Force Field based Investigations on Structure and Dynamics of RNA Molecules

DISSERTATION
zur Erlangung des akademischen Grades
Doctor rerum naturalium

Vorgelegt der
Fakultät für Naturwissenschaften und Mathematik
der Universität Wien

von
Dipl. Ing. Wolfgang A. Svrcek-Seiler

im Januar 2003

An dieser Stelle möchte ich mich bei all jenen bedanken, die direkt oder indirekt zur Entstehung dieser Arbeit beigetragen haben.

Professor Peter Schuster danke ich für die Aufnahme in diese Arbeitsgruppe und für die wissenschaftliche Unterstützung. Diskussionen mit ihm halfen immer wieder, den Überblick zu bewahren und das Wesentliche nicht aus den Augen zu verlieren.

Professor Peter Stadler stand oft bereitwillig zur Verfügung, um mathematische und physikalische Fragen zu erörtern. Auch seine Erklärung zur Etymologie des Terminus "Force Field" wird mir in Erinnerung bleiben.

Dr. Ivo Hofacker gebührt besonderer Dank für unzählige interessante Diskussionen, die viel zu meinem Verständnis beigetragen haben. Dies gilt auch für Dr. Kurt Grünberger und Dr. Christoph Flamm. Mag. Roman Stocsits war und ist eine unerschöpfliche und geduldige Quelle biologischen Wissens.

Alle anderen Kolleginnen und Kollegen aus dieser Arbeitsgruppe trugen und tragen permanent zum angenehmen Arbeitsklima in unserer Gruppe bei, das letztendlich auch einen Teil des Reizes zu forschen ausmacht.

Dr. Stefan Boresch hat durch seine stete Bereitschaft, seine Erfahrung über auf Kraftfeldern basierende Methoden zu teilen und diesbezügliche Probleme zu diskutieren, ebenfalls viel zur Entstehung dieser Arbeit beigetragen. Als Kurzzeitgast in der Arbeitsgruppe Molekulardynamik und Biomolekulare Simulation habe ich mich stets willkommen gefühlt.

Dr. Martin Zacharias, Dr. Alexey Onufriev und Professor David Case verhalfen mir durch per e-mail geführte Diskussionen zu vielen wesentlichen Einsichten.

Ohne die Unterstützung durch meine Familie wäre mein Studium nicht möglich gewesen. Der ihr gebührende Dank geht weit über das hinaus, was sich in wenigen Zeilen ausdrücken läßt.

Meiner Freundin Teresa Egger danke ich für ihre Geduld und Toleranz und dafür, daß sie verzeiht, wenn ich köperlich anwesend, aber geistig bei der Arbeit bin.

Abstract

Experimental investigations on three-dimensional structures of biomolecules are currently costly both in terms of time and material resources. Besides, they provide mostly time averaged static structures, lacking detailed information about the thermal motion of these molecules.

In this thesis, the dynamics of frequently occurring and evolutionarily conserved structural motifs of RNA, namely GNRA and UUCG tetraloops, are investigated in detail by the use of molecular dynamics simulations based on the AMBER force field. The results show excellent agreement with experimental results and yield insights that can not be obtained by current experimental methods.

The molecular dynamics based investigations are extended to an RNA helix containing an adenosine bulge. For this structure, the observed transitions between different conformations of the bulged base partially elucidate the pathway between environment-dependent structural differences observed in experiment. The results obtained on this structural motif are in good agreement with similar studies on adenosine bulges in DNA double helices.

An extensive search for GCAA tetraloop conformations of low energies outlines the accuracy limit of current force fields since conformations of lower energy than any experimentally determined structure were found. The false optimal conformations are a valuable source of information for the improvement of force field parameters.

Two small pseudoknot structures are presented, that were modeled based on constraints inferred from sequence and base pairing pattern, as well as assumptions based on knowledge gained from other RNA structures. After force field based extensive optimization, the latter pseudoknot, whose experimentally determined structure is available, shows good qualitative agreement with the experimental reference structure. Deviations of the resulting modeled structure from the experimentally determined reference structure provide useful hints for future improvement of the modeling process.

Zusammenfassung

Die experimentelle Strukturbestimmung von Biopolymeren ist sowohl zeit- als auch materialaufwändig und liefert als Ergebnis hauptsächlich zeitgemittelte, statische Strukturen.

Im Rahmen der vorliegenden Dissertation wird das dynamische Verhalten von häufig vorkommenden und evolutionär konservierten RNA - Strukturmotiven im Detail untersucht. Die Untersuchungen basieren auf Molekulardynamik-Simulationen unter Verwendung des AMBER Kraftfeldes. Die Ergebnisse zeigen nicht nur gute Übereinstimmung mit experimentell ermittelten Strukturen. Sie liefern auch neue und detaillierte Einsichten in Übergänge zwischen deutlich unterschiedlichen Konformationen in gegenwärtig experimentell nicht zugänglicher Auflösung.

Weitere Molekulardynamik-Simulationen einer RNA Doppelhelix, die ein ungepaartes Adenosin Nucleotid (Adenosin-Bulge) enthält, liefern detaillierte Informationen über einen Teil des Übergangs zwischen umgebungsabhängig unterschiedlichen Konformationen von Adenosin Bulges. Diese Ergebnisse stimmen gut mit jenen anderer Gruppen über Adenosin Bulges in DNA Doppelhelices überein.

Die Ergebnisse aus ausfürlicher Konformationssuche nach energetisch bevorzugten GCAA Tetraloops zeigen deutlich die Grenzen von auf Kraftfeldern basierenden Vorhersagemethoden auf. Es wurden Loopkonformationen gefunden, die sich deutlich von jeder experimentell ermittelten Struktur unterscheiden, aber vom Kraftfeld energetisch besser bewertet werden. Diese Strukturen sind von grossem Interesse für die Weiterentwicklung der Parametrisierung von Kraftfeldern.

Es werden weiters zwei Pseudoknotenstrukturen vorgestellt, deren Modellierung auf deren Sequenz und Sekundärstruktur, sowie auf strukturellen Einschränkungen beruht, die aus allgemeinen Erkenntnissen über die dreidimensionale Struktur von RNA molekülen abgeleitet wurden. Für die letztere der beiden Strukturen existiert eine röngtenkristallographisch ermittelte Referenzstruktur. Nach sehr zeitintensiver und ausführlicher, auf dem AMBER Kraftfeld basierender Minimierung resultiert gute qualitative, wenn auch nicht gänzlich zufriedenstellende Übereinstimmung zwischen modellierter und experimentell bestimmter Struktur. Die Abweichungen im Detail liefern nützliche Hinweise zur künftigen Verbesserung der Modellierung.

Contents

1 Introduction

1.1 General Context

The biopolymers which are fundamental in molecular genetics are nucleic acids, DNA and RNA, and proteins. Originally, the flow of information during gene expression was viewed as essentially unidirectional, with DNA being copied to both itself and to RNA, and RNA being subsequently decoded for protein synthesis. Proteins were assumed to be the only biomolecules with catalytic properties.

Within the last decades, this view has given way to a more detailed understanding due to several important discoveries. The study of viruses having a RNA genome showed that RNA, like DNA, can serve as a primary information-encoding molecule. The process of transcription from RNA to cDNA is termed reverse transcription. In addition, and more importantly, various types of RNA molecules possessing catalytic properties have been found, the first ones independently by the groups of Cech and Altman in the 1980s [23,49,50]. Recently, DNA molecules exhibiting catalytic behavior (deoxyribozymes) have also been discovered [20]. Last but not least, aptamers, RNA molecules able to specifically bind to substrate molecules have been discovered by *in vitro* selection [40, 82].

The function of biomolecules is determined by their structure, which led to a growing interest in the three-dimensional shapes of RNA molecules. At a coarse grained (two-dimensional) level, RNA structure can be described and predicted by computational methods. This level is termed *secondary structure*: During the formation of the three-dimensional thermodynamically most stable structure, RNA folds back onto itself, and the unfavorable interaction between hydrophobic bases and polar solvent is minimized by the formation of base pairs that stack on each other. The secondary structure is then defined as the pattern of base pairs. Since not all combinations of bases can form base pairs that fit into a Watson-Crick helix, the secondary structure is determined by the sequence of bases. In RNA, there are four types of bases, adenine (**A**), guanine (**G**), cytosine (**C**) and uracil (**U**). The hydrogen bonds which determine the geometry of pairing can be formed between the complementary Watson-Crick pairs G-C and A-U, as well as the less stable G-U wobble pair. Secondary structures are a very useful coarse grained representation, since they account for most of the free energy of folding and, at this level, the interaction is binary and digital, in the sense that two bases

either do or do not form a base pair. The thermodynamically most stable secondary structures, called minimum free energy (mfe) structures, can be calculated efficiently by the use of *dynamic programming* algorithms [122].

However, as mentioned above, the function of biomolecules is determined by their three dimensional shape or tertiary structure. Furthermore, a full description of biomolecules at the most detailed level must also include their dynamic behavior.

To investigate static (energetic) and dynamic properties of biomolecules *in silico*, they are described by approximate models based on the principles of classical mechanics. This approach is termed molecular mechanics and is based on force fields, which provide an approximation of the self- and interaction energies of molecules. Since force fields yield intra- and intermolecular forces, they allow for the simulation of the thermal motion of molecules via Newton's second law. This approach is termed molecular dynamics. Molecular dynamics simulations can yield detailed insights into energetic and dynamic aspects of biomolecules and in addition, force field based refinement is an integral part of experimental structure determination.

In this work, detailed analyses of the simulation of mostly small but functionally important structural motifs of RNA are presented with emphasis on structural transitions. It is also shown that, at least for nucleic acids, force fields are not yet accurate enough to be of use for structure prediction without additional restraints based on experimental results. While this finding might be deemed unsatisfactory, the false optimal structures found are valuable for the future improvement of force fields.

1.2 Outline of this Thesis

In the next chapter, we shall give a description of the chemical structure of nucleic acids, in particular RNA. In addition, a more detailed description of the hierarchy of structural representations is presented.

Chapter 3 shows a brief overview of the experimental techniques for the determination of the tertiary structure of biomolecules.

In chapter 4, the computational methods used throughout this work are described in detail. After a brief introduction to methods for structure creation, force fields, as they are implemented for the simulation of large biomolecules, are outlined. Special emphasis is laid on the generalized Born approximation for the treatment of solvent polarization effects. The understanding of this method and its limitations is of utmost importance for the

assessment of obtained results. As a 'technical demonstration', it is shown how this method might be improved by the use of surface integrals. In this chapter we also discuss the process of structure optimization by energy minimization, with emphasis on the technical consequences of the roughness of the energy landscape defined by force fields. This is followed by an introduction to molecular dynamics and methods to constrain the lengths of chemical bonds during molecular dynamics simulations. Finally, a newly conceived and implemented method for the truncation of the Coulomb potential is outlined.

Chapter 5 describes our implementation of principal component analysis (PCA) of molecular structures. Within this work, PCA is used as a heuristic, but nonetheless valuable tool for the coarse grained analysis of molecular dynamics trajectories.

In chapter 6, we outline different methods for conformational search devised and applied in this work.

Chapter 7 provides a short overview of the software packages used in this work, as well as as of numerous extensions to them that were newly developed and implemented by us.

The results are presented and discussed in chapter 8.

2 Nucleic Acid Structure

2.1 Chemical Structure of Nucleic Acids

Nucleic acids are linear heteropolymers consisting of a sequence of monomers called nucleotides. These nucleotides are adenylic acid, guanylic acid, cytidylic acid and uridylic acid. In DNA, uridylic acid is replaced by a functionally equivalent thymidylic acid. Each nucleotide consists of three molecular constituents:

Phosphate Group The phosphates are linking the nucleotides. Carrying an electron charge, they are responsible for the polyanionic character of nucleic acids.

Ribose The ribose unit is of furanoside type (β-D-ribose in RNA and β-D-2'-deoxyribose in DNA). It is phosphorylated at the 5'-position. Ribose and base are linked by the glycosidic bond between atom C1' of the pentose and base atoms N1 of purines or N9 of pyrimidines. The replacement of the 2'-OH in RNA by a hydrogen in DNA has important consequences. The presence of the 2' hydroxil in RNA renders it more rigid than DNA and specific interactions between the hydroxil group and other parts of the molecule lead to the stabilization of structural motifs like hairpin loops.

Heterocyclic bases The heterocyclic bases are the purine bases adenine (A) and guanine (G) and the pyrimidine bases cytosine (C) and uracil (U) (replaced by thymine in DNA).

Figure 1 shows a short strand of RNA containing four nucleotides containing the bases adenine (A), and guanine (G), cytosine (C) and uracil (U). The constituents described above are shown in different colors (green - phosphate, blue - ribose, black - purine (G, A) and pyrimidine (C, U) bases). All four monomers are connected to a single strand, which is directional and starts at the 5′-end (top left of figure 1) and ends at the 3′-end (bottom of figure 1). It should be noted that several naturally occurring modified nucleotides exist besides those containing the bases A, G, C and U. Such modified nucleotides for example stabilize the secondary and tertiary structure of tRNA.

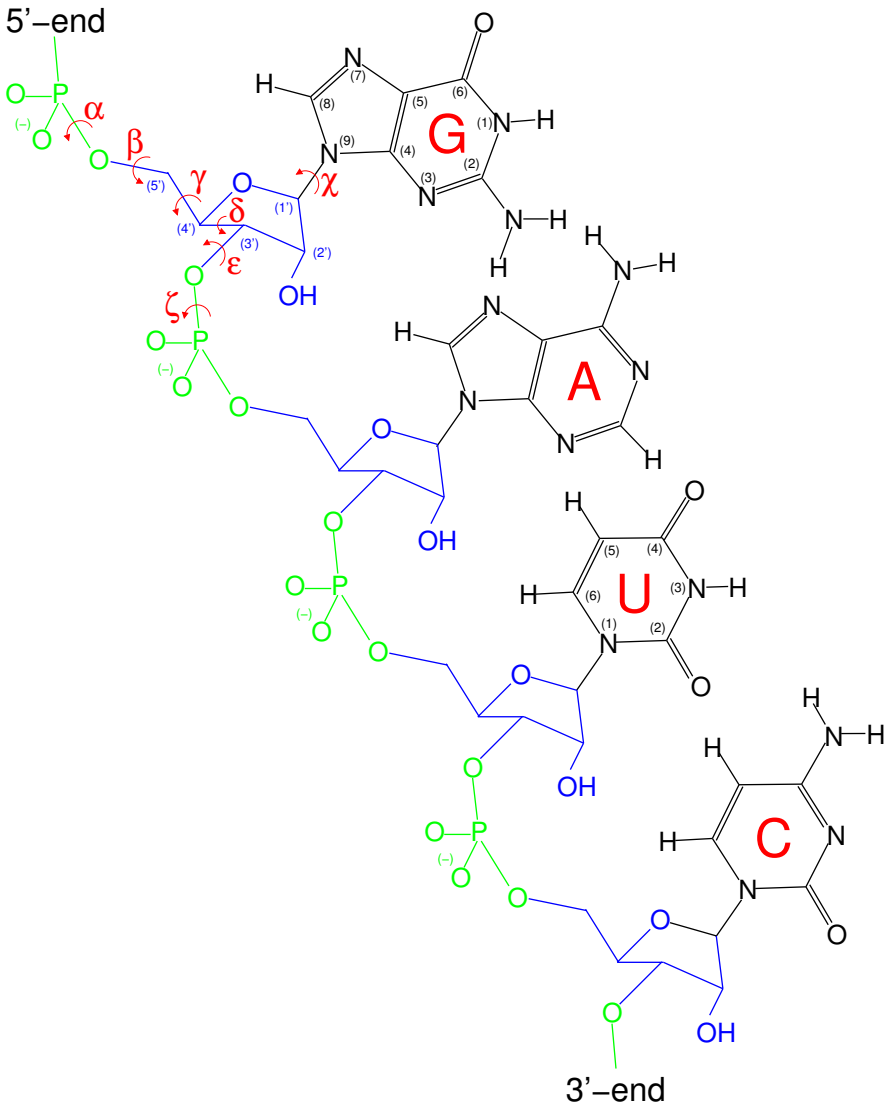


Figure 1: Atomic structure of RNA: (green - phosphate, blue - ribose, black - purine (G, A) and pyrimidine (C, U) bases)

Also shown are the backbone torsion angles α to ζ and the glycosidic torsion angle χ. The nucleobases are heterocyclic ring compounds and therefore rigid substructures. The conformation of the pentoses can be described by the phase of the sugar pucker defined further down. Therefore, the overall flexibility of nucleic acid molecules is due to changes of the torsion angles shown and the three-dimensional structure of RNA molecules can be represented by the torsion angles and the phase of the sugar pucker. The six backbone torsion angles are defined as follows:

$$\begin{array}{lcccccccc}
\alpha & = & \angle\ \mathrm{O3'} & - & \mathrm{P} & - & \mathrm{O5'} & - & \mathrm{C5'} \\
\beta & = & \angle\ \mathrm{P} & - & \mathrm{O5'} & - & \mathrm{C5'} & - & \mathrm{C4'} \\
\gamma & = & \angle\ \mathrm{O5'} & - & \mathrm{C5'} & - & \mathrm{C4'} & - & \mathrm{C3'} \\
\delta & = & \angle\ \mathrm{C5'} & - & \mathrm{C4'} & - & \mathrm{C3'} & - & \mathrm{O3'} \\
\epsilon & = & \angle\ \mathrm{C4'} & - & \mathrm{C3'} & - & \mathrm{O3'} & - & \mathrm{P} \\
\zeta & = & \angle\ \mathrm{C3'} & - & \mathrm{O3'} & - & \mathrm{P} & - & \mathrm{O5'}
\end{array}$$

Angle χ ($\angle$ O1'-C1'-N9-C4 in purines and $\angle$ O1'-C1'-N1-C2 in pyrimidines) describes the orientation of the heterocycle with respect to the sugar ring. Its values are restricted to two distinct regions, around 0 and around 180 degrees. If the heterocycle is rotated towards the C5'-atom ($\chi \approx 0°$), the conformation is called *syn*, if the heterocycle is turned away from the C5'-atom ($\chi \approx 180°$), the conformation is called *anti*. In naturally occurring RNA, the *syn* conformation is rare for purine bases and has, to our knowledge, not at all been observed for pyrimidine bases. Angle δ is determined by the conformation of the furanose ring and can therefore be substituted by the sugar pucker angle P.

The sugar ring conformation is best described using the concept of pseudorotation. Since a five membered ring cannot adopt planar geometry, one or two of the ring atoms must lie above or below the plane defined by the other three atoms. If the atom is on the same side of the plane as atom C5', the conformation is called *endo*, if it is on the opposite side, the conformation is called *exo*. Figure 2 shows the relation between the sugar pucker angle P and the ribose conformation and the two favored pucker modes in RNA, C2'-end and C3'-endo.

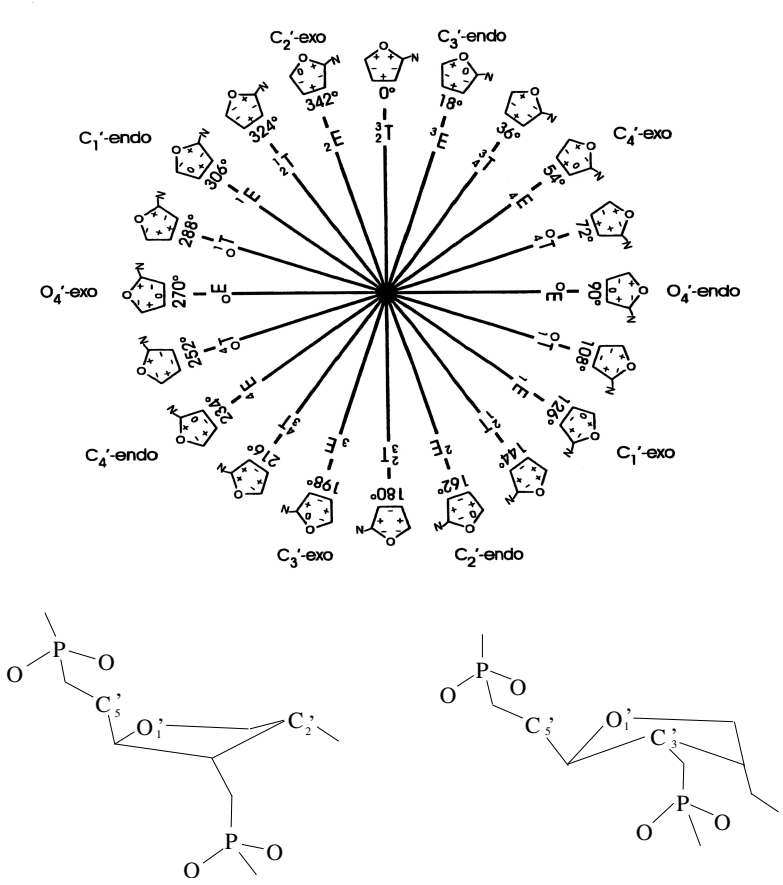


Figure 2: Sugar pucker wheel and favored pucker modes in RNA, C2'-endo - lower left and C3'-endo - lower right.

C3'-endo is the 'standard' sugar conformation in A-RNA helices. The C2'-endo conformation mostly occurs in regions where the backbone direction changes (e.g. small loops), since it implies an elongation of the backbone. The sugar pucker P is defined through the 5 consecutive endocyclic torsion angles of the pentose ring:

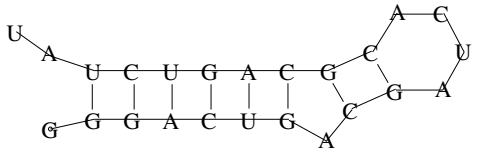
$$\tan P = \frac{\nu_4 + \nu_1 - \nu_3 - \nu_0}{2\,\nu_2(\sin\frac{\pi}{5} + \sin\frac{2\pi}{5})} \tag{1}$$

2.2 Three Levels of Structure Representation

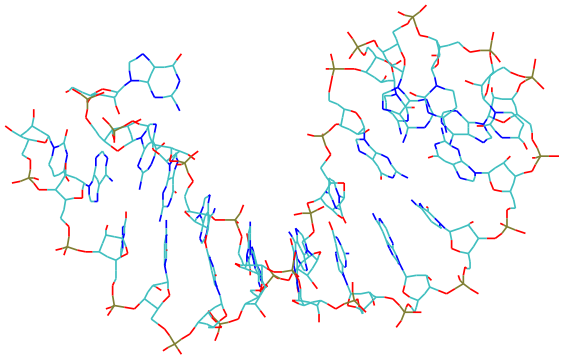
RNA structure is generally described at three different levels of detail. These levels are termed primary, secondary and tertiary structure. Figure 3 shows primary, secondary and tertiary structure of a stem-loop structure with an adenosine bulge in the stem, taken from `1tfn.pdb` [65].

5′-GGGACUGACGAUCACGCAGUCUAU-3′

(a) Primary structure



(b) Secondary structure



(c) Tertiary structure

Figure 3: Primary structure (sequence) extracted from `1tfn.pdb`, secondary structure calculated with RNAfold from the Vienna RNA package [58, 74, 122] (minimum free energy structure) and tertiary structure from the NMR model with PDB identifier `1tfn`.

The most coarse grained description of RNA molecules is given by the declaration of the nucleotides' order in the direction from 5' to the 3' end of the sequence. This results in a string of the four letters **A**, **G**, **C** and **U**. This string is termed *primary structure*.
RNA can form stable double helices of complementary strands. Since RNA usually occurs single stranded, formation of double helices is accomplished by the molecule folding back onto itself to form Watson-Crick ($\mathbf{G} \equiv \mathbf{C}$ and $\mathbf{A} = \mathbf{U}$) base pairs or the slightly less stable $\mathbf{G} = \mathbf{U}$ pairs. The *secondary structure* can be depicted as shown in the middle of figure 3, but this representation can be misleading. Though some distance constraints for the three dimensional structure of an RNA molecule can be inferred from the secondary structure, it is only a list of paired or unpaired bases and not a two dimensional depiction of the actual (three-dimensional) structure. It is, in other words, a topological description, not a geometric one. In slightly more rigorous terms, the secondary structure is a planar vertex-labeled graph with the bases representing the vertices and the edges representing linkages between adjacent nucleotides and base pairs. Based on this definition, the prediction of minimum free energy secondary structures is possible by the use of dynamics programming algorithms [122].

Primary and secondary structure provide a tractable model system to investigate genotype-phenotype relationships and extensive investigations on this model system have provided many insights into the process of evolutionary adaptation [43].

RNA secondary structures can be built up from a small set of building blocks or motifs shown in figure 4. These motifs are loops and external elements. The loops are specified by their degree, i.e. the number of stack terminating base pairs they contain and their size, defined as the number of unpaired bases inside the loop. External elements are unpaired regions that are not part of a loop, for example segments connecting substructures or free 5' and 3' ends. According to this definition, **hairpin loops** are loops of degree one and arbitrary size. **Internal loops** are loops of degree two and arbitrary size and bulges can be viewed as internal loops with unpaired bases occuring only on one side of a stem. **Stacks** consist of consecutive base pairs and form a loop of degree two and size 0. **Multiloops** are internal loops of degree greater than 2 and arbitrary size. Based on these definitions, each nucleotide of a secondary structure, as shown for example in figure 4 can be uniquely assigned to one of these motifs.

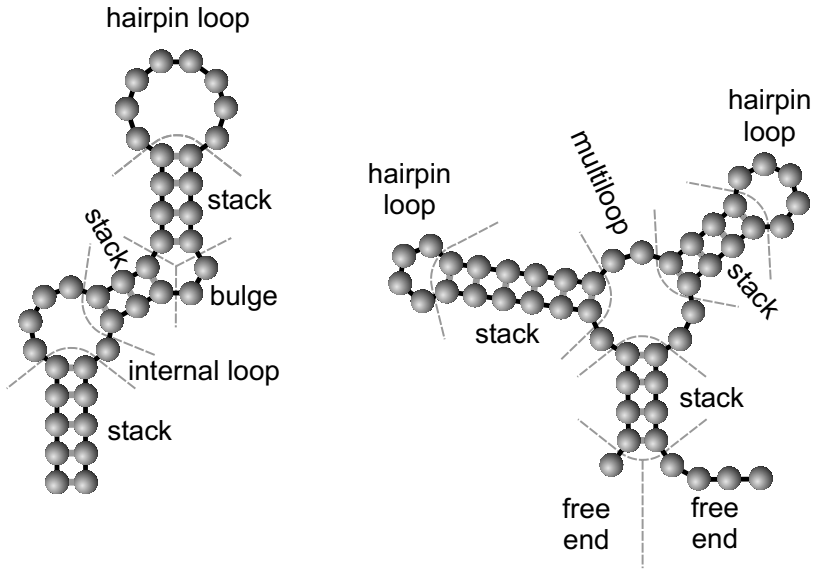


Figure 4: RNA secondary stuctures built from the motifs described in the text.

The most detailed representation of RNA structure is the so called *tertiary structure*, representing the molecular structure as a three-dimensional object. Assuming parts of the molecule (for example bond lengths and angles and the heterocycles) as rigid, at least six torsion angles and the conformation of the ribose (the sugar pucker) of each nucleotide are necessary to completely specify the three dimensional structure of a RNA molecule. The most detailed description of the tertiary structure is given by the Cartesian coordinates of all atoms contained in the molecule. For *in silico* investigations on the structure and dynamics of RNA molecules, it is necessary to take into account all atomic coordinates. This is, however, more accuracy than experimental investigations can yield. Several structural motifs can only be described in terms of the tertiary structure. These are for example base-triples [19,22,98], G-quartets [66], A-platforms [22] and pseudoknots [84].

3 Experimental Techniques

NMR spectroscopy and X-ray crystallography are currently the most successful techniques for determining the structure of biological macromolecules like proteins and nucleic acids at atomic resolution.

Structure determination by X-ray crystallography involves the analysis of the X-ray diffraction pattern produced when a beam of X-rays is directed onto a well-ordered crystal. The diffraction pattern is recorded on a detector and analyzed on the basis of Bragg's law. Based on the diffraction pattern obtained by X-ray scattering from the periodic assembly of molecules or atoms in the crystal, the electron density can be reconstructed. Additional phase information must be extracted either from the diffraction data or from supplementing diffraction experiments to complete the reconstruction. Both the crystallization process and the necessary analysis of the diffraction data make structure determination by X-ray crystallography a hard task. However, the continuing advances in the applied methods lead to highly refined biomolecular structures with resolutions at or even below two angstroms [77, 104]. It should be noted that the electron density at best only provides a 'blurred' picture of the actual structures. Additional knowledge, like sequence, bond lengths and angles must be used to construct atomic precision models from the data obtained by X-ray crystallography. The last stage of refinement is done by means of molecular mechanics and restrained molecular dynamics calculations.

In NMR (nuclear magnetic resonance) spectroscopy, strong magnetic fields and high frequency electromagnetic waves ($10^7 - 10^9$ Hz) are applied to probe the magnetic environment of the nuclei. The local environment of the nuclei determines the frequency of the resonance absorbtion: For a nucleus with spin 1/2, there are two possibilities for the orientation of the spin, either aligned with the external field (lower energy) or pointing in the opposite direction (higher energy). Since those two states have different energies, transitions from lower to higher energy state can be induced by the input of external energy from an oscillating electromagnetic field. To cause a transition in the spin state, the energy of the absorbed photons $h\nu$ must match the energy difference between the two spin orientations. The absorbtion (resonance) frequency is determined by the strength of the static external field and by the type of nucleus as well as small local perturbations due to its chemical environment. These perturbations cause the local magnetic field with which the spin interacts to differ from the applied external field, shifting

the resonance frequency. This shift is termed chemical shift. NMR spectra can be extended over two or more dimensions by replacing the single pulse with a sequence of pulses, separated by varying time intervals. Extension of the spectra to two dimensions leads to a reduction in spectral overlap and therefore to more accurate measurements. An indispensable technique for the determination of through-space inter-proton distances is nuclear Overhauser effect spectroscopy (NOESY). NOESY produces signals by transfer of magnetization via dipole-dipole interaction between nuclei which are close in distance but not connected through bonds. Since the strength of the obtained signal is proportional to the inverse sixth power if distance, distances of up to 5 Å can be measured. The through-space correlations obtained from NOESY measurements provide a dense network of distance constraints which provide the basis for structure determination of biomolecules.

Extending the measurements to heavier nuclei (^{13}C, ^{15}N and ^{31}P) leads to clearer spectra and allows the determination of much more interatomic distances. This technique is termed heteronuclear NMR.

Since NMR spectroscopy directly yields information mostly about relatively short interatomic distances, much additional input is again necessary to finally arrive at three dimensional molecular structures. The NMR data provide constraints and three dimensional structures are built so as to satisfy these as well as other known constraints like bond lengths, angles and additional chiral and distance constraints that can be inferred from other sources. One successful method for building molecular structures from distance constraints is distance geometry (see section 4.1.1). The initial models are then refined by minimizing an energy function comprising force field energy as well as the NMR constraints on distances and dihedral angles as energetic penalties [70]. Standard methods for minimizing these energy functions and thereby refining the structures are simulated annealing and simulated tempering, i.e. restrained molecular dynamics at alternating high and low temperatures. Depending on the amount of NMR data as well as actual structural variability, different structures having a low force field energy as well as constraint violation penalty can be obtained as final result. Therefore, NMR determined entries in the `PDB` data base often contain several different structures.

4 Computational Methods

For the *in silico* creation of molecular models in atomic detail, knowledge about sequence, secondary structure, geometric properties of regular substructures like nucleic acid helices and experimental constraints are combined to build model structures. The obtained structures are subsequently refined by methods based on molecular mechanics and molecular dynamics. In the following, we shall give an overview of the methods applied in this process.

4.1 Methods for Structure Creation

The most straightforward methods to create molecular models are based on rigid body transformation. Model structures are built piece-wise, using for (for example) single nucleotides as building blocks. These building blocks are combined using knowledge about the relative orientation of consecutive nucleotides. A widely used program based on this principle is MC-SYM (Macromolecular Conformations by SYMbolic programming) [45,46]. Within MC-SYM, RNA molecules are described in terms of functional constraints. Then these constraints are used to generate structures that are consistent with that description. MC-SYM structures are built from a small library of conformers for each of the four nucleotides and transformation matrices describing their relative orientation. This approach has been successfully applied in modeling GNRA tetraloops and their GNYA mutants [72]. However, if the structures to be modeled are larger, the number of models created by MC-SYM becomes prohibitively large and their order is arbitrary, but not random.

A more general approach, routinely used to build structures from experimental data is *distance geometry* [54].

