A 3D-model for coarse grained structure prediction of RNA

DISSERTATION
zur Erlangung des akademischen Grades
Doktor rerum naturalium

Vorgelegt der
Fakultät für Naturwissenschaften und Mathematik
der Universität Wien

von
Mag. Kurt Grünberger

im Juli 2002

Abstract

One of the greatest challenges of biophysics is the prediction of molecular structures. Biomolecules, especially proteins and nucleic acids, are in the focus of numerous efforts, because three dimensional structures of these molecules are essential for their function. The follow up of genome sequencing, often called proteomics, requires simultaneous determination of the structures of thousands of biopolymers. Despite substantial progresses in structural analysis of biopolymers by X-ray crystallography and NMR spectroscopy three dimensional structure determination is often time consuming or expensive at the current state of the art. Therefore determination of three dimensional structures based on theoretical models gained in importance. Although the main research focus is still on proteins, three dimensional investigations on nucleic acids increased due to the discovery of new functions and aspects of RNA. RNA is also challenging from a theoretical point of view, since secondary structures not only can be calculated by efficient algorithms, but also are relevant as folding intermediates.

This thesis aims to set a first step to overcome the gap between secondary structure and three dimensional structure of RNA. A coarse grained model of RNA was designed, based upon off-lattice force field models. Each nucleotide is represented by a composition of seven to nine pseudo atoms. The classical force field potentials were used to define the bonds, angles and torsion angles between these pseudo atoms. Using statistical data derived from a PDB data set the parameters were estimated and dependences were checked. To describe the hydrogen bonding and stacking interaction a new angle dependent potential was designed. The created model was successfully applied on two simple test sets (helices and tetra loops). The designed program was capable to describe helices and tetra loops in an adequate way. Finally, for detecting limits and possible improvements of this simplified model, a more complex structured molecule (the hammerhead ribozyme) was calculated. Again, the program was capable to describe many structural features of this ribozyme in an appropriate way. Comprising the data, the followed approach resulted in a prototype representing the base for further coarse grained RNA structure models.

Zusammenfassung

Eine der größten Herausforderungen der Naturwissenschaften ist zweifellos die Vorhersage von Molekülstrukturen. Speziell Biomoleküle, wie Proteine und Nukleinsäuren, sind Ziel intensiver Forschungsaktivität, weil ein bestimmter Teil der dreidimensionalen Struktur immer essentiell für das Verständnis der Funktion ist. Der aktuelle Fortschritt der Genomsequenzierung erfordert die gleichzeitige Bestimmung von tausenden Biopolymerstrukturen. Trotz der großen Fortschritte bei der Untersuchung von Biopolymeren durch die Röntgenstrukturanalyse und NMR Spektroskopie ist die experimentelle Bestimmung der dreidimensionalen Struktur aufwendig und oft mit enormen Kosten verbunden. Daher wird die Bestimmung der dreidimensionalen Struktur mit dem Computer immer wichtiger. Obwohl der Focus solcher Untersuchungen nach wie vor bei Proteinen liegt, erlangen diesbezügliche Untersuchungen bei RNA Molkülen zunehmend an Bedeutung. Nichtzuletzt weil viele neue Funktionen und Aspekte der RNA erkannt wurden. Die RNA ist auch aus theoretisch chemischer Sicht interessant, weil ihre Sekundärstruktur durch Algorithmen berechnet werden kann.
Ziel dieser Dissertation war es, ein Modell zu entwickeln, dass zwischen der Vorhersage von Sekundärstruktur und der dreidimensionalen RNA Struktur vermittelt. Durch die Entwicklung eines vereinfachten Modells, das auf off-lattice Kraftfeldern basiert, wird jedes Nukleotid durch sieben bis neun Pseudoatome charakterisiert. Durch klassische Kraftfeldpotenziale wurden Bindungen, Winkeln und Torsionen definiert. Die Konstanten dieser Potenziale wurden durch statistische Auswertungen gemessener Strukturen abgeschätzt. Um die Stacking Wechselwirkung und die Wasserstoffbrücken Bindung adequat zu beschreiben, wurde ein neues winkelabhängiges Potenzial entwickelt. Dieses Programm wurde an zwei einfachen Testsets (Helices und einem Tetraloops) erfolgreich geprüft. Es konnte gezeigt werden, dass das entwickelte Programm im Stande ist, Helices und Tetraloops in entsprechender Weise zu beschreiben. Um die Grenzen und Verbesserungmöglichkeiten dieses Ansatzes zu erkennen, wurde mit diesem Programm abschließend eine komplexe Molekülstruktur (das Hammerhead Ribozym) berechnet. Es zeigte sich, dass viele strukturelle Merkmale des Ribozyms durch dieses Modell auf geeignete Weise dargestellt werden können. Zusammenfassend läßt sich sagen, dass durch den verfolgten Ansatz der Grundstein gelegt wurde, um eine Brücke zwischen der zweidimensionalen und dreidimensionalen Welt herzustellen.

Contents

1 Introduction 1

2 Structure motifs of RNA 4

2.1 Chemical structure and building blocks of RNA 4

2.2 Definition of primary, secondary and tertiary structure of RNA 6

2.2.1 Primary structure 7

2.2.2 Secondary structure 7

2.2.3 Tertiary structure 9

2.3 Forces shaping RNA structure 9

2.3.1 Base pairing 9

2.3.2 Stacking 12

2.3.3 Backbone flexibility 14

3 Experimental structure determination of RNA 16

3.1 X-ray diffraction 16

3.1.1 Introduction 16

3.1.2 RNA and X-ray diffraction 18

3.2 NMR 19

3.2.1 Introduction 19

3.2.2 RNA and NMR 21

4 Computational structure determination of RNA 23

4.1 Force fields 23

4.1.1 Energy Calculation 25

4.1.2 Structure optimation 28

4.1.3 Molecular dynamics 30

4.1.4 Programs 32

4.2 Conformation build-up programs 37

4.2.1 MC-SYM 38

4.3 A short glance on protein structure prediction 42

4.3.1 Structure of proteins 42

4.3.2 Modeling of proteins 42

4.3.3 Comparison between proteins and RNA 45

5 The program `toyRNA` 48

5.1 General aspects . . . 48
5.2 Building blocks . . . 48
5.3 Types of potential functions . . . 49
    5.3.1 Bond, angles and torsion potentials . . . 50
    5.3.2 The nonbonded interactions . . . 54

6 Calculations 61

6.1 Helix . . . 61
    6.1.1 Energy minimization . . . 61
    6.1.2 Molecular dynamic simulations . . . 66
6.2 GNRA tetra loop . . . 68
    6.2.1 Energy minimization . . . 68
    6.2.2 Molecular dynamic simulations . . . 70
6.3 Hammerhead ribozyme . . . 72
    6.3.1 Energy minimization . . . 73
    6.3.2 Molecular dynamic simulation . . . 73

7 Conclusion 80

8 Outlook 81

A Statistics 83

1 Introduction

The prediction of molecule structure is probably one of the greatest tasks of chemistry. This is not surprising, because the structure of a molecule always includes information about its properties and functions. Beside the development of various experimental techniques, such as NMR and X-ray crystallography, chemists try to calculate molecule structures. The first approaches in the 1960s , restricted by computer technology in those days, were of very simple nature and based upon Newtonian mechanics using minimization of 'force fields'. Today, after a tremendous growth of computational power, there are two different strategies of using this enormous potential. One strategy is heading for the most exact results possible based upon methods of quantum chemistry. Such methods are restricted to small molecules, because time consuming algorithms are used. The other strategy aims at larger structures but has to omit detailed information in order to make this calculations feasible.

Biomolecules seem to be attractive targets for such calculations, because the three-dimensional structure of at least part of the molecule is essential for molecular function. Nearly all biomolecules are linear polymers, consisting of covalently bounded monomers. They are so called “hetereopolymers” hence the sequence is built up by a handful of different monomers. For example 20 amino acids form the building blocks for proteins, four nucleotides are those for RNA and DNA. The simplest description of such linear molecules, the declaration of the monomer's order, is named sequence.

One vision of the theoretical chemists was to calculate the natural conformation of any biomolecule only with the information of it's sequence. The disillusionating facts given by the first calculations destroyed this intention rapidly. The large number of atoms and the immense dimension of the conformational space of these molecules lead to non solved problems in calculation. To get any impression of the biomolecule's energetic and dynamic aspects, scientists created coarse-grained models of biomolecules. Depending on biomolecules' nature, different attempts were made.

For proteins the simplest approach presents each monomer by one point in a three-dimensional grid [27] and allows only specific movements on this grid. The corresponding approach for RNA molecules is the prediction of the secondary structure. Although these models include considerable abstractions, important energetic aspects of the folding procedure, especially the shape of the conformational landscape was elucidated [28]. The fundamental dif-

ference between these simple models is the dimensional validity. While the protein grid model is based on spaciousness, no three dimensional aspects are included in RNA's secondary structure.
The importance of proteins was never questioned and due to the close interrelation between function and conformation also three dimensional structure was of early scientific attention. Therefore attempts of spacial prediction have been focused on these biomolecules quite continuously and for a long time.
The history of research activity on nucleic acid structure is characterized by permanent "up and downs". The milestone for nucleic acids' three dimensional structure took place in 1953, when Watson and Crick [140, 139] identified the structure of DNA as a double helix. Although this structure was one of the first solved for a biomolecule, the interest in nucleic acid structure declined in the 1950s and 1960s after the various types of RNA in cells an their biological functions were understood. DNA was identified as the storage of the genetic information while RNA considered solely as a passive transporter of the genetic code. Recently, we have learned that RNA plays even wider and unexpected roles. RNAs are an extremely versatile class of molecules actively participating in all steps of gene expression: RNAs that specifically recognize substrate molecules, e.g. aptamers [33, 100], and RNAs that catalyze chemical reactions, e.g. ribozymes [16, 78, 88], have been discovered both in nature and by *in vitro* selection. With this new knowledge the interest on information of three dimensional structure, dynamics and conformational energetics of RNA grows extremely.
During the last years fundamental improvements of experimental techniques result in growth on solved RNA structures. Theoretical methods, especially a new generation of force fields and efficient algorithms for calculating long range electrostatic forces, render in accurate calculations for small nucleic acids. But nevertheless the problems due to the large number of atoms and the conformational variety remain. There are still no methods available for obtaining RNAs' native three dimensional structure only with the knowledge of it's secondary structure. At time, the gap between the efficient secondary structure prediction methods and three dimensional RNA models seems insuperable. Due to the impossibility of overcoming this gap by a single step, it is necessary to split the transformation procedure.

The objective of this thesis is creating a simplified spacial model of RNA, that corresponds to an intermediate stage connecting the world of secondary structure and the world of three dimensional geometry of RNA molecules. The approach is based on the assumption that a reduction of atoms subsequently reduces the conformational space and simplifies the search for relevant RNA conformations. If the prediction of spacial structure is established on such simplified level, it will be possible to predict every three dimensional structure based on a given RNA sequence.

2 Structure motifs of RNA

2.1 Chemical structure and building blocks of RNA

RNAs are linear polymers of defined sequences build from a very limited range of monomers, called nucleotides. These nucleotides are adenosine (**A**), guanosine (**G**), cytidine (**C**), and uridine (**U**) (in DNA uridine is replaced by the functionally equivalent thymidine (**T**)). Each monomer of a nucleotide consists of three molecular blocks:

- **PHOSPHATE GROUP**
  The phosphates are linking the nucleotides. This group is also responsible for one of the typical features of nucleic acids, their polyionic character.

- **PENTOSE**
  The pentose is of furanoside-type (β-D-ribose in RNA or β-D-2$'$ - desoxyribose in DNA). It is phosphorylated in 5$'$ position and substituted at C1$'$ by one of the purine or pyrimidine groups attached by a β-glycosyl C1$'$-N linkage. Replacement of the 2$'$-OH of RNA by a hydrogen in DNA has major ramifications. DNA is much less susceptible to hydrolysis than RNA and cannot have branched polynucleotides involving the 2$'$-, 3$'$- and 5$'$-OH groups. Lacking the steric hindrance of the 2$'$ hydroxyl, DNA has more conformational flexibility than RNA. Otherwise this extra hydrogen bonding site gives RNA greater possibilities for specific interactions.

- **PURINE OR PYRIMIDINE GROUP**
  The hetero cycles are the purine bases adenine (A) and guanine (G) and the pyrimidine bases cytosine (C) and uracil (U, uracil is replaced in DNA by the functionally equivalent thymine - 5-methyluracil)

Figure 1 shows a short strand of RNA containing the four usual nucleotides adenine (A), and guanine (G), cytosine (C) and uracil (U). The building blocks described above are shown in different colors (green - phosphate, blue - ribose, black - purine (G, A) and pyrimidine (C, U) nitrogen 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). Numerous naturally occurring modified nucleotides exist beside these four bases: Many of them have antibiotic activity, e.g. the important class of arabinosides, nucleosides with β-D-arabinose instead of β-D-ribose.

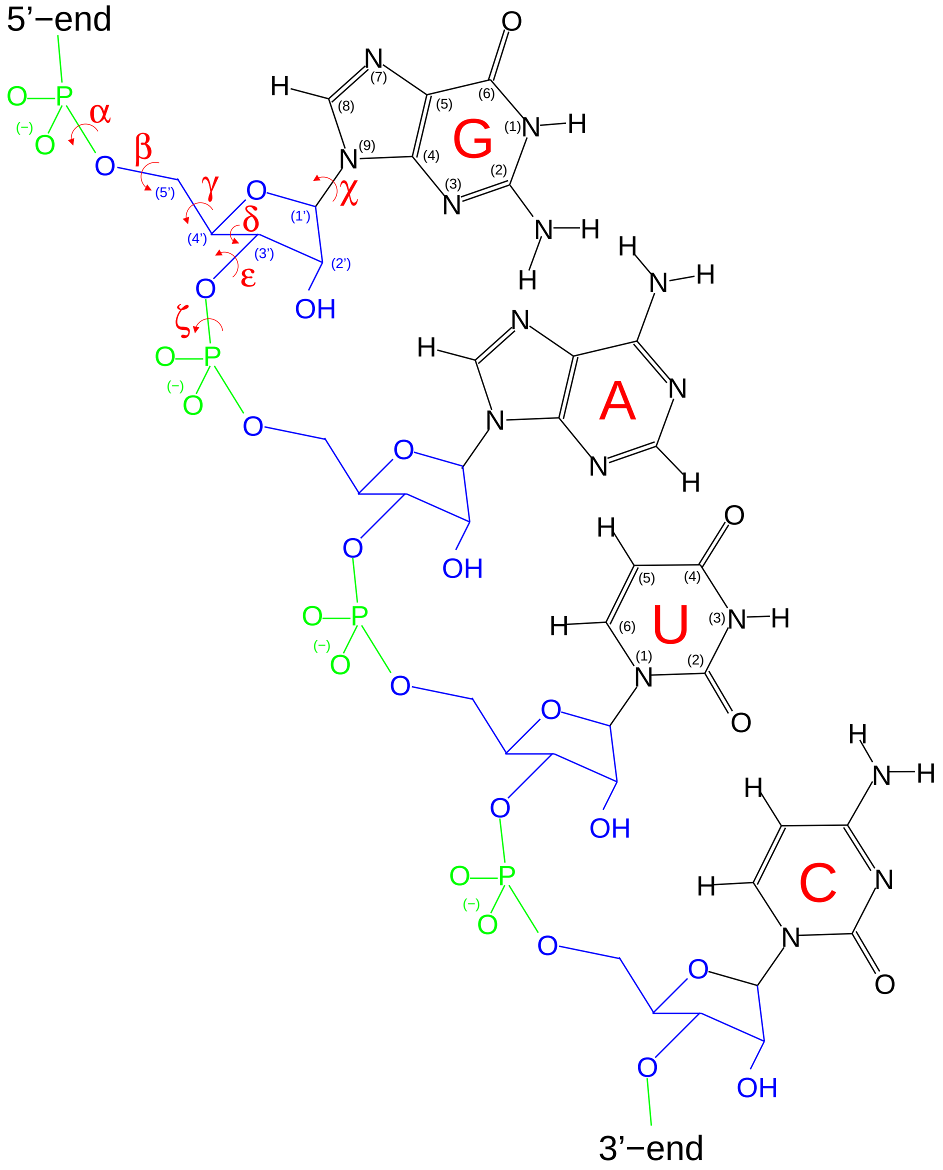


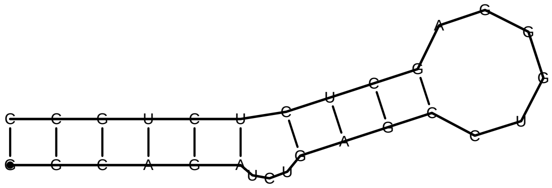
Figure 1: Atomic structure of RNA: (green - phosphate, blue - ribose, black - purine (G, A) and pyrimidine (C, U) nitrogen bases)

2.2 Definition of primary, secondary and tertiary structure of RNA

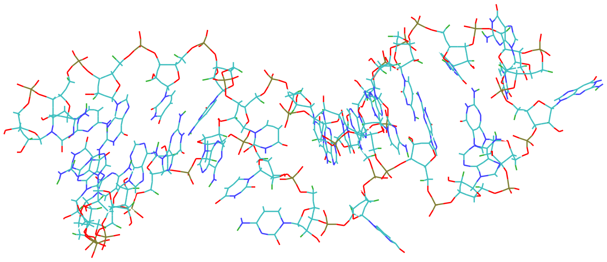
RNA structure can be described in three different ways depending on the desired information. The descriptions are named primary, secondary and tertiary RNA-structure and include different resolutions of structural details. Figure 2 shows the primary, secondary and tertiary structure of a cis-acting RNA regulatory element [1].

`5′-GGCAGAUCUGAGCCUGGGAGCUCUCUGCC-3′`

(a) Primary structure



(b) Secondary structure



(c) Tertiary structure

Figure 2: Primary structure extracted from the PDB-file 1anr.pdb, secondary structure calculated with RNAfold from the Vienna RNA package [56, 89, 151] (minimum free energy structure), tertiary structure based one of the models from PDB-file 1anr.pdb.

2.2.1 Primary structure

The simplest way of characterizing RNA is the declaration of the nucleotides' order. As every sequence has two ends, there are two possible starting points for enumeration. There is a difference between the so called $5'$ and $3'$ ends, because of the nucleotides binding types. By convention the starting-point is the $5'$-end. The result of this approach is a string of the four letters **A,G,C,U**, the so-called **Primary structure**.
In spite of the significant simplification RNA's and DNA' s primary structure is a very useful representation of nucleic acids. Primary structure is mostly the information molecular biologists and genetics receiving and using for their experiments. The excellent mathematical manageability of primary structure is exploited in biophysics for comparison of different nucleic acids by its sequence, the so called sequence alignment [31, 138].

2.2.2 Secondary structure

One of the main features of the nucleotides is the ability to form hydrogen-bond mediated base pairs. Therefore RNA, much like DNA, can form double helices of complementary strands. Since RNA usually occurs single stranded, formation of double helical regions is accomplished by the molecular folding back onto itself to form base pairs. While base pair patterns are very restricted in DNA, a large variety occurs in RNAs. Starting with Watson Crick types G-C and A-U (in analogy to DNA) RNA occurs in different geometries to G-U pairs and even more uncommon types like G-A, G-G or A-C. The declaration of the base-base connectivity pattern results in the so-called **Secondary structure**.
Secondary structure can be classified in very few types of structural motifs. The most abundant of these motifs are the so-called hairpins consisting of a double-stranded part (the 'stem') and a connecting single-stranded part (the loop). Other motifs are the bulge (unpaired bases on one side of the stem), or the multi-loop (several stems connected by short unpaired regions). Unpaired regions at the end of a strand are called 'dangling ends' (see Figure 3).
In the mathematical point of view the secondary structure of RNA represents a graph. Efficient algorithms [98, 151], including dynamic programming techniques [9] and experimentally measured energy parameters [37, 48, 59, 132], are available to calculate RNA's secondary structure of given

sequence under distinct restrictions (for example allowing only secondary structure motifs presented in Figure 3 and G≡C, A=U and wobble G=U base pairs). Using these handy tools the sequence to secondary structure map for RNA [36, 35, 115] and its consequences for evolutionary adaption [57] have been characterized in detail.

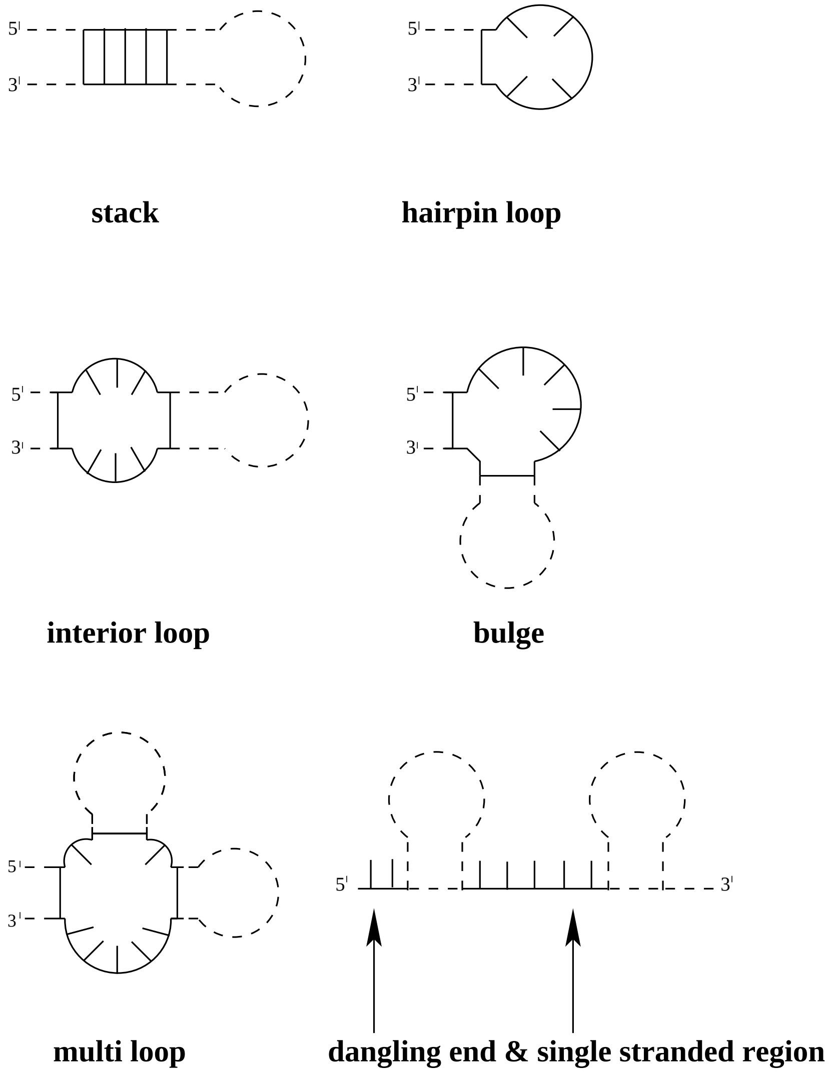


Figure 3: Secondary structure motifs in RNA

2.2.3 Tertiary structure

As mentioned before representations of RNA as strings or graphs can be handled in an efficient way, but ignore the fact that RNAs are three dimensional objects. This three dimensional arrangement of RNA is the so-called **Tertiary structure**.
Tertiary structure can be defined in different resolutions. At lowest resolution all the relative positions of the secondary structure elements are given with respect to each other. At the highest resolution the position of each atom is known.
The transition between secondary and tertiary structure is attended by a change of methods. Strings and graphs were substituted by three dimensional objects, handled by vector analysis. Interactions between base pairs, which are binary and digital in secondary structure (that means "only *two* bases can *interact or not*") are substituted by continuous potentials between atoms. These changes cause a tremendous grow of data and the awkwardness of experimental and theoretical prediction increases proportional to it. Nevertheless several structure motifs can only be identified if the tertiary structure is taken into consideration (see section 2.3.1). Border structure motifs between secondary and tertiary structure are the so-called pseudo knots [103]. Although expressible in secondary structure's notation, pseudo knots are mostly assigned to the tertiary structure of RNA, due to the problems arising when included in classic computational secondary structure determination. With several restrictions (allowing only distinct types of pseudo knots) it is also possible to calculate pseudo knots on the secondary structure level [47, 107, 108].

2.3 Forces shaping RNA structure

Three factors have dominant influence on the conformation of RNA: (1) *base pairs*, formed by hydrogen bonds between the nitrogen bases, (2) *base stacking*, which tents to minimize the repulsive interactions between the polar solvent and the nitrogen bases, and (3) the *flexibility of the backbone.*

2.3.1 Base pairing

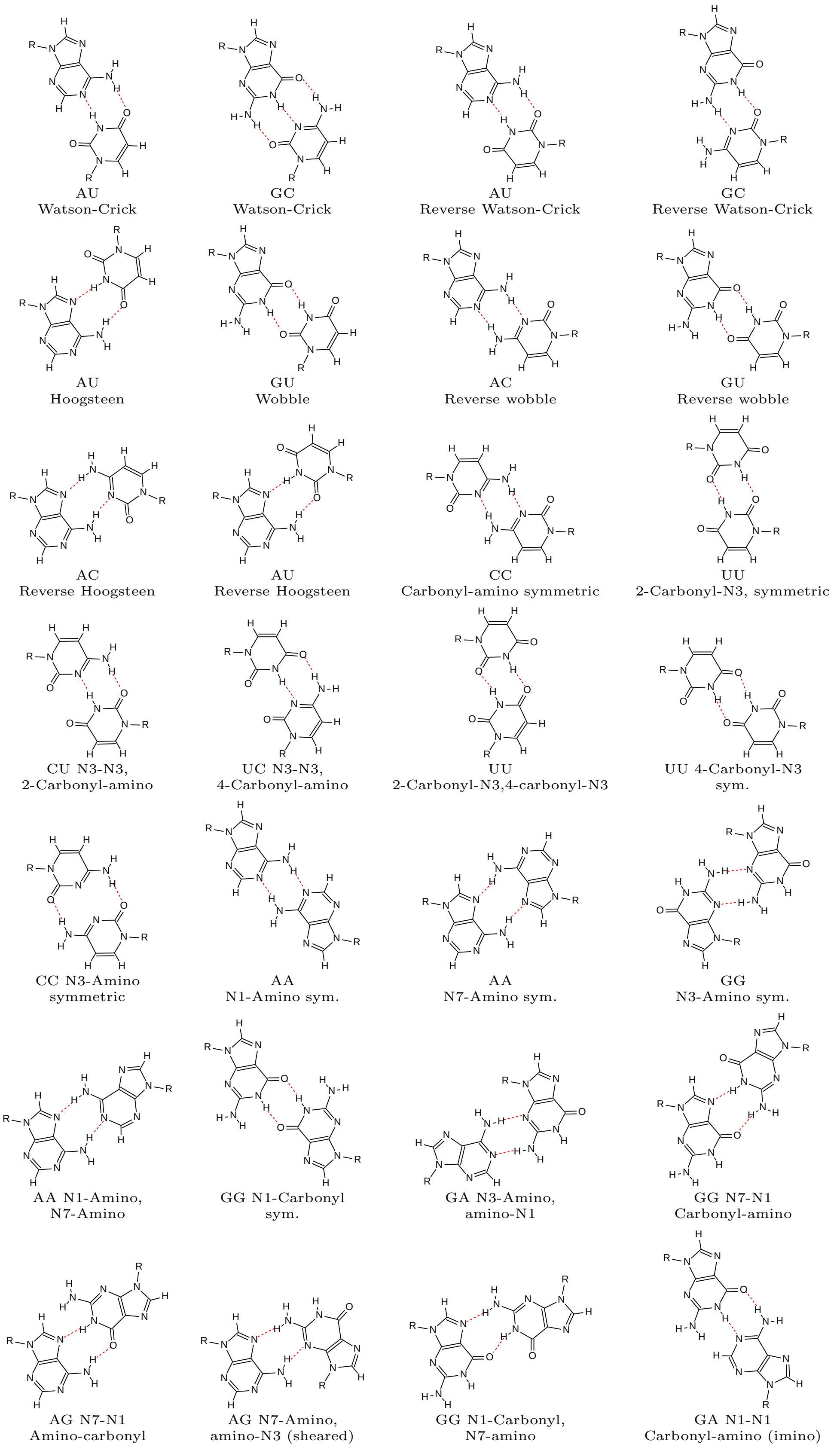
Base pairs are edge-to-edge complexes between purine and pyrimidine bases, mediated by hydrogen bonding between complementary arrays of electrically polarized atoms. The most important examples of base pairs are the standard or canonical Watson-Crick pairs. They are characterized by their remarkable isostericity, which gives rise to the regular A-form double helix, and allow each of the four combinations to substitute for any of the others without distorting the three-dimensional helical structure. Therefore these base pairs are usually found in the double-stranded DNA and RNA. However, the canonical Watson-Crick pairs represent only one of various possible edge-to-edge interactions. If a continuous helix of arbitrary sequence is not required, many other hydrogen-bonded base pairs are possible. Feasible base pairs with at least two hydrogen bonds are shown in figure 4.

