INVESTIGATIONS ON THE THREE-DIMENSIONAL STRUCTURE OF RNA MOLECULES USING FORCE FIELD CALCULATIONS

DISSERTATION

zur Erlangung des
akademischen Grades

Doktor rerum naturalium

vorgelegt von

Mag. rer. nat. Herbert F. Kratky

Wien, im Juni 1996

am Institut für Theoretische Chemie und Strahlenchemie
der Universität Wien

Never worry about theory as long as the machinery does what it is supposed to do !

Robert A. Heinlein

Before we begin: thank you, thank you, thank you . . .

Diese Arbeit entstand zwischen Dezember 1993 und Juni 1996 am Institut für Theoretische Chemie, und es ist mir ein Bedürfnis allen jenen zu danken, die mich während dieser Zeit unterstützt haben.

Professor Peter Schuster hat zur Entstehung dieser Arbeit durch seine wissenschaftliche Betreuung und Förderung wesentlich beigetragen, indem er Hilfestellung gegeben hat, wo es notwendig war, aber gleichzeitig genug Freiraum und Verständis für eigene Kreativität und Experimentierlust gelassen hat.

Dr. Eberhard von Kitzing vom Max-Planck-Institut in Heidelberg und Dr. Heinz Sklenar vom Max-Delbrück-Centrum in Berlin haben mich freundlich in ihre Arbeitgruppen aufgenommen und mir beim Kennenlernen neuer Methoden und Programme sehr geholfen.

Die Kollegen am Institut für Theoretische Chemie, allen voran Bärbel Krakhofer, Tanja Ukrainczyk, Ronke Babajide, Ivo Hofacker, Robert Happel, Alexander Renner, Thomas Griesmacher, Josef Frömcke, Walter Fontana, Peter Stadler, Erich Bornberg-Bauer und nicht zuletzt Judith Jakubetz haben dafür gesorgt, daß am Institut neben der hehren Wissenschaft auch der Schmäh nicht zu kurz kommt. Auch den Kollegen der Arbeitsgruppen für Molekulardynamik und Quantenchemie sei für die gute Zusammenarbeit gedankt. Wolfgang Bitomsky verdanke ich die regelmäßige Versorgung mit den neuesten Programmversionen und Patches.

Die Kollegen am Institut für Molekulare Biotechnologie in Jena aus den Arbeitsgruppen von Prof. Schuster und Prof. Brown, insbesondere Christian Forst, Christian Reidys, Stephan Kopp, Oliver Ohlenschläger, Christian Sich und Janos Palinkas haben in zahlreichen Diskussionen wertvolle Anregungen zu dieser Arbeit geliefert und mir bei meinen Besuchen in Jena das Gefühl vermittelt, ein gern gesehener Gast zu sein.

Meine Eltern haben durch ihre Unterstützung dieses Studium erst möglich gemacht. Dank gebührt ihnen aber auch für ihr Interesse an meiner Arbeit und für die Tatsache, daß sie mir nie das Gefühl gegeben haben, ihnen auf der Tasche zu liegen.

Meiner Freundin Petra Wimmer danke ich für ihre Geduld und ihre Unterstützung beim Zustandekommen dieser Arbeit und vor allem dafür, daß Sie einfach da ist, wenn ich sie brauche.

Abstract

Two families of small RNA hairpins were investigated using the molecular mechanics programs AMBER 4.0 and JUMNA together with several optimization algorithms for internal energy minimizations. In both cases the three-dimensional geometry of different loop sequences were calculated with the aim of comparing the results to experimental data.

- NUN-triloops: 10 different NUN-triloops were geometry optimized starting only from the secondary structure to compare different optimization algorithms. Though the minimum geometries are different, some common features like extended stacking in the loop and hydrogen bonding between bases in the loop and backbone atoms were found. Since NMR-studies indicated that NUN-triloops in solution form dimers at higher concentrations, duplex structures were included as well. The results of the optimization of monomer and dimer structures in solution were able to reproduce experimental stability data in qualitative terms.
- GNNA-tetraloops: Hairpins with GNRA consensus sequence are common, they show exceptional thermodynamic stability, and they are conserved in evolution. NMR spectroscopic investigations revealed a stem-closing G-A base pair as the most prominent structural feature of GNRA hairpins. This base pair and other structural regularities were also found in the computer generated conformations with lowest internal energies determined without structural constraints derived from experimental data. GNYA-loops were included in this study to clarify the unusual stability of their GNRA relatives. 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 on the sequences of the middle bases. 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 the GNYA loops.

Zusammenfassung

Die dreidimensionale Geometrie zweier Sequenzfamilien von RNA hairpins wurden mittels den Molecular Mechanics Programmen AMBER 4.0 und JUMNA durch Energieminimisierung optimiert. In beiden Fällen sollten die Resultate mit experimentellen Ergebnissen verglichen werden.

- NUN-Triloops: 10 verschiedene NUN-Triloopsequenzen wurden geometrieoptimiert, wobei von der Sekundärstruktur ausgegangen wurde, um verschiedene Optimierungsalgorithmen vergleichen zu können. Obwohl die Minimumsgeometrien der einzelnen Sequenzen natürlich verschieden sind, konnten einige gemeinsame, sequenzunabhängige Strukturmerkmale wie das Fortsetzen des Stackings in die Loopregion oder Wasserstoffbrückenbindungen zwischen Basen im Loop und dem Backbone gefunden werden. Da NMR-Untersuchungen gezeigt haben, daß NUN-Triloops in Lösung bei höheren Konzentrationen als Dimere vorliegen, wurden auch Duplexstrukturen untersucht. Eine Optimierung von Monomeren und Dimeren in Lösung konnte die Stabilitätsdaten der NMR qualitativ reproduzieren.
- GNNA-Tetraloops: Hairpins vom Sequenztyp GNRA zeichnen sich durch große Häufigkeit, hohe thermodynamische Stabilität und Erhaltung während der Evolution aus. NMR-Untersuchungen haben gezeigt, daß ein G-A Basenpaar in der Loopregion gebildet wird. Dieses Basenpaar und andere strukturelle Merkmale wurden auch in den Konfomeren mit der niedrigsten freien Energie wiedergefunden, obwohl während der Optimierung keinerlei Einschränkungen gegeben waren. Um die erhöhte Stabilität der GNRA-Loops gegenüber der GNYA-Loops zu klären, wurden auch die GNYA-Loops geometrieoptimiert. Ein Vergleich der Minimumsstrukturen aller GNNA-loops zeigt überraschend wenig Veränderung der Geometrie des G-A Paares bei einer Variation der mittleren Basen. Eine „Öffnung" des Loops zur Aufhebung der Ringspannung und eine nachfolgenden Optimierung der „offenen" Strukturen zeigt jedoch eine höhere Stabilität der GNRA-Loops gegenüber den GNYA-Sequenzen.

1 Introduction

The prediction of experimental data by theoretical means is probably one of the oldest goals in modern science. In chemistry many topics invite computational investigation, like the simulation of reaction kinetics, the evaluation and visualization of experimental data and, last but not least, the prediction of molecular structures. Two strategies can be distinguished in the field of structure prediction: One of them is heading for the most exact results possible based upon the methods of quantum chemistry. They require enormous amounts of computer resources and time, and consequently they are restricted to very small systems. The other strategy aims at larger structures but has to omit detailed information in order to make calculations feasible. Every strategy and method has its pros and cons but for the simulation of molecules of biochemical relevant size there is currently only one type of technique based on simple Newtonian mechanics using minimizations of 'force-fields'.

Biochemistry nowadays is dealing with two dominant classes of molecules: proteins and nucleic acids and interactions between them. The importance of nucleic acids for biochemistry and for life itself, however, was not recognized completely until 1944, when Avery, MacLeod and McCarty [1] discovered that a '*nucleic acid of the deoxyribose type is the fundamental unit of the transforming principle of Pneumococcus Type III*'. Up to that time it was believed that genetic information was carried by proteins and that nucleic acids play only a secondary role. But it was not until 1953, when Watson and Crick [2][3] presented their model for the DNA double helix, that gene function in molecular terms could be understood. Since then many other functionalities of nucleic acids, of DNA as well as RNA, were found, the most recent example being the discovery of catalytic function of RNA, the so-called 'ribozymes' [4][5].

For all biochemical active molecules a certain three-dimensional structure of at least part of the molecule is essential to the correct functioning. The investigation of the so-called 'active sites' of molecules and their structures is one of the most important tasks of biochemistry. In this work the objective was to investigate the three-dimensional structure of small RNA molecules by computational means. The molecular systems chosen are of relatively small size since the aim of this work was not the refinement of existing experimental data but on the contrary the *de novo* prediction of three-dimensional structures on an atomic level starting only from the sequence and the secondary structure. This study can be seen as the beginning of the second step of a two-step-process: From sequence to secondary structure and then on to the three-dimensional geometry of RNA molecules. The first step, the prediction of RNA secondary structure starting from its sequence, is a main topic in our group for several years, to sample the possibilities and problems of the second step was the essential goal of the thesis in hand.

2 Structural investigations on biopolymers

2.1 Structural features of nucleic acids

2.1.1 Definition of nucleic acid structure

RNA and DNA are both biopolymers built from a very limited range of monomers. In the case of proteins there are 20 amino acids, in the case of nucleic acids only four essentially different monomers are used. Each monomer, or nucleotide, consists of three molecular fragments: sugar, heterocycle, and phosphate. The sugar 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. 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). Figure 2.1 shows a short strand of RNA (sequence AGUC). Four monomers are connected to a single strand, which is directional and starts at the 5$'$-end (top left of figure 2.1) and ends at the 3$'$-end (bottom right of figure 2.1). RNA structure can be defined in three consecutive ‘steps’ each adding more structural details and information. The first of these steps is the sequence which simply gives the order of the nucleotides starting at the 5$'$-end and ending at the 3$'$-end. The next step is the so-called secondary structure which shows which bases are paired to others and which are unpaired.

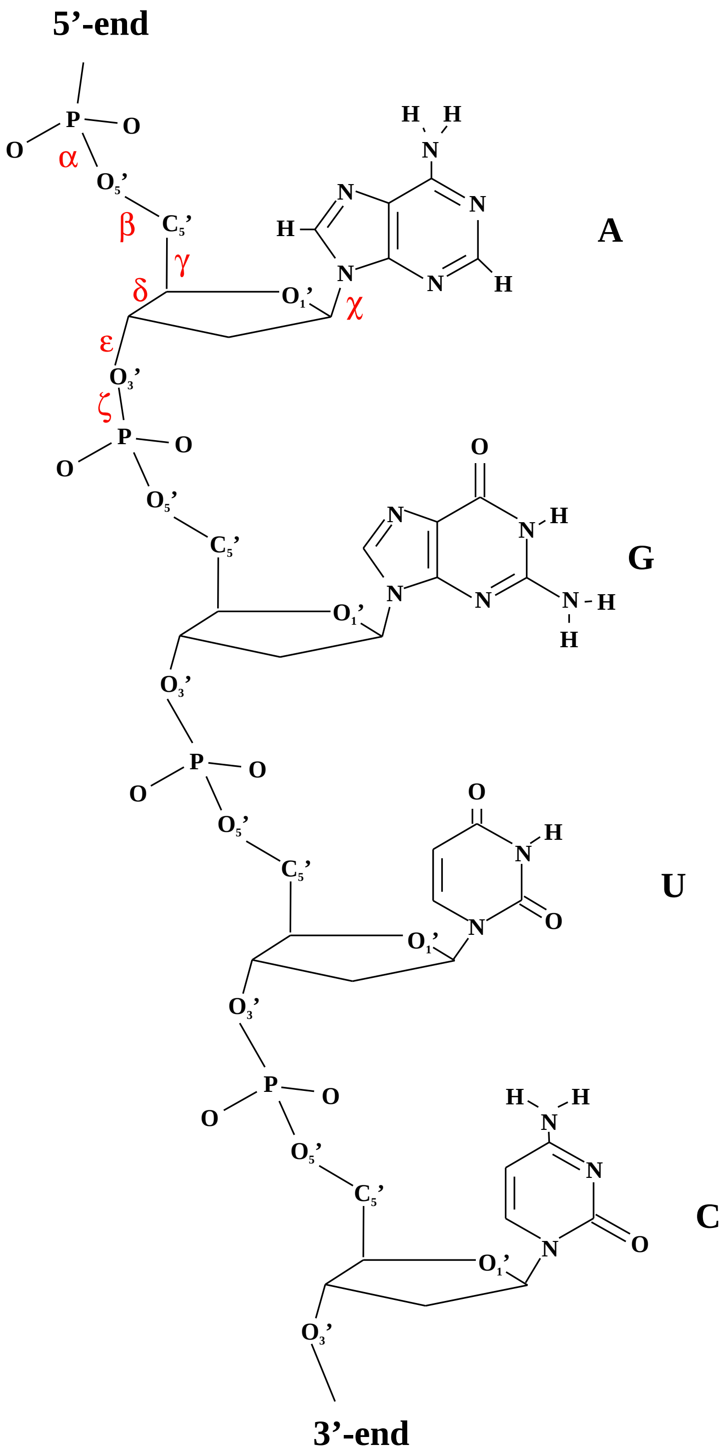


Fig. 2.1: Atomic structure of an RNA.

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 2.2). The most abundant of these motifs are the so-called hairpins (figure 2.2 b) 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'.

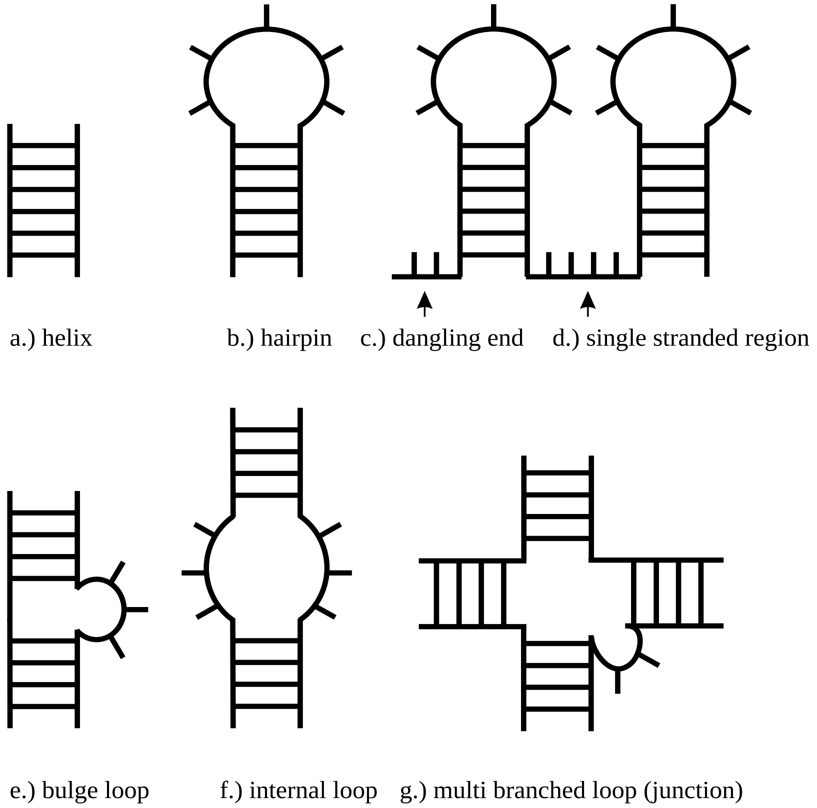


Fig. 2.2: Secondary structure motifs in RNA.

The next step in improving structural information is the definition of a tertiary structure. In essence the tertiary 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 tertiary structure is considered. The most prominent examples are pseudoknots [6][7], base triples [8]-[13] and most recently G-quartets [14]-[17]. The range of possible secondary and tertiary structural elements is rather large, proving that RNAs are very flexible molecules. Since both the sugars and - even more - the heterocycles are very rigid structures, most of the conformational flexibility comes from the backbone. In figure 2.1 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 three-dimensional 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 2.3 shows two of the most frequent sugar-puckers in RNA, C2′-endo (left-hand-side of figure 2.3) and C3′-endo (right-hand-side of figure 2.3). 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.

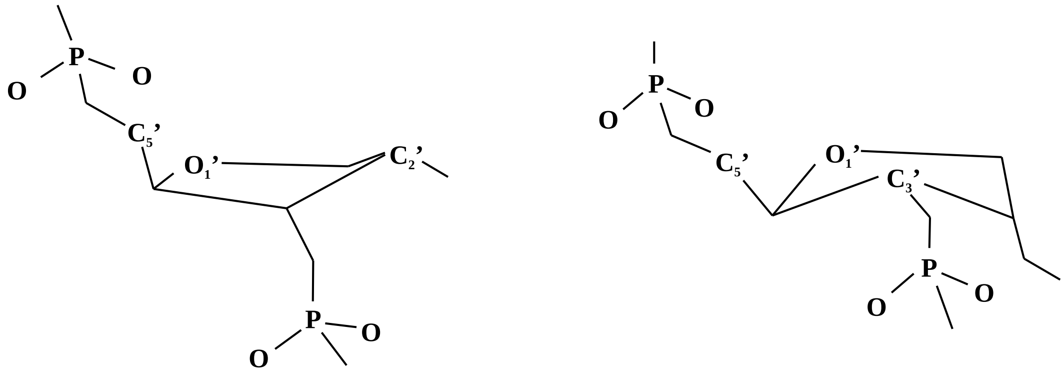


Fig. 2.3: Major puckering modes of sugars in RNA (left-hand-side: C2′-endo, right-hand-side: C3′-endo).

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 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 Wolfram Saenger's 'Principles of Nucleic Acid Structure' [18].

2.1.2 Energy landscapes

The energy of a molecule is a function of the positions of its atoms; the hyper-dimensional plot of the energy versus the positions of the atoms is referred to as an energy landscape. Without any approximation the number of degrees of freedom in a non-linear molecule is $3N - 6$, with N being the number of atoms. This leads to very complex energy landscapes even in the case of small molecules, for RNA molecules of a reasonable size some kind of approximation has to be found. As shown in the previous section most of the flexibility of an RNA molecule stems from the torsions along its backbone and from the orientation of the heterocycle with respect to the sugar. A simple example shows the large number of possible conformations even under very restrictive conditions. For the calculation of a four-membered hairpin loop (a so-called 'tetraloop') it is assumed that the stem is completely rigid and only the loop area is flexible. This leads to 28 degrees of freedom (seven for each of the four bases) for the whole molecule. Even if each of the torsional angles is allowed to adopt only two different values this leads to $(2^7)^4 = 268,435,456$ possible conformations. Of course not all of these solutions would fulfill the criteria of loop closure, but on the other hand allowing only two values per torsional angle is very restrictive.

Because of the great number of degrees of freedom energy landscapes of nucleic acids are highly complex. The internal energy of the molecule is not only a function of many variables, it can also possess the properties of a so-called *rugged* landscape known from many other optimization problems in molecular biology or physics, where so-called ‘cost functions’ are considered. The cost function of a rugged landscape is characterized by the fact that neighboring vectors x and $x + \delta x$ may belong to very different values of the cost function. Similar problems can be found for example in the theory of spin glasses or the theory of biological evolution. The energy landscapes of proteins were studied in great detail by Hans Frauenfelder et al. [19]. The investigation of Myoglobin and its binding reaction to O_2 showed several interesting results. It was found that Myoglobin is not active in only one specific conformation, but rather in several active conformations which were slightly but distinctly different regarding its binding rate binding rate. These conformations or microstates are referred to as ‘substates’. Further investigation of the highly complex energy landscape of Myoglobin showed that within each substate there are more ‘tiers’ of further substates, their number becoming so large that they cannot be characterized individually anymore, but must be described using distribution functions. The uppermost substates can still be seen as different energy levels of the molecule, the lower the considered tier lies, the more a near-continuum of energy states is reached. Several tiers could be observed by examining the inhomogeneous broadening of lines in the infrared spectra and from ‘hole burning’ experiments. Frauenfelder et al. compare this perspective of successive tiers of substates within a substate to the description of terrestrial landscapes where the higher peaks and deeper valleys are worthy of separate names (corresponding to the upper-most tier of substates), but where less specific terms such as ‘roughness’ or ‘difficulty’ characterize the myriad of smaller scale features. The experimental results were also corroborated by molecular dynamics simulations, however the presently accessible time-scales of computation do not allow a complete sampling of protein substates and infrequent transitions between substates are a major source of statistical error in current molecular dynamics simulation [19]. Though these experiments were done on proteins there is no reason why the results should not also be true for other biological macromolecules and especially for nucleic acids. A detailed knowledge of the energy levels and the laws governing them would be essential to perform dynamics on these complex energy landscapes.

2.2 Experimental techniques

The biological and biochemical importance of RNA has long been seen only in storage and propagation of genetic information on the way from DNA to proteins. This view has changed drastically with the findings of Cech et al. [4] who discovered catalytic activity in a precursor of ribosomal RNA of Tetrahymena. Afterwards many other examples for catalytic active RNAs were found (see for example [20]-[24]). The functional diversity of RNA reflects diversity in its three-dimensional structure: knowledge of the three-dimensional structure and general rules for RNA folding will be invaluable for deducing more detailed mechanisms of all RNA functions. In principle the prediction of the three-dimensional structure of RNA is a two step process: from the primary structure (the sequence) to the secondary structure (the folding pattern) and from there to the tertiary structure. Whereas for the first step a variety of experimental methods like cross-linking, footprinting or gel-electrophoresis as well as theoretical methods (secondary structure prediction using combinatorial [25], Monte-Carlo [26][27], or dynamic programming algorithms [28]) is available, the choice of methods for determination and prediction of the tertiary structure is much more limited. Results obtained with the two most important experimental methods to determinate the tertiary structure of RNA, X-ray diffraction and NMR-methods are described shortly in sections 2.2.1 and 2.2.2, respectively, the prediction of RNA structure by computational methods is described in more detail in chapter 3.

2.2.1 X-ray crystallography

The principle of x-ray crystallography is rather simple, basically three components are needed: a source of x-rays, a crystal and a detector. X-rays can for example be produced by accelerating electrons against a copper target. A narrow beam of x-rays is directed on the crystal where it is scattered by the electrons in the molecule. The scattered (of diffracted) beams can be detected by x-ray film or by a solid state electronic detector. Usually the most difficult part in doing an x-ray investigation on a biochemical macromolecule lies in obtaining the crystal itself. This is especially true for RNAs (see next paragraph). From the diffraction pattern an electron density map can be calculated from which the atomic structure - provided the resolution is high enough - can be derived. The calculation of the electron density map can be done in two ways, namely isomorphous replacement and molecular replacement. With isomorphous replacement a heavy atom derivative (e.g. made by replacing cytosine with bromocytosine or replacing magnesium with lead) of a molecules that crystallizes in the same unit cell as the original molecule is used. Molecular replacement is a computational technique where a model or part of the structure is randomly oriented within the unit cell until an optimum match with the diffraction data is achieved [29]. The rest of the structure is then solved iteratively. Molecular replacement is particularly appropriate for RNA, since the stems of RNA are usually close to the canonical A-form. As already mentioned above the critical factor in x-ray crystallography is resolution, it is determined by the 'quality' of the crystal meaning its size and its purity (the lack of solvent inclusions). The higher the resolution the more detail of the molecule becomes 'visible'; a resolution of 6 Å for example would only reveal the course of the backbone, in a resolution of 3 - 4 Å groups of atoms would be visible, whereas a resolution of 1.5 - 2.0 Å allows single atoms to be positioned. Another problem of x-ray crystallography is the fact that the three-dimensional structure of an RNA molecule in solution and in the crystal might be different as a comparison between NMR and x-ray structures for pGGAC(UUCG)GUCC seems to indicate [30].