4.1.1 Distance Geometry

The only necessary input for distance geometry are distances and, in the later stages of the procedure, chiral constraints. If *all* $N(N-1)/2$ interpoint distances from a set of N points are known, these distances can be embedded in $\mathbb{R}^3$ or -in other words- converted to a three dimensional object through the following procedure: The squared distance from each point to the centroid of the whole set (d_{0i}^2) is calculated from the interpoint distances

(see e.g. reference [70], pp. 428):

$$d_{0i}^2 = \frac{1}{N}\sum_{j=1}^{N} d_{ij}^2 - \frac{1}{N^2}\sum_{j=2}^{N}\sum_{k=1}^{j-1} d_{jk}^2$$

Then, defining the centroid as the origin of the coordinate system, the *metric matrix* $\mathbf{M}$ can be obtained by:

$$\mathbf{M}_{ij} = \mathbf{r}_i \cdot \mathbf{r}_j = \frac{1}{2}(d_{0i}^2 + d_{0j}^2 - d_{ij}^2) \tag{2}$$

The metric matrix $\mathbf{M}$ is positive semidefinite and of rank three (if the vectors $\mathbf{r}_i$ are three-dimensional). After diagonalization and sorting the eigenvalues by their size, $\mathbf{M}$ can be rewritten as

$$\mathbf{M} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^T \tag{3}$$

with $\mathbf{X}$ containing the normalized eigenvectors of $\mathbf{M}$ as columns and $\mathbf{\Lambda}$ being a diagonal matrix containing the corresponding eigenvalues. Comparing equations 2 and 3 shows that the matrix $\mathbf{X}\sqrt{\mathbf{\Lambda}}$ contains the Cartesian coordinates of the N points in the first three columns. The orientation of the embedded structure is such that the direction in which most coordinate variance occurs is parallel to the x-axis, in analogy with principal component analysis. It should be noted, that distances (and, equivalently, the metric matrix) contain no information about overall chirality. Of course, a three dimensional structure comprised of N points is strongly over-determined by all $N(N-1)/2$ interatomic distances.

The gist of distance geometry is that random models satisfying a possibly insufficient set of distance constraints can be generated. This usually (as implemented for example in the NAB [71] or TINKER [85] molecular modeling packages) happens in the following steps:

A bounds matrix is generated using all available structural information that can be expressed through interatomic distances. Interatomic distances for covalently bound atoms are set to the corresponding ideal bond length. 1–3 distances are set to the distances defined by the two known bond lengths and the corresponding bond angle. For atoms connected through a dihedral angle, lower and upper bound are set corresponding to the corresponding cis and trans 1–4 distances. The lower distance bound for every non-bonded atom pair is set to the sum of their Van der Waals radii. Further distance information must be inferred from chemical principles or experimental data.

Subsequently, the bounds matrix is refined by adjusting lower and upper distance bounds for every point triple so that the three corresponding distance bounds satisfy the triangle inequality. This procedure is termed triangle smoothing.

From this refined bounds matrix a distance matrix is generated by randomly choosing distance values between the lower and upper bound for each atom pair where no exact distance is available.

Since a distance matrix obtained this way is usually not embeddable in three dimensions, the distance matrix is embedded in four-dimensional space. The four-dimensional structure is then refined by minimizing a penalty function, penalizing violations of the input distance constraints as well as the fourth-dimension components of the individual atomic position vectors. Since the landscape defined by the distance violation penalty is very rough, the minimization is done by simulated annealing or simulated tempering.

The quality of the resulting models strongly depends on the number of consistent distance bounds taken as input. When, in the absence of experimental data, the input distance constraints are chosen based on educated guess or distance data bases (which are available within the NAB distribution), the amount of computer time necessary to achieve satisfactory models is tremendous. When modeling structures containing helical regions, it was found that modeling the helical regions as regular A-form helices and setting all interatomic distances within the helices to the values obtained from the A-form models generally improves the performance of this method.

4.2 Molecular Mechanics

4.2.1 Introduction

Molecular mechanics is a computational technique used to model the conformational behavior and energetic properties of molecules. While the most accurate description of those properties can in principle be achieved by the use of quantum mechanical methods, these are computationally too demanding - especially in the case of biopolymers, where the molecules of interest typically contain several hundreds (or even thousands) of atoms. Therefore, molecular properties are approximately described by empirical potentials, which are termed *force fields*, but are actually *potential* fields. A force field maps the *relative* positions of all atoms in the system considered onto an energy value. Furthermore the dynamics of molecules are assumed to be

governed by the laws of classical (Newtonian) rather than quantum mechanics. The accuracy of these approximations crucially depends on the validity of the Born-Oppenheimer approximation, within which the Hamilton operator for a (set of) molecule(s) is separated into two terms, describing the electronic wave function and the nuclear wave function respectively. The nuclear wave function is derived by solving an effective Schrödinger equation containing only nuclear coordinates.

4.2.2 Force Fields

Force Fields are built up by a number of analytical expressions that yield the potential energy of a (set of) molecule(s) as a function of its conformation. These analytical expressions are chosen in consideration of the underlying physics as well as computational efficiency. Contributions to the total energy can be separated into bonded and non-bonded parts. The bonded parts contain penalty terms for the deviation of bond lengths and bond angles from their respective reference values as well as torsional terms that represent the energy barrier(s) associated with rotating an atom about a covalent bond.

Bond Stretching
The most common analytical expression for a covalent bond is:

$$E_{bond}(l) = \frac{k_b}{2}(l - l_0)^2 \tag{4}$$

This potential is just the Taylor-Series expansion of the 'true' bond stretching energy up to order 2 and can therefore only describe bonded interactions when the actual length l is close to the reference length l_0. A function describing the bond stretching energy for a wider range of inter-atomic distances is the Morse potential

$$E_{bond}(l) = D_e(1 - e^{-a(l-l_0)})^2 \tag{5}$$

This expression shows better (exponentially repulsive) behavior for bonded atoms 'too close' (i.e when their inner orbitals overlap) as well as for larger distances, where the bond breaks. However, the Morse potential has two drawbacks: First, there are two adjustable parameters per bond (D_e, a) beside the reference length. Second, the exponential is costly in computational terms. Therefore, the most common force fields make use of expression 4. Generally, treating covalent bonds as completely rigid during energy

minimization (as in JUMNA [69]) or dynamics (constraining them to their reference lengths using , RATTLE [5], LINCS [56] or SHAKE [89]) yields technical advantages while not deteriorating overall accuracy.

Angle Bending
The deviation of angles from their reference value is also usually described by a harmonic potential:

$$E_{ang}(\theta) = \frac{k_\theta}{2}(\theta - \theta_0)^2 \tag{6}$$

Again, the accuracy of the E_{ang} can be improved by taking into account higher order terms and/or terms accounting for the coupling between angle bending and bond stretching, but the force fields most commonly used for modeling proteins and nucleic acids only use equation 6 and, in case of the CHARMM force field, an additional *Urey Bradley* term, i.e. a harmonic term that, technically, is equivalent to a soft covalent bond:

$$E_{bond}(r_{1,3}) = \frac{k_{1,3}}{2}(r_{1,3} - r^0_{1,3})^2 \tag{7}$$

Torsion Angles
Bond stretching and angle bending are 'hard' degrees of freedom, in that substantial forces are needed to cause significant deviations from their reference values. Most of structural variation in molecules stems from changes of the torsion angles and changes in relative energies from torsional and non bonded force-field terms. The torsional potential is most commonly expressed as a cosine series expansion:

$$E_{tors}(\phi) = \sum_{n=0}^{N} \frac{V_n}{2}(1 + cos(n\phi - \gamma) \tag{8}$$

Torsion terms are also commonly used for penalizing out-of-plane bending or for maintaining chirality. These are termed *improper* torsions, since they apply a torsion potential to a quadruple of atoms not necessarily bonded in the sequence 1-2-3-4:

$$E_{improper}(\phi) = \frac{V}{2}(1 - cos2\phi) \tag{9}$$

Technically, equation 9 is only a special case of equation 8.

Non-bonded Interactions
Atoms and molecules also interact through non-bonded forces, i.e. forces that *in principle* [1] act between every pair of atoms. Therefore, whereas the evaluation of the bonded contributions to the intramolecular energy requires $O(N)$ calculations, the evaluation of the non-bonded interactions is an $O(N^2)$ task. As a result, even for a system containing only a few hundred atoms, the calculation of the non-bonded terms accounts for more than 99% of total CPU time.

The Van der Waals interaction contains an attractive and repulsive term. The attractive term stems from the dispersive force between instantaneous dipoles that arise due to fluctuations of electron clouds and can be approximated by a power series expansion with the leading term

$$E(r_{ij}) = -\frac{C_{ij}}{r_{ij}^6} \tag{10}$$

The repulsive term is due to Pauli principle, which prohibits the overlap of closed electron shells, and is best approximated by

$$E(r_{ij}) = C_{ij} \cdot exp(-D_{ij} r_{ij}) \tag{11}$$

The most frequently used function for the Van der Waals interaction of two non-bonded atoms is the Lennard-Jones potential:

$$E(r_{ij}) = 4\epsilon_{ij} \left[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^6 \right] \tag{12}$$

The twelfth power in the repulsive term was originally chosen because it can be calculated from the sixth power by a single multiplication. The parameter ϵ is the 'well depth', σ is derived from the Van der Waals radii of atoms i and j. The exact definition varies between different force fields.

Differences in electronegativity give rise to an unequal distribution of charge in molecules. This unequal charge distribution can be modeled in a number of ways, the most straightforward of which is the assignment of partial atomic charges, located at the nuclear centers. The electrostatic Energy

[1] In some force fields (e.g. AMBER), the non bonded interactions between atoms connected by a covalent bond or through a bond angle are omitted and those between atoms connected through a dihedral angle are scaled by an empirical factor.

for a pair of atoms carrying the partial charges q_i and q_j is given by the Coulomb law:[2]

$$E_{Coul}(r_{ij}) = \frac{q_i \cdot q_j}{\epsilon r_{ij}} \tag{13}$$

Herein, ϵ is the dielectric constant of the medium between the charges. The total potential is given by the sum of all the individual terms and is shown here in the original form and type-setting for the AMBER force field:

$$\begin{aligned} E_{\text{total}} &= \sum_{\text{bonds}} K_r(r - r_{eq})^2 \\ &+ \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 \\ &+ \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \\ &+ \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right] \end{aligned} \tag{14}$$

Though the underlying formulas of equations 4 - 14 are (deceptively) simple, the parameterization is an ongoing process [25, 109]. Also, calculation of the angle dependent terms and their Cartesian gradients has recently received renewed attention, in order to overcome possible numerical instabilities due to coordinate-system dependent singularities [15].

4.3 Electrostatics - A closer Look

The most accurate description of the intra- and intermolecular electrostatic interactions of biomolecules presently available entails the explicit consideration of a large number of solvent molecules surrounding the solute of interest. In this case, equation 13 with $\epsilon = 1$ is used to calculate the total electrostatic Energy. However, despite the availability of sophisticated algorithms based on the Ewald summation method [32,114], that allow the calculation of E_{Coul} within $O(N \cdot \log(N))$ instead of $O(N^2)$ operations, this approach is still costly in computational terms. The more so, since for each solute structure the surrounding solvent must be properly equilibrated before any simulation

[2] In the field of molecular mechanics, charges are scaled so that Coulomb's law as given in equation 13 yields the electrostatic energy in kcal/mol when the distance r_{ij} is measured in Ångstroms.

can commence. To get around this limitation, the damping of intramolecular electrostatic interactions due to solvent permittivity can be modeled by modifying Coulomb's law, making the dielectric 'constant' a function of the inter-charge distance. A relatively successful [72, 119] example for such functions was found by Ramstein et al. [87]:

$$\epsilon(r) = D - \frac{D-1}{2}((rS)^2 + 2rS + 2))\exp(-rS) \tag{15}$$

where D is the bulk dielectric constant and S is a parameter specifying the rate at which $\epsilon(r)$ asymptotically approaches D. Relatively successful means that, combining an appropriate value of S and a scaling of the phosphate charges, experimentally determined nucleic acid structures are structurally close to local minima of the force field with a Coulomb term modified according to equation 15 [72]. However, equation 15 is a crude model, especially when applied to highly charged molecules like nucleic acids, and therefore cannot be used for reliably estimating the relative stability of different conformers.

4.4 Solvent Polarization Effects

Charges inside a cavity surrounded by a dielectric medium polarize the medium, the medium produces a *reaction potential* Φ_{RF} and the charges interact with this field. The 'total electrostatic energy' G_{el} of a solvated molecule is the sum of the Coulomb interactions of its atoms and its free energy of solvation (also termed solvation energy for the sake of brevity).

$$G_{el} = \sum_{j>i} \frac{q_i q_j}{\epsilon\, r_{ij}} + \frac{1}{2}\sum_i q_i \Phi_{\mathrm{RF}}(\mathbf{r}_i) \tag{16}$$

The most accurate method to calculate G_{el} is solving the (finite difference) Poisson (PO) equation

$$\nabla[\epsilon(\mathbf{r})\nabla\Phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}) \tag{17}$$

Once for an exterior dielectric $\epsilon_{\mathrm{ex}} = 1$ corresponding to vacuum, which yields Φ_{vac} and a second time for $\epsilon_{\mathrm{ex}} = \epsilon_{\mathrm{sol}}$, resulting in Φ_{sol}. The difference of these two results is the reaction field, $\Phi_{\mathrm{RF}} = \Phi_{\mathrm{sol}} - \Phi_{\mathrm{vac}}$. Analytic solutions to equation 17 are only known for highly symmetric solute geometries. In

general, the Poisson equation is solved by finite difference (FD) methods [13, 33]. These numerical methods are very demanding in computational terms, even when the derivatives with respect to atomic positions are not calculated.

4.4.1 The Generalized Born Approximation

One promising approach to efficiently calculate the free energy of solvation of molecules that has received much attention in recent years is the generalized Born (GB) approximation. The starting point for this approximation is the solvation energy for a point charge q at the center of a spherical cavity of radius a (a model ion) surrounded by a dielectric medium specified through its dielectric constant ϵ, that was found by Born [18]:

$$\Delta G_{Born} = -\frac{q^2}{2a}\left(1 - \frac{1}{\epsilon}\right) \tag{18}$$

For a model molecule consisting of charges q_i embedded in spheres of radii a_i, with interatomic distances r_{ij} sufficiently large compared to the radii, the solvation energy is,

$$\Delta G_{Sol} = -\sum_{i=1}^{N} \frac{q_i^2}{2a_i}\left(1 - \frac{1}{\epsilon}\right) - \frac{1}{2}\sum_{i,j\neq i}^{N} \frac{q_i q_j}{r_{ij}}\left(1 - \frac{1}{\epsilon}\right) \tag{19}$$

Where the first sum stems form the individual Born terms (eq. 18) and the second sum gives the difference between Coulomb interactions in the solvent and in vacuo. Equation 18 can - for a charge centered inside a spherical cavity - be obtained by integrating over the energy density of the dielectric displacement field outside the cavity, once in vacuum and once in a dielectric medium, and taking the difference.

$$\Delta G_{Sol} = -\frac{q^2}{8\pi}\left(1 - \frac{1}{\epsilon}\right)\int_{\text{Solvent}} \frac{\mathrm{d}V}{r^4} \tag{20}$$

Equation 20 approximately [12] holds if the cavity inside which the charge q_i is located is not of spherical shape [92]. Comparing equations 19 and 20 gives rise to the following definition of the inverse *effective* Born radius of atom i

$$a_{\text{eff i}}^{-1} = \frac{1}{4\pi}\int_{\text{Solvent}} \frac{\mathrm{d}V}{|\mathbf{r} - \mathbf{r}_i|^4} \quad . \tag{21}$$

Put in words, the effective Born radius of an atom is ideally the radius of the sphere at whose center it would have to be placed to yield the same solvation energy as it actually has as part of the non-spherical molecule. Several methods have been developed to efficiently calculate the effective Born radii as continuously differentiable functions of interatomic distances [12, 91, 97]. Provided the effective Born radii are known, Still et al. [97] have introduced the following approximate formula for the solvation energy of a molecule consisting of N atoms with the effective Born radii a_i , separated by the interatomic distances r_{ij}:

$$\Delta G_{Sol} = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{f_{\mathrm{GB}}(r_{ij}, a_i, a_j)} \tag{22}$$

Where

$$\begin{aligned} f_{\mathrm{GB}}(r_{ij}, a_i, a_j) &= \sqrt{(r_{ij}^2 + a_i a_j \exp(-D))}, \\ D &= \frac{r_{ij}^2}{4 a_i a_j} \end{aligned} \tag{23}$$

This choice of f_{GB} yields the original Born expression for $i = j$, a value close to the solvation energy for a dipole inside a spherical cavity if $r_{ij} < \sqrt{a_i a_j}$ as well as the correct damped electrostatic interaction if $r_{ij} > 2.5\sqrt{a_i a_j}$ [97]. The GB approximation is several thousand times faster than FD methods for determining Φ_{RF} and yields not only the approximate reaction field energy, but also its gradients with respect to atomic positions, which has given rise to a considerable amount of implicit solvent molecular dynamics studies in recent years [94, 101, 112, 121]. However, the quality of the approximation heavily depends on large parameter sets that are continually being improved by several groups [35, 63]. These optimization procedures entail the comparison between numerical Poisson calculations and GB calculations for large training sets of molecules. Alternatively, Tsui and Case employed snapshot structures taken from molecular dynamics simulations of nucleic acid molecules as training set [101]. However, it is not yet established whether discrepancies between PO and GB results can be overcome by ever improving parameterization or slight modifications that still lie within the framework of pairwise descreening [80]. With the reaction field contribution included as shown above, the total electrostatic energy, including the reaction field

contribution for a molecule containing N atoms has the following form:

$$E_{elec} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{q_i q_j}{r_{ij}} - \frac{1}{2}(1 - \frac{1}{\epsilon}) \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{q_i q_j}{f_{GB}(r_{ij}, a_i, a_j)} \tag{24}$$

4.4.2 Pairwise Descreening

To calculate the approximate effective Born radii a_i according to equation 21, a volume integral over the *outside* of the molecular structure (ideally the volume outside the Connolly surface that is generally assumed to constitute the dielectric boundary) has to be evaluated for each atom. Hawkins et al. [55] proposed a procedure termed *pairwise descreening* to obtain the effective Born radii as an analytical function of interatomic distances. This works as follows: Without loss of generality placing atom i at the origin, the integral over the outside of the molecule is written as an integral over the outside of atom i minus the integral over the volume of all other atoms j. Let V_i denote the volume of atom i and r_i the radius of atom i, then

$$\int_{\text{Solvent}} \frac{\mathrm{d}V}{r^4} = \int_{r>r_i} \frac{\mathrm{d}V}{r^4} - \sum_{j \neq i}^{N} \int_{V_j \setminus V_i} \frac{\mathrm{d}V}{r^4} \tag{25}$$

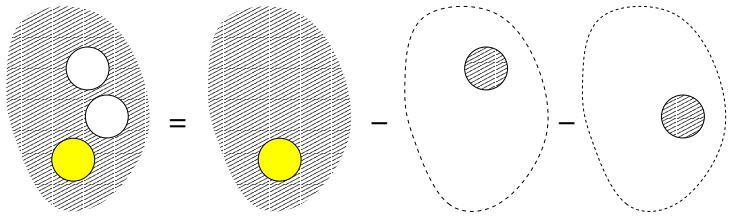


Figure 5: Diagrammatic illustration of equation 25.

With s_j being radius of atom j reduced by an atom-type dependent correction factor to account for the partial overlap of atoms,this leads to the following expression for the inverse effective Born radius of atom i:

$$\frac{1}{a_i} = \frac{1}{r_i - \beta} - \frac{1}{4\,\pi} \sum_{j \neq i} I(d_{ij}; r_i; s_j) \tag{26}$$

With

$$I(d_{ij}; r_i; s_j) = \int_{V_j \setminus V_i} \frac{\mathrm{d}V}{|\mathbf{r} - \mathbf{r}_i|^4} \tag{27}$$

The function $I(d_{ij}; r_i; s_j)$ was originally given by Schaefer and Froemmel [92]. Its derivation is shown in detail in the appendix. Being continuous and differentiable, it allows the calculation of the gradient of the approximate solvation energy. It is important to note that the introduction of the offset β as well as the atom-type dependent scaling of the radii of the "descreening" atoms makes the absolute value of the solvation energy (that should ideally be equal to the corresponding result from finite-difference Poisson calculations) heavily parameter-dependent while leaving the relative energetic ranking of different conformations unchanged and in good agreement with PO results.

While the generalized Born approximation in conjunction with the pairwise descreening method is able to yield solvation energies as well as individual charge-charge interactions in very good agreement with solutions to the PO equation for small molecules, it leads to an underestimation of the effective Born radii of deeply buried atoms inside larger molecules (again, when compared to reference values calculated via solving the PO equation) [80]. This is mainly due to the fact, that the pairwise descreening method consists essentially of an integration over the Van der Waals volume, treating small intramolecular crevices as being filled with the solvent. As can be seen from equations 23 and 24, an underestimation of the effective Born radii leads to too large contributions to the solvation energy and thus to an underestimation of the total electrostatic interaction. The resulting error in the calculation of solvation energies is often acceptable since the atoms contributing most strongly to solvent polarization are the ones at the surface and are treated most accurately [80]. With respect to nucleic acids, this means that the generalized Born approximation is well suited for yielding insights about stem-loop or duplex structures. The observed structural stability of e.g. compact pseudoknots or tRNA structures in implicit solvent MD simulations in the absence of experimentally observed cations might well be due to the aforementioned underestimation of the effective Born radii of deeply buried atoms [79].

Within the framework of the GB approximation, there are two points of attack for trying to improve the agreement with PO-results:

The first one is to improve the performance of the pairwise descreening method, thus better approximating the volume integral from equation 21. The second one would then be to find a better replacement for the effective interaction distance f_{GB} [81].

It has only recently been demonstrated [81], that combining the GB expression (equation 22) in combination with effective Born radii obtained from

solving the PO equation leads to a significant improvement of calculated solvation energies, for nucleic acid structures as well as for proteins. As will become clear below, the so called "diagonal" terms, i.e. the contributions to the solvation energy proportional to the squared individual charges, are, in this case, identical by construction. It is the excellent agreement of the "cross" terms, i.e. the interactions between different charges, which shows that finding a way to calculate effective Born radii in better agreement with the PO approach is a desirable goal.

As stated previously, the effective Born radius of an atom is the radius of the sphere at whose center has it has to be placed to yield its correct self-energy contribution to the solvation energy within the respective molecular structure. Thus, the radius a_i is obtained by calculating the solvation energy for the molecule the atom is part of, setting the charges of all other atoms to zero. Upon having obtained this energy value, the radius is calculated by

$$a_i = -\frac{1}{2}\left(1 - \frac{1}{\epsilon}\right)\frac{q_i^2}{\Delta G_{\mathrm{Sol}\ i}} \tag{28}$$

This procedure is straightforward, but extremely demanding in computational terms, since to obtain the effective Born radii of all N atoms of the structure in consideration, the Poisson equation must be solved N times to get the individual solvation energies. As an example, a single evaluation of the solvation energy of a molecule containing 390 atoms (a RNA tetraloop) takes more than three minutes on the fastest CPU available in our group at the time of this writing. Once having calculated the effective Born radii for a test structure, they can be compared with those obtained by standard the pairwise descreening method, as well as those obtained by converting the Volume integral from equation 21 into a *surface integral* via Gauss' theorem:

$$\frac{1}{a_i} = \frac{1}{4\pi}\int\limits_{\mathrm{ext}} \frac{\mathrm{d}V}{|\mathbf{r} - \mathbf{r}_i|^4} = -\oint\limits_{O(V)} \frac{(\mathbf{r} - \mathbf{r_i}) \cdot \mathrm{d}\mathbf{f}}{|\mathbf{r} - \mathbf{r}_i|^4} \tag{29}$$

For the comparison shown below, the surface integral over the molecular surface of the test structure was approximated using a triangulation. This triangulation was obtained from the program `surf`, which is part of the `VMD` molecular visualization program [62]. The number of triangles was increased until convergence of the resulting surface area. Then, with $\mathbf{c_i}$ denoting the

center of triangle i, A_i denoting its area and $\hat{\mathbf{n}}_i$ denoting its normalized normal vector pointing outside, the approximate inverse effective radius $\frac{1}{a_i}$ is obtained as follows:

$$\frac{1}{a_i} \approx \frac{1}{4\pi} \sum_{\text{triangles}} \frac{(\mathbf{c}_i - \mathbf{r}_i)\hat{\mathbf{n}}_i A_i}{|\mathbf{c}_i - \mathbf{r}_i|^4} \tag{30}$$

Figure 6 shows a plot of approximate effective Born radii over their ideal reference values obtained from the numerical solution of the Poisson equation. The ideal case (perfect agreement) is indicated by a dotted line. The dashed curve shows an ideal fit obtained by physical intuition and and the `xmgrace` plotting program. The inverse of the fitting function can be used to obtain effective radii in very good agreement with the reference results from those calculated via equation 30. The effective radii obtained via the pairwise descreening method are shown in black, those obtained from the surface integral in blue and those calculated by applying the inverse fitting function in red. The standard pairwise descreening method shows very good agreement between approximate and optimal radii as long as they are small,

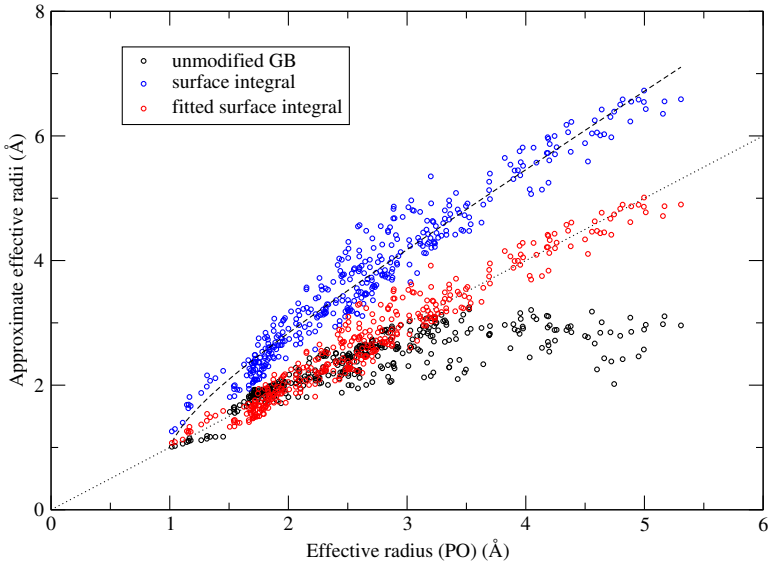


Figure 6: Approximate over optimal effective Born radii for a RNA tetraloop structure. The radii obtained via pairwise descreening are shown in black, those obtained directly via the surface integral in blue and those obtained by a heuristic fit in red.

corresponding the atoms lying at the molecular surface. Since those atoms carry the largest partial charges, this explains the success of the generalized Born approximation when applied to molecules with small interior regions. The radii of atoms not in contact with the molecular surface are underestimated, which makes the GB approximation in its standard form less suited for molecules containing larger interior regions, be it proteins or larger nucleic acid molecules [12, 80, 81]. The under-estimation of the effective radii for buried atoms leads to an over-estimation of their solvation energies and, as can be inferred from equation 24, to an underestimation of their total electrostatic interaction. Calculating the effective radii via the surface integral leads to a systematic over-estimation. This is due to the Coulomb approximation: Equation 20 is only exact for a point charge located at the center of a spherical cavity. For point charges in off-center positions, the volume integral leads to a systematic underestimation of the solvation energy. Therefore (cf. equation 28) the corresponding radii are too large. However, the deviation is systematic and allows the calculation of 'nearly perfect' effective radii via a heuristic fit.

The purpose of the foregoing discussion is to illustrate physical background and limitations of the generalized Born approximation at the current state of development. As long as no computationally efficient analytical representation of the molecular surface is available, a combination of discretized surface integral and equation 19 is pretty cool, but can at best only be used for single point energy evaluations. Therefore, most room for improvement seems to lie with further developing the pairwise descreening method.

4.5 Structure Optimization

Given a molecular structure $\mathbf{R} = (x_1, y_1, z_1, \ldots, x_N, y_N, z_N)^{\mathrm{T}} = (\zeta_1, \ldots, \zeta_{3N})^{\mathrm{T}}$ the numerical value of $E_{pot}(\mathbf{R})$ has no meaning by itself. The foremost requirement for *stable* structures $\mathbf{R_0}$ is that $\mathbf{R_0}$ corresponds to a local minimum of E_{pot}, requiring that the gradient $\nabla E_{pot}(\mathbf{R_0})$ vanishes and that the Hessian matrix

$$H_{ij}(\mathbf{R}) = \frac{\mathrm{d}^2 E_{pot}}{\mathrm{d}\zeta_i \mathrm{d}\zeta_j}$$

is positive (semi-)definite if $\mathbf{R} = \mathbf{R_0}$. Due to numerical limitations, it is impossible to exactly reach a local minimum. In practice, local minimum refers to a point on the energy surface, where the applied minimization procedure cannot further reduce the function value. Depending on the quality of

the minimization method and the number if degrees of freedom, this is commonly the case when the root mean square gradient (g_{rms}) is in the range of 10^{-5} to 10^{-6} kcal/(mol Å) for *conjugate gradient* methods. *Newton Raphson* methods, that employ the Hessian matrix to obtain optimal downhill search directions can further reduce the function value until $g_{rms} \lessapprox 10^{-9}$ to 10^{-11} kcal/(mol Å). Such stringent minimization, however, is only necessary if the detailed shape of the energy landscape in the close vicinity of a local minimum is of interest, e.g. for *normal mode* calculations.

If the reduction of the function fails when g_{rms} is substantially larger than the approximate values given for conjugate gradient methods, this commonly points to either a function that is not everywhere differentiable, or worse, to an error in the calculation of the gradient. It should also be noted that minimization methods based on the Hessian matrix are very costly in computational terms as they require storing the matrix as well as its inversion[3]. Additional computational cost arises from the fact that the analytical computation of the Hessian matrix is impractical (though possible in principle) for some parts of the force field energy, most notably for the solvation energy.

The energy landscape defined by a force field is so rough, that the energy value at convergence of gradient based (downhill) minimization methods still only allows a crude estimate of energetic ranking. By numerical experiments we found that minimization starting from the same structure but using different minimization methods can converge at different local minima. The roughness of the energy landscape and the limited significance of energy values obtained from evaluating the force field energies is illustrated by the following example:

Minimization of 1000 consecutive snapshots taken every picosecond from a molecular dynamics simulation of a RNA tetraloop (390 atoms) leads to convergence in 671 different stationary points. These are not necessarily local minima, but can in principle also be saddle points. This is, however, highly improbable, since the conjugate gradient minimization protocol used throughout this work entails three restarts of the minimization upon first reaching convergence. The probability that the disturbance induced by restarting the minimization still leads to the same saddle point is minimal [95]. Principal component analysis of the minimized structures (see section 5) shows that there are -in a coarse grained sense and for all practical

[3]The Hessian matrix for a force field energy is actually singular, so additional steps have to be taken in order to properly remove the six "external" degrees of freedom

purposes- three or four essentially different minimum structures. This is in sharp contrast to 671 energy values obtained upon minimization, differing by up to 10 kcal/mol.

Each molecular dynamics snapshot corresponds to a point 'somewhere uphill' the energy landscape, since each internal degree of freedom is on the average excited by $\frac{kT}{2}$. The landscape is so rough that most downhill paths found by the minimization routine end up in local minima above the closest *relevant* minimum. Figure 7 shows the first principal component of each consecutive snapshot and the distribution of force field energies obtained from the minimized structures. Inset in the lower graph, a rough one dimensional energy function containing several irrelevant and one 'coarse grained' relevant minimum is shown as illustration. The example given above shows

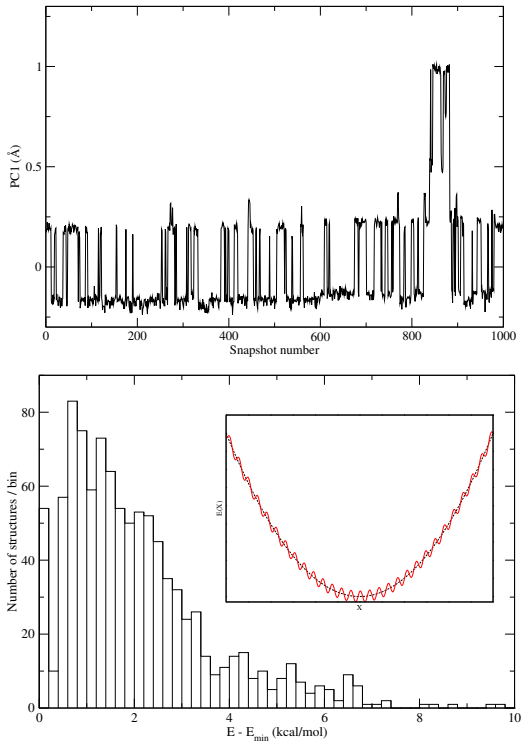


Figure 7: First principal component of 1000 consecutive minimized MD snapshots (upper graph) and distribution of force field energies at convergence of minimization (lower graph). Inset in the lower graph a qualitatively similar one dimensional energy function is shown.

that a reliable energetic ranking of different conformers based on 'downhill only' minimization methods is not possible, since the possible error due to convergence in an irrelevant local minimum is significantly larger than kT at physiological temperature. The situation is different when the molecular conformation is represented in internal coordinates, as for example in the JUMNA program. The significantly reduced number of internal degrees of freedom in internal coordinate representation leads to a smoothing of the energy landscape and thus to a much smaller number of local minima. Minimization of the 671 different minimum structures using the JUMNA program leads to only fifteen different minimum structures. On the other hand, the AMBER force field (on which JUMNA is also based) is parameterized for the use of Cartesian coordinates and it is hard (or impossible) to reliably estimate the error due to the reduction of the number of degrees of freedom. This shows that even if a fictitious perfect force field were available, the energies obtained after minimization are to be interpreted with a grain of salt. Is is also obvious that any minimization procedure in Cartesian coordinates *must* contain a combination of 'uphill' moves and gradient based methods even if the purpose is only finding a relevant local minimum. Force field energies given within this work were obtained by a procedure we term 'cautious' molecular dynamics based simulated tempering, consisting of repeated heating the trial structure to 300 Kelvin, slow cooling by viscous drag (i.e. contact to a Langevin heat bath of zero temperature) and subsequent conjugate gradient minimization until no further reduction of the energy value could be observed.

