lapaplme, E. Fillion, and R. Cedergren. The combination of symbolic and numerical

computations for three-dimensional modelling of RNA. *Science*, 253(5025):1255–1260, 1991.

[93] R. Mans, C. Pleij, and L. Bosch. Transfer RNA-like structures: Structure, function and evolutionarz significance. *Eur J Biochem*, 201:303–324, 1991.

[94] R. Mans, M. H. V. Steeg, P. Verlaan, C. Pleij, and L. Bosch. Mutational Analysis of the Pseudoknot in the tRNA-like Structure of Turnip Yellow Mosaic Virus RNA. Aminoacylation Efficiency and RNA Pseudoknot Stability. *J.Mol.Biol.*, 223:221–232, 1992.

[95] W. B. Mattes, J. A. Hartley, and K. Kohn. DNA sequence selectivity of Guanine-N7 alkylation by nitrogen mustards. *NAR*, 14:2971–2987, 1986.

[96] O. Melefors, U. Lundberg, and A. v. Gabain. RNA processing and degradation by RNase K and RNase E. In J. G. Belasco and G. Brawerman, editors, *Control of messenger RNA stability*, pages 53–70. Academic Press Inc. New Yor, 1993.

[97] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087, 1953.

[98] F. Michel, A. D. Ellington, S. Couture, and J. W. Szostak. Phylogenetic and genetic evidence for base-triples in the catalytic domain of group I introns. *Nature*, 347:578–580, 1990.

[99] F. Michel and E.Westhof. Modelling of the three-dimmensional architecture of group I catalytic introns based on comparative sequence anaysis. *J.Mol.Biol.*, 216:585–610, 1990.

[100] W. A. Miller and S. L. Silver. Alternative tertiary structure attenuates self-cleavage of the ribozyme in the satellite RNA of barley yellow dwarf virus. *Nucleic Acids Research*, 19(19):5313, 1991.

[101] S. R. Mirmira and I. Tinoco. NMR structure of a bacteriophage T4 RNA hairpin involved in translational repression. *Biochemistry*, 35:7664–7674, 1996.

[102] D. Moazed and H. F. Noller. Transfer RNA shields specific nucleotides in 16S ribosomal RNA from attack by chemical probes. *Proc Natl Acad Sci USA*, 47:985–994, 1986.

[103] F. A. Momany, L. M. Carruthers, R. F. McGuire, and H. A. Scheraga. Intermolecular potentials from crystal data. 3. determination of empirical potentials and application to the packing configurations. *J. Phys. Chem*, 78:1595–1620, 1974.

[104] P. B. Moore. Recent RNA structures. *Curr. Opin. Struct. Biol.*, 3:340–344, 1993.

[105] F. L. Murphy and T. R. Cech. GAAA tetraloop and conserved bulge stabilize tertiary structure of group I intron domain. *J.Mol.Biol.*, 236(1):49–63, 1994.

[106] S. R. Niketic and K. Rasmussen. *The Consistent Force Field*. Springer: New York, 1977.

[107] R. Nussinov. Conserved quartets near 5′intron junctions in primate nuclear pre-mRNA. *Journal of Theoretical Biology*, 133:73–84, 1988.

[108] H. Ogata, Y. Akiyama, and M. Kanehisa. A genetic algorithm based molecular modeling technique for RNA stem-loop structures. *Nucl.Acids Res.*, 23:419–426, 1995.

[109] M. Orita, F. Nishikawa, T. Shirnayama, K. Taira, Y. Endo, and S. Nishikawa. High-resolution NMR study of a synthetic oligoribonucleotide with a tetranucleotide GAGA loop that is a substrate for the cytotoxic protein, ricin. *Nucl. Acid. Res.*, 21:5670–5678, 1993.

[110] D. A. Pearlman, D. A. Case, J. C. Caldwell, G. L. Seibel, C. Singh, P. Weiner, and P. A. Kollman. AMBER 4.0, University of California, San Francisco. 1991.

[111] C. Philippe, C. Portier, M. Mougel, M. Grunberg-Manago, J. P. Ebel, B. Ehresmann, and C. Ehresmann. Target site of escherichia coli ribosomal protein S15 on its messenger RNA. *J.Mol.Biol.*, 211:415–426, 1990.

[112] C. W. A. Pleij. Pseudoknots a New Motiv in the RNA Game. *Trends Biochem Sci*, 15:143–147, 1990.

[113] C. W. A. Pleij, K. Rietveld, and L. Bosch. A new principle of RNA folding based on pseudoknotting. *Nucl.Acid.Res.*, 13:1717, 1985.

[114] H. W. Pley, K. M. Flaherty, and D. B. McKay. Model for an RNA tertiary interaction from the structure of an intermolecular complex between GAAA tetraloop and an RNA helix. *Nature*, 372:111–113, 1994.