Though x-ray diffraction has provided many structures at atomic resolutions for proteins as well as DNAs, the number of results for RNAs is very small. Structures of single crystals of RNA have so far be obtained for ApU [31] and GpC [32], for several tRNAs (e.g. $tRNA^{Phe}$ [33][34], $tRNA^{Asp}$ [35], $tRNA^{Gly}$ [36], $tRNA^{Fmet}$ [37][38], and $tRNA^{Imet}$ [39]), and most recently for a hammerhead ribozyme [40]. Especially the single-crystal structures of the tRNAs have provided a wealth of stereo-chemical information and have shown, that RNA can fold into a complex tertiary structure similar to that of a protein. The overall shape of tRNA was found to be the now familiar L-shape, with the anticodon loop and the acceptor-end at opposite ends. The four helixes agree largely with average helical parameters deduced from diffraction studies of A-form fibers [41][42]. The secondary structure motifs found include helixes with Watson-Crick base pairs but also a G-U mismatch, hairpin loops, and a multi-loop. Other results included the coaxial stacking of helixes, several unusual base pairs like G-A, a Hoogsteen A-T, or various base triples. It was also found that out of 76 bases 72 are involved in base stacking. Many of these qualitative results can seen as ‘patterns’ or structural motifs for RNA (e.g. extended base stacking, coaxial positions of helixes, formation of unusual base pairs) which in turn can be used as an ‘educated guess’ when RNA molecules are modeled.

2.2.2 NMR-spectroscopy

NMR (**N**uclear **M**agnetic **R**esonance) techniques measure distances with through-space interactions (nuclear Overhauser effect, or NOE) as well as dihedral angles using through-bond interactions (J-coupling) [43]. Measurements of a sufficient number of these interactions can define a structure nearly as well as x-ray crystallography [44]. Optimally, NMR samples are 0.5 ml of solution at as high of a concentration as possible (usually around 2 mM) and of high purity. The need for relatively high concentrations can lead to problems with dimerization, especially in small RNA molecules. Several nuclei can be used for NMR investigations in natural abundance these are ^{1}H, ^{19}F, and ^{31}P, in enriched samples ^{2}H, ^{3}H, ^{13}C, and ^{15}N can be measured. These nuclei have characteristic peaks or resonances, which are sensitive to their environment. An introduction to NMR of biological macromolecules can be found in [43], and Varani and Tinoco give a review of RNA NMR techniques [45]. The highest resolution RNA NMR structures published, which have most of their protons assigned, are UUCG- [44] and GNRA-hairpins [46]. Published medium-resolution NMR structures in which many protons are not assigned include Helix I [47], and Loop E from 5S rRNA [48][49], a pseudoknot [50], a hairpin containing an A-C pair in the stem [51], the three base bulge from HIV-I TAR stem [52], an RNA G-quartet [16], and a minor groove triple [11].

Proton NMR: Nucleic acids have two types of protons, nonexchangeable and exchangeable. Nonexchangeable protons are bound to carbon, whereas exchangeable protons are bound to either nitrogen or oxygen and exchange rapidly with the surrounding water thus broadening the corresponding peaks. The rate of exchange of protons with solvent is pH dependent and can be influenced by the use of buffer species, which act as catalysts for proton exchange. Furthermore the rate of exchange is influenced by hydrogen bonding or decreased accessibility by the solvent so that the exchange rate is an important probe of structure and dynamics. Each base contains 10 or 11 protons, of which the H1′, the aromatic and amino, and the imino protons all have their own distinct regions of chemical shift, whereas the five remaining sugar protons resonate together in a relatively small region (approx. 1 ppm). Naturally these regions become crowded for larger molecules, but base composition and the three-dimensional structure of the RNA influence the chemical shift dispersion further. For example the GGAC(UUCG)GUCC hairpin structure investigated by Varani et al. [44] stacks the C7H5$''$ above the G8 base, causing a significant chemical shift of 1.5 ppm above the normal region.

Chemical shift dispersion problems can be solved at least partly by using multi-dimensional NMR techniques, like for example NOESY. NOESY (**NOE S**pectroscop**Y**) measures dipole-dipole interactions through space. The volume of a NOESY cross peak is related to the time the two proton dipoles interact (mixing time). The build-up rate σ is the slope of a plot of the NOESY cross peak volume *versus* mixing time. Distances are measured by comparing an unknown buildup rate σ_u with the buildup rate of two protons at a known and fixed distance, like for example that of a pyrimidine H5-H6 pair. Since the buildup rate is linear with r^{-6}, r being the distance between the two protons, its use is limited to a distance of 5Å . The NOE and the NOESY techniques are described for example by Neuhaus and Williamson [53].

Through-bond interactions (J-coupling) occur between nuclei separated by one or more chemical bonds. J-couplings are measured using 2D COSY (**CO**herence **S**pectroscop**Y**) experiments [43][54]. The size of the coupling constant depends sinusoidally on the dihedral angle between the two nuclei [55]. The dihedral angle is calculated from the coupling constant using an empirical equation (called a Karplus relation) fitted on model compounds. Three-bond coupling is the most common in RNA proton NMR, it can define the backbone torsion angle γ, the sugar pucker (the pseudorotation angle or the equivalent torsion angle δ), and the sugar pucker amplitude.

Heteronuclear NMR: The major problem of NMR studies of RNAs with more than 25 nucleotides is spectral overlap, distinct regions of the NMR spectra become so crowded that signal assignment is impossible. Similar problems occur in NMR measurements of proteins, but 3D and even 4D studies of ^{13}C and ^{15}N isotopically enriched proteins (80 - 150 amino acids) have shown a significantly reduced spectral overlap [56] - [58]. Isotopic enrichment is expected to produce similar advances in NMR studies of RNA [59][60]. Multidimensional heteronuclear NMR experiments are useful for a number of reasons: Chemical shifts are dispersed, so that spectral overlap is reduced and larger RNA molecules can be investigated, selective enrichment (e.g. only one strand is enriched) can simplify signal assignment, and the measurement of heteronuclear J-couplings like ^{1}H -^{31}P, ^{1}H -^{13}C, and ^{13}C -^{31}P allows the determination of backbone and glycosidic torsion angles. The lack of heteronuclear studies of RNA has been due to the lack of sample availability, however methods are now available [61][62] for preparing uniformly ^{13}C- and ^{15}N-enriched nucleoside triphosphates (NTPs) from nucleoside monophosphates (NMPs) derived from the RNA of cells grown on ^{13}C- and/or ^{15}N-enriched media [63][64]. Though the synthesis of enriched nucleotides is expensive both in cost and synthetical effort, this method will yield greatly improved structural information for larger RNA molecules than conventional NMR-techniques would allow.

3 Computational methods

3.1 Molecular Mechanics

3.1.1 Introduction

The term molecular mechanics is usually used synonymously for force field calculations, since the underlying principles are coming from Newtonian mechanics rather than from quantum chemistry. The basic idea is that bond length, valence angles and torsional angles have ‘natural’ values depending on the respective atoms and that molecules try to adjust their geometries to adopt these values as closely as possible. In addition, steric and electrostatic interactions, mainly represented by van der Waals and Coulomb forces, respectively, are included in the so-called ‘potential’. The basic ideas for these calculations go back to the work of Andrews in 1930 [65], the first serious applications of force field methods date back to 1946 [66]-[68].

One of the basic principles of molecular mechanics is the Born-Oppenheimer approximation [69] known from quantum mechanics, which states that for most cases the Schrödinger equation for a molecule can be separated into an electronic and a nuclear term. Within this approximation the motion of the electrons is independent from the nuclear motion. The energy is therefore a function of nuclear coordinates forming a so-called ‘potential surface’ on which each point corresponding to an energy minimum is referred to as a 'conformer'. A typical force field contains a set of several potential functions which themselves contain adjustable parameters that are optimized to obtain the best fit of calculated or experimental values, as geometries, conformational energies, heats of formation or spectroscopic properties. This implies that parameters and force constants can be transferred from one molecule to the other. It is important to notice that (1) this assumption cannot be proven and (2) force fields are usually only parameterized for a limited set of molecular properties (i.e. geometry and conformational energies or spectroscopic properties but usually not for both).

In contrast to the so-called 'ab-initio-methods' derived directly from the Schrödinger equation and the exact Hamiltonian, molecular mechanics is an empirical method for calculating molecular properties, so that experimental data is needed for a parameterization. These experimental values are usually taken from x-ray and electron diffraction measurements, but can also stem from quantum mechanical calculations of smaller molecules or molecular fragments. A good overview and introduction into the topic of molecular mechanics and force fields can be found for example in [70].

3.1.2 Force fields

Energy calculation: Force fields were originally developed for vibrational analysis of molecules. The Taylor series for the potential energy of a molecule with n atoms and $3n$ degrees of freedom can be seen in equation 3.1.1. Since the first term V_0 can be set to zero as it has a constant value for any given molecule and the second term is also zero since it is assumed that the molecule is in a potential minimum the only terms left are of quadratic or higher order.

$$\begin{aligned}V_{pot} = V_0 + \sum_{i=1}^{3n} (\frac{\delta V}{\delta x_i})_0 \Delta x_i + \frac{1}{2} \sum_{i,j=1}^{3n} (\frac{\delta^2 V}{\delta x_i \delta x_j})_0 \Delta x_i \Delta x_j + \\ + \frac{1}{6} \sum_{i,j,k=1}^{3n} (\frac{\delta^3 V}{\delta x_i \delta x_j \delta x_k})_0 \Delta x_i \Delta x_j \Delta x_k + higher\ terms\end{aligned} \tag{3.1.1}$$

For sufficiently small displacements from the minimum which are usually treated in vibrational analysis the terms of higher than quadratic order are neglected which yields the so-called 'harmonic approximation'. By replacing the second derivatives with their symbol f_{ij} a simple relationship for a harmonic force field is found (see equation 3.1.2).

$$V_{pot} = \frac{1}{2} \sum_{i,j=1}^{3n} f_{ij} \Delta x_i \Delta x_j \tag{3.1.2}$$

This harmonic formula is used in internal coordinates for deviation of the bond lengths, the valence angles, and the torsional angles in a molecule from the 'equilibrium' values (see equation 3.1.3). Here r_i, θ_k, and ω_l stand for the current value of the bonding variables, whereas r_{0i}, θ_{0k}, and ω_{0l} stand for the equilibrium values of these variables, depending on the types of atoms involved. The $f_{r,i}, f_{\theta,k}$, and $f_{\omega,l}$ are the corresponding force constants again depending on the atoms forming the bond.

$$V = \frac{1}{2} \sum_i f_{r,i}(r_i - r_{0i}) + \sum_k f_{\theta,k}(\theta_k - \theta_{0k}) + \sum_l f_{\omega,l}(\omega_l - \omega_{0l}) \tag{3.1.3}$$

Since van der Waals contributions are also important for the torsional terms in most force fields a term using a cosine function is used, resulting in an energy function with minima at staggered conformations and maxima at the eclipsed geometries (see equation 3.1.4).

$$V_{tor} = k_\omega(1 - \cos 3\omega) \tag{3.1.4}$$

All types of forces mentioned above can be summarized as so-called ‘bonded interactions’ since they are in effect along axes where one would draw a bond in a structural formula. The simplest model potentials are the harmonic potentials already mentioned above and the so-called Morse-potential (see equation 3.1.5), D_{eq} and β are two parameters which depend on the nature of the bonded atoms. The approximation of harmonic potentials are only valid in a very small region around the equilibrium value, since they do not give correct results, neither for $x \to 0$ nor for $x \to \infty$. This can be improved by the use of the Morse potential (after P. M. Morse) which shows the correct behavior at least for $x \to \infty$, but since it is more time consuming to calculate exponential functions and most molecules treated in molecular mechanics are assumed to be very close to equilibrium geometry, harmonic potentials are used in most force field programs.

$$U_{r-r_0} = D_{eq}(1 - e^{-\beta(r-r_0)})^2 \tag{3.1.5}$$

Another type of forces which do not act along bonds are the so-called ‘non-bonded-interactions’ which are in effect between atoms that are spatial near to each other but which are not necessarily bond together. These forces mainly consist of the van der Waals and the Coulomb interactions. Equation 3.1.6 shows the Lennard-Jones-potential for the van der Waals interaction whereas equation 3.1.7 gives the formula for Coulomb interaction between charged particles.

$$V_{vdW} = \epsilon\big[\big(\frac{r_0}{r}\big)^{12} - 2\big(\frac{r_0}{r}\big)^{6}\big] \tag{3.1.6}$$

$$V_{Col} = \frac{q_i q_j}{D r_{ij}} \tag{3.1.7}$$

The terms for non-bonded interactions are of great importance to the quality of a force field for two reasons. First these terms are responsible for most of the computational time needed since they are quadratic with the number of atoms. The second reason is that the correct simulation of electrostatic interactions is essential for a 'realistical' minimum structure, since these forces going with $\frac{1}{r_{ij}}$ are very long ranged and errors in cutting them of too soon can have serious influences on the molecular geometry. This is especially important in highly charged molecules like nucleic acids. Most of the molecules of biochemical importance have to be considered within a solvent rather than in vacuum. In principle this can be done in two ways; either the solvent molecules are treated explicitly or the dielectric constant has to be modified to mimic the presence (and thus the screening effect) of a solvent. Another possibility is to use a dielectric function that varies with distance as for example in the force field program JUMNA (see chapter 3.3).

In some force fields additional terms are added to account for dipole-dipole interactions or for hydrogen bonding. The total energy of a molecule is thus calculated by the formula seen in equation 3.1.8:

$$V_{tot} = \underbrace{V_{bond} + V_{ang} + V_{tors}}_{\text{bonded interactions}} + \underbrace{V_{vdW} + V_{Col} + \big(V_{HB} + V_{dipol} + \ldots\big)}_{\text{non-bonded interactions}} \tag{3.1.8}$$

Parameterization: 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 3.1.4 - 3.1.7 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 [71] -[76] 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. [77], [78]) 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.

3.1.3 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.

Minimization methods: The global minimum problem is known for a long time since it 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 [79] or analytically (see for example [80]). 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. [81]). 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

$$p < e^{-\frac{E_1 - E_0}{kT}}$$

where 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 [82]. 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 [83], but it can be adopted to geometry optimization of molecules as it was done by Eberhard von Kitzing for the AMBER force field [84][85].

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 3.1. Figure 3.1 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,

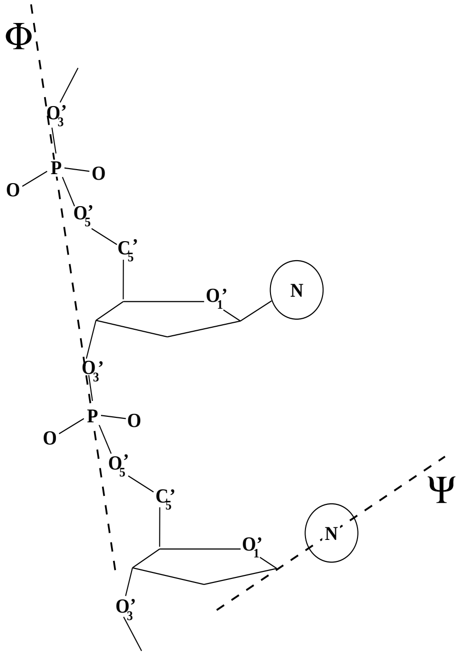


Fig. 3.1: Definition of rotational axes for the Bremermann method.

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 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 Molecular dynamics

Molecular dynamics (MD) can be seen as an extension to the concept of molecular mechanics in that sense that one is not only interested in the structure of a molecule at a given time but also in the structural changes of a molecule as a function of time. The underlying force field (the potential function) can be the same as for simple geometry optimizations, only an additional variable - time - has to be considered.

The most exact and detailed information is provided by molecular dynamics simulations in which Newton’s equation of motion are solved for the atoms of the system and any surrounding solvent (see for example [86]). The simulated time-span depends on the size of the system and the available computer-power but is generally in the range between pico- and nanoseconds. To begin a dynamic simulation an initial set of atomic coordinates and velocities are required. The coordinates can be obtained by using X-ray crystallography or NMR methods, or by model-building (for example a molecular structure optimized using the methods described above). The structure is first refined using an iterative minimization algorithm to relieve local stresses due to overlap of atoms or bond length distortions. This is especially important for structures based on experimental data since usually the position is not known for all of the atoms so that some have to be added by the MD-program. Next, atoms are assigned velocities v_i taken at random from a Maxwellian distribution for a low temperature and a simulation is performed for a few pico-seconds. This is done by finding the acceleration a_i of atom i from Newton's law $F_i = m_i a_i$. F_i, the force on the atom i is computed from the derivatives of equation 3.1.8 with respect of the position and m_i is the atomic mass. The accelerations a_i are then introduced in equation 3.1.9 to compute the position r_i of atom i at the time $t + \Delta t$.

$$r_i(t + \Delta t) = r_i(t) + v_i \Delta t + \frac{1}{2} a_i (\Delta t)^2 + \ldots \tag{3.1.9}$$

The equilibration is continued by alternating new velocity assignments, chosen from Maxwellian distribution for temperatures that are successively increased to some chosen value, with intervals for dynamical relaxation. The temperature T of the system is measured by the mean kinetic energy (see equation 3.1.10, N is the number of atoms in the system, $\langle v_i^2 \rangle$ is the average velocity squared for the ith atom and k_B is the Boltzmann constant).

$$\frac{1}{2} \sum_{i=1}^{N} m_i \langle v_i^2 \rangle = \frac{3}{2} N k_B T \tag{3.1.10}$$

The equilibration period is considered finished when the temperature is stable for longer than about 10 ps, the atomic momenta obey a Maxwellian distribution and different regions of the molecule have the same average temperature. For reviews on MD-methods see for example [86] or [87].

3.2 AMBER 4.0

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 [88] in the versions 3.1 and 4.0 (mainly), written by Kollman, Weiner et al. AMBER 4.0 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. AMBER 4.0 is comprised of several modules that fulfill specific tasks; figure 3.2 shows the flow of information between the different AMBER modules.

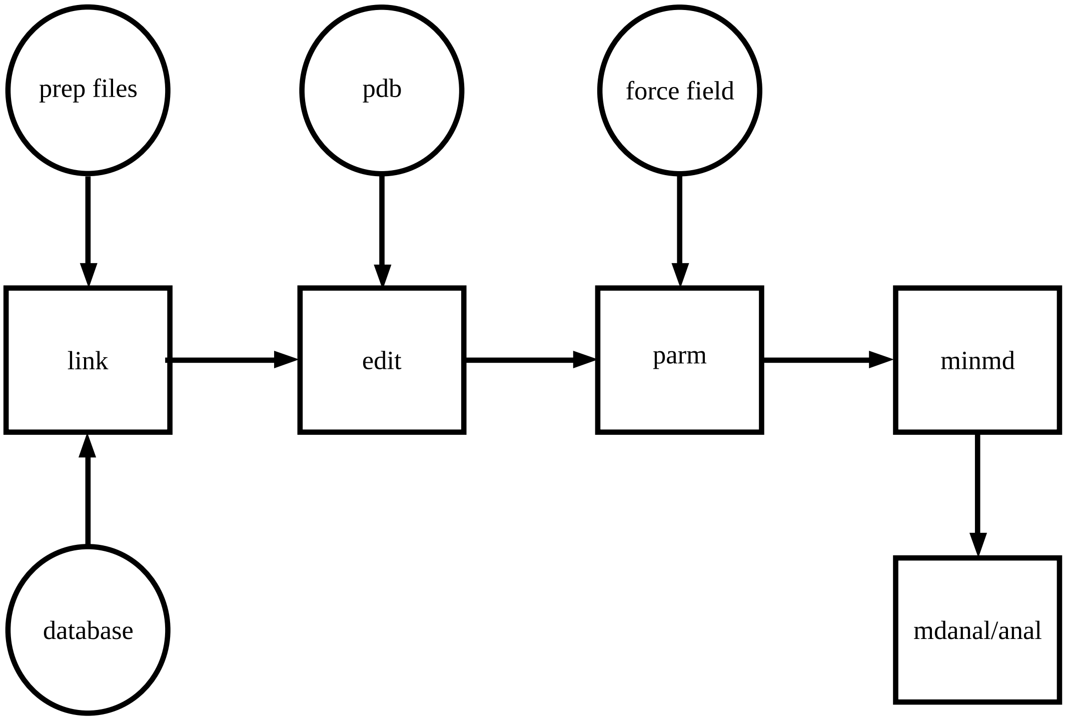


Fig. 3.2: Basic information flow in AMBER 4.0.

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', 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 3.2 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 molecular dynamics 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 (which atoms is connected to which) and positional information makes AMBER ideal for experiment with new optimization algorithms as has been done in the Bremermann method or with conformational sampling (see section 3.5.1).

3.3 JUMNA 7

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 [89] - [92] 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 [89, 93, 94]) 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 [95]). 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 experimental 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.

3.4 MC-SYM

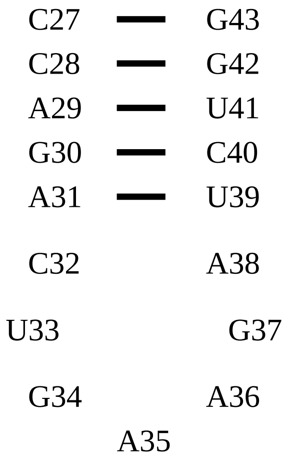
MC-SYM stands for **M**acromolecular **C**onformations by **SYM**bolic programming and is not a force field program but a tool to obtain three-dimensional nucleic acid structures which are in accordance to a list of input constraints. The program was written and tested by the group of Cedergren and Gautheret (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 [96, 97], for tRNAs [98], or for the Rev-binding site of HIV-1 [99].

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 3.3). The molecule modeled is the anticodon stem-loop of a tRNA.

The secondary structure shown on the left-hand side of figure 3.3 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 3.3. 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.



```
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
```

Fig. 3.3: Input file for MC-SYM for a simple stem-loop structure.

- *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 3.3 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.

3.5 Conformational sampling

3.5.1 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 the author 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 AMBER 4.0 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 3.4).

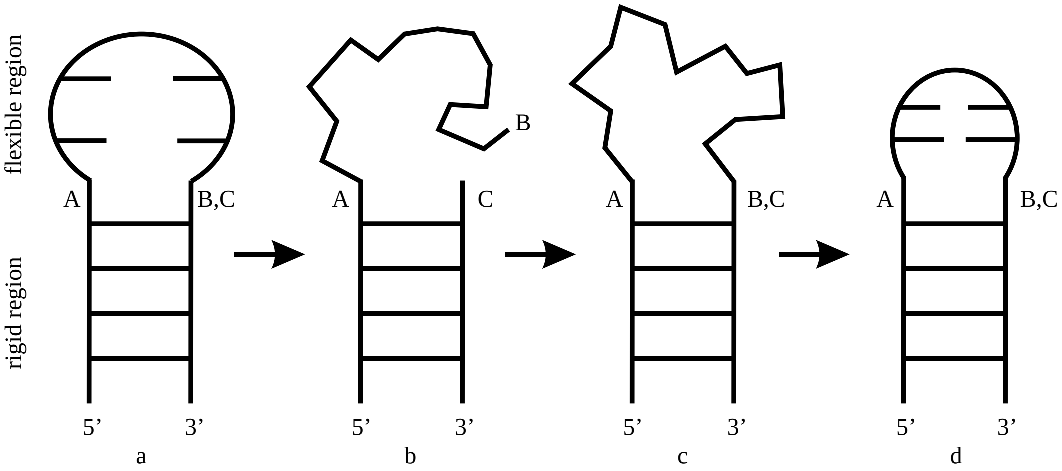


Fig. 3.4: Schematic representation of an optimization using *randstruct*.

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 3.4. 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 3.5 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). Only if the energy after this short minimization is lower than a given value (also supplied via command line; in figure 3.5 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

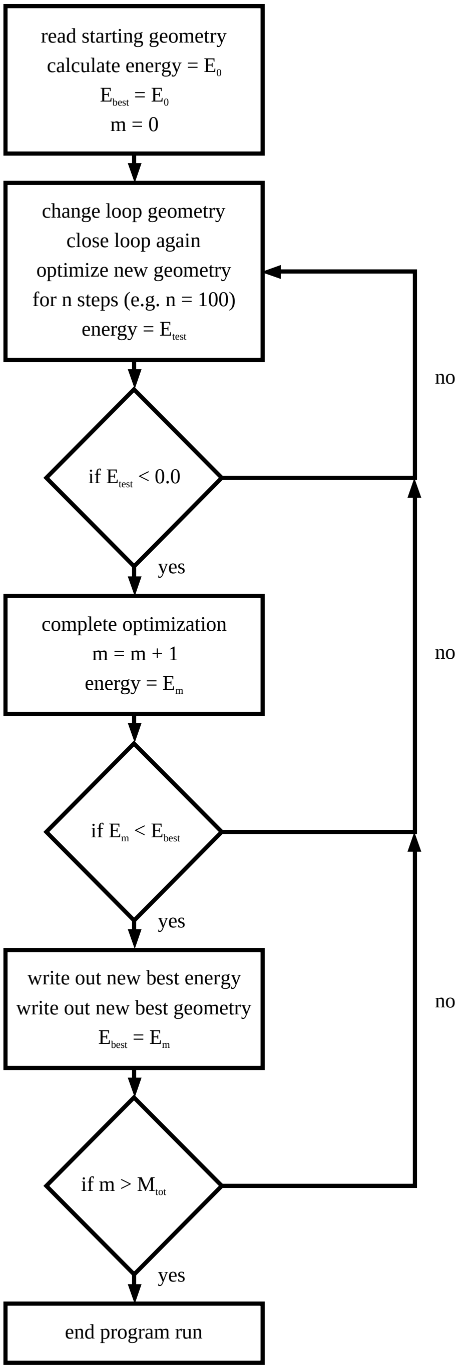


Fig. 3.5: Flowchart of the program *randstruct*.