4.6 Molecular Dynamics

Biomolecules at physiological temperature are far from static due to their thermal motion. While the potential energy surface itself yields some information about local differences in flexibility via the normal mode analysis [115], more detailed insight can only be gained by simulating the temporal evolution of molecular structures. This technique is referred to as molecular dynamics (MD). In molecular dynamics, the temporal evolution of a system is simulated by integrating Newton's second law of motion:

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i \tag{31}$$

Where m_i is the mass of atom i, $\mathbf{r}_i$ its position and $\mathbf{F}_i$ is the total force acting on atom i. The force is related to the potential energy by

$$\mathbf{F}_i = -\nabla_{\mathbf{r}_i} E_{pot} \tag{32}$$

Given the positions and velocities of all atoms, as well as the total force acting on each atom, the equations of motion can be integrated numerically, applying finite difference methods. These methods are based on a Taylor series expansion of the coordinates:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{(\delta t)^2}{2}\mathbf{a}(t) + \frac{(\delta t)^3}{6}\mathbf{b}(t) + O((\delta t)^4) \tag{33}$$

The most popular integration method in molecular dynamics is the Verlet algorithm [108], that is simple to implement, numerically stable and of modest memory requirements. The basic iteration follows directly from equation 33:

$$\mathbf{r}(t+\delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + \mathbf{a}(t)(\delta t)^2 + O((\delta t)^4) \tag{34}$$

A numerically equivalent integration scheme that directly yields coordinates and velocities at the cost of additional memory requirement is the Velocity Verlet algorithm:

$$\begin{aligned}
\mathbf{r}(t+\delta t) &= \mathbf{r}(t) + \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)(\delta t)^2 \\
\mathbf{v}(t+\frac{\delta t}{2}) &= \mathbf{v}(t) + \frac{1}{2}\mathbf{a}(t)\delta t \\
\mathbf{a}(t+\delta t) &= -\frac{1}{m}\nabla V(\mathbf{r}(t+\delta t)) \\
\mathbf{v}(t+\delta t) &= \mathbf{v}(t+\frac{\delta t}{2}) + \frac{1}{2}\mathbf{a}(t+\delta t)
\end{aligned} \tag{35}$$

This iteration is self starting, i.e. the integration of the equations of motion can be started at time t_0 without backwards extrapolation. Furthermore it yields positions and velocities at the same time. This in turn allows the accurate computation of the total energy. Another advantage of the velocity Verlet algorithm is that it can easily be combined with restraints (e.g. on bond lengths) via the RATTLE algorithm and that the extension to *stochastic dynamics* is straightforward [105, 106].

The total energy of any system for which equations 31 and 32 hold is theoretically exactly conserved. The numerical integration of the equations of motion introduces two artifacts, which are due to the finite time step: First, there ar e short-time fluctuations, the size of which depends on the time step and on the integration algorithm. Second, there is the so called drift, the slow growth or decline of total energy over longer periods of time. While the Verlet algorithm leads to larger fluctuations than more sophisticated predictor-corrector algorithms [2], it does in itself not lead to drifts. For a harmonic oscillator, the fluctuations are themselves harmonic, which is shown in the appendix.

The temperature of a system comprised of N atoms subject to N_c constraints is related to its kinetic energy by the following equation, that follows directly from the *equipartition theorem*:

$$E_{Kin}(t) = \sum_{i=1}^{N} \frac{m_i \mathbf{v}(t)_i^2}{2} = \frac{kT(t)}{2}(3N - N_c) \tag{36}$$

Kinetic energy and temperature are time dependent quantities and only their time averages are relevant to the analysis of a molecular dynamics simulation.

For implicit solvent molecular dynamics, if the initial structure corresponds to well defined local minimum of the fore field, the fastest way assign a desired temperature to a molecule is to sample the initial atomic velocities from a Maxwell-Boltzmann distribution at *twice* the target temperature. Owing to the equipartition theorem, this leads to a redistribution of the initially available kinetic energy among all available degrees of freedom and an average temperature close to the desired target value. In practice, the initial atomic velocities are commonly sampled from a Maxwell-Boltzmann distribution at less than the desired temperature and the system is subsequently heated through the contact to a *heat bath*.

4.7 Heat Baths

A common way to achieve or maintain a desired average temperature, that is standard in the AMBER [21] program suite and NAB [71], is the Berendsen method [16]. In this method, the rate of energy exchange between the heat bath and the simulated system is proportional to their temperature difference, leading to a rescaling of all atomic velocities at every time step by the

following factor:

$$\lambda(t) = \sqrt{1 + \frac{\delta t}{\tau}\left(\frac{T_0}{T(t)} - 1\right)} \tag{37}$$

where τ is a coupling constant, regulating the strength of the thermal coupling. The advantage of the Berendsen method lies in its simplicity and computational efficiency. On the other hand, trajectories calculated with the a heatbath of the Berendsen type do not sample from any well defined ensemble. Furthermore, velocity rescaling methods violate energy equipartition, redistributing kinetic energy from high-frequency to low-frequency motions. If center of mass motion is not restrained, this in turn leads to a slow but steady increase in center of mass velocity and overall angular momentum at the cost of internal excitation. Harvey et al. descriptively call this phenomenon the 'flying icecube' [52].

A different approach to temperature regulation was proposed by Nosè [78] and extended to the following form by Hoover [59]. Newton's equations of motion 31 are augmented by an additional degree of freedom ζ, so that the time dependent behavior of the system system is described by the following equations of motion:

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i - \zeta m_i \mathbf{v}_i \tag{38}$$

$$\frac{d\zeta}{dt} = \frac{f k_B}{Q}(T_0 - T) \tag{39}$$

Herein, f is the number of degrees of freedom and Q is a parameter regulating the strength of the coupling between the heat bath and the simulated system. A system governed by equation 38 samples from a canonical ensemble [59].

The third approach, that also leads to sampling from a canonical ensemble [106], is replacing equation 31 by a Langevin equation, representing the heat bath as a viscous medium at the desired temperature. In this case, the equations of motion are given by:

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i - m_i \gamma_i \mathbf{v}_i + \mathbf{R}_i \tag{40}$$

With γ_i a friction constant and $\mathbf{R}_i$ a random force with zero autocorrelation time, specified through its autocorrelation function [24]:

$$\langle \mathbf{R}_i(0) \cdot \mathbf{R}_j(\tau) \rangle = 6\, m_i\, k_B\, T_0\, \gamma_i\, \delta_{ij} \delta(\tau) \tag{41}$$

Equation 41 is based on the *assumption* that the random force on particle i, $\mathbf{R}_i(t)$ is independent from its values at previous times as well as from the random forces acting on other particles. Throughout this work, a discretization of equation 40 given in [105] and [106] respectively, that has originally been implemented in the TINKER molecular modeling package has been used for integrating stochastic equations of motion.

It should be noted that the Langevin approach to simulating molecular motion provides a readily available physical explanation for the time dependent behavior of the simulated system in dependence of the assumed friction constant(s). In contrast to the two other thermostats briefly described here, Temperature enters the equations of motion only via the variance of the random forces. It can, however, only be a crude approximation at best to the interaction (i.e. power transfer) between solute and surrounding solvent in simulations where the latter is considered explicitly. The friction-constant best reproducing solute-solvent power transfer is about $100\,\mathrm{ps}^{-1}$ [31], scaled by the solvent accessible surface area of each atom. This leads to a 'diffusion like' overall molecular motion, slowing down conformational changes as well as to significant computational overhead due to re-calculating individual friction constants at every (or every few) timestep(s). Therefore, unless explicitly stated otherwise, a small uniform friction constant ($\gamma < 1\ \mathrm{ps}^{-1}$)[4] for all atoms was used in the molecular dynamics calculations throughout this work. The use of small a small friction constant leads to 'enhanced sampling' [31, 111, 112], i.e. a faster exploration of accessible conformations from a start structure at a given temperature, and has even been proposed as a means of preventing systematic long-time drifts of total energy in conjunction with multiple timestep methods [11]. Furthermore, equilibrium properties are independent of the friction constant.

4.8 Constraints in Dynamics and Minimization

The time step in molecular dynamics simulations is limited by the highest frequency motion present in the system. The upper limit for systems containing hydrogen atoms bonded to heavy atoms is generally *assumed* to be one femtosecond (10^{-15} seconds). To overcome this limit and thereby increase computational efficiency without noticeable loss in accuracy, algorithms for constraining bonds to their reference values (i.e. treating them as rigid)

[4] Exact values are given individually.

have been developed. The most commonly used procedures are SHAKE [89] or RATTLE [5]. The choice of the method for imposing the constraints is dictated by the integration method in use. Integrators where velocities do not explicitly occur in the integration can be combined with the SHAKE procedure, in which only position constraints are imposed. When the integration explicitly involves velocities, it can be combined with the RATTLE procedure, where velocities are also constrained.

The constraints to be satisfied for two atoms i, j at positions $\mathbf{r}_i$ and $\mathbf{r}_j$ connected through a bond of reference length d_{ij} are given by:

$$(\mathbf{r}_i - \mathbf{r}_j)^2 - d_{ij}^2 = 0 \tag{42}$$

and

$$(\mathbf{r}_i - \mathbf{r}_j) \cdot (\mathbf{v}_i - \mathbf{v}_j) = 0. \tag{43}$$

In practice, the constraint satisfaction procedures work iteratively, applying corrections to the positions or velocities of each constrained atom pair until all constraint violations are below a given tolerance. For a single atom pair i, j , the correction applied at each step can be written in the following form [5, 89] for the positions and similarly for the velocities:

$$\begin{aligned} \mathbf{r}_i &= \mathbf{r}_i + \mathbf{r}_{ij} m_i^{-1} f(m_i, m_j, \mathbf{r}_i, \mathbf{r}_j) \\ \mathbf{r}_j &= \mathbf{r}_j - \mathbf{r}_{ij} m_j^{-1} f(m_i, m_j, \mathbf{r}_i, \mathbf{r}_j) \end{aligned} \tag{44}$$

The above equations suggest that the number of iterations necessary to satisfy all constraints to a given tolerance should increase if the function $f(m_i, m_j, \mathbf{r}_i, \mathbf{r}_j)$ is multiplied by a factor $\lambda \in (0, 1)$. This assumption was verified by tests on different molecular structures (data not shown). Of more practical value was the subsequent finding, that by slightly exaggerating the iterative correction, i.e. choosing a factor $\lambda > 1$, the number of iterations until convergence can be reduced by (roughly) 50 percent! Numerical experiments showed that the optimal value of λ is structure dependent. However, by choosing $\lambda = 1.25$ a significant speed increase for all structure tested could be obtained. While this was found in the process of adopting the RATTLE routine contained in the TINKER package to fit into the molecular dynamics routines within NAB, we became aware that Barth et al. [10] proposed the same method for speeding up the convergence of the SHAKE iteration.
Upon our notice, this simple but efficient trick was integrated into the official TINKER distribution (version 3.9).

If bond lengths are to be constrained during a molecular dynamics simulation, these constraints should also be imposed during the preceding energy minimization. Duan et al. [36] showed a procedure to iteratively correct the gradient in a way similar to equations 44 so as to satisfy imposed distance constraints. While no details of implementation are given in the reference cited above, we found the following procedure to allow constrained (conjugate gradient) minimization with satisfactory convergence:

- Call the RATTLE or SHAKE routine to reset atomic coordinates to fulfill the imposed constraints within the required tolerance.
- Call the force field routine yielding the energy and its gradient for the constrained structure.
- Correct the gradient as shown by Duan et al.

A favorable side effect of constraining *all* bond lengths during minimization is that, starting from the same (unminimized) structure, the number of minimization steps needed until convergence[5] is significantly reduced compared to minimization without constraints, without changing the relative energetic ranking of minimized conformers. There are two possible explanations for this behavior:

Either the roughness of the constrained energy landscape leads to artificial stationary points where the minimization routine converges, or the acceleration is due to the significant reduction in the number of degrees of freedom. While we are not aware of a way to prove that either explanation is true, numerical experiments (minimization of several molecular dynamics snapshots with as well as without constraints - not shown), indicate that for all practical purposes the second explanation can be accepted.

4.9 A Note on Potential Truncation

When simulating large molecular systems, it is desirable to truncate long range interactions in order to increase computational efficiency unless methods based on the Ewald summation [32] are available. The most straightforward way to implement a cutoff is to simply ignore all non-bonded interactions for atom pairs whose distance r_{ij} is greater than the cutoff. However,

[5]By convergence, we mean that the root mean square gradient is below a required threshold.

while this is sufficient for the Lennard-Jones potential (which drops off like r_{ij}^{-6}), it leads to severe problems when applied to the Coulomb interaction. During molecular dynamics simulations, all interatomic *forces* should be continuous and differentiable ("smooth") functions of interatomic distance for *all* values of r_{ij} [96]. Leach [70] suggests multiplying the pairwise potential $V(r_{ij})$ with a switching function $S(r_{ij})$. By choosing an appropriate switching function $S(r_{ij})$, the switched potential function $V(r_{ij}) \cdot S(r_{ij})$ can be made to obey any requirements with respect to differentiability. However, applying a switching function implies changing the potential over a relatively short distance, leading to force artifacts that have been shown to lead to readily noticeable errors in MD calculations [9]. With this in mind, we devised a *force derivative shifted* Coulomb potential. Let r be the distance between two charges q_i and q_j and let c be the cutoff distance. Then

$$V_{fds}(r) = \begin{cases} q_i q_j \left(\frac{1}{r} + \left(\frac{9}{8}\right) \frac{r^{16}}{c^{17}} - \frac{r^{17}}{c^{18}} - \left(\frac{9}{8}\right) \frac{1}{c} \right), & r < c \\ 0, & r \geq c \end{cases} \tag{45}$$

This modified potential as well as its first and second derivative with respect to the distance r_{ij} vanish at the cutoff distance, satisfying the numerical requirements of standard integration algorithms. A comparison between switched and force derivative shifted Coulomb potential is shown in figure 8 for the Coulomb energy (upper graph) and force (lower graph) of a pair of monovalent cations as function of their distance. In this example, the switching region spans the distance range between 13 and 17 Å, the cutoff distance equals 17 Å for both potentials. The relatively sudden change of the switched potential function between 13 and 17 Å leads to a strong amplification of the coulomb force inside in this region. The force due to the potential defined in equation 45 shows no artifacts (except the underestimation of the true Coulomb interaction with increasing distance). A similar *force shifted* Coulomb potential was proposed by Steinbach et al. [96]. All long-range modifications investigated by Steinbach et al. lead to a cusp in the Coulomb force at the cutoff distance, which in turn leads to drifts in total energy. The modification given above leads to conservation of total energy indiscernible from simulations with unmodified non-bonded potentials. The large difference in the absolute values of the modified coulomb energy visible in figure 8 does not affect system behavior during MD simulations.

However, it should never be forgotten that *any* cutoff scheme, no matter how sophisticated, is unphysical in principle. For the calculations where V_{fds}

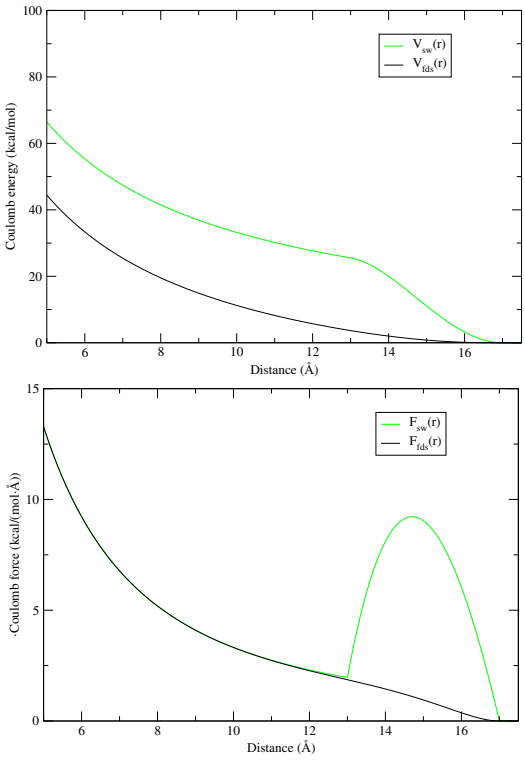


Figure 8: Comparison of switched (gray) and force derivative shifted (black) Coulomb potential (upper graph) and the resulting forces (lower graph).

was used in this work, a cutoff of 25 Å was chosen as the default, leading to a distance dependent underestimation of the true Coulomb *forces* as shown in the following table:

Distance (Å)	Deviation (%)
15	0.2
17.5	1.5
20.0	10.0
22.5	45.0
25.0	100.0

4.10 Droplet Dynamics

The choice of a force field is often conditioned by the use of the associated molecular mechanics software [9]. So is the choice of the simulation methods. To complement the implicit solvent molecular dynamics simulations performed throughout this work, a non bonded force field routine allowing the inclusion of explicit solvent (TIP3P waters and neutralizing counter ions) was implemented. For truncating long range electrostatic interactions, the force derivative shifted Coulomb potential described previously was used with a default cutoff of 25 Ångstroms. The Lennard Jones potential was shifted to zero at the same distance. To prevent evaporation at the solvent/vacuum boundary, a semi-harmonic maximum pairwise distance restraint was used. With $d_{\max}$ denoting the maximal allowed pairwise distance, the restraint potential has the following form:

$$V_{\text{rest}}(r_{ij}) = \begin{cases} w(r_{ij} - d_{\max})^2, & r_{ij} > d_{max} \\ 0, & r_{ij} \leq d_{\max} \end{cases}$$

This is clearly unphysical, but only affects isolated atom pairs at the edge of the solvent droplet. Furthermore it has the advantage of being translation invariant and does not constitute a rigid boundary. The restraint constant w can be chosen arbitrarily. Droplet dynamics simulations are limited to small solute molecules (e.g. RNA tetraloops or short helices) since to minimize boundary artifacts, the distance between any solute atom and the droplet boundary must be several Ångstroms and a large cutoff (in conjunction with proper truncation) is necessary. A second limitation is that the initial shape of the solvent droplet must be approximately spherical. It was found that non-spherical droplets tend to assume spherical shape, deforming solute molecules. Despite several potential problems due to the finite system size, the 'droplet dynamics' simulations performed throughout this work show very good agreement with results reported from simulations based on Ewald summation techniques and periodic systems (see section 8.1.3 and references there).

5 Principal Component Analysis of molecular Structures

A molecular structure is represented by N points in three-dimensional space or -equivalently- one point in $3N$-dimensional space. Given a set of S structures defined as points $\mathbf{R}_i$, with $\langle \mathbf{R} \rangle$ denoting their average, we define the corresponding covariance matrix by

$$\mathbf{C}_{ij} = \frac{1}{N} \left(\mathbf{R}_i - \langle \mathbf{R} \rangle \right) \cdot \left(\mathbf{R}_j - \langle \mathbf{R} \rangle \right). \tag{46}$$

Upon diagonalization, the matrix $\mathbf{C}$ can be written in the following form:

$$\mathbf{C} = \left(\mathbf{X}\sqrt{\mathbf{\Lambda}} \right) \left(\mathbf{X}\sqrt{\mathbf{\Lambda}} \right)^T \tag{47}$$

Here, $\mathbf{\Lambda}$ is is a diagonal matrix, containing the eigenvalues of $\mathbf{C}$ sorted in descending order and $\mathbf{X}$ is an orthonormal matrix, containing the corresponding eigenvectors of $\mathbf{C}$, termed principal axes, as columns. The principal axes span a subspace of $\mathbb{R}^{3N}$ such that the structural variances along these axes are sorted according to their contribution to the total variance. The matrix $\mathbf{X}\sqrt{\mathbf{\Lambda}}$ contains the projection of the vectors $\mathbf{R}_i$ onto the eigenvectors of $\mathbf{C}$, termed *principal components*. Definition 46 is different from the one given by Amadei et al. [3, 4]:

$$\mathbf{C}^{\text{conv}}_{ij} = \frac{1}{S} \sum_{k=1}^{S} \left(x_i - \langle x_i \rangle \right) \cdot \left(x_j - \langle x_j \rangle \right) \tag{48}$$

where the x_i are individual atomic coordinates. Both definitions yield (up to a constant factor due to different normalization) the same spectrum of (positive) eigenvalues: Let M be a $3N \times S$ matrix containing the vectors $(\mathbf{R}_i - \langle \mathbf{R} \rangle)$ as columns. Then eigenvectors $\mathbf{x}$ and eigenvalues λ fulfill

$$\mathbf{M}^T\mathbf{M}\mathbf{x} = \lambda \mathbf{x}. \tag{49}$$

Premultiplication with $\mathbf{M}$ leads to

$$(\mathbf{M}\mathbf{M}^T)(\mathbf{M}\mathbf{x}) = \lambda(\mathbf{M}\mathbf{x}). \tag{50}$$

Therefore $\mathbf{M}^T\mathbf{M}$ and $\mathbf{M}\mathbf{M}^T$, proportional to $\mathbf{C}$ and $\mathbf{C}^{\text{conv}}$ respectively, have the same spectrum of positive eigenvalues and the eigenvectors of $\mathbf{M}\mathbf{M}^T$

can, if necessary, be obtained from those of $\mathbf{M}^T\mathbf{M}$ by a simple matrix-vector multiplication. The normalization of the entries of matrix $\mathbf{C}$ is chosen so that the sum of the squared principal components describing an individual structure $\mathbf{R}_i$ is equal to the mean square distance between this structure and the average structure $\langle\mathbf{R}\rangle$.

As long as the number of structures S is smaller than the number of individual coordinates $3N$, our definition is of advantage, since the diagonalization of the covariance matrices is the computationally most demanding part of the calculation and $\mathbf{C}$ is a $S \times S$ matrix, whereas $\mathbf{C}^{\text{conv}}$ is of size $3N \times 3N$. In order to retain only the relevant, i.e. internal motions in the analysis, all structures are in the first step superimposed onto a reference structure, commonly the first one within the analyzed trajectory.

PCA proved to be a helpful method for locating and coarsely visualizing structural transitions during MD simulations. As an example, figure 9 shows the (purely hypothetical) structural transitions during an MD run of a polyalanine, starting from an extended α-helix, as reflected by the first principal component over the time course from one to ten nanoseconds.

The first step in applying principal component analysis to a set of structures consists of superimposing all structures onto a common reference structure in order to remove the contribution of overall translation and rotation. Therefore, the superposition program `suppose` that is part of the NAB program package was used as the basis for PCA. For matrix diagonalization, the corresponding routines from the `ccmath`-library [8] were used. Our implementation also allows choosing only a subset of atoms by their names (following the NAB naming convention), which has proved useful when investigating localized structural transitions.

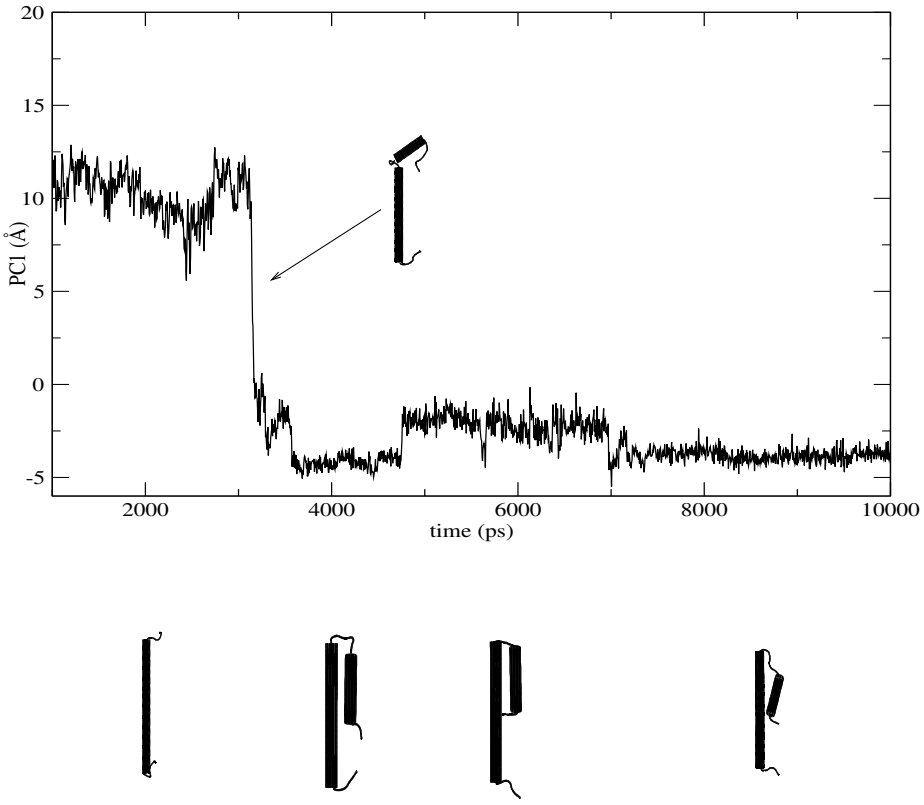


Figure 9: Large scale transitions of a hypothetical polyalanine structure as represented by the first principal component. α-Helical regions are depicted as cylinders. The structures shown are snapshots taken at times corresponding to their position. The inset snapshot corresponds to the middle of the sharp drop of PC1 around $t \approx 3000$ps.

6 Conformational Search Methods

Even for restricted systems like tetraloops, the number of sterically allowed structures is by far too large to be sampled exhaustively, despite the currently available computational resources. Conformational search for low energy GNRA tetraloop conformers based on distance geometry yielded the first indication that loop conformers of distinctly lower energy than minimized experimental structures can be found, but at prohibitive computational cost (see section 8.1.4). The lowest energy conformers obtained from distance geometry based conformational search had to be treated by simulated tempering to arrive at the first 'false optimal' conformer.

Conformational search methods based on molecular dynamics are a slow but safe way to find low energy structures. Since the energetic barriers between different conformers (in a coarse grained sense) are generally too high to be crossed within an acceptable time scale at physiological temperature, there are -in principle- two ways of facilitating barrier crossing.

Barrier crossing can be facilitated by repeated heating and cooling (simulated tempering). In this approach, the choice of of the temperature of the different heat baths as well as the number of time steps of 'hot' and 'cold' phase must be chosen based on experience and educated guess. A general rule is that, with increasing temperature of the 'hot' heat bath, the time of contact with this heat bath must become shorter in order not to obtain extended structures, that are entropically favored, but do not have a favorable force field energy. One successful combination of parameters was: A time step of 2 femtoseconds (with hydrogen masses set to 5 atomic mass units to maintain numerical stability). 700 time steps in contact with a Langevin heat bath of 700 Kelvin alternating with 1500 timesteps in contact with a Langevin heat bath at 300 Kelvin, both heat baths with friction constants of 1 ps^{-1} These values are meant as examples. Probably other combinations of values could be found that show a higher rate of favorable transitions.

The second possibility is to modify the potential energy, assigning an energetic penalty to visited structures to accelerate conformational transitions (*Local Elevation*)[6].

[6]The term 'Local Elevation' was coined by Huber et al. [61] for a slightly different method, but also fits the method described here.

To achieve this, the following scheme was devised: A molecular dynamics run is started and after N timesteps, the *current* (sub-)structure is assigned a rmsd based energetic penalty, arbitrarily chosen to be of Gaussian shape. With $\{\mathbf{c}_i\}$ denoting the positions of the N_c restrained atoms, λ denoting the 'inverse squared width' of the Gaussian penalty and $\{\mathbf{c}_{0\,i}\}$ denoting the currently penalized positions, The restraint potential has the following form:

$$\begin{aligned} V_p(\{\mathbf{c}_i\}) &= V_0 \exp(-\lambda\, d^2), \\ d^2 &= \frac{1}{N_c} \sum_{i \in N_c} (\mathbf{c}_i - \mathbf{c}_{0\,i})^2 \end{aligned} \tag{51}$$

The derivative of the restraint potential with respect to an individual coordinate (the negative restraint force component) shows the effect of such a potential:

$$\frac{\partial V_p}{\partial x_j} = -2\lambda V_p (x_j - x_{0\,j}) \tag{52}$$

The restraint potential results in a repulsive force, driving each restrained atom away from its reference position. The amount of restraint force depends on the overall structural change. Such a potential can be interpreted as a mountain on the energy landscape. After N time steps allowing relaxation and equilibration (i.e. absorption of the excess kinetic energy gained by moving away from the mountain), the structure is stored on disk for minimization and energetic evaluation, and the restraint is shifted from the previous to the current structure.

Since it was found that this procedure often leads to repeated back and forth switching between two distinct structures, it was augmented by a memory function: The last M structures are assigned a penalty of the functional form shown in equation 51. After the generation of $M + 1$ structures, the first penalized structure in memory is replaced by structure $M + 1$ and so on. This can be interpreted as placing a mountain range consisting of M mountains on the energy landscape.

These method again depends on a large number of parameters, number, height and width of the repulsive mountains, the number of relaxation timesteps, temperature, and the choice of the restrained atoms. Furthermore, the most direct way for a molecular structures to relax the restraints is overall rotation or translation. Therefore, the local elevation methods require constraining a subset of atoms to fixed positions. The local elevation

methods were applied to RNA loop structures containing a tetraloop and a four base pair stem. In this case, residues 1 to 3 and 9 to 12 were constrained to their initial positions.

The conformational search procedures described here reproducably led to loop conformations of lower energy than any minimized experimental reference structure, but only once - and therefore irreproducably - to a structure in close agreement with the experimental reference structure (see section 8.1.4). Since the occurrence of low energy conformers constitutes a rare event in all methods described above, it is impossible to tell which one is the most efficient.

7 Program Packages

7.1 Nucleic Acid Builder

Nucleic Acid Builder (NAB) [71] is a freely available programming language developed specifically for molecular modeling tasks involving macromolecules. The most prominent features of NAB are:

> A subset of the `C` programming language.
>
> A unique form of regular expressions termed 'atom regular expressions', allowing for a convenient way to select arbitrary parts of molecules.
>
> An interface to distance geometry routines, covering every step from building up distance matrices to embedding them in tree dimensional space.
>
> An implementation of the AMBER force field including the generalized Born approximation, as well as a minimization routine and a molecular dynamics routine.
>
> Compatibility with programs and file formats contained in the commercially available AMBER molecular modeling program suite.

The main advantage of NAB is not the provided routines, but its extensibility. The core force field routines contained in NAB are written in the `C` programming language. Extensions, such as for example new types of restraint potentials, advanced modifications of the generalized Born approximation or alternative molecular dynamics routines can be readily built into the existing program package. Time critical additional functions or functions requiring parts of the `C` programming language not within the range of NAB can be linked with NAB programs.

7.1.1 Additions to NAB

The most useful additions to NAB developed within this work are:
• A flag-toggled replacement of harmonic position restraints by 'funnel shaped' (tightly tipped hyperbolic) restraint potentials. Funnel shaped restraint potentials allow for greater flexibility during restrained molecular dynamics simulations. Since the forces due to hyperbolic potentials are approximately

constant when the restrained distance is large compared to the radius of curvature at the potential minimum, potentials of this type allow 'targetted' molecular dynamic and steered transitions, which is possibly useful for localizing transition states.

A molecular dynamics routine based in the Beeman integration algorithm [14], as well as two stochastic dynamics routines, one based on the discretization given by Chandrasekar [24], and a more generally applicable one adopted from the TINKER package and based on a discretization scheme developed by van Gunsteren et al. [105, 106]. The latter was used in all implicit solvent simulations reported in this work

An augmented force field energy function, combining the AMBER force field with freely selectable harmonic or hyperbolic pairwise distance restraints. This function proved very useful for restraining known distances during simulated tempering minimizations.

• The RATTLE algorithm for constraining either bonds between hydrogen atoms and heavy atoms or all bonds to their equilibrium lengths.

• The LCPO (Linear Combination of Pairwise Overlaps) method [110] for approximate inclusion of non-polar solvation effects via a penalty term proportional to the solvent accessible surface area. This routine was adopted using the original implementation, that is part of the AMBER molecular modeling program suite, and the reference given previously.

