Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

A. BOIT, F. CORDES

RNA 3D-Modeling

ZIB-Report 03-48 (December 2003)

RNA 3D-Modeling

A. Boit, F. Cordes

Abstract

This article presents a new computational approach to the three-dimensional (3D) modeling of ribonucleic acid (RNA) sequences with unknown spatial structure. The main concept is a mapping of the query sequence onto the 3D structures of a suitable template RNA molecule. This technique called *threading* has originally been developed for the modeling of protein 3D structures. The application to RNA systems bridges the information gap between the growing mass of RNA sequence data and the relatively limited number of available 3D structures. The new RNA threading method is demonstrated on a tRNA model system because sufficient representative 3D structures have experimentally been elucidated and deposited in the public databases. Nevertheless, the method is in principle transferable on all other RNA species. Algorithms are developed that decompose these template structures into their secondary structure elements and gather this information in a specific template database. The best template is chosen with public alignment and secondary structure prediction tools which are integrated in the RNA modeling module. The structural information gathered from the template and the best alignment is combined to establish a comprehensive 3D model of the query sequence. A range of complete tRNA structures has successfully been modeled with the RNA threading method. The prototype module visualizes the models and provides convenient access to the proposed 3D structures. Therefore, the method could give new insight into a variety of RNA systems which in the recent years have become increasingly important as potential new pharmaceutical agents.

Key words. threading, alignment, secondary structure, loop decomposition, template library, relaxation.

Mathematics subject classification. 92C40, 62P10.

1 Introduction

The coherence between a biopolymer sequence and its 3D structure is one of the most intriguing and challenging biochemical issues of our time. Although most of the research on tertiary folding traditionally attends to proteins, the same questions can be posed for RNA molecules. Following in the wake of modern RNA technologies, a variety of new RNA species has emerged whose functionality reaches far beyond their traditionally acknowledged coding capacity(for an overview, see [1, 2, 3, 4, 5, 6, 7]). Though the majority of RNA molecules does not form 3D architectures as intricate as

proteins do, RNA folds are on a comparable level of complexity and RNA-RNA, RNA-DNA and RNA-protein interactions regulate a wide range of cellular activities. Since structure conservation is one of the basic principles in molecular biology, identifying fold relationships by computational means holds promising potential for modeling still unclarified 3D structures of RNA sequences. Modeling 3D RNA structures *in silico* is at the heart of the matter because reliable fold prediction paves the way to understanding the various functions RNA exhibits.

Threading methods are *Fold Recognition Techniques* which aim at modeling 3D structures of proteins on the basis of a sequence-to-structure comparison [8, 9, 10, 11, 12]. A *threading* defines a particular alignment between query sequence and template structure which has been selected from the large number of possible alignments due to energy-related criteria. This way the term *threading* specifies the more general term *alignment* to a mapping of a sequence that is being arranged on a template structure.
Threading techniques are characterized by a common pattern:

- Known template structures provide a set of coordinates in 3D space.
- The coordinates combined with the annotated backbone trace provide potential positions for the residues of the query sequence.
- The query sequence is threaded onto the template structure, while loops and coiled regions are modeled separately.
- The mapping of a certain query sequence to each template results in different candidate threadings (alignments).
- A score function distinguishes valuable threadings from the decoys and identifies the best possible template structure.

According to this scheme, a threading method has been developed and applied to a model system of tRNA query sequences and template structures. It provides an environment for the modeling of a 3D structure to a given RNA query sequence.
Their specific properties predestine RNAs to be successfully folded by threading algorithms [13, 14, 15]: RNA folding is hierarchial in the way that secondary structure is much more stable than tertiary folding. The energies involved in the stabilization of secondary structure elements are much larger than those involved in the tertiary interactions. The folding pathway is predominantly unidirectional so that tertiary interactions form after the stems and loops have arranged themselves in an energetically favourable position.
The following three paragraphs introduce external software packages which have been used throughout the development of the new RNA Threading method. Comprehensive background information about the employed

algorithms can be gathered from the corresponding publications.
The alignment tool associated with the RNA Threading module is a self-contained package called *RAGA (RNA Sequence Alignment by Genetic Algorithm)* developed by C. Notredame, E. A. O'Brien and D. G. Higgins: [23, 24, 25].[1]. The main emphasis of RAGA is on identifying conserved base pairs, thereby mapping the query sequence both on the template sequence and on its secondary structure. RAGA employs a genetic algorithm (GA) derived from the GA described by D. E. Goldberg to optimize the score of the sequence-to-structure alignment [26].
Furthermore the program *RNAfold*, a secondary structure prediction tool within the *Vienna package* [16, 21][2], has been integrated in the RNA Threading module. It determines the base pairs of unaligned regions in the query sequence that do not match with the main template structure at all (Figure 1).
The last step of the RNA Threading algorithm is a force-field based energy minimization which refines the initial 3D model structures. The potential function is described by the *Merck Molecular Force Field (MMFF94)* [27]. The refinement process is performed by the program zibMol® which has been developed at the ZIB for Conformation Dynamics simulations.

2 Methods and Algorithms

Concept
The RNA Threading algorithm involves all structural levels of RNA architecture (Figure 1). While the input sequence is known at the primary structure level, the folds of the different 3D templates have been verified experimentally. Two databases comprising sequence and secondary structure information of different template structures have been established. Whereas the main template database consists of the data of entire tRNA molecules, the loop database provides the structures of small hairpin loop motifs.
Structural information is exchanged at the secondary structure level. For this purpose, the template structures are decomposed into their secondary structure elements. The secondary structure of the query sequence is derived by a combination of algorithms operating at the secondary structure level. The query sequence is read into a sequence-to-secondary structure alignment tool (RAGA) referring to the main template database. The highest-scoring output alignment assigns base pair positions to the query sequence and specifies the best template structure. Loop regions which occur only in the query sequence are submitted to a secondary structure prediction tool (Vienna Package). Both the loop sequence and its predicted secondary structure are

[1](http://igs-server.cnrs-mrs.fr/)
[2](http://www.tbi.univie.ac.at/ivo/RNA/)

used in a second sequence-to-secondary structure alignment of the region resorting to the loop template database. The complete secondary structure of the query sequence is mapped onto the 3D fold of the best possible template resorting to the coordinates of the template atoms. The result is a 3D model of the query sequence.