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.

3.5.2 Randloop

An even simpler approach to investigate the geometry of small RNA molecules and especially loop regions is made by the program *randloop* (from **rand**om **loop**). The very basic requirement for the formation of a loop in an RNA molecule is that the backbone of the loop region is able to connect the 5′- and 3′-sides of the stem, respectively. This situation is shown in figure 3.6.

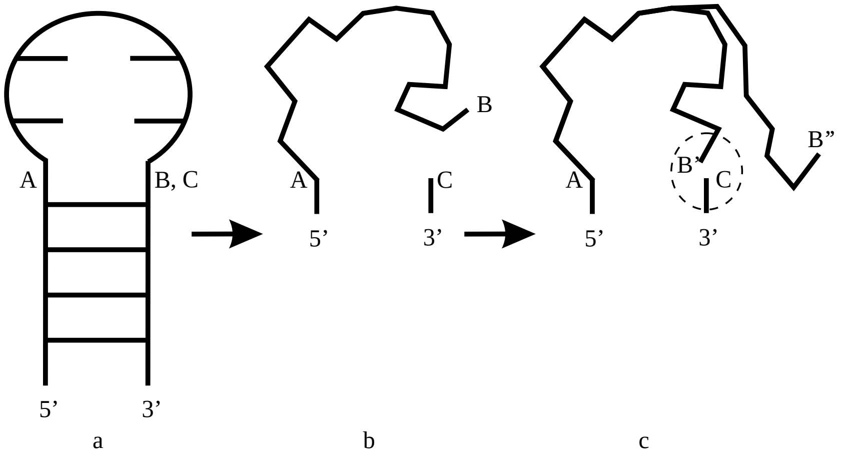


Fig. 3.6: Molecule representation in the program *randloop*.

Points A and C are the end and the beginning of the paired region, respectively, and can be considered as two points in three-dimensional space at a given distance. To facilitate a loop closure the backbone must form a series of vectors ending in point B that connects A with C. For the purpose of this program the rigid region of the molecule can be reduced to two parallel vectors at a given distance (see figure 3.6 b). This stem distance and the number of connecting bonds between A and B are parameters of the program as well as the bond length and the valence angles between the bonds. Both bond length and valence angles can either be the same for all bonds or can be defined individually for each bond to be able to include 'biochemical' values. Around the target (C) a sphere with a certain 'acceptance radius' (shown in figure 3.6 c as a dotted circle) is defined, since the chance to hit C exactly with a random set of vectors is infinitesimal small. The size of this radius is again a parameter of the method. Using random values for the torsional angles a vector starting from A is constructed. If the end point of this vector (B' or B'' in figure 3.6 c) lies in the acceptance sphere the vector is counted as a solution and the corresponding torsional angles are stored. In figure 3.6(c) an accepted solution (ending in B') and a not accepted vector (ending in B'') are shown. The program *randloop* was made to get an overview over the possible range of torsional angles and to see whether these angles are restricted in very confined systems by the requirement of loop closure. The program ends after a given number of tries and writes out the number of successful vectors, and the distribution of each torsion over the range of possible angles.

4 Results: GNNA-tetraloops

4.1 Introduction

Tetraloops are one variation of the structural motif of a hairpin. Hairpins with four bases in the loop are by far the most common type of stem-loop structures. Phylogenetic studies showed that in ribosomal RNA tetranucleotide loops constitute 55 % of all hairpins [100][101], forming a highly conserved region. The vast majority of tetraloops is covered by a very small number of sequences, among which GCAA, UUCG and - to a lesser degree - CUUG are the most prominent. GCAA- and UUCG-, or more generally GNRA- and UNCG- (N stands for any base and R for a purine base, either G or A) loops together make up 70 % of all tetraloops in 16S rRNA. Apart from the loop sequence the closing pair seems also to be important, since most of the conserved CUUG loops tend to have a G-C pair at the end of the stem, whereas in the case of UUCG loops the closing pair tends to be C-G. The sequence of C(UUCG)G seems to be particularly stable [102][103].

Three interrelated factors potentially influence the sequence of a loop: the physical stability of the hairpin structure as such, interactions of the loop with other parts of the RNA molecule (or other molecules), and the degree of selective pressure associated with a given sequence [100]. Therefore it was concluded that the loop sequences mentioned above do not only possess a high thermodynamic stability but also an important biochemical function. 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. It was suggested [104][105] 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. At least two instances of this behavior exist in the conserved core of group I self-splicing introns [104]. Jaeger et al. [106] published a series of experiments that support this model for a GUGA-loop in a self-splicing group I intron. Recently a model for the interaction between a GAAA-loop and a RNA-helix was published [107].

A very important factor for both the high stability and the biochemical functions of tetraloops are of course their three-dimensional structures. Experimental investigations concentrated mainly on the loop sequences mentioned above. Most available studies are a combination of NMR-methods and distance geometry calculations. Examples can be found in [30], [44], or [103] for the UUCG and the UUUG loops, respectively, and in [46] for the GNRA-loops. The x-ray structure of the hammerhead ribozyme published by Pley et al. [40] also contains a GAAA-loop. 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 actual loop size 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 very time consuming 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 quite severe steric constraints caused by the additional base pair in the loop. The first investigation of RNA three-dimensional structure at atomic level without using experimental data was done by Kajava and Rüterjans [108]. They investigated the conformations of the 16 possible NUUN tetraloops to examine the structure of the 'new' pair in the loop. For most of the tetraloops in question the molecular modeling approach yields a few energetical equivalent 3D-structures, so that a 'family' of conformations is obtained rather than a unique minimum energy geometry.

For the conformational analysis of this study a computational approach was chosen to examine the structure and stability of the GNRA-loops and their pyrimidine relatives. By calculating the lowest energy conformers for all 16 GNNA-loops it is intended to clarify the issue whether or not there are any specific structural and/or energetic features that give rise to the high stability of the GNRA-loops and that can explain their preferences over the GNYA-species (Y stands for pyrimidine bases).

4.2 Minimum structures

For the first part of this investigation the JUMNA program with the FLEX force field as described in chapter 3.3 was used. The starting point was not the secondary structure but the NMR-study by Heus and Pardi [46] and the three-dimensional structures of the GCAA- and GAAA-loops given therein. One of the experimental results was the formation of an additional base pair in the loop as described above, which in this case is the rather uncommon G-A pair. Starting from the qualitative description of the three-dimensional structure given in the literature, the molecular graphics program QUANTA 7 of 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. The sequence used as a starting point was GGGC(GCAA)GCCU (see figure 4.1) as described in [46] with just the dangling end on the 3′-side of the stem omitted.

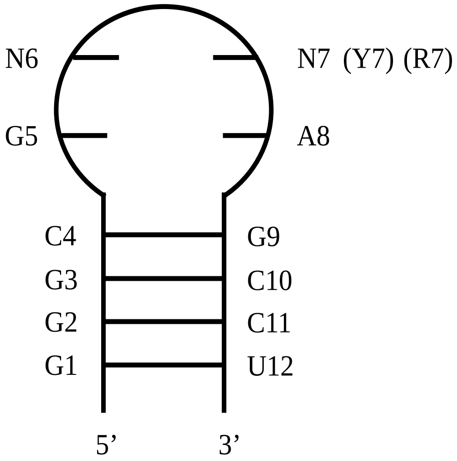


Fig. 4.1: Schematic representation of the GNNA-hairpin structure.
N = A, C, G, U
R = G, A
Y = C, U

This structure was optimized to relax any close contacts that have occured during the modeling process. Subsequently the sugars belonging to the four bases in the loop (G5, C6, A7, and A8) were forced to adopt C3′-endo, O1′-endo, and C2′-endo conformations by using the appropriate constraints for the amplitude and pseudo-rotation angle. The systematic construction of all possible combinations yields $3^4 = 81$ different conformations which were all energy minimized. The five conformations with the best (i.e. lowest) energies were chosen for further investigation. They are denoted in the following with A, B, C, D, and E, in order of increasing energy. Subsequently the geometries for the other 15 possible loop sequences were created by replacing

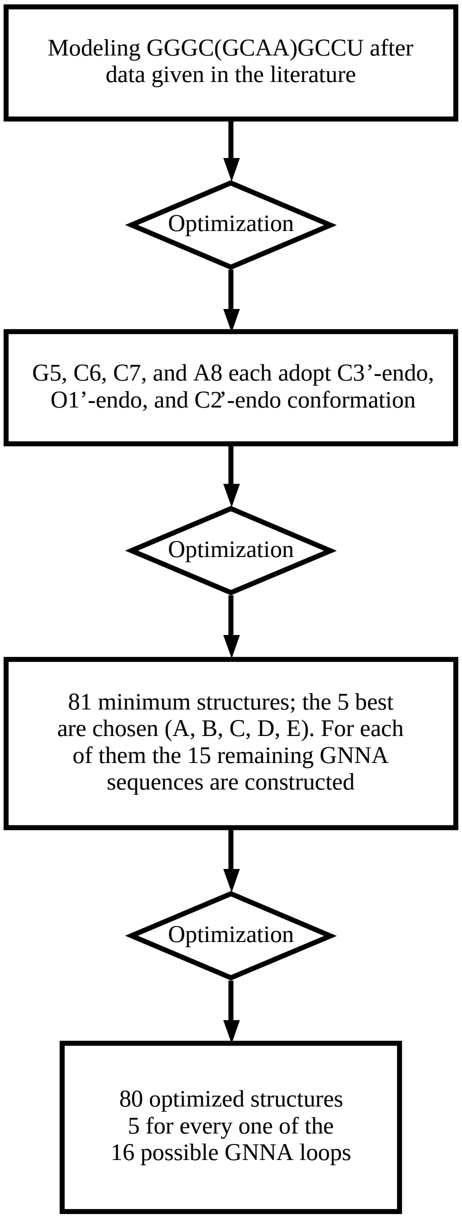

Modeling GGGC(GCAA)GCCU after data given in the literature
Optimization
G5, C6, C7, and A8 each adopt C3'-endo, O1'-endo, and C2'-endo conformation
Optimization
81 minimum structures; the 5 best are chosen (A, B, C, D, E). For each of them the 15 remaining GNNA sequences are constructed
Optimization
80 optimized structures 5 for every one of the 16 possible GNNA loops

Fig. 4.2: Flowchart for obtaining the GNNA minimum structures.

the appropriate bases in the minimum structures A - E. This can be done very easily in JUMNA because of the implemented sequence-independent representation of RNA-structure. The resulting structures were again optimized yielding the 80 different minimum geometries that are analyzed in the following section. Figure 4.2 gives an schematic overview of how the minimum structures were obtained.

4.3 Results and discussion

4.3.1 Structural features

Figure 4.3 shows the wireframe model of the GCAA-loop whereas figure 4.4 shows the same molecule in a space filling representation. Both figures exhibit structural details that are representative to all GNNA-loop geometries investigated and an analysis of the minimum geometries shows that there are several structural features that are common to all molecules regardless of their sequence.

- The G5–A8 base pair is conserved in all of the structures though the actual geometries of the base pairs may be quite different. In each sequence at least one of the hydrogen bonds is formed (whether or not a hydrogen bond was formed was decided via distance and angle criteria).
- Stacking is continued in the loop so that the G–A pair stacks on the closing pair of the stem. Furthermore bases N6 and N7 are positioned over the G-A pair, so that stacking is also continued from the 3′-end of the stem.
- The stem conformation is unaffected by the formation of the G-A pair; it is retaining the standard A-RNA helix.
- There is additional stabilization by forming a hydrogen bond between the HO2′-group and the N7 (or N4) of the third base in the loop.
- Sugar puckers are predominantly of N-type, if S-type sugars occur they are in the N6 position to stretch the backbone and to enable the closure of the loop.
- All nucleotides including those in the loop have anti-conformation.

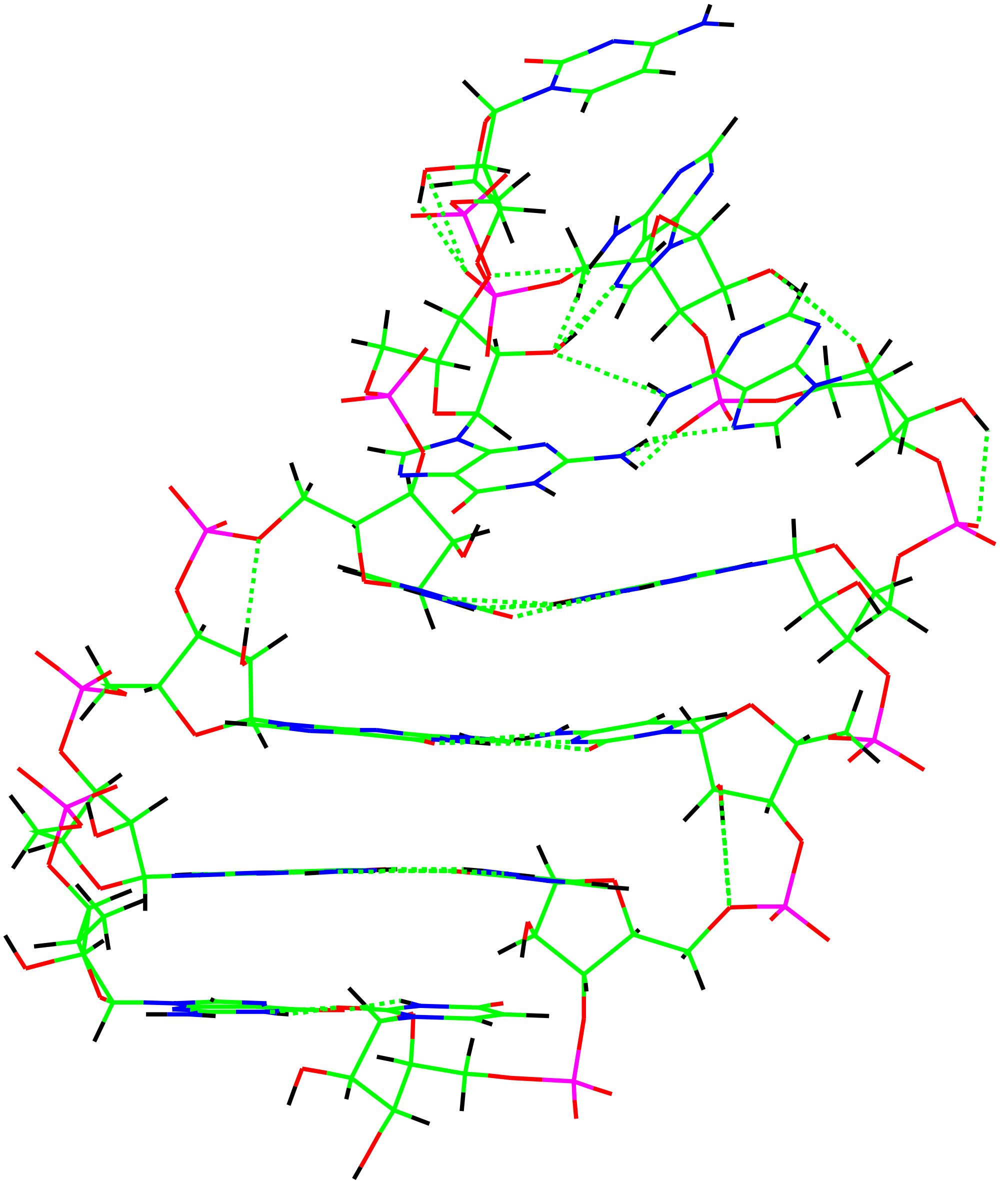


Fig. 4.3: Wireframe model of the optimized GCAA-structure. Dotted lines show hydrogen bonds.

These results are in strong agreement with the data presented in [46] and with qualitative results found for four-membered loops with other sequences. The overall dependence of the structure on a variation of the middle bases

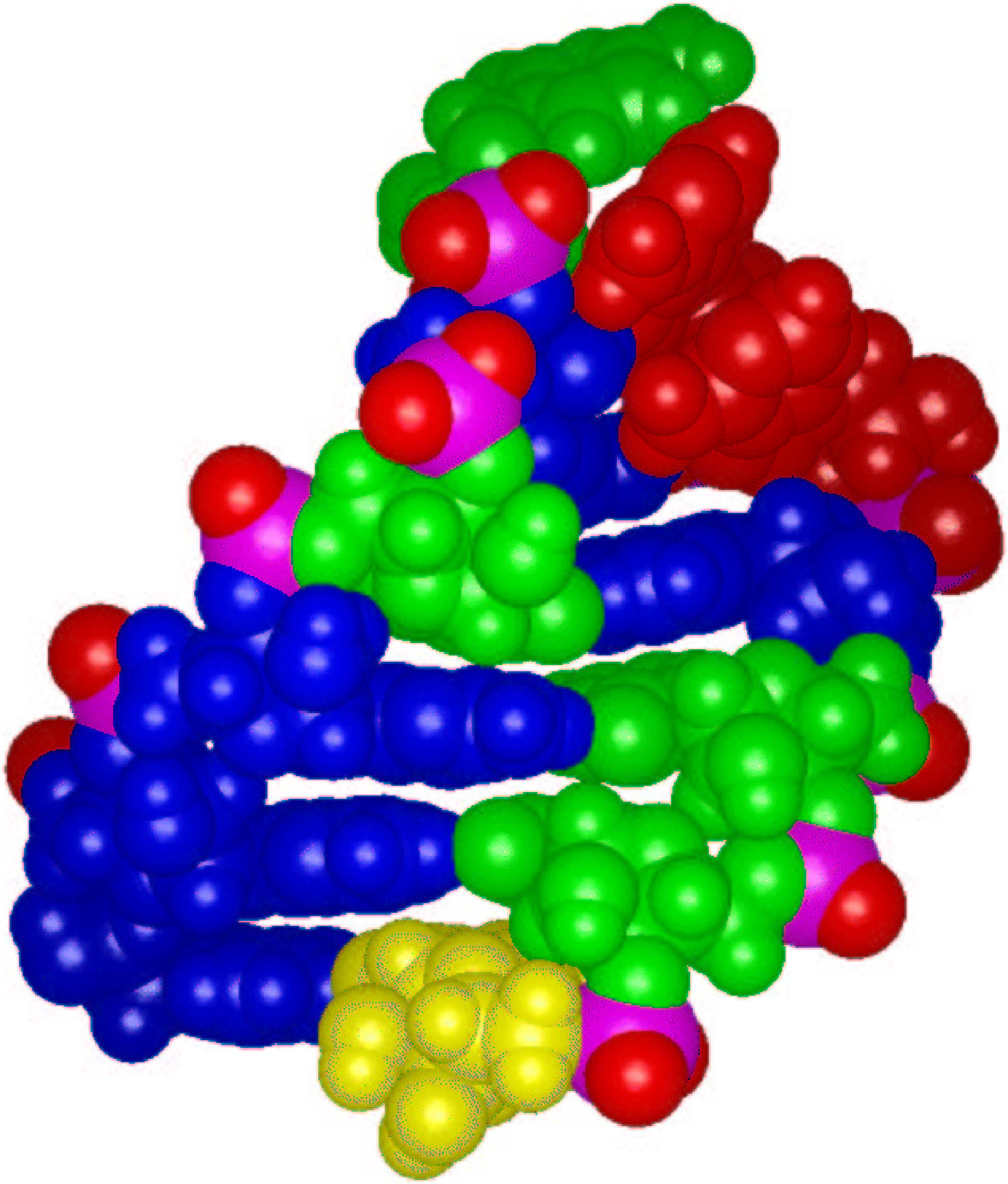


Fig. 4.4: CPK-model of the optimized GCAA-structure.

is surprisingly low, even in cases were pyrimidine nucleotides were replaced with purines. This of course is probably due to the fact, that the formation of the G-A pair makes the largest contribution to the stability of 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 interaction does not seem to be of significant magnitude.

The G-A base pair: since the initial identification of potential guanine-adenine interactions in ribosomal RNA [109][110], G-A pairs have been recognized as one of the most common non-canonical structural elements in rRNA [111]. Their three-dimensional structure has been extensively studied in DNA [112] - [117] and RNA [46], [48], or [117] - [119]. Several examples for G-A pairs can be found in the recently published x-ray structure of the hammerhead ribozyme. G-A pair functions are best characterized in the context of GNRA tetraloops, where they stabilize the loop [120] and expose functional groups to the solvent in a specific pattern that can be involved in protein recognition [121] or tertiary interactions [104][105]. Gautheret et al. [122] give a comprehensive overview over occurrences and structures of G-A mismatches in ribosomal RNA. They list five possible G-A pair structures which can be found in either helices or in loop-helix-junctions.

One of the most interesting features of the GNNA-loops is the formation of such a G-A base pair in the loop. The formation of an additional base pair within a tetraloop is also known for example from the CUUG- and the UNCG-loops, but in case of the GNNA-loops the loop formed is not of the common ‘pyrimidine-purine’ variety but rather between two purines. Since the space requirement of two purines is definitely larger than that of a canonical base pair and hence the steric strain on the loop must be strong it is even more surprising that comparing the minimum structures for all sequences shows that the G-A pair is conserved in all of the structures, although this was not forced through constraints on the base positions. Obviously the energy gained from the hydrogen bonds in the loop and from additional stacking is enough to make up for the somewhat crowded loop. Among the five different structures per sequence which were examined, only two distinct types of G-A pairs could be observed (see figures 4.5 and 4.6, both types of base pairs are shown from the side and from above), which will be referred to in the following as base pairs of type I and II, respectively.

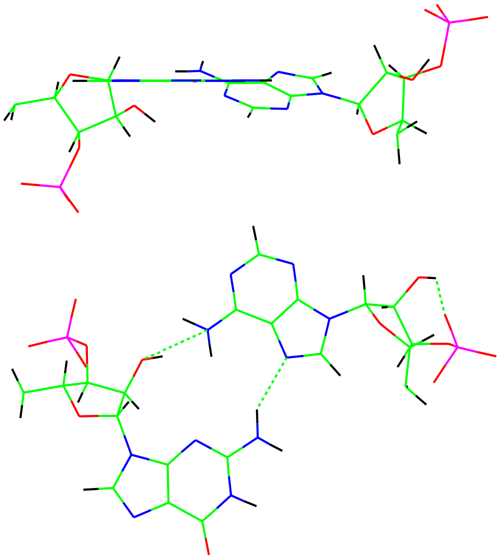


Fig. 4.5: G-A base pair of type I.

The structural differences can be seen from the respective figures: in a base pair of type I, the bases are in nearly coplanar positions and the base pair is ‘opened’ to one side; in a base pair of type II the nucleotides are obviously twisted with respect to each other.