• A wrapper routine allowing the imposition of bond length constraints during minimization. This routine is loosely based in the 'Gradient SHAKE' algorithm developed by Duan et al. [36], but seems to show better convergence than reported by the authors.

• A non bonded energy routine for 'droplet dynamics' simulations. In this routine, the Coulomb potential is replaced by the force derivative shifted Coulomb potential described in section 4.9. This routine allows for multi nanosecond simulations of NVE ensembles without any discernible drift of total energy.

• The principal component analysis functionality described in the previous section is based on the versatile superposition program `suppose`, that is part of the NAB distribution.

Additions marked by a • are publicly available within the latest release (version 4.5) of NAB.

7.2 TINKER

The tinker molecular modeling software package [85] contains the implementation of more different force fields than any other software package know to us. Like NAB, the source code is freely available. Implicit solvent simulations based one the TINKER package were used to cross-check with results obtained through simulations based on the corresponding NAB routines and showed good agreement without exception.

7.3 JUMNA

JUMNA (Junction Minimization of Nucleic Acids) is a molecular mechanics program developed by Heinz Sklenar and Richard Lavery [68,69]. It features two underlying force fields, AMBER 94 [30] and FLEX [67]. The JUMNA version used in this work is based on the AMBER 94 force field and includes the generalized Born approximation for the implicit treatment of solvent polarization effects. The GB approximation was implemented and kindly supplied by Martin Zacharias. In JUMNA, the conformation of nucleic acid molecules is described in sequence-independent internal coordinates. This is achieved by splitting nucleic acid molecules into a set of 3'-mono-phosphates via cutting the O5'-C5' bond of the phosphodiester backbone. These nucleotides are positioned with respect to a local helical axis with a set of 6 local helical parameters. These are the three translations Xdisp, Ydisp and Rise and the three rotations Inclination, Tip and Twist, according to the Cambridge convention [34]. The essential advantage of JUMNA when used for energy minimization of nucleic acid structures is the significant reduction of internal degrees of freedom when compared to 'all atom' minimization. This leads to a significant coarse graining and smoothing of the potential energy landscape (see discussion of figure 7) and to faster convergence of the minimization. The sequence independent representation of nucleic acid structures also facilitates the introduction of mutations.

One drawback of the internal coordinate representation is that small deviations from ideal bond lengths and valence angles present in the input structure can lead to large errors during the conversion to internal coordinates. Therefore, structures defined in Cartesian coordinates must be minimized in this representation before further minimization using JUMNA is safe. This holds for 'imperfect' structures, i.e. experimentally determined structures, molecular dynamics snapshots and structures generated by rigid

body transformations or distance geometry. Another drawback is the impossibility to estimate the error introduced by the reduction of the number of internal degrees of freedom. In general, however, the agreement between the energetic rankings based on JUMNA and all atom results is close and JUMNA is a valuable tool for minimization and preliminary energetic ranking of large numbers of conformers obtained by various conformational search procedures.

7.4 MEAD

MEAD (Macroscopic Electrostatics with Atomic Detail) [13] is a freely available software package for electrostatic investigations on macromolecules, including the calculation of solvation energies by numerically solving the Poisson Boltzmann equation. The flexibility of the programs contained in the MEAD package makes it ideally suited for advanced tasks like calculating optimal effective Born radii to eventually improve the pairwise descreening approximation. All solvation energies reported in the results section were obtained by using the `solvate` program contained in the MEAD package. A grid spacing of 0.2 Ångstroms and a grid size of 1.5 times the maximum solute interatomic distance was used in all calculations.

7.5 VMD

VMD (Visual Molecular Dynamics) [62] is a freely available, versatile and powerful program for the visualization of three dimensional molecular structures. It allows a large number of representations and the generation of movies from molecular dynamics trajectories, which often yield more insights than torsion angle analysis. All depictions of three dimensional molecular structures shown in this work were created using VMD. Last but not least, the VMD distribution contains the program `surf`, which is intended as an auxiliary tool for the visualization of molecular surfaces but can easily 'misused' to calculate surface integrals via a triangulation.

8 Results

8.1 Tetraloops

8.1.1 Conformational Transitions of a GCAA Tetraloop

Hairpins containing tetraloops are extremely common in biologically active RNAs. Three groups of tetraloop sequences appear most frequently: The GNRA, the UNCG and the CUUG sequences, where N can be any nucleotide and R is a purine, either G or A. Thermodynamic studies have shown that these frequently occurring RNA tetraloops are more stable than other four-nucleotide loops with the same stem [6, 7]. These stable hairpin loops may provide nucleation sites to ensure proper folding of secondary and tertiary structure in large RNA molecules [103]. GNRA hairpins are the most abundant family of tetraloops, and several members of of this tetraloop family have been investigated by biochemical, molecular modeling, X-ray and NMR techniques [57, 64, 72, 84, 99, 107, 113]. Nearly all experimentally determined GNRA tetraloop structures share the same overall features, whether the loop is part of an oligonucleotide [64] or of a large molecule and involved in tertiary interactions [22]. The most prominent common features are:

- The loop delimiting bases G and A stack onto the stem and form a sheared base pair.
- The second and third loop-bases stack onto each other, the third base stacks on to the loop-terminating A.
- The loop has a major change in the direction of the backbone between the first and second nucleotide, mostly due to the torsion angle α of the second loop nucleotide being in the +anticlinal range. The four atoms defining this angle are highlighted in figure 10.

A notable exception to the features listed above are two of the ten tetraloop structures with the sequence GGGCGCAAGCCU, determined by Jucker et al. [64] using NMR spectroscopy. Among the ten structures available under the PDB ID `1zih`, two (conformers 2 and 6) show the base C6 protruding into the solvent instead of stacked onto base A7. The same non-canonical orientation of base C6 has already been observed as a structural alternative to the standard geometry by Heus and Pardi in 1991 [57]. This seems to be

a structural alternative present only if the loop sequence is GCAA, since all other tetraloops investigated by Heus and Pardi (GAAA) as well as by Jucker et al. (GAGA and GAAA) exhibit the canonical loop shape with only small sequence dependent variations in the backbone torsion angles, consistent with the data obtained from other GNRA tetraloops. Figure 11 shows conformers

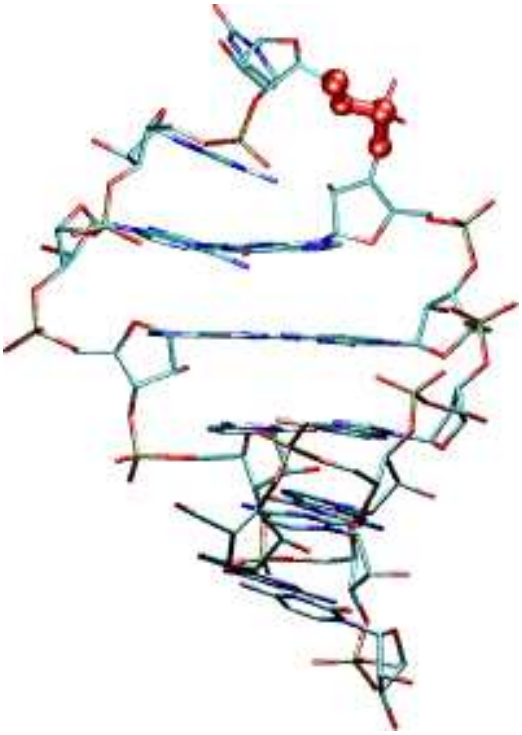


Figure 10: NMR structure of a GCAA tetraloop. The site of the backbone inversion is highlighted by red spheres.

one and two from `1zih.pdb`, optimally superimposed with respect to the stem-atom positions. Interestingly, all backbone torsion angles of residues G1 to G5 and A7 to U12 lie in the same ranges for both conformers, whereas for residue C7, all torsion angles except γ-C6 (55.8° vs. 58.3°) significantly differ. The sugar puckers for residue C6 are of C3'-endo type (conformer 1) and C2'-endo (conformer 2).

Since there are two distinct conformations, conversions between them should be observable if the free-energy barrier between them is not too high to be crossed within the time scale accessible within MD simulations. To investigate the temporal evolution of the tetraloop, two implicit solvent molecular dynamics simulations, differing only in the initial random seed, were set up in the following steps:

Conformer 1 from `1zih.pdb` (in closed loop conformation) was minimized until convergence. All bond lengths were constrained to their reference values using our implementation of the gradient-shake algorithm [36] for minimization, and the RATTLE algorithm [5] during the MD simulation. The electrostatic effects due to solvent polarization were included via the generalized Born approximation with all respective parameters chosen as given by Tsui and Case [101]. The electrostatic screening due to the presence of monovalent cations was implicitly included via a Debye-Huckel length of 9.62 Å, corresponding to 100 mMol concentration of monovalent salt [12, 101]. The effects of non-polar solvation were included using the LCPO method with a surface tension parameter of 5 cal/(mol Å^2). The time step chosen was 2 fs and no (i.e. infinite) cutoff was used.

The simulation commenced with 40 ps in contact with a Langevin heat bath at $T = 50K$ and a uniform friction constant of 0.5 ps^{-1}, slowly heating the structure. Equilibration continued with two intervals of 40 ps each at temperatures of $T = 200K$ and $T = 300K$. During the subsequent production run, the trajectories were stored on disk every picosecond for overall fifteen nanoseconds.

Figure 12 shows the first principal components, calculated from the positions of the heavy atoms (C,N,O,P) of the residues C4 to G9 of consecutive snapshots over both entire trajectories. Even without comparing PCs and corresponding snapshots, two results can be derived: In a coarse-grained sense, three different states are visited in both trajectories, the most populated one corresponding to PC1 fluctuating around -0.8 Å for the first trajectory and around around -0.3 Å for the second.

Although both trajectories span a time interval of 15 nanoseconds, they do not converge. This means that the time averages over both trajectories are different and that the principal axes derived from both trajectories do not point into the same directions. Therefore, the same structure will be represented by different coordinates in the space spanned by the principal axes, depending on whether it is part of the first or the second trajectory.

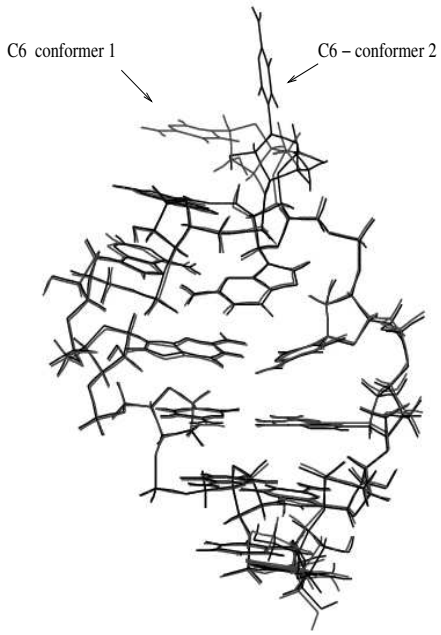


Figure 11: Conformers 1 and 2 from `1zih.pdb` superimposed. Conformer 1 exhibits the canonical loop-shape whereas base C6 of conformer 2 protrudes into the solvent.

Comparing the principal components with representative snapshots from the corresponding time intervals yields further insights: For both trajectories, the snapshots corresponding to PC1 $\lesssim -0.5$ Å and PC1 $\lesssim 0.0$ Å respectively, are of the 'open loop' type, close to conformers 2 and 6 from `1zih.pdb`. The highest peaks for both trajectories correspond to a short lived 'collapsed loop' state, where the sheared G5-A8 base pair is temporarily open and base C6 points downward towards the minor groove of the stem. This state is incompatible with the experimentally determined structures. It is either too short-lived to be observed in NMR measurements or an artifact due to the necessarily approximate character of implicit solvent MD simulations. The 'closed loop' conformation also appears short-lived in the second trajectory, but is stable over the time interval from approximately 11 to 15 nanoseconds in the first trajectory. Figure 13 shows the minimized averaged structures over the time intervals corresponding to the close (canonical), open and collapsed state.

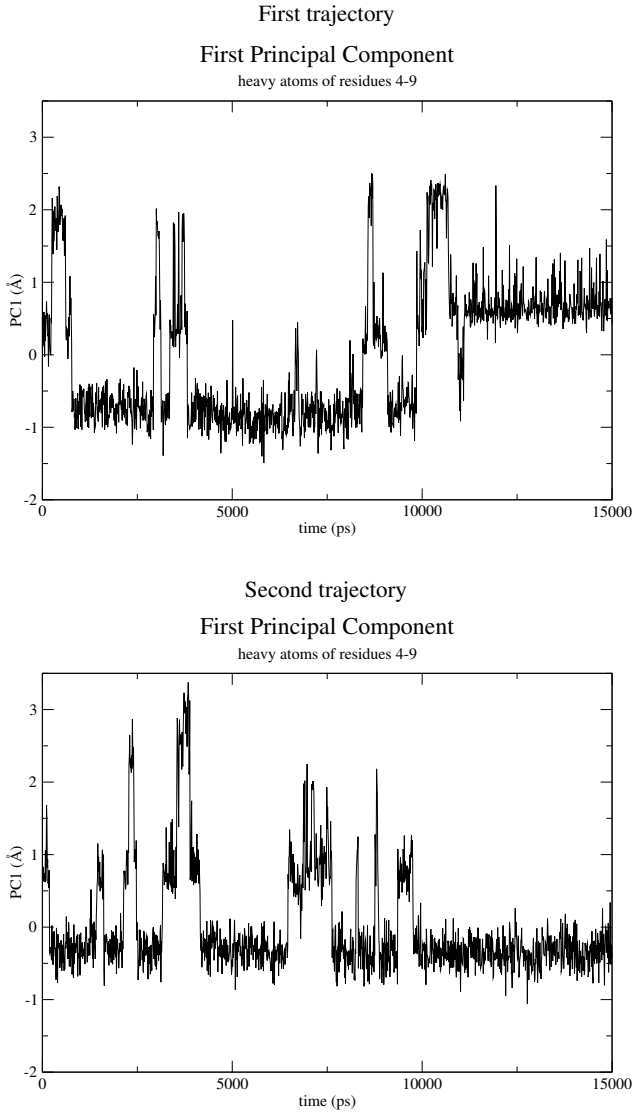


Figure 12: First principal component calculated from the coordinates of the heavy atoms of residues C4 to G9 over both entire trajectories.

The most detailed information about the transition between the two states in accordance with experimental data is obtained by analyzing the backbone torsion angles during the transitions. Figure 14 shows those backbone torsion angles undergoing significant and rapid changes during the transition from the open-loop state to the closed-loop state. Figures 15 and 16 show the

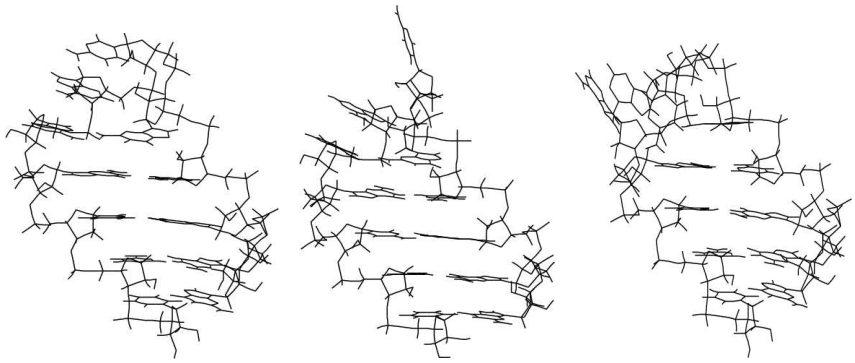


Figure 13: Minimized averaged structures over the three different states as derived from the first principal component.

time course of those backbone torsion angles changing significantly during the closed-loop to open-loop transition. Interestingly, only four torsion angles and the sugar pucker of residue C6 undergo a concerted rotation during a time interval of approximately 25 picoseconds. Most notable is the transient change in angle ζ-G5 that is seemingly necessary to provide the 'rotational freedom' for the concerted rotation to occur. The sugar re-puckering and the related change in angle δ-C6 occur most rapidly while the much slower switching of angle ϵ-C6 seems to anticipate the conformational change. The dashed lines in figures 15 and 16 show the standard values for the respective backbone torsion angles for ideal A-RNA helices as given by Saenger [90]. As expected, all angles except α-C6, the site of the 'canonical GNRA backbone inversion', shift towards the ideal A-RNA values during the transition. The reverse transition shows the same (time reversed) behavior of the torsion angles, except for the transient shift of ζ-G5 being of much shorter duration and the switching of ϵ-C6 occurring this time as suddenly as the conversion of the sugar pucker P-C6 back from C3'-endo to C2'-endo (data not shown).

The fact that both trajectories do not converge even after fifteen nanoseconds each prohibits a reliable estimate of the free energy difference between the three observed substates open, closed and collapsed. According to the force field energy (enthalpy), using the same parameters as during the simulation, the three representative minimized averaged structures are close to degenerate, as can be seen in table 1. For the behavior during the MD simulation, the energy values computed replacing the GB approximation by a

Conformer	Energy (GB)	Energy (PO)
open	-2849.0	-2921.0
closed	-2850.0	-2923.0
collapsed	-2850.0	-2925.0

Table 1: Force field energies of the open, closed and collapsed loop structures.

more accurate Poisson calculation are irrelevant. Interestingly, the collapsed state is definitely less populated than the open and closed state despite the nearly equal force-field energies and the open state *appears* to be the most favored state despite the slightly higher potential energy.

A possible explanation for the occurrence of the collapsed state, preceded by a complete opening of the G5-A8 base pair can be found in [64]: In GCAA and GAAA tetraloops, the G5-N3 to A8 amino proton distance is too large for the formation of a direct hydrogen bond, indicating that for these two loop sequences, the G5-A8 base pair is stabilized by one direct and one water-mediated hydrogen bond. This is an effect no continuum solvation approximation can account for. Despite the inherent inaccuracy of single-point corrections, it should be noted that the structure incompatible with experimental data is favored when the solvation energy is calculated via the numerical solution of the Poisson equation. It is, however, a success that the implicit solvent MD simulations of a GCAA tetraloop show repeated transitions between experimentally determined alternative structures, permitting a detailed analysis of the transitions in terms of concerted changes of a relatively small number of backbone torsion angles.

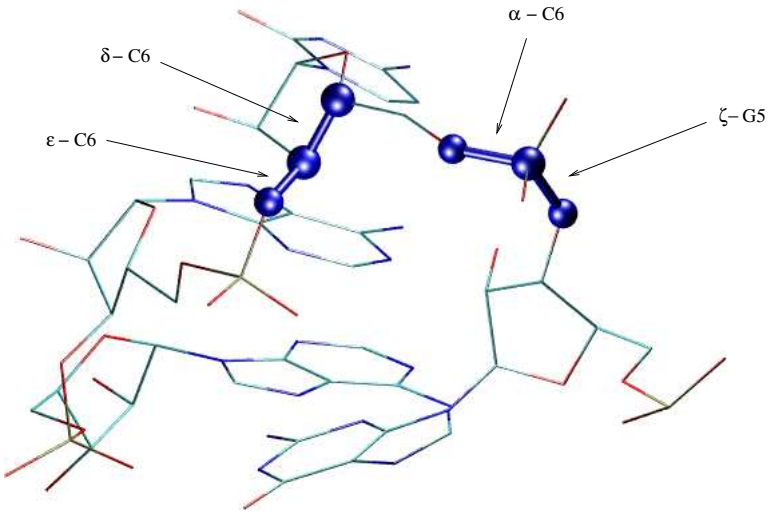


Figure 14: The loop bases GCAA. Stem bases and hydrogen atoms are omitted for clarity. The central axes (atoms 2 and 3) of those torsion angles depicted in figures 15 and 16 are highlighted.

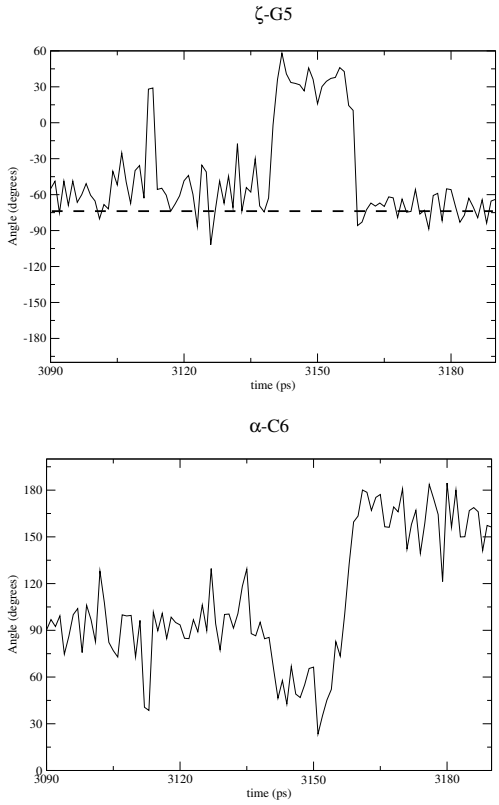


Figure 15: Backbone torsion angles ζ-G5 and α-C6 during the transition from open loop state to closed loop state. The dashed line shows the reference value from an ideal A-RNA helix when within plotting range.

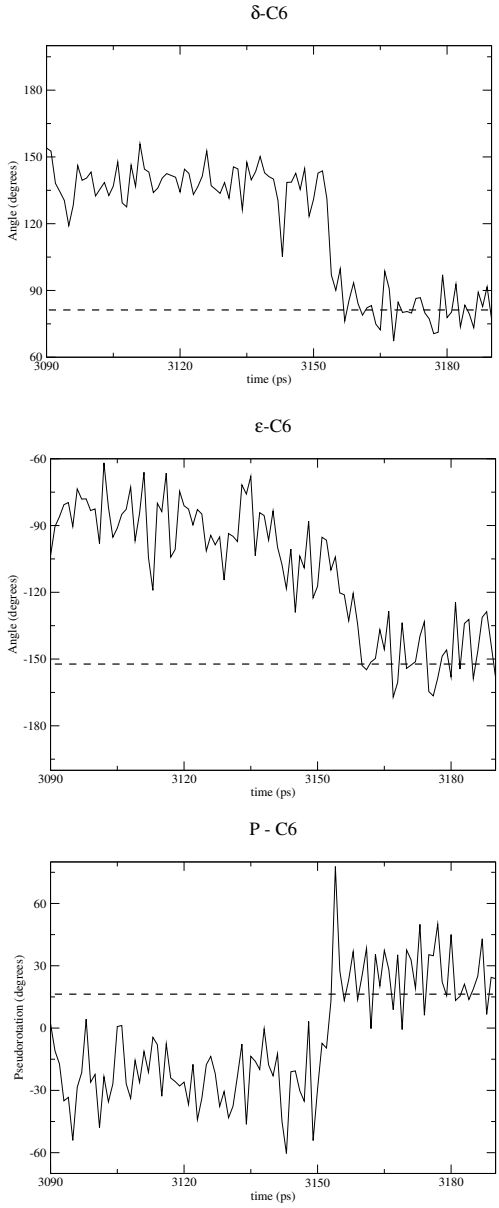


Figure 16: Backbone torsion angles δ-C6 and ϵ-C6 and sugar pucker P-C6 during the transition from open loop state to closed loop state. The dashed lines show the reference values from an ideal A-RNA helix.

8.1.2 Investigations on a GCCA Tetraloop

In contrast to GNRA tetraloops, their pyrimidine relatives of loop sequence GNYA occur rarely in nature [72]. This indicates that there is a high selective pressure to preserve GNRA sequences, presumably because the GNYA sequences do not fold into a well defined tertiary structure, necessary for sequence-specific tertiary interactions [84]. Static analyzes, based on the energetic ranking of different GNYA loop conformers showed that the stabilizing effect of the reaction field of the solvent is significantly stronger for GNRA loops than for their pyrimidine mutants. Another contribution to the reduced stability might be the absence of a possible hydrogen bond between atom N7 (or N4) of purine R7 and the HO2' group of base G5. Since molecular dynamics simulations can yield insights into the transitions between different loop conformations compatible with experimental results, we investigated whether the reduced stability of a GNYA loop is observable during a molecular dynamics simulation.

As a first step, the first conformer from `1zih.pdb` was mutated from sequence GGGCGCAAGCCU to sequence GGGCG**C**CAGCC**C**. This mutation and subsequent minimization were carried out using the JUMNA program [69], modified to include the electrostatic effects of solvation via the generalized Born approximation and kindly supplied by M. Zacharias. The mutation from U12 to C12 was introduced to stabilize the stack-terminating base pair during accompanying short simulations at temperatures higher than $300K$ (data not shown). The resulting mutant structure was then retranslated so as to be compatible with the standard AMBER representation. The simulation was set up following exactly the same protocol as in the simulations described in the previous section, only the duration was limited to 10 nanoseconds.

A first impression of the results can again be obtained from a principal components analysis of the coordinates of the heavy atoms of residues C4-G9. Figure 17 bears a striking similarity to the principal components shown in figure 12 and indeed, the mutant structure also visits open, closed and collapsed state, with the open state seemingly preferred. Repeated simulations differing only in different initial random seeds corroborate the three state picture (data not shown). This indicates, that the canonical GNRA conformation is at least meta-stable also for a GCCA tetraloop. If alternative structures exist, they are separated from the canonical initial structure by a free energy barrier too high to be overcome within the time scale accessible

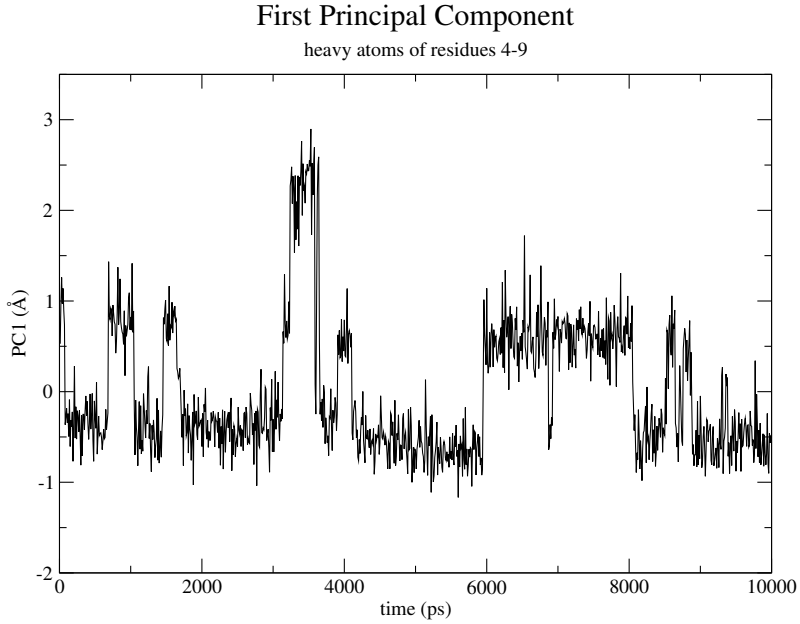


Figure 17: First principal component calculated from the coordinates of the heavy atoms of residues C4 to G9 over the entire trajectory.

to MD simulations. Due to the lack of experimental data, it is a matter of speculation whether a RNA molecule with sequence GGGCGCCAGCCC would fold into the canonical GNRA tetraloop shape at all.

To investigate the effect of explicit solvent, we also carried out a 'droplet dynamics' simulation of the mutated tetraloop, using the force derivative shifted Coulomb potential as given in section 4.9 and a cutoff of 25 Å. The initial structure of the explicit solvent simulation was neutralized by eleven sodium ions and solvated by a water sphere of 25 Å radius, containing 1753 TIP3P water molecules, using the `tLeap` program. The resulting system was minimized for 100 steps of conjugate gradient minimization, constraining all RNA atoms, allowing only counterions and water molecules to relax any initial steric strain. As a next step, the pre-minimized system was equilibrated with the solute atoms and counter-ions constrained. The TIP3P waters were heated to 380 K using a Langevin heat bath with a friction constant of 5 ps^{-1} over 1000 timesteps (2 picoseconds). This short heating period was followed by cooling down the solvent to 300 K over 2000 timesteps. Evaporation at the water-vacuum boundary was prevented by a semi-harmonic maximum

distance restraint for every atom pair, penalizing any interatomic distance larger than 55 Å as described in section 4.10 with 5.0 kcal/(mol Å^2).

Equilibration continued with ten picoseconds of stochastic dynamics (friction constant $\gamma = 0.1$ ps $^{-1}$) restraining the heavy atoms of the RNA molecule and the sodium ions to their previous positions with a harmonic potential of 5 kcal/(mol Å^2). The production run commenced with the same parameters, but without any restraints for 3.5 nanoseconds. To prevent any drifts due to the 'sudden' removal from of the harmonic restraints from influencing the results, only the trajectory after the first nanosecond was considered in the subsequent analysis. It should be noted that in this simulation the Langevin heat-bath was present, resulting in an extremely close agreement between target temperature ($300K$) and the average temperature ($299.92K$), while causing slight deviations from Newtonian dynamics.

Again, a principal component analysis of the coordinates of the heavy atoms of residues C4 to G9 over the trajectory shows a major transition of the loop structure. Figure 18 shows the ubiquitous first principal component and, inset, the averaged structures with averages taken over the time intervals corresponding to before and after the transition. A noteworthy feature of the loop transition is that it takes place over a time-scale comparable to the corresponding transition observed in the implicit solvent simulations. Figures 19 and 20 show those torsion angles whose concerted rapid change corresponds to the transition of the loop conformation from closed to open state, ordered according to their occurrence along the backbone in 5' - 3' direction. Where the standard values from A-RNA helices are within plotting range, they are shown as dashed lines. Probably due to solvent friction, the torsion angle transitions do not all occur at the same time. The transition is initiated by a simultaneous change of ϵ-C6, ζ-C6 and β-C7. The angles α-C6, δ-C6 and the sugar pucker change roughly 20 picoseconds later, but the sugar pucker shows a strong transient fluctuation, seemingly as a consequence of the change of the first three torsion angles. The change of β-C7 is remarkable because this torsion angle is not noticeably affected during the corresponding transition between the experimentally observed GCAA structures.

Figure 21 shows first and second principal component, calculated from all 101 snapshots available during the time interval considered in the previous torsion angle analysis. Since significant changes of backbone torsion angles were only observed in residues C6 and C7, only the heavy atom coordinates of these residues were taken into account. The first PC is shown by a full line, the second by a dashed line. Interestingly, PCA yields a clear separation

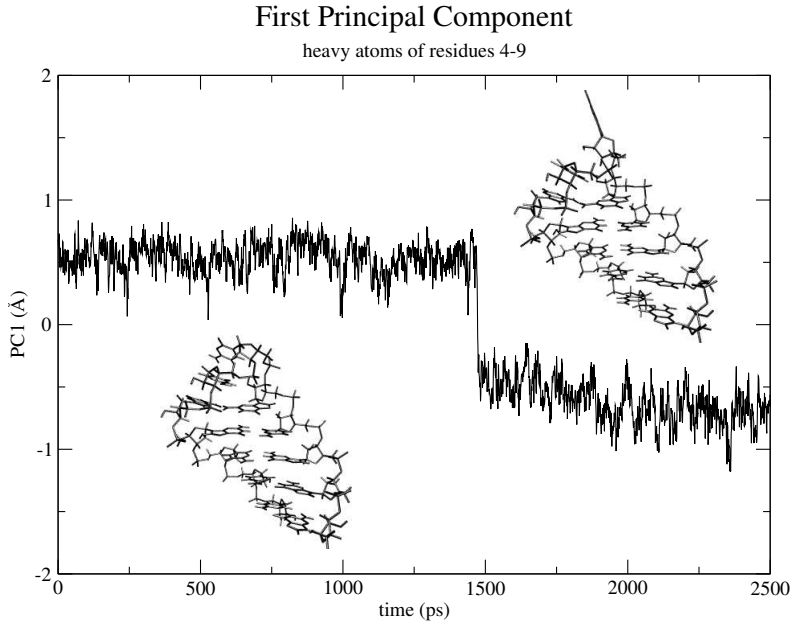


Figure 18: First principal component from the coordinates of the heavy atoms of residues C4 to G9. The presence of explicit solvent reduces short time fluctuations but also leads to a transition.

between the two concerted backbone transitions visible in figures 19 and 20. The initial change of ϵ-C6, ζ-C6 and β-C7 corresponds to a relatively sharp drop of the second PC around $t \approx 1455$ picoseconds, whereas the total rearrangement of the loop bases due to a concerted change of α-C6, δ-C6 and the sugar pucker P around $t \approx 1475$ picoseconds leads to a sharp drop in the first principal component. While the first concerted backbone rearrangement seems to initiate the overall change in loop structure, it contributes little to the change in loop geometry, since the second PC accounts for only about 6 percent of the total structural variance during the analyzed time interval, whereas the first PC accounts for about 78 percent of the total structural variance.