Beside the interactions shown in figure 4, bases can also pair in a multitude of other ways. For example, ionization of nucleotides provides further opportunities for base pairing patterns. In addition base interactions are not confined to pairs: "Base triplets" occur regularly, as several examples known from the crystal structures of tRNAs demonstrate (see figure 5a). "Quartets" of Guanine and Adenine platforms represents examples of interactions between four bases with regular structures [18].

A further extension is caused by non-standard hydrogen-bonds. So called bifurcated (or more appropriately 'chelated') hydrogen bonds have been observed recently in high-resolution structures of rRNA's loop E and previously in the lower resolution of tRNAs (see figure 5b). Such bifurcated hydrogen bonds are often elements of base multiplets such as 'base quadruples', which found to stabilize the loop of an RNA pseudo knot [124].

Some hydrogen bond patterns including one or more water bridges between the bases (see figure 5c) were also observed. In crystal structures $\text{C-H}\cdots\text{N/O}$ systems often show hydrogen bond like geometries [60]. These motifs were shown to be stable in molecular dynamics simulations [4]. The significance of these results are controversially interpreted.

The classical sight of base pairs as edge-to-edge complexes between purine and pyrimidine bases was recently expanded by Westhof [74, 142], who identified three edges as possible linkers between two nucleotides: the Hoogsteen edge, the Sugar-edge and the Watson-Crick edge (see figure 6). The sugar edge, which includes the 2'-OH group of the ribose, takes account to new structure elements such as ribose zippers, which can be found for example in the crystal structures of P4-P6 domain of group I ribozymes [18] and of the HDV ribozyme [34].



AU
Watson-Crick

GC
Watson-Crick

AU
Reverse Watson-Crick

GC
Reverse Watson-Crick

AU
Hoogsteen

GU
Wobble

AC
Reverse wobble

GU
Reverse wobble

AC
Reverse Hoogsteen

AU
Reverse Hoogsteen

CC
Carbonyl-amino symmetric

UU
2-Carbonyl-N3, symmetric

CU N3-N3,
2-Carbonyl-amino

UC N3-N3,
4-Carbonyl-amino

UU
2-Carbonyl-N3,4-carbonyl-N3

UU 4-Carbonyl-N3
sym.

CC N3-Amino
symmetric

AA
N1-Amino sym.

AA
N7-Amino sym.

GG
N3-Amino sym.

AA N1-Amino,
N7-Amino

GG N1-Carbonyl
sym.

GA N3-Amino,
amino-N1

GG N7-N1
Carbonyl-amino

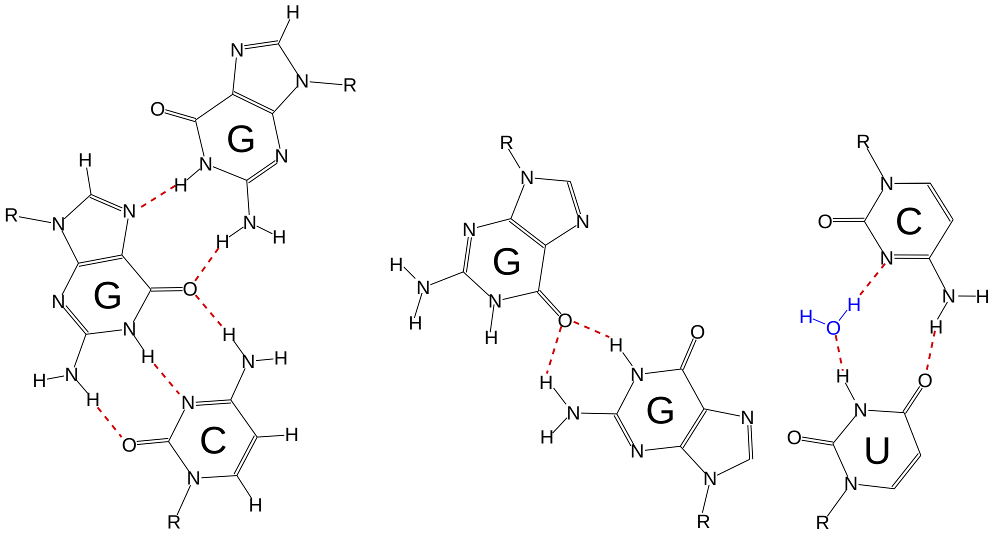
AG N7-N1
Amino-carbonyl

AG N7-Amino,
amino-N3 (sheared)

GG N1-Carbonyl,
N7-amino

GA N1-N1
Carbonyl-amino (imino)

Figure 4: Base pairs with at least two hydrogen bonds (Saenger [112])



(a) (b) (c)

Figure 5: Non standard hydrogen-bonding patterns: (a) Basetriple in tRNA^{Phe}(b) Bifurcated base pair (c) Water-mediated base pair

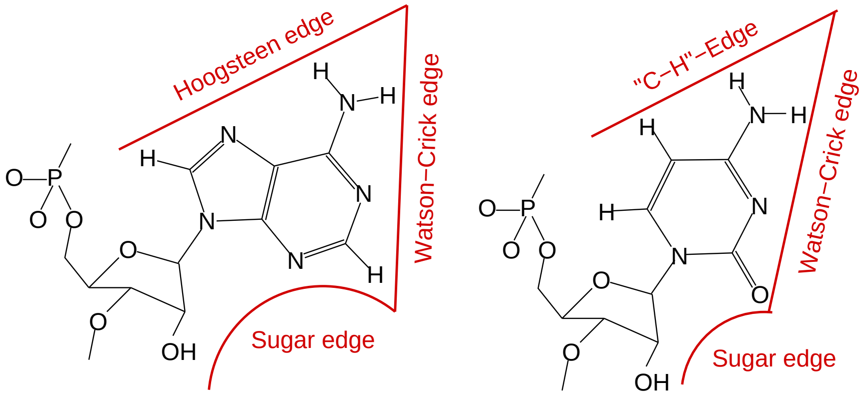
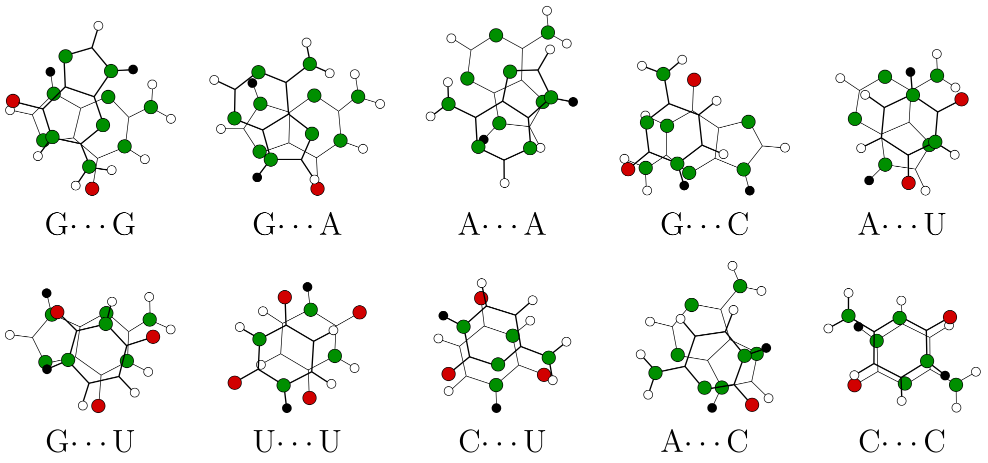


Figure 6: Three edges for possible interaction of purine (A or G, indicated by "R") and pyrimidine (C or U, indicated by "Y") as recommended by Westhof and Leontis.

2.3.2 Stacking

Whereas hydrogen bonds provide directionality and specificity, energetics of nucleic acids are dominated by base and base pair stacking. The stacking interaction between the nitrogen bases mediated by their π-electron systems, is often considered as the nucleic acids analogue of the hydrophobic interaction in proteins. This is not true, since base pair stacking is an enthalpy driven process, that means there are attractive forces between the aromatic systems [116]. However, the driving force for stacking is by no means less sophisticated than that for hydrophobic aggregation. The detailed molecular mechanism involves water structures in both cases: base or base pair stacking is not observed in non aqueous media like, for example, chloroform. The formation of stacks in solution can be described well by cooperative stack thermodynamics as expressed by $K(n) = \sigma \cdot s^n$ where K is the macroscopic equilibrium constant for stack formation, s is the microscopic constant for conversion of a coil element into a segment of the double helix, σ is the nucleation parameter, and n is the length of the stack [105]. Lots of data are available for such a treatment of stack formation. Nevertheless, these macroscopic approaches condense all interactions and prevent an appropriate identification of the interactions' physical nature. Therefore gas phase experiments are necessary for analyzing the attractive interactions between the nitrogen bases in absence of solvent. Unfortunately, the gas phase experiments for stacking obstructed by the fact, that the stacked configurations do not represent global minima on the gas phase free energy surface of base dimers. Thus, formation of hydrogen bonded assemblies is prevalent. The only sources for data are detailed *ab initio* quantum chemical calculations [122, 54]. Results show that there are three contributions: dispersion attraction, short range repulsion, and electrostatic interactions. The stabilization of base stacking is dominated by the dispersion attraction, which is rather isotropic and proportional to the geometrical overlap of the heteroaromatic systems. The distance between stacked bases is determined by the balance between dispersion attraction and short range repulsion present between the adjacent nucleobases. Finally, the mutual orientation of bases and their displacement are primarily determined by the electrostatic attractions.

Stacking of bases, contributing significantly to the stability of RNA architectures, occurs predominantly between consecutive residues within one strand [112]. In a number of RNA three-dimensional structures, cross strand stacking of bases belonging different strands is also observed.



G···G G···A A···A G···C A···U

G···U U···U C···U A···C C···C

Figure 7: The optimal stacking geometries of ten stacked nitrogen base dimers, obtained by ab initio calculations [54] (Cycle colouring code: blue = oxygen, green = nitrogen, white = hydrogen, black = hydrogen instead of sugar at the glycosidic link)

2.3.3 Backbone flexibility

Since both, the sugar and - even more - the heterocycles are very rigid structures, most of the conformational flexibility is induced by the backbone. In figure 1 seven torsional angles are designated by Greek letters (following IUPAC recommendations [58]): Six of them are along the backbone, and starting from the 5′-end of the molecule their definition is as follows:

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

Angle χ (O1′ - C1′ - N9 - C4 in purines and O1′ - C1′ - N1 - C2 in pyrimidines), the seventh torsional angle, is of major importance to the three dimensional structure. As a very good assumption these seven internal degrees of freedom per monomer unit can be used to define the whole conformational space of an RNA molecule.

Two of the seven torsional angles are of special interest as they occupy only very specific values:

- **δ :** This torsion angle lies within the sugar ring system and is restricted by a ring closure criterion. Since a five-membered ring has no flat geometry, one or two atoms are lying above or below the plain defined by the other four or three atoms. If the atom is on the same side of the plain as the C5′, the conformation is called *endo*, if it is on the opposite side it is called *exo*. This behavior is also called 'sugar-puckering'. Figure 8 shows two of the most frequent sugar-puckers in RNA: C2′-endo (left-hand-side of figure 8) and C3′-endo (right-hand-side of figure 8). Nucleotides in the standard A-RNA helix are of C3′-endo conformation, C2′-endo conformations occur mostly in small loops, because of their tendency to elongate the backbone. Apart from the two major types other conformations occur mainly in non helical regions.

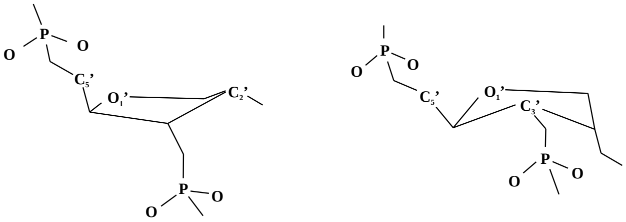


Figure 8: Major puckering modes of sugars in RNA (left-hand-side: C2′-endo, right-hand-side: C3′-endo)

- **χ :** This angle determines the position of the heterocycle with respect to the sugar ring. Though this torsional angle is not involved in a ring system, its values are nevertheless restricted to two distinct regions, one around 0 degrees and the other around 180 degrees. If the heterocycle is rotated towards the C5′-atom ($\chi = 0°$) the conformation is called *syn*, if the heterocycle is in opposite position (away from the C5′-atom, $\chi = 180°$) the conformation is called *anti*. In standard A-RNA-helices all bases occupy the *anti*-conformation, *syn*-conformations can be found in loop regions and in some non-Watson-Crick base pairs.

All other torsional angles prefer also certain ranges, but with an extended variability. For more detail W. Saenger's book "Principles of Nucleic Acid Structures" [112] offers a comprehensive introduction to nucleic acid structure.

3 Experimental structure determination of RNA

The determination of the three-dimensional structure is not a trivial task. It is a step by step process starting with the determination of the primary structure (the sequence) followed by the analysis of the secondary structure (folding pattern). Finally the tertiary structure has to be solved. Whereas for the first two steps a variety of experimental methods like cross-linking, foot-printing or gel-electrophoresis as well as theoretical methods (e. g. secondary structure prediction; for an overview see [152]) are available, the choice of methods for determination and prediction of the tertiary structure is much more imitated. X-ray diffraction and NMR-methods are the two most important experimental methods for determination of RNAs' tertiary structure. A short introduction to these two methods is given in section 3.1 and 3.2, respectively. The prediction of RNA structure by computational methods is described in more detail in chapter 4.

3.1 X-ray diffraction

3.1.1 Introduction

The principle of X-ray crystallography is very simple. X-rays (wavelength 0.1 to 100 Å), are scattered when they pass through a non-homogeneous distribution of electrons. As atoms in matter provide localized concentrations of electron density, X-radiation is deflected at different angles. In orderly systems like crystals (or fibers), constructive and destructive interference of the scattered radiation cause a typical diffraction pattern, which can be used to determine the structure of the crystal and the crystallized molecule. The basic equations relating structure and diffraction are

$$F(hkl) = \sum_{j=1}^{n} f_j exp[2\pi i(hx_j + ky_j + lz_j)] \qquad I(hkl) \propto |F(hkl)|^2$$

where $I(hkl)$ is the intensity of a diffracted X-ray beam coming out from the crystal at an angle described by three integers (Miller indices, or coordinates of the diffraction maxima in reciprocal space) h, k, l. $F(hkl)$ is the structure

factor and f_j is the X-ray scattering factor of atom j whose Cartesian coordinate relative to the unit cell axes are (x_j, y_j, z_j). The summation runs over all atoms in the unit cell.
To work in opposite direction, i.e. when one has the diffraction pattern and wishes to calculate the structure (more exactly the electron density map), this formula can be inverted. Thereby the so called 'phase problem' occurs. There are two ways to solve this problem for large molecules, namely isomorphous replacement and molecular replacement. *Isomorphous replacement* uses heavy atom derivatives of molecules (e.g. made by replacing cytosine with bromocytosine or replacing magnesium with lead). The derivates have to crystallize in the same unit cell as the original molecule. *Molecular replacement* is a computational technique where a molecules' model or part of its' structure is randomly oriented within the unit cell until an optimum match with the diffraction data is archived. The rest of the structure is solved iteratively. Molecular replacement is particularly appropriate for RNA, since the stems of RNA are usually close to the canonical A-form. Once initial phases are determined and the first electron density map is created, a computer model of the biopolymer can be built, which can be starting-point for structure refinement algorithms. The goodness of fit of a proposed structure with the observed diffraction pattern is generally summarized in terms of the R factor

$$R = (\sum_j ||F_{obs.}| - |F_{calc.}||)/\sum_j |F_{obs.}|$$

where $F_{obs.}$ and $F_{calc.}$ are the observed and calculated structure factors, respectively. In general R is about 15%, for a correct oligonucleotide's structure solved at 2Å resolution. Structures, less well resolved or incorrect, may have R about 25%.

Several facts can limit the operation area of X-ray investigation on biochemical macromolecules. As obtaining single crystals is very difficult and unpredictable, it is one of the limiting steps in X-ray crystallography. This is specially true for nucleic acids. Crystallography reveals the structure of only those atoms that are fixed at a given position in most of the molecules throughout the entired crystal. Therefore three dimensional structure of RNA molecules in solution and in crystal might be different as comparison of NMR and X-ray structures seems to indicate (for example [21, 11]).
These problems and the longing for new structures of larger biomolecules

have led to several fundamental improvements in crystallographic analyzes, including the development of large area detectors, improved synchrotron radiation sources, enhanced computer capabilities, cyro-crystallography techniques and breakthroughs in obtaining crystals with improved diffraction limits. In the near future free-electron lasers, providing X-ray flashes with enormous peak brilliance, may allow structural studies of single biomolecules without the need to amplify scattered radiation through Bragg reflections [96, 43].

3.1.2 RNA and X-ray diffraction

Though X-ray diffraction has provided many structures at atomic resolution for proteins and DNA, the number of results for RNAs is still small. Recently the analysis of large RNAs and their complexes with proteins increased dramatically, due to the fundamental technical improvements and the special interest of finding new features of RNA. Landmarks of RNAs' crystallographic structure determination are the first single crystal structures of ApU [118] and GpC [109] in 1976, the tRNAs (e.g. $tRNA^{Phe}$ [64, 125], $tRNA^{Asp}$ [95], $tRNA^{Gly}$ [114], $tRNA^{Fmet}$ [147, 148], and $tRNA^{Imet}$ [7]),the hammerhead ribozyme [104, 117], the P4-P6 domain of the self splicing group I intron from *Tetrahymena thermophila* [18] , the genomic ribozyme from hepatitis delta virus (HDV) [34], the small ribosomal subunit from *Thermus Thermophilus* at 3Å resolution [145], and the large ribosomal subunit from *Haloarcula marismortui* [6]. These complex structures have shown new intricated three-dimensional structures where specific interaction sites are grafted into helices. Such sites comprise structurally conserved modules which can be classified as follow: (1) variations of the Watson Crick base-pairing scheme, i.e. mismatches; (2) triples and quadruples of interacting bases; (3) platforms with pairing between consecutive bases within one strand; (4) bulged-out residues; (5) alternate cross-stacking between bases of different strands, i.e. “interdigitation”; and (6) recurring hydrogen-bonding pattern between riboses of consecutive nucleotides in two strands, i.e. “ribose zipper”.

3.2 NMR

3.2.1 Introduction

NMR (**N**uclear **M**agnetic **R**esonance) based on the absorption of electromagnetic (10 - 1000 MHz) radio-frequency by nuclei in an applied magnetic field B_0, reflects the realignment of the nuclear magnetic moments from low to high energy state. Several nuclei can be used for biomolecular NMR investigations. In natural abundance these are predominantly $^1H,^{19}F$, and ^{31}P. In enriched samples also $^2H,^3H,^{13}C$, and ^{15}N can be measured. An introduction to NMR of biological macromolecules can be found in [149].

Proton - NMR: The exact resonant frequency of each nucleus depends on the strength of the applied magnetic fields, and local perturbations due to the magnetic effects induced by the surroundings named shielding. This perturbation is described in terms of chemical shift δ, since it is a function of the chemical environment of the nucleus. Through-bond interactions (spin or $\mathcal{J}$-coupling) occur between nuclei separated by one or more chemical bonds. The size of the coupling constant depends sinusoidally on the dihedral angle between the two nuclei [62]. This relation can be used to calculate the dihedral angle. Since a biomolecule includes hundreds of hydrogen atoms, the spectrum is very crowded, even at high frequencies.
NMR spectra can be extended over two dimensions, by replacing the single pulse with a sequence of pulses, separated by varying time intervals. The results are transformed into a two dimensional spectrum. Compared with the one dimensional spectra, those spectra are not so crowded.
Two modes of interaction are commonly measured for proton spectroscopy:

- **COSY** (**CO**rrelation **S**pectroscop**Y**) detects sets of protons interacting through bonds. Interacting means that the protons are linked to adjacent bonded pairs of C or N atoms. As results COSY cross peaks allow tracing of the network of protons closely coupled through bonds. **TOCSY** (**TO**tal **C**rrelation **S**pectroscop **Y**) is an extension of COSY based on a modified pulse sequence which gives cross peaks for all protons linked in a J-coupled network, not just for those on adjacent atoms.

- **NOESY** (**N**uclear **O**verhauser **E**ffect **S**pectroscop**Y**) produces signals by transfer of magnetization via dipole-dipole interaction between the nuclei which are close in distance but not through bonds. The volume of a NOESY cross peak is related to the time the two proton dipoles interact (mixing time). The buildup rate σ is the slope of a plot of the NOESY cross peak volume versus mixing time. Distances can be measured by comparing an unknown buildup rate σ_u with the buildup rate of two protons at a known fixed distance (e.g. H5-H6 in pyrimidine). Since the buildup rate is proportional to r^{-6} (r being the distance between the two protons), r can be measured up to distances of 5Å.

Heteronuclear and multidimensional NMR: Nucleic acids above 25 nucleotides show crowded proton (homo nuclear) NMR spectra even in two dimensions. By extending to other nuclei, crowding problems can be reduced or eliminated. Chemical shifts of ^{13}C, ^{15}N, ^{31}P vary over a wider range than protons, increasing the spread of the spectrum. To enhance fraction of the rare isotopes ^{13}C and ^{15}N in the biomolecule, it can be expressed in bacterial culture, use artificially ^{13}C and ^{15}N labeled samples ($^{15}NH_4Cl$, ^{13}C-glucose, ^{13}C-methanol, ^{13}C-acetate depending on the bacteria). This method, the so-called *uniform isotope enrichment* technique, furnishes totally ^{13}C and ^{15}N labeled biomolecules, and has substituted the former difficult chemical labeling techniques. With these labeled biomolecules and complex pulse sequences the spectra can further be spread into three or four dimensions to reduce data density and eliminate ambiguities in assigning NOE and TOSCY cross peaks.

Recently there have been significant advances in NMR solution structure determination [94]. At *residual dipolar coupling* methods, the biomolecule is placed in dilute-crystalline media, which exhibit partial alignment that results in incomplete averaging of their anisotropy properties, such as dipolar coupling and chemical shift anisotropy. These properties yield oriented structural information, rather than distance-based constraints, typical for NMR [129]. A potential advantage of dipolar couplings is their r^{-3} distance dependence, which can allow the observation of longer proton-proton distances (up to 7.4Å) as NOE interactions. *TROSY (Transverse relaxation-optimized spectroscopy)* improves the poor resolution of the important ^{1}H-^{13}C and ^{1}H-^{15}N correlation spectra and has the potential to size of biomolecules amenable to solution NMR studies. Several new methods have been established for direct detection of hydrogen bonds ($\mathbf{N}{-}\mathrm{H}\cdots\mathbf{N}$ with $^{2h}\mathcal{J}_{NN}$ or $\mathbf{N}{-}\mathrm{H}\cdots\mathrm{O{=}C}{-}\mathbf{N}$ with $^{4h}\mathcal{J}_{NN}$ couplings via $^{15}N^{15}N$-COSY experiments)

[29, 146]. These methods are especially valuable for RNA structure determination.

Although NMR is an experimental method, there is a large portion of theoretical input, which is used to determine the required conformation. An initial three-dimensional structure model is first computed with a random structure, and then incrementally adjusted to fit the experimentally determined constraints, given by the data of various experiments mentioned above. Molecular dynamics and simulated annealing computations allow the structure to come to equilibrium with respect to standard inter atomic force fields. This process is repeated using several variations of the initial random model. If successful, all final structures converge to a single conformation with model to model variations of less than 0.4Å rms per atom.

3.2.2 RNA and NMR

In contrast to X-ray structure determination, where the studied biomolecules have reached biological relevant size, the molecules' size that can be investigated by NMR is small. This is especially true for RNA molecules. The largest RNAs amenable to NMR structure studies barely reach 15 kDa, a limit protein NMR passed long ago. Therefore most NMR studies are only done with relevant fragments of the investigated RNA structures. Nevertheless, NMR investigations on nucleic acids are of fundamental importance, because they furnish native molecule conformations in solution and not in crystals, and dynamic data at different temperatures, pH and ionic strength.

Nucleic acids possess *exchangeable* protons (those attached to nitrogen or oxygen) and *not exchangeable* (those attached to carbon). Exchangeable means, that they are rapidly exchanged with the surrounding water. The rate of exchange varies with pH-value and is influenced by hydrogen bonding or decreased accessibility by the solvent. This exchangeability excludes the use of deuterated water (D_2O) as solvent, because the rapid exchange of ^1H with ^2D would delete the corresponding signals. Therefore H_2O must be used in order to characterize the resonances of these protons and specialized techniques are applied to suppress the solvent signal (for example WATERGATE pulse sequence [102] with water flip back pulses [79]). Nevertheless, the exchange with the solvent leads to a broadening of the corresponding peaks. Both, exchangeable and not exchangeable protons can be used to

track along the chain via NOE contacts. Once the nucleotides are assigned, the other NOE contacts can be used to determine the tertiary structure.
^{31}P resonances can be used to link the experimental data of the connected nucleotides via hetero-COSY and hetero-TOCSY [63] and hetero-TOCSY-NOESY experiments. ^{13}C and ^{15}N can be observed at natural abundance in small nucleic acids. With ^{1}H detected heterocorrelation experiments it is straightforward to carry over proton assignments to directly bonded heteronuclei. There are several examples of RNA NMR structures published, which have used these and related experiments. Highest resolution structures (which have most of their protons assigned) are the UUCG- [135] and GNRA-hairpins [52]. Published medium-resolution NMR structures (in which many protons are not assigned) include Loop E from 5S rRNA [144], the sarcin/ricin loop from 28S RNA [127], and a pseudo knot [106].
Nowadays ^{13}C and ^{15}N labeling (isotope enrichment) is extensively used in NMR experiments of RNAs. The first experiments used ^{15}N labeled bases and biosynthetic incorporation into tRNA to identify resonances. Identification was done by direct observation of the $^{15}N -^{1} H$ splitting in the spectrum, and also by establishing $^{15}N -^{1} H$ heteronuclear correlations [40]. In the first experiments only parts of the molecule were labeled and allowed a fast determination of neighbor correlations of these marked nucleotides [113]. Meanwhile, uniform isotope labeling is state of the art, and spreads the possible NMR techniques for RNA extremely. All types of experiments developed for protons can now be expanded to ^{13}C and^{15}N. As mentioned above, $\mathcal{J}_{NN}$ Hetero-$^{15}N^{15}N$-COSY and related methods are of special importance for nucleic acids, because these techniques allow the direct detection of hydrogen bonds between nucleotides via scalar coupling. Such techniques were used to detect the reverse Hoogsteen base pairs of the E-loop of E.coli 5S rRNA [146].
Residual dipolar coupling of RNA is most frequently employed in filamentous phage Pf1, an aligning medium, which has shown an optimal performance for nucleic acids [45, 46]. One application of this technique was the determination of the relative orientation of helical stems in *Escherichia coli* tRNA^{Val} using a small number of residual $^{15}N -^{1} H$ dipolar couplings [93]. The rotational helical parameters can be determined directly from residual dipolar coupling without prior knowledge of the refined structure [130].