Both types show two hydrogen bonds between G and A, one being formed between G5(H2N2) and A8(N7) and the other being formed between G5(N3) and A8(H2N6). In both types of the G-A pair an additional hydrogen bond is formed between the A8(HO2′) and an oxygen from the phosphorus group. Structures with a G-A mismatch of type II were found to be generally higher in energy than those of type I. The reason might be that the twisting

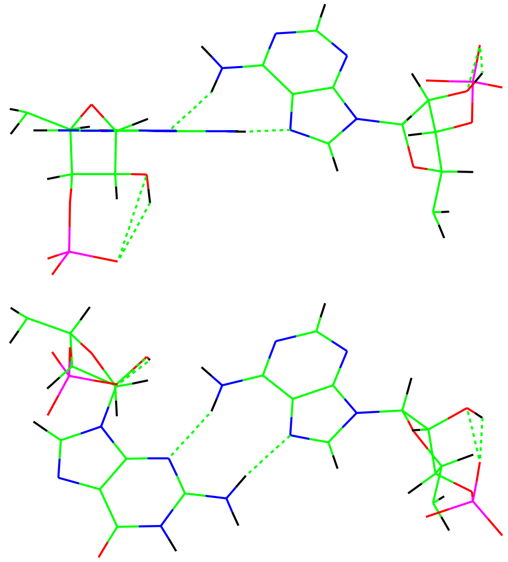


Fig. 4.6: G-A base pair of type II.

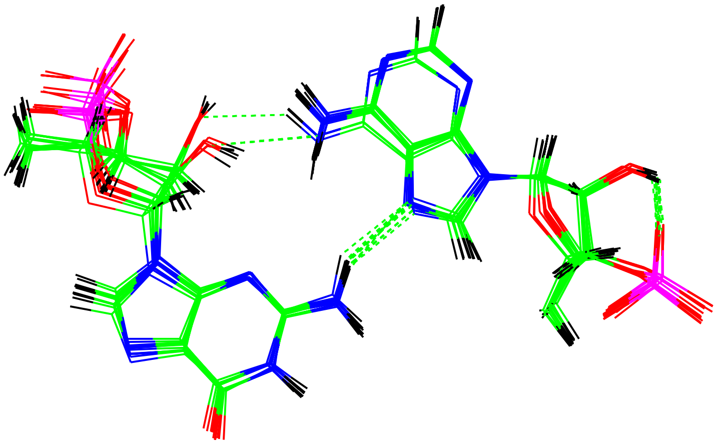


Fig. 4.7: Overlay of 16 G-A pairs of type I stemming from starting geometry A.

of the bases, that is necessary to facilitate a more compact loop, imposes stronger steric constraints on the two middle bases in the loop and weakens the hydrogen bonds. For both types of base pairs the somewhat surprising result is the fact that their geometry is virtually independent of the nature of

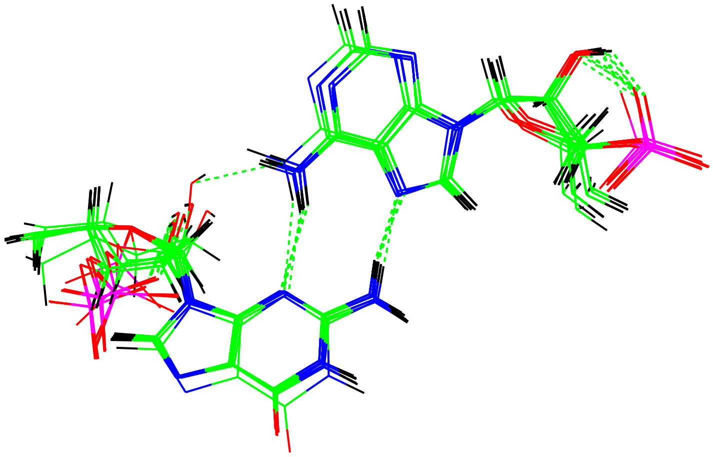


Fig. 4.8: Overlay of 16 G-A pairs of type I stemming from starting geometry E.

the middle bases. This can be seen from figures 4.7 and 4.8, in which 16 G-A pairs from minimum structures obtained from starting structures A (figure 4.7) and E (figure 4.8) are overlaid. A classification of the G-A pairs using the nomenclature introduced by Saenger [18] shows that both types belong to the class of asymmetric hetero purine base pairs (type X).

The middle bases: figures 4.9 and 4.10 show the loop structures of the minimum geometries for the GNRA-loops (figure 4.9) and the GNYA-loops (figure 4.10). In each figure the 8 minimum geometries obtained from starting structure A are displayed. As can be seen from these figures, bases N6, N7, and A8 show stacking on each other to some extent while there is virtually no interaction between N6 and G5. The observed tendency to maximize stacking from the stem into the loop was also found for example in the structures of the UNCG- and CUUG-loops and seems to represent a more general pattern, or rule, which is applicable to a great number of hairpin loops different in size and sequence. Again the variations in the structure with different sequences are surprisingly small, though they seem to be larger in GNYA-loops than in the GNRA-loops. The most obvious effect can be seen in the position of N6.

Clearly this base possesses the most flexible position because of its place on 'top' of the loop where there is little chance of interaction with other parts of the molecule. This explains also why the nature of base N6 is not essential to the overall structure. Figures 4.9 and 4.10 refer to the respective minimum geometries of each sequence containing a G-A pair of type I.

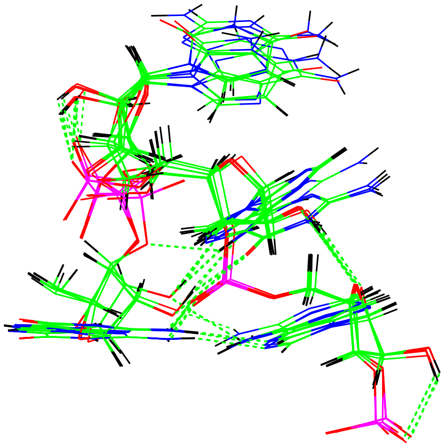


Fig. 4.9: Minimum structures for the GNRA-loops for G-A pair type I.

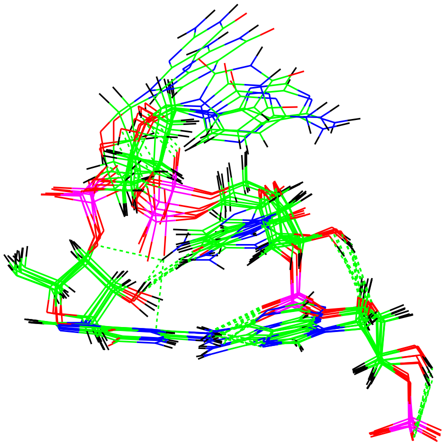


Fig. 4.10: Minimum structures for the GNYA-loops for G-A pair type I.

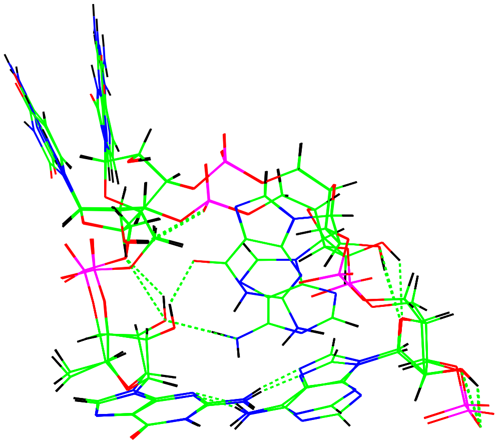


Fig. 4.11: Minimum structures for the GNRA-loops for G-A pair type II.

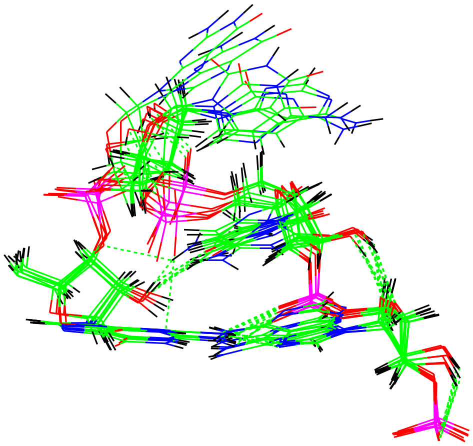


Fig. 4.12: Minimum structures for the GNYA-loops for G-A pair type II.

Figures 4.11 and 4.12 depict the loop structures for geometries with a G-A pair of type II. In this case the structural variations with different sequences are somewhat larger but nevertheless the most serious structural changes take place at the N6 position; the base in the second position in the loop is standing almost perpendicular to the plain of the G-A pair, so that stacking between N6 and N7 is not longer possible. In GNRA structures derived from a G-A pair of type II, two distinct groups of conformations can be observed depending on the nature of base N7 (see figure 4.11). Two groups, one for base C7 and one for base U7 can clearly be distinguished. The exposed positions of bases N6 and N7 and the higher conformational flexibility of these nucleotides may be of importance for forming intra-molecular contacts in the context of a larger molecule. The corresponding GNYA structures do not show a similar preference.

4.3.2 Energies

In total there are eighty different structures to be considered; a comparison of the minimum energies shows quite clear results: in the case of the GNRA- and the GNUA-sequences the lowest energy conformation is always reached from starting geometry A, only for the GNCA-sequences the lowest energy conformation is reached from starting geometry B (see table 4.1).

The energy differences between the geometries with the lowest and the highest energy within a given sequence are in the range of 6 kcal/mol for the GNRA-molecules but are significantly higher - 9 kcal/mol - for the GNYA-species, resembling the stronger structural variations of the sequences. Interesting dependences of minimum energies on the loop sequence can be observed by comparing the energies of G-A types I and II within a given sequence. For the GNRA-loops G-A pairs of type I are usually between 4.2 and 6.8 kcal/mol more stable than those of type II. For GNUA-loops this stabilization lies in the range from 8.4 to 11.8 kcal/mol whereas for GNCA the differences vary only between 2.0 and 3.7 kcal/mol (see table 4.2). A detailed energy analysis that splits the total minimum energy in its contributions of structural terms (bond stretching and valence angles), dihedral potentials, van-der-Waals interactions and electrostatic terms, shows that

Sequence	E(A)	E(B)	E(C)	E(D)	E(E)
GCAA	**-101.53**	-100.37	-98.88	-97.61	-95.39
GAAA	**-97.74**	-96.18	-97.30	-93.80	-92.81
GAGA	**-99.28**	-92.34	-98.26	-94.48	-93.74
GCGA	**-102.80**	-96.57	-100.13	-98.07	-97.32
GGAA	**-97.67**	-94.31	-97.32	-93.72	-93.52
GGGA	**-100.32**	-92.53	-98.16	-94.02	-93.48
GUAA	**-99.79**	-96.19	-99.57	-95.86	-94.41
GUGA	**-101.72**	-94.95	-100.51	-97.00	-96.04
GACA	-90.51	**-96.26**	-92.00	-86.22	-94.32
GAUA	**-96.64**	-92.67	-91.90	-91.65	-84.87
GCCA	-95.00	**-99.86**	-93.42	-93.71	-96.13
GCUA	**-97.85**	-96.38	-93.40	-92.87	-87.38
GGCA	-90.81	**-97.14**	-92.12	-86.84	-95.07
GGUA	**-94.03**	-93.25	-91.96	-89.33	-85.68
GUCA	-91.45	**-97.25**	-94.07	-87.54	-95.90
GUUA	**-94.77**	-93.94	-94.05	-89.81	-86.42

Minimum energies for each sequence are in bold face.

Table 4.1: Minimum energies for the GNNA loops in kcal/mol.

Sequence	E(G–A I)	E(G–A II)	ΔE
GCAA	-101.53	-95.39	6.14
GAAA	-97.74	-92.81	4.93
GAGA	-99.28	-93.74	5.54
GCGA	-102.80	-97.32	5.48
GGAA	-97.67	-93.52	4.15
GGGA	-100.32	-93.48	6.84
GUAA	-99.79	-94.41	5.38
GUGA	-101.72	-96.04	5.68
GACA	-96.26	-94.32	1.94
GAUA	-96.64	-84.87	11.77
GCCA	-99.86	-96.13	3.73
GCUA	-97.85	-87.38	10.47
GGCA	-97.14	-95.07	2.07
GGUA	-94.03	-85.68	8.35
GUCA	-97.25	-95.90	1.35
GUUA	-94.77	-86.42	8.35

Table 4.2: Energy differences between conformations of base pair types I and II in kcal/mol.

the large energy differences stem mostly from the contribution of the electrostatic energies. One reason for this is an additional hydrogen bond in the GNCA-loops between the O2′ and C7(N2H4) atoms. The energy differences between two conformations with the same base pair type on the other hand cannot be attributed to a single energy term, they are rather distributed over all energy terms.

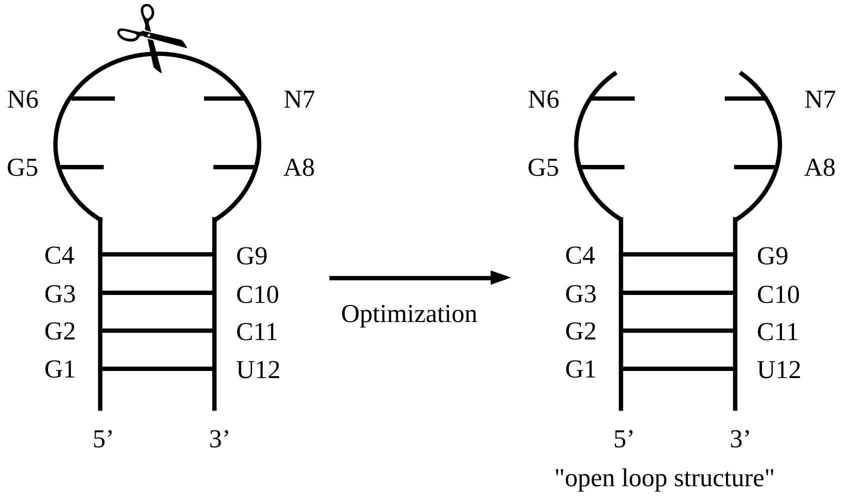


Fig. 4.13: Schematic representation for generating ‘open loop structures’.

A question that is even more interesting than the comparison of energies of different conformations belonging to the same sequence is the comparison of energies of molecules with different sequences. Normally this cannot be done since the energy of a molecule calculated by molecular mechanics depends on the number and the kind of bonds that occur, so that only the energies of different conformations can be compared directly. The obvious solution to this problem is the definition of a reference state, which makes it possible to discuss energy differences with respect to this reference state rather than comparing absolute energies. After some trials with different reference states such as single or double helices, an ‘open loop structure’ was chosen (see figure 4.13). To construct these loop structures, the P-O3′-bond between bases N6 and N7 was 'cut', the missing hydrogen and oxygen atoms inserted, and both halves of the molecules were treated as single strands that happen to form several base pairs. Another optimization was performed to allow for relaxation of the 'open loop'. In figure 4.14 the three-dimensional structure of a closed and an 'open loop' can be seen side by side for comparison. It is obvious that the structural changes caused by opening the bond are not dramatic.

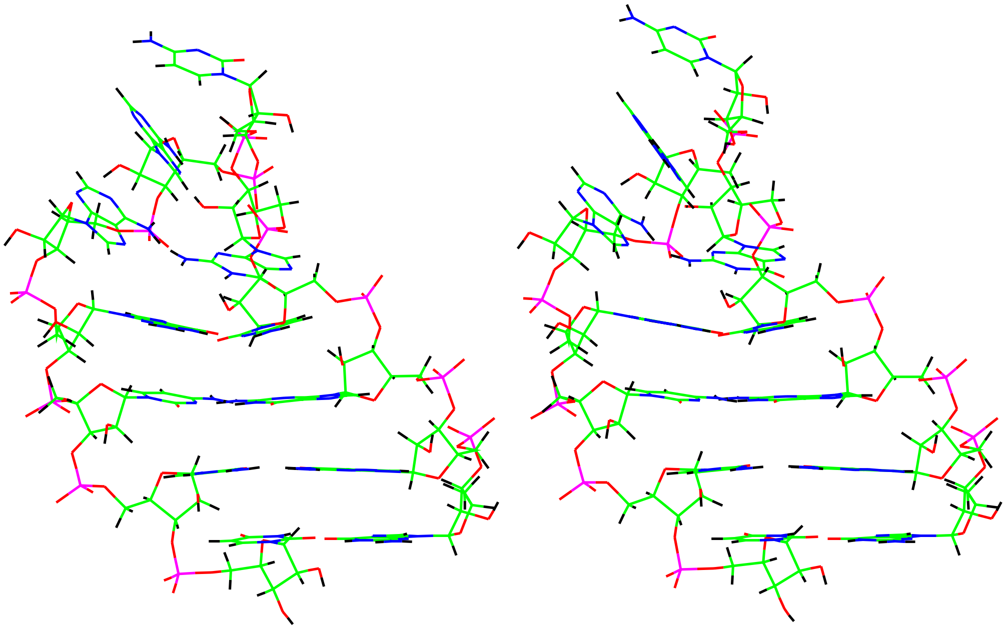


Fig. 4.14: 3D-structures for the GCAA-loop. On the left hand side the complete loop can be seen, the right hand side depicts the 'open loop structure'.

The release of the closing condition allows a relaxation in the loop structures so that the two 'halves' of the former loop are able to move away from each other. This means a decrease of steric constraints so that the 'open loops' are all more stable (i.e. have lower energies) than the closed ones.

The energy differences between 'open' and closed structures can be seen in table 4.3, they can be regarded as 'stabilization energies'. The greater the stabilization energy is, the greater is the tendency of the loop structure to release its steric strains.

Sequence	$\Delta E(A)$	$\Delta E(B)$	$\Delta E(C)$	$\Delta E(D)$	$\Delta E(E)$
GCAA	4.78	5.03	3.68	5.47	1.09
GAAA	4.66	6.00	3.66	4.61	1.12
GAGA	5.14	2.19	3.85	5.50	3.01
GCGA	3.24	2.39	2.15	4.94	2.96
GGAA	4.05	6.06	3.65	4.01	1.16
GGGA	3.10	4.29	3.95	5.29	3.12
GUAA	4.14	5.46	3.68	4.06	1.11
GUGA	2.54	1.94	2.25	5.24	3.01
GACA	12.45	8.28	2.03	13.74	8.79
GAUA	8.47	5.99	4.32	8.55	18.29
GCCA	9.71	7.44	2.04	4.61	17.65
GCUA	9.02	5.56	4.33	5.23	18.05
GGCA	7.83	6.54	6.22	14.68	7.09
GGUA	8.40	5.32	4.32	8.86	17.92
GUCA	5.38	5.12	2.04	11.83	6.54
GUUA	9.59	4.96	4.38	15.95	17.10

Table 4.3: Energy differences between open and closed loop structures in kcal/mol.

The magnitude of the stabilization energies shows a distinct correlation with the loop sequence (see figure 4.15). In figure 4.15 stabilization energy is plotted against the 16 possible GNNA sequences, the vertical dashed line separating the GNRA- from the GNYA-region. For GNRA-loops the stabilization energies have a mean value of 3.7 kcal/mol whereas the mean stabilization energy GNYA-loops is 8.6 kcal/mol. Obviously the steric strain in loops of type GNYA is much greater than it is in GNRA-loops. Of course this does not imply that the GNYA sequences are adopting a conformation in which the loop is broken and two independent strands are formed, but it is a sign that GNYA-loops are more susceptible to conformational changes that destroy the loop structure and release the steric constraints.

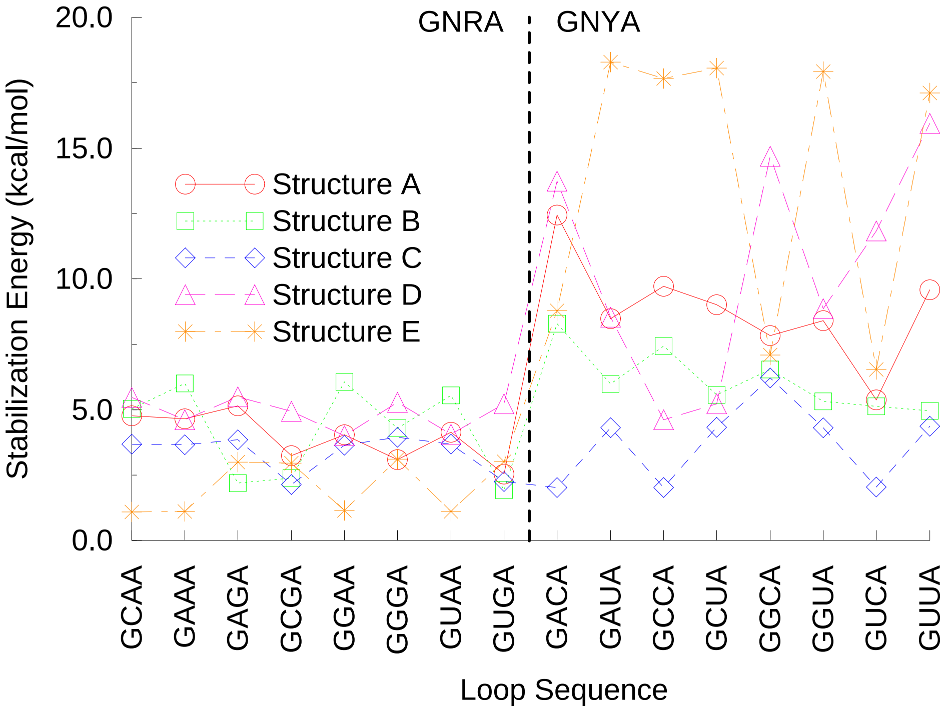


Fig. 4.15: Stabilization energies plotted against the GNNA sequences.

4.4 Conformational search

To complete the investigation on GNNA-tetraloops it was decided to make a more complete conformational search for few selected sequences. The sequences selected were two of GNRA type (GCAA and GAAA) and one of GNYA type (GCUA), a selection was necessary because of the tremendous amount of computation time needed for the conformational search. The program MC-SYM was used to generate starting structures which were then optimized using the JUMNA program. MC-SYM tested 810,000 different conformations for each of the above mentioned sequences and yielded 1459 GCAA-, 1435 GAAA-, and 1457 GCUA-loops, respectively. All of these structures were optimized using the JUMNA program. These optimized geometries were then put into AMBER and optimized again to be able to compare results between both force fields and the different parameter sets. For technical reasons it was necessary to change the sequence from GGGC(GNNA)GCCU to GGGC(GNNA)GCCC, i.e. the last G-U pair was replaced by an additional G-C pair, thus making the stem more stable. In consequence the energies in the following tables are significantly lower than those achieved with the original sequence however it does not influence the loop geometries in any way. The resulting minimum geometries from both force fields were compared and their root mean square deviation (RMS) was calculated. The results for the three different loop sequences are shown in tables 4.4, 4.5, and 4.6. Each table shows the results for the twenty best (i.e. lowest in energy) geometries for each sequence. Each table is divided by a double line, on the left hand side the energies and RMS-values are ordered by JUMNA-energies, on the right-hand side they are ordered by AMBER-energies. The first column in each part of a table gives a number which is unique to this geometry and refers to the starting geometry created by MC-SYM. The following two columns show JUMNA- and AMBER-energies, respectively and the fourth and fifth columns give the corresponding RMS-deviations between the optimized JUMNA and AMBER structures (J-A) and between the geometry resulting from the JUMNA-optimization and the best result from the previous calculations (J-J$_o$, for JUMNA$_{old}$), presented in sections 4.3.1 and 4.3.2. Only geometries marked with an asterisk (*) correspond to geometries which are similar in their three-dimensional structures to those calculated above, i.e. they form a G-A pair and the two remaining nucleotides in the loop are in similar positions to the experimental structure. This structure is referred to in the following as the 'experimental structure' since its main features correspond to the results published by Heus and Pardi.

Some results are common to all sequences investigated:

- RMS-deviation between geometries optimized with JUMNA and AMBER are not too large, usually they are not greater than 1, the mean value is 0.7.
- Energy differences between the 'best' and the 'worst' structure are much larger in AMBER (mean value is 24.75 kcal/mol) than they are in JUMNA (mean value is 9.72 kcal/mol).
- All structures listed in tables 4.1 - 4.3 are more stable (have lower energies) than those found with the original method of varying the sugar conformation.
- Ordering of the geometries after JUMNA- and AMBER-energies usually gives not the same sequence of structures. AMBER results prefer geometries that are closer to the experimental structure.
- A detailed energy analysis shows that the energy differences between geometries that are close to the experimental structure and others lie mainly in the van-der-Waals term. This holds true for the AMBER as well as the JUMNA results.
- All structures show the correct secondary structure (four base-pairs in the stem and a four-membered loop). This means that additional stabilization is not achieved by opening the closing pair and thus releasing steric strain.