It is a matter of speculation whether a GCCA tetraloop would fold into the 'canonical' shape and therefore, whether the transitions observed during the MD simulations described above would also occur in nature, but it has been demonstrated that implicit solvent MD simulations and their counterparts where the solvent molecules are treated explicitly are in good qualitative agreement, not only with respect to the 'conservation' of experi-

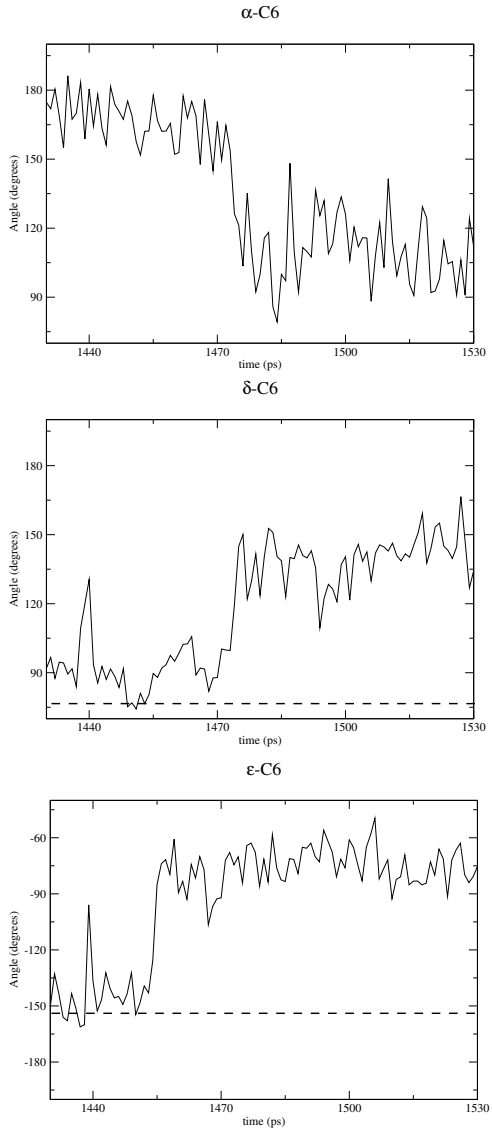


Figure 19: Backbone torsion angles α, δ and ϵ of residue C6 during the transition from closed to open loop conformation.

mentally determined structures but also with respect to transitions between alternative structures in agreement with experimental data.

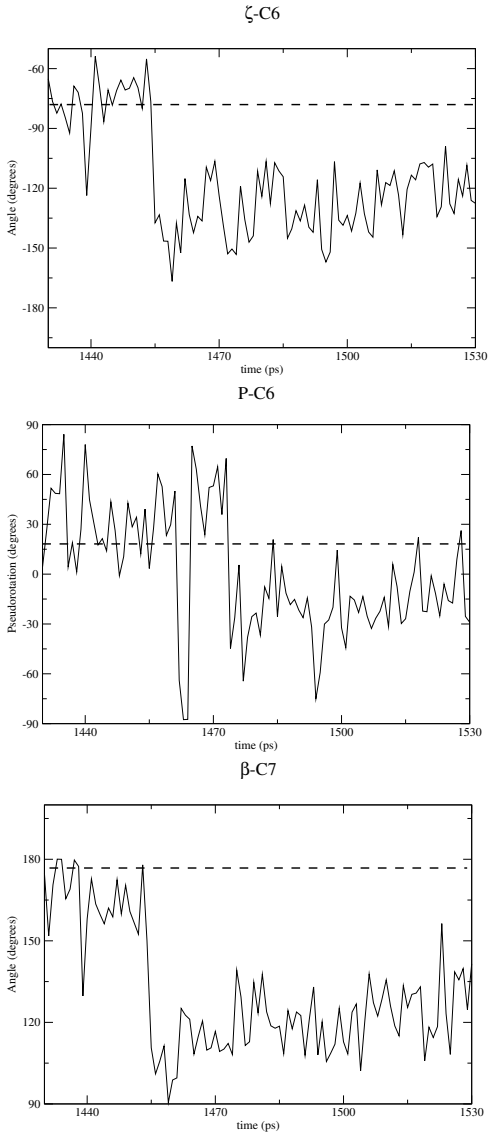


Figure 20: Backbone torsion angles ζ, and sugar pucker P of residue C6 and β of residue C7 during the transition from closed to open loop conformation.

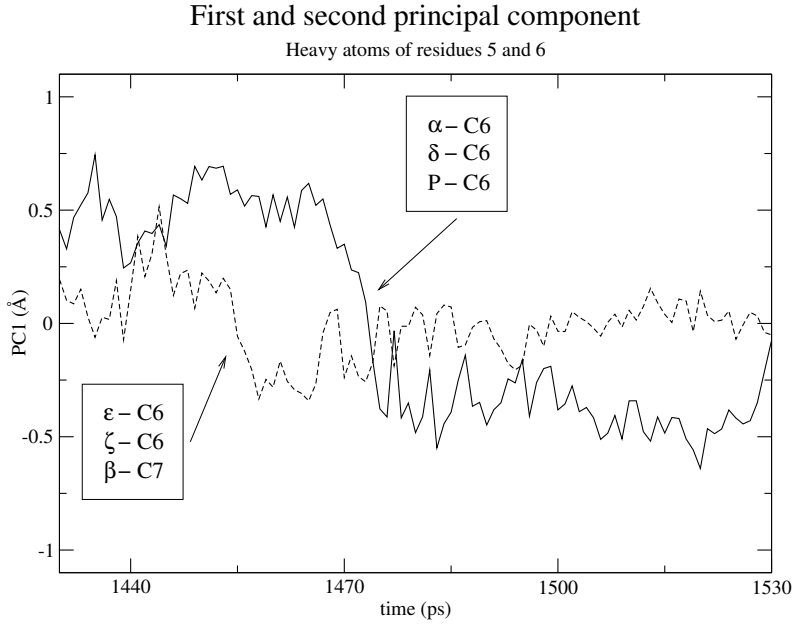


Figure 21: 'Fine grained' principal component analysis of the closed to open loop transition. The principal components were calculated from all heavy atom coordinates of the residues C6 and C7.

8.1.3 Simulations of a UUCG Tetraloop - a Comparison

Among the frequently occurring extrastable tetraloops, UUCG tetraloops have been found to the most stable ones [6]. Their exceptional stability has led to their use in experimental studies. For example, they are used as markers for NMR structure determination [1, 76] and as caps to shorten helical regions in X-ray [84] and NMR [60] investigations. So their stability does not only lead to their evolutionary conservation, but also to their use as a tool in experimental investigations in RNA structure and function. Numerous experimental [27, 38] as well theoretical studies [75, 111, 112] were aimed at finding an explanation for the exceptional stability, but besides an unusual hydrogen bond between the U1 hydroxyl and atom O6 of loop base G4 no complete and detailed explanation for the stability could be inferred from structural analysis. Figure 22 shows the loop bases of a UUCG tetraloop, extracted from `1c0o.pdb` [29] from two perspectives. The three possible hydrogen bonds contributing to the loop stability are shown in black.

Molecular dynamics simulations of UUCG tetraloops have been performed by Williams et al. in the first reported implicit solvent MD study of a non-helical RNA motif [112] and by Miller et al. using and explicit representation of the solvent and the Particle Mesh Ewald method for the treatment of long range electrostatic interactions. Here, an implicit solvent simulation and its counterpart with explicit representation of the solvent are compared directly. To our knowledge, this is the first direct comparison of both methods except the landmark study of Tsui et al. [101].

Both simulations were started from a 12 nucleotide tetraloop structure extracted from `1c0o.pdb` with sequence G**GU**CUUCGG**GU**C. The G-U base pairs G2-U11 and U3-G10 are highly unstable and account for the observed flexibility of the stem. The implicit solvent simulation was performed employing exactly the same protocol as given in section 8.1.1, but lasted overall 25 nanoseconds of production run. For the explicit solvent simulation, the starting structure was solvated by a sphere of 50 Å diameter, containing 1741 TIP3P water molecules and 11 sodium ions for neutralization. Minimization and equilibration were performed as given in section 8.1.2. For the smooth truncation of the long range electrostatic interaction, the force derivative shifted Coulomb potential as shown in section 4.9 was used with a 25 Å cutoff. During the production run, evaporation at the water-vacuum boundary was prevented by penalizing any interatomic distance greater than 57.5 Å by a semi-harmonic penalty of 5.0 kcal/(mol Å^2). In this simulation,

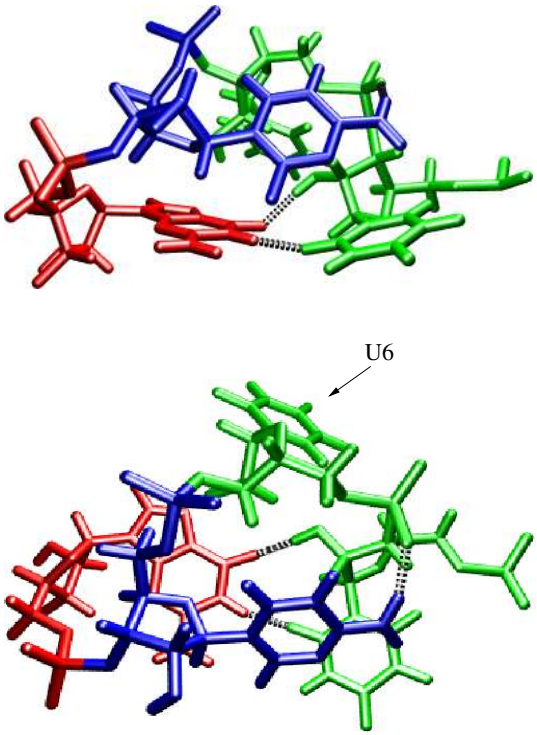


Figure 22: The four loop bases of the UUCG tetraloop structure, shown from up front and from above. Uracils are colored green, the cytosine in blue and the guanine in red. Three stabilizing hydrogen bonds are shown by grey dashed lines. The flexible base U6 is marked by an arrow.

the heat bath was switched off after equilibration, leading to an excellent conservation of total energy, corroborating the favorable numerical properties our long-range modification of the Coulomb potential. Table 2 shows the mean total energy of the solute-solvent system and its standard deviation over each nanosecond time-interval from first to seventh nanosecond, from data stored every second timestep. No drift can be observed. The excellent conservation of total energy is even more remarkable when considering the large timestep of 2 femtoseconds, which is generally considered the upper limit for molecular dynamics simulations with bond length constraints.

The system equilibrated at an average temperature of 304.1 Kelvin, by far close enough to the average temperature of the stochastic dynamics simulation (300.3 Kelvin) to allow a detailed comparison. Figure 23 shows the rmsd distance between individual snapshots and and the minimized NMR refer-

Time interval	$\langle E_{\mathrm{Tot}} \rangle$	$\sqrt{\langle (E_{\mathrm{Tot}} - \langle E_{\mathrm{Tot}} \rangle)^2 \rangle}$
1	-15851.60	0.454
2	-15851.58	0.451
3	-15851.58	0.450
4	-15851.60	0.450
5	-15851.62	0.453
6	-15851.61	0.450
7	-15851.57	0.448

Table 2: Average total energies, calculated over seven consecutive nanosecond time intervals.

ence structure over the first seven nanoseconds. The peaks in both graphs in figure 23 correspond to strong short lived fluctuations of the stem, due to the instability induced by the two G-U base pairs. However, due to the absence of solvent friction, these fluctuations are more frequent and of shorter duration in the implicit solvent simulation. The average rmsd between individual snapshots and the minimized NMR structure is virtually identical in both simulations (Explicit solvent - 0.95 Å versus implicit solvent - 0.94 Å), and compares well to state of the art simulations with explicit solvent and no truncation of electrostatic interactions [75].

Figure 24 shows the distribution of root mean square deviations with respect to the NMR structure over both entire trajectories. Here, some overall differences become visible. The implicit solvent trajectory samples more structures strongly deviating from the NMR structure. The explicit solvent trajectory is shorter than the implicit solvent one due to computational limitations and solvent friction significantly slows down conformational transitions. So this difference can either be due to insufficient sampling [26] or to the limitations of implicit solvent simulations. The 'bulge' in the distribution function of the explicit solvent trajectory (marked by an arrow in figure 24) stems from the high percentage of snapshots from the explicit solvent trajectory with the highly flexible loop base U6 in syn-conformation. This in agreement with the results of the PME simulations conducted by Miller and Kollman [75] and of the GB/SA simulation done by Williams and Hall [111]. To our knowledge, however, there is no direct experimental evidence for the possibility of base U6 adopting *syn* as well as *anti* conformation, or rather for the *syn*-conformation significantly contributing to the ensemble. Our

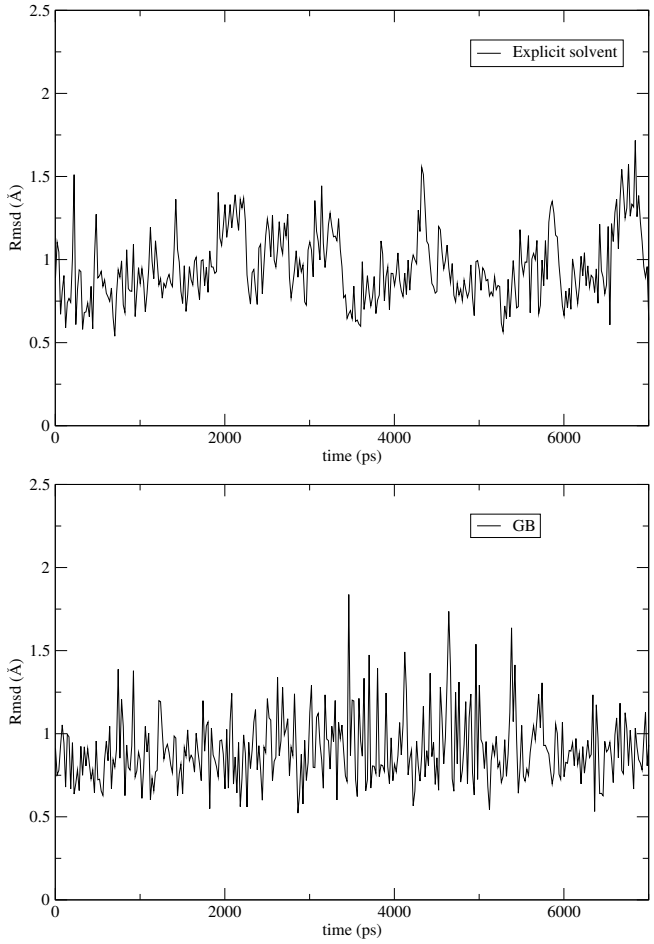


Figure 23: Time course of the root mean square deviation of the heavy atoms of residues G2 to U11. To facilitate direct comparison, both graphs show the same time window.

finding that the *syn* conformation of base U6 hardly at all contributes to the ensemble in the implicit solvent simulation is in better agreement with experimental data, but contradicts the results of our own explicit solvent simulation as well as the other simulation results cited here. This again indicates that even slight differences in simulation protocol and parameterization can significantly shift conformational preferences. Figure 25 shows the time course of the glycosidic torsion angle χ-U6 during the first seven nanoseconds of both simulations and their distribution. It must be noted that no

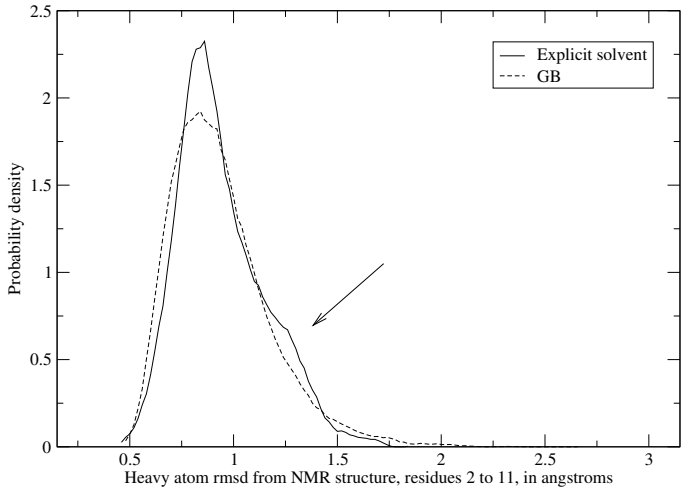


Figure 24: Distribution of rmsd deviations between individual snapshots and the NMR structure. The distribution function for the explicit solvent simulation is represented by a full line, that of the implicit solvent simulation by a dashed line.

equilibrium properties the χ-U6 distribution can be deduced from the angle distribution of the explicit solvent simulation, since the number of distinct rotations of base U6 around the glycosidic bond is small despite the relatively long simulation time of 7 nanoseconds. Individual atomic fluctuations around the respective average positions are in good agreement for both simulations conducted throughout this work and also with the results of Miller and Kollman [75]. Figure 26 shows the atomic root means square fluctuations of all atoms of the tetraloop structure relative to their average positions. The peaks at the 5' and 3' end stem from the inherent flexibility of the terminal base pair. The exceptionally high peak at the 5' end for the implicit solvent simulation is due to the two conformations adopted by base G1 and shortly discussed further down. All other significant structural variation observed in both simulations is due to the extreme flexibility of the stack, caused by the two G-U base pairs. This flexibility leads to large short time deviations from the NMR structure (up to 2.2 Å rmsd during the explicit solvent simulation and up to 2.9 Å rmsd with implicit solvent). These large short-time fluctuations, however, contribute little to the ensemble and the agreement between the minimized NMR structure and the average structures from both trajectories is excellent. The following table gives the pairwise heavy-atom root mean square distances between the minimized NMR structure and the

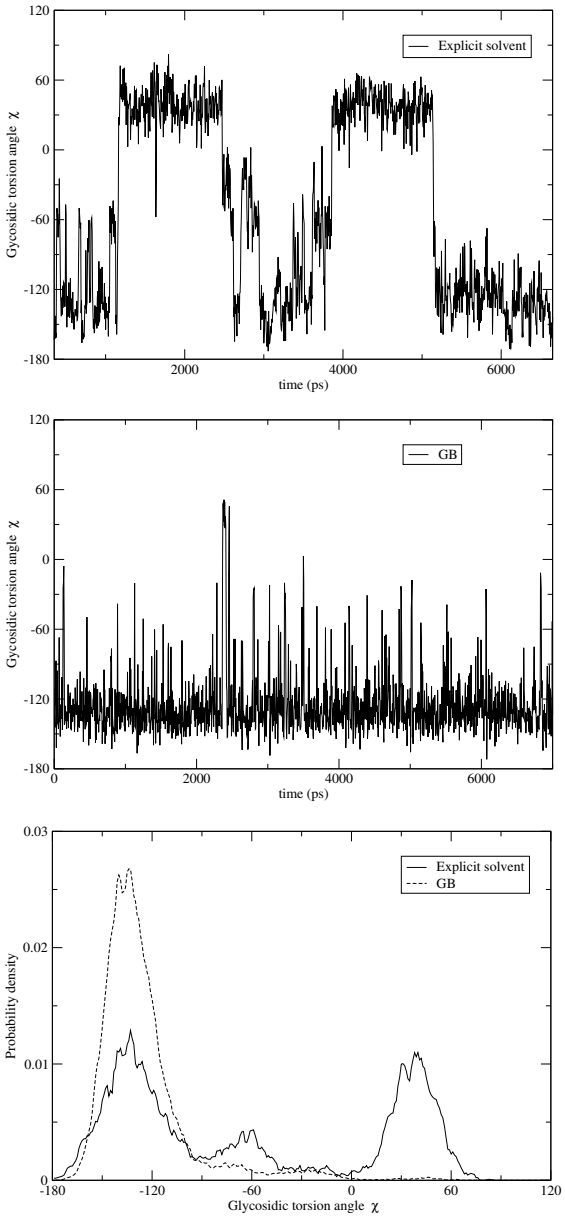


Figure 25: Time course and distribution of the glycosidic torsion angle χ-U6 during the explicit solvent simulation, upper graph, and of the implicit solvent simulation, middle graph. The lower graph shows the distribution of angle χ-U6 from both simulations.

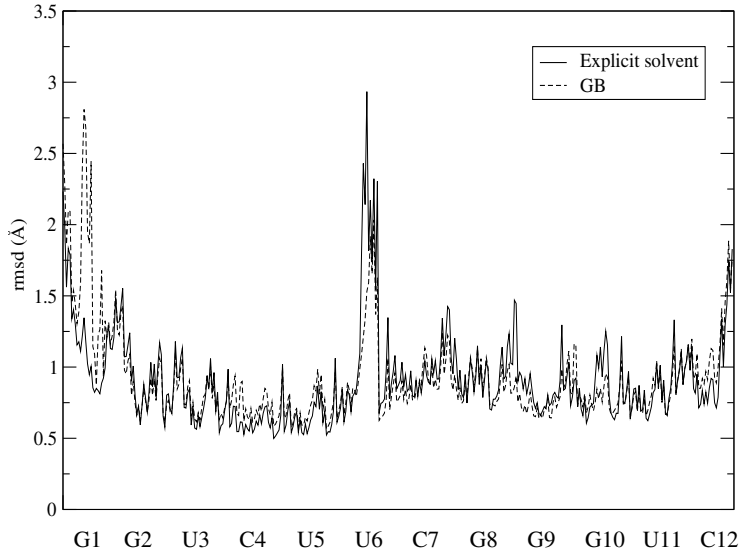


Figure 26: Individual atomic rms fluctuations for all atoms relative to their average positions. Data obtained from the explicit solvent simulation are given by the full line, those from the implicit solvent simulation by the dashed line.

two average structures from the simulations, termed 'EXP' and 'GB' for the sake of brevity. The overall similarity between the average structures and

Structure 1	Structure 2	Rmsd
NMR	GB	0.6
NMR	EXP	0.4
GB	EXP	0.6

Table 3: Pairwise rmsd between minimized NMR and average structures from the simulations.

the NMR structure is the more remarkable because, due to the two distinct (meta)stable conformations of base U6, the averaged atom positions of this base do not correspond to a sterically possible structure. This also applies to base G1. The lower stack-terminating base pair G1-C12 opened and closed with base G1 in syn-conformation after about 12 nanoseconds and stayed in this unusual conformation for roughly 3.7 nanoseconds. Therefore, the average atom positions of base G1 also do not correspond to a sterically possible conformation. Due to the positioning of the superimposed average

structures, this is not visible in figure 27. Figure 27 shows the minimized NMR structure in green and the superimposed average structures, from the implictic solvent simulation in red and from the explicit solvent simulation in blue.

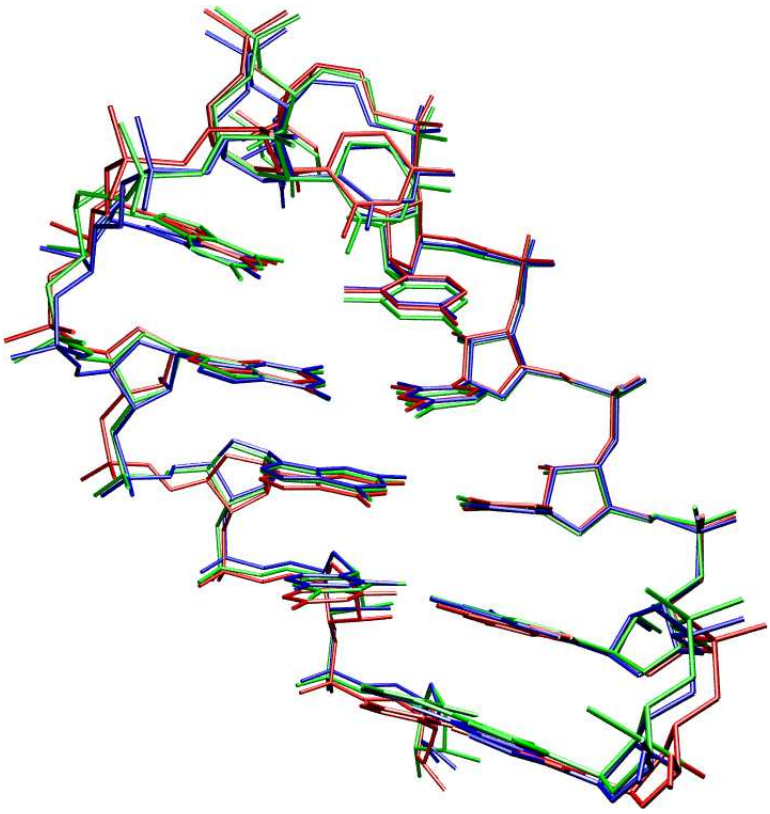


Figure 27: Minimized NMR structure (green) and the superimposed average structures from the implicit (red) and explicit solvent (blue) simulations.

The results obtained from the simulations of a UUCG tetraloop illustrate most of all the exceptional overall agreement between implicit and explicit solvent simulations. The increased flexibility of the loop base U6 observed in the explicit solvent simulation can be interpreted in two ways:

(i) The complete rotation of the flexible base U6 about the glycosidic bond observed twice in the explicit solvent simulation is a 'frozen accident' and much longer simulations would lead to similar distributions of χ-U6 in both simulations. The simulations of Miller and Kollman give evidence to the contrary. While they show neither the time course of χ-U6 nor the distribution of this torsion angle, they report a standard deviation of ± 60 degrees, which clearly indicates a roughly bivariate distribution.

(ii) The differences between experimental evidence and our implicit solvent simulation on the one side and the explicit solvent simulations on the other side are inherent in the different applied approximations and molecular dynamics simulations are at the current state of development not accurate enough to decide such subtle questions.

8.1.4 Structure Prediction by Energy Minimization ?

The exceptional stability or structural well-definedness of GNRA tetraloops indicates that the native conformation corresponds to a set of closely related minima of the free energy landscape. In the case of especially stable tetraloops, the magnitude of the entropic contributions can be expected to lie within the range of uncertainty of the force field parameters and the treatment of the dielectric solvent as a continuum [72]. Indeed, tetraloop conformers in excellent agreement with experimentally determined GNRA structures were found as those of lowest enthalpy from among a restricted set of trial conformers generated by the MC-SYM program [45,73], when the reaction field energy was included in the final energetic evaluation. The study of Maier et al. also showed that for a proper energetic ranking of conformers, the inclusion of the reaction field energy is of crucial importance, and that simple linear or sigmoidal dielectric damping do not lead to a correct energetic ranking of conformers, even among the restricted set considered. Since, with the generalized Born approximation, a computationally efficient and completely analytical -albeit approximate- treatment of the solvation energy contribution has become available, the question arises whether at least (sub)structures like loop conformations could be predicted by a combination of conformational sampling and energetic ranking. At the outset, the following must be considered:

- In the absence of a target structure determined by experiment, any conformational search is an open-ended process.
- If a target structure exists, the evaluation of its force field energy introduces a certain amount of arbitrariness, since the target structure itself must be minimized to obtain a reference value. This minimization must be subject to restraints in order to prevent the reference structure from deviating too far from the original structure. The choice of restraints, as well as the minimization protocol can influence the reference value.
- All methods for conformational search devised and applied throughout this work -except plain molecular dynamics at higher than physiological temperature- depend on relatively large sets of parameters that have to be chosen by educated guess. The occurrence of conformers with energies close to or lower than the reference value obtained from the lowest energy conformer obtained from experimental data was found to

be a rare event. Therefore an optimization of the parameter sets based on their success in yielding low energy conformers is not feasible.

Due to the abundance of structural data available, a GCAA tetraloop structure with sequence GGGCGCAAGCC**C** was chosen. The structures were taken from `1zih.pdb` [64] and the terminating GU base pair was mutated to a GC base pair using the JUMNA program. As might be expected, mutations of the stack-terminating base pair never noticeably influenced the energetic ranking of different loop conformations. All ten (mutated) structures were minimized using restrained simulated annealing and subsequent conjugate gradient minimization to convergence. All heavy atoms were restrained to their reference positions with a semi-harmonic restraint function penalizing deviations d of more than 1.5 Å with a potential of the following form:

$$V_p(\mathbf{r}-\mathbf{r}_0) = \begin{cases} w(|\mathbf{r}-\mathbf{r}_0|-d)^2, & |\mathbf{r}-\mathbf{r}_0| > d \\ 0, & |\mathbf{r}-\mathbf{r}_0| \leq d \end{cases} \tag{53}$$

with a restraint strength of $w = 5.0$ kcal/(mol Å^2) The lowest energy conformer obtained this way was chosen as the reference structure. Due to the unexpected outcome of this investigation described below, restrained simulated annealing of the experimental structures was repeated several times, with different initial temperatures as well as different cooling schedules so as to minimize the probability of narrowly missing 'the' optimal structure. Since all applied protocols yielded agreeing results, details about them are not given here. The sampling methods shortly described in the following only showed that at the current stage of development, despite the recent advances in the treatment of electrostatic solvation effects, the AMBER force field is unable to discern native conformers from well minimized non-native decoy structures. Therefore, the different methods are shortly described in the following and representative conformers will be described afterwards.
The conformational search methods applied were:
Distance Geometry: The stack of the reference structure was assumed rigid. A hexanucleotide structure with sequence CGCAAG was generated, using the 1–2, 1–3 and 1–4 distances as described in section 4.1.1, with all interatomic distances of residues C1 and of residue G6 taken from the residues C4 and G9 of the reference structure. Each random hexanucleotide structure obtained from the chosen set of distances, again as described in section 4.1.1, was optimally superimposed onto the reference structure with respect to the base atoms of C1(C4) and G6(G9) and the coordinates of

the new loop coordinates replaced the old loop coordinates. This process is illustrated in figure 28. The conformers generated by the use of distance geometry are completely random and sterically feasible (since one ubiquitous distance criterion is that no non-bonded pair of atoms may be closer than the sum of their Van der Waals radii), but without additional assumptions formulated as distance constraints, the number of possible conformers is too large be handled, even with the currently available computational resources. However, after generating 10000 trial conformers as described above and further minimization of those 30 (!) being inside a range of 10 kcal/mol from the lowest initial energy found, a family of conformers with lower energy than any of the minimized NMR structures from Jucker et al. was found, showing no similarity whatsoever with the well known canonical loop motif.

Molecular dynamics at 340 Kelvin: To investigate whether a tetraloop structure with sequence GGGCGCAAGCCC, built using the JUMNA program and the torsion angle pattern of a UUCG tetraloop (extracted from PDB entry `1COO` [29]), would re-fold into the correct canonical GNRA loop shape, an implicit solvent molecular dynamics simulation was performed. The protocol applied and parameters used were the same as given in section 8.1.1, except for the absence of any constraints on bond lengths and, therefore, a smaller timestep of one femtosecond. The target temperature was $340K$ to accelerate the transition between different conformations. The UUCG conformation was left after about two nanoseconds and between $t \approx 2$ nanoseconds and $t \approx 7$ nanoseconds, the loop structure fluctuated around a

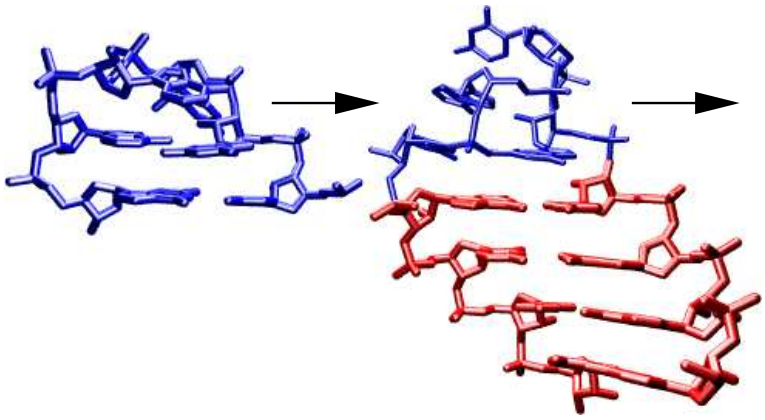


Figure 28: A randomly generated new loop structure replaces the previous one.

deep local minimum of the potential energy (again 'better' than anything in the vicinity of the experimental reference structure), before adopting another (energetically) less favorable conformation with base A8 protruding into the solvent until the end of the simulation. Figure 29 shows three representative loop conformations from $t = 0$ nanoseconds, $t = 7$ nanoseconds and $t = 10$ nanoseconds. Stems and hydrogen atoms are omitted for clarity. Due to the unphysiologically high temperature, the terminal GC base pair opened and closed several time during the course of the simulation.