4 Computational structure determination of RNA

Conformations and interactions of biomolecules can be most rigorously studied using quantum mechanical methods. These methods solve for the electronic structure of molecules and thus derive the effective Born-Oppenheimer potential for nuclear motion from first principles. *Ab initio* methods solve for the energy and wave functions with the "correct" Hamiltonian. *Semiempirical* quantum mechanical methods simplify this process by leaving out much of the time-consuming part of the calculation, the evaluation of electron repulsion integrals, and making appropriate empirical adjustments to other terms in the Hamiltonian to compensate.

However, none of these quantum mechanical methods are currently able to furnish results for larger biomolecules because the calculations on such systems are either time consuming, or rather accurate, when carried out at an approximate level of quantum mechanical theory. Instead, force field methods, that ignore the electronic motions are used to calculate the energy of systems as a function of the nuclear positions only.

4.1 Force fields

The principles of force fields (also known as molecular mechanics) are based upon Newtonian mechanics. The basic idea is that bond lengths, valence and torsional angles have "natural" values depending on the involved atoms and that molecules try to adjust their geometries to adopt these values as closely as possible. Additionally, steric and electrostatic interactions, mainly represented by van der Waals and Coulomb forces, are included in the so-called potential. Basic ideas for these calculations go back to the work of Andrews in 1930 [3], the first serious applications of force field methods date back to 1946 [53, 30].

The basis of molecular mechanics derives from the accuracy of the Born-Oppenheimer approximation [13], in which one describes the motion of the nuclei of molecules on a so-called "potential surface", caused by the electronic structure. The Born-Oppenheimer approximation works because electrons are so much lighter than nuclei that they respond rapidly to changes in nuclear positions.

A typical force field contains a set of several potential functions which themselves contain adjustable parameters. These parameters are optimized to obtain the best fit of experimental values, as geometries , conformational energies and spectroscopical properties. It is important to realize that force fields are usually parameterized for a limited set of molecular properties and a specific set of molecules. If some parameters are not experimental available, quantum mechanical calculation of representative fragments can be used to obtain the desired values.

In principle there are two basic methods that can be used to actually obtain the parameters. The first approach is *"parametrization by trial and error"*, in which the parameters are gradually refined to give better and better fits to data. Problems arise when the dataset is large, because it is difficult to simultaneously modify a large number of parameters. Therefore it is usual to perform the parametrization in stages, for example starting with the van der Waals parameters, continuing with the electrostatic interactions and ending with the torsion potentials (bond and angles parameters, the so called hard degrees of freedom, are usually transferred from one force field to another without modification). Of course, it may be necessary to modify any of the parameters at any stage should the results be inadequate and so parametrization is invariably an iterative procedure.
The second approach to parametrization, pioneered by Lifson and coworkers in the development of their *"consistent" force fields*, is to use least-squares minimization [41, 42, 77, 97]. The objective is to change the force field parameters for minimizing the error, as the sum of squares of the differences between the observed and calculated value for a given set of parameters. The advantage of this approach is a well defined precise and automated optimation. Nevertheless there are several disadvantages like the enormous computational effort, due to the large data sets and most important the fact that least square optimation depends on all variables being measured in the same units. Therefore the method is easily modified to enable various weighting factors to be assigned to the different pieces of experimental data [141], so that for example the thermodynamic data could give greater importance than vibrational frequencies. Probably the best way for parametrization is a combination of both methods, using "intuition" to get reasonable starting values and numerical methods when huge amounts of data are involved.

4.1.1 Energy Calculation

Many of the molecular modeling force fields in use today can be interpreted in terms of a relatively simple four component picture of intra- and intermolecular forces within the system.

$$E_{total} = E_{bond} + E_{angle} + E_{torsion} + E_{non-bonding}$$

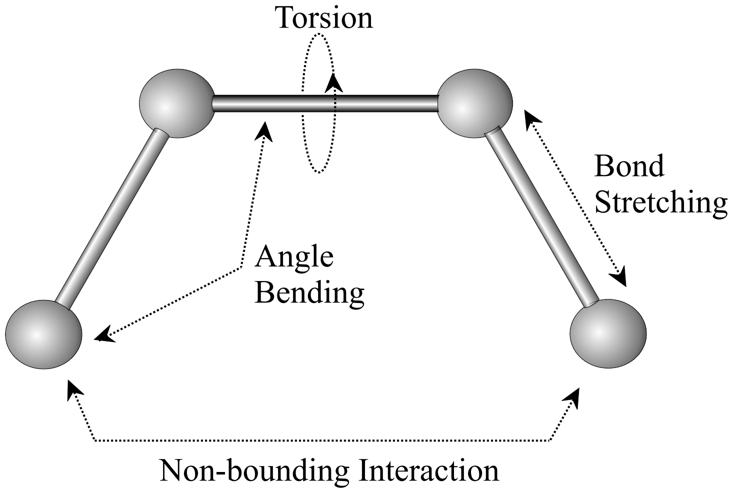


Figure 9: Four component picture of inter- and intramolecular forces

In the simplest approach the energy terms are in detail :

Bond - Energy:
The energy between two bonded atoms increases, when the bond is compressed or stretched. The potential is described by an equation based on Hooke's law for springs.

$$E_{bond} = \sum_{bonds} k_b(r - r_0)^2$$

whereby k_b is the force constant, r is the actual bond length and r_0 the equilibrium length. This quadratic approximation fails as the bond is stretched towards the point of dissociation.

Angle Energy:
Energy increases if the equilibrium bond angles are bent. Again the approximation is harmonic and uses Hooke's law.

$$E_{angle} = \sum_{angles} k_\theta(\theta - \theta_0)^2$$

k_θ controls the stiffness of the angle, θ is the current bond angle and θ_0 the equilibrium angle. Both, the force and equilibrium constant have to be estimated for each triple of atoms.

Torsion Energy:
Intra-molecular rotations (around torsions or dihedrals) require energy as well:

$$E_{torsion} = \sum_{torsions} \frac{V_n}{2}(1 + cos(n\omega - \gamma))$$

V_n controls the amplitude of this periodic function, n is the multiplicity, and γ the so-called phase factor, shifts the entire curve along the rotation angle axis ω. Again the parameters V_n, n and γ for all combinations of four atoms have to be determined.

Non-bonding Energy:
The simplest potential for non-bonding interactions includes two terms, a Van der Waals and a Coulomb term.

$$E_{non-bonding} = \underbrace{\sum_i \sum_{j>i} \left(\frac{A_{ij}}{r_{ij}^6} - \frac{B_{ij}}{r_{ij}^{12}} \right)}_{\text{Van der Waals}} + \underbrace{\sum_i \sum_{j>i} \frac{q_i q_j}{r_{ij}}}_{Coulomb}$$

The Van der Waals term accounts for the attraction and the Coulomb term for electrostatic interaction. The shown approximation for the van der Waals energy is of the Lennard-Jones 6-12 potential type.

These simple terms mentioned above can be expanded to adjust the potentials better to the experimental results (e.g Morse potential for bonds, Taylor expansions with higher terms, cross-terms between the potentials), but with the disadvantage of higher calculational effort. That is the reason why biomolecular force fields usually do not include refinement terms for the

bond, angle and torsion potential.
Sometimes force fields include additional potential terms for specific interactions, such as hydrogen bonding or dipole-dipole interaction. Typical example is the hydrogen bonding term in the AMBER force field (see section 4.1.4).
The most critical term, even for biomolecules, are the non-bonded interactions. First the number of non-bonded interactions in a molecule grows as $\frac{n(n-1)}{2}$, where n is the number of atoms in the molecule. Choosing a complex term for the non-bonding interactions results in a tremendous increase of computational effort. Second this non-bonded interaction term must include the solvatation effects, because biomolecules are usually solvated in water. This solvatation has a major influence on the electrostatic forces. The most accurate way for describing this solvatation is including the solvent and counter-ions explicitly. Such an "explicit solvent" approach increases the number of particles considerably, because a lot of solvent molecules are need for an accurate description of solvation.
Other approaches, named "implicit models", represent the environment (counter-ion, solvent) around macromolecules as a continuum. Such models must describe the damping of the electrostatic interaction by the solvent in an appropriate way. The simplest way to model damping effects is to increase the permittivity, most easily by using an appropriate value for the relative permittivity in the Coulomb's law equation (i.e. $\varepsilon_{eff} = \varepsilon_0 \varepsilon_r$). A better description of the dielectric damping can be introduced by a distant dependent dielectric function $\varepsilon_{eff}(r)$. This dependence can be simple linear or most common sigmoidal. One example of such a function is [121]:

$$\varepsilon_{eff}(r) = \varepsilon_r - \frac{(\varepsilon_r - 1)}{2}\left[(rS)^2 + 2rS + 2\right] e^{-rS}$$

The value of $\varepsilon_{eff}(r)$ varies from 1 at zero separation to ε_r (the bulk permittivity of the solvent) at large distances, in a manner determined by the parameter S (which is typically given a value between 0.15Å and 0.30Å).

Another approach to incorporate solvent effects, is the *generalized Born equation*, which has been widely used to represent the electrostatic contribution to the free energy of solvation. The electrostatic term is expanded by a second term, which contributes the solvent effects.

$$E_{elec} = \underbrace{\sum_i^N \sum_{j=i+1}^N \frac{q_i q_j}{r_{ij}}}_{Coulomb} \underbrace{- \frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right) \sum_i^N \sum_{j=i+1}^N \frac{q_i q_j}{f(r_{ij}, a_{ij})}}_{\text{generalized Born}}$$

$f(r_{ij}, a_{ij})$ depends on the inter-particle distances r_{ij} and the Born radii a_i. A variety of expressions is possible for function f; that one proposed by Still and co-workers [123] was:

$$f(r_{ij}, a_{ij}) = \sqrt{r_{ij}^2 + a_{ij}^2 e^{-D}}, \text{where } a_{ij} = \sqrt{a_i a_j} \text{ and } D = r_{ij}^2/(2a_{ij})^2$$

This form of the function f can be physically justified, and has the advantage that it can be differentiated analytically, thereby enabling the solvation term to be included in gradient-based optimization methods and molecular dynamics simulations.

4.1.2 Structure optimation

Calculating the energy with respect to a given conformation is only one part of optimizing the structure of molecules. For improving the structure it is necessary to change the geometry in such a way, that the total energy is lowered. This process is repeated iteratively so that an energy minimization corresponds to a geometry optimation. The potential function is a function of a large number of variables which specify the molecule's geometry either in internal or Cartesian coordinates. The ideal solution for geometry optimation would be the **global minimum** of this function corresponding to the molecule in a state of minimal free energy. Since there is no method available to determine the global minimum of a function of many variable, optimation algorithms are usually trapped in a **local minimum**. This behavior is often referred to as the 'global minimum problem'. As a consequence of ending the optimation procedure in a local minimum, different optimized structures will be achieved, depending one the starting geometry. Therefore it is usually necessary to use different starting geometries and compare the obtained structures to get lower energies.

Minimization methods: Minimization problems are known for a long time since they occur in various fields of science. The great interest of general optimation procedures resulted in a great variability of algorithm available (for an excellent overview see Leach chapter 4 [72]).

Probably the most frequently used optimation algorithms in molecular modeling are first-order minimization methods, namely the **steepest decent** algorithm and the **conjugate gradient** method. Both techniques use the

first derivative of the potential function. While the steepest decent method changes the geometry of the molecule along the largest energy gradient, the conjugate gradient method calculates a path, which is a combination of the current gradient and the gradient of the point before. The advantage of the conjugate gradient method is the faster convergence (especially in narrow valleys of the energy landscape).
Second-order methods do not only use the first derivatives but also the second derivatives to locate the minimum. The great advantage of second-order methods is a faster convergence (even in the vicinity of a minimum). The second derivatives can be calculated numerically or analytically. The most favorite example of these methods is the **Newton-Rawson** method.

All techniques mentioned so far are based on pure analytical calculations and trap usually in one of the local minimums near the starting point. A useful approach to overcome this problem is the implementation of some kind of randomness (namely stochastic techniques) as it is done by method of **simulated annealing**. Simulated annealing is a widely used optimation procedure that originally came from the field of statistical physics (e.g.[65]). In effect it tries to simulate the cooling and the crystallization process occurring in a heated solid. Starting point is the configuration space Ψ and a so-called energy function U, which is defined as follows: $U : \Psi \rightarrow \mathbb{R}$. In the case of molecular mechanics U corresponds with the potential function whereas Ψ is the conformational space constructed from all possible conformations of the molecule. Beginning from a starting geometry the energy E_0 of the molecule is calculated. That is followed by a random step in conformational space, which equals a random change of the molecular geometry. Again the energy is calculated resulting in energy E_1. Now there are two possibilities: If $E_1 < E_0$, the random-step is accepted in any case. If $E_1 > E_0$ is only accepted with the probability p:

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

p is the probability of accepting the new conformation as a new starting structure, and k and T are the Boltzmann constant and the temperature in Kelvin, respectively. This criteria is also known as the Metropolis algorithm [90]. It ensures that the optimation cannot be trapped in a local minimum since higher energies are accepted with a certain probability so that energetic

barriers can be overcome. If n is the number of simulated annealing steps the global minimum is always found for $n \rightarrow \infty$. A typical simulated annealing procedure starts at high temperature T to warrant that the random walk overcomes the highest barriers and reaches most of the conformational space. Then the temperature is lowered by a certain scheme (the so-called cooling schedule) and the molecule is trapped in the conformation, it has entered most often.
Simulated annealing is most useful for systems that are not very restricted Usually good results are obtained using a high computer effort.

4.1.3 Molecular dynamics

In molecular dynamics, successive configurations of a system are generated by integrating Newton's laws of motion. The result is a trajectory that specifies how the positions and velocities of the particle in the system vary with time. Molecular dynamics was initiated by Alder and Wainwright 1957 [2].

$$F_i(t) = m_i a_i(t) = m_i \frac{\partial^2 r_i(t)}{\partial t^2}, \text{ whereas } F_i(t) = -\frac{\partial E_{tot}}{\partial r_i}$$

The forces of the atoms are the negative gradient of the potential energy E_{tot}. Under the influence of a continuous potential the motions of all the particles are coupled together, giving rise to a many-body problem that cannot be solved analytically. Therefore the equations of motion are integrated using a *finite difference method*. As basic idea the integration is broken down into small stages, each separated in time by a fixed time δt. The accelerations a_i of the particles are available from the force F_i, calculated from E_{tot}. The accelerations a_i are then combined with the positions and velocities at a time t to calculate the positions and velocities at a time $t + \delta t$. There are several algorithms for integrating the equations of motion using finite difference methods, e.g the Verlet [136], the leap-frog [55], the velocity Verlet [126], and the Beeman algorithm [8]. Choosing an appropriate time step δt is essential for a successful molecular dynamics simulation: If δt is too small the trajectory will cover only a limited part of the phase space. If δt is too large instabilities may arise in the integration algorithm due to high energy overlaps between atoms. Typical δt for all-atom force fields with no constraints is 1 femtosecond. As the process of folding takes place in a millisecond scale, the simulation of biomolecular folding is not within the reach of present day computers.

Since the time step δt of a molecular dynamics simulation is dictated by the highest frequency motion (e.g. bond vibrations), it is possible to enlarge the time step, through fixing bond and angles at a specific value via constraints. The most commonly used method for applying constraints, particularly in molecular dynamics, is the SHAKE algorithm [110].

The most time consuming element of a molecular dynamics simulation is usually the calculation of the forces, and therein the calculation of the non-bonding interactions. There are several approaches to reduce this calculation effort. A very crude method is the use of a *cut-off* distance beyond which atoms are no longer considered to interact. Shifting and switching functions smooth the gradient over this distance. Another technique is the so-called *united atom method*, where atom groups with non-polar hydrogen atoms are treated as an ensemble.

The inclusion of the solvent can be done explicitly where the solute is immersed in a cubic box of solvent molecules. The use of non-rectangular periodic boundary conditions, stochastic boundaries and "solvent shell" can help to reduce the number of solvent molecules required and therefore accelerate the molecular dynamic simulation.

When using implicit solvent models in molecular dynamics simulations, there are two additional effects to bear in mind. The solvent also influences the dynamical behavior of the solute via (1) random collisions, and by (2) imposing a frictional drag on the motion of the solute through the solvent. While explicit solvent calculations include these effects automatically, it is also possible to incorporate these effects of solvent without requiring any explicit specific solvent molecules to be present. The *Langevin equation* of motion is the starting point for these *stochastic dynamics* models.

$$\frac{m_i \partial^2 r_i(t)}{\partial t^2} = \mathbf{F}_i(r_i(t)) - \gamma_i m_i \frac{\partial r_i(t)}{\partial t} + \mathbf{R}_i(t)$$

The first component is due to interactions between the particle and other particles. The second force arises from the motion of the particle through the solvent and is equivalent to the frictional drag on the particle due to the solvent. γ_i is often referred to as the friction coefficient. The third contribution, the force $\mathbf{R}_i(t)$ is due to random fluctuations caused by interactions with solvent molecules.

By approximately 1995, stable nanosecond-length molecular dynamics simulations of nucleic acids' structures in solution were becoming routine. Before, the simulations were plagued by instabilities, owing largely to the application - necessitated by limits in computational power - of approximation methods, which lack the rigor required to reasonable represented highly charged systems, such as nucleic acids. Simulations were characterized by distortion of duplex structures, broken base pairing, and misrepresented sequence-specific fine structures. Tricks were often applied to generate stable simulations (such as addition of artificial Watson Crick base pair restraints, reduced phosphate charges, etc). With the increase in computational power, the development of more reliable force fields, and an accurate treatment of electrostatic interactions (Ewalds methods, atom-based force shifting or modern implicit solvent models), it is now possible to carry out molecular dynamics simulations of RNA without these tricks.

One of the first modern simulations, applying Ewalds methods and OPLS force field, was done 1995 by Zichi [150]. In this simulation, a hairpin was stable over nanoseconds. Also the force field of Cornell et al. (AMBER) showed the expected stabilization of A-form RNA, when tested by Cheatham and coworkers [20]. One limitation of the Cornell et. al. force field however is B-RNA↔A-RNA transitions. Using this force field [20] B-RNA (i.e. RNA in a B-DNA-like conformation) is stable over more than 10 nanoseconds, despite no experimental evidence of this conformation.

Excellent representations of larger RNA structures such as the hammerhead ribozyme [50, 51] and the tRNA^{Asp} (Anticodon hairpin [5] and the whole [5]) have been achieved by using the modern explicit solvent model simulation methods in combination with force field methods. Simulations with the implicit solvent models, Poisson-Boltzmann and generalized Born methods have also furnished reliable results [131, 143].

4.1.4 Programs

There are several force field program packages available for biomolecular computation. General to all these force fields are simple approaches for bond, angle and torsion potentials (as described in section 4.1.1) to reduce the calculation time for the energy function and the gradient. For the important nonbonded interactions, several implicit and explicit solvent techniques, mentioned above, are implemented.

The most prominent of these force fields is the Cornell force field of AMBER, which is not only used in the AMBER packages, but is also included in various other program packages (e.g. NAB, JUMNA). Other examples of force fields

are CHARMM (**C**hemistry at **HAR**vard **M**olecular **M**echanics) [15, 82] and GROMOS (**GRO**ningen **MOL**ecular **S**imulation System) [133, 134]. The potentials in AMBER, CHARMM and GROMOS have the same basic structure as described in section 4.1.1. Only AMBER has an additional energy term for an adequate description of hydrogen bonds. Beside these force fields there are some other packages for simulating biomolecules including other energy term expressions, like DREIDING [87] and Tripos 5.2 [22].
The next two paragraphs describe two prominent force fields packages, AMBER and JUMNA respectively, AMBER is a general force field package for biomolecules, whereas JUMNA is especially created for nucleic acids.

AMBER One of the most widely used force fields is AMBER (**A**ssited **M**odel **B**uilding with **E**nergy **R**efinement) [101]. It is suitable for the calculation of the two most important types of macromolecules in biochemistry, namely peptides and nucleic acids. There is a difference between the AMBER program package and the so called AMBER force field, which is implemented in the AMBER package, but also in various other programs. The force field is public domain, whereas the package is distributed under license agreement. The current version of the package, AMBER 6.0 [17] is comprised of several modules that fulfill specific tasks. Figure 10 illustrates the information flow between the AMBER 6.0 modules. Inputs supplied by the user are represented by circles, whereas modules drawn as a box stand for the actual programs.
There are four major input data to AMBER modules:

(1) Cartesian coordinates for each atom in the system

(2) "Topology": connectivity, atom names, atom types, residue names and charges

(3) Force field: Parameters for all of the bonds, angles, dihedrals and state parameters desired

(4) Commands: The user specifies the procedural option and state parameters desired

The modules in figure 10 can be divided into three categories:

- **Preparatory programs:** LEaP is the primary program to create the amber specific topology file prmtop and the coordinate file prmcrd.

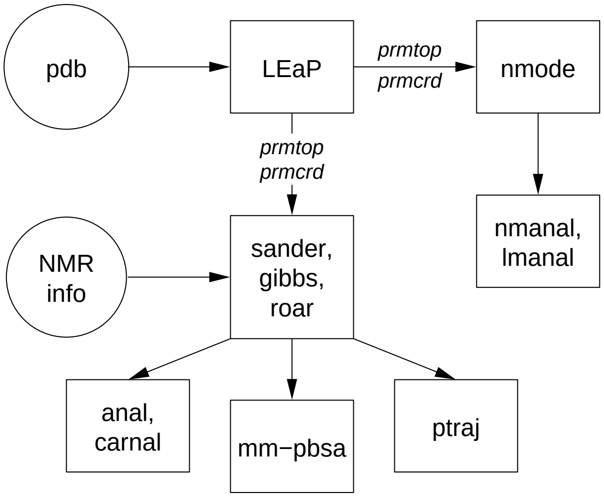


Figure 10: Basic information flow in AMBER 6.0

- **Energy programs:** SANDER is the energy minimizer and molecular dynamics module, GIBBS the free energy perturbation program, NMODE the normal mode analysis program and ROAR a module, where parts of the molecule can treated quantum mechanically and others with molecular mechanics.
- **Analysis programs:** ANAL is created for analyzing single conformations, CARNAL to examine molecular dynamics simulations.

The AMBER force field, or better the Cornell force field, consists of five potential terms

$$\begin{aligned} E_{total} &= \sum_{bonds} K_b (r - r_0)^2 + \\ &+ \sum_{angles} K_\theta (\theta - \theta_0)^2 + \\ &+ \sum_{torsions} \frac{V_n}{2} (1 + cos(n\omega - \gamma)) + \\ &+ \sum_i \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\epsilon r_{ij}} \right] + \\ &+ \sum_{H-bonds} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{10}} \right] \end{aligned}$$

The commonly used parameter set for the AMBER force fields was published by Cornell *et al.* 1995 [23], slight reparametrizations were done to adjust to nucleic acids 1999 [19, 137].

Another very useful program package, which includes AMBER force field, is NAB (Nucleic Acid Builder) written by David Case and Thomas Macke [83]. It's source code is free available via anonymous ftp at `ftp.scripps.edu`. NAB is a high-level description language that facilitates manipulation of macromolecules and their fragments. Further included are simple conformation build-up procedures via distance geometry or rigid-body transformation, molecular mechanics and dynamics methods. The force field works in an AMBER like environment (using coordinate and topology files [prmtop]), with the option to use the pairwise generalized Born model for solvation.

JUMNA 10 stands for **Ju**ction **M**inimization of **N**ucleic **A**cids and is a molecular mechanics program that was designed by Richard Lavery and Heinz Sklenar [69, 66, 68, 71] especially for dealing with nucleic acid structures. JUMNA differs from AMBER not only in the specialization to nucleic acids but also in a different force field (JUMNA uses the FLEX [67, 69, 70]) [or as additional option the AMBER] force field) and in a different description of molecular structure.

The starting point of the JUMNA algorithm is to split nucleic acid fragments into a collection of 3′-mono-phosphates (with the exception of the 3′-termini which are simple nucleosides). This division is achieved by cutting the O5′-C5′ bonds of the phosphodiester backbone. These nucleotides are positioned with respect to a local helical axis with a set of 6 helicoidal parameters (according to the Cambridge convention [25]). These helicoidal variables consist of three translations (Xdisp, Ydisp, and Rise) and three rotations (Inclination, Tip, and Twist).

The internal flexibility of each nucleotide (see figure 11) is represended by the sugar ring flexibility and dihedrals at the glycosidic link χ and within the phosphodiester backbone ϵ and ζ. To allow conformational changes in the ribose, the sugar ring is broken at the C4′-O1′ bond to get a linearized system with 5 degrees of freedom, three valence angles ($\nu1$: O1′-C1′-C2′, $\nu2$: C1′-C2′-C3′, $\nu3$: C2′-C3′-C4′,) and two dihedral angles ($\tau1$: O1′-C1′-C2′-C3′, $\tau2$: C1′-C2′-C3′-C4′). Valence angles outside the sugar moiety and bond length within the nucleotide are taken to be fixed.

The structure of the fragment can then be energy optimized in terms of helicoidal parameters plus variables describing the internal conformation of each nucleotide (glycosidic angle (χ), sugar torsions ($\tau1$, $\tau2$) and valence angles ($\nu1$,$\nu2$,$\nu3$) and two backbone torsions ϵ and ζ). The remaining back-

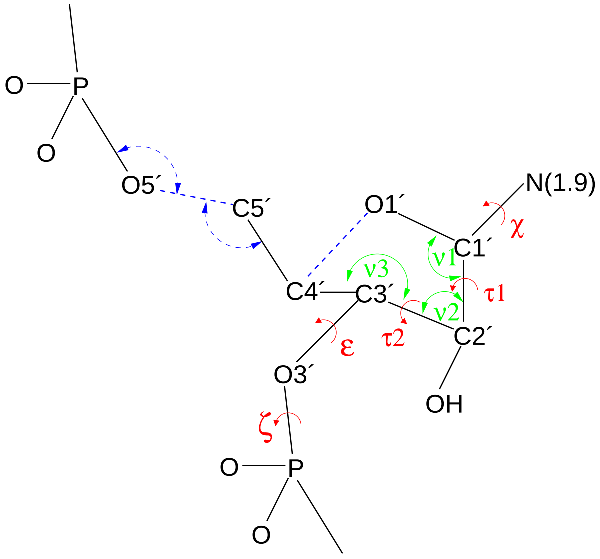


Figure 11: Structure of the internal variables in JUMNA:
green = valence angles of the broken sugar ring
red = dihedrals
blue (dashed) = harmonic constraints

bone torsions are treated as dependent variables. Four harmonic constraints ensure that the sugar rings and the phosphodiester junctions between successive nucleotides close properly during energy minimization (blue dashed lines in figure 11). One distance constraint, O5′-C5′, and two angle constraints P-O5′-C5′ and O5′-C5′-C4′, are used per nucleotide junction, and one distance constraint, C4′-O1′ for the sugar ring closure. This approach leads to an important reduction in the number of variables required compared to classical molecular mechanical algorithms and also gives more control over the conformations which are generated. Dielectric conditions can be varied through the use of a sigmoidal distance dependent dielectric function of variable slope and plateau, the use of a chosen fixed dielectric constant or the function $\epsilon = nr$. The net charge on each phosphate group can also be varied to mimic counter-ion screening. Explicit mobile counter-ions or water molecules can also be included through a ligand option.