A common feature for all structures which do not correspond to the experimental geometry is the total absence of the G-A pair. The energy contribution from the now missing base pair is more than made up by hydrogen bonds between loop bases and the backbone and by improved stacking of the loop bases (see figure 4.16). Figure 4.16 depicts the GAAA-geometry with the lowest energy (number 583), the perfectly coplanar positions of A6 and A7 can be seen clearly. The effect of additional stabilization from improved stacking is most prominent in the GAAA-loop since here two purines are stacking. This might explain why the GAAA-loop is the only sequence

#	JUMNA	AMBER	RMS(J-A)	RMS(J-J_o)	#	JUMNA	AMBER	RMS(J-A)	RMS(J-J_o)
705	-119.519	-86.604	0.69088	4.62268	*508	-115.591	-110.783	0.95701	1.34526
915	-119.405	-104.097	0.71120	3.39948	*850	-114.571	-107.871	1.07277	1.33721
171	-118.139	-92.070	0.58458	4.29668	*853	-113.327	-107.388	0.89281	1.56141
454	-117.865	-90.323	0.78506	4.36023	*857	-115.009	-105.775	0.88589	1.44586
770	-116.969	-94.467	0.74637	4.22184	*464	-114.952	-105.365	0.74787	1.46620
796	-116.509	-102.979	1.42118	2.45796	613	-114.594	-104.487	1.21419	2.44998
*623	-115.699	-100.570	0.66629	1.42540	915	-119.405	-104.097	0.71120	3.39948
*508	-115.591	-110.783	0.95701	1.34526	*851	-114.715	-103.598	0.78349	1.37800
*857	-115.009	-105.775	0.88589	1.44586	796	-116.509	-102.979	1.42118	2.45796
*464	-114.952	-105.365	0.74787	1.46620	854	-114.598	-102.551	0.98184	3.45810
*851	-114.715	-103.598	0.78349	1.37800	*623	-115.699	-100.570	0.66629	1.42540
854	-114.598	-102.551	0.98184	3.45810	188	-113.957	-100.218	0.79872	5.63660
613	-114.594	-104.487	1.21419	2.44998	630	-113.648	-9.039	1.58506	3.82370
*850	-114.571	-107.871	1.07277	1.33721	658	-113.416	-95.050	1.02809	2.14872
188	-113.957	-100.218	0.79872	5.63660	770	-116.969	-94.467	0.74637	4.22184
630	-113.648	-9.039	1.58506	3.82370	164	-113.323	-92.731	0.50885	4.30976
658	-113.416	-95.050	1.02809	2.14872	171	-118.139	-92.070	0.58458	4.29668
*853	-113.327	-107.388	0.89281	1.56141	454	-117.865	-90.323	0.78506	4.36023
164	-113.323	-92.731	0.50885	4.30976	705	-119.519	-86.604	0.69088	4.62268
317	-113.069	-83.023	0.68022	4.63675	317	-113.069	-83.023	0.68022	4.63675

Table 4.4: Minimum energies and RMS deviations for GCAA-loop structures in kcal/mol. Energy of the 'best' GCAA-loop so far: -113.664 kcal/mol.

#	JUMNA	AMBER	RMS(J-A)	RMS(J-J_o)	#	JUMNA	AMBER	RMS(J-A)	RMS(J-J_o)
583	-118.445	-112.700	1.38253	3.34459	583	-118.445	-112.700	1.38253	3.34459
484	-117.735	-9.133	0.79927	4.71454	575	-114.919	-110.782	1.26919	2.54277
416	-115.391	-106.146	0.99578	2.57950	*419	-112.313	-107.720	0.77968	1.53230
575	-114.919	-110.782	1.26919	2.54277	*418	-114.155	-107.471	0.81198	1.57223
403	-114.617	-9.039	0.91817	3.55292	*405	-110.730	-107.450	0.79466	1.21725
*418	-114.155	-107.471	0.81198	1.57223	802	-113.485	-106.783	1.37301	3.49001
*404	-113.827	-105.769	0.84538	1.20707	416	-115.391	-106.146	0.99578	2.57950
802	-113.485	-106.783	1.37301	3.49001	*404	-113.827	-105.769	0.84538	1.20707
588	-112.773	-97.269	0.78247	2.91092	*799	-111.689	-105.454	0.72779	1.55491
414	-112.756	-92.288	0.88760	3.83385	408	-112.267	-104.440	1.29803	2.55199
*419	-112.313	-107.720	0.77968	1.53230	576	-111.454	-104.253	0.90643	3.47474
408	-112.267	-104.440	1.29803	2.55199	*587	-109.571	-102.312	0.71662	1.85716
*799	-111.689	-105.454	0.72779	1.55491	420	-110.855	-101.849	1.49245	3.75030
576	-111.454	-104.253	0.90643	3.47474	484	-117.735	-9.133	0.79927	4.71454
420	-110.855	-101.849	1.49245	3.75030	403	-114.617	-9.039	0.91817	3.55292
*405	-110.730	-107.450	0.79466	1.21725	467	-109.012	-97.771	1.13814	3.32033
*587	-109.571	-102.312	0.71662	1.85716	588	-112.773	-97.269	0.78247	2.91092
620	-109.304	-95.487	1.12554	2.13571	620	-109.304	-95.487	1.12554	2.13571
668	-109.071	-95.329	0.92251	3.33522	668	-109.071	-95.329	0.92251	3.33522
467	-109.012	-97.771	1.13814	3.32033	414	-112.756	-92.288	0.88760	3.83385

Table 4.5: Minimum energies and RMS deviations for GAAA-loop structures in kcal/mol. Energy of the 'best' GAAA-loop so far: -109.858 kcal/mol.

#	JUMNA	AMBER	RMS(J-A)	RMS(J-J_o)	#	JUMNA	AMBER	RMS(J-A)	RMS(J-J_o)
127	-127.462	-89.943	0.83316	5.00593	*460	-118.354	-105.565	1.04933	1.13495
*445	-121.108	-103.914	0.94459	1.09644	*619	-117.848	-105.382	0.84926	1.15547
177	-120.771	-87.209	0.58766	5.34104	*445	-121.108	-103.914	0.94459	1.09644
797	-119.681	-101.832	0.99599	2.33622	*612	-118.256	-103.770	0.93906	1.58806
*460	-118.354	-105.565	1.04933	1.13495	610	-114.627	-102.347	0.82432	3.48348
*612	-118.256	-103.770	0.93906	1.58806	*833	-114.689	-102.117	0.86638	1.76703
*619	-117.848	-105.382	0.84926	1.15547	797	-119.681	-101.832	0.99599	2.33622
467	-116.861	-84.661	0.79124	4.47728	*502	-116.529	-101.642	0.73969	1.19054
609	-116.563	-95.450	0.99659	2.98094	776	-115.255	-97.962	1.07614	2.59180
*502	-116.529	-101.642	0.73969	1.19054	*613	-116.435	-97.229	0.74910	1.43887
*613	-116.435	-97.229	0.74910	1.43887	609	-116.563	-95.450	0.99659	2.98094
184	-115.285	-79.378	0.62957	5.42980	495	-114.709	-93.403	1.16237	2.72719
776	-115.255	-97.962	1.07614	2.59180	906	-114.861	-92.752	0.83086	4.09793
469	-115.232	-92.503	0.69451	2.93497	469	-115.232	-92.503	0.69451	2.93497
906	-114.861	-92.752	0.83086	4.09793	175	-114.764	-91.897	0.79929	5.43622
175	-114.764	-91.897	0.79929	5.43622	127	-127.462	-89.943	0.83316	5.00593
495	-114.709	-93.403	1.16237	2.72719	177	-120.771	-87.209	0.58766	5.34104
*833	-114.689	-102.117	0.86638	1.76703	686	-114.180	-85.339	0.98439	3.97498
610	-114.627	-102.347	0.82432	3.48348	467	-116.861	-84.661	0.79124	4.47728
686	-114.180	-85.339	0.98439	3.97498	184	-115.285	-79.378	0.62957	5.42980

Table 4.6: Minimum energies and RMS deviations for GCUA-loop structures in kcal/mol. Energy of the best GCUA-loop so far: -110.030 kcal/mol.

where the structure with minimum energy in AMBER does not correspond to the experimental geometry. Obviously there are structures that do not correspond to the experimental geometry but still have a lower energy so that the reason for a preference of the 'experimental' GNRA-loop geometries might lay in the better accessibility of bases N6 and R7 that are necessary to form tertiary contacts within a larger molecule, which would result in an additional stabilization.

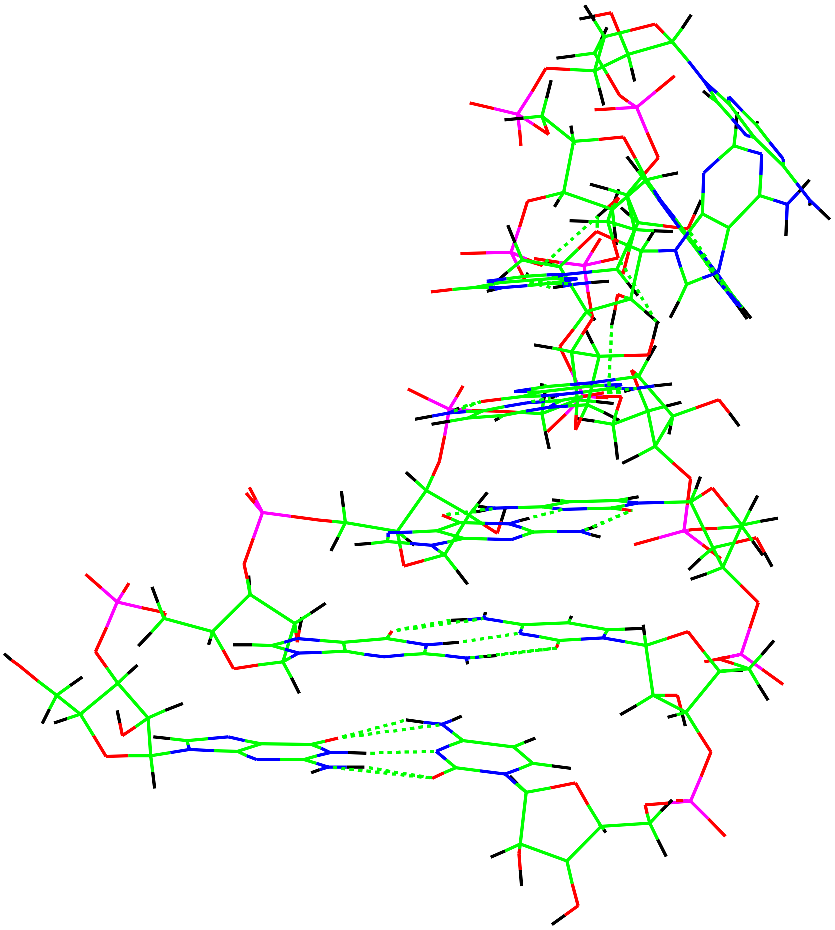


Fig. 4.16: Wireframe model of the minimum GAAA-loop conformation.

Figure 4.17 shows two CPK-models of GAAA-loops side by side: on the right hand side is the experimental structure with the ‘upright’ loop, whereas the left-hand side depicts the GAAA minimum structure shown in 4.16. The energy difference between both geometries is 6.13 kcal/mol in JUMNA and 4.98 kcal/mol in AMBER. Obviously the structure on the left hand side is more compact, thus making bases N6 and N7 less accessible than in the geometry on the right hand side. Unfortunately the influence of possible intermolecular interactions cannot be verified by molecular mechanics calculation because of the size of the molecules involved, however it would help to explain the data presented above.

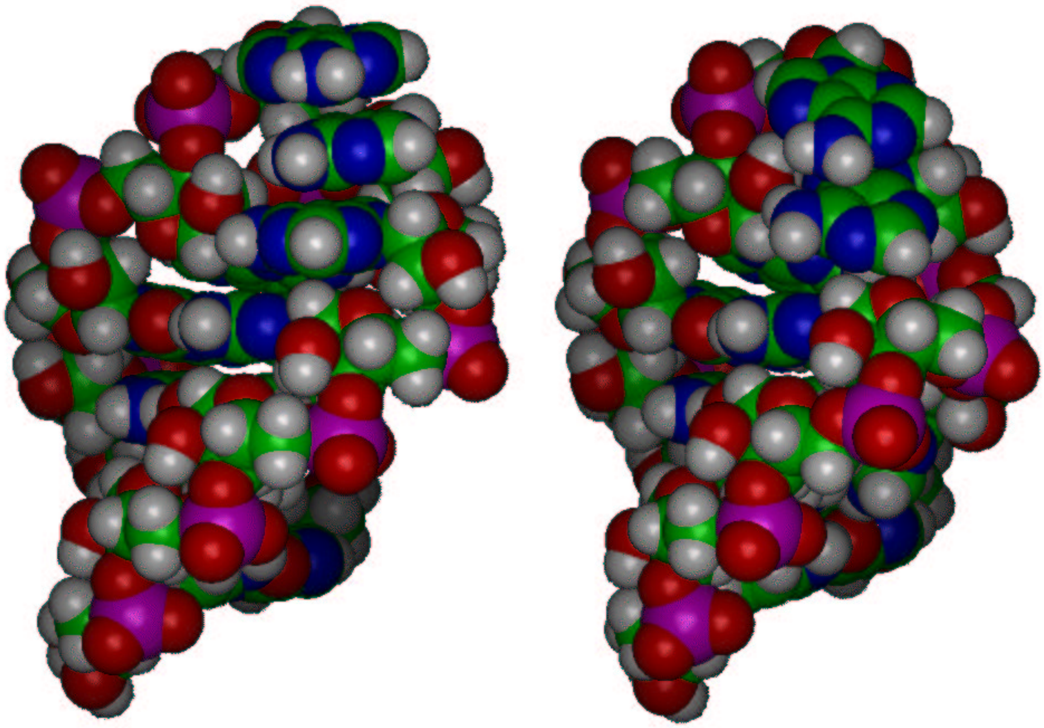


Fig. 4.17: CPK-models of two GAAA-loop geometries.

5 Results: NUN-triloops

5.1 Introduction

It was already mentioned in chapter 2 that RNA secondary structure can be classified by the use of a small number of motifs. 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 to 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 interesting 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 [100][123] where they replace their more abundant four-membered relatives [100]. The energy difference between the three- and the four-membered hairpins is of the order of 1 kcal/mol in favor of the tetraloops [101]. Their small size and the high rigidity makes them an ideal starting point for experimental as well as computational approaches.

At the beginning of this work it was planned to treat the problem of triloop structure from two sides: on one hand triloops should be synthesized from ^{13}C- and ^{15}N-enriched nucleoside triphosphates to investigate their structure by heteronuclear NMR, and on the other hand a computational approach should start only from the secondary structure to be able to compare the results afterwards. Since synthesizing full isotopic labeled molecules is not only expensive in cost but also in experimental effort it was necessary to choose molecules that are as small as possible but that still make ‘biochemical sense’. Almost the same arguments apply for the computational approach. A small size makes for a lower computational effort and the high rigidity makes for fewer accessible conformations. Both facts are important since it was tried to calculate the correct three-dimensional structure without the use of experimental data starting only from the secondary structure. So far very few studies have been done on triloops, among them the rCGC(UUU)GCG- [124] and the rGCGAUU(UCU)GACCGCC-hairpin [51]. 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.

5.2 Minimum structures

The first step to a systematic investigation of triloop structure was the choice of sequences. Phylogenetic studies on eubacterial 16S-RNA [125] showed that triloops with sequence UNU occur most often, followed by the ANU and UNA loops. In all of the above mentioned systems N can stand for any base but G. Therefore five different loop sequences (UUU, AUU, GUU, UUA, UUG) were chosen. To examine a possible influence of the nature of the closing pair on the triloop structure the five loop sequences were combined with two closing pairs (G-C and C-G). The stem sequence was chosen for maximum stability (GGCG or GGCC); a combination of the possible loop and stem sequences results in a total of 10 different molecules. The nomenclature used for these sequences can be seen in figure 5.1. Since synthesizing various sequences is far easier done on a computer than in a laboratory only two of the ten sequences were selected for experimental investigation, namely sequences A and E.

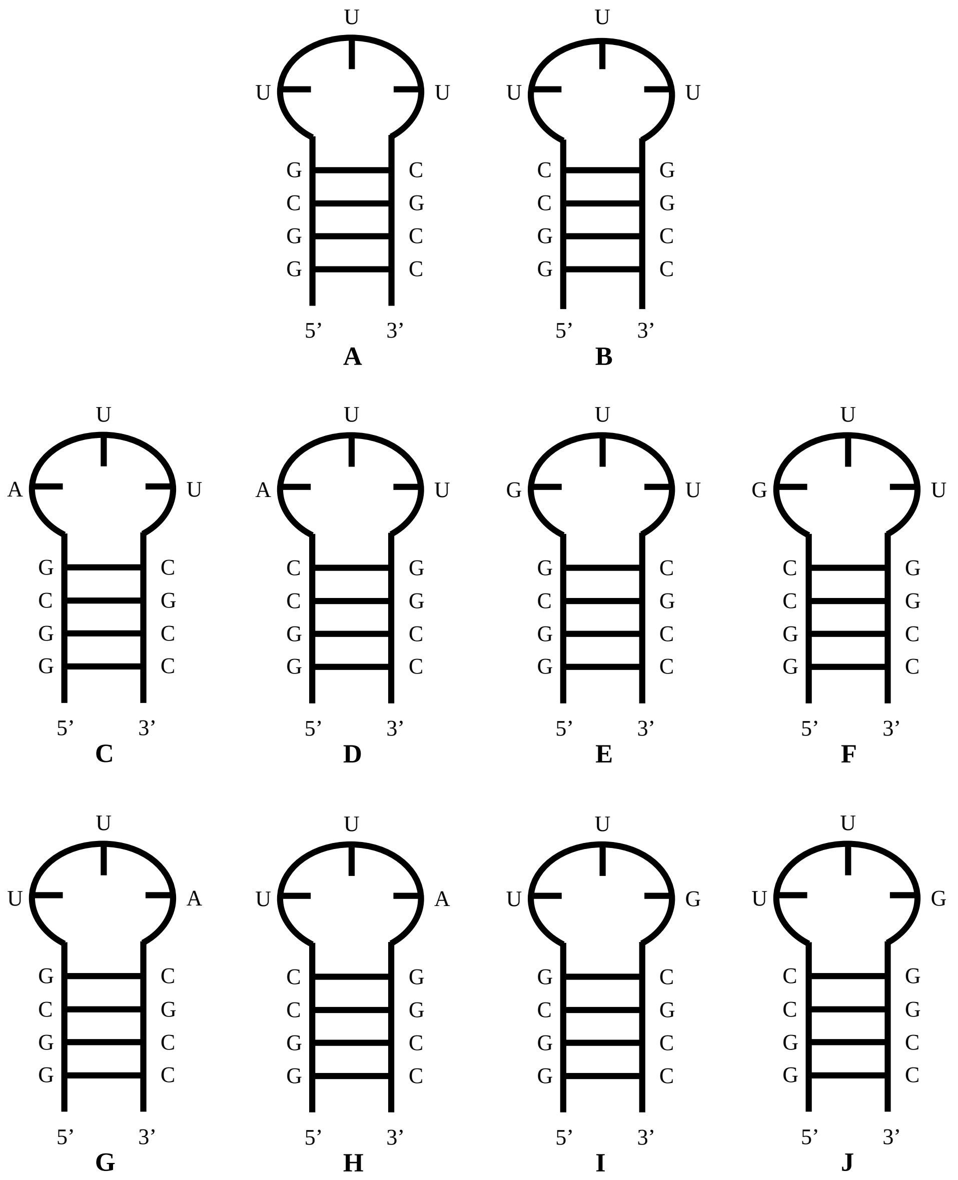


Fig. 5.1: Nomenclature for the NUN-triloops.

As a starting point all molecules were modeled using the INSIGHT II program of Biosym Inc. During modeling it was assumed that the stems adopt RNA-A-helix conformation, whereas several different geometries were used for the loop region. Optimization was done by using the Bremermann algorithm described in section 3.1.3. The definition of the rotational axes for the Bremermann method can be seen from figure 5.2.

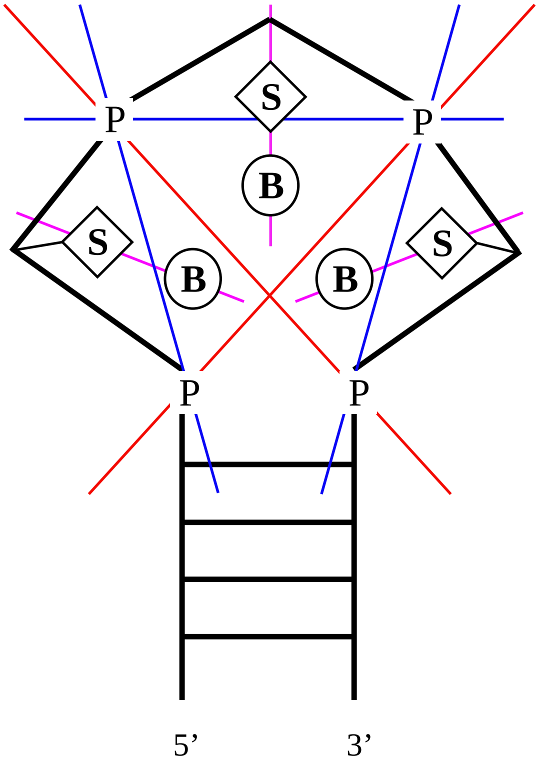


Fig. 5.2: Schematic representation of Bremermann axes in the triloops.

In figure 5.2 ‘P’ stands for phosphorus atoms, a ‘B’ in a circle stands for base, and a ‘S’ stands for the sugar. In total 8 rotational axes were defined: three between adjacent phosphorus atoms (blue lines), two between phosphorus atoms that have two bases between them (red lines), and three at the connection between sugar and heterocycle which corresponds to the glycosidic bond or a variation of the χ-angle (violet lines). The definition of these axes allows for considerable conformational freedom in the loop region, since the bases can change between syn- and anti-conformation and whole nucleotides can be moved.

5.3 Results and discussion

5.3.1 Energies

The use of the Bremermann algorithm led to a distinct improvement (i.e. a lowering of the energy) of the molecular geometries. Table 5.1 shows the energies for each sequence before (in column 'Model.') and after (in column 'Bremer.') several runs of the Bremermann-optimization. The Bremermann method is also terminated via an energy criterion using the energy difference between the last two geometries that were accepted. Since this optimization involves random elements it makes sense to make several sequential Bremermann runs, each starting with the optimized geometry obtained from the previous run. In typical cases between 3 and 5 sequential optimizations were made. The energies shown in column 'Modeling' correspond to structures which were obtained by 'constructing' molecules with the given stem-loop structure which were relaxed with a simple optimization in the AMBER force field.

Apart from simply connecting both sides of the stem it was tried to make an educated guess for the loop structure. As is known from other loop geometries bases in the loop try to maximize two stabilizing effects, namely base pairing and stacking. Since base pairing seems out of question in a three-membered loop it was tried to optimize stacking in the loop as far as possible. Nevertheless structures were changed quite drastically and energies

Seq.	Model.	Bremer.	ΔE	$\Delta E(\%)$
A	-65.08	-71.68	-6.60	-10.14
B	-53.60	-69.44	-15.84	-29.55
C	-60.49	-80.44	-19.95	-32.98
D	-61.50	-73.23	-11.73	-19.07
E	-58.47	-85.04	-26.57	-45.44
F	-60.99	-77.82	-16.83	-27.59
G	-62.75	-76.53	-13.78	-21.96
H	-54.15	-75.37	-21.22	-39.19
I	-68.42	-77.29	-8.87	-12.96
J	-70.97	-78.73	-7.76	-10.93

Table 5.1: Minimum energies of triloops optimized using the Bremermann method

improved significantly by use of the Bremermann optimization (see table 5.1). The two right most columns give the energy improvement in absolute (kcal/mol) and relative (percentage) units.