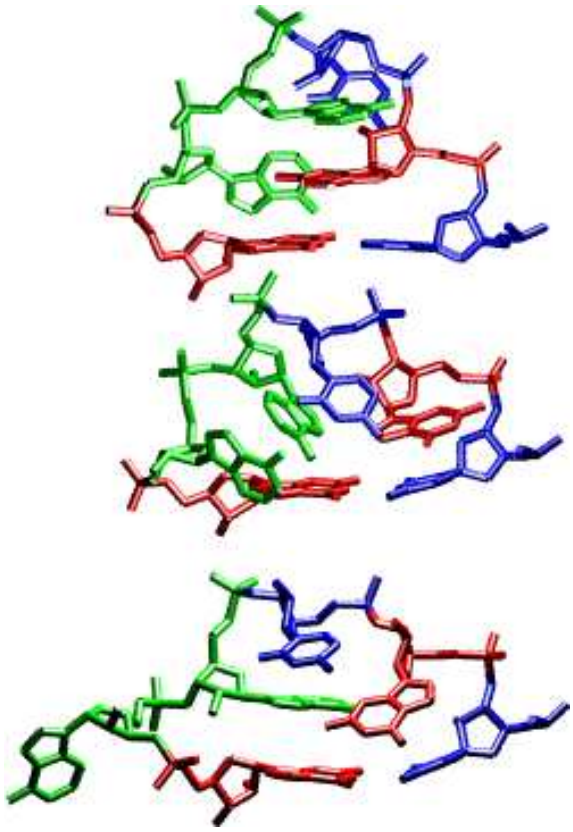


Figure 29: Three representative loop snapshots as described in the text, showing residues C4 to G9. cytosines are colored in blue, guanines in red and adenosines in green.

Enhanced sampling based on local elevation: In this approach, molecular dynamics was used simply to facilitate structural relaxation in the presence of repulsive restraints. All atoms of the first and last three stack bases were constrained to their starting positions, while the heavy atoms of the loop bases G5-A8 were restrained with a root mean squared distance (rmsd) based energy penalty as described in section 6. In short, the last ten structures visited during the conformational search runs were penalized, preventing them from switching back and forth between closely related successive shapes. With $\{\mathbf{c}_i\}$ denoting the positions of the restrained atoms, $\{\mathbf{c}_{0i}\}$ denoting the penalized reference coordinates and N_c their number, each of the last ten visited structure was assigned a penalty potential of the form:

$$V_p(\{\mathbf{c}_i\}) = V_0 \exp(-\lambda d^2) \tag{54}$$

with

$$d^2 = \frac{1}{N_c} \sum_{i \in N_c} (\mathbf{r}_i - \mathbf{r}_{\mathbf{0}i})^2 \tag{55}$$

Here, the parameters V_0 specifying the height of the repulsive 'mountains' as well as their 'inverse squared width' λ and the number of penalized structures kept in memory are more or less arbitrary parameters. So are the friction constant for the stochastic dynamics and the timestep, that can in this case chosen as large possible, as long as the integration is numerically stable. Height and width were chosen as $V_0 = 30$ kcal/mol, $\lambda = 1.0$ Å^{-2}. The friction constant was chosen to be $\gamma = 10$ ps^{-1}, to assure that the kinetic energy gained by the relaxing structures is absorbed by the heatbath between updates of the penalized structures. At the end of relaxation and before updating the list of penalized structures (every 3000 timesteps), snapshots were stored on disk. These snapshots were minimized so as to be compatible with the internal coordinate representation of the JUMNA program. The pre-minimized snapshots were further minimized using the JUMNA program with solvent polarization effects included via the GB approximation. For consistency, those conformers having lowest energy according to JUMNA were finally retranslated to be compatible with the standard AMBER representation and evaluated using the AMBER99 forcefield, again with with the reaction field contribution approximated via the GB method. It must be noted, that the outcome of local elevation sampling runs depends on the starting structure and 'frozen accident' or -in other words- the 'right' random seed at the beginning of the runs. Different runs were performed starting from experimentally

determined structures (conformers one and ten from `1zih.pdb`), the UUCG like structure that served as starting point of the previously described simulation at 340 Kelvin and the lowest energy conformer found by distance geometry based sampling and simulated tempering.

All methods for conformational search described above were aimed at testing whether it is feasible to predict a tetraloop structure by sampling of conformations and energetic ranking. In principle a negative answer to this question can be derived from the existence of the short lived 'collapsed loop' state occurring during the simulation of the GCAA tetraloop discussed in section 8.1.1. The search for low energy loop conformers was, however, continued in order to investigate whether other low energy conformers with entirely different arrangements of the loop bases exist. It is reassuring that the number of different conformers of lower energy than the experimental reference structure is so small that they could be assigned to different families by visual inspection, although the aforementioned search procedures yielded overall several thousands of structures. In the following , the lowest energy conformers are shown according to their energetic ranking in figure 30. Figure 31 shows the loop structure found closest to the reference structure and the reference structure itself. The lowest energy conformers **Ia** and **Ib** were found using distance geometry and subsequent simulated annealing and again during the high temperature simulation described previously. Conformers close to **Ia** and **Ib** were also repeatedly obtained during conformational search runs based on local elevation, starting from experimentally determined structures as well as from the UUCG like conformation. Conformer **II** was found once in a local elevation search run starting from the experimentally observed structure and bears a striking similarity to the base arrangement of UUCG tetraloops. It contains a nearly perfectly planar regular GA N3-amino-amino-N1 base pair [90], forming a "diloop". Structures **IIIa** and **IIIb** were also found during the 'local elevation' searches, starting from either the UUCG conformation or from the experimentally determined structure.

Table 4 shows the result of the final energetic evaluation of the conformers **I** to **IV** and of the reference structure. The angle and dihedral contributions as well as the Coulomb and solvation terms are given separately, since they show a systematic difference between experimental and artificial structures. Total energies (for AMBER and JUMNA) are rounded to the closest integer value, since any more accuracy is obviously below the error threshold inherent in energetic evaluations based on force fields. All energy values are given in kcal/mol. JUMNA, based on the AMBER94 force field and AMBER (using the AMBER99 force field) are roughly consistent in the resulting

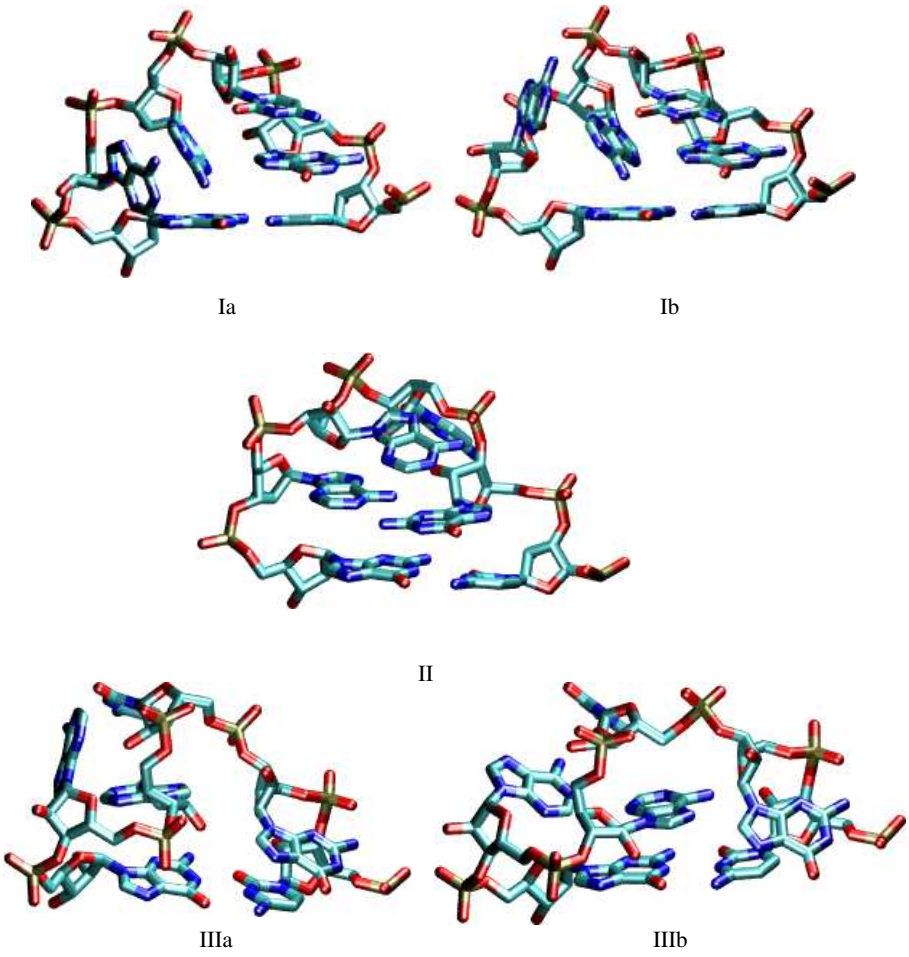
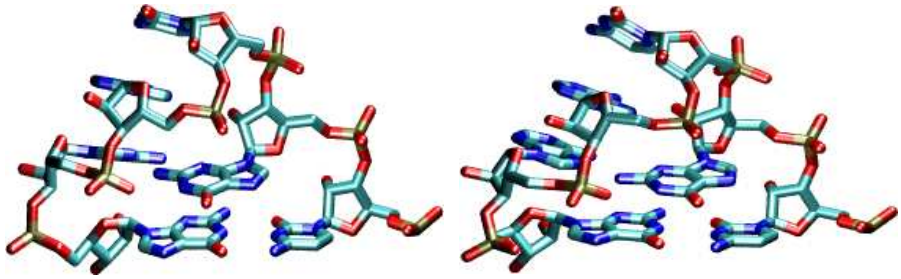


Figure 30: GCAA loop conformers of lower energy than the reference structure. Only the uppermost stem and the four loop residues are shown.



IV Reference

Figure 31: The conformer found closest to the reference structure (**IV**) and the reference structure.

energetic ranking. As can be seen in table 4, all 'false best' conformers show

Conformer	E_{Total}	$E_{ang+dihed}$	$E_{Coulomb}$	$\Delta G_{sol(GB)}$	E_{JUMNA}
Ia	-2948	314.98	-923.70	-2211.71	-1475
Ib	-2948	315.65	-942.07	-2195.87	-1476
II	-2944	318.04	-937.00	-2198.10	-1475
IIIa	-2944	313.19	-895.41	-2233.3	-1474
IIIb	-2943	316.18	-882.66	-2244.47	-1474
IV	-2938	320.95	-850.06	-2278.03	-1471
REF	-2939	320.18	-789.42	-2336.20	-1471

Table 4: Force field energies of false optimal and the experimental reference conformers.

a lower contribution from the angle and dihedral terms, which is somewhat surprising, since in the reference structure, the only torsion angle significantly deviating from the ideal A-RNA helix value is α-C6. Another systematic discrepancy between the reference structure and conformers **Ia** to **IVb** is that *all* artificial structures show a significantly less favorable solvation energy than the reference structure, which is in each case compensated by a more favorable Coulomb contribution. The performance of the GB approximation in the evaluation of differences in solvation energies is at least disconcerting when unrealistic structures (e.g. conformers **I** to **IIIb**) are concerned. It has been amply documented that the GB approximation yields differences in solvation energies in very good agreement with computationally more de-

manding finite difference Poisson calculations if the analyzed structures are similar, e.g. different snapshots from MD trajectories or closely related loop conformers [35, 37, 63, 86, 101, 120]. However, as can be seen in table 5, the GB approximation fails to yield correct differences in solvation energies when the analyzed structures are dissimilar and, in addition, novel and 'unnatural' structural motifs like the relative positions of bases C4 and G5 in conformers **Ia** and **Ib** are present. Replacing the GB results with the PO results when

Conformer	$\Delta G_{sol(GB)}$	$\Delta G_{sol(PO)}$	$\Delta G_{sol(GB)}$-$\Delta G_{sol(PO)}$
Ia	-2211.71	-2287.53	75.82
Ib	-2195.87	-2274.30	78.43
II	-2198.10	-2277.84	79.74
IIIa	-2233.35	-2305.41	72.06
IIIb	-2244.47	-2311.20	66.73
IV	-2278.03	-2346.04	68.01
REF	-2336.20	-2404.82	68.62

Table 5: Comparison of solvation energies evaluated with the GB approximation and from PO calculations.

evaluating the total energy additionally favors conformers **Ia** to **IIIa** over the reference structure. This fact indicates that single point evaluations, i.e. energetic evaluations of structures previously minimized with a different energy function, bear an inherent additional inaccuracy. One might speculate that if the forces due to the reaction field could be included during the minimization, the discrepancies would at least decrease. Conformers **Ia** to **IIIb** all contain at least one of the loop bases in the rarely occurring *syn*-conformation. Base G5 of conformers **Ia** and **Ib** is neither in *syn* nor in *anti* conformation with $\chi \approx 0$ degrees. Since bases in *syn*-conformation rarely occur in natural nucleic acid molecules, the AMBER force field, being parameterized to optimally describe regular RNA and DNA structures, possibly treats bases in this conformation too favorably [116]. All low energy loop structures shown except the reference structure also contain unusual sugar puckers. In conformers **Ia** and **Ib**, the ribose ring is in C1'-exo and C3'-exo conformation respectively. In conformers **IIIa** and **IIIb**, the sugars of the three consecutive loop bases C6 to A8 are all in C2'-endo conformation. Figure 32 shows the absolute deviations of the backbone torsions angles α-G5 to ζ-A8 ordered according to their occurrence along the backbone in direction from 5' and to

3' end. Figure 32 shows that the majority of all loop backbone torsion angles of conformers **Ia** to **IIIb** significantly deviate from their common values in A-form helices. Most deviations lie around 60, 90, 120 and 150 degrees, indicating that the respective angles adopt 'locally optimal' values according to the periodicities of the corresponding force field energy terms (cf. section 4). Figure 33 shows the torsion angle deviations from the values in A-form helices for conformer **IV** and the reference structure. A conspicuous trait of the reference structure is that only α-C6 significantly deviates from the A-form helix value, corresponding to the single large peak in the right graph of figure 33. This deviation pattern is a characteristic of the ideal canonical GNRA loop shape. Other torsion angle patterns corresponding to the same overall shape but having a slightly higher energy are also compatible with NMR data [64] and were also found by Maier et al. [72]. Conformer **IV** shows

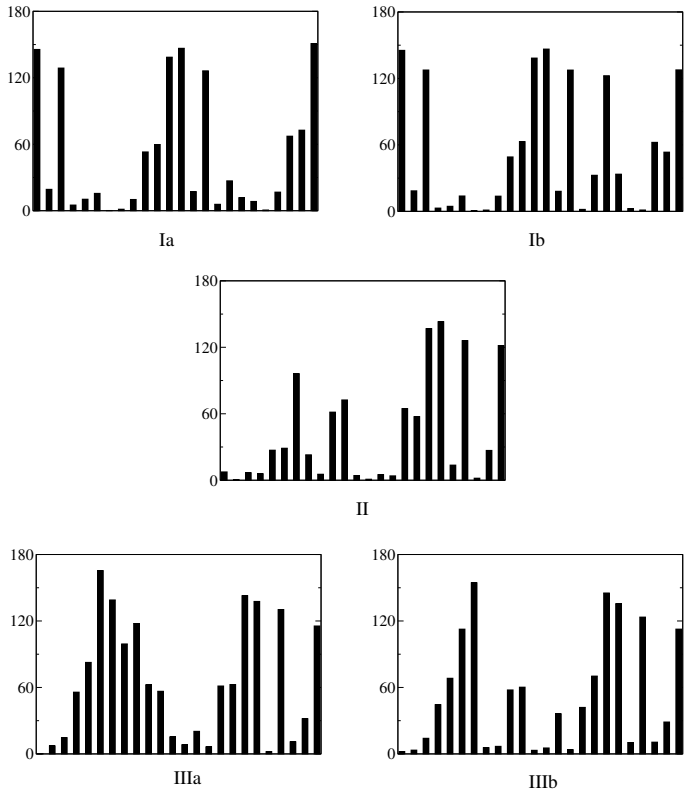


Figure 32: Conformers **Ia** to **IIIb**: Absolute deviations of the backbone torsion angles α to ζ of residues G5 to A8 from the ideal A-helix values in degrees.

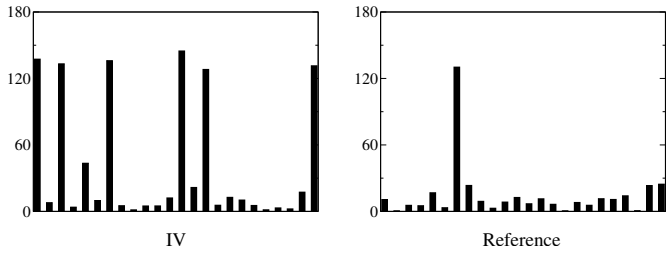


Figure 33: Conformers **IV** and **Reference**: Absolute deviations of the backbone torsion angles α to ζ of residues G5 to A8 from the ideal A-helix values in degrees.

the same structural motif when analyzed by visual inspection. Nonetheless, though being the conformer found closest to the experimentally determined structures, it contains six backbone torsion angles in disagreement with the reference structure.

For a combination of conformational search and energetic evaluation being a feasible method for structure prediction, two conditions must hold: First, the force field including the approximate solvation energy must be able to identify the correct structure as that of lowest energy. This condition is not fulfilled by the AMBER force field and most probably not by any other force field currently available. The severe discrepancies between the solvation energies yielded by the generalized Born approximation and finite difference Poisson calculations show that despite the partial success of the GB approximation, there is still ample need for improvement.

Second, the search space must be 'small' enough, so that the available computational resources allow enough different conformations to be sampled, minimized and evaluated. The only search method that guarantees the sampling of completely random conformations is distance geometry. However, the number of sterically possible but unrealistic structures obtained when no additional knowledge based distance constraints are imposed is beyond even the currently available computational resources. Conformational search based on 'local elevation' repeatedly yielded conformers of very low energy belonging to the 'family' of structures **Ia** and **Ib**. One might speculate that this 'family' represents the minimum energy conformation of GCAA tetraloops in a coarse grained sense. This can, however, not be proven and since all sampling experiments based on this method were started from only three different classes of structures, is impossible to estimate the bias introduced by the choice of the starting structure. Conformer **IV** was found only

once in a search run starting from the GNRA like conformation. This shows that 'brute force' conformational search *can* yield shapes close to the experimentally determined ones and of low energy (The heavy loop atom root mean square distance between conformer **IV** and the reference structure is only 0.75 Å). But, being a singular event it is inherently irreproduceable. Besides, conformer **IV** was not found due to its favorable energy, but during a search among *all* generated conformers for those with the lowest root mean square deviation from the reference structure.

These results do not invalidate the AMBER forcefield nor the generalized Born approximation. They only show that neither the force field nor the approximate treatment of solvent polarization effects are at the current state of development accurate enough to have any predictive capability in the absence of experimental input. Still, the relative energetic ranking of different conformers that are similar in a structural sense shows close agreement between the GB approximation and and the results obtained from the solution of the Poisson equation. As shown in sections 8.1.2 and 8.1.3 and, for example, by Williams et al. [112] and Tsui et al. [101], implicit solvent simulations do yield insights into the behavior and conformational preferences of nucleic acids at a fraction of the computational cost of MD simulations including the solvent explicitly. However, the results clearly indicate that attempts to observe the folding of RNA molecules in prolonged implicit solvent MD simulations starting from unfolded states ('the open chain') are bound to fail at the current status of development of implicit solvation models.

While the results presented here are somewhat disconcerting for force field users, they are of large interest for the force field developers and were gratefully accepted by the head of the group developing the AMBER force field.

8.2 Transitions of an Adenosine Bulge

Bulge containing RNA structures occur in biologically significant RNA molecules. Bulged bases in RNA helices are thought to play important roles in specific protein and drug binding to the RNA and RNA-RNA interactions [17, 28, 44, 100] and bulge induced kinks in RNA helices are significant in the tertiary folding of RNA [47, 117]. Structural studies using X-ray crystallography or NMR indicate that bulge nucleotides can adopt a variety of conformations, depending on bulge and neighboring nucleotides, solution or crystallization conditions and the presence of ligands. adenosine bases are the most commonly occurring unpaired bases in double helical RNA secondary structures [100]. NMR studies of DNA duplexes containing a bulged adenosine base have revealed that the extra adenosine stacks into the duplex [88]. In the genomic RNA of bacteriophage R17 a bulged adenosine base is stacked into the RNA duplex [17]. However, X-ray crystallographic studies on RNA revealed a looped out conformation of the bulged aenosine [22, 39]. Energy based conformational analysis of single base bulges in DNA and RNA conducted by Zacharias and Sklenar [118] also showed that a stacked conformations the energetically most favorable form for the adenosine bulge. However, this analysis also showed that a conformation of very low energy with the bulge base and the adjacent (5') base pair forming a base triple exists. Transitions between stacked in and looped out conformations of adenosine bulges have, to our knowledge, not been observed yet.

In order to investigate transitions between stacked in and looped out conformations, a duplex containing an unpaired adenosine with sequence r(GCGGCAC-CUGCC):r(GGCAG**A**GUGCCGC) was modeled based on the information available from the abstract of the publication of Thiviyanathan et al. [100]. After the end of the simulation described in the following, the complete publication and also the NMR structure in PDB format became available. From both it became obvious that the experimental structure had been synthesized with the strands in reverse order. This does, however, not influence the qualitative insights into the transitions of the bulged base. An additional simulation starting from the NMR structure also showed transitions between significantly distinct conformations of the bulged adenosine in general agreement with the results shown below.

To obtain a regular helix with a stacked-in adenosine bulge, a regular A-form helix with sequence r(GCGGCAC**U**CUGCC):r(GGCAG**A**GUGCCGC) was built using the `fd_helix` function contained in the NAB molecular mod-

eling language. The modeled structure was minimized to convergence, the placeholder residue U8 of strand one was deleted and the molecular topology was subsequently rebuilt so as to describe the target sequence. Subsequently the resulting structure was again minimized to convergence with distance restraints applied to all Watson-Crick base pairs to prevent structural distortions due to the strong local forces caused by the deletion of the placeholder residue. The final structure was minimized to convergence without any constraints. Starting from the resulting structure, an implicit solvent MD simulation was set up with all parameters as given for the tetraloop simulations described previously. The only difference in protocol were the use of a 20 Å cutoff necessary due to the computational demands inherent in simulations of large molecules, a 1 femtosecond timestep without any bond length constraints and a friction constant of 0.05 ps^{-1}, leading to slow but unbiased equilibration and an overall quasi-harmonic motion. Cutoffs in the range between 15 and 20 Å are commonly used in implicit solvent simulations based on the AMBER forcefield and the artifacts introduced by them seem to be below the inaccuracy inherent in implicit solvent simulations [35, 101]. Bond length constraints were not used since the RATTLE algorithm had not been implemented at the time the investigation was started. Figure 34 shows secondary and tertiary structure of the helix containing the bulged adenosine in stacked-in conformation. The bulge base is highlighted in red. The first five nanoseconds were somewhat arbitrarily defined as equilibration

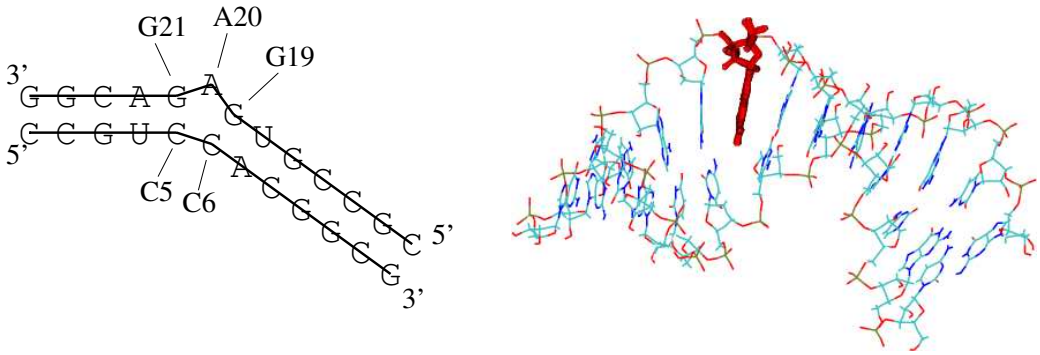


Figure 34: Secondary and tertiary structure. The bases most involved in the structural transitions are marked.

period to allow the relaxation of artifacts introduced by the modeling process. During this relaxation period, the stacked-in conformation changed to a partially looped out conformation with the bulge adenosine A20 forming a base triple with the paired bases G19 and C6. This base triple, shown in figure 35 is not perfectly planar, but is stabilized by several possible hydrogen bonds, shown as dashed lines. This metastable conformation is in agreement with the results of the energy based analysis of Zacharias and Sklenar [118], but has not been found either in NMR or X-ray investigations. During the subsequent twenty nanoseconds, again arbitrarily termed production run, several transitions between (partially) looped out and stacked-in conformation occurred. To localize these transitions, principal component analysis was applied to the snapshots. Due to the large amount of data, this was done in two steps. As a first step, the first principal component was calculated from the heavy atom positions of the bulged adenosine and the flanking bases along the entire production run trajectory, taking into account snapshots taken every 10 picoseconds. After identifying the time span between $t = 9980$ and $t = 13800$ nanoseconds as 'region of interest' the principal component analysis was repeated taking into account every snapshot inside this time interval. This process is shown in figure 36. It should be noted

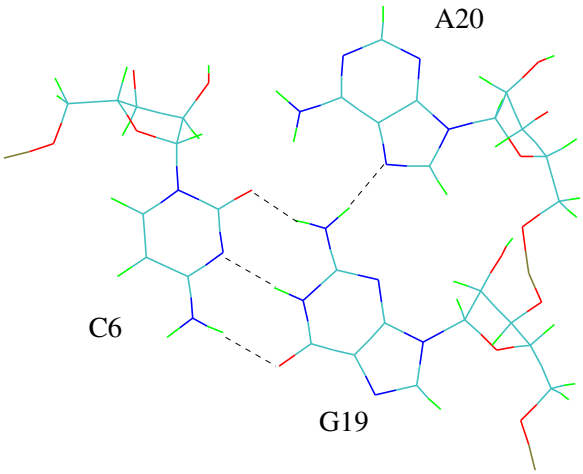


Figure 35: Base triple consisting of the bases C6, G19 and A20, taken from the minimized average structure over the time interval between $t = 12.5$ and $t = 13.0$ nanoseconds. Possible hydrogen bonds are shown by dashed lines.

that the principal components for both graphs were calculated from different sets of structures spanning *different* coordinate systems. Thus, the same substructures (defined by the heavy atom coordinates of residues G19, A20 and G21) do not correspond to the same coordinates in both graphs.

During the whole trajectory, the base-triple conformation is dominating and corresponds therefore to a first principal component close to zero in the upper graph. During the time interval visualized in the lower graph, base-triple and stacked-in conformation occur with roughly equal frequency and therefore correspond to first principal components with approximately equal absolute values, but of opposite sign. In both graphs shown, the 'upper state' represents the base-triple conformation, the 'lower state' to the stacked-in conformation. As can be seen from the coarse grained principal component analysis, the base-triple state is the more stable one *under the simulated conditions*.

The 'flexible' phase of the trajectory, visualized via the first principal component in the lower graph shown in figure 36 is preceded by an alpha/gamma flip in residue G19 at $t \approx 10.5$ nanoseconds. By this transition, the backbone torsion angles α-G19 and γ-G19 achieve values close to those found in ideal A-form helices. While this concerted torsion angle transition seems to initiate the subsequent transitions betweens base-triple and stacked-in conformation, it is not part of the actual transition of interest, since it occurs only once during the time interval between $t \approx 10.5$ nanoseconds and the end of the simulation at $t = 20$ nanoseconds.

Figure 37 shows the time course of α-G19 and γ-G19. Analysis of the time course of all backbone torsion angles of residues G19 to G21 leads to an interesting insight. The transition between base-triple and stacked-in conformation is essentially due to concerted shifts of the angles ϵ and ζ of residues G19 and A20. During the transitions, the phosphate groups of both residues stay in BI conformation ($-160^{\circ} < \epsilon - \zeta < +20^{\circ}$). Figure 38 shows residue A20 and the flanking base pairs G19-C6 and G21-C5, in base triple and stacked conformation. The backbone atoms defining the torsion angles ϵ and ζ of residues G19 and A20 are highlighted in blue. It is somewhat surprising that the distinctly different conformations shown in figure 38 are, apart from the highlighted angles, characterized by virtually the same torsion angle pattern. Bases C5 and C6 in the opposite strand show changes mainly in the glycosidic torsion angles, but these occur less rapidly than those around the bulged adenosine. Figure 39 shows the time course if these transitions, again during time span between $t = 9980$ ps and $t = 13800$ ps. The representa-

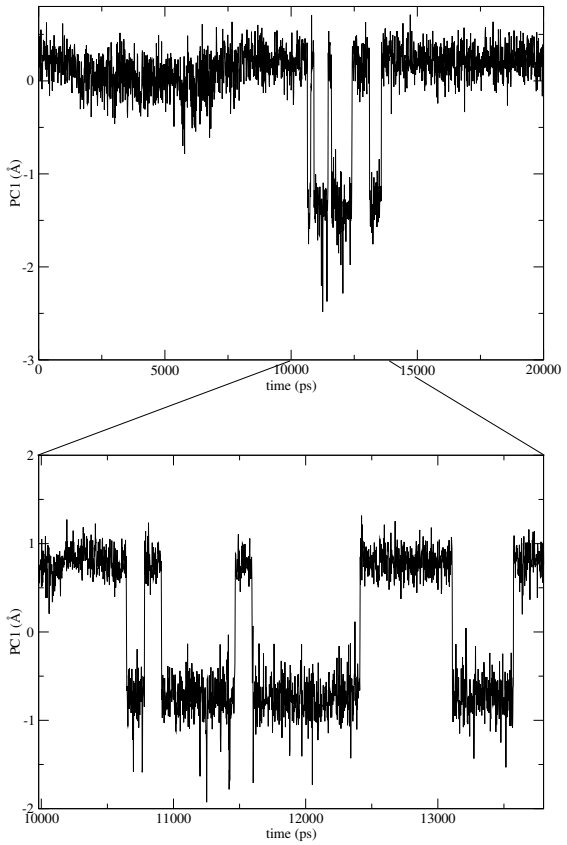


Figure 36: Two steps of principal component analysis, visualizing the transition between base-triple and stacked-in conformation. In both graphs, the larger values correspond to the base triple conformation, the smaller values the stacked-in conformation.

tion of the torsion angles is in this case smoothed by a running average over ten consecutive values to allow a clear distinction between short term fluctuations and significant changes. This does, however, render the transitions smoother than in the simulation. The sugar puckers of residues G19 to G21, C5 and C6 remained in C3'-endo conformation throughout the analyzed part of the trajectory. This is in agreement with the NMR structure submitted by Thiviyanathan et al. (`1k8s.pdb`), but not with the corresponding publication, where C2'-endo conformation is reported for the bulge adenine and the flanking bases. The stacked and base-triple conformations are in qualitative agreement with the results from energy based analysis [118], and with results

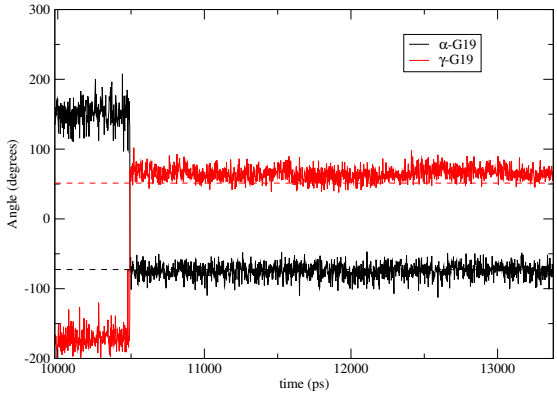


Figure 37: Time course of backbone torsion angles α-G19 and γ-G19. The sharp concerted transition marks the beginning of the repeated interconversion between stacked in and looped out conformation.

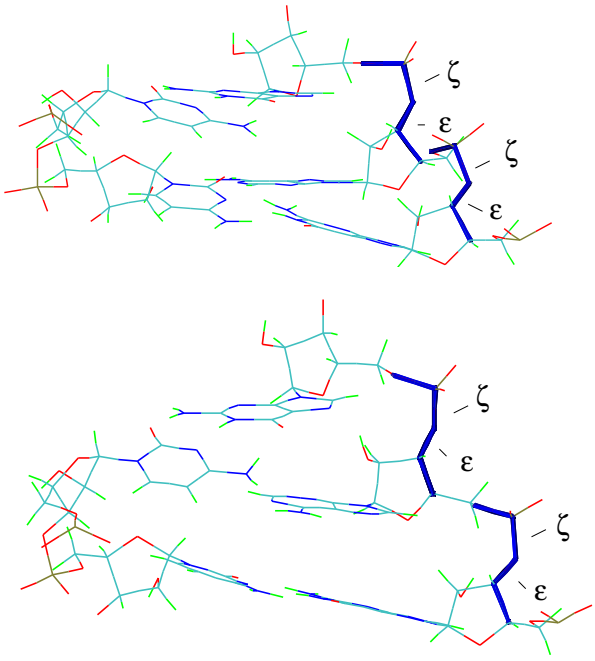


Figure 38: Adenine A20 and flanking base pairs G19-C6 and G21-C5. The bonds defining the backbone torsion angles ϵ and ζ are highlighted.