JUMNA can build, manipulate and energy optimize fragments of DNA or

RNA having up to 4 strands. Many structural features can be fixed during minimization and certain global or local features can be constrained such as base opening angle, average twist or rise per base step, radius of curvature, sugar phase and amplitude, atom pair distances, and torsion and valence angles. This ensures an easy use of experimental data like atom-atom distances determined by NMR. The simple use of constraints and the representation of the molecule in terms of helicoidal and backbone parameters are the most powerful features of JUMNA, since the description of molecular geometry is thus sequence independent, so that the effect of sequence changes can be tested quite easily.

4.2 Conformation build-up programs

Because of the global minimum problem, force field programs need accurate three-dimensional starting structures. There are two sources for relevant starting conformations: experimental structures at atomic resolution (obtained by X-ray crystallography and NMR methods) and conformation build-up programs. These conformation build-up programs construct the biomolecules with the knowledge of the sequence and several constraints, such as basepair patterns, stacking interactions etc. (see example in paragraph of MC-SYM). Build-up programs take this information and build up the molecule with modules (nucleotides, in the case of nucleic acids). There are several algorithms for building molecules under different constraints. For example MC-SYM creates structures from a small library of conformers for each of the four nucleotides based on transformation matrices for each base. Building up conformers from these starting blocks can quickly generate a very large tree of structures. This tree can be pruned. In a related approach, Erie *et al.* [32] used a Monte-Carlo procedure based on sets of low energy dinucleotide conformers.

Another very general approach for finding conformers, which fulfill several constraints, is *distance geometry*. The whole structural information, bonds, angles, torsions and distance constraints are translated in a distance bounds matrix. This matrix contains the maximum and minimum values permitted to each interatomic distance in the molecule. A procedure called triangle smoothing is then used to refine the initial set of distance bounds via triangle inequation. Then random values are assigned to all interatomic distances between the upper and lower bounds to give a trial distance matrix. This matrix is then subjected to a process called embedding, in which the distance matrix is converted to a set of atomic Cartesian coordinates.

4.2.1 MC-SYM

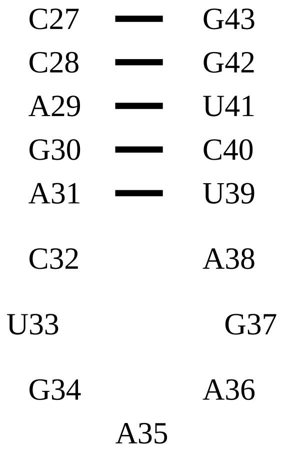
MC-SYM stands for **M**acromolecular **C**onformations by **SYM**bolic program and is a high-level molecular description language used to describe single stranded RNA molecules in terms of functional constraints. It is the most prominent bild-up program for nucleic acids and was written and tested by the group of Cedergren and Gautheret [38, 85, 111]. The current version, MC-SYM 3.1, is distributed under license agreement. The version described here is the older, free available version of the program.

A backtracking algorithm in MC-SYM searches the conformational space of an RNA molecule and all geometries that fulfill the constraints are returned in PDB-format to be optimized by a force field program. The conformational space explored is determined by the choice of pre-computed nucleotide conformations and transformations. MC-SYM has been successfully used for RNA hairpins [85, 38], for tRNAs [84], or for the Rev-binding site of HIV-1 [73].

The program input for MC-SYM consists of a simple ASCII-file divided into two sections. The first section, the so-called "sequence-section", defines the sequence and secondary structural information of a macromolecule. It lists all the nucleotides and fragments that compose the RNA and information on how these parts are connected or related to others. The second section, the "constraints-section" consists of additional constraints which might be local (i.e. they are valid for just one base or a base pair) or global (i.e. they are valid for all nucleotides).

The following example shows the description of a simple stem-loop structure (RNA hairpin) and was taken from the MC-SYM manual (see figure 12). The modeled molecule is the anticodon stem-loop of a tRNA. The secondary structure shown on the left-hand side of figure 12 indicates that bases C27 to A31 form base pairs with G43 to U39. It is assumed that bases A38 to G34 are stacked and as a first attempt C32 over A31 and U33 over C32 are stacked as well (following a quite common strategy in RNA modeling that tries to maximize stacking). These assumptions lead to the input file shown in figure 12. In the first section of the input file a typical line consists of several entries of the following format:

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



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

Figure 12: Input file for MC-SYM for a simple stem-loop structure

•) *nucleotide-type*: gives the sequence of the molecule and can be one of rA, rC, rG, or rU.
•) *nucleotide-identifier*: a unique number identifying a certain nucleotide.
•) *connection-function*: a keyword that specifies the position of the current nucleotide relative to another. Keywords can be chosen from a wide range of possibilities such as all kinds of base-pairs (Watson-Crick, Hoogsteen, reverse Hoogsteen, Wobble, unusual base pairs like G-A, base pairs with different numbers of hydrogen bonds, . . .), standard RNA or DNA helix forms, stacking, or simple connections between two adjacent bases.
•) *reference-nucleotide*: the number of an already defined nucleotide which the connection-function refers to.
•) *conformational-set*: a set of pre-computed conformations and transformations which is taken from a database. This set comprises the "allowed" movements for the given nucleotides. The "allowed" movements range from a simple "type_A", which stands for a base in C3′-endo conformation taken from an A-RNA helix, to the keyword "sample+", which represents a total of 59 different conformations and transformations. The total number of conformations in the example is 6561 (= 3^8). This stems from the combination of 9 A-type nucleotides ("type_A", 1 conformation) and 8 A-type nucleotides stacked over other A-type bases ("stk_AA", 3 conformations).

Whereas the first part of the input file specifies the largest possible search tree for the MC-SYM run, the following section (starting with keyword "ADJACENCY") reduces the number of possible conformations significantly by introducing a number of constraints. The "ADJACENCY" keyword refers to the O3′-P bonds in the molecule and is used when MC-SYM detects a loop-construction (i.e. when unpaired bases are not at the end of a stem, but between paired regions). In the given example this distance may vary between 1 and 4 Å. Adding the "ADJACENCY" section to the input file reduces the number of conformations to 645. In the following "CONSTRAINT" section an example for a local constraint can be seen. It is specified that the distance between atoms O3′ of U33 and P of G34 must be larger than 1 Å and must not be greater than 3 Å , thus reducing the number of possible conformations to 56. The last section, labeled "GLOBAL", is for definition of global constraints that are valid for all nucleotides in the molecule. Exemplified for figure 12 only conformations in with P and C1′ atoms at least 3.5 Å apart are acceptable. This reduces the total number to 52 different geometries.

MC-SYM is a very handy tool useful for finding possible molecular geometries in cases the secondary structure and some additional data are available. For small molecules it can also be used to generate a “pool” of starting geometries only based on known secondary structure. These starting geometries can then be minimized by a force field program and the “best” geometries (in terms of energy) can then be selected for further optimization.

4.3 A short glance on protein structure prediction

Before creating a simplified three dimensional model of RNA it can be very useful to take a short glance on three dimensional structure prediction of proteins. This should be done for several reasons:

- Proteins are the most extensively studied biopolymers in science. Therefore a lot of experimental and theoretical approaches have been made.
- Proteins are linear heteropolymers like nucleic acids.
- Protein structure prediction must always include some kind of three dimensional information.
- Several simplified models have been used to study the protein folding process.

4.3.1 Structure of proteins

Proteins are build up from monomers, the amino acids. Proteins are assembled by a chain of amino acids and are therefore linear polymers. The 20 fundamental amino acids are shown in figure 13. Monomers are connected via amide bonds between the α-amino group of one amino acid and the carboxylgroup of the neighboring amino acid. Each amide bond is characterized by a planar structure involving the carboxyl carbon and its oxygen plus the α-nitrogen and its hydrogen. This characteristic is caused by the π-bonded electron cloud that extends along the $O - C - N$ set of atoms (see figure 14). Therefore the only relevant torsion angles of the backbone are the angles between Φ and Ψ.
It is easily possible to define a primary, secondary and tertiary structure of proteins. The term primary structure is used for the protein sequence (starting at the amino group side), secondary structure includes structure motifs (e.g. α-helices and β-sheets), and tertiary structure, gives additionally information about the arrangements of these secondary structure motifs.

4.3.2 Modeling of proteins

Structure calculations of proteins at atomar resolution use the same force field machinery as nucleic acids' calculations. Actually most of these force fields

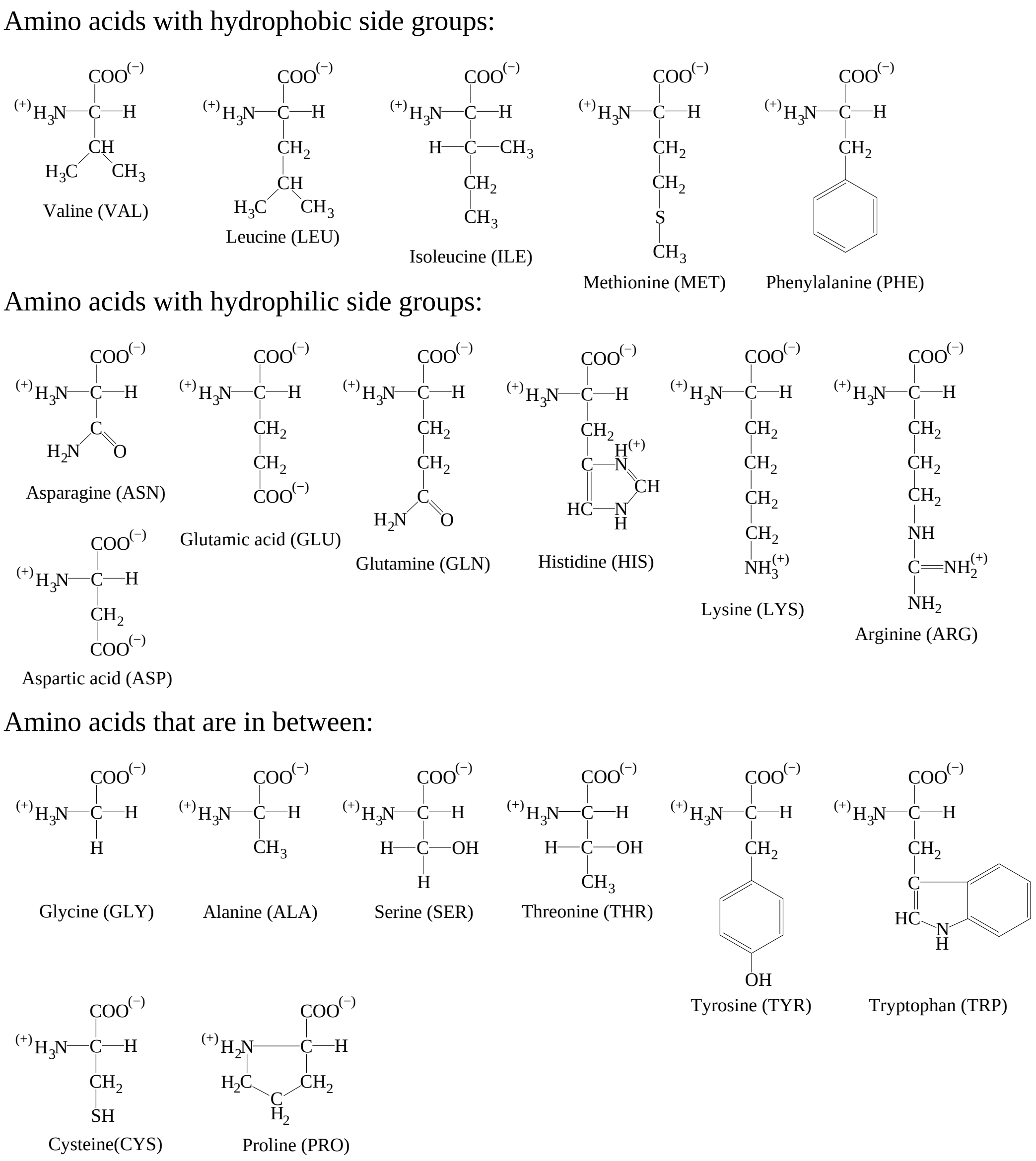


Figure 13: The structure of the 20 fundamental amino acids, classified into three groups depending on the character of their side chains

are originally invented for protein structure determination and not for nucleic acid conformational analysis. Nevertheless similar problems arise, when calculating proteins or nucleic acids at this level. As long range electrostatics

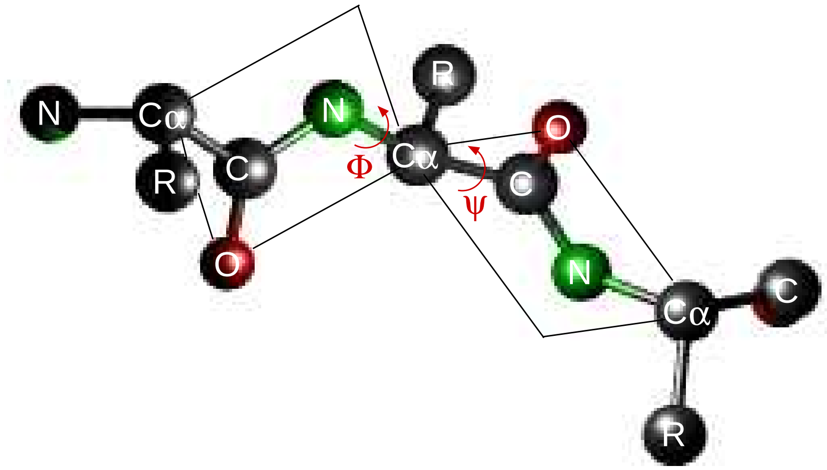


Figure 14: Structure of the protein backbone, including the characteristic planar amide-bond (represented by rectangle). Φ and Ψ are the only relevant torsion angles of the backbone.

are more important for nucleic acids additional difficulties occur. One of the crucial problems is the large number of degrees of freedom, which leads to the global minimum problem, as mentioned before. Consequently it is still impossible to determine the minimum energy structure for larger proteins based on the knowledge of only their sequence. To get rid of this problem, many approaches have been made to reduce the conformational space. Most approaches work with reduced amino acid representations. The simplest approaches use only one representative pseudo atom per amino acid (mostly Cα sometimes Cβ), extended versions include additional pseudo atoms for the side chains.

A further simplification can be achieved by the use of *lattice models*, where the pseudo atoms can only cover specific points in a lattice grid. There are two types of lattice model simulations, aiming at two distinct objectives. One was designed to understand the basic physics governing the protein folding process. The key feature of this lattice type is its simplicity. The energy evaluation on such a lattice model can be achieved quite efficiently. Based on this type of models, methods involving exhaustive searches of the available conformational space became feasible. However, most of these models are unable to describe subtle geometric aspects of proteins' conformation. Prominent examples of this type are the models of Go and coworkers [39] and the HP-model of Dill [26, 27].

Lattice models by Skolnick et al. [120], Miyazawa and Jernigan [91, 92] belong to the second category of lattice models. These models are geared towards realistic folding of real proteins. They are parametrized using measured protein structures. By statistical sampling of such available structures model templates are created. The resulting potentials are often referred to as statistical potentials. Works by Crippen [24], Eisenberg at al. [14], and Sippl et al. [49] are further examples of this category.

Both approaches can be uncoupled from the lattice condition, resulting in the so called *off-lattice models*. The origin of off-lattice models can be found in the works of Warshel and Levitt [76, 75]. In the simplest approaches the protein is represented by a chain of balls (amino acids) connected via stiff bonds. All energy functions used in lattice models have also been used in off-lattice models (e.g. [99, 119]). Meanwhile various extended versions of off-lattice models representing each amino acid with more pseudo atoms have been invented an extensively studied (e.g. [75, 80, 81]).

4.3.3 Comparison between proteins and RNA

After this short glance on the various models created for determing the proteins' conformation and enlightening the proteins' folding process one might think that it is possible to transfer these ideas into the RNA world. A closer view shows that this can be done only with several restrictions. Although both biomolecules show some similarities, there are fundamental differences according to their chemical structure, the conformation constructing forces and the folding process:

(1) The structural differences between the single nucleotides are small and only found in different purine and pyrimidine structures. In contrast the side chains of proteins show remarkable varieties (see figure 13).

(2) The existence of secondary structure elements for proteins (α-helices, parallel and anti-parallel β-sheets, β-turns, random coils, etc.) is contextual, i.e. these elements are formed and stable in context of the proteins' environment. But they are not formed when they are isolated

in solution. Contrary to proteins, there are only four basic secondary structure elements in RNA (helices, loops, bulges, and junctions). The helices are A-form Watson-Crick duplexes; the loops, bulges and junctions arise in non-Watson-Crick regions terminated by one or more helices. Because the energies involved in the secondary structure formation are larger than those involved in tertiary interaction, secondary structure elements can be formed and be stable by themselves. Thus, the energetics of secondary and tertiary structural elements are separable. Consequently it is possible to treat the energy of tertiary interactions as a perturbation on the energy stabilizing the secondary structure.

(3) There are considerable differences in the character of the backbone. First, the backbone of the nucleic acids is multiple charged, leading to a strong influence of the conformation with respect to counter ions, pH and ionic strength of the solvent. Second, the ribose-phosphate backbone of nucleic acids is more flexible than the polyamid backbone of proteins. While two dihedrals (Φ, Ψ) are adequate to describe proteins' backbone flexibility, six dihedrals are necessary for an accurate description of a nucleic acids' backbone (see section 2.3.3).

(4) The non-bonded interactions between nucleotides show characteristic structure patterns, not only for adjacent nucleotides. These patterns are mainly generated through hydrogen bonds between the nitrogen bases and the $2'$-hydroxy group of the ribose (see section 2.3.1). These patterns allow a classification of the interaction between two nucleotides. Even the more unspecific stacking interaction between two nitrogen bases has a definable character depending whether stacking takes place or not. In contrast to these clear arrangements in nucleic acids, proteins show diffuse interactions between amino acids. Adjacent amino acids show hydrogen bond patterns resulting in typical secondary structure motifs. Contacts between non adjacent amino acids are mostly not as sharply defined as nucleotides' interactions. But these diffuse structured interactions facilitate the treatment of protein interactions with statistical methods.

Summarizing these points it becomes clear that transforming the ideas of simplified protein models to nucleic acids must be exercised with caution.

Interactions between nucleotides are much more specific and require other approaches than aminoacids' interactions.

5 The program `toyRNA`

5.1 General aspects

The large variability of possible simplified models for proteins (as described in subsection 4.3) necessitates an exact analysis of the objective target of a simplification. The aim of this study was the creation of `toyRNA`, a program for linking the secondary and three dimensional RNA structure. Therefore the non spacious information of RNA secondary structure must be converted in a three dimensional model including enough information to reconstruct the atomar arrangement. Additionally such a model must describe nucleotides' interaction in more detail. Most of the interactions, described in section 2.3, should be feasible with such a model. Since minimalist lattice protein models equate with RNA secondary structure prediction, a new model should cover additional aspects of RNA conformation. Even the second type of lattice models, which bears in mind conformational aspects, is unusable for an accurate description of RNA conformation. Therefore three dimensional RNA structure must base on off-lattice models. One point representations such as done in YAMMP [128, 86] lack structural information for an adequate transformation of secondary structure into three dimensional conformation. Two, three and four pseudo atom models are also inadequate linkers between secondary and three dimensional RNA structure, because various basepair patterns explained in section 2.3.1 can not be described in an accurate way. These considerations lead to the invention of a coarse grained model. This new approach is presented in the following section.

5.2 Building blocks

Each nucleotide in `toyRNA` consists of seven to nine pseudo atoms, three for the backbone and four to six for the nitrogen base.

- **Backbone pseudo atoms**
  Three pseudo atoms represent the backbone of each nucleotide: one pseudo atom at the phosphor represents the phosphate group, two pseudo atoms, one at $C4'$ and one at $C1'$ position of the ribose, represent the sugar moiety. The hydroxy group at $C2'$, capable of hydrogen bonding, is included in the character of the $C1'$ pseudo atom. Connecting the P and $C4'$ pseudo atoms generates the basic backbone of

the `toyRNA`. To connect C1′ to the backbone, there is a further bond between the C4′ - C1′ pseudo atoms.

- **Nitrogen base pseudo atoms**
  The selection of the pseudo atoms for the purine or pyrimidine group depends on the type of nucleotide. Each atom able to act as hydrogen bond acceptor or donator is taken as a pseudo atom. Additionally, pseudo atoms at N1 in case of pyrimidine or at N9 in case of purine act as a linker between the nitrogen base and the backbone. The selected atoms of each nucleotide are summarized in Table 1. C1′ acts as a nitrogen base carrier, and is therefore connected to N9 in case of purine and N1 in case of pyrimidines (see figure 15). The pseudo atoms of each nitrogen base itself are connected via bonds generating cyclic polygons.

Adenosine	N9	N7	N6	N1	N3	
Guanosine	N9	N7	O6	N1	N2	N3
Uridine	N1	O4	N3	O2		
Cytidine	N1	N4	N3	O2		

Table 1: Atoms representing the nitrogen bases of the four nucleotides

5.3 Types of potential functions

After choosing the level of simplification of appropriate potentials have to be chosen. Beside the four classical terms a new term was designed to describe the effect of base stacking and hydrogen bond interaction. To check the introduced simplifications and to estimate the energy constants of the potentials, statistics of a representative set of three dimensional RNA structures became necessary. The used set of structures are taken from the Brookhaven Protein Data Bank [10] and a list of the file names are summarized in Appendix A. The statistical evaluation was done for all relevant parameters (bond lengths, angles, dihedrals). Also the dependence of these parameters from the sugar pucker effect was studied. The resulting plots for the most important parameters can be found in Appendix A.

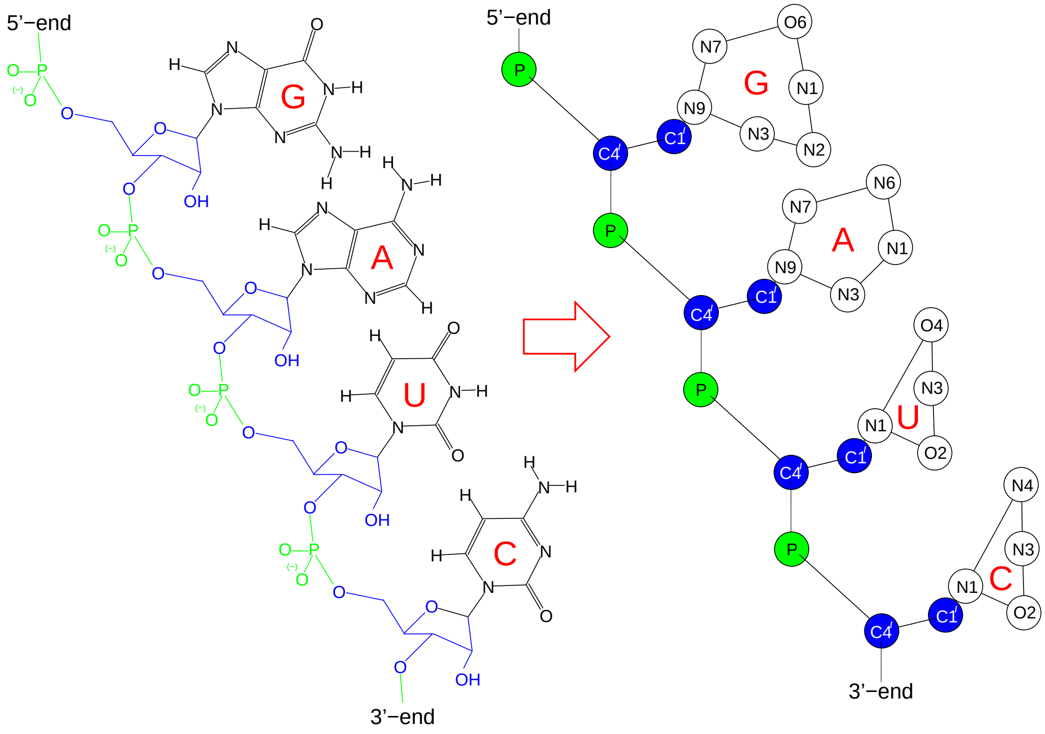


Figure 15: Conversion of a GAUC strand into the simplified representation of `toyRNA`

5.3.1 Bond, angles and torsion potentials

A. The bond and angle potential
Both hard degrees of freedom, the bonds and the angles are described via harmonic potentials:

$$E(r) = k_b(r - r_0)^2 \text{ and } E(\theta) = k_\theta(\theta - \theta_0)^2$$

k_b, k_θ are the spring constants. r_0, θ_0 define the distance and the angle at equilibrium state.

The **bond length** in the polygons describing the purine and pyrimidines heterocycles are naturally sharply defined due to the stiffness of the aromatic system. The first four blocks of Table 2 summarize the chosen bond lengths of the nitrogen bases. For all these pseudo bonds k_{bond} is set to 400.0 energy units.

The bond lengths of the backbone are equal for all four nucleotides. Including the whole backbone flexibility, they are not that stiff. The length distribution of the backbone bonds derived from the PDB test set are shown in Figure 32 (Appendix A). Each of the three left plots of Figure 32 displays a large peak at a distance, that assures optimal stacking with the neighbor nucleotides. Moreover a tailing to shorter distances is found. This phenomenon is caused by the simplification of the real atomic system. As pseudo bonds connect also atoms, that are normally separated by more than two bonds compression can easily take place by conformational changes. Nevertheless a harmonic potential was chosen to describe the bond between pseudo atoms.
Of particular interest is the bond between the C4′ - C1′. This bond represents the sugar moiety including the sugar pucker effect. The statistical results of the C4′ - C1′ bond lengths were evaluated in correlation to the sugar pucker effect.
As shown in the right plots of figure 32 (Appendix A) there are two culmination points, both representing the major sugar pucker modes, C2′-endo and C3′-endo, respectively. All three plots show a small correlation between the sugar pucker and the distances of the backbone, but again both spots show a tailing to shorter distances.

Equivalent to the bonds the **angles** of the simplified purine and pyrimidines heterocycles are sharply defined. The first four blocks of table 3 summarize the chosen angles of the nitrogen bases. k_θ for all these angles is set to 100.0 energy units, except for the angles between the nitrogen base and the sugar entity. In these cases k_θ is set to 90.0 energy units. Four of the angles have equilibrium angles θ_0 close to 180°. To prevent program instabilities these equilibrium angles θ_0 are set 180° and for the calculations another potential is used instead of the classical harmonic potential. The chosen potential for these angels is the well-behaved form used in DREIDING [87] and MMFF [44].

$$E(\theta) = k_\theta(1 + cos(\theta))$$

To assure an analogical potential behavior near the equilibrium state, the k_θ values for this potential are doubled , compared to the harmonic one.