Corresponding to the drastic improvements in energy strong geometrical changes could be observed in all molecules. A typical example is shown in figure 5.3. On the left side of figure 5.3 is the 'starting structure' for triloop A (loop sequence G(UUU)C), the right-hand side shows the geometry after a Bremermann optimization. The energy difference between both structures is in the range of 7 kcal/mol. The most obvious differences lie in the optimized stacking (parallel orientation of the bases is improved) and the release of several close contacts. The second base in the loop is already moved to a position parallel to the axis of stem, where it can be stabilized by additional hydrogen bonding. Later investigations show that this is a typical feature of triloop structures.

The next step in the search for optimized loop geometries was a combination of the AMBER and JUMNA programs. On advantage of the JUMNA package is the ease with which certain internal coordinates can be subjected to constraints. This is especially true for the conformation of the sugar in a

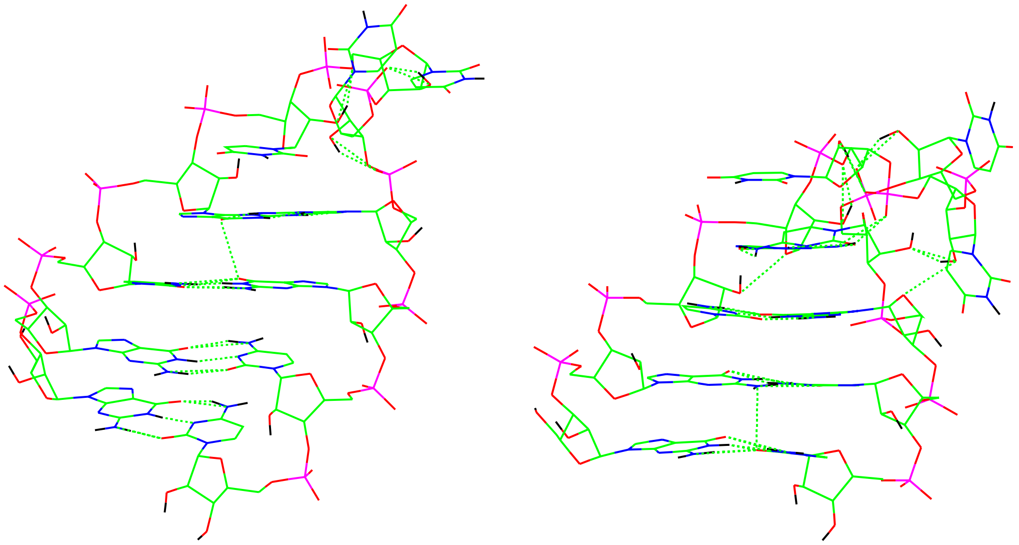


Fig. 5.3: Geometry of the UUU-triloop before and after a typical Bremermann optimization.

nucleotide, where it is possible to simply specify an amplitude and a phase and thus force the sugar pucker in a given conformation. Each of the three bases in the loop was forced to adopt C3′-endo, O1′-endo, and C2′-endo conformations, thus yielding $3^3 = 27$ different geometries for each sequence. The resulting geometries were optimized in AMBER using the Bremermann method and then the conformational search with JUMNA was repeated. This iterative process of optimizations in JUMNA and AMBER was repeated until no energy improvement could be achieved. Table 5.2 shows the results for this optimization procedure. Here the column 'Bremer.' shows the results from the simple Bremermann optimization described above (see table 5.1) and column 'Optim.' shows the final energies for the iterative use of JUMNA and AMBER.

Energy differences are usually largest where the variation of the sugar puckers switched from a N-type sugar to a S-type sugar, thus making the

Seq.	Bremer.	Optim.	ΔE	$\Delta E(\%)$
A	-71.68	-82.55	-10.87	15.16
B	-69.44	-79.96	-10.52	15.15
C	-80.44	-88.96	-8.52	10.59
D	-73.23	-84.37	-11.14	15.21
E	-85.04	-88.64	-3.60	4.23
F	-77.82	-84.22	-6.40	8.22
G	-76.53	-81.78	-5.25	6.86
H	-75.37	-76.96	-1.59	2.11
I	-77.29	-84.65	-7.36	9.52
J	-78.73	-82.19	-3.46	4.40

Table 5.2: Minimum energies of triloops optimized iteratively in JUMNA and AMBER force fields.

backbone longer and decreasing steric strain in the molecule. The next step in optimizing triloop structures was to subject the minimum geometries obtained above to optimization using the program *randstruct* (see section 3.5.1) which resulted again in an energy improvement (see table 5.3).

Seq.	Optim.	Randstr.	ΔE	$\Delta E(\%)$
A	-82.55	-91.44	-8.89	10.77
B	-79.96	-83.78	-3.82	4.78
C	-88.96	-91.19	-2.23	2.51
D	-84.37	-93.50	-9.13	10.82
E	-88.64	-93.25	-4.61	5.20
F	-84.22	-95.82	-11.60	13.77
G	-81.78	-93.02	-11.24	13.74
H	-76.96	-79.83	-2.87	3.73
I	-84.65	-97.84	-13.19	15.58
J	-82.19	-89.93	-7.74	9.41

Table 5.3: Minimum energies of triloops optimized by using *randstruct*.

Energy improvements lie again in the range of 2.5% to 15%. The resulting minimum structures can be seen on the following pages (figures 5.4 - 5.8).

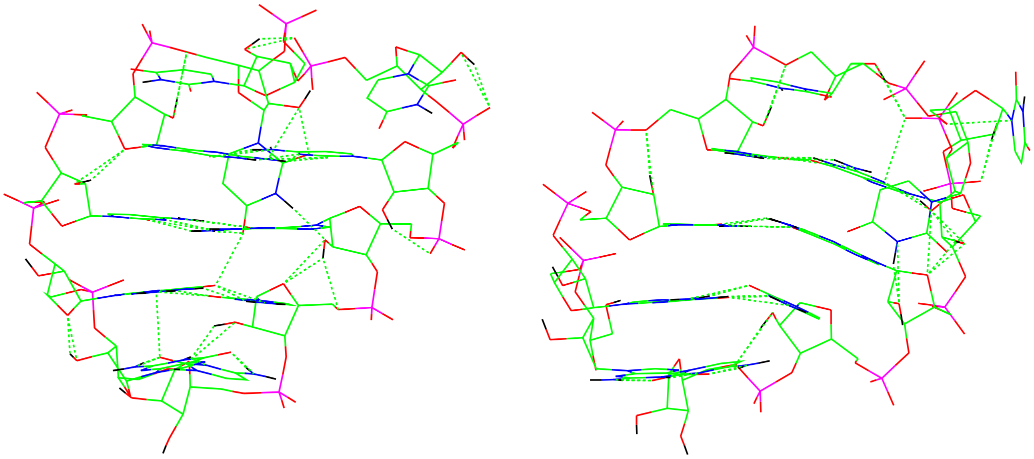


Fig. 5.4: Minimum geometries of sequences A and B.

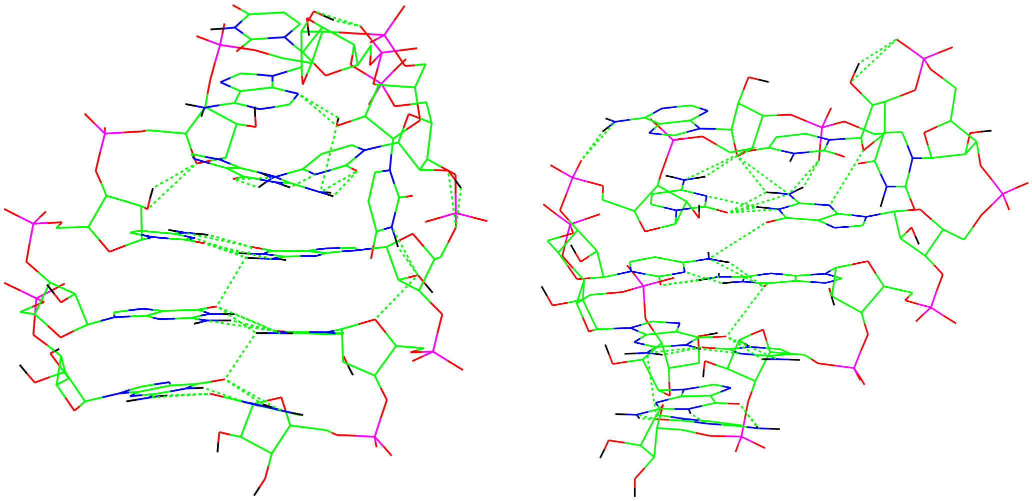


Fig. 5.5: Minimum geometries of sequences C and D.

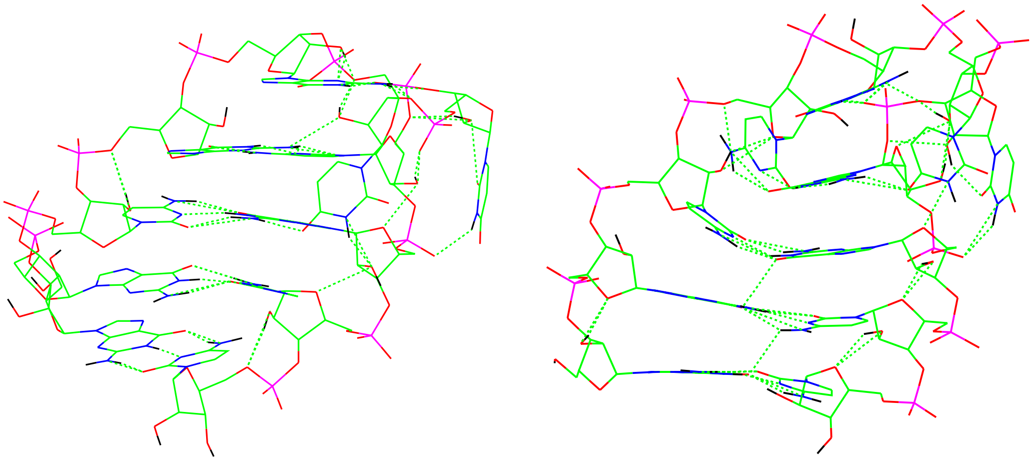


Fig. 5.6: Minimum geometries of sequences E and F.

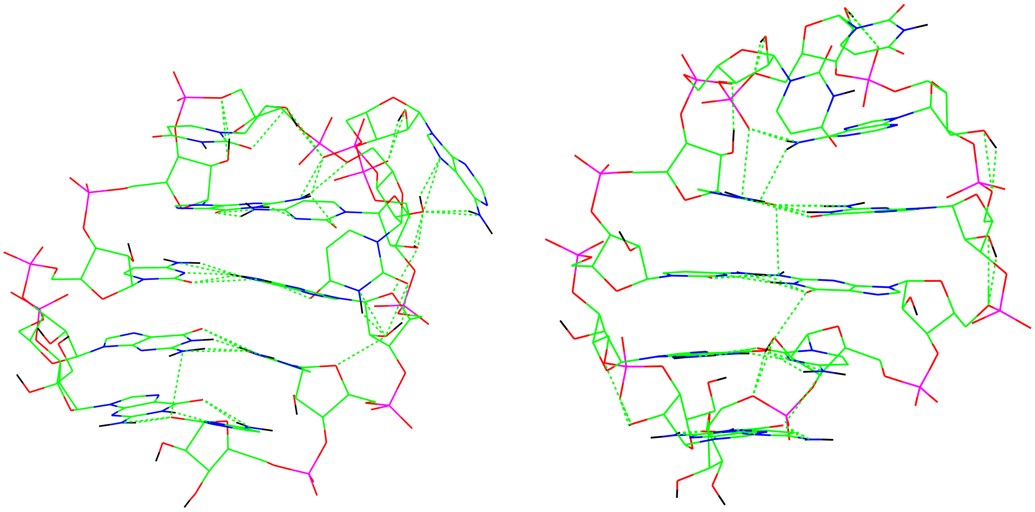


Fig. 5.7: Minimum geometries of sequences G and H.

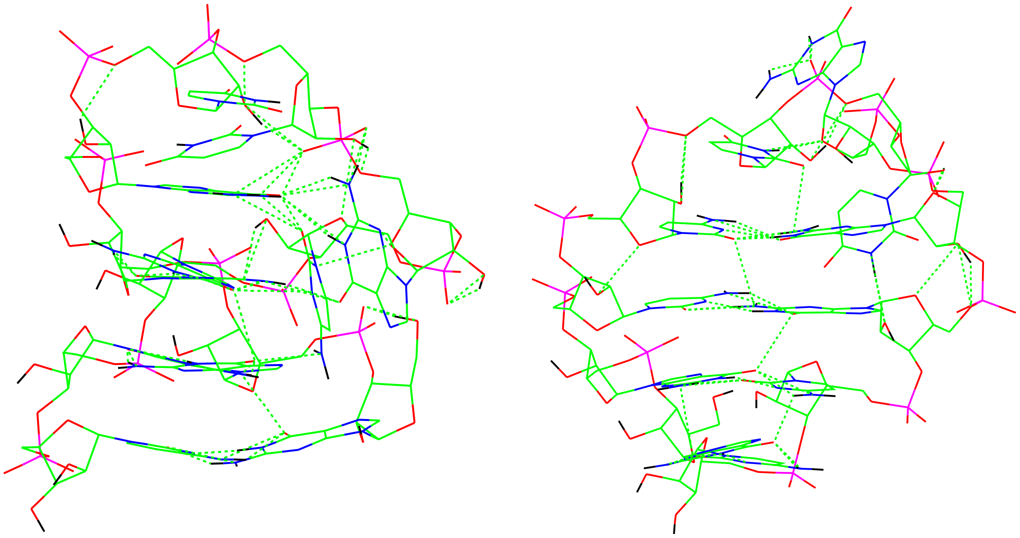


Fig. 5.8: Minimum geometries of sequences I and J.

Description of the different minimum geometries:

- Sequence A: the stem structure is conserved, the geometries of the four base pairs correspond 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 (dotted green lines). The second nucleotide in the loop is stacked on the closing pair on the 5′-side of the stem. The position of the third base is given by the steric constraints of the two other nucleotides which are in energetically favorable positions (stacking and/or hydrogen bonding). The structure described above can also be found in some though not in all of the following sequences. The general principles of structure formation seems to be based on three goals: maximum amount of stacking, maximum number of hydrogen bonds, and very compact structures to minimize the surface between the hydrophobic nucleotides and the solvent.
- Sequence B: a similar structure to that found for sequence A, though the stem geometry is more distorted than in the previous case.
- Sequence C: only three base pairs in the stem retain a perfect RNA-A-helix

conformation, the nucleotides in the closing pair are both tilted in direction of the loop. The distortion of the closing pair makes it possible that the first and the third base in the loop stack on each other, whereas the middle base lies again parallel to the stem axis.

- Sequence D: all base pairs in the stem are distinctly twisted so that a fifth base pair - consisting of the first and second nucleotide in the loop - can be formed, stacking on top of the closing pair. Of course the geometry of this last pair is far from perfect but even so it is an impressive example for the flexibility in the RNA backbone.
- Sequence E: the overall structure of this sequence looks similar to those of molecules A and B, but in this case the first nucleotide stacks on top of the closing pair. The reason for this is obviously the larger stabilization by stacking a purine (G) rather than a pyrimidine base (U), even more so since the purine base does not stack on an individual base but rather ‘lies’ across the closing pair. The second and third base in the loop are again parallel to the stem axis and they are stabilized by additional hydrogen bonds.
- Sequence F: again the geometry of the closing pair is heavily distorted to accommodate stacking between the first base in the loop and the purine base of the closing pair. The second and third bases in the loop can again be found parallel to the axis of the stem.
- Sequence G: the overall structure corresponds to that of sequence E. Interesting in this case is the fact that not the purine in the loop but rather a pyrimidine base is stacking on the closing pair.
- Sequence H: this sequences shows another type of possible structures. In this case bases 2 and 3 in the loop stack on each other and on the closing pair on the $3'$-side. The reason for this behavior is probably that again a purine-purine stacking can be facilitated.
- Sequence I: again the stem structure is reduced to three base pairs, though on the $5'$-side a total of 5 nucleotides are stacked. In addition two nucleotides

- the third base in the loop and the 3′-part of the closing pair are stacking on each other and are positioned parallel to the axis of the stem.

• Sequence J: this structure corresponds to that already found for sequence E with the exception that the third base lies on top of the loop instead of being stabilized to the backbone by hydrogen bonding.

Though the overall structures of the minimum geometries are different, some features can be found that all 10 sequences have in common:

• At least three of the four base-pairs in the stem are conserved, the geometry of the closing pair can be subjected to various degrees of distortion.

• Sugar puckering: sugar puckers in the stem are of N-type. In the loop sugar puckers consist mainly of S-types or at least O1'-endo types to enable both loop closure and maximum stacking in the loop.

• Base conformations : All bases have anti-conformation.

• At least one base is moved to the rear side (small groove) of the molecule were hydrogen bonds to the backbone offer a source for further stabilization.

An alternative route to optimized molecular geometries was the use of the program MC-SYM described in chapter 3.4. Here the loop region was defined to be as flexible as possible (keyword 'sample+', which corresponds to 59 conformations per nucleotide). This resulted in $59^3 = 205,379$ initial structures of which approximately 500 were chosen by MC-SYM as the fulfilled the loop closing criteria. All these structures were optimized using AMBER 4.0 and their geometries were further improved by use of the program *randstruct*. Since optimizing so many structures is very demanding in terms of computational time the investigation was restricted to those sequences which are also of experimental interest, namely structures A and E. In case of sequence A MC-SYM found 497 valid solutions, in case of sequence E 523 starting structures were optimized.

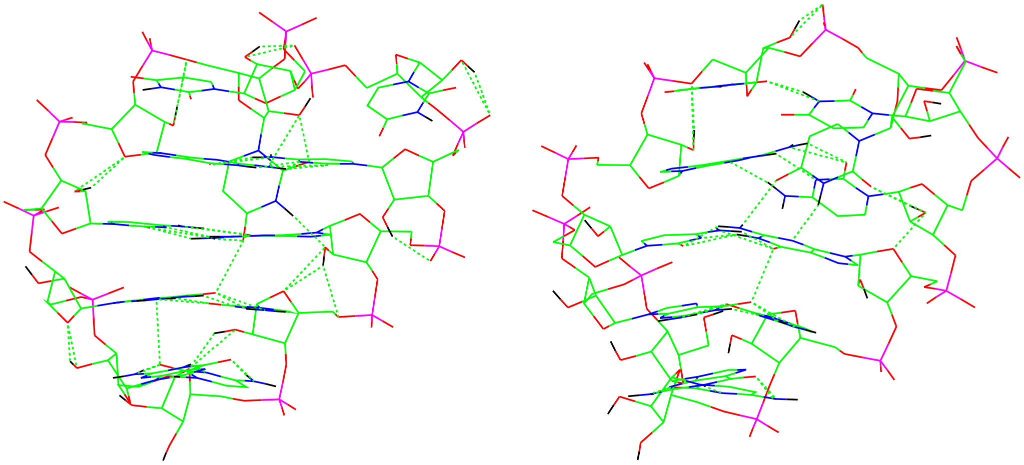


Fig. 5.9: Different minimum geometries of sequence A.

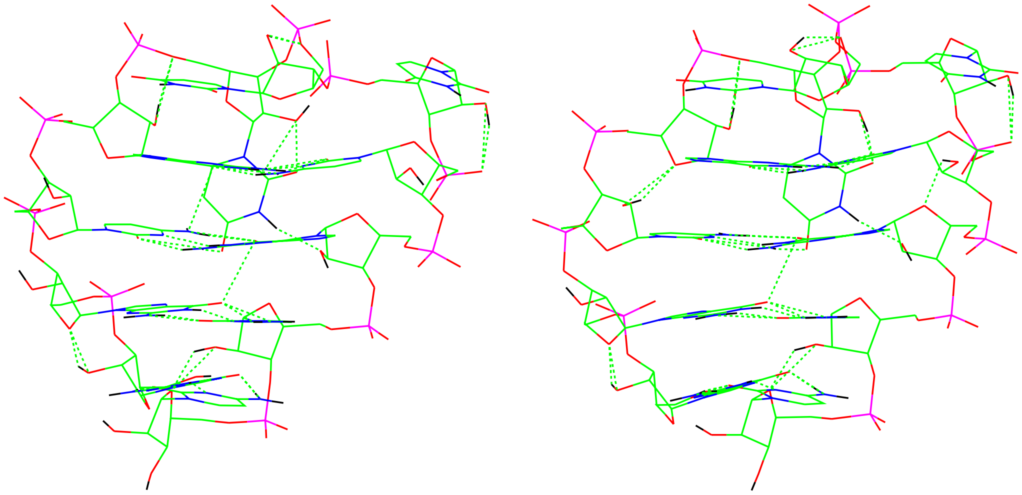


Fig. 5.10: Different minimum geometries of sequence A.

Though this procedure did not yield any 'new' minimum geometries with the assumed secondary structure it showed that there can be several different geometries with radically different structural features but with nearly the same energies. This is true for both sequences, the examples shown (figures 5.9 and 5.10) depict four different geometries for the UUU-triloop which differ energetically by only 3 kcal/mol. Among the new minimum structures were also 'pentaloop', i.e. geometries where the closing G-C pair is opened and a five-membered loop is formed. This result was also indicated by NMR-experiments done in Larry Browns group [126].

5.3.2 Dimer structures

During experimental investigations on sequences A and E it was found that these triloops tend to form dimers in solution rather than to remain monomers. This is at least true for the comparatively high concentrations needed for NMR-experiments. Figure 5.11 shows a schematic drawing of the process of dimer formation.

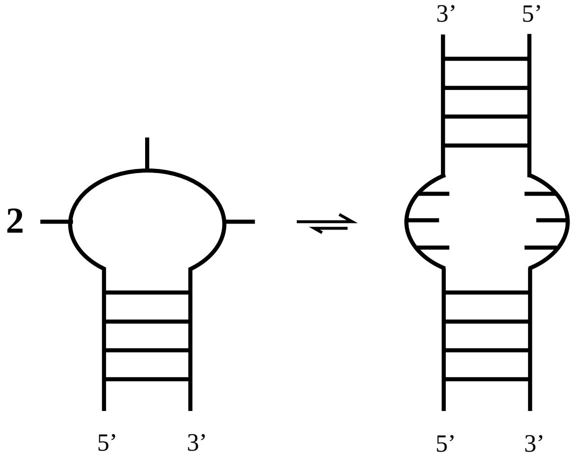


Fig. 5.11: Schematic representation of the dimer forming process.

The simplest assumption on the structure of the dimer is the formation of a standard RNA-A-helix. With the loop sequences under consideration (UUU and GUU) there should be no problem in accommodating the newly formed base pairs (U-U and G-U). It is assumed that the dimer molecule forms a long double-stranded stem consisting of 8 canonical base pairs and three newly formed ones (see figure 5.11). An example for such a structure can be seen in figure 5.12 which shows the dimer of the UUU-triloop. Figure 5.12 shows the optimized structure of this dimer, the colored ribbons indicate the backbone which obviously is very close to pure RNA-A-helix conformation. Figure 5.13 shows the minimum structure for the GUU-dimer.

Again deviation from standard RNA helix is minimal, the most obvious structural change is a slight bend introduced at the position of the former loop region. In this case the structural investigation was less interesting than the energetical comparison between monomer and dimer geometries. The simplest approach for this evaluation is to compare the energy of the monomer times two with the dimer minimum energy. Using this simple approach an energy difference of 20 kcal/mol is found for the UUU-triloop and of 5 kcal/mol for the GUU-loops in favor of the monomer species. According to these results the monomer-dimer equilibrium should be shifted distinctly to the monomer side. Experimental results, however showed that this is not true, at least not for higher monomer concentrations. To improve the results of the molecular mechanics calculation both the monomer and the dimer structures were immersed in a solvent, in this case water. In terms of the simulation this means the addition of a number of water molecules around the RNA molecule. Therefore two new types of interaction are introduced in the calculation of the minimum energy, namely RNA-water and water-water interactions. These new contributions to the total energy make it impossible to compare monomer and dimer energies directly since it cannot be avoided to count some of the water-water interactions twice. This dilemma can be solved with the following method: starting from the molecules in quasi vacuum different numbers of water molecules are added to both the monomer and the dimer (see figure 5.14).

On the left-hand-side of figure 5.14 the molecule is in a quasi-vacuum state. By adding more and more water molecules the energy begins to drop.