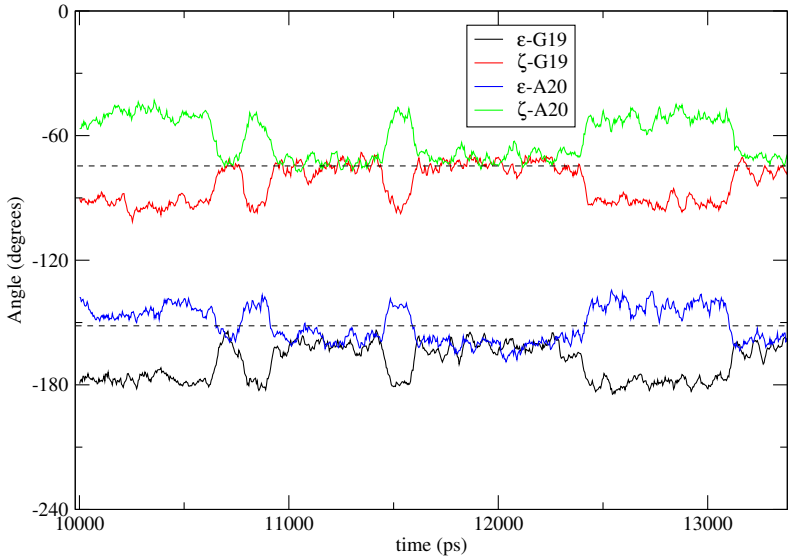


Figure 39: Time course of the concerted changes of the backbone torsion angles ϵ and ζ. The reference values from ideal A-form helices are shown by dashed lines.

obtained from the simulation of an adenine bulge in a DNA octamer [41] using the AMBER forcefield and explicit solvent. However, the finding of the base-triple state as the more stable one contradicts experimental evidence from NMR measurements, where single base bulges are found to be in stacked conformation [83,100]. Here, subtle differences between forcefields appear to have a strong influence on the results, since simulations performed with the CHARMM27 forcefield [42] showed a preference for the stacked conformation, at least in DNA [41]. Experimental data on DNA and double stranded RNA molecules with single bulge bases distinguish between structures where the bulge base is in stacking conformation (in solution, measured with NMR) and completely looped out with the flanking bases stacking on each other (in crystals, measured by X-ray crystallography). The transition from stacked conformation apparently favored in solution to an extended conformation, e. g. in nucleic acid-protein complexes is not yet well understood. While being the most stable state in the the simulation presented here, the base-triple state is actually probably a meta-stable transition state between the stacked and fully looped out state [41]. Therefore, the simulation presented here can be viewed as another valuable step towards the better understanding of nucleic acids and their interactions with other biomolecules.

The situation presented here can qualitatively compared to the simulations of tetraloop transitions presented previously: While the force field and the continuum approximation of the solvent may not be accurate enough to determine which of two energetically close states is the preferred one, detailed insights in the transitions between them can still be obtained at a fraction of the computational cost of explicit solvent simulations. In the case of the bulged adenine, the culprit for the discrepancy between experimental and simulation results is most probably not the continuum solvent approximation, since a similar preference was reported by Feig and Zacharias in simulations with explicit solvent and Particle Mesh Ewald treatment of long range electrostatic interactions.

After the release of the NMR structure described by Thiviyanathan et al., with sequence r(GGCAG**A**GUGCCGC): r(GCGGCAC-CUGCC) and A6 in stacked conformation, a simulation similar to the one described above in detail was set up as a consistency check. Indeed this simulation also showed a preference for the base-triple state with return to the stacked state a rare event (once in on over 20 nanoseconds simulation time). During this transition, bulge base A6 rotated about the glycosidic bond so as to enter the stack in *syn*-conformation. Since we are not aware of any experimental evidence for nucleic acids in quasi-regular helices (except Z-DNA) being in *syn*-conformation, this seems to add to the evidence that the AMBER force field does not assign correct energies to bases in *syn*-conformation. This potential 'problem' with bases in *syn*-conformation is somehow not surprising, since it is demanding enough to develop a set of force field parameters leading to behavior in accordance with experimental results for 'regular' nucleic acids [25, 30].

8.3 Modeling RNA pseudoknots

The initial motivation for modeling pseudoknot structures was to visualize the *possible* three dimensional structure of pseudoknot structures compatible with the base pairing pattern predicted by the dynamic programming algorithms developed and implemented by C. Haslinger [53]. Due to the computational demands of distance geometry calculations and force field based simulated annealing, the shortest sequence found to fold into a H-type pseudoknot was chosen. Sequence and predicted minimum free energy base pairing pattern were:

```
ACGGAUUGUGUCCGUAAUCACA
((((..[[[[.))))...]]]]
```

Figure 40 shows the base pairing pattern in a representation developed by Han et al [51]. What can be inferred from base pairing pattern and steric

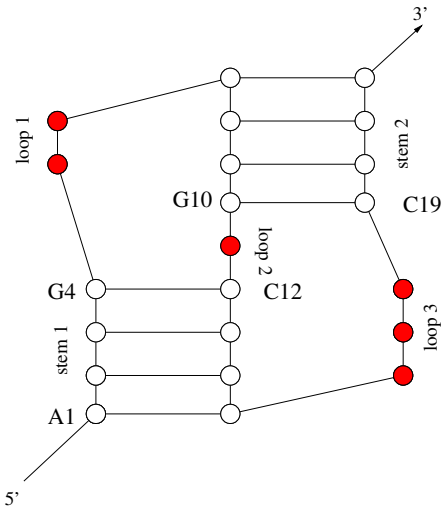


Figure 40: Annotated minimum free energy base pairing pattern of the RNA sequence r(ACGGAUUGUGUCCGUAAUCACA).

considerations? Assuming coaxial stacking, base C12 must stack on base G10 and base and, by the same token, base C19 must stack on base G4. Due to the short length of the single stranded regions (loop 1 and loop 3 in Han's notation), severe length constraints are imposed on the distances between bases G4 and U6, U15 and C19. The following procedure was chosen to obtain a *feasible* model: As a first step, all 1–2, 1–3 and 1–4 distance were

chosen as described in section 4.1.1. The two stacks were modeled separately as regular A-form helices using the `fd_helix` function included in the NAB programming language. All interatomic distances within each stack were set to the values from the ideal model helices. In the following, the two helices were *manually* positioned so as be in coaxial stacking position as inferred from the base pairing pattern. For each atom pair a_i, a_j with a_i being part of stem 1 and a_j being part of stem 2, the corresponding distance d_{ij} was set to lie within 80% and 120% of the distance taken from the manually aligned model structure to allow some certain variation in the relative stack positions. From this set of distances, 100 structures were generated as described in section 4.1.1. The trial structures were each minimized for 1000 steps of conjugate gradient minimization, using the AMBER99 force field and the generalized Born approximation[7]. The ten structures of lowest initial energy were further refined by five cycles of simulated tempering with heavy atoms of the stack bases constrained to their respective starting positions to allow mainly the single stranded parts to relax any initial local mis-configurations. The simulated tempering cycles consisted of 100 timesteps in contact with a Langevin heat bath at 350 Kelvin followed by a cooling phase of 500 timesteps at 200, 150 and 100 Kelvin each, followed again by 100 steps of conjugate gradient minimization.

The lowest energy structure obtained was subsequently refined by greedy simulated tempering, this time with the heavy atoms of the stacked bases restrained to their initial positions by a harmonic potential with harmonic restraints of 10.0 kcal/(mol Å^2), until no further reduction of force field energy could be observed. The final structure is shown in figure 41, from three different perspectives. The second and third representations correspond to the first one rotated by -90° and -180° around the positive z-axis (pointing toward the top of the page). The final structure is compact, shows coaxial stacking of stem 1 and stem 2 as well as stacking interaction in a single stranded region (bases A17 and G18 from loop three, cf. figure 40) and preserved base pairing and stacking pattern during 100 picoseconds of implicit solvent MD.

While the model described above shows characteristic features of experimentally determined pseudoknot structures, it cannot be claimed that the

[7]It should be noted that for *generated* molecular structures, the force field energy obtained after 1000 steps of CG minimization can only give a crude estimate at best for the quality of the structure, but in the absence of a corresponding experimental structure, it is the only measure available.

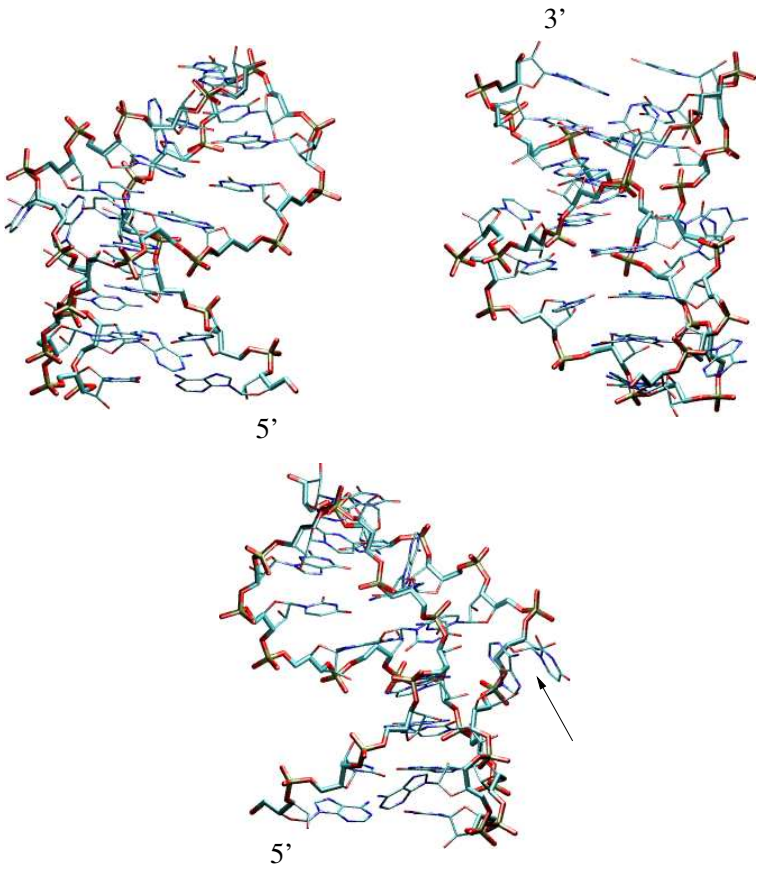


Figure 41: Pseudoknot model corresponding to the minimum free energy base pairing pattern of sequence r(ACGGAUUGUGUCCGUAAUCACA) shown from three perspectives. The 5' and 3' ends of the backbone are annotated where they do not overlap with other parts of the structure. Stacking bases A17 and U18 are marked by an arrow.

process described above leads to successful structure *prediction*. One possible flaw is the position of bases A17 and U18. These bases point towards the outside of the molecule and the phosphate group of base U18 points towards the inside of the molecule, leading to an unfavorable closeness of the backbones of loop 3 and stem 1. One might speculate that in the actual structure, the backbone of loop 3 is exposed to the solvent and bases A17 and U18 possibly stack in an orientation similar to the bases of stem 1.

To test, how close a pseudoknot structure modeled using common sense, experience and distance geometry can come to an experimentally determined structure, the crystal structure of a ribosomal viral frame-shifting pseudoknot, determined by Su et al. [98] (PDB ID `437d`) was chosen as reference. The base pairing pattern predicted by the dynamic algorithm developed by Christian Haslinger differed from the actual pattern found in the crystal structure. Therefore in the following, the sequence, the predicted base pairing pattern, and the one used as modeling constraint are shown:

Sequence	`GCGCGGCACCGUCCGCGGAACAAACGG`
Predicted	`.(((((..[[[[)))))......]]]]`
Model constraint	`.(((((..[[[.)))))......]]]`

It should be noted that assigning a base pairing pattern to the crystal structure is somewhat ambiguous, since bases C7, G11 and C25 actually form a base triple. Figure 42 shows the pseudoknot in Han's representation. The triple interaction between bases C7, G11 and C25 is signified by dashed lines. Distance restraints inferred from base pairing pattern, but assuming no knowledge of the crystal structure were: All consecutive nucleotides from stems one and and two adopt A-form conformation and base U13 stacks on base G11. Furthermore, it was assumed that the *inner* nucleotides of the long single stranded region termed loop 3 also stack on each other in 'close to A-form' conformation. Therefore base-base distances available from the distance geometry data base included in the NAB program package for regular A-form helices were imposed on the heavy atoms of every consecutive pair of bases from residue A19 to A23. All other distances were set as described in section 4.1.1. From these distances, 100 trial structures were generated by embedding. Since harmonic position restraints were found to hinder relaxation during simulated tempering in the previously described modeling process, a modified molecular mechanics energy function was implemented and used. This function allows the flexible addition of arbitrary harmonic (parabolic) or funnel shaped (hyperbolic) distance restraints to the force field energy and was built into the NAB program package.

During the subsequent simulated tempering runs, the base pairing and stacking distance restraints were this way added to the force field energy, allowing MD based conformational search while preventing the disruption of known or assumed stacks and base pairs. The 100 trial structures were each minimized by restrained simulated annealing and 100 steps of conjugate gradient minimization. The lowest energy conformation obtained this way

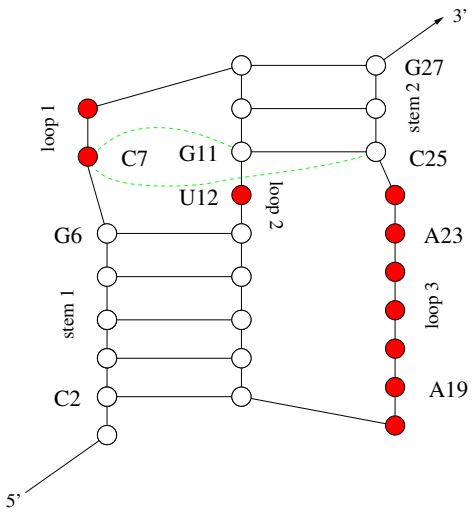


Figure 42: Representation of the minimum free energy base pairing pattern of sequence r(GCGCGGCACCGUCCGCGGAACAAACGG). The base triple formed by bases C7, G11 and C25 is signified by dashed lines.

served as the starting structure for the final simulated tempering run. For this simulated tempering run, the mass of hydrogen atoms was set to 5 atomic mass units to allow a *nominal* timestep of 2 femtoseconds. All interatomic distances for pairs of heavy atoms being part of the same base pair were restrained to their initial distances by harmonic restraints of 50 kcal/(mol Å), rendering the base pairs rigid while preserving overall stack flexibility. For 1600 cycles, the restrained structure was brought in contact with alternating Langevin heat baths of *nominal* temperatures of 400 and 200 Kelvin followed by 1000 steps of unrestrained conjugate gradient minimization. After minimization, each structure was stored on disk.

While each of the obtained structures is close to a local minimum of the forcefield, they are 'shock frozen' molecular dynamics snapshots and the final energies therefore again give only a crude estimate of the quality of the structures. For technical reasons (see section 4.5), the resulting structures were again minimized using the JUMNA program (with the GB method for approximate treatment of solvent polarization effects) for final energetic evaluation. The success of the simulated tempering protocol in leading to a low energy structure in at least partial agreement with the X-ray structure is surprising. The orientation of bases G1, U12 and G18 is probably due to the crystal environment, as is the conspicuous propeller-twist and tilt of the

base pairs in stems 1 and 2 [98]. Bases G1, U12 and G18 are depicted in black in figure 43. However, omitting these residues while optimally superimposing X-ray and lowest energy structure, the heavy atom rmsd over all other residues is only 2.8 Å. Furthermore, the bases of loop 3 adopted stacking conformation during the minimization process in agreement with experiment, albeit with a slightly different orientation with respect to stem one and with several (among them the *pyrimidine* (!) C21) in syn-conformation.

This is the more surprising when comparing the best minimized structure with the initial structure, where the bases of stem 3 show no stacking and the relative position of stems one and two is in strong disagreement with both final and experimental reference structure. The starting structure, the final structure after minimization and the X-ray structure are shown in figure 43. Finally, figure 44 shows the X-ray structure and the lowest energy structure optimally superimposed while neglecting residues G1, U12 and G18.

While leading to good qualitative agreement with the experimental structure *in the case shown*, the applied approach has serious drawbacks:

Since the AMBER force field is generally used to model solution conditions, the choice of an X-ray structure as reference was not optimal. However, structure `437d` was the smallest pseudoknot available in the PDB data base at the time of this study.

Due to computational limitations, building all-atom molecular models using distance geometry is limited to structures of sizes not much larger then the molecule shown here.

The applied simulated annealing protocol was chosen based on experience and educated guess. The number of possibly relevant parameters and the CPU time necessary to estimate the success of the protocol prohibit a systematic optimization of parameters.

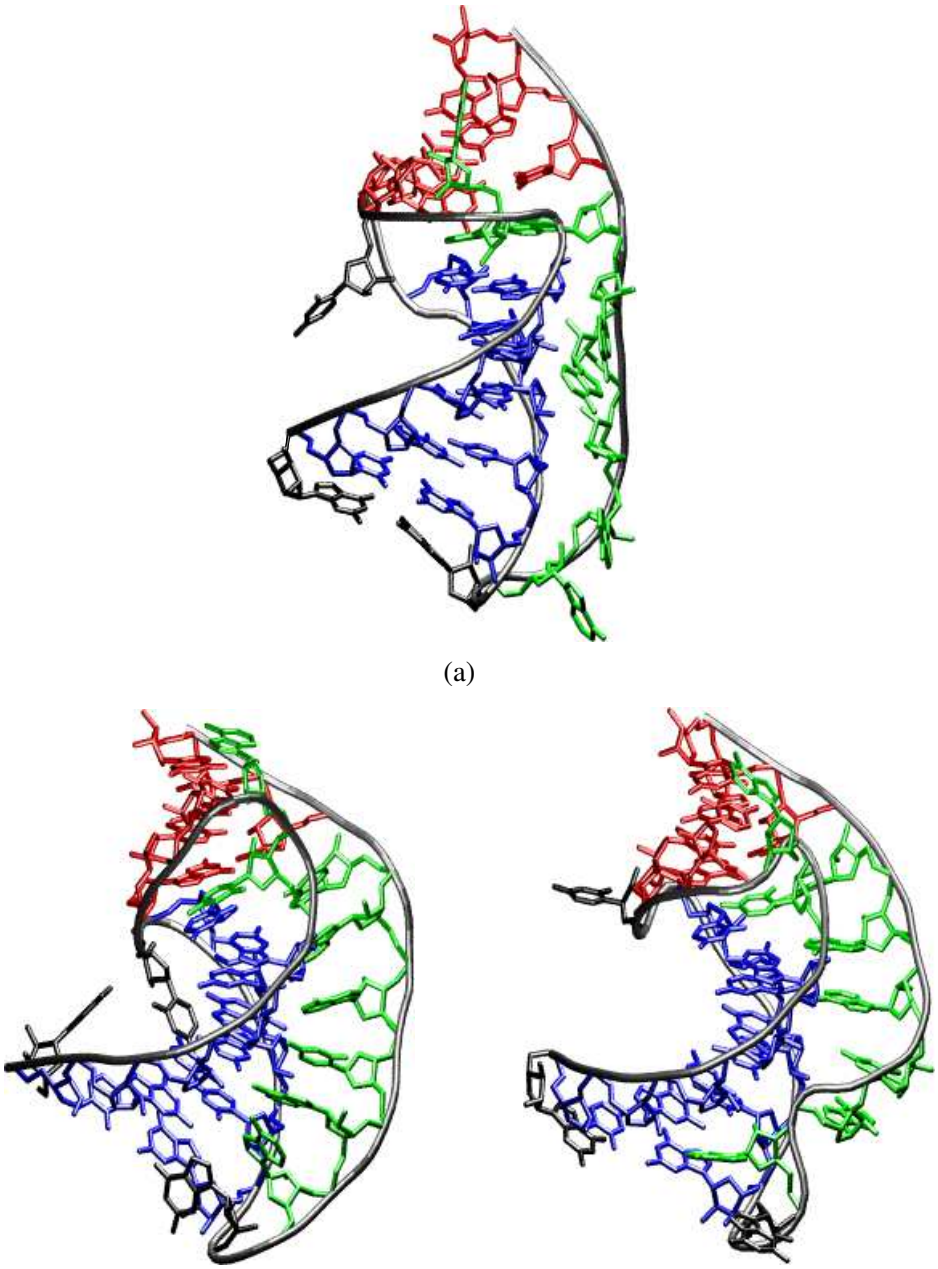
The failure of the AMBER force field to assign a uniquely low energy to the 'right' GNRA tetraloop structure shows that even successful minimization does not necessarily lead to structures in agreement with experiment.

The generalized Born approximation is at the current state of development not accurate enough to yield correct differences of solvation energies for molecules containing deeply buried atoms.

The most striking drawback is the necessary amount of CPU time. The 1600 cycles of simulated tempering and preliminary minimization took about three months (!) of CPU time on an AMD Athlon processor running at 800 MHz. Based on the reasoning leading to the initial distance constraints, a semi-manual approach using a molecular editing program and an iterative

protocol combining manual intervention and force field relaxation might have led to a result at least equally close to the experimental structure in shorter time.

In the final analysis, the above drawbacks notwithstanding, it should be reiterated that the applied protocol led to good qualitative agreement between modeled and experimental structure. The experience gained from this study may well be useful in future modeling efforts.



(a)

(b) (c)

Figure 43: Initial structure from the simulated tempering run (a), lowest energy structure found (b) and crystal structure from `437d.pdb` (c). Hydrogen and phosphate oxygen atoms are omitted for clarity. The course of the backbone is signified by a spline.

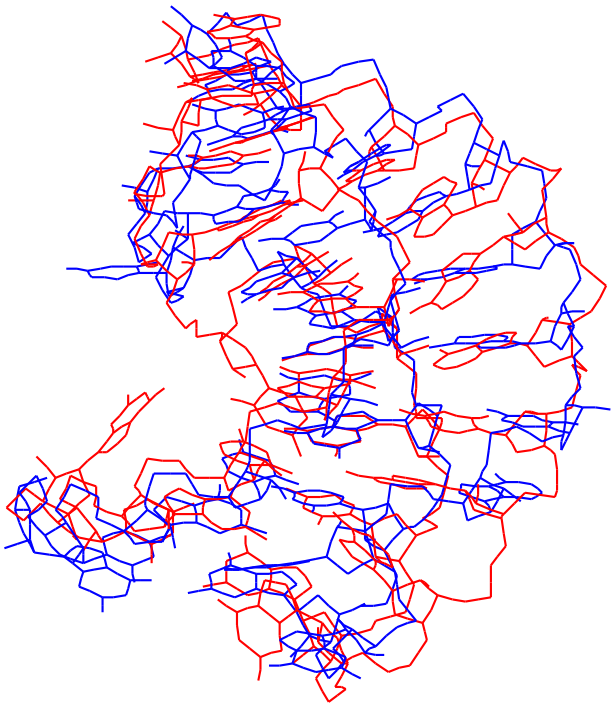


Figure 44: Optimally superimposed X-ray structure (in blue) and modeled structure (in red).

9 Conclusion and Outlook

9.1 Conclusion

In this thesis, structure and dynamics of some functionally important and evolutionarily conserved structural motifs, namely GNRA and UUCG tetraloops, were investigated by molecular dynamics simulations based on the AMBER force field. The simulations were performed with the generalized Born implicit solvation model, as well as with explicit consideration of the surrounding solvent and neutralizing counterions.

Simulations based on the implicit solvent model allow for the computation of longer trajectories due to reduced computational demands, as well as a faster sampling of accessible conformations due to the absence of solvent friction. Only this relatively young technique made the observation of repeated transitions between distinctly different experimentally determined conformations possible.

On the other hand, simulations with explicit inclusion of surrounding solvent in principle provide a higher level of accuracy. Reassuringly, some structural transitions observed occurred with as well as without explicitly included solvent. The simulation results reported for a UUCG tetraloop comprise, to our knowledge, the first direct and detailed comparison of implicit and explicit solvent simulations for a non-helical RNA structural motif. The comparison between both simulations and the experimental reference structure showed excellent agreement.

Further investigations were undertaken to find out whether the conformation of an extrastable structural motif, namely a GCAA tetraloop, could be predicted by conformational search and enthalpic ranking. The results from this investigation are both unsatisfactory as well as of practical use: Some distinct classes of loop conformations with force field energies lower than any conformation found in the structural vicinity of any experimentally determined reference structure were found. This shows the limits of currently available force fields, but is, by the same token, valuable for the further improvement of the AMBER force field.

Simulations of a helix containing an adenosine-bulge partially elucidated the transition between environment dependent different conformations of the bulged base. The results from these simulations are in good agreement with results reported from DNA helices containing adenosine bulges. However, in DNA and RNA the partially looped out transition state with the bulged

adenosine forming a base triple with the adjacent base pair in 5' direction is favored over the 'stacked in' conformation, as observed in solution. This again points to the limitations of the AMBER force field and has been reported to its developers.

Finally, two pseudoknot structures were modeled. The modeling process was based on sequence, secondary structure and 'experience', i.e. knowledge gained from the visual inspection of experimentally determined structures. During structure creation, distance geometry as well as manual intervention for the approximately correct relative orientation of coaxially stacking helices were used for the first structure, of which the actual structure is unknown. Since the second structure had been previously determined by X-ray crystallography, no manual intervention was used in the modeling process. After extensive force field based minimization by restrained simulated tempering, a conformation in surprisingly good *qualitative* agreement with the experimentally determined reference structure was obtained.

During this work, many routines useful for force field aided molecular modeling, molecular dynamics and coarse grained analysis of molecular dynamics trajectories were implemented. Some of them are based on algorithms taken from literature, some of them were adopted from other molecular modeling software packages, again some of them newly devised. A significant part of these routines has become part of the official distribution of the freely available and widely used Nucleic Acid Builder molecular modeling software package, version 4.5, maintained by the head of the AMBER force field development group.

9.2 Outlook

Force fields and methods relying upon them are under continuous development. We are looking forward to an improvement of parameterization, as well as to an improvement of the generalized Born approximation, so as to allow a more accurate implicit consideration of solvent polarization effects. Some preliminary analytical results that may be of use for improving the accuracy of the generalized Born approximation are given in the appendix.

As for structure prediction, we believe that there are several approaches to be tested:

(i) The development of a computer program allowing interactive manipulation of molecular (sub)structures *and* force field based structure optimization iteratively.

(ii) The improvement of promising coarse grained models, as developed for example by Kurt Grünberger during his PhD thesis [48]. Since this *pseudo atom* model contains significantly less atoms than 'real' RNA molecules, many approaches currently in use in conjunction with 'all atom' models, such as for example distance geometry, require less computational resources and allow the refinement of larger structures.

Last but not least, the ever increasing number and quality of structures determined by experimental methods will without doubt increase our understanding of structure and function of RNA molecules as well as aid further efforts aimed at ultimately *predicting* the three dimensional structure of RNA.

A Calculations

Energy Conservation by Verlet Integration

A single harmonic oscillator of mass m, 'spring constant' C and angular eigenfrequency $\omega = \sqrt{\frac{C}{m}}$ obeys the following equation of motion:

$$\ddot{x}(t) = -\omega^2 x(t) \tag{56}$$

Velocity $v(t)$ and acceleration $a(t)$ are denoted by $\dot{x}$ and $\ddot{x}$ respectively. Further, we denote the finite timestep by Δ. Then, the velocity Verlet recursion for the numerical integration of equation 56 takes the following form:

$$\begin{aligned} x(t+\Delta) &= x(t) + \dot{x}(t)\Delta + \ddot{x}(t)\frac{\Delta^2}{2} \\ \dot{x}(t+\Delta) &= \dot{x}(t) + \frac{\Delta}{2}(\ddot{x}(t) + \ddot{x}(t+\Delta)) \end{aligned} \tag{57}$$

Defining a state vector $\mathbf{X} = (\omega x, \dot{x})^T$, the total energy of the oscillator can be written as

$$E(t) = \frac{m}{2}\mathbf{X}(t)^T\mathbf{X}(t) \tag{58}$$

and equations 56 and 57 lead to the following relation between the state vectors $\mathbf{X}_\mathrm{n}$ and $\mathbf{X}_{\mathrm{n+1}}$:

$$\mathbf{X}_{n+1} = \mathbf{M}\mathbf{X}_n \tag{59}$$

with

$$\mathbf{M} = \begin{pmatrix} 1 - \frac{\omega^2\Delta^2}{2} & \omega\Delta \\ -\omega\Delta + \frac{\omega^3\Delta^3}{4} & 1 - \frac{\omega^2\Delta^2}{2} \end{pmatrix} \tag{60}$$

Defining the angle $\phi = \arccos(1 - \frac{\omega^2\Delta^2}{2})$, $\mathbf{M}$ can be rewritten as

$$\mathbf{M} = \begin{pmatrix} \cos(\phi) & \frac{\sin(\phi)}{\cos(\frac{\phi}{2})} \\ -\sin(\phi)\cos(\frac{\phi}{2}) & \cos(\phi) \end{pmatrix} \tag{61}$$

$\mathbf{M}$ can again be rewritten as:

$$\mathbf{M} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & \cos(\frac{\phi}{2}) \end{pmatrix}}_{\mathbf{D}} \underbrace{\begin{pmatrix} \cos(\phi) & \sin(\phi) \\ -\sin(\phi) & \cos(\phi) \end{pmatrix}}_{\mathbf{R}} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\cos(\frac{\phi}{2})} \end{pmatrix}}_{\mathbf{D^{-1}}} \tag{62}$$

Thus, the state $\mathbf{X}_n$ after n discrete time-steps can be obtained from the initial state vector $\mathbf{X}_0$ by

$$\mathbf{X}_n = \mathbf{D}\mathbf{R}^n\mathbf{D}^{-1}\,\mathbf{X}_0 \tag{63}$$

and the total energy after n timesteps is given by:

$$\begin{aligned} E_n &= \frac{m}{2}\,\mathbf{X}_n^T\mathbf{X}_n = \\ &= E_0 - \frac{m}{2}\left(\sin^2(n\phi)\left(\omega^2 x_0^2\sin^2(\frac{\phi}{2}) - \dot{x}_0^2\tan^2(\frac{\phi}{2})\right)\right) + \\ &\quad + \frac{m}{2}\,\omega x_0\dot{x}_0\frac{\sin^2(\frac{\phi}{2})}{\cos(\frac{\phi}{2})}\sin(2n\phi) \end{aligned} \tag{64}$$

Therefore, the total energy of the harmonic oscillator is a periodic function of the number of steps n and *its average* is exactly conserved to machine precision, when the equation of motion is integrated via the velocity Verlet algorithm. It should be noted that in general, conservation of total energy by Verlet integrators is a direct consequence of the time reversibility of the iteration defined by equation 57 [93], but the example presented above allows an analytical description of the short-time fluctuations.

Pairwise Descreening - Derivation and Generalization

Within the pairwise descreening approximation, inverse effective Born radii of atoms are obtained as follows: Let R_i be the intrinsic (e.g. Van der Waals) radius of atom i and S_j the intrinsic radius of atom $j, j \neq i$. Then the inverse effective Born radius a_i^{-1} of atom i is defined by:

$$\frac{1}{a_i} = \frac{1}{R_i - \beta} - \sum_{j \neq i} \frac{1}{4\pi} \underbrace{\int_{V_j \setminus V i} \frac{\mathrm{d}V}{|\vec{r} - \vec{r}_i|^4}}_{\mathbf{I_{ij}}} \tag{65}$$

The resulting integral $I_{ij}(d; R_i; S_j)$ is a continuous differentiable function of

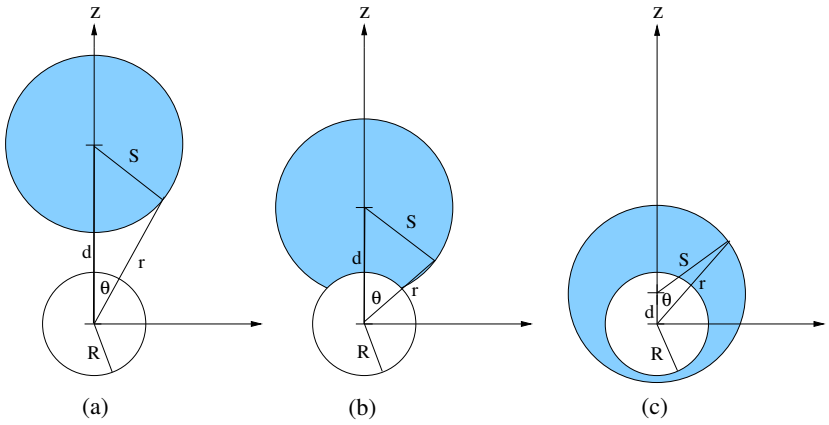


Figure 45: The geometric situation for the descreening integrals. The integration volumes are shown in gray.

the distance d, depending in the radii R_i and S_j as parameters. The result was given by Schaefer and Froemmel [92] without derivation. Preliminary studies on the improvement of the agreement between effective radii obtained by Poisson Boltzmann calculations and those obtained by a modified pairwise descreening method indicate that it might be preferable to represent the descreening atoms by a non-uniform density function $\rho(\vec{r})$. Therefore, the following derivation is given for the descreening atoms represented by a density function $\rho(\vec{r})$, that is symmetric with respect to the center of sphere j. In the following, sphere i is without loss of generality placed on the origin of the coordinates system and sphere j is places with its center at $(0, 0, d)^T$.