Again, the angles of the backbone are of greater interest than the angles of the nitrogen bases. Figures 33 and 34 (Appendix A) show the distribution

	Adenine					
Bond	N9 - N7	N7 - N6	N6 - N1	N1 - N3	N3 - N9	
r_0	2.24	3.08	2.31	2.39	2.43	
k_b	400.0	400.0	400.0	400.0	400.0	

	Guanine					
Bond	N9 - N7	N7 - O6	O6 - N1	N1 - N2	N2 - N3	N3 - N9
r_0	2.25	3.09	2.27	2.29	2.31	2.42
k_b	400.0	400.0	400.0	400.0	400.0	400.0

	Uridine					
Bond	N1 - O4	O4 - N3	N3 - O2	O2 - N1		
r_0	4.04	2.27	2.28	2.29		
k_b	400.0	400.0	400.0	400.0		

	Cytidine					
Bond	N1 - N4	N4 - N3	N3 - O2	O2 - N1		
r_0	4.04	2.31	2.26	2.27		
k_b	400.0	400.0	400.0	400.0		

	Backbone (for all nucleotides)					
Bond	P-C4′	C4′-P	C4′ - C1′	C1′ - N(1,9)		
r_0	3.90	3.90	2.35	1.48		
k_b	250.0	250.0	300.0	350.0		

Table 2: The bond parameters for the nitrogen bases and the backbone, r_0 in Ångstroem, k_b values in arbitrary units

of the backbone bond angles of the PDB test set. Since the angles of the backbone show remarkable flexibility, all k_θ values are set lower to allow the system to adopt the different settings in the various base pairs.

B. The dihedral angle potential

The dihedral angle potential implemented in `toyRNA` is the standard term of the AMBER force field:

$$E(\omega) = \frac{k_{tor}}{2}(1 + cos(n\omega - \gamma))$$

k_{tor} controls the amplitude of this periodic function, n is the multiplicity,

	Adenine							
Angle	C1′-N9-N7	N9-N7-N6	N7-N6-N1	N6-N1-N3	N1-N3-N9	N3-N9-N7	C1′-N9-N3	
θ_0	160.9	118.9	85.4	123.9	112.1	99.7	99.1	
k_θ	90.0	100.0	100.0	100.0	100.0	100.0	90.0	

	Guanine							
Angle	C1′-N9-N7	N9-N7-O6	N7-O6-N1	O6-N1-N2	N1-N2-N3	N2-N3-N9	N3-N9-N7	C1′-N9-N3
θ_0	161.2	117.7	85.2	175.1	62.2	169.6	100.3	97.9
k_θ	90.0	100.0	100.0	100.0	100.0	100.0	100.0	90.0

	Uridine							
Angle	C1′-N1-O4	N1-O4-N3	O4-N3-O2	N3-O2-N1	O2-N1-O4	C1′-N1-O2		
θ_0	180.0	29.5	180.0	61.8	87.3	90.8		
k_θ	180.0	100.0	200.0	100.0	100.0	90.0		

	Cytidine							
Angle	C1′-N1-N4	N1-N4-N3	N4-N3-O2	N3-O2-N1	O2-N1-N4	C1′-N1-O2		
θ_0	180.0	30.5	180.0	62.8	88.1	89.2		
k_θ	180.0	100.0	200.0	100.0	100.0	90.0		

	Backbone (for all nucleotides)							
Angle	P-C4′-P	C4′-P-C4′	P-C4′-C1′	C1′-C4′-$(\mathrm{P})_{3'}$	C4′-C1′-N(1,9)			
θ_0	97.0	103.0	113.0	98.0	122.5			
k_θ	45.0	60.0	60.0	60.0	60.0			

Table 3: The angle parameters for the nitrogen bases and the backbone, θ_0 in degree, k_θ in arbitrary units, $(\mathrm{P})_{3'}$ is the next phosphor in $3'$-direction

and γ, the so-called phase factor, shifts the entire curve along the rotation angle axis ω. Figures 35 and 36 show the distribution of these torsion angles within the PDB test set. All distributions show (1) remarkable peaks at the torsion angles, which are adopted when they are in helical structures and (2) no noticeable dependence regarding the sugar pucker. Beside these peaks all torsion angles of the basic backbone (the C4′-P chain) are not restricted to special torsion angles. Therefore no torsion potentials for these chains are calculated. The torsion angles P - C4′ - C1′ - N(1,9) and $(P)_{3'}$ - C4′ - C1′ - N(1,9), which include the sugar moiety (Figure 35 third row), do not show such a behavior. Therefore two torsion potentials with multiplicity 1 are set to characterize the favored torsion angle (see table 4). The torsion angle regarding the C1′- N(1,9) bond is also not restricted. Consequently it is also not calculated by an torsion potential.

Torsion angle	k_{tor}	n	γ	remark
P - C4′ - C1′ - N(1,9)	7.5	1	-2.880	sugar-pucker
$(P)_{3'}$ - C4′ - C1′ - N(1,9)	7.5	1	+1.571	sugar-pucker

Table 4: The torsion parameters for `toyRNA`: k_{tor} is the force constant (given in energy units), n is the multiplicity (given in integers) and γ is the phase shift factors (given in degree)

5.3.2 The nonbonded interactions

While hard degrees of freedom can be handled in an equivalent way as force fields in atomic resolution, coarse graining the structure results in more complex non bonded interaction terms. The potential is often not longer radial symmetrical, but depends on angle dependent terms. In the case of RNA the atoms of the nitrogen bases show different behavior due to the relative position of the two pseudo atoms. They can stack upon each other or they can act as hydrogen bond donator or acceptor. To take these facts into account angle dependent terms must be introduced into the force field.

To handle this characteristic feature of pseudo atoms, non bonded potentials of `toyRNA` are subdivided into two groups: The “classical” radial symmetrical non bonded potentials (Coulomb and Lenard-Jones) and angle dependent potentials describing the stacking and the hydrogen bonding behavior. All terms are now introduced in a more detailed way.

Classical nonbonded terms:

A. The Coulomb potential

Since RNAs are polyionic molecules, the Coulomb interaction plays an important part in the conformational behavior of these molecules. The determination of the partial charges in atomistic resolution is not a trivial task. For pseudo atoms this determination is even more complicated, because they show multipole behavior. For simplification only Coulomb interactions of the phosphate pseudo atoms are taken into account. No extended dielectric constants types are used ($\varepsilon = 78$). q_1 and q_2 are set to the simple elementary charge value.

$$E(r) = \frac{1}{4\pi\varepsilon_0\varepsilon} \cdot \frac{q_1 q_2}{r}$$

B. The Lenard-Jones potential

The classical Lenard-Jones 6-12 potential is used to describe the dispersive forces between two atoms. The potential used in toyRNA is expressed in the following way:

$$E(r) = E_{min}\left(\left(\frac{\sigma}{r}\right)^{12} - 2\left(\frac{\sigma}{r}\right)^{6}\right)$$

E_{min} is the energy minimum of the potential and σ the distance between the pseudo atoms at the energy minimum. Since pseudo atoms equal atom clusters with no spherical behavior, the values of σ are estimated by van der Waals radii taken from Bondi [12].

pseudo atom type	estimated $\sigma_{0.5}$ [Å]
P	3.0
$C4', C1'$	2.0
nitrogen base	1.9

σ is simply the sum of the two $\sigma_{0.5}$, whereas E_{min} equals 0.2 energy units for all pseudo atom pairs, except the pairs between the $N1_{Pyr}$, $N9_{Pur}$ and the nitrogenbase atoms. For these pairs E_{min} is set to 0.3 energy units. $N1_{Pyr}$, $N9_{Pur}$ can show stacking abilities but no hydrogen bond sites. Therefore no angle dependent potential is calculated for them.

Angle dependent potentials:

To describe different interactions between two pseudo atoms of two different nitrogen bases (hydrogen-bond or stacking interaction) an angle dependent potential is introduced. These pseudo atoms have no longer radial symmetry, they have a preferred interaction axis. In the case of the nitrogen base atoms this axis is the perpendicular of the plane spanned by the base atoms. The connecting vector $\vec{r}$ and the perpendicular vectors, given by the cross product from the $\vec{a} \times \vec{b}$ and $\vec{c} \times \vec{d}$, respectively, define two angles α and β. These angles are used to define the angle dependent potential (see figure 16).

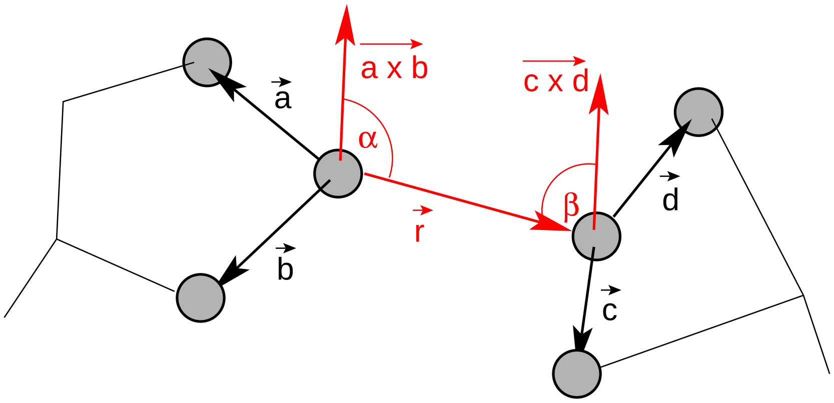


Figure 16: Schematic description of the angle dependent potential between two nitrogen bases' pseudo atoms.

If $\vec{a} \times \vec{b}$ and $\vec{c} \times \vec{d}$ are chosen in an appropriate way (if necessary, the inverse of them must be chosen), α and β lie in between the interval 0 and $\frac{\pi}{2}$. Therefore the sum of α and β lies between 0 and π. Considering the sum of α and β in a molecular context three special cases can be distinguished:

(1) $\alpha + \beta = \pi$ that means $\angle(\vec{r}, \vec{a}) = \angle(\vec{r}, \vec{b}) = \frac{\pi}{2}$. This is the typical situation of hydrogen bond interaction. It is obvious, that stacking do

not contribute to the energy between the pseudo atoms in this case.

(2) $\alpha+\beta = \frac{\pi}{2}$. In this case neither hydrogen bond nor stacking interactions between the two pseudo atoms can be expected.

(3) $\alpha+\beta = 0$ that means $\vec{a}$ and $\vec{b}$ are parallel and $\angle(\vec{r},\vec{a}) = \angle(\vec{r},\vec{b}) = 0$. This is the typical case of large stacking interactions. In this situation hydrogen bonds do not contribute to the energy between the pseudo atoms.

Figure 17 shows schematically the three border cases of this potential (it is important to mention once again, that the perpendicular vectors must be chosen in an appropriate way!). Since the sum of $\alpha+\beta$ changes the behavior of the potential, a new variable $\kappa := 2\alpha+2\beta-\pi$ is introduced. κ lies in between the interval $[-\pi,\pi]$ and has the advantage, that κ equals zero in border case (2), where a change of interaction type takes place (either stacking or hydrogen bonding).

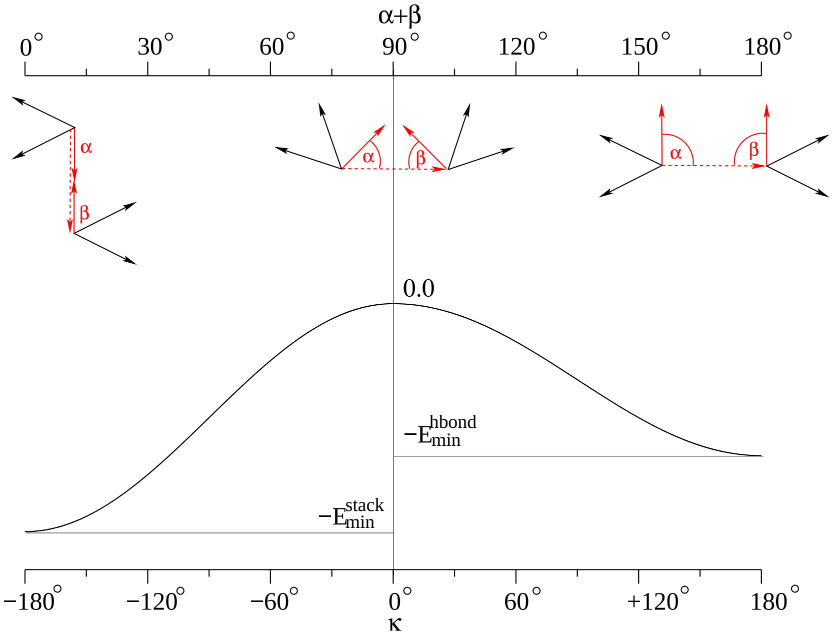


Figure 17:

With this new variable κ, two new potentials for the stacking and the hydrogen bond interaction can be expressed. Both potentials consist of two terms, an angle dependent term included ε_{HB} and ε_{ST}, and an radial term $(\frac{\sigma}{r})^m$. m equals 3 for the hydrogen bond potential, characterizing the dipole-dipole interaction, whereas m equals 6 in case of stacking (describing an van der Waals like interaction). Combining these terms result in the following potentials:

- *Hydrogen bond term:* $E = \varepsilon_{HB}(\kappa)(\frac{\sigma}{r})^3$
- *Stacking term:* $E = \varepsilon_{ST}(\kappa)(\frac{\sigma}{r})^6$

As a next step functions for ε_{HB} and ε_{ST} must be defined . These functions have to fulfill the following restriction within the intervals of κ:

	kappa				
	$-180°$	$[-180°, 0°]$	$0°$	$[0°, +180°]$	$+180°$
ε_{HB}	0	$0 \rightarrow 0$	0	$0 \searrow E_{min}^{hbond}$	E_{min}^{hbond}
ε_{ST}	E_{min}^{stack}	$E_{min}^{stack} \nearrow 0$	0	$0 \rightarrow 0$	0

For $\kappa = -180°$ to $\kappa = 0°$ ε_{HB} is zero. For $\kappa = 0°$ to $\kappa = +180°$ the function monotony decreases till E_{min}^{hbond} is reached at $+180°$. ε_{ST} adopts the value E_{min}^{hbond} at $\kappa = -180°$ and increases till κ is zero. For $\kappa = 0°$ to $\kappa = +180°$ ε_{ST} remains zero. The further selection of the function is somehow arbitrary, but two restrictions reduce the number of possible function forms: First the chosen function must be monotony and second, this function should be simple differentiable to allow the calculation of the potential's gradient. A simple linear function has the problem of differentiability at $\kappa = 0°$. Considering harmonic functions showed a third restriction: The slope of the function must be zero at the border minimums. Otherwise a non zero gradients exist at these minimums. A function, which fulfills all restrictions, is the cosine of kappa. The resulting potentials are as follows:

- *Hydrogen bond potential:*

$$E = \begin{cases} 0 & for \quad -180° < \kappa \leq 0 \\ -\frac{E_{min}^{hbond}}{2}(1 - \cos\kappa)(\frac{\sigma}{r})^3 & for \quad 0 < \kappa < +180° \end{cases}$$

- *Stacking potential:*

$$E = \begin{cases} -\frac{E_{min}^{stack}}{2}(1-\cos\kappa)(\frac{\sigma}{r})^6 & for \quad -180^\circ < \kappa \leq 0 \\ 0 & for \quad 0 < \kappa < +180^\circ \end{cases}$$

Figure 17 displays the resulting functions for ε_{HB} and ε_{ST} in the intervals, where they are not zero

Combining all non bonded potentials leads to an energy surface, which depends on the distance r between the two pseudo atoms and the angle κ of the angle dependent potentials. Figure 18 shows such a surface for arbitrary values. There is a saddle point at $kappa = 0^\circ$ where the system can fall into the hydrogen bond or the stacking interaction minimum.
All pseudo atom combinations are described by classical non bonded terms. The angle dependent potential is only executed, when both pseudo atoms are of nitrogen base type. Exceptions are the N1 in purines and N9 in pyrimidines due to their lacking hydrogen bonding capacity. The pseudo atoms, which are required for calculating the normal vectors, are in most cases the neighboring pseudo atoms of the polygon. To prevent problems in calculating the vector product pseudo atoms, that are placed in the center of linear angles are omitted and the next but one pseudo atom is taken instead.

To complete this potential the force constants E_{min}^{hbond} and E_{min}^{stack} must be set. Since experimental data for pseudo atoms are not available, this can be done only more or less arbitrary. E_{min}^{stack} has the same value for all pseudo atoms in the purine heterocycle and in the pyrimidine heterocycle. It is set to 0.5 energy units. (A short estimation of the constants' magnitude was done, by making molecular dynamic simulation of a `GCGCGC` duplex (Settings: $T = 300^\circ K$, 1 time step one femtosecond and a friction constant $\gamma = 0.0001$). Both values were continuously enlarged until the the duplex was stable more than 2 nanoseconds). These pseudo atoms (except $N1_{Pyr}$, $N9_{Pur}$) must be subdivided into two groups: Acceptors and donators with regards to hydrogen bonding.

nucleotide	donators	acceptors
Adenosine	N6	N7,N1,N3
Guanosine	N1,N2	O6,N7,N3
Uridine	N3	O4,O2
Cytidine	N4	O2,N3

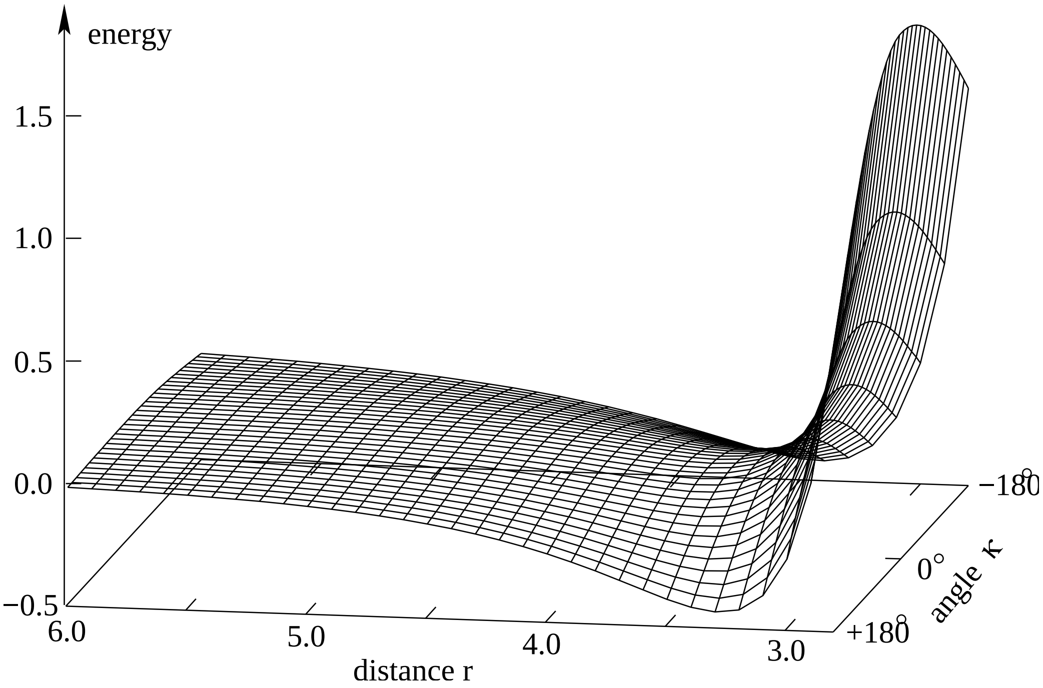


Figure 18: Energy surface of the non bonded potentials (Classical and angle dependent potential). The new variable κ and the distance r between the two pseudo atoms are plotted against the energy. The values for the minima and σ are arbitrarily chosen. $\sigma = 3.2, E^{lj}_{min} = 0.6, E^{hbond}_{min} = 0.7, E^{stack}_{min} = 0.9$ (Coulomb interaction are not considered)

E^{hbond}_{min} is set zero for pairs of two donators or acceptors. For mixed pairs, where hydrogen bonds are possible, $E^{hbond}_{min} = 0.75$ energy units.

6 Calculations

6.1 Helix

6.1.1 Energy minimization

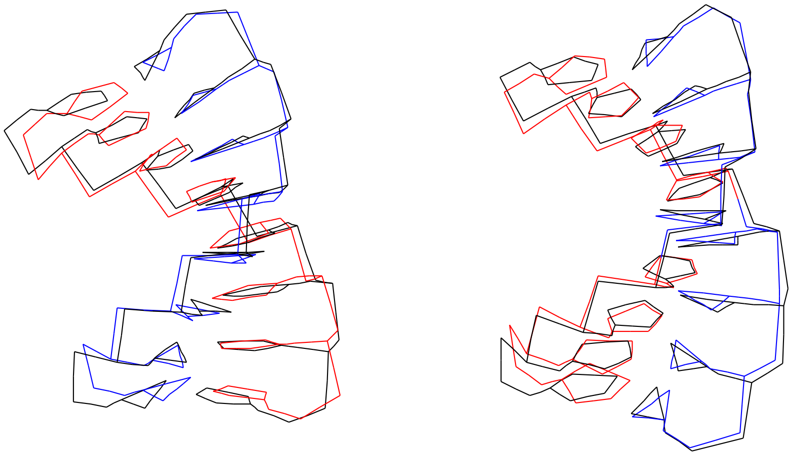
The simplest and most important structure motif of RNA molecules is the helix. Therefore the first calculation with `toyRNA` must show, that the designed forcefield reproduces the helix in an adequate way. Ten helices with increasing complexity were constructed and optimized with conjugate gradient optimation. The results were compared with the original structures via root mean squared distance (rmsd) on the simplified level and relative energies. The original helices were created with the very useful program package NAB (Version 4.4) [83] and minimized with the conjugate gradient routine in this package. The used potential is AMBER with the AMBER99 parameters and generalized Born electrostatics with Debeye Hückel screening (10mmol salt concentration) as suggested by Tsui and Case [131]. The results of the calculations are shown in Table 5. Figures 19, 20 and 21 present the superposed structures for each original and optimized helix. The starting structure optimized on classical atomic level is colored black. This structure is the initial point for the optimation on the simplified level.
Analyzing the data of Table 5, two facts are noticeable: First, the good agreement of the minima for helices containing only adenosine and uridine (now shortly termed AU-helices) and the bad matching of the minimized structures for the helices containing only guanosines and cytidines (now shortly termed GC-helices). The root mean square gradients for AU-helices are all smaller than 1.0, whereas the GC-helices have values greater than 1.0 . Both energies of the helices, the classical energy derived by the atomic NAB force field and the one according to `toyRNA`, show similar patterns. For both kind of calculations, the energy values for GC-helices are higher than for AU-helices. The last two rows of Table 5 show the values for mixed helices containing all possible nucleotides. Remarkable is the high root mean square gradient of the `AAGGCCUU` duplex. The derived value of 1.505 is much higher than expected for such a system. The value of the other mixed helix is as low as the of the tested AU-helices. The reason for this effect lies in the structure of the `AAGGCCUU` system. Neighboring GG pairs change the structure of the system, whereas systems with no neighboring GG pairs show better rmsd values (e.g. `GCGCGCGC` duplex). Generally `toyRNA` favors structures with al-

Sequence	$Energy_{NAB}$	$Energy_{toyRNA}$	$rmsd$
`AAAAAAAA` `UUUUUUUU`	-2803.0	-128.9	0.968
`AAAAUUUU` `UUUUAAAA`	-2803.2	-128.7	0.699
`AAUUAAUU` `UUAAUUAA`	-2803.3	-130.9	0.718
`AUAUAUAU` `UAUAUAUA`	-2805.5	-133.4	0.621
`GGGGGGGG` `CCCCCCCC`	-4037.8	-191.0	1.877
`GGGGCCCC` `CCCCGGGG`	-4038.0	-190.8	1.725
`GGCCGGCC` `CCGGCCGG`	-4035.1	-191.1	1.623
`GCGCGCGC` `CGCGCGCG`	-4030.7	-192.5	1.018
`AAGGCCUU` `UUCCGGAA`	-3423.0	-162.6	1.505
`AUGCAUGC` `UACGUACG`	-3417.2	-163.0	0.698

Table 5: Energies and root mean square distances of ten RNA helices. $Energy_{NAB}$ in $kcal/mol$, $Energy_{toyRNA}$ in arbitrary energy units

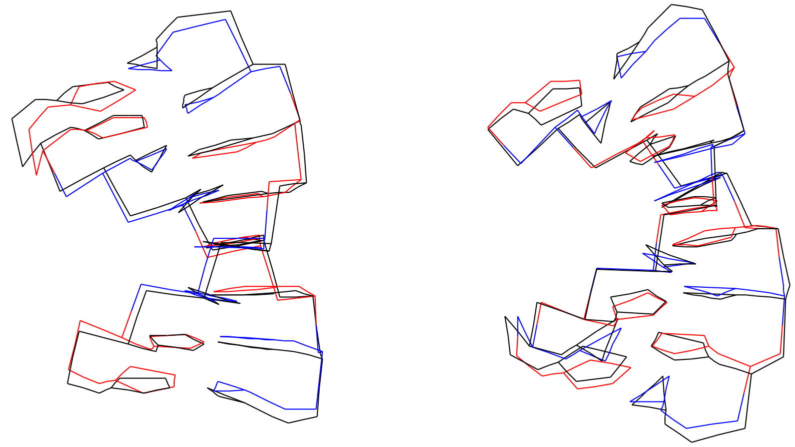
ternated purine and pyrimidine patterns.