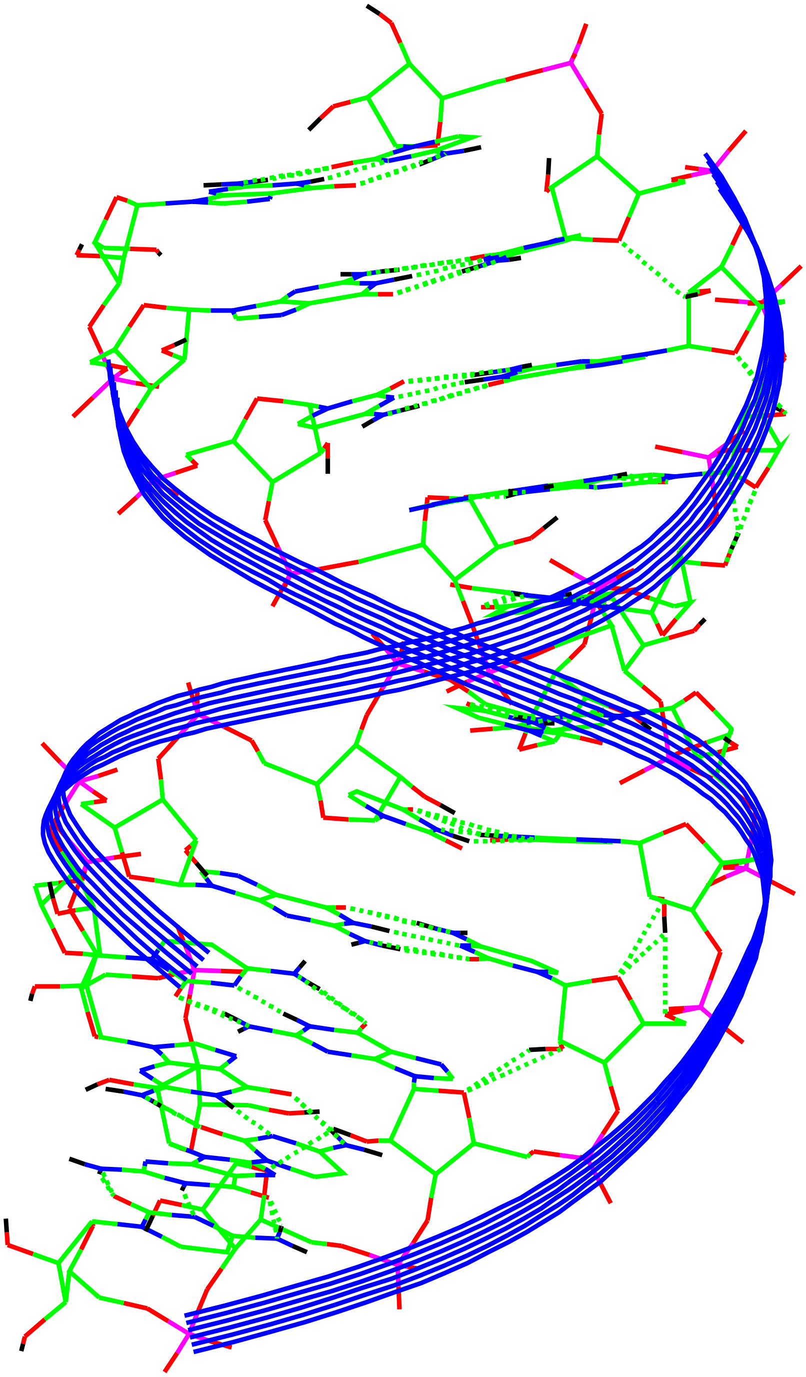


Fig. 5.12: Minimum structure of the UUU-dimer. Blue ribbons show the course of the backbone.

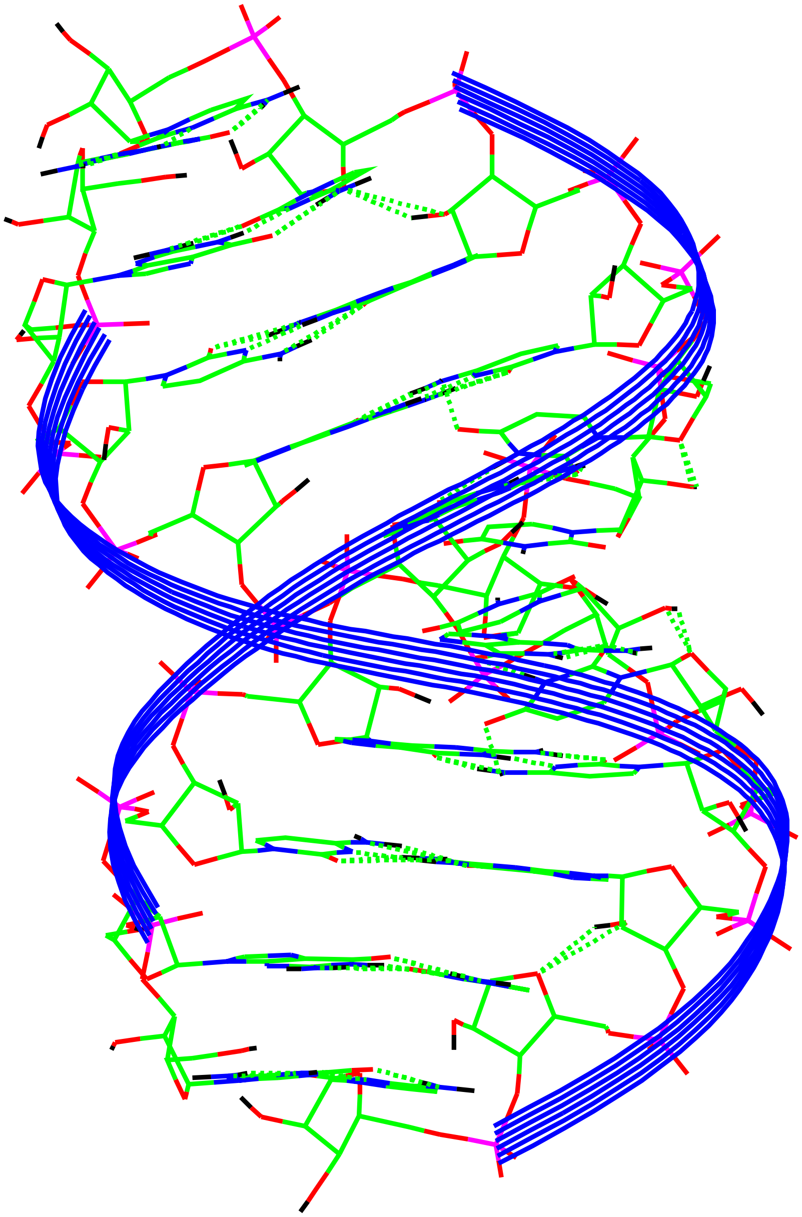


Fig. 5.13: Minimum structure of the GUU-dimer. Blue ribbons show the course of the backbone.

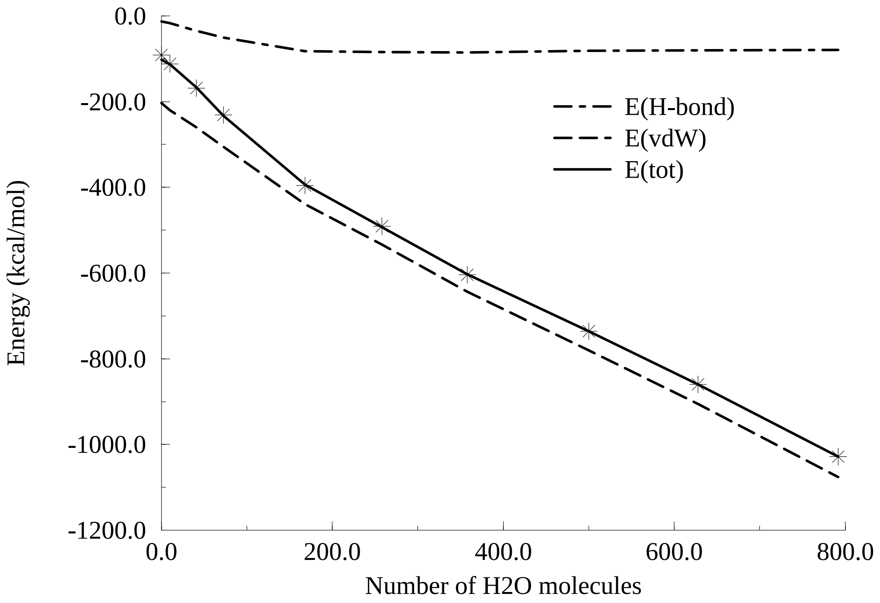


Fig. 5.14: Minimum energies of the UUU-triloop with different numbers of water molecules.

This is understandable from the fact that the first sphere of solvent molecules is constructed and hydrogen bonds are formed. The contributions that are responsible for this decrease in energy are shown in the figure: these are the van-der-Waals interaction and the hydrogen bonds. The other energy contributions stay constant and were therefore omitted. By adding more and more water molecules the energy drops further, but at some point the curve makes a rather sharp bend. At this point the first solvent sphere is completed and the additional energy effects come mainly from new water-water interactions rather than from new RNA-water interactions. The energy curve is thus divided into two linear parts; by extrapolating the second part it should be able to gain the energy for a molecule in the ‘right’ environment (i.e. in aqueous solution) but independent from the number of water molecules. Applying this method to both the monomer and the dimer structures should yield energies that are directly comparable again and that include solvent effects. Figures 5.15 and 5.16 show these results for the UUU- and the GUU-molecules. Details and exact values for the minimum energies and the number of water molecules are given in tables 5.4 (UUU-sequences) and 5.5 (GUU-sequences). The left-hand side of each table contains the results for the monomer (index $_m$), the right-hand side shows the dimer results (index $_d$). The last row in each table (in bold face) gives the estimated minimum energies from the extrapolation.

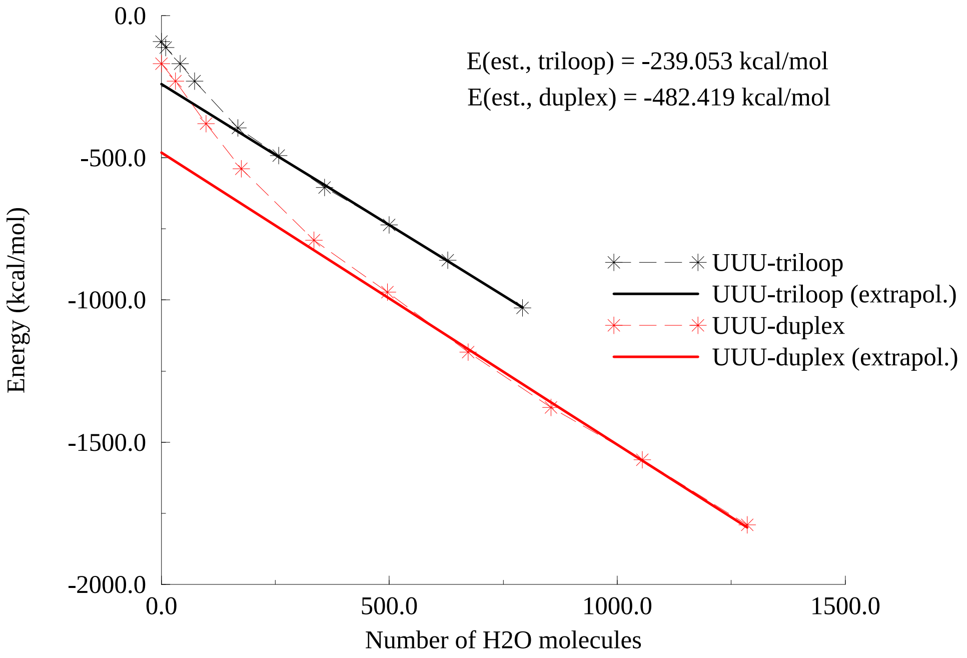


Fig. 5.15: Minimum energies for UUU-triloops and dupleces with different numbers of water molecules.

In each of figures 5.15 and 5.16 there are two dashed curves and two bold lines. The upper curve and line (black) correspond to the monomer, whereas the lower curve and line (red) show the results of the dimer. The curves show the energies of molecules surrounded with different numbers of

(# H_2O)$_m$	E_m	(# H_2O)$_d$	E_d
0	-91.46	0	-168.87
10	-112.22	31	-230.41
41	-169.00	98	-380.00
73	-230.79	176	-538.04
168	-395.65	335	-790.17
258	-491.41	496	-972.91
358	-603.78	673	-1183.45
500	-735.88	855	-1377.64
628	-895.68	1055	-1562.09
792	-1028.23	1285	-1790.47
(0)	**-239.05**	**(0)**	**-482.42**

Table 5.4: Minimum energies and number of water molecules for the UUU-triloop and its duplex. Energies are in kcal/mol.

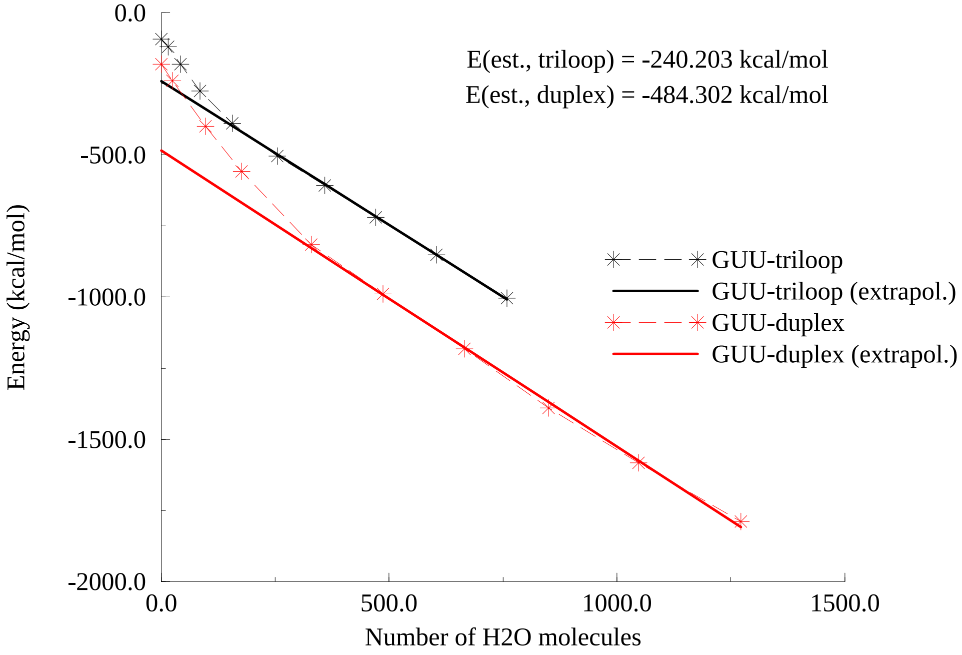


Fig. 5.16: Minimum energies for GUU-triloops and dupleces with different numbers of water molecules.

water molecules and lines are the extrapolations of the lower (and linear) part

$(\# H_2O)_m$	E_m	$(\# H_2O)_d$	E_d
0	-93.46	0	-180.33
15	-119.57	25	-238.68
42	-180.55	97	-399.77
85	-275.90	177	-557.93
156	-389.12	329	-814.83
255	-504.00	487	-988.54
359	-607.07	666	-1181.35
471	-720.22	850	-1390.28
604	-850.70	1048	-1582.21
759	-1004.35	1272	-1789.44
(0)	**-240.20**	**(0)**	**-484.30**

Table 5.5: Minimum energies and number of water molecules for the GUU-triloop and its duplex. Energies are in kcal/mol.

of each curve. By multiplying the estimated energies E_{est} of the monomer and comparing them to the estimated dimer energies it can be seen that in both cases the monomers are approximately 4 kcal/mol higher in energy than the dimers. The exact energy differences are 4.313 kcal/mol for the UUU-sequences and 3.896 kcal/mol for the GUU-sequences. Of course the absolute value has to be treated with some caution, but still these results correspond to those found experimentally. An interesting point is the fact that the stabilization of both sequences is almost the same, as it also would be plausible that the GUU-dimer is even more stable than the UUU-species since two additional G-U pairs are formed. An explanation for the higher stability of the dimers are of course the additional hydrogen bonds in the newly formed base pairs and improved stacking of all bases since the dimers form one long double helix. It is remarkable that the former loop regions do not form bulges, but that the unusual U-U base pair is found in the dimer.

The next logical step in verifying the experimental results would be the calculation of the various minimum geometries not only in the presence of water molecules but also including counter ions to mimic the Coloumb-forces more accurately. Such a configuration can be seen from figure 5.19

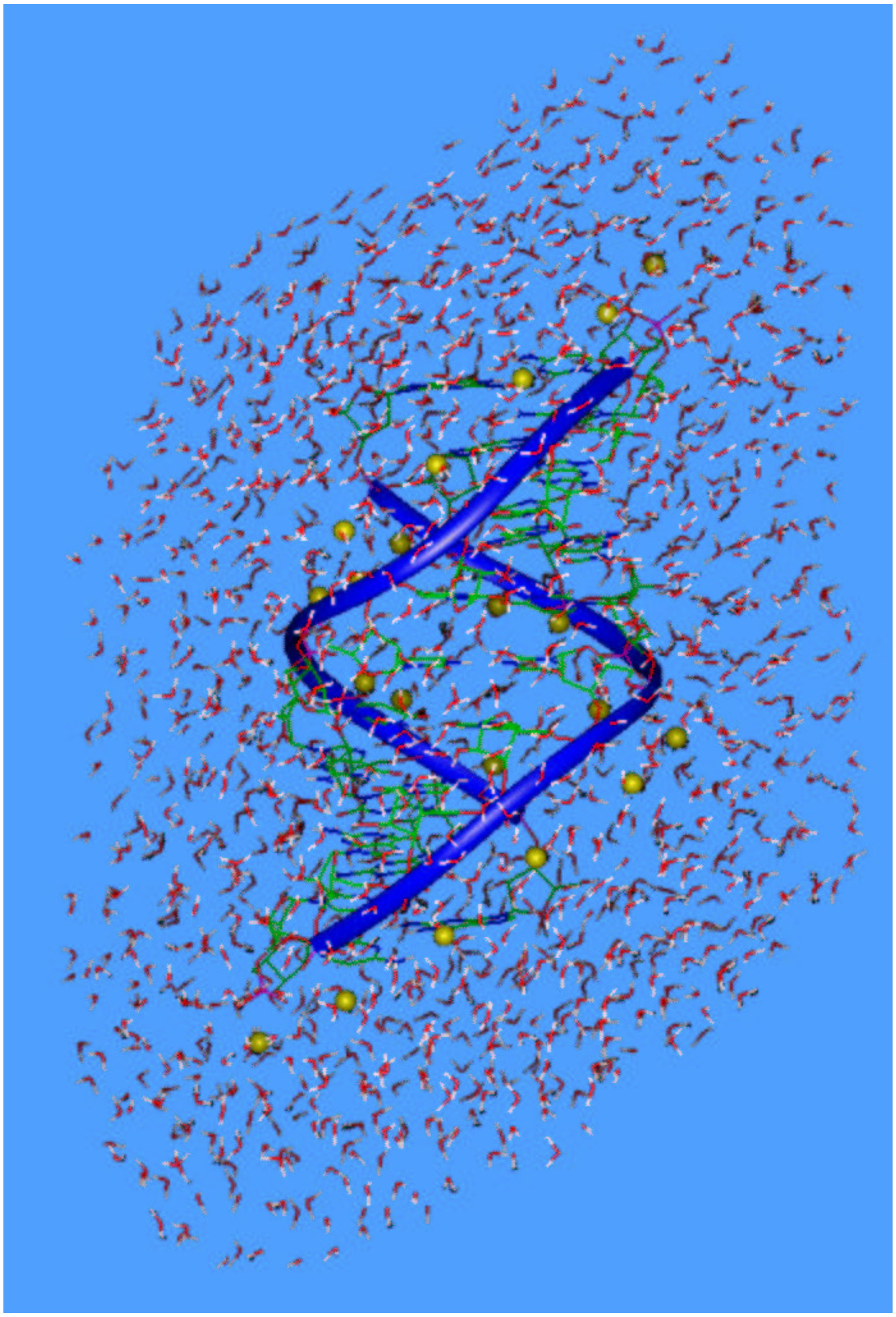


Fig. 5.17: Minimum structure of the UUU-triloop in a solvent 'blob' and surrounded by counter-ions (sodium ions shown as yellow spheres).

which shows the UUU-triloop surrounded by solvent-molecules (water) and counter-ions (in this case sodium, shown as yellow spheres). Again the blue

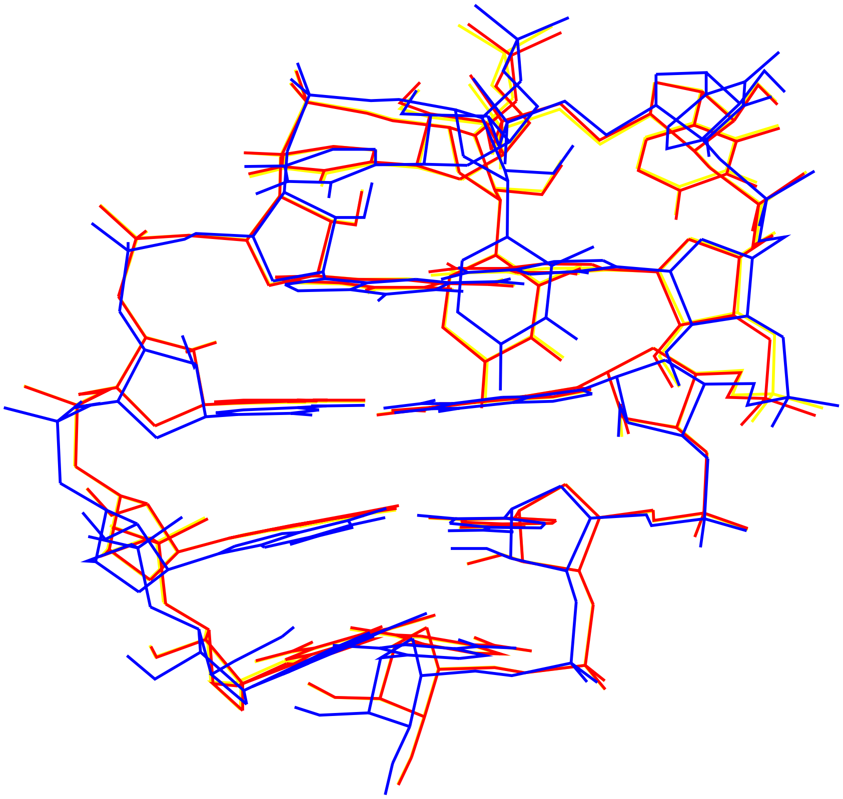


Fig. 5.18: Overlay of three different minimum structures of the UUU-triloop.

ribbons show the backbone of the molecule.

Optimization was done for both the UUU-triloop and dimer structures but it was found that the extrapolation-method described above could not be applied here because of an additional variable, namely the number of counter-ions. However, minimization in the presence of counter-ions yielded another interesting result which can be derived from figure 5.19. Figure 5.19 shows an overlay of three different UUU-triloop geometries. The three molecules in figure 5.18 (blue for optimization in quasi-vacuum, red for optimization in water and yellow for optimization in water and in the presence of counter-ions) are virtually the same, the RMS-deviation between them being not larger than 0.5 Ångstrøms. Consequently the overall structure is not influenced greatly by the presence of solvent or counter-ions, the energetical ordering of different conformation on the other hand is changed by these factors. In other words good geometries that were found in quasi-vacuum will still be good geometries in solution, but the best geometry from quasi-vacuum will not necessarily be the best geometry in solution. Thus it seems possible to make the largest part of the optimization in quasi-vacuum (which makes for a far smaller computational effort) and add solvent only for the purpose of energetical comparisons.