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

[116] T. Powers and H. F. Noller. A functional pseudoknot in 16S ribosomal RNA. *EMBO*, 10:2203–2214, 1991.

[117] M. Prabhakaran and S. C. Harvey. Molecular dynamics anneals large-scale deformations of model macromolecules: stretching the DNA double helix to form an intercalation site. *J. Phys. Chem.*, 89:5767–5769, 1985.

[118] S. L. Price and N. J. G. Richards. On the representation of electrostatic fields around ab initio charge distributions. *J. Comput.-Aided Mol. Design*, 5:41–54, 1991.

[119] J. D. Puglisi, J. R. Wyatt, and I. Tinocco. RNA Pseudoknots. *Acc Chem Res*, 24:152–158, 1991.

[120] J. D. Puglisi, J. R. Wyatt, and I. Tinoco. Solution conformation of an RNA hairpin loop. *Biochemistry*, 29:4215–4226, 1990.

[121] J. D. Puglisi, J. R. Wyatt, and I. Tinoco. RNA pseudoknots. *Accounts of Chemical Research*, 24:153, 1991.

[122] A. Pullman and B. Pullman. Electrostatic effect of the macromolecular structure on the biochemical reactivity of the nucleic acids. significance for chemical carcinogenesis. *Int. J. Quantum Chem. QBS*, 7:245–259, 1980.

[123] A. L. N. Rao, T. W. Dreher, L. E. Marsch, and T. C. Hall. Telomeric Function of the tRNA-like Structure of Brome Mosaic Virus RNA. *Proc. Natl. Acad. Sci. USA*, 86:5335–5339, 1989.

[124] S. N. Rao and P. A. Kollman. Simulations of the B-DNA molecular dynamics of d(CGCGAATTCGCG)2 and d(GCGCGCGCGC)2: an analysis of the role of initial geometry and a comparison of united and all-atom models. *Biopolymers*, 29:517–532, 1990.

[125] S. N. Rao, U. C. Singh, P. A. Bash, and P. A. Kollman. Free energy perturbation calculations on binding and catalysis after mutating Asn 155 in subtilisin. *Nature*, 328:551–554, 1987.

[126] G. Ravishanker, S. Swaminathan, D. L. Beveridge, R. Lavery, and H. Sklenar. Conformational and helicoidal analysis of 30 ps of molecular dynamics on the d(CGCGAATTCGCG) double helix: "curves" and dials and windows. *J. Biomol. Struct. Dyn.*, 6:669–699, 1989.

[127] I. Rechenberg. Cybernetic solution path of an experimental problem. *Royal Aircraft Establishement (U. K.)*, 1965.

[128] I. Rechenberg. *Evolutionsstrategie: Optimierung Technischer Systeme nach Prinzipien der Biologischen Evolution.* Frommann-Holzboog(Stuttgart), 1973.

[129] C. A. Reynolds, G. G. Ferenczy, and W. G. Richards. Methods for determining the reliability of semiempirical electrostatic potentials and potential derived charges. *J. Mol. Struct. (Theochem)*, 256:249–269, 1992.

[130] F. M. Richards. Die Faltung von Proteinmolekülen. *Sp. d. Wiss.*, 3:72–81, 1991.

[131] J. A. C. Rullmann and P. T. V. Duijnen. Potential energy models of biological macromolecules: A case for ab initio quantum chemistry. *Reports in Molecular Theory 1*, pages 1–21, 1990.

[132] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motions of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23:327–341, 1977.

[133] W. Saenger. *Principles of Nucleic Acid Structure*. Springer Verlag New York, 1984.

[134] T. Sakata, H. Hiroaki, Y. Oda, T. Tanaka, M. Ikehara, and S. Uesugi. Studies on the structure and stabilizing factor of the CUUCG hairpin RNA using chemically synthesized oligonucleotides. *Nucleic Acids Research*, 18(13):3831–3839, 1990.

[135] P. Schimmel. RNA Pseudoknots that Interact with Components of the Translation Apparatus. *Cell*, 58:9–12, 1989.

[136] S. Schulze-Kremer. *Genetic algorithms for protein tertiary structure prediction.* Parallel Problem Solving from Nature 2. North Holland, 1992.

[137] P. Schuster. Evolutionary biotechnology - theory, facts, and perspectives. *Acta Biotechnol*, 16:3–17, 1996.

[138] H. P. Schwefel. *Evolutionsstrategie und numerische Optimierung.* PhD thesis, Technische Universität Berlin, 1975.

[139] W. G. Scott, J. T. Finch, and A. Klug. The crystal structure of an all-RNA hammerhead ribozyme: A proposed mechanism for RNA catalytic cleavage. *Cell*, 81:991–1002, 1995.