Figures 19, 20 and 21 emphasize the results of Table 5. Figure 19 shows the superimposed structure for the AU-helices. A good agreement between the NAB minimized structure and the minimized structure of `toyRNA` can be seen. The difference in the backbone conformation is noticeable, but not very large. The distances between the neighboring nitrogen bases are nearly identical.
Figure 20 visualizes clearly, why the root mean square distances of the GC-helices are that high. In contrast to the AU-helices, the backbone conformation is contorted. The difference between NAB and the `toyRNA` structure



AAAAAAAA
UUUUUUUU

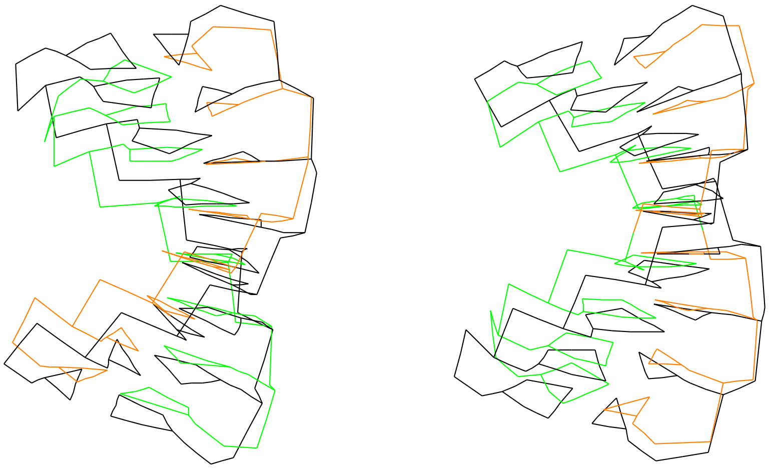
AAAAUUUU
UUUUAAAA



AAUUAAUU
UUAAUUAA

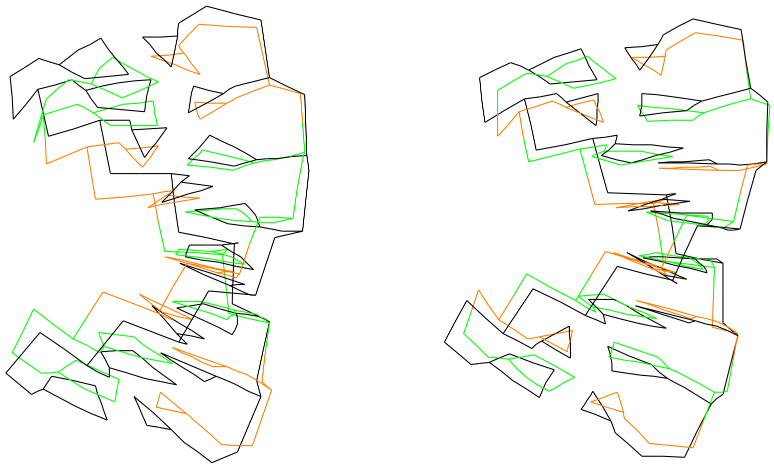
AUAUAUAU
UAUAUAUA

Figure 19: Superposed structures of the original and optimized helices containing only adenosines and uridines. (color code: black-starting structure, red-adenosine, blue-uridine)



`GGGGGGGG`
`CCCCCCCC`

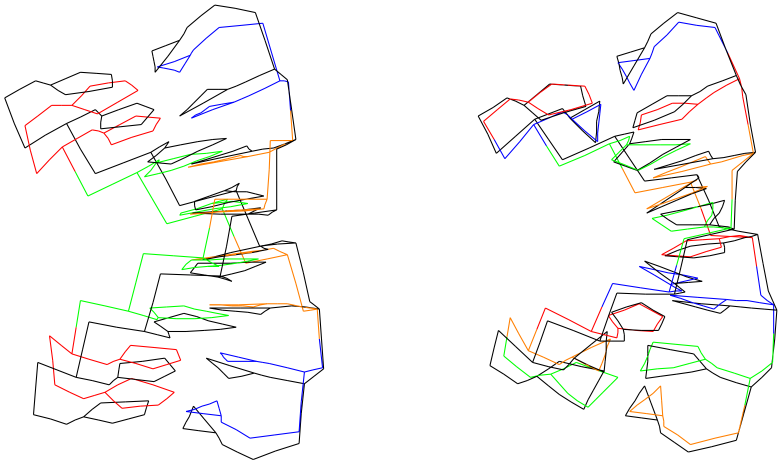
`GGGGCCCC`
`CCCCGGGG`



`GGCCGGCC`
`CCGGCCGG`

`GCGCGCGC`
`CGCGCGCG`

Figure 20: Superposed structures of the original and optimized helices containing only Guanosine and Cytidine. (color code: black-starting structure, green-guanosine, orange-cytidine)



`AAGGCCUU`
`UUCCGGAA`

`AUGCAUGC`
`UACGUACG`

Figure 21: Superposed structures of the original and optimized helices containing all types of nucleotides.(color code: black-starting structure, red-adenosine, green-guanosine, blue-uridine, orange-cytidine)

increases in case of two neighboring guanosines in one strand. Even the distances between the neighboring nitrogen bases are not identical. This leads to a compression of the helices. Once again the neighboring guanosines are the reason for this effect. They reduce the distance between the pseudo atoms of the nitrogen base (eg. `GGGGCCCC` in contrast to `GCGCGCGC` duplex).
This effect can also be observed in the mixed helices of Figure 21. The `GGCC` system in the middle of the `AAGGCCUU`-helix compresses the whole system, resulting in a shift of the border AU base pairs to the middle of the helix. In contrary, the NAB and the `toyRNA` optimized structures of the other mixed helix are in good accordance.

6.1.2 Molecular dynamic simulations

Due to the local minimum problem any designed force field cannot be qualitatively proofed by minimizing a given structure. All algorithms are trapped in the next local minimum and this local minimum is mostly not the global minimum of the structure. To overcome this problem, molecular dynamic simulations have to be done. The conformational space around a minimum structure was analyzed on four helices. For these simulations Langevin dynamics (see section 4.1.3) with an velocity verlet algorithm have been used in the following setting: $T = 300°K$, time steps lasting one femtosecond and a friction constant $\gamma = 0.0001$. The mol masses of all pseudo atoms are set to $40g$. All molecular dynamic simulations are done over a time period of 4 nanoseconds[1] Table 6 summarizes the different root means square distances of all snap shots (written every 5 picoseconds) for these five simulations:

Sequence	$rmsd_{pair}$	$rmsd_{mean}$	$rmsd_{min}$
`GGGGGGGG` `CCCCCCCC`	1.255	0.877	0.536
`GGGGCCCC` `CCCCGGGG`	1.237	0.875	0.442
`GCGCGCGC` `CGCGCGCG`	1.225	0.866	0.379
`AAGGCCUU` `UUCCGGAA`	4.272	3.059	4.302
`AUGCAUGC` `UACGUACG`	1.475_1 2.274_2	1.041_1 1.612_2	0.655_1 2.722_2

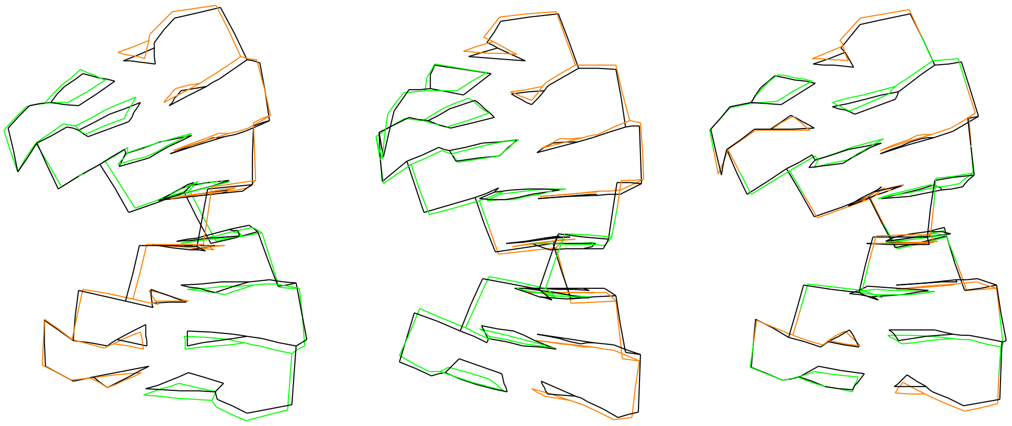
Table 6: Root means square distances (rmsd) of the molecular dynamics' structure snap shots (written every 5 picoseconds) for the five helices, that are used for the molecular dynamics simulations. $rmsd_{pair}$ is the rmsd between the structures, $rmsd_{mean}$ the rmsd difference from the mean structure, $rmsd_{min}$ is the rmsd between the mean structure and the minimum structure of the helix calculated in subsection 6.1.1.$_1$ calculated for the first nanosecond, $_2$ calculated for the second nanosecond

The first three GC-helices are stable over the whole simulation period. The

---

[1] Due to the simplification, time, temperature and friction constant are not of realistic character

root mean square distances are quite low and indicate that the conformation does not change during the four nanoseconds considered. The difference between the mean structure and the minimum structure is small. Figure 22 compares the mean structure of all molecular dynamics' structure snap shots and minimized structures of these three GC-helices:



`GGGGGGGG`
`CCCCCCCC`

`GGGGCCCC`
`CCCCGGGG`

`GCGCGCGC`
`CGCGCGCG`

Figure 22: Comparison of the mean structure of all molecular dynamics' structure snap shots and minimized structure for all three GC-helices (color code: black-minimum structure, red-mean structure)

The situation for the two mixed helices is quite different. The `AAGGCCUU` duplex does not decompose in the four nanoseconds simulation, but the two AU base pairs on both ends of the helix open, and do not close till the end of the considered time period. The four GC base pairs stabilize the system and prevent the system from decomposition. Figure 23 shows two typical snap shots of the molecular dynamics trajectory. The left figure is a snap shot made shortly after the start of the simulation. The helical structure is still recognizable, but the AU base pairs are destroyed. The right figure shows a typical situation: As there are no torsion potentials in the C4′ - P4 - C4′ backbone, this system is very flexible and allows the stack between three adenosines (red polygons) beside the GC basic stack. This unsatisfying result indicates, that the backbone system is maybe to unrestricted.

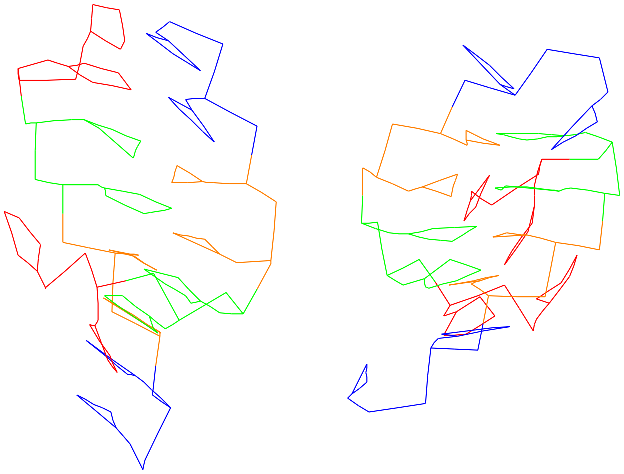


Figure 23: Two typical snap shots of the molecular dynamic trajectory of the `AAGGCCU` duplex (color code: red-adenosine, green-guanosine, blue-uridine, orange-cytidine)

The other helix $\binom{\mathsf{AUGCAUGC}}{\mathsf{UACGUACG}}$ including all four possible nucleotide types shows a different behavior. In the first nanosecond the helix is stable showing only temporary base pair openings of the terminal AU base pairs. In the second nanosecond of the simulation partial decomposition of the helical structure takes place. After two nanoseconds the system decomposes completely and the two strands diffuse away. The two values in Table 6 show root mean square distances for these two stages.

6.2 GNRA tetra loop

6.2.1 Energy minimization

The second type of structure motif, which is investigated in more detail, is the loop, especially the tetra loop of the GNRA type. The procedure of evaluation is quite the same as for helices, with the only difference that in this case, the reference structures are taken from experimental structures. Target object in this subsection is a GCAA tetra loop investigated by Pardi et al. [61]. The experimental data file (**1zih.pdb**) is taken from the Brookhaven Protein Data Bank [10]. Since these data are derived from nuclear resonance experiments, more than one structure is suggested in this file (In this special case ten structures fulfill the restriction of the measured NMR data). Therefore the given file is splitted into ten files, each containing one of the conformers. Consequently each of these files is translated in the simplified model and optimized with conjugate gradient method. Again, the root mean square distances (rmsd) on the simplified level are used as an indicator for similarity. Additionally to the similarity between the starting structure and the optimized structure and the similarities between the conformers among each other are investigated. This information is required to estimate the simplification of the conformational landscape done by the `toyRNA` model.

Conformere	$Energy_{toyRNA}$	$rmsd$
zih1	-113.7	0.989
zih2	-111.6	0.993
zih3	-113.7	0.879
zih4	-112.1	0.722
zih5	-112.4	0.685
zih6	-111.8	0.995
zih7	-113.7	0.928
zih8	-112.9	0.810
zih9	-113.7	0.955
zih10	-112.4	0.758

Table 7: Energies and rmsd of the experimental and optimized structures of the ten GCAA tetra loop conformers extracted from the Brookhaven Protein Data Bank file 1zih.pdb. $Energy_{toyRNA}$ values in abitrary energy units

The energy values of the optimized conformers and the rmsd of these structures in relation to their starting structure are summarized in Table 7. The energies differ only from -113.7 and -111.6. The rmsd between the starting structure and the minimized structure is quite large, showing a bad accordance of these structures. The next question is, in what direction the potential of `toyRNA` guides the conformers. A look at table 8 illuminates the rather large conformational change. Nearly all rmsd values between the conformers on the simplified level are smaller than the equivalent values on the experimental structures' level. These results indicate that `toyRNA` effect an extreme simplification of the energy landscape. Consequently there are conformers, which have an identical conformation after minimizing them with toyRNA (e.g. zih1 and zih3). This effect can also be found, comparing the pairwise rmsd between all structures ($rmsd_{pair}$) within one set. $rmsd_{pair}$ is 0.881 for the experimental set and 0.526 in case of the `toyRNA` optimized set.

	zih1	zih2	zih3	zih4	zih5	zih6	zih7	zih8	zih9	zih10
zih1	0.000	1.521	0.442	0.800	0.692	1.250	0.571	0.789	0.661	0.831
zih2	0.692	0.000	1.285	1.358	1.285	0.866	1.481	1.347	1.385	1.134
zih3	0.013	0.687	0.000	0.621	0.446	1.140	0.639	0.693	0.567	0.476
zih4	0.550	0.965	0.543	0.000	0.461	0.984	0.572	0.575	0.327	0.557
zih5	0.405	0.900	0.401	0.220	0.000	1.092	0.664	0.558	0.467	0.449
zih6	0.557	0.334	0.573	0.846	0.787	0.000	1.103	0.985	1.042	1.119
zih7	0.068	0.667	0.055	0.509	0.383	0.559	0.000	0.713	0.475	0.815
zih8	0.452	0.837	0.452	0.695	0.593	0.668	0.457	0.000	0.583	0.752
zih9	0.048	0.681	0.036	0.519	0.388	0.564	0.019	0.454	0.000	0.547
zih10	0.485	0.702	0.477	0.370	0.373	0.672	0.440	0.673	0.452	0.000

Table 8: Root mean square distances between the ten GCAA tetra loops conformers extracted from the Brookhaven Protein Data Bank file. The values above the diagonal are the rmsd for the experimental data. The values below are the rmsd of the simplified structures, which are formerly optimized with the conjugate gradient method of `toyRNA` zih.pdb. $Energy_{toyRNA}$ values in abitrary energy units

Figure 24 illustrates this effect in an impressive way. The bunch of the experimental conformers' structures is reduced to a nearly sharp structure by the `toyRNA` model. Only the cytidine in the loop shows is not reduced in its variability.

6.2.2 Molecular dynamic simulations

Next, a molecular dynamic simulation of the tetra loop was done under the same conditions as described for the helices. This simulation indicates clearly possible benefits but also existing problems of this new designed model. All typical properties of the tetra loop system are found in the model's simulation. For example the terminal GU base pair opens and closes during

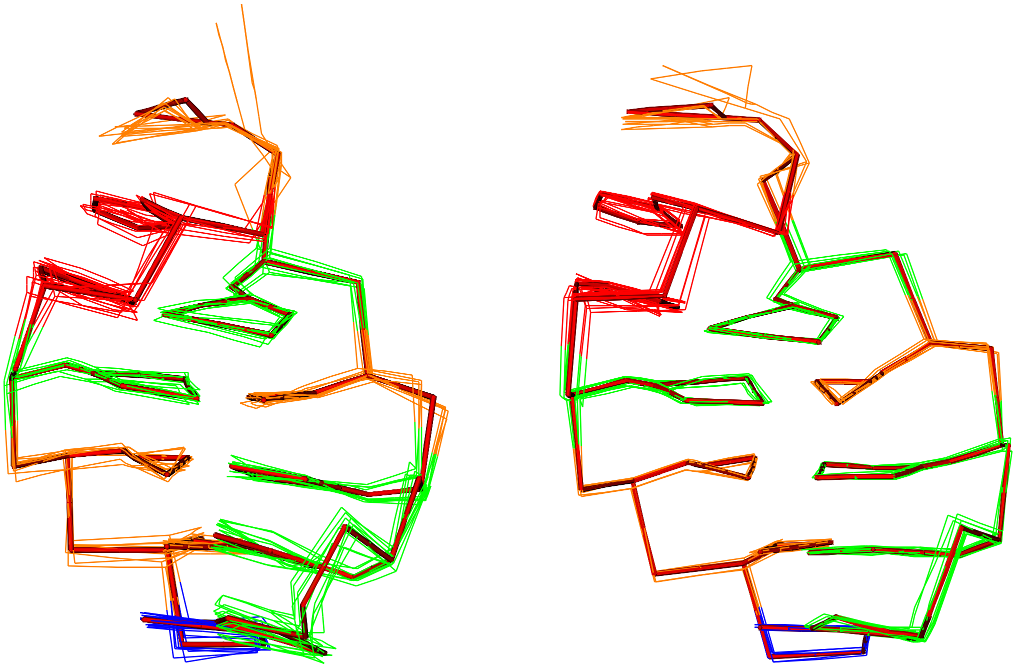


Figure 24: Left: All conformers of the 1zih.pdb file and the mean structure of them. Right: All conformers of the 1zih.pdb file, optimized on the simplified level with `toyRNA` (the mean structure is highlighted in red color sticks).(color code: red-adenosine, green-guanosine, blue-uridine, orange-cytidine)

the ongoing simulation. On the other hand the model is not able to represent the loop region in a realistic way. Although the G-A base pair is the typical feature of GNRA tetra loops, it cannot be found in the snap shot structures of the simulation until it is once open. Probably the chosen force constants (specially the angle between C1′-C4′-$(P)_{3'}$) are not correctly tuned and therefore an unrealistic conformational variety occurs (see Figure 25).

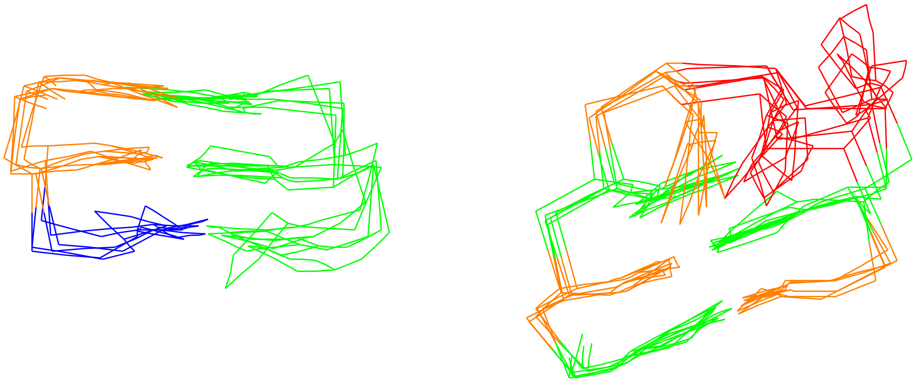


Figure 25: Left: Open-close sequence of the terminal GU base pair. Right: problematic flexibility inside the loop region (color code: red-adenosine, green-guanosine, blue-uridine, orange-cytidine)

6.3 Hammerhead ribozyme

The last target for testing the simplified `toyRNA` model is the hammerhead ribozyme. The hammerhead ribozyme is one of the few catalytic RNA motifs that has been identified and characterized [104, 117]. It is a complex RNA structure containing helical regions, a tetra loop and a multi loop. Figure 26 shows the secondary and the three dimensional structure of the two strands forming the hammerhead ribozyme. Beside the three stems (I, II, III) with canonical base pairs (green lines), various non-canonical base pairs govern the structure of the multi loop (blue arcs). This multi loop, a highly conserved region, contains two special RNA structural motifs, the so called "U-turn" (a turn formed by a specific hydrogen bond pattern of the nucleotides C6, U7, G8, A9) and a duplex containing tandem GA mismatched base pairs (G11-A22, A12-G21). The catalytic reaction of the hammerhead ribozyme, the cleavage of the $3',5'$-phoshodiester bond between nucleotides C37 and A38, is accomplished through this special multi loop pattern. Aim of this calculation is to check, if `toyRNA` tend to destroy the system, especially the specific hydrogen bond patterns of the catalytic structure motifs within the multi loop. Starting point of the calculation is the experimental data file (**299d.pdb**) taken from the Brookhaven Protein Data Bank [10]. The data are based upon the investigations of Scott et al. [117].

6.3.1 Energy minimization

Since the experimental data are derived from an X-ray diffraction experiment, only one structure is suggested in this file. This file is translated in the simplified model structure and optimized with conjugate gradient method. The root mean square distance (rmsd) between the starting structure and the optimized structure is 1.160. Figure 27 shows the superimposed experimental and `toyRNA`-optimized structures. Again, the typical effects, that have been found earlier, can be investigated; all stems remain stable during the optimation and stem II with the GGCC pattern shows the compression effect observed in GC helices (see section 6.1.1). The multi loop region shows no remarkable changes. All non-canonical base pairs inside the multi loop of the hammerhead ribozyme still exist. Also the U-turn can be clearly identified after optimation. This is of special interest, because the `toyRNA` cannot describe the hydrogen bond pattern in the U-turn (hydrogen bonds including OH2′ cannot be calculated).

6.3.2 Molecular dynamic simulation

The molecular dynamic simulation of the hammerhead ribozyme provides a lot of information about this new force field `toyRNA`. Again, the 2 nanosecond simulation was done under the same conditions as described for the helices. The first impressions of the results are a little bit confusing. The changes in the molecules structure seem tremendous, and it is difficult to summarize all observable effects. As changes often happen instantaneously, an adequate description is rather complicated.

The first effect, which can be observed easily, is the *compression* of the whole molecule structure (see Figure 28). Stem II and I get in close contact to each other. Thereby the hammer like shape is transformed to a drop like one. Reasons for this global change of the molecular shape are the multi loop region (see later) and the compression effect in the GC containing stem II.

The *stability of the three stems* differs depending on the number of GC base pairs inside. Stem II remains stable during the whole simulation time. In stem I the G1-C42 and 2U-41A base pairs opens after one 1 nanosecond. While the nucleotides 41A and 42C mainly remain stacking on the G3-C39 base pair, the nucleotides 1G and 2U move around and interact with the

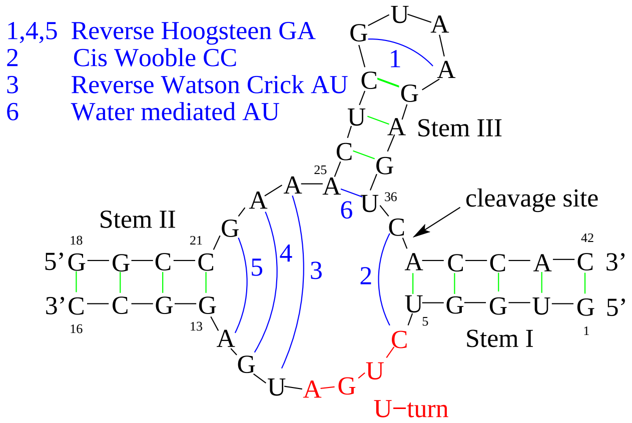


Figure (1)

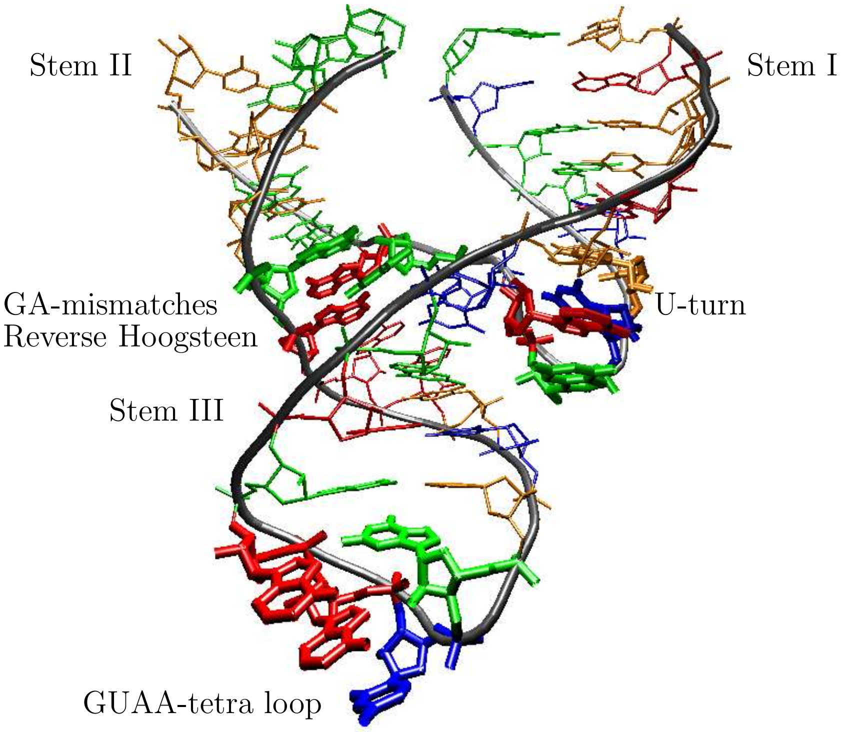


Figure (2)

Figure 26: Secondary and tertiary structure of the hammerhead ribozyme. (1) Green lines represent classical Watson Crick base pairs. Blue arcs indicate possible non-canonical base pairs inside the tetra and the multi loop. The red colored region shows the nuclotides involved in the U-turn (2) Location of the structural motifs in the three dimensional structure (color code: red-adenosine, green-guanosine, blue-uridine, orange-cytidine).

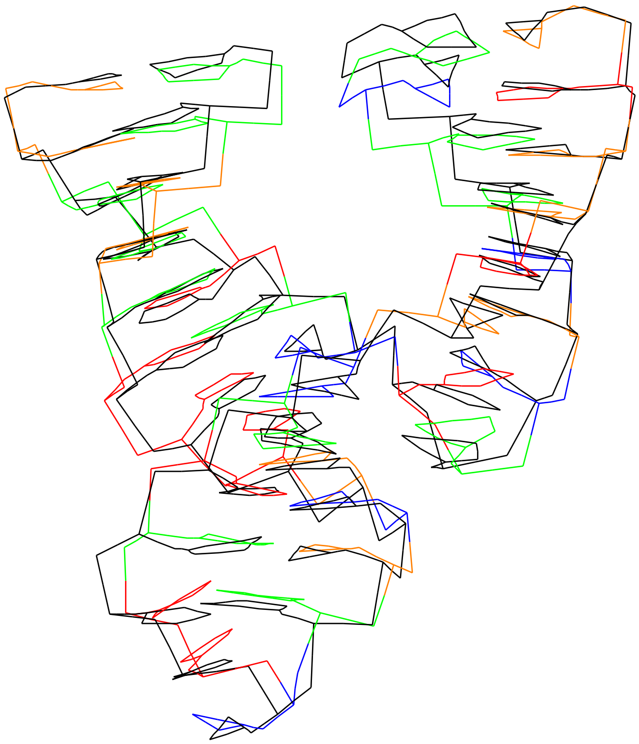


Figure 27: Superimposed experimental and `toyRNA`-optimized structures of the hammerhead ribozyme (color code: black-experimental structure, red-adenosine, blue-uridine, green-guanosine, orange-cytidine)

nucleotides of stem II (see Figure 29 (1)). These interactions are favored by the compression of the molecule and the G21-A12 mismatch, that possess a free Watson-Crick edge for hydrogen bond formation (see Figure 29 (2)). The results of this simulation and the former helix simulations indicate, that the helix stability in this simple model is mainly caused by GC base pairs. Stem III with the GUAA tetra loop is stable during simulation time except the water mediated A24-U35 base pairs, which cannot be described in an adequate way by this model.