6 Conclusions

6.1 GNNA-tetraloops

The structures of a family of very common and unusually stable RNA hairpins were determined by molecular modeling. GNRA hairpins are known for their extraordinary thermodynamic stability, which is presumably caused by several specific interactions in the loop. Firstly, all bases except N6 stack on other bases, a G-A pair of type X (see W. Saenger [18]) is formed which stacks upon the closing pair of the loop and base N7 stacks on A8. Secondly, because of the G-A pair the tetraloop resembles a 'diloop' and the resulting steric constraints force base N6 in a position on top of the loop. Further stabilization comes from additional hydrogen bonds between the bases in the loop and the backbone. These features, however, are not restricted to GNRA-loops; they appear also in GNYA-loops. Two distinct types of the G-A pair were 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 the variation of bases N6 and N7 the relative stabilization energies calculated between 'open' and 'closed' loop structures show clear differences between GNRA and GNYA loops in the order of 5 kcal/mol. They might be responsible for the preference of GNRA over GNYA hairpin loops in nature.

The predicted 3D-structures of tetraloop hairpins demonstrate a 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 for base stack formation, particularly in double helical geometries. It is continued in 3D-structure formation. Non-Watson-Crick closing pairs of the stems 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. The same principles can also be observed on a larger scale in the 'stacking' of two stems on each other in tRNAs or in the hammerhead ribozyme.

A conformational search shows that there are more stable loop geometries than the one indicated by Heus and Pardi. These structures lack the G-A pair but are even more compact than those corresponding to the experimental structure. GNRA-loops occuring in nature seem to have this G-A pair, because it forces the two middle bases in positions that are both flexible and accessible, which is important for forming intra- (formation of pseudoknots) as well as inter-molecular (protein-recognition) tertiary contacts. A comparison between structures optimized within the AMBER and the JUMNA force field shows similarity of the optimized structures. The energetical order, however, may be completely different. This is probably due to a different relative importance of stacking and hydrogen bonding in the empirical parameters of the two force fields.

6.2 NUN-triloops

The structure of several members of the triloop sequence type NUN were investigated using a molecular modeling approach. The loop sequences UUU, AUU, GUU, UUA, and UUG are the most abundant triloop sequences that occur in nature. It was tried to optimize the geometry of small hairpin molecules containing these sequences without the use of experimental data and to compare the results with a NMR-study that was done in parallel. Typically not only one minimum structure per sequence is found, but several different ‘types’ of structures are energetically roughly equivalent. Though the resulting minimum geometries are different - depending on the sequence of the loop and on the closing pair - some common features in the loop geometries became apparent. Four effects seem to be of importance for the structure of a small RNA loop: maximization of base pairing, maximization of hydrogen bonding, maximization of stacking, and minimization of the interacting surface of the hydrophobic nucleotides and the aqueous solvent.

- Maximization of base pairing: even in small three-membered loops it is possible to have an additional base pair in the loop, though the formation of this pair distorts the structure of the stem.
- Maximization of hydrogen bonding: in most minimum structures at least one nucleotide is in a position parallel to the axis of the stem, where it can be stabilized by additional hydrogen bonds between nucleotide and the backbone atoms.
- Maximization of stacking: in all minimum structure some kind of additional stacking can be found. This happens either by stacking of a loop-base on the closing pair or - in case of a possible purine-purine stacking - the closing pair is opened and the triloop structure becomes a pentaloop.
- Minimization of the surface to the solvent: all minimum structures present very compact shapes in which the hydrophobic parts of the molecules - the nucleotides - are on the inside, whereas the hydrophilic parts - the backbone, mainly the phosphorus and oxygen atoms - are on the outside, towards the solvent.

Of course, not all of these effects can be realized to their full extent in a typical triloop structure, the minimum geometry rather is governed by a combination of various amounts of these features. Which effects are the dominant ones is determined by the loop- and/or the closing pair-sequence. In fact, all of the effects listed above are already known from experimental investigations on RNA structures, but it is interesting that they are reproduced rather well by a standard force field. This indicates that it is possible to make reasonable structure predictions for small RNA molecules using molecular modeling and heuristic methods of conformation space sampling, though usually not one designated minimum geometry is found but rather several families of different structures, which should be easily distinguished by additional experiments.

The experimental investigation on RNA triloops done in parallel to the calculations presented in this thesis showed that for small stem sizes and high concentrations triloop structures tend to form dimers and that the monomer structure is rather a pentaloop than a triloop. Thus it was tried to reproduce these results by comparing molecular modeling data of monomers as well as dimers. Optimization in quasi-vacuum showed the monomer to be distinctly more stable than the dimer, thereby contradicting the experimental results. By adding explicit solvent molecules to the nucleic acid and re-optimization it was possible to roughly reproduce the experimental values, which show the dimer to be more stable by at least 5 kcal/mol. The minimum structure however was not changed by the presence of solvent molecules and/or counter-ions. This seems to indicate that while quasi-vacuum might be sufficient for simple geometry optimizations, it is necessary to include the solvent explicitly in the calculations for energetical comparisons. Both experiment and energy minimizations come to the result that the structure of the dimer resembles a long helix containing the unusual U-U pair.

6.3 Outlook

Some general results found in this work are of general importance for future calculations of three-dimensional structures of small RNA hairpins. Obviously it is not feasible to make exhaustive conformational searches for molecules with flexible regions larger than three nucleotides. This automatically excludes all RNA molecules of biochemical interest including even tetraloops with exceptionally high stability. Another typical result is the fact that usually there is not one minimum structure which is distinctly more stable than all others, but that there are several structurally different families of minimum geometries with nearly the same energy. Though the correct structure is usually among these minima it cannot be easily distinguished from others by all current force fields. This could probably be changed by the introduction more detailed force fields which account for stacking or for the 'compactness' of the geometry by specific terms.

The results shown in this thesis indicate that the modeling of RNA three-dimensional structure should rather be using a knowledge-based algorithm in combination with heuristical elements or an iterative process between theoretical calculations and experiments. The knowledge-basis in this case might be rather general rules like 'Reduce the actual loop size by forming new base pairs in the loop region', 'Reduce the surface of the molecule by forming more compact structures', or 'Continue stacking in the loop whenever possible'. Rules like this could be used to find reasonable starting structures very quickly, which then could be optimized by algorithms using random elements like that presented in section 3.5.1. By a combination of these two methods it should be possible to produce a pool of 'probable' families of minimum structures, which are nearly equivalent in energy, but which have fundamentally different geometries. A distinction between these different geometries could than be made with experiments (e.g. NMR-measurements) which test for specific structural features so that the time-consuming process of determining the complete three-dimensional structure with experimental methods would be avoided.

7 Literature

[1] O. T. Avery, C. M. MacLeod, and M. McCarty. *Journal of Experimental Medicine*, 79:137 – 158, 1944.

[2] J. D. Watson and F. H. C. Crick. *Nature*, 171:737 – 738, 1953.

[3] J. D. Watson and F. H. C. Crick. *Nature*, 171:964 – 967, 1953.

[4] A. J. Zaug and T. R. Cech. *Science*, 231:470 – 475, 1986.

[5] T.R. Cech and B. L. Bass. *Annual Review of Biochemistry*, 55:599, 1986.

[6] J. D. Puglisi, J. R. Wyatt, and I. Tinoco(Jr.). *Accounts of Chemical Research*, 24:153, 1991.

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

[8] D. Gautheret, S. H. Damberger, and R. R. Gutell. *Journal of Molecular Biology*, 248:27 – 43, 1995.

[9] R. Klinck, J. Liquier, E. Taillandier, C. Gouyette, T. Huynh-Dinh, and E. Guittet. *European Journal of Biochemistry*, 233:544 – 553, 1995.

[10] R. Green and J. W. Szostak. *Journal of Molecular Biology*, 235:140 – 155, 1994.

[11] M. Chastain and I. Tinoco (Jr.). *Biochemistry*, 31:12733 – 12741, 1992.

[12] M. Chastain and I. Tinoco (Jr.). *Biochemistry*, 32:14220 – 14228, 1993.

[13] F. Michel, A. D. Ellington, S. Couture, and J. W. Szostak. *Nature*, 347:578 – 580, 1990.

[14] R. Nussinov. *Journal of Theoretical Biology*, 133:73 – 84, 1988.

[15] D. Sen and W. Gilbert. *Biochemistry*, 31:65 – 70, 1992.

[16] C. Cheong and P. B. Moore. *Biochemistry*, 31:8406 – 8414, 1992.

[17] G. Awang and D. Sen. *Biochemistry*, 32:11453 – 11457, 1993.

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

[19] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. *Science*, 254:1598 – 1603, 1991.

[20] C. Guerrier-Takada and S. Altman. *Science*, 223:285 – 286, 1984.

[21] H. F. Noller, V. Hoffarth, and L. Zimniak. *Science*, 256:1416 – 1419, 1992.

[22] J. A. Piccirilli, T. S. McConnell, A. J. Zaug, H. F. Noller, and T. R. Cech. *Science*, 256:1420 – 1424, 1992.

[23] T.R. Cech. *Science*, 236:1532 – 1539, 1987.

[24] A. C. Foster and R. H. Symons. *Cell*, 49:211 – 220, 1987.

[25] C. Papanicolaou, M. Gouy, and J. Ninio. *Nucleic Acids Research*, 12:31 – 44, 1984.

[26] H. M. Martinez. *Methods in Enzymology*, 183:306 – 317, 1990.

[27] J. P. Abrahams, M. v. d. Berg, E. v. Batenburg, and C. Pleji. *Nucleic Acids Research*, 18(10):3035 – 3044, 1990.

[28] M. Zuker. *Science*, 244:48 – 52, 1989.

[29] A. T. Brünger. *Acta Crystallographica A*, 45:46 – 57, 1990.

[30] C. Cheong, G. Varani, and I. Tinoco(Jr.). *Nature*, 346:680 – 682, 1990.

[31] N. C. Seeman, J. M. Rosenberg, F. L. Suddath, J. J. P. Kim, and A. Rich. *Journal of Molecular Biology*, 104:109, 1976.

[32] J. M. Rosenberg, N. C. Seeman, R. O. Day, and A. Rich. *Journal of Molecular Biology*, 104:145, 1976.

[33] S.-H. Kim, G. J. Quigley, F. L. Suddath, A. McPherson, and D. Sneden. *Science*, 179:285 – 288, 1973.

[34] J. L. Sussman, S. R. Holbrook, R. W. Warrant, G. M. Church, and S.-H. Kim. *Journal of Molecular Biology*, 123:607 – 630, 1978.

[35] D. Moras, M. B. Comarmond, J. Fischer, R. Weiss, and J. C. Thierry. *Nature*, 288:669, 1980.

[36] R. W. Schevitz, A. D. Podjarny, Krishnanachari, J. J. Hughes, P. B. Sigler, and J. L. Sussman. *Nature*, 278:188, 1979.

[37] N. H. Woo, B. A. Roe, and A. Rich. *Nature*, 286:346 – 351, 1980.

[38] H. T. Wright, P. C. Manor, K. Beurling, R. L. Karpel, and J. Fresco. In P. R. Schimmel, D. Soll, and J. N. Abelson, editors, *Transfer RNA: Structure, Properties, and Recognition*, pages 145 – 160. Cold Spring Harbor, NY Cold Spring Harbor Lab. Press, 1979.

[39] R. Basavappa and P. B. Sigler. *The EMBO Journal*, 10(10):3105, 1991.

[40] H. W. Pley, K. M. Flaherty, and D. B. McKay. *Nature*, 372:68 – 74, 1994.

[41] S. Arnott, D. W. L. Hukins, and S. D. Dover. *Biochim. Biophys. Res. Commun.*, 48:1392 – 1399, 1972.

[42] S. Arnott, D. W. L. Hukins, S. D. Dover, W. Fuller, and A. R. Hudgson. *Journal of Molecular Biology*, 81:107 – 122, 1973.

[43] K. Wüthrich. *NMR of Proteins and Nucleic Acids.* New York: Wiley, 1986.

[44] G. Varani, C. Cheong, and I. Tinoco(Jr.). *Biochemistry*, 30:3280 – 3289, 1991.

[45] G. Varani and I. Tinoco(Jr.). *Quarterly Reviews of Biophysics*, 24(4):479 – 532, 1991.

[46] H. A. Heus and A. Pardi. *Science*, 253:191 – 194, 1991.

[47] S. A. White, M. Nilges, A. Huang, A. T. Brünger, and P. B. Moore. *Biochemistry*, 31:1610, 1992.

[48] B. Wimberly, G. Varani, and I. Tinoco(Jr.). *Biochemistry*, 32:1078 – 1087, 1993.

[49] G. Varani, B. Wimberly, and I. Tinoco(Jr.). *Biochemistry*, 28(19):7761, 1989.

[50] J. D. Puglisi, J. R. Wyatt, and I. Tinoco(Jr.). *Journal of Molecular Biology*, 214:437, 1990.

[51] J. D. Puglisi, J. R. Wyatt, and I. Tinoco(Jr.). *Biochemistry*, 29:4215 – 4226, 1990.

[52] J. D. Puglisi, J. R. Wyatt, and I. Tinoco(Jr.). *Science*, 257:76 – 80, 1992.

[53] D. Neuhaus and M. Williamson. *The Nuclear Overhauser Effect in Structural and Conformational Analysis.* New York: VCH, 1989.

[54] A. E. Derome. *Modern NMR Techniques for Chemistry Research*. New York: Pergamon, 1987.

[55] M. Karplus. *Journal of Chemical Physics*, 30:11 – 15, 1959.

[56] G. M. Clore and A. M. Gronenborn. *Science*, 252:1390 – 1399, 1991.

[57] R. Powers, D. S. Garrett, C. J. March, E. A. Frieden, and A. M. Gronenborn. *Science*, 256:1673 – 1677, 1992.

[58] L. P. McIntosh and F. W. Dahlquist. *Quarterly Reviews of Biophysics*, 23:1 – 38, 1990.

[59] E. P. Nikonowicz and A. Pardi. *Nature*, 355:184 – 186, 1992.

[60] E. P. Nikonowicz and A. Pardi. *Journal of the American Chemical Society*, 114:1082 – 1083, 1992.

[61] R. T. Batey, M. Inada, E. Kujawinski, and J. D. Puglisi. *Nucleic Acids Research*, 20:4515 – 4523, 1992.

[62] E. P. Nikonowicz, A. Sirr, P. Legault, F. M. Jucker, and L. M. Baer. *Nucleic Acids Research*, 20:4507 – 4513, 1992.

[63] E. S. Simon, S. Grabowski, and G. M. Whitesides. *Journal of the American Chemical Society*, 111:8920 – 8921, 1989.

[64] E. S. Simon, S. Grabowski, and G. M. Whitesides. *Journal of Organic Chemistry*, 55:1834 – 1841, 1990.

[65] D. H. Andrews. *Physics Reviews*, 36:544, 1930.

[66] T.L. Hill. *Journal of Chemical Physics*, 14:465, 1946.

[67] I. Dostrovsky, E. D. Hughes, and C. K. Ingold. *Journal of the Chemical Society*, page 173, 1946.

[68] P. de la Mare, L. Fowden, E. D. Hughes, C. K. Ingold, and J. Mackie. *Journal of the Chemical Society*, page 3200, 1955.

[69] M. Born and J. R. Oppenheimer. *Annalen der Physik*, 84:457, 1927.

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

[71] S. Lifson and A. Warshel. *Journal of Chemical Physics*, 49:5116, 1968.

[72] O. Ermer. *Structural Bonding (Berlin)*, 27:161, 1976.

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

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

[75] A. Warshel and S. Lifson. *Journal of Chemical Physics*, 53:582, 1970.

[76] A. T. Hagler and S. Lifson. *S.Acta.Crystallogr., Sect B*, 30:619, 1974.

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

[78] D. H. Wertz and N. L. Allinger. *Tetrahedron*, 35:3, 1979.

[79] R. H. Boyd. *Journal of Chemical Physics*, 49:2574, 1968.

[80] N. L. Allinger. *Advances in Physical Organic Chemistry*, 13:1, 1976.

[81] S. Kirkpatrick, C. Gelatt Jr., and M. Vecchi. *Science*, 220:671 – 680, 1983.

[82] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. *Journal of Chemical Physics*, 21:1087, 1953.

[83] H. Bremermann. *Mathematical Biosciences*, 9:1 – 15, 1970.

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

[85] E. v. Kitzing. *Methods in Enzymology*, 211:449 – 467, 1992.

[86] M. Karplus and G. A. Petsko. *Nature*, 347:631 – 639, 1990.

[87] W. F. van Gunsteren and H. J. C. Berendsen. *Angewandte Chemie*, 102:1020 – 1055, 1990.

[88] D. A. Pearlman, D. A. Case, J. C. Caldwell, G. L. Seibel, C. Singh, P. Weiner, and P. A. Kollman. 1991.

[89] R. Lavery, H. Sklenar, K. Zakrzewska, and B. Pullman. *Journal of Biomolecular Structure and Dynamics*, 3:989 – 1014, 1986.

[90] R. Lavery. In W. K. Olson, M. H. Sarma, R. H. Sarma, and M. Sundaralingam, editors, *Structure & Expression Volume 3 : DNA Bending and Curvature*, pages 191 – 211. Adenine, Schenectady, New York, 1987.

[91] R. Lavery and H. Sklenar. *Journal of Biomolecular Structure and Dynamics*, 6:655 – 667, 1989.

[92] R. Lavery. *Laboratorie de Biochimie Theorique CNRS, Institute de Biologie Physico-Chimique, Paris*, 1992.

[93] R. Lavery, K. Zakrzewska, and A. Pullman. *Journal of Biomolecular Structure and Dynamics*, 4:443 – 461, 1986.

[94] R. Lavery, I. Parker, and J. Kendrick. *Journal of Biomolecular Structure and Dynamics*, 4:443 – 461, 1986.

[95] 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. *Journal of Molecular Biology*, 205:787 – 791, 1989.

[96] F. Major, M. Turcotte, D. Gautheret, G. Lapaplme, E. Fillion, and R. Cedergren. *Science*, 253(5025):1255 – 1260, 1991.

[97] D. Gautheret, F. Major, and R. Cedergren. *Journal of Molecular Biology*, 229(4):1049 – 1064, 1993.

[98] F. Major, D. Gautheret, and R. Cedergren. *Proceedings of the National Academy of Sciences*, 90:9408 – 9412, 1993.

[99] F. Leclerc, R. Cedergren, and A. E. Ellington. *Nature:Structural Biology*, 1:293 – 300, 1994.

[100] C. R. Woese, S. Winker, and R. R. Gutell. *Proceedings of the National Academy of Sciences*, 87:8467 – 8471, 1990.

[101] D. R. Groebe and O. C. Uhlenbeck. *Nucleic Acids Research*, 16:11725 – 11735, 1988.

[102] C. Tuerk, P. Gauss, C. Thermes, D. R. Groebe, M. Gayle, N. Guild, G. Stormo, Y. DÁubenton-Carafa, O. C. Uhlenbeck, I. Tinoco(Jr.), E. N. Brody, and L. Gold. *Proceedings of the National Academy of Sciences*, 85:1364 – 1368, 1988.

[103] T. Sakata, H. Hiroaki, Y. Oda, T. Tanaka, M. Ikehara, and S. Uesugi. *Nucleic Acids Research*, 18(13):3831 – 3839, 1990.

[104] F. Michel and E. Westhof. *Journal of Molecular Biology*, 216:585 – 610, 1990.

[105] F. L. Murphy and T. R. Cech. *Journal of Molecular Biology*, 236(1):49 – 63, 1994.

[106] L. Jaeger, F. Michel, and E. Westhof. *Journal of Molecular Biology*, 236:1271 – 1276, 1994.

[107] H. W. Pley, K. M. Flaherty, and D. B. McKay. *Nature*, 372:111 – 113, 1994.

[108] A. Kajava and H. Rüterjans. *Nucleic Acids Research*, 21(19):4556 – 4562, 1993.

[109] C. R. Woese, L. J. Magrum, R. Gupta, R. B. Siegel, D. A. Stahl, J. Kop, N. Crawford, J. Brosius, R. Gutell, J. J. Hogan, and H. F. Noller. *Nucleic Acids Research*, 8:2275 – 2293, 1980.

[110] W. Traub and J. L. Sussman. *Nucleic Acids Research*, 10:2701 – 2708, 1982.

[111] R. R. Gutell, N. Larsen, and C. R. Woese. *Microbiological Reviews*, 58:10 – 26, 1994.

[112] T. Brown, G. A. Leonard, E. D. Booths, and J. Chambers. *Journal of Molecular Biology*, 207:455 – 457, 1989.

[113] X. Gao and D. J. Patel. *Journal of the American Chemical Society*, 110:5178 – 5182, 1988.

[114] A. Lane, S. R. Martin, S. Ebel, and T. Brown. *Biochemistry*, 31:12087 – 12095, 1992.

[115] Y. Li, G. Zon, and D. Wilson. *Proceedings of the National Academy of Sciences*, 88:26 – 30, 1991.

[116] K. Maskos, B. M. Gunn, D. A. LeBlanc, and K. M. Morden. *Biochemistry*, 32:3583 – 3595, 1993.

[117] J. SantaLucia(Jr.) and D. H. Turner. *Biochemistry*, 32:12612 – 12623, 1993.

[118] E. Westhof, P. Romby, P. J. Romaniuk, J.-P. Ebel, C. Ehresmann, and B. Ehresmann. *Journal of Molecular Biology*, 207(2):417, 1989.

[119] A. A. Szewak, P. B. Moore, Y.-L. Chan, and I. G. Wool. *Proceedings of the National Academy of Sciences*, 90:9581, 1993.

[120] J. SantaLucia(Jr.), R. Kierzek, and D. H. Turner. *Science*, 256:217 – 219, 1992.

[121] C. Zwieb. *Journal of Biological Chemistry*, 267(22):15650, 1992.

[122] D. Gautheret, D. Konnings, and R. R. Gutell. *Journal of Molecular Biology*, 242:1 – 8, 1994.

[123] J. Wolters. *Nucleic Acids Research*, 20:1843 – 1850, 1992.

[124] P. W. Davis, W. Thurmes, and I. Tinoco(Jr.). *Nucleic Acids Research*, 21(3):537 – 545, 1993.

[125] D. Konnings. *private communications*, 1994.

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

8 Curriculum Vitæ

Name : Herbert Friedrich Kratky
Geburtsdatum und -ort : 5. Juni 1968 in Wien
Staatsbürgerschaft : Österreich
Wohnort : 1100 Wien; Laxenburgerstraße 128/27/10

Schulbildung :

1974 – 1978 : Volksschule in Wien
1978 – 1986 : Realgymnasium Pichelmayergasse 1 (BRG X), mathematischer Zweig
4. Juni 1986 : Matura mit Auszeichnung bestanden

Studium :

September 1986 : Immatrikulation an der Universität Wien (Studienrichtung Chemie; Studienzweig Chemie)
5. Juli 1989 : 1. Diplomprüfung
Juli 1992 : Beginn der Diplomarbeit am Institut für Theoretische Chemie bei Prof. Dr. Hans Lischka
Oktober 1993 : 2. Diplomprüfung mit Auszeichung bestanden
November 1993 : Sponsion zum Magister der Naturwissenschaften
Beginn der Dissertation bei Prof. Dr. Peter Schuster

Index

1 Introduction 3

2 Structural investigations on biopolymers 5
2.1 Structural features of nucleic acids 5
2.1.1 Definition of nucleic acid structure 5
2.1.2 Energy landscapes 10
2.2 Experimental techniques 12
2.2.1 X-ray crystallography 13
2.2.2 NMR-spectroscopy 14

3 Computational methods 18
3.1 Molecular Mechanics 18
3.1.1 Introduction 18
3.1.2 Force fields 19
3.1.3 Structure optimization 24
3.1.4 Molecular dynamics 28
3.2 AMBER 4.0 30
3.3 JUMNA 7 33
3.4 MC-SYM 34
3.5 Conformational sampling 38
3.5.1 Randstruct 38
3.5.2 Randloop 42

4 Results: GNNA-tetraloops 44
4.1 Introduction 44
4.2 Minimum structures 46
4.3 Results and discussion 49
4.3.1 Structural features 49
4.3.2 Energies 58
4.4 Conformational search 63

5 Results: NUN-triloops 71
5.1 Introduction 71
5.2 Minimum structures 72
5.3 Results and discussion 75
5.3.1 Energies 75
5.3.2 Dimer structures 85

6 Conclusions 96
6.1 GNNA-tetraloops 96
6.2 NUN-triloops 97
6.3 Outlook 100

7 Literature 102

8 Curriculum Vitæ 109