[140] G. L. Seibel, U. C. Singh, and P. A. Kollman. A molecular dynamics simulation of double-helical B-DNA including counterions and water. *Proc. Natl. Acad. Sci. USA*, 82:6537–6540, 1985.

[141] D. Sen and W. Gilbert. Novel DNA superstructures formed by telomere-like oligomers. *Biochemistry*, 31:65–70, 1992.

[142] M. J. Serra, M. H. Lyttle, T. J. Axenson, C. A. Schadt, and D. H. Turner. RNA hairpin loop stability depends on closing pair. *NAR*, 21(16):3845–3849, 1993.

[143] B. A. Shapiro. An algorithm for comparing multiple RNA secondary structures. *CABIOS*, 4:381–393, 1988.

[144] B. A. Shapiro. Comparing multiple RNA secondary structures using tree comparison. *CABIOS*, 6:309–318, 1990.

[145] L. X. Shen, Z. Cai, and I. Tinoco. RNA structure at high resolution. *FASEB*, 9:1023–1033, 1995.

[146] C. Sich, O. Ohlenschläger, R. Ramachandran, M. Görlach, and L. Brown. Structure of an RNA hairpin loop with a 5'-CGUUUCG-3' loop motif by heteronuclear NMR spectroscopy and distance geometry. *Biochemistry*, 36:13989–14002, 1997.

[147] S. B. Singh and P. A. Kollman. Understanding the thermodynamic stability of an RNA hairpin and its mutant. *Biophys.J.*, 70:1940–1948, 1996.

[148] U. C. Singh and P. A. Kollman. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.*, 5:129–145, 1984.

[149] H. Sklenar, C. Etchebest, and R. Lavery. Describing protein structure: A general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins: Struct. Funct. Genet.*, 6:46–60, 1989.

[150] H. Sklenar, R. Lavery, and B. Pullman. The flexibility of nucleic acids (I) "SIR" a novel approach to variation of polymer geometry in constrained systems. *J.Struct.Dyn.*, 3(5):967–986, 1986.

[151] M. Y. Song and M. S. Jhon. Molecular dynamics study of the effect of ion concentration on the B-DNA and Z-DNA and DNA-Daunomycin complex. *J. Mol. Struct. Theochem*, 89:33–47, 1992.

[152] J. Srinivasan, F. Leclerc, A. D. Ellington, and R. Cedergren. A docking and modelling strategy for peptide-RNA complexes: Applications to BIV-Tat-TAR and HIV Rev-RBE. *Folding & Design*, 1:463–472, 1996.

[153] S. Swaminathan, G. Ravishanker, and D. L. Beveridge. Molecular dynamics of B-DNA including water and counterions: A 140-ps trajectory for d(CGCGAATTCGCG) based on the GROMOS force field. *J. Biomol. Struct. Dyn.*, 6:669–699, 1989.

[154] K. N. Swamy and E. Clementi. Hydration structure and dynamics of B- and Z-DNA in the presence of counterions via molecular dynamics simulations. *Biopolymers*, 26:1901–1927, 1987.

[155] B. Tidor, K. K. Irikura, B. R. Brooks, and M. Karplus. Dynamics of DNA oligomers. *J. Biomol. Struct. Dyn.*, 1:231–252, 1983.

[156] M. Turcotte, G. Lapalme, and F. Major. Exploring the conformations of nucleic acids. *J.Functional Programming*, 5:443–460, 1995.

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

[158] T. H. Tzeng, C. L. Tu, and J. A. Bruenn. Ribosomal frameshifting requires a pseudoknot in the saccharomyces cerevisiae double-stranded RNA virus. *J Virus*, 66:999–1006, 1992.

[159] G. Varani, C. Cheong, and I. Tinoco. Structure of an unusually stable RNA hairpin. *Biochemistry*, 30:3280–3289, 1991.

[160] R. Walczak, E. Westhof, P. Carbon, and A. Krol. A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. *RNA*, 2:367–379, 1996.

[161] Y. X. Wang, S. Huang, and D. E. Draper. Structure of a U-U pair within a conserved ribosomal RNA hairpin. *NAR*, 24:2666–2672, 1996.

[162] A. Warshel. Semiempirical methods of electronic structure calculation. In S. G. A., editor, *Modern Theoretical Chemistry Vol. 7*, page 133. Plenum, New York, 1977.

[163] A. M. Weiner and N. Maizels. tRNA-like structures tag the 3' ends of genomic RNA molecules for replication: Implications for the origin of protein sybthesis. *Proc. Natl. Acad. Sci. USA*, 84:7383–7387, 1987.

[164] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784, 1984.

[165] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem*, 7:1230–1252, 1986.

[166] D. H. Wertz. PhD thesis, University of Georgia, 1974.

[167] E. Westhof and L. Jaeger. RNA pseudoknots. *Current Opinion in Structural Biology*, 2:327, 1992.