The *GUAA tetra loop* shows the same unrealistic conformational variety as

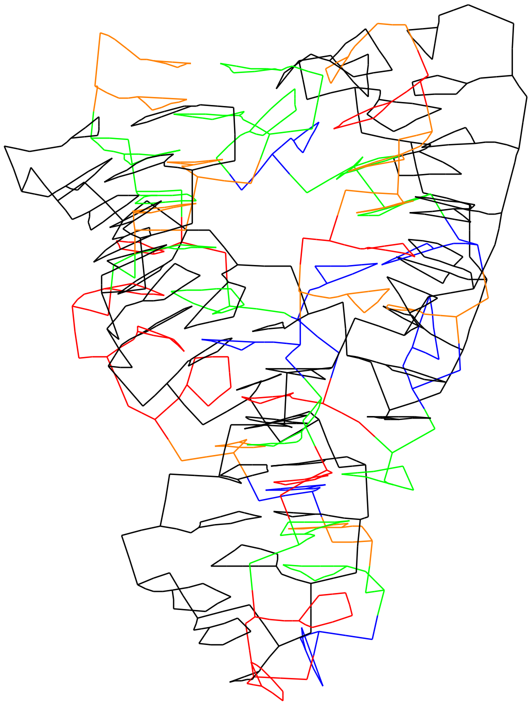


Figure 28: Superimposition of the experimental hammerhead ribozyme structure and a typical snap shot of the molecular dynamic trajectory shows the remarkable compression of the system (color code: black-experimental structure, red-adenosine, blue-uridine, green-guanosine, orange-cytidine)

seen in the simulation of the GCAA tetra loop in section 6.2.2. Again the behavior indicates a strain in the ring system that causes the opening of the reversed Hoogsteen base pair G29-A32.

The most complex system to be analyzed was the *multi loop* and its non-canonical base pairs. The two neighboring reverse Hoogsteen base pairs (G11-A23, A12-G22) remain stable during simulation, with short temporary opening sequences. This shows clearly, that `toyRNA` is also able to describe non-canonical base pairs as good as canonical ones. The Adenosines A23 and A24 beside the GA mismatched pairs show high flexibility. First the

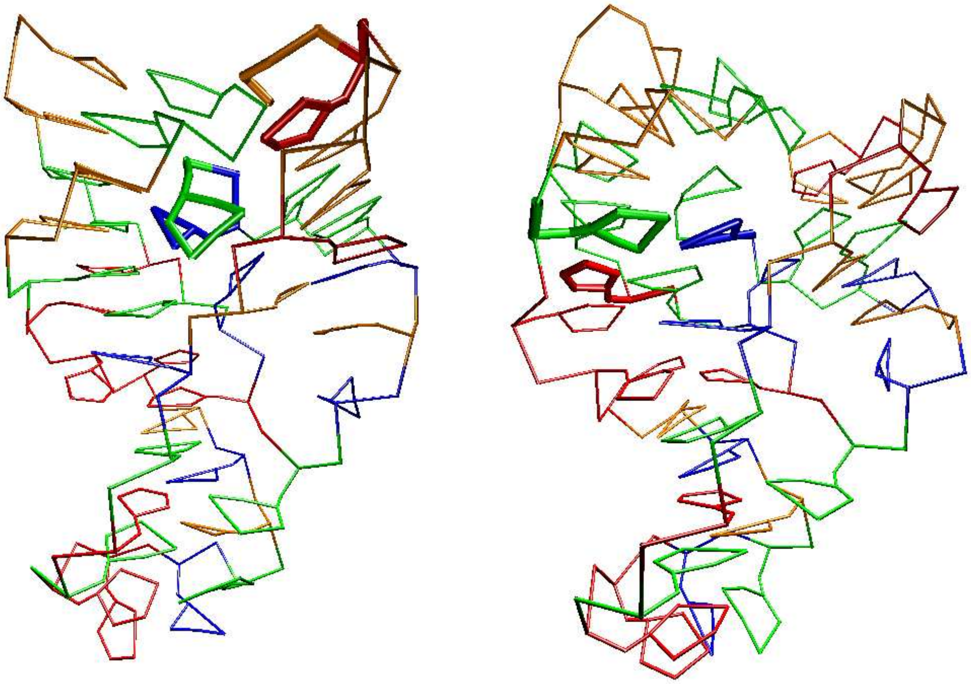


Figure (1) 1.50ns Figure (2) 0.35ns

Figure 29: Snap shots of the hammerhead ribozyme molecular dynamic simulation: (1) Opening of the Stem I base pairs. (2) base triple between the G21-A12 mismatch and U2 (color code: red-adenosine, blue-uridine, green-guanosine, orange-cytidine)

Cis Watson Crick base pair U10-A23 opens and the A23 interacts with U35, while A24 forms a base pair with A9. This constructs a fully stacked helix structure between stem II and III at 0.35 ns (Stem II, G21-A12, A22-G11, A23-U35, A24-A9, Stem III). Figure 29 (1) shows the stack depicted with sticks. Later this system collapses with an out-of-stack move of A23. Such out of stack movements of A23 happen several times during the molecular dynamic simulation. A movement of U35 outside the stacked region allow the system to form a new shorter interconnection stack between stem II and III after 1.75ns (Stem II, G21-A12, A22-G11, A24-A9, Stem III) (see Figure 30 (2)). By opening of the Cis Watson Crick base pair U10-A23 the U10 moves to stem I and interacts with the G4 and A38 (see Figure 31 (1)) and other

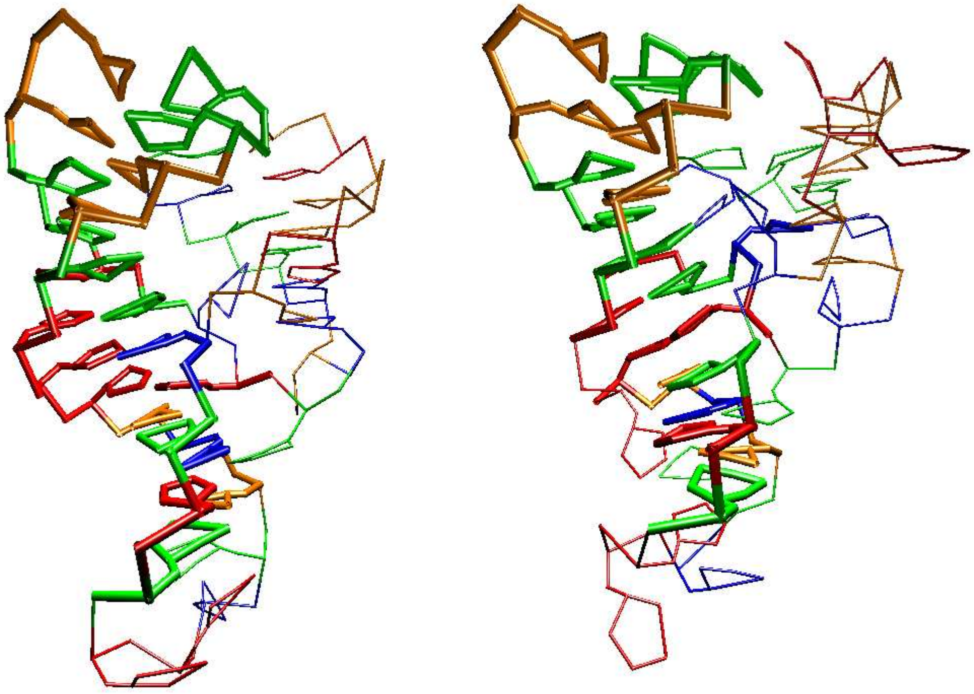


Figure (1) 0.25ns Figure (2) 1.9ns

Figure 30: Snap shots of the hammerhead ribozyme molecular dynamic simulation: Formation of stacked helical regions between stem II and III. (color code: red-adenosine, blue-uridine, green-guanosine, orange-cytidine)

nucleotides around this area.

Since `toyRNA` is unable to describe hydrogen bond interactions including OH2′ of the ribose, it has to be expected that the U-turn will be destroyed during simulation. The only question was the size of the system's damage. The simulation shows that the cis wobble C5-C37 base pair first opens and a rearrangement of the turn takes place. The G8 of the U-turn leaves the stacked position and forms a base triple with the U26-A33 base pair of stem III (see Figure 31(2)). This triple remains stable during the whole simulation time and preserve an U-turn like structure.

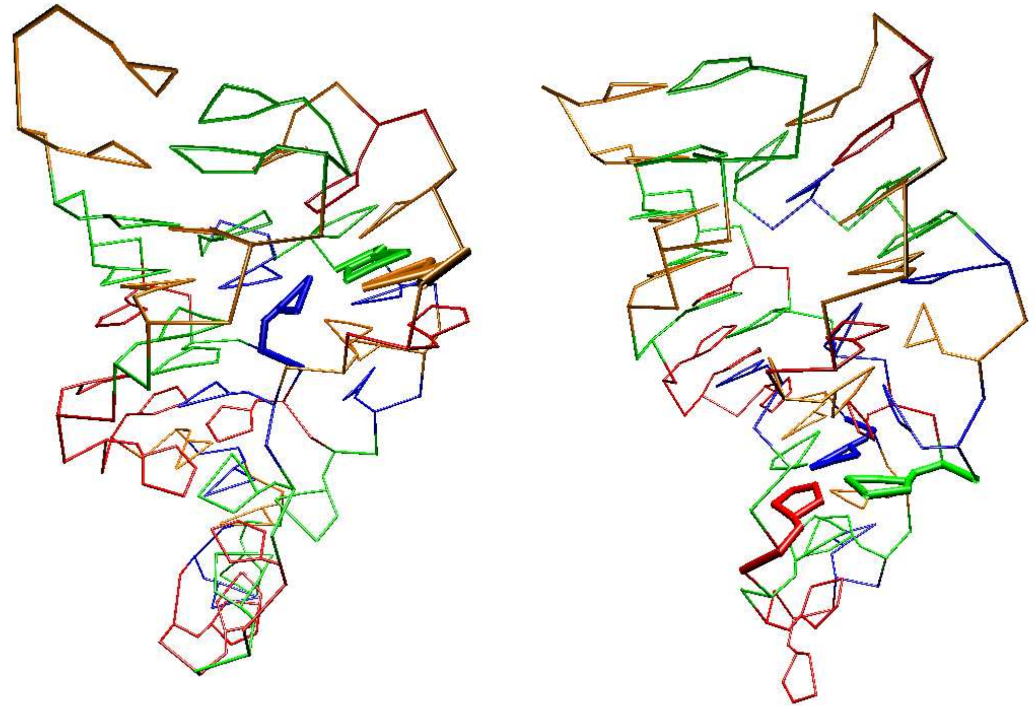


Figure (1) 1.30ns

Figure (2) 0.25ns

Figure 31: Snap shots of the hammerhead ribozyme molecular dynamic simulation: (1) Interaction of C10 and Stem I nucleotides (2) Stabilization of the U-turn with a base triple. (color code: red-adenosine, blue-uridine, green-guanosine, orange-cytidine)

7 Conclusion

This thesis aimed to create a simplified spatial model of RNA that corresponds to an intermediate stage, connecting the world of RNA's secondary structure with its three dimensional conformation. Based on off-lattice force field models the simplified `toyRNA` model was designed and implemented. In this coarse grained model each nucleotide is represented by a composition of seven to nine pseudo atoms. Classical force field potentials are used to define bonds, angles and dihedrals between these pseudo atoms. To describe hydrogen bonding and stacking interaction a new angle dependent potential was designed. The parameters were estimated as far as statistical data were available.

To test the capabilities of the created model, it was applied to three different types of RNA structures: several helices, a hairpin with a tetra loop and the hammerhead ribozyme as an example for a more complex system. Even with the estimated parameters `toyRNA` showed the potential to represent these RNA structures quite adequately. Despite the simplifications the general energy landscape showed no substantial change. The new angle dependent potential was able to describe the base stacking and hydrogen bond interactions in an appropriate way. Also the complex hammerhead ribozyme's structure turned out to be stable in a molecular dynamics simulation.

Another positive effect shown by all calculations is the resulting simplification of the conformational landscape. Thus searching conformational space becomes more easy. Nevertheless is has to be checked, whether there are distortions in the energy landscape caused by the coarse grained potential.

In spite of these good results, some problems of `toyRNA` also become evident. The determination of the parameters is the major problem for all simplified RNA models, and therefore also a problem in `toyRNA`. While the constants of bond and angle potentials can be estimated with the help of statistical data, there are no such data available for torsion and non bonded potentials. These potentials, however, influence strongly molecular conformations. In all cases the selection of parameters for the potentials remains an open problem.

In general, calculations of simplified RNA three dimensional structures are more difficult than for proteins. The forces between the RNA elements are more specific and therefore several restriction have to be met. It was a great challenge to create an appropriate set of potentials, which are able to describe these interactions (The non bonded potential of toyRNA represents the final stage of an extensive search). Describing the stacking interaction in an accurate way is the main problem of the simplification procedure. Each approach for overcoming this problem leads to complex mathematical expressions (e.g. in case of toyRNA the gradient of the angle dependent potential resulted in a calculation of a fivefold cross product). As the non bonded potentials are the time consuming factor in force field calculations these complex potentials slow down the calculation rate. In the case of large molecules this fact might lead to problems with computer time. At this stage the performance of `toyRNA` is not optimized. Therefore no serious prediction about its efficiency can be made. The main advantage of the designed model is the ability of a distinct separation of the three structure determining interactions (backbone, hydrogen bonding, stacking).

Taken together, the results obtained data strongly suggest, that the created model features the typical behavior of RNA structures. Nevertheless, it is still not clear, whether or not the simplified model `toyRNA` has a sufficient precision for determining the three dimensional structure of RNA from its secondary structure.

8 Outlook

The program `toyRNA` is a prototype of a simplified three dimensional RNA force field. Therefore many features of the performance can be improved based upon the prototype. The next step is clearly defined in optimizing the parameters of the model. As there are no experimental data available, these constants must be estimated by statistical methods or alternative approaches. Especially the parameters for the torsion and the non bonded potentials require innovative approaches. One suggestion might be a parameter optimization for example, by means of a genetic algorithm. At present state the polar and ionic character of pseudo atoms except phosphate is ignored. An upgraded model might include Coulomb potentials to direct the nitrogen base plates to the correct relative positions.

Beside the optimation of the parameters also the general form of the potentials can be a goal of further modifications. One candidate for such a modification is the sugar moiety. In `toyRNA` this complex system is described by a simple bond only. Also the problem of bond compression, as mentioned in section 5.3.1, may be a point for future improvements.

Beside the optimization of the force field, other factors can be improved. Such efforts only make sense, when the force field development is in an advanced state. `toyRNA` was designed as a flexible program for testing new potentials. This flexibility reduces the speed of the program. A new improved implementation in a high level language, for example C, will certainly accelerate the computational performance of the approach.

The aim of this thesis was to create a linker between secondary structure of RNA and three dimensional structure. Thus an efficient program must be designed and implemented for converting the secondary structure into the toy spacial structure. With a forthcoming program, which converts the coarse grained structure into atomic resolution, the gap between two and three dimensional representations would be filled.

A Statistics

For the statistical check of the simplified model, 35 structures from the Brookhaven Protein Data Bank [10] were analized. In alphabetical order these files are:

`1a3m.pdb 1a4d.pdb 1a60.pdb 1aqo.pdb 1ato.pdb 1atv.pdb 1atw.pdb`
`1bj2.pdb 1bvj.pdb 1cql.pdb 1ebq.pdb 1guc.pdb 1kis.pdb 1mis.pdb`
`1qc8.pdb 1rau.pdb 1rna.pdb 1rng.pdb 1rnk.pdb 1rrr.pdb 1tfn.pdb`
`1u2a.pdb 1uuu.pdb 1vop.pdb 1zif.pdb 255d.pdb 259d.pdb 280d.pdb`
`283d.pdb 28sr.pdb 2a9l.pdb 373d.pdb 3php.pdb 420d.pdb 433d.pdb`

These files were controlled for irregularies and include the broad spektrum of strukture motivs in RNAs.

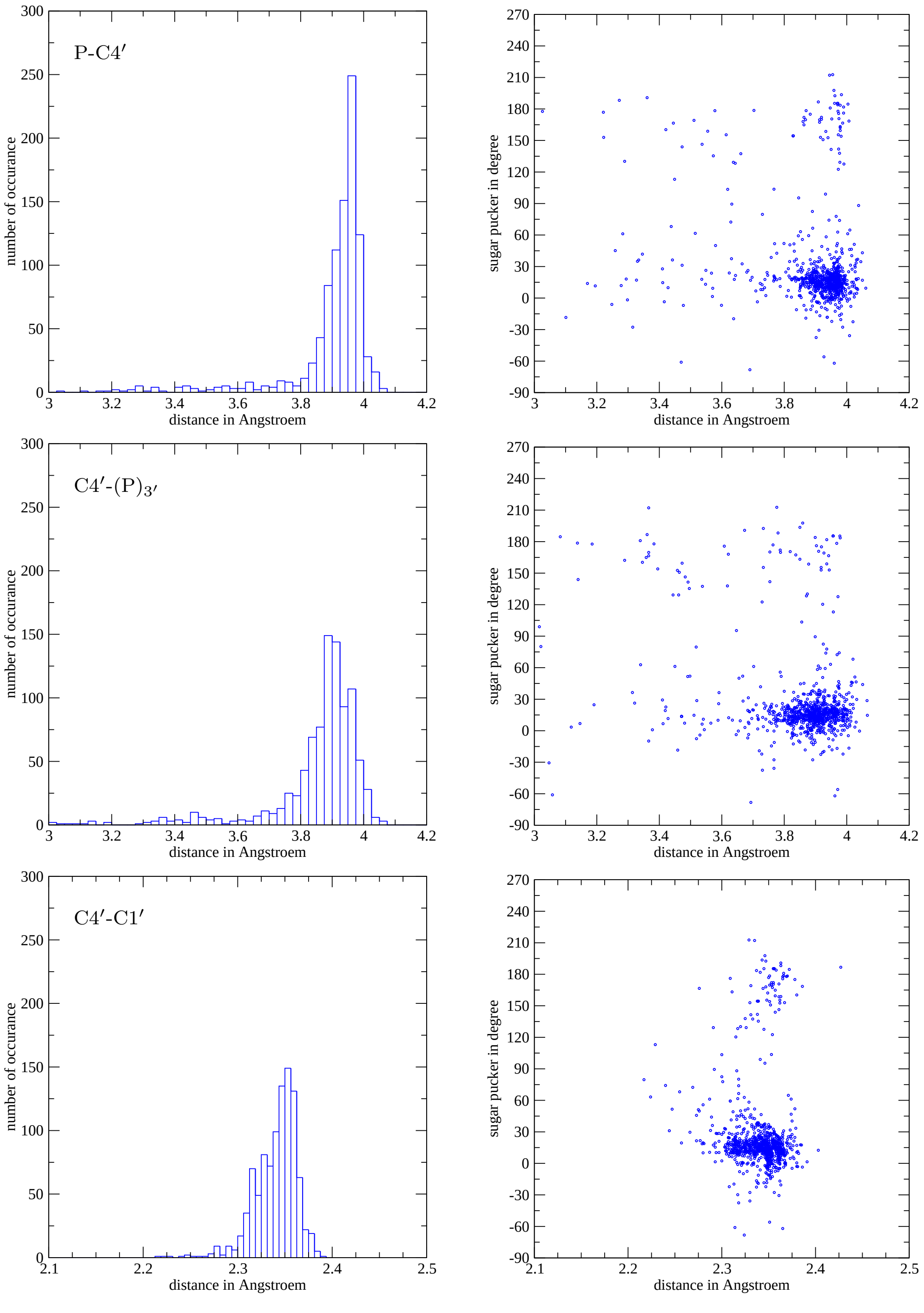


Figure 32: Left: Statistics of the backbone bond distances between the pseudo atoms of `toyRNA` Right: Statistics of the correlation between the bond distances and the sugar pucker

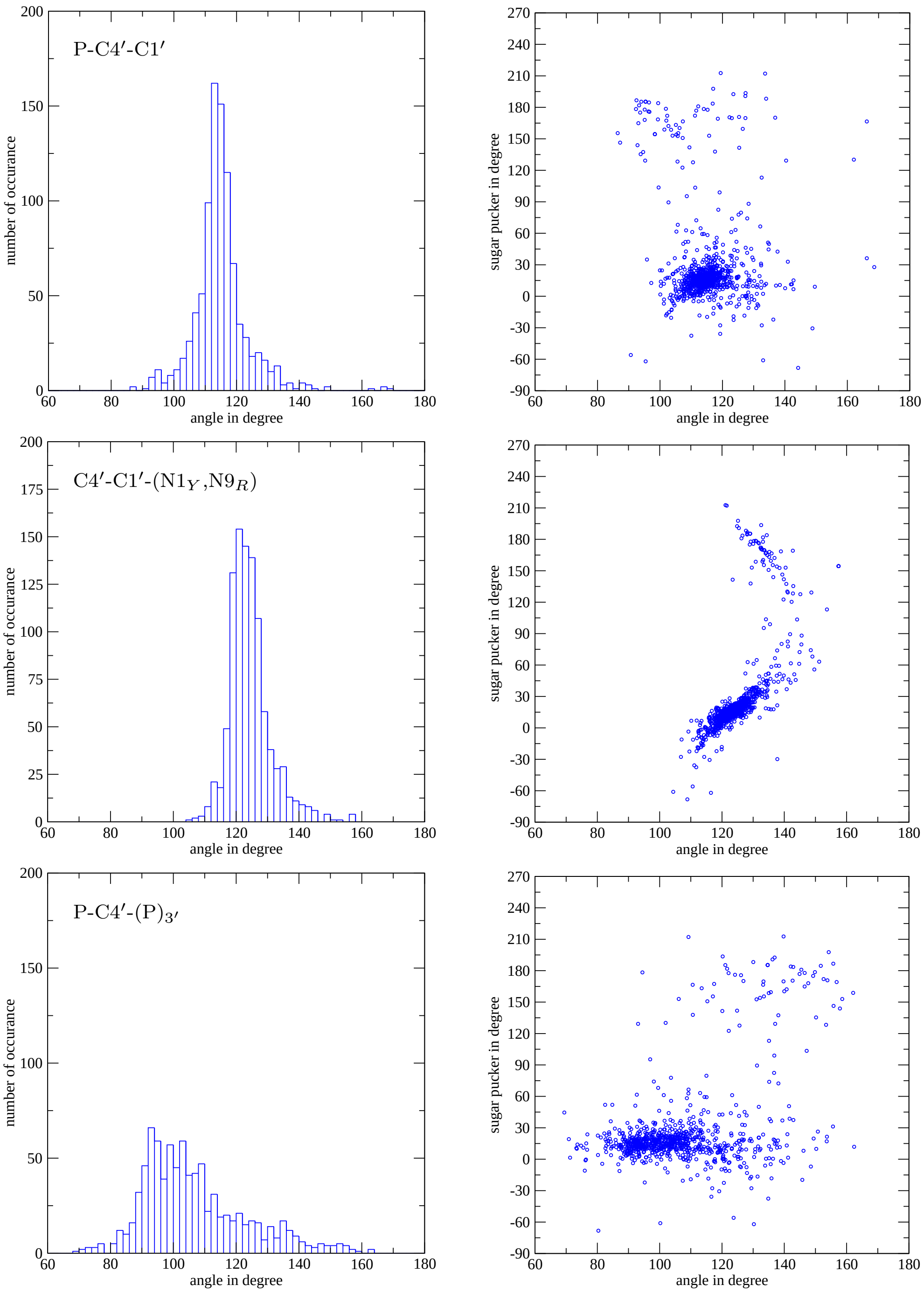
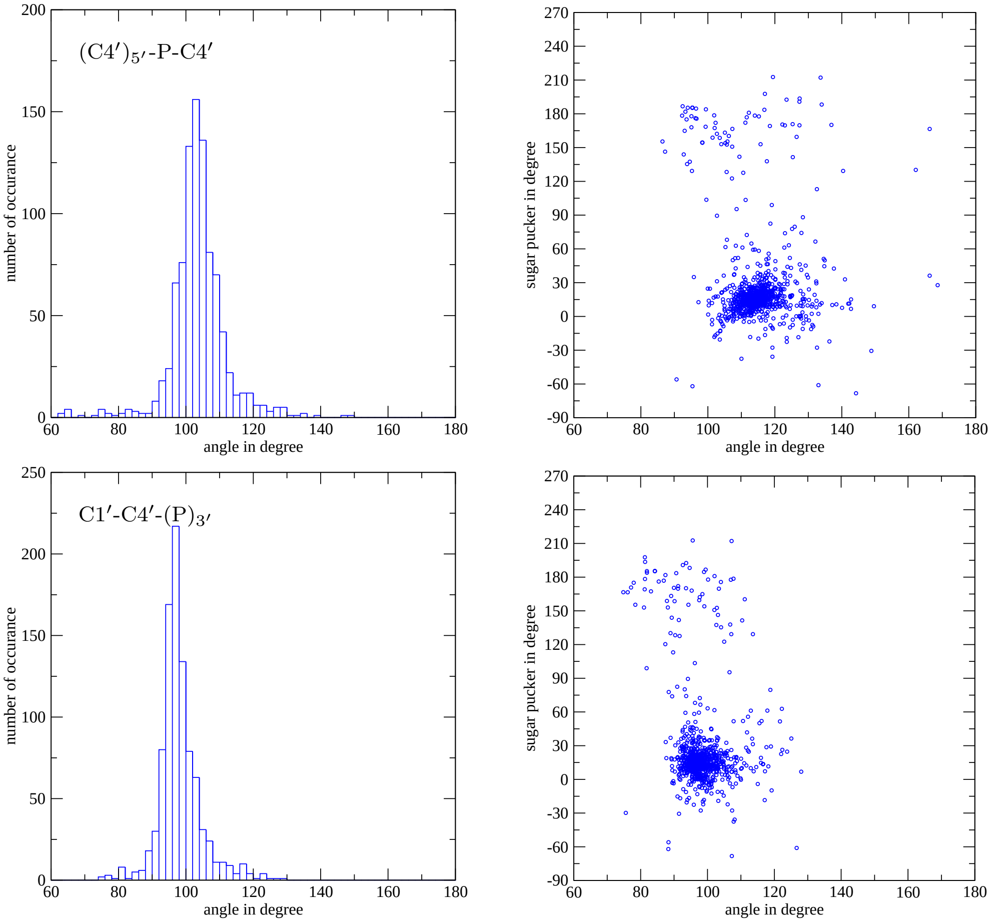


Figure 33: Left: Statistics of the backbone angles between the pseudo atoms of `toyRNA` Right: Statistics of the correlation between these angles and the sugar pucker



$(C4')_{5'}$-P-C4'

C1'-C4'-$(P)_{3'}$

Figure 34: Left: Statistics of the backbone angles between the pseudo atoms of `toyRNA` Right: Statistics of the correlation between these angles and the sugar pucker