Then, the law of cosines yields

$$\cos(\theta(r)) = \frac{r^2 + d^2 - S^2}{2dr}$$

and for case (a) shown in figure 45, I_{ij} can be written as

$$\begin{aligned} I_{ij}(d) &= \int\limits_{V_j \setminus V_i} \frac{\rho(\vec{r})}{r^4} \,\mathrm{d}V = \\ &= 2\pi \int\limits_{d-S}^{d+S} \left(\int\limits_{0}^{\theta(r)} \frac{\rho(r,\theta)}{r^4} \,\sin(\theta)\,\mathrm{d}\theta \right) r^2 \mathrm{d}r = \\ &= 2\pi \int\limits_{d-S}^{d+S} \left(\int\limits_{\frac{r^2+d^2-S^2}{2dr}}^{1} \frac{\rho(r, \arccos(u))}{r^2} \,\mathrm{d}u \right) \mathrm{d}r. \end{aligned} \tag{66}$$

In the case of a uniform density function $\rho(\vec{r}) = 1$, this simplifies to

$$\begin{aligned} I(d) &= 2\pi \int\limits_{d-S}^{d+S} \frac{1}{r^2} - \frac{d^2 + r^2 - S^2}{2dr^3} \,\mathrm{d}r = \\ &= 2\pi \left(\frac{S}{d^2 - S^2} + \frac{1}{2d} \log \frac{d-S}{d+S} \right). \end{aligned} \tag{67}$$

For case (b) shown in figure 45, the derivation is similar, only the bounds of the radial integral extend from R to $d + S$. This results in

$$I(d) = 2\pi \left(\frac{1}{R} - \frac{1}{d+S} + \frac{1}{4d} \frac{d-S}{d+S} - \frac{d^2 - S^2}{4dR^2} + \frac{1}{2d} \log \frac{R}{d+S} \right) \tag{68}$$

Case (c) requires a reverse order of integration [8]. In this case, the law of cosines yields

$$r(\theta) = d\cos(\theta) + \sqrt{S^2 - d^2 \sin(\theta)^2}$$

[8] Or, alternatively, splitting the integration volume, which requires more imagination, but leads to more simple intermediate expressions

and the required integral is given by

$$I(d) = \int_0^{\pi} \left(\int_R^{r(\theta)} \frac{\rho(\vec{r})}{r^2} \, \mathrm{d}r \right) \sin(\theta) \, \mathrm{d}\theta \tag{69}$$

Again, for a uniform density ($\rho(\vec{r}) = 1$), the result is:

$$I(d) = 2\pi \left(\frac{2}{R} + \frac{S}{d^2 - S^2} + \frac{1}{2d} \log \frac{S-d}{S+d} \right) \tag{70}$$

Due to steric limitations on interatomic distances and intrinsic atomic radii, case (c) never occurs within the unmodified pairwise descreening model. For any pair if radii R and S and all possible distances d, The above results can be summarized as follows, ordered with increasing distance d:

$$I(d;R;S) = \begin{cases} 2\pi \left(\frac{2}{R} + \frac{S}{d^2 - S^2} + \frac{1}{2d} \log \frac{S-d}{S+d} \right), \\ (R < S) \wedge (d < S - R) \\ \\ 2\pi \left(\frac{1}{R} - \frac{1}{d+S} + \frac{1}{4d} \frac{d-S}{d+S} - \frac{d^2 - S^2}{4dR^2} + \frac{1}{2d} \log \frac{R}{d+S} \right), \\ (d \geq |R - S|) \wedge (d < S + R) \\ \\ 2\pi \left(\frac{S}{d^2 - S^2} + \frac{1}{2d} \log \frac{d-S}{d+S} \right), \\ d \geq R + S \end{cases} \tag{71}$$

The expressions given in equation 71 are costly in computational terms and they, as well as their derivatives with respect to the interatomic distance d, have to be computed for each atom pair, making implicit solvent calculations significantly slower than 'vacuum' calculations with a possibly modified Coulomb law. The function $I(d;R;S)$ has a rather serious drawback: its second derivative with respect to d is discontinuous where spheres i and j 'touch' each other, i.e. at $d = R + S$ and $d = S - R, S > R$. Tsui and Case reported a slight upwards drift in total energy during implicit solvent simulations of an NVE ensemble and ascribed the drift to problems with the Verlet integration [102]. The origin of the drift, however, is the presence of kinks in the interatomic forces, due to the properties of $I(d)$. Figure 46 shows $I(d)$ and its derivative $I'(d)$. As an example of a non-uniform density, we choose

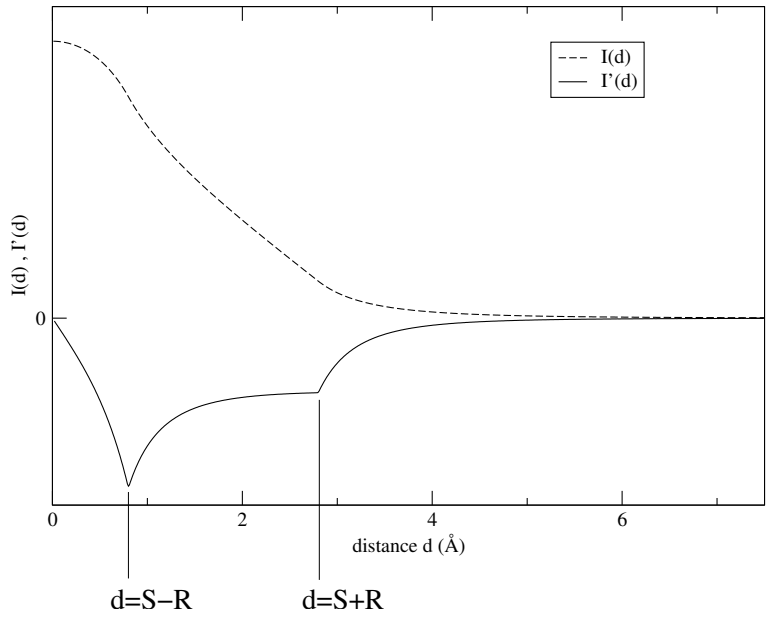


Figure 46: Function $I(d)$ (dashed line) and its derivative (full line) for arbitrary radii R and S, $S > R$.

a parabolic density for representing the descreening atoms. Let s be the distance to the center of the 'descreening' sphere with radius S (see figure 45). Then, The density function to be integrated over is given by

$$\rho(s) = \begin{cases} \lambda(1 - \frac{s^2}{S^2}), & s \leq S \\ 0, & s > S \end{cases}$$

Choosing a geometric setting as shown in figure 45, the law of cosines yields

$$s^2 = r^2 + d^2 - 2dr\cos\theta$$

and the actual integration is analogous to the uniform density case, but slightly more tedious. The resulting function $I(d)$ is shown below in equation 72:

$$I(d;R;S) = \begin{cases} 2\pi\lambda\left(\frac{3d^2-S^2}{2dS^2}\log\frac{S-d}{S+d} + \frac{2}{RS^2}(R^2+S^2) - \frac{2d^2}{RS^2} - \frac{3}{S}\right), \\ (R<S)\wedge(d<S-R) \\ \\ \frac{\pi\lambda}{4dR^2S^2}\left((d^2-S^2)^2 - 8d^3R - R^4 + 4dR(2R^2-3RS+2S^2)\right) - \\ -\frac{\pi\lambda(3d^2-S^2)}{dS^2}\log\frac{R}{d+S}, \\ (d\geq|R-S|)\wedge(d<S+R) \\ \\ 2\pi\lambda\left(\frac{3d^2-S^2}{2dS^2}\log\frac{d+S}{d-S} - \frac{3}{S}\right), \\ d\geq R+S \end{cases} \tag{72}$$

The results shown above have not yet been tested in conjunction the the AMBER force field and the GB approximation. The complexity of the expressions in equation 72 is not a limiting factor with respect to computational speed, since only a small fraction of all atom pairs is closer than the sum of their respective radii and the third expression in equation 72 is hardly more complex than the corresponding expression for the uniform density case (cf. equation 71). The ultimate goal of choosing a non-uniform density and extending the radius of the integration volume is to better approximate the volume integral over the molecular interior by filling the intermolecular crevices left out by the integration over the Van der Waals volume with the overlapping densities. Preliminary investigations have shown that this approach is worth pursuing in future investigations. Figure 47 show the function $I(d)$ and its derivative $I'(d)$ for arbitrarily chosen radii R and S , $S > R$. As might be inferred from physical intuition, the vanishing of the density function at the boundary of the integration volume yields a twice continuously differentiable function $I(d)$, which is -in principle- better suited for force field applications than the standard form shown in equation 71 and figure 46. Since no optimization with respect to radius S and parameter λ has yet been attempted, the scaling is arbitrary.

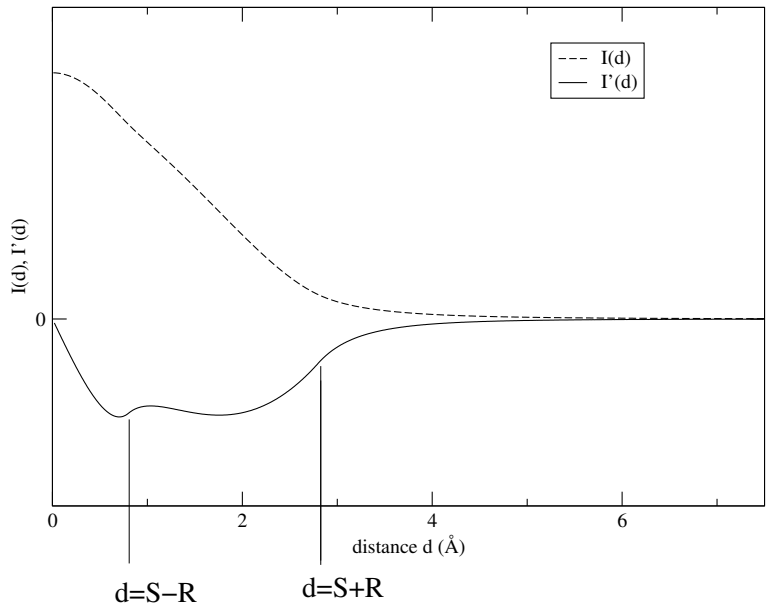


Figure 47: Function $I(d)$ (dashed line) and its derivative (full line) for arbitrary radii R and S, $S > R$. The use of a parabolic density function leads to a smooth derivative.

List of Figures

List of Tables

References

[1] F. H. T. Allain and G. Varani. Structure of the P1 helix from group I self-splicinig introns. *J. Mol. Biol.*, 250:333–353, 1995.

[2] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Clarendon Press, Oxford, 1987.

[3] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen. Essential dynamics of proteins. *PROTEINS: Struct. Func. Gen.*, 17:412–425, 1993.

[4] A. Amadei, A. B. M. Linssen, D. L. de Groot, D. M. F. van Aalten, and H. J. C. Berendsen. An efficient method for sampling the essential subspace of proteins. *J. Biomol. Str. Dyn.*, 13:615–625, 1996.

[5] H. C. Andersen. Rattle: a 'velocity' version of the SHAKE algorithm for molecular dyamics calculations. *J. Comp. Phys.*, 52:24–34, 1983.

[6] V.P. Antao, S.Y. Lai, and I. Tinoco. A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucl. Acid Res.*, 19:5901–5905, 1991.

[7] V.P. Antao and I. Tinoco. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucl. Acid Res.*, 20:819–824, 1992.

[8] D. A. Atkinson. Ccmath: A mathematics library. Online, 2 2000. URL: `http://freshmeat.net/redir/ccmath/1083/url_tgz/ccmath-2.2.1.tar.gz` .

[9] P. Auffinger and E. Westhof. *Encyclopedia of Computational Chemistry*, volume 5, chapter Molecular Dynamics: Simulations of Nucleic Acids, pages 1629–1640. Wiley and Sons, 1998.

[10] E. Barth, K. , Kuczera, B. Leimkuhler, and Skeel R. D. Algorithms for constrained molecular dynamics. *J. Comp. Chem.*, 16(10):1192–1209, 1995.

[11] E. Barth and T. , Schlick. Overcoming stability limitations in biomolecular dynamics. i. combining force splitting via extrapolation with langevin dynamics ln. *J. Chem. Phys.*, 109(5):1617–1632, 1998.

[12] D. Bashford and D. A. Case. Generalized born models of macromolecular solvation effects. *Ann. Rev. Phys. Chem.*, 51:129–152, 2000.

[13] Donald Bashford. An object-oriented programming suite for electrostatic effects in biological molecules. In Yutaka Ishikawa, Rodney R. Oldehoeft, John V. W. Reynders, and Marydell Tholburn, editors, *Scientific Computing in Object-Oriented Parallel Environments*, volume 1343 of *Lecture Notes in Computer Science*, pages 233–240, Berlin, 1997. ISCOPE97, Springer.

[14] D. Beeman. Some multistep methods for use in molecular dynamics calculations. *J. Comp. Phys.*, 20:130–139, 1976.

[15] H. Bekker. *Molecular Dynamics Simulation Methods Revised*. PhD thesis, Rijksuniversitit Groningen, 1997.

[16] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and A. Di Nola. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[17] P. N. Borer, Y. Lin, S. Wang, M. W. Roggenbruck, J. M. Gott, O. C. Uhlenbeck, and I. Pelczer. Proton NMR and structural features of a 24-nucleotide RNA-hairpin. *Biochemistry*, 34:6488–6503, 1995.

[18] M. Born. Volumen und hydratationswärme der ionen. *Z. Phys.*, 1:45, 1920.

[19] J.W. Brown, J. M. Nolan, E. S. Haas, M. A. T. Rubio, F. Major, and N. R. Pace. Comparative analysis of ribonuclease R RNA using gene sequences from natural microbial populations reveals tertiary structural elements. *Proc. Natl. Sci. USA*, 93:3001–3006, 1996.

[20] Nir Carmi, Shameelah R. Balkhi, and Ronald R. Breaker. Cleaving DNA with DNA. *Proc. Natl. Acad. Sci. USA*, 95:2233–2237, 1998.

[21] D. A. Case, D. A. Pearlman, J. W. Caldwell, T. E. Cheatham, W. S. Ross, C. L. Simmerling, T. A. Darden, K. M. Merz, R. V. Stanton, A. L. Cheng, J. J. Vincent, M. Crowley, V. Tsui, R. J. Radmer, Y. Duan, I. Pitera, J. Massova, G. L. Seibel, U. C. Singh, P. K. Weiner, and P. A. Kollman. AMBER 6. Technical report, Univercity of California, San Francisco, 1999.

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

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

[24] S. Chandrasekar. Stochastic problems in physics and astronomy. *Rev. Mod. Phys.*, 15:2–87, 1943.

[25] T. E. Cheatham, P. Cieplak, and P. A. Kollman. A modified version of the cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.*, 16:845–862, 1999.

[26] T. E. Cheatham III and B. R. Brooks. Recent advances in molecular dynamics simulation towards the realistic representation of biomolecules in solution. *Theor. Chem. Acc.*, 99:279–288, 1998.

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

[28] J. Christiansen, S. R. Douthwaite, A. Christiansen, and R. A. Garrett. Does unpaired adensine-66 from helix II of E. coli 5s RNA bind to protein L18? *EMBO J.*, 4:1019–1024, 1985.

[29] G. Colmenarejo and I. Jr. Tinoco. Structure and thermodynamics of metal binding in the P5 helix of a group I intron ribozyme. *J. Mol. Biol.*, 290:119–135, 1999.

[30] W. D. Cornell, P. Cieplak, C. I. Bayly, K. M. Merz Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins and nucleic acids. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[31] C. J. Cramer and D. G. Truhlar. Implicit solvation models: Equilibria, structure, spectra and dynamics. *Chem. Rev.*, 99:2161–2200, 1999.

[32] T. Darden, D. York, and L. Pedersen. Particle mesh ewald: an N*log(N) method for computing ewald sums. *J. Chem. Phys. Comm.*, 98:173, 1993.

[33] M.E. Davis, J.D. Madura, B.A. Luty, and J.A. McCammon. Electrostatics and diffusion of molecules in solution: Simulations with the university of houston brownian dynamics program. *Comp. Phys. Comm.*, 62:187–197, 1991.

[34] 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, Shaked Z., Sklenar H., D. M. Soumpasis, Tung C. S., von Kitzing E., A. H. J. A. Wang, and V. B. Zhurkin. Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, 205:787 – 791, 1989.

[35] B.D. Dominy and C.L. Brooks III. Development of a generalized born model parametrization for proteins and nucleic acids. *J. Phys. Chem. B*, 103:3765–3773, 1999.

[36] Y. Duan, S. Kumar, J. M. Rosenberg, and Kollman P. A. Gradient SHAKE: An improved method for constrained energy minimization in macromolecular simulations. *J. Comp. Chem.*, 16(11):1351–1356, 1995.

[37] S. R. Edinger, C. Cortis, P. S. Shenkin, and R. A. Friesner. Solvation free energy of peptides:comparison of approximate continuum solvation models with accurate solution of the Poisson-Boltzmann equation. *J. Phys. Chem. B*, 101:1190–1197, 1997.

[38] E. Ennifar, A. Nikulin, S. Tishchenko, A. Serganov, N. Nevskaya, M. Garber, B. Ehresmann, C. Ehresmann, S. Nikonov, and P. Dumas. The crystal structure of UUCG tetraloop. *J. Mol. Biol.*, 304:35–42, 2000.

[39] E. Ennifar, M. Yusupov, P. Walter, R. Marquet, B. Ehresmann, C. Ehresmann, and P. Dumas. The crystal structure of the dimerization initiation site of genomic HIV-1 RNA reveals an extended duplex with two adenine bulges. *Structure Fold. Des.*, 7:1439–1444, 1999.

[40] M. Famulok. Oligonucleotide aptamers that recognize small molecules. *Curr. Opin. Struct. Biol.*, 9:324–329, 1999.

[41] M. Feig, M. Zacharias, and B. M. Pettitt. Conformations of an adenine bulge in a DNA octamer and its influence on DNA structure from molecular dynamics simulations. *Biophys. Jour.*, 81:352–370, 2001.

[42] N. Folloppe and A. D. MacKerrell. All-atom empirical force field for nucleic acids. I˙ parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comp. Chem.*, 21:86–104, 2000.

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

[44] D. Fourmy, M. I. Recht, S. C. Blanchard, and J. D. Puglisi. Structure of the A site of the Escherichia coli 16 S ribosomal RNA complexed with an aminoglycoside antibiotic. *Science*, 274:1367–1371, 1995.

[45] D. Gautheret and R. Cedergren. Modeling the three-dimensional structure of RNA. *FASEB J.*, 7(1):97, 1993.

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

[47] C. Gohlke, A. I. H. Murchie, D. M. J. Lilley, and R. M. Clegg. Kinking of DNA and RNA helices by bulged nucleotides observed by fluorescence energy transfer. *Proc. Natl. Acad. Sci. USA*, 91:11660–11664, 1994.

[48] K. Grünberger. *A 3D-Model for coarse grained Structure Prediction of RNA*. PhD thesis, University of Vienna, 2002.

[49] C. Guerrier-Takada and S. Altman. Catalytic activity of an RNA molecule orepared by transcription in vitro. *Science*, 223:285–286, 1984.

[50] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman. The RNA moiety of ribonuclease p is the catalytic subunit of the enzyme. *Cell*, 35:849–857, 1983.

[51] K. Han, Y. Lee, and W. Kim. Pseudoviewer: automatic visualization of RNA pseudoknots. *Bioinformatics*, 18:321–328, 2002.

[52] S. C. Harvey, R. K. Z. Tan, and T. E. III Cheatham. The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition. *J. Comput. Chem.*, 19:726–740, 1998.

[53] C. Haslinger. *Prediction Algoritms for Restricted RNA Pseudoknots*. PhD thesis, University of Vienna, 2001.

[54] T. F. Havel, I. D. Kuntz, and G. M. Crippen. The theory and practice of distance geometry. *Bull. Math. Biol.*, 45:665–720, 1983.

[55] D. H. Hawkins, C. J. Cramer, and D. G. Truhlar. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Letters*, 246:122–129, 1995.

[56] B. Hess, H. Bekker, H.J.C. Berendsen, and J. G. E. M. Fraaije. LINCS: a linear constraint solver for molecular simulations. *J. Comp. Chem.*, 18:1463–1472, 1997.

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

[58] I. L. Hofacker. *The rules for the evolutionary Game for RNA: A Statistical Characterization of the Sequence to structure Mapping in RNA*. PhD thesis, University of Vienna, 1994.

[59] W. G. Hoover. Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A31*, pages 1695–1697, 1985.

[60] Y. M. Hou, X. Zhang, Holland J. A., and D. R. Davis. An important 2'OH group for and RNA-protein interaction. *Nucl. Acids Res.*, 29(4):976–985, 2001.

[61] T. Huber, A. E. Torda, and W. F. van Gunsteren. Local elevation: A method for improving the searching properties of molecular dynamics. *J. Comp.-Aided Mol. Design*, 8:695–708, 1994.

[62] W. Humphrey, A. Dalke, and K. Schulten. VMD - Visual molecular dynamics. *J. Molec. Graphics*, 14:33–38, 1996.

[63] B. Jayaram, D. Sprous, and D. L. Beveridge. Solvation free energy of biomacromolecules: Parameters for a modified generalized born model consistent with the AMBER force field. *J. Phys. Chem. B*, 102:9571–9576, 1998.

[64] F. M. Jucker, P. F. Heus, E. Moors, and A. Pardi. A network of heterogenous hydrogen bonds in GNRA- tetra loops. *J. Mol. Biol.*, 264:968–974, 1996.

[65] D. J. Kerwood and P.N. Borer. Structure refinement for a 24-nucleotide RNA hairpin. *Magnetic resonance in Chemistry*, 34:136, 1996.

[66] I. Kuzmine, P.A. Gottlieb, and C. T. Martin. Structure in nascent RNA leads to termination of slippage transcription by T7 RNA polymerase. *Nucl. Acids Res.*, 29(12):2601–2606, 2001.

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

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

[69] R. Lavery, K. Zakrzewska, and H. Sklenar. JUMNA (junction minimization of nucleic acids). *Comp. Phys. Commun.*, 91:135 – 158, 1995.

[70] Andrew R. Leach. *Molecular Modelling: Principles and Applications*. Pearson Education Limited, 2001.

[71] T. Macke and D. A. Case. *Molecular Modeling of Nucleic Acids*, chapter Modeling unusual nucleic acid structures, pages 379–393. Washington, DC: American Chemical Society, 1998.

[72] A. Maier, H. Sklenar, H. Kratky, A. Renner, and P. Schuster. Force field based conformational analysis of RNA structural motifs: GNRA tetraloops and their pyrimidine relatives. *Eur. Biophys. J.*, 28:564 – 573, 1999.

[73] M. Major, M. Turcotte, D. Gautheret, G. Lapaplme, E. Fillion, and R. Cedergren. The combination of symbolic and numerical computations for three-dimensional modelling of RNA. *Science*, 253(5025):1255 – 1260, 1991.

[74] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers*, 29:1105–1119, 1990.

[75] J. L. Miller and P. A. Kollman. Theoretical studies of an exceptionally stable RNA tetraloop: Observation of convergence from an incorrect NMR structure to the correct one using unrestrained molecular dynamics. *J. Mol. Biol.*, 270:436–450, 1997.

[76] M. Molinaro and I. Jr. Tinoco. Use of ultra stable UNCG tetraloop hairpins to fold RNA structures: thermodynamic and spectroscopic applications. *Nucl. Acids Res.*, 23:3056–3063, 1995.

[77] S. Neidle. New insights into sequence-dependent DNA structure. *Nature Struc. Biol.*, 5:754–756, 1998.

[78] S. Nosè. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, 52:255–268, 1984.

[79] A. Onufriev. personal communication, 2002.

[80] A. Onufriev, D. Bashford, and D. A. Case. Modification of the generalized born model suitable for macromolecules. *J. Phys. Chem. B*, 104:3712–3720, 2000.

[81] A. Onufriev, D. A. Case, and D. Bashford. Effective born radii in the generalized born approximation: The importance of being perfect. *J. Comp. Chem.*, 23:1297–1304, 2002.

[82] D. J. Patel. Structural analysis of nucleic acid aptamers. *Curr. Opin. Chem. Biol.*, 1:32–46, 1997.

[83] D. J. Patel, S. A. Kozlowski, L. A. Marky, J. Rice, C. Broka, K. Itakura, and K. J. Breslauer. Extra adenosine stacks into the self complementary d(CGCAGAATTCGCG) duplex in solution. *Biochemistry*, 21:445–451, 1982.

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

[85] J. W. Ponder. TINKER - software tools for molecular design. Online, 12 2001. URL: `http://dasher.wustl.edu/tinker` .

[86] D. Qui, P. S. Shenkin, F. P. Hollinger, and W. C. Still. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate born radii. *J. Phys. Chem.*, 101:3005–3014, 1997.

[87] J. Ramstein and R. Lavery. Energetic coupling between DNA bending and base pair opening. *Proc. Natl. Sci. USA*, 85:7231–7235, 1988.

[88] M. A. Rosen, D. Live, and D. J. Patel. Comparative NMR study of A-bulge loops in DNA duplexes: Intrahelical stacking of A, A-A and A-A-A bulge loops. *Biochemistry*, 31:4004–4014, 1992.

[89] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constrains: Molecular dynamics of n-alkanes. *J. Comp. Phys.*, 23:327, 1977.

[90] Wolfram Saenger. *Principles of Nucleic Acid Structure*. Springer-Verlag New York Inc., 1984.

[91] M. Scarsi, J. Apostolakis, and A. Caflisch. Continuum electrostatic energies of macromolecules in aqueous solutions. *J. Phys. Chem. A*, 101:8098–8106, 1997.

[92] M. Schaefer and C. Froemmel. A precise analytical method for calculating the electrostatic energy of macromolecules in aqueous solution. *J. Mol. Biol.*, 216:1045–1066, 1995.

[93] R. D. Skeel, G. Zhang, and T. Schlick. A family of symplectic integrators: Stability, accuracy and molecular dynamics applications. *SCIAM J. Sci. Comp.*, 18(1):203–222, 1997.

[94] E.J. Sorin, M. A. Engelhart, D. Herschlag, and V. S. Pande. RNA simulations: Probing hairpin unfolding and the dynamics of a GNRA tetraloop. *J. Mol. Biol.*, 317:493–506, 2002.

[95] P. F. Stadler. personal communication, 2001.

[96] P. J. Steinbach and B. R. Brooks. New spherical-cutoff methods for long-range forces in macromolecular simulation. *J. Comp. Chem.*, 15(7):667–683, 1993.

[97] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990.

[98] L. Su, L. Chen, M. Egli, J. M. Berger, and A. Rich. Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot. *Nature Struct. Biol.*, 6:285–292, 1999.

[99] A. A. Szewczak, P. B. Moore, Chan Y-L., and I. G. Wool. The conformation of the sarcin/ricin loop from 28S ribosomal RNA. *Proc. Natl. Acad. Sci.*, 90:9581–9585, 1993.

[100] V. Thiviyanathan, A. B. Guliaev, N. B. Leontis, and D. G. Gorenstein. Solution conformation of a bulged adenosine base in an RNA duplex by relaxation matrix refinement. *J. Mol. Biol.*, 300:1143–1154, 2000.

[101] V. Tsui and D. A. Case. Molecular dynamics simulations of nucleic acids using a generalized Born solvation model. *J. Am. Chem. Soc.*, 122:2489–2498, 2000.

[102] V. Tsui and D. A. Case. Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers: Nucleic Acid Sciences*, 56:275–291, 2001.

[103] O. C. Uhlenbeck. Nucleic-acid structure - tetraloops and RNA folding. *Nature*, 346:613 – 614, 1990.

[104] I. Uson and G.M. Sheldrick. Advances in direct methods for protein crystallography. *Curr. Opin. Struct. Biol.*, 9:643–648, 1999.

[105] W. F. van Gunsteren and H. J. C. Berendsen. Algorithms for brownian dynamics. *Mol. Phys.*, 45(3):637–647, 1982.

[106] W. F. van Gunsteren, H. J. C. Berendsen, and J. A. C. Rullmann. Stochastic dynamics for molecules with constraints. brownian dynamics of n-alkanes. *Mol. Phys.*, 44(1):69–95, 1981.

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

[108] L. Verlet. Computer 'experiments' on classical fluids. I. theroynamical properties of Lennard Jones molecules. *Phys. Rev.*, 159:98–103, 1967.

[109] J. Wang, P. Cieplak, and P. A. Kollman. How well does a RESP(restrained electrostatic potential) model do in calculating the conformational energies of organic and biological molecules. *J. Comp. Chem.*, 21:1049–1074, 2000.

[110] J. Weiser, P. S. Shenkin, and W. C. Still. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J. Comp. Chem.*, 20:217–230, 1999.

[111] D. J. Williams and K. B. Hall. Unrestrained stochastic simulations of the UUCG tetraloop using an implicit solvation model. *Biophys. J.*, 76:3192–3205, 1999.

[112] D. J. Williams and K. B. Hall. Experimental and theoretical studies of the effects of deoxyribose substitutions on the stability of the UUCG tetraloop. *J. Mol. Biol.*, 297:251–265, 2000.

[113] B. Wimberly, G. Varani, and Tinoco I. Jr. The conformation of loop E of eukaryotic 5S ribosomal RNA. *Biochemistry*, 32:1078 – 1087, 1993.

[114] D. York, H. Yang, H. Lee, L. Darden, and L. Pedersen. Toward the accurate modeling of DNA: the importance of long-range electrostatics. *J. Am. Chem. Soc.*, 117:5001–5002, 1995.

[115] M. Zacharias. Comparison of molecular dynamics and harmonic mode calculations on RNA. *Biopolymers*, 54:547–560, 2000.

[116] M. Zacharias. personal communication, 2001.

[117] M. Zacharias and P. J. Hagerman. Bulge induced bends in RNA: Quantification by transient electric birefringence. *J. Mol. Biol.*, 247:486–500, 1995.

[118] M. Zacharias and H. Sklenar. Conformational analysis of single-base bulges in A-form DNA and RNA using a hierarchical approach and energetic evaluation with a continuum solvent model. *J. Mol. Biol.*, 289:261–275, 1999.

[119] M. Zacharias and H. Sklenar. Conformational deformability of RNA: A harmonic mode analysis. *Biophys. J.*, 78(5):2528–2542, 2000.

[120] Martin Zacharias. Conformational analysis of DNA-trinucleotide-hairpin-loop structures using a continuum solvent model. *Biophys. J.*, 80:2350–2363, 2001.

[121] B. Zagrovic, E.J. Sorin, and V. Pande. β-Hairpin folding simulations in atomistic detail using an implicit solvent model. *J. Mol. Biol.*, 117(00):1–19, 2000.

[122] M. Zuker and P. Stiegler. Optimal computer folding of large RNA dequences using thermodynamic and auxiliary informations. *Nucl. Acid. Res.*, 9:133–148, 1981.

Curriculum vitae

Dipl. Ing. Wolfgang A. Svrcek-Seiler

Geboren	17. Mai 1969 Wien, Österreich
Staatsbürgerschaft	Österreichisch
1975 – 1979	Volksschule, Wien
1979 – 1987	Realistisches Gymnasium BG XIII, Wien
Juni 1987	Matura mit ausgezeichnetem Erfolg
1987 – 1997	Studium der Technischen Physik, Technische Universität Wien. Zweite Diplomprüfung mit Auszeichnung bestanden. Diplomarbeit bei Prof. Dr. Harald Markum und Prof. Dr. Frank Rattay in der Arbeitsgruppe TU-BIOMED. Titel: *Einfluß der Brownschen Bewegung auf die Wahrnehmung schwacher akustischer Signale*
Juni 1997	Sponsion zum Diplom-Ingenieur, Technische Physik
1997 – 1998	Zivildienst
Oktober 1998 – Januar 2003	Dissertation am Institut für theoretische Chemie und molekulare Strukturbiologie der Universität Wien bei Prof. Dr. Peter Schuster. Titel: *Force Field based Investigations on Structure and Dynamics of RNA Molecules*

Publikationen

W. A. Svrcek-Seiler, I. C. Gebeshuber, F. Rattay , T. Biro and H. Markum Micromechanical models for the Brownian motion of hair cell stereocilia. *J. Theor. Biol.* 193:623-630, 1998

I. C. Gebeshuber, A. Mladenka, F. Rattay and W. A. Svrcek-Seiler. Brownian motion and the ability to detect weak auditory signals, Chaos and Noise in Biology and Medicine (Eds. M. Barbi and S. Chillemi), *World Scientific Press*, 230-237

I. C. Gebeshuber, A. Mladenka, F. Rattay and W. A. Svrcek-Seiler. Brownian motion enhances the ability to detect low level auditory signals, *Medical & Biological Engineering & Computing*, Vol. 35, Supplement part I, 302