[168] E. Westhof, D. Weslowski, and S. Altman. Mapping in three dimensions of regions in a catalytic RNA protected from attack by an Fe(II)-EDTA reagent. *J.Mol.Biol.*, 258:600–613, 1995.

[169] N. Wills, R. F. Gesteland, and J. F. Atkins. Evidence that a downstream pseudoknot is required for translational readthrough of the moloney murine leukemia virus gag stop codon. *Proc. Natl. Acad. Sci. USA*, 88:6991–69995, 1991.

[170] C. R. Woese and R. R. Gutell. Evidence for several higher order structural elements in ribosomal RNA. *Proc Natl. Acad. Sci. USA*, 86:3119–3122, 1989.

[171] C. R. Woese, S. Winker, and R. R. Gutell. Architecture of ribosomal RNA: Constraints on the sequence of "tetra-loops". *Proc Natl. Acad. Sci. USA*, 87:8467–8471, 1990.

[172] J. Wolters. The nature of preferred hairpin structures in 16S-like RNA variable regions. *NAR*, 20:1843–1850, 1992.

[173] D. A. Zichi. Molecular dynamics of RNA with OPLS force field. aqueous simulation of a hairpin containing a tetranucleotide loop. *J.Am.Chem.Soc.*, 117(11):2958–2969, 1995.

[174] T. J. Zielinski and M. Shibata. A molecular dynamics simulation of the (dG)6.(dC)6 minihelix including counterions and water. *Biopolymers*, 29:1027–44, 1990.

[175] T. J. Zielinski, M. Shibata, and R. Rein. High propeller twist and unusual hydrogen bonding patterns from the MD simulation of (dG)6.(dC)6. *FEBS Lett.*, 236:450–454, 1988.

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

Curriculum vitae

Name:	Alexander RENNER
Date of Birth:	30.6.66
Home Address:	Fasholdgasse 6, A - 1130 Vienna, AUSTRIA
Citizenship:	Austrian
Language Skills:	Fluent in German, Swedish, English
-1985	Studentexamen - Deutsches Gymnasium Stockholm
	(German High School, Stockholm, Sweden)
1985-1986	Service in Austrian Armed Forces
since 1986	University of Vienna, Austria
	Major: Biochemistry
1992-1993	SANDOZ GesmbH, Austria
	Scientific EDV, System Administrator
1994-1995	Diploma Theses *:
	Self Avoiding Walks And Lattice Polymers
since 1995	Phd Theses *:
	Algorithms for the Prediction of 3D RNA Molecules
Summer Jobs:	
1987	WISTAR Institute; Philadelphia, USA
	Lab Assistant, Molecular Biology
1988	Boehringer Ingelheim Braknell, UK
	Lab Assistant, Quality Control, Pharmaceuticals
1989	Bender Vienna, Austria
	Lab Assistant, Molecular Biology
1990	Genentech, Inc., San Francisko, USA
	Lab Assistant, Molecular biology
	project: Atrial Natriuretic Factor
1991	Karolinska Institute, Stockholm, Sweden
	Dpt of Medical Chemistry
	Laboratory of Molecular Neurobiology
	project: Cloning of Neurotrophins related to NGF
1993	Boehringer Ingelheim, Ridgefield, Conneticut, USA
	Dpt of Molecular Biology
	Nevirapine Project, etc
1997	Santa Fe Institute New Mexico, USA
	Complex System Summer School
	project: Genetic Algorithms for 3D Structure Prediction

* at the Institue of Theoretical Chemistry, Vienna - AUSTRIA

Publications

A.Renner and E. Bornberg-Bauer. Exploring the fitness landscapes of lattice proteins. Proceedings of the Pacific Symposium on Biocumputing 1997, p. 361-372. L. Hunter and T. Klein (Eds), World Scientific, London.

A.Renner, E. Bornberg-Bauer, I.L. Hofacker, P. Schuster and P.F. Stadler. Self-avoiding walk models for non-random heteropolymers, preprint 1997.

J. Cupal, C. Flamm, A. Renner and P.F. Stadler. Density of States, Metastable States, and Saddle Points. Exploring the Energy Landscape of an RNA Molecule. Proceedings of ISMB-97, p. 88-91, T. Gaasterland, P. Karp, K. Karplus, Ch. Ouzounis, Ch. Sander, & A. Valencia (Eds), AAAI Press.

P. Schuster, P.F. Stadler and A.Renner. RNA structures and folding: from conventional to new issues in structure predictions. Current Opinion in Structural Biology 1997, 7:229-235.

A. Maier, H. F. Kratky, A. Renner, H. Sklenar and P. Schuster. Predicting RNA structural motifs by conformational search: GNRA tetraloops and their pyrimidine relatives. preprint 1998.