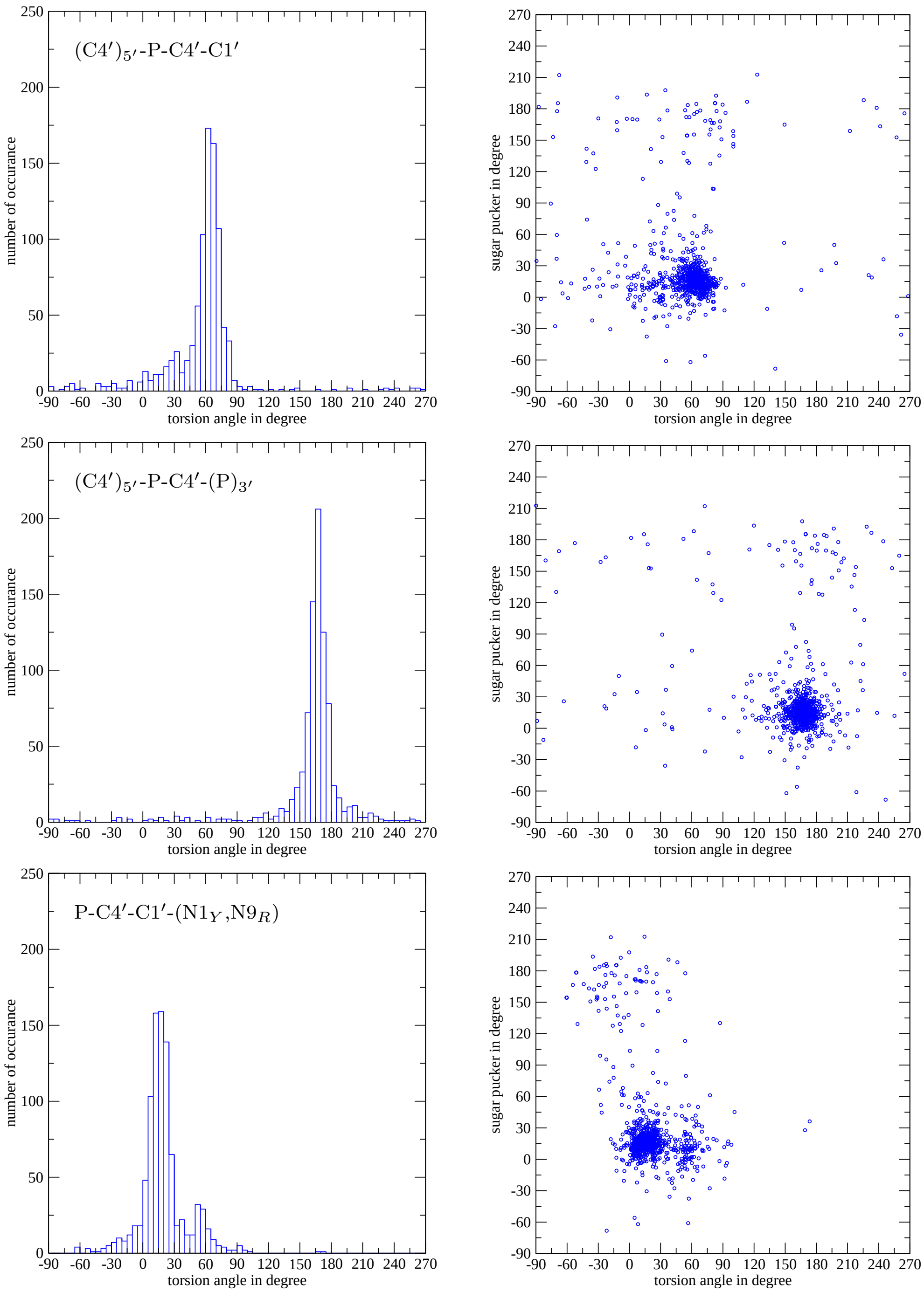


Figure 35: Left: Statistics of the backbone torsion angles between the pseudo atoms of `toyRNA` Right: Statistics of the correlation between these angles and the sugar pucker

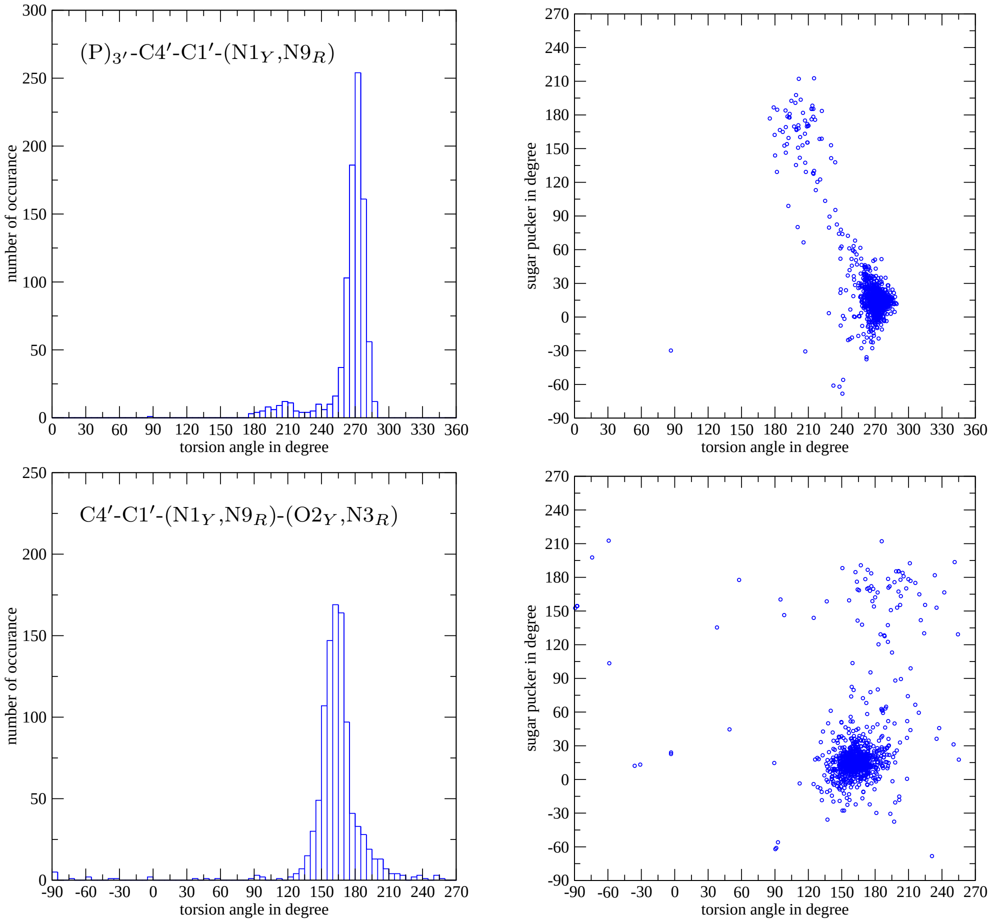


Figure 36: Left: Statistics of the backbone torsion angles between the pseudo atoms of `toyRNA` Right: Statistics of the correlation between these angles and the sugar pucker

List of Figures

1 Atomic structure of RNA 5
2 Primary, secondary and tertiary Structure of an RNA 6
3 Secondary structure motifs in RNA 8
4 Base pairs with at least two hydrogen bonds 11
5 Non standard hydrogen-bonding patterns 12
6 Three edges for possible interaction as recommended by Westhof 12
7 Optimal stacking geometries 14
8 Major puckering modes of sugars in RNA 15
9 Four component picture of inter- and intramolecular forces . . 25
10 Basic information flow in AMBER 6.0 34
11 Structure of the internal variables in JUMNA 36
12 Input file for MC-SYM for a simple stem-loop structure 39
13 20 essential amino acids 43
14 Structure of the amide-bond 44
15 Transformation to the simplified model 50
16 Schematic description of the angle dependent potential 56
17 Geometrical base-base potential 57
18 Energy surface of the non bonded potentials 60
19 Optimized helices structures with adenosines and uridines . . . 63
20 Optimized helices structures with guanosines and cytidines . . 64
21 Superposed structures of the original and optimized helices . . 65
22 Comparison of the mean and minimized structure 67
23 Two snap shots of the `AAGGCCU` duplex MD trajectory 68
24 Experimental and optimized conformers of 1zih.pdb 71
25 Two group of snap shots of the tetra loop MD trajectory . . . 72
26 Secondary and tertiary structure of the hammerhead ribozyme 74
27 Optimized structure of the hammerhead ribozyme 75
28 Compression of the hammerhead molecule structure 76
29 Snap shots of the hammerhead molecular dynamic simulation I 77
30 Snap shots of the hammerhead molecular dynamic simulation II 78
31 Snap shots of the hammerhead molecular dynamic simulation III 79
32 Statistics of the backbone bond distances 84
33 Statistics of the backbone angles (Part I) 85
34 Statistics of the backbone angles (Part II) 86
35 Statistics of the backbone torsion angles (Part I) 87

36 Statistics of the backbone torsion angles (Part II) 88

List of Tables

1 Atoms representing the nitrogen bases of the four nucleotides . 49
2 Bond parameters 52
3 Angle parameters 53
4 Torsion parameters 54
5 Energies and root mean square distances of ten RNA helices . 62
6 RMSD of the molecular dynamics' structure snap shots 66
7 Energies and rmsd of ten GCAA tetra loops conformeres . . . 69
8 Energies and rmsd of ten GCAA tetra loops conformeres . . . 70

References

[1] F. Aboul-ela, J. Karn, and G. Varani. Structure of HIV-1 TAR RNA in the absence of ligands reveals a novel conformation of the trinucleotide bulge. *Nucl. Acids Res.*, 24:3974–3981, 1996.

[2] B. J. Alder and T. E. Wainwright. Phase transition for a hard-sphere system. *J. Chem. Phys.*, 27:1208–1209, 1957.

[3] D. H. Andrews. The relation between the raman spectra and the structure of organic molecules. *Phys. Rev.*, 36:544–554, 1930.

[4] P. Auffinger, S. Louise-May, and E. Westhof. Hydration of C-H groups in tRNA. *Faraday Discuss.*, 103:151–173, 1996.

[5] P. Auffinger and E. Westhof. RNA hydration: Three nanoseconds of multiple molecular dynamics simulation of the solvated tRNA^{Asp} anticodon hairpin. *J. Mol. Biol.*, 269:326–341, 1997.

[6] N. Ban, P. Nissen, J. Hansen, P.B. Moore, and T. A. Steitz. The complete atomic structure of the large ribosomal subunit at 2.4Å resolution. *Science*, 289:905–920, 2000.

[7] R. Basavappa and P. B. Sigler. The 3Å crystal structure of yeast initiator tRNA: functional implications in initiator-elongator discrimination. *EMBO J.*, 10(10):3105, 1991.

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

[9] R. Bellman. On the theory of dynamic programming. *Proc. Natl. Acad. Sci. USA*, 38:716–719, 1952.

[10] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliand, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[11] S. C. Blanchard and Pluglisi J. D. Solution structure of the A loop of 23S ribosomal RNA. *Proc. Natl. Acad. Sci.*, 98(7):3720–3725, March 2001.

[12] A. Bondi. van der waals volumes and radii. *J. Phys. Chem.*, 68:441, 1964.

[13] M. Born and R. Oppenheimer. Zur Quantumtheorie der Molekeln. *Ann. Phys. (Leipzig)*, 84:457–484, 1927.

[14] J. U. Bowie, R. Lüthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–169, 1991.

[15] B. R. Brooks, R. E. Bruccoleri, B. D Olafson, D. J. States, S. Swaminathan, and M Karplus. CHARMM: A program for macromolecular energy minimization and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.

[16] C. Carola and F. Eckstein. Nucleic acid enzymes. *Curr. Opin. Struct. Biol.*, 3:274–283, 1999.

[17] 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.

[18] 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.

[19] 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.

[20] T. E. Cheatham III and P. A. Kollman. Molecular dynamics simulations highlight the structural differences among DNA:DNA, RNA:RNA and DNA:RNA hybrid duplexes. *J. Am. Chem. Soc.*, 119:4805–4825, 1997.

[21] 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.

[22] M. Clark, R. D. Cramer III, and N. van Obdenbusch. Validation of the general purpose Tripos 5.2 force field. *J. Comp. Chem.*, 10:982–1012, 1989.

[23] 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.

[24] G. M. Crippen. Easily searched protein folding potentials. *J. Mol. Biol.*, 260(3):467–475, 1996.

[25] 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.

[26] K. A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24:1501–1509, 1995.

[27] K. A. Dill, S. Bromberg, K. M. Fiebeg, D. P. Yee, P. D. Thomas, and H. S. Chan. Principles of protein folding - a perspective from simple exact models. *Protein Sci.*, 4:561–602, 1995.

[28] K. A. Dill and H. S. Chan. From levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4(1):10–19, 1997.

[29] A. J. Dingley, J. E. Masse, M. Peterson R. D. Barfield, J. Feigon, and S. Grzesiek. Internucleotide scalar coupling across hydrogen bonds in watson-crick and hoogsten base pairs of a dna triplex. *J. Am. Chem. Soc.*, 121:6019–6027, 1999.

[30] I. Dostrovsky, E. D. Hughes, and C. K. Ingold. The role of steric hindrance. Section G. magnitude of steric effects, range of occurance of steric and polar effects, and place of the wagner rearrangement in

nucleophilic substitution and elimination. *J. Chem. Soc.*, page 173, 1946.

[31] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis. Probabilistic models of proteins and nucleic acids.* Cambridge University Press, 1998.

[32] D. A. Erie, K. J. Breslauer, and W. K. Olson. A monte carlo method for generating structures of short single stranded dna sequences. *Biopolymers*, 33(1):75–105, 1993.

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

[34] A. R. Ferré-D'Amaré, K. Zhou, and J. A. Doudna. Crystal structure of a hepatitis delta virus ribozyme. *Nature*, 395:567–574, 1998.

[35] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[36] W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Taranzona, E. D. Weinberger, and P. Schuster. RNA folding and combinatory landscapes. *Phys. Rev. E*, 47:2083–2099, 1993.

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

[38] 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.

[39] N. Go. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.*, 12:183–210, 1983.

[40] R. H. Griffey, C. D. Poulder, A. Bax, B. L. Hawkins, Z. Yamaizumi, and S. Nishimura. Multiple quantum two dimensional ^{1}H-^{15}N nuclear magnetic resonance spectroscopy: Chemical shift correlation maps for exchangeable imino protons of *E.Coli* tRNAfMet in water. *Proc. Natl. Acad. Sci. USA*, 80:5895–5897, 1983.

[41] A. T. Hagler, E. Huler, and S. Lifson. Energy functions for peptides and proteins. I. derivation of a consistent force field including the hydrogen bond from amide crystals. *J. Am. Chem. Soc.*, 96:5319–5327, 1977.

[42] A. T. Hagler and S. Lifson. Energy functions for peptides and proteins. II. the amide hydrogen bond and calculation of the amide crystal properties. *J. Am. Chem. Soc.*, 96:5327–5335, 1974.

[43] Janos Hajdu. Single-molecule X-ray diffraction. *Curr. Opin. Struct. Biol.*, 10:569–573, 2000.

[44] T. A Halgren. Merck molecular force field. I basis, form scope, parametrization, and performancd of mmff94. *J. Comput. Chem.*, 17:490–519, 1996.

[45] M. R. Hansen, P. Hanson, and A. Pardi. Filamentous bacteriophage for aligning RNA, DNA and proteins for measurement of nuclear magnetic resonance dipolar coupling interactions. *Enzymol.*, 317:220–240, 2000.

[46] M. R. Hansen, P. Hanson, and A. Pardi. Pf1 filamentous phage as an alignment tool for generating local and global structural information in nucleic acids. *J. Biomol. Struct. Dyn.*, 11:365–369, 2000.

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

[48] L. He, Kierzek. R., J. SantaLucia, A. E. Walter, and D. H. Turner. Nearest-neighbour parameters for G-U mismatches. *Biochemistry*, 30:11124, 1991.

[49] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl. Identification of native protein folds amongst a large number of incorrect models — the calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.*, 216:167–180, 1990.

[50] T. Hermann, P. Auffinger, W. G. Scott, and E. Westhof. Evidence for a hydroxide ion bridging two magnesuim ions at the active site of the hammerhead ribozyme. *Nucleic Acids Res.*, 25:3421–3427, 1997.

[51] T. Hermann, P. Auffinger, and E. Westhof. Molecular dynamics investigations of hammerhead ribozyme RNA. *Eur. Biophys. J.*, 27:153–165, 1998.

[52] 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.

[53] T. L. Hill. Steric Effects. *J. Chem. Phys.*, 14:465, 1946.

[54] P. Hobza and J. Sponer. Structure, energetics, and dynamics of the nucleic acid base pairs: Nonempirical *ab initio* calculations. *Chem. Rev.*, 99:3247–3276, 1999.

[55] R. W. Hockney. The potential calculations and some applications. *Methods in Computational Physics*, 9:136–211, 1970.

[56] Ivo L Hofacker, Walter Fontana, Peter F Stadler, L Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

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

[58] IUPAC-IUB. Recommended nomenclature and definitions are given in abreviations and symbols for the description of conformation of polynucleotide chains. *Eur. J. Biochem.*, 131:9–15, 1983.

[59] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA*, 86:7706–7710, 1989.

[60] G. A. Jeffrey. *An Introduction to Hydrogen Bonding.* Oxford University Press, Oxford, UK, 1988.

[61] 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.

[62] M. Karplus. Contact electron-spin coupling of nuclear magnetic moments. *J. Chem. Phys.*, 30:11–15, 1959.

[63] G. W. Kellogg and B. I. Schweitzer. Two and three-dimensional ^{31}P-driven RME procedures for complete assignment of backbone resonances in oligodeoxyribosenucleotides. *J. Biomol. NMR*, 3:577–595, 1993.

[64] S.-H. Kim, G. J. Quigley, F. L. Suddath, A. McPherson, and D. Sneden. Three- dimensional structure of yeast phenylalanine transfer RNA at 3.0Å resolution. *Science*, 179:285 – 288, 1973.

[65] S. Kirkpatrick, C. Gelatt, and M. Vecchi. Optimation by simulated annealing. *Sciene*, 220:671–680, 1983.

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

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

[70] R. Lavery, K. Zakrzewska, and B. Pullman. The flexibility of nucleic acids (II): The calculation of internal energy and applications to mononucleotide repeat DNA. *J. Biomol. Struct. Dyn.*, 4:989 – 1014, 1986.

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

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

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

[74] Neocles B. Leontis and Eric Westhof. Geometric nomenclature and classification of rna basepairs. *RNA*, 7:499–512, 2001.

[75] M. Levitt. A simplified represenation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*, 104:59–107, 1976.

[76] M. Levitt and A. Warshel. Computer simulation of protein folding. *Nature*, 253:694–698, 1975.

[77] S. Lifson and A. Warshel. Consistent force field for calculations of conformation, vibrational spectra and enthalpies of cycloalkane and n-alkane molecules. *J. Chem. Phys.*, 49:5116–5129, 1968.

[78] D. M. J. Lilley. Structure, folding and catalysis of small nucleolytic ribozymes. *Curr. Opin. Struct. Biol.*, 9:330–338, 1999.

[79] G. Lippens, C. Dhalluin, and J.-M. Wieruszeski. Use of a water flip-back pulse in the homonuclear NOESY experiment. *J. Biomol. NMR*, 5:327–331, 1995.

[80] A. Liwo, S. Ołdziej, M. R. Pincus, J. R. Wawak, S. Rackovsky, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comp. Chem.*, 18(7):850–873, 1997.

[81] A. Liwo, M. R. Pincus, J. R. Wawak, S. Rackovsky, S. Ołdziej, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by z-score optimation. *J. Comp. Chem.*, 18(7):874–887, 1997.

[82] A. D. Mac Kerell Jr., B. Brooks, C. L Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus. CHARMM: The energy function and its parametrization with an overview of the program. In P. v. R. Schleyer, editor, *The Encyclopedia of Computational Chemistry*, volume 1, pages 271–277. John Wiley & Sons: Chichester, 1998.

[83] 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.

[84] M. Major, D. Gautheret, and Cedergren R. Reproducing the three-dimensional structure of a transfer RNA molecule from structural constraints. *Proc. Natl. Acad. Sci. (USA)*, 90:9408–9412, 1993.

[85] 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.

[86] A. Malhotra, H. A. Gabb, and S. C. Harvey. Modeling large nucleic acids. *Curr. Opin. Struct. Biol.*, 3:241–246, 1993.

[87] S. L. Mayo, B. D. Olafson, and W. A. Goddard III. DREIDING: A generic force field for molecular simulations. *J. Phys. Chem.*, 94:8897–8909, 1990.

[88] D. B. Mc Kay and J. E. Wedekind. *The RNA World*, chapter Small ribozymes: The role of metal ions in RNA biochemistry, pages 265–286. Cold Spring Habor Labortory Press, Cold Spring Habour, NY, 2nd edition, 1999.

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

[90] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller. Equation of state calculations by fast computing maschines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[91] S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules*, 18:534–552, 1985.

[92] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. *J. Mol. Biol.*, 256:623–644, 1996.

[93] E. T. Mollowa, M. R. Hansen, and A. Pardi. Gobal structure of RNA determined with residual dipolar coupling. *J. Am. Chem. Soc.*, 122:11561–11562, 2000.

[94] Emilia T. Mollowa and Arthur Pardi. NMR solution strucutre determination of RNAs. *Curr. Opin. Struct. Biol.*, 10:298–302, 2000.

[95] D. Moras, Comarmond M. B., J. Fischer, R. Weiss, and J. C. Thierry. Crystal structure of yeast tRNA(asp). *Nature*, 288:669, 1980.

[96] R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, and J. Hadju. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature*, 406:752–757, 2000.

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

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

[99] H. Nymeyer, A. E. Garcia, and J. N. Onuchic. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl. Acad. Sci.*, 95:5921–5928, 1998.

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

[101] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Commun.*, 91:1–41, 1995.

[102] M. Piotto, V. Saudek, and V. Sklenár. Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR*, 2:661–665, 1992.

[103] C. W. Pleij. RNA pseudoknots. *Curr. Opin. Struct. Biol.*, 4:337–344, 1994.

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

[105] D. Pörschke. *Chemical Relaxation in Molecular Biology. Molecular Biology, Biochemistry, and Biophysics*, volume 24, chapter Elementary steps of base recognition and helix-coil transitions in nucleic acids, pages 191–218. Springer-Verlag Berlin, 1977.

[106] J. D. Puglisi, J. R. Wyatt, and Ignatio Tinoco Jr. Conformation of an RNA pseudoknot. *J. Mol. Biol.*, 214:437–453, 1990.

[107] E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, 1999.

[108] E. Rivas and S. R. Eddy. The language of RNA: a formal grammar that includes pseudoknots. *Bioinformatics*, 16:334–340, 2000.

[109] J. M. Rosenberg, N. C. Seeman, R. O. Day, and A. Rich. RNA double-helical fragments at atomic resolution. *J. Mol. Biol.*, 104:145, 1976.

[110] J. P. Ryckaert, G. Cicotti, and H. J. C. Berensden. Numerial integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23:327–341, 1977.

[111] W. Saenger, M. Turcotte, G. Lapalme, and F. Major. Exploring the conformation of nucleic acids. *J. Funct. Program.*, 5:443–460, 1995.

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

[113] J. S. Santa Lucia Jr., L. X. Shen, Z. Cai, H. Lewis, and I. Tinoco Jr. Synthesis and NMR of RNA with selective isotope enrichment in the base moieties. *Nucl. Acids Res.*, 23:4913–4921, 1995.

[114] R. W. Schevitz, A. D. Podjarny, Krishnanachari, J. J. Hughes, P. B. Sigler, and J. L. Sussman. Crystal structure of a eukaryotic initiator tRNA. *Nature*, 278:188, 1979.

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

[116] P. Schuster and P. Wolschann. Hydrogen bonding: From small clusters to biopolymers. *Monatshefte für Chemie*, 130:947–960, 1999.

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

[118] N. C. Seeman, J. M. Rosenberg, F. L. Suddath, J. J. P. Kim, and A. Rich. RNA double-helical fragments at atomic resolution. *J. Mol. Biol.*, 104:109, 1976.

[119] J.-E. Shea, Y. D. Nochomovitz, Z. Guo, and C. L. Brooks III. Exploring the space of protein folding Hamiltonian: The balance of forces in a minimalist β-barrel model. *J. Chem. Phys.*, 109:2895–2903, 1998.

[120] J. Skolnick and A. Kolinski. Simulations of the folding of a globular protein. *Science*, 250:1121–1125, 1990.

[121] P. E. Smith and B. M. Pettit. Modelling solvent in biomolecular systems. *J. Phys. Chem.*, 98:9700–9711, 1994.

[122] J. Sponer, I. Berger, N. Spackova, J. Leszczynski, and P. Hobza. Aromatic base stacking in DNA: From *ab initio* calculations to molecular dynamics simulations. *J. Biomol. Struct. Dyn.*, 21:383–407, 2000.

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

[124] 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.

[125] J. L. Sussman, S. R. Holbrook, R. W. Warrant, G. M. Church, and Kim S.-H. Crystal structure of yeast phenylalanine transfer RNA. *J. Mol. Biol.*, 123:607 – 630, 1978.

[126] W. C. Swope, H. C. Anderson, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium contants for the formation of pysical clusters of molecules: Application to small water clusters. *J. Chem. Phys.*, 76:637–649, 1982.

[127] 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.

[128] R.K.Z. Tan and S.C. Harvey. Yammp: Development of a molecular mechanics program using the modular programming method. *J. Comp. Chem.*, pages 455–470, 1993.

[129] N. Tjandra and A. Bax. Direct mesurement of distances and angles in biomolecules by NMR. *Science*, 278:1111–1114, 1997.

[130] L. Trantirek, Urbášek M., R. Štefl, J. Fegion, and Sklenár V. A method for direct determination of helical parameters in nucleic acids usng residual dipolar coupling. *J. Am. Chem. Soc.*, 122:10454–10455, 2000.

[131] 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.

[132] D. H. Turner, N. Sugimoto, and S. Freier. RNA structure prediction. *Ann. Rev. Biophys. Biophys. Chem.*, 17:167–192, 1988.

[133] W. F. van Gunsteren and H. J. C Berendsen. Groningen Molecular Simulation (GROMOS) Library Manual. Biomos, Nijenborgh 16, Groningen, NL, 1987.

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

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

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

[137] 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.

[138] Michael S. Waterman. *Introduction to Computational Biology*. Chapman & Hall/CRC, 1995.

[139] J.D. Watson and F.H.C. Crick. Genetic implications of the structure of deoxyribonuclec acid. *Nature*, 171:964–967, 1953.

[140] J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids. A structure for deoxyribonuclec acid. *Nature*, 171:737–738, 1953.

[141] D. H. Wertz and N. L. Allinger. Conformational analysis-129 ; Heats of formation and thermodynamic parameters for hydrocarbons, calculated by molecular mechanics method including the effect of molecular vibrations. *Tetrahedron*, 35(1):3–12, 1979.

[142] Eric Westhof and Valerie Fritsch. Rna folding: beyond Watson-Crick pairs. *Structure*, 8(3):R55–R65, 2000.

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

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

[145] B. T. Wimberly, D. E. Broderson, W. M. Clemons, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. Structure of the 30S ribosomal subunit. *Nature*, 407:327–339, 2000.

[146] J. Wöhnert, A. J. Dingley, M. Stoldt, Görlach M., S. Grzesiek, and L. R. Brown. Direct identification of NH$\cdots$N hydrogen bonds in non canonical base pairs of RNA by NMR spectroscopy. *Nucleic Acids. Res.*, 27:3104–3110, 1999.

[147] N. H. Woo, B. A. Roe, and A. Rich. Three-dimensional structure of escherichia coli initiator tRNA(met). *Nature*, 286:346 – 351, 1980.

[148] H. T. Wright, P. C. Manor, K. Beurling, R. L. Karpel, and J. Fresco. In *Transfer RNA: Structure, Properties, and Recognition.*

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

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

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

[152] Michael Zuker. Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, 10:303–310, 2000.

Curriculum vitae

Mag. Kurt Grünberger
1968-12-31

Schulbildung: Volksschule in Wien
Integrierte Gesamtschule
Bundesrealgymnasium 17 in Wien
Matura am 13.Juni 1989

Studium: Beginn Oktober 1989
1.Diplomprüfung am 6.Juli 1994
2.Diplomprüfung am 28.November 1997

Diplomarbeit: Organische Naturstoffsynthese
(Februar 1996 – November 1997 bei Prof. Edda Gössinger)
Titel: Versuche zur enantiomerenreinen Darstellung von
5,6-Dihydro-6-allyl-2*H*-pyran-2,4(3*H*)-dion
im Rahmen der Totalsynthese der Cochleamycine

Dissertation: Theoretische Chemie
Beginn 1.Oktober 1998 bei Prof. Peter Schuster
Titel: A 3D-Model for coarse grained structure prediction of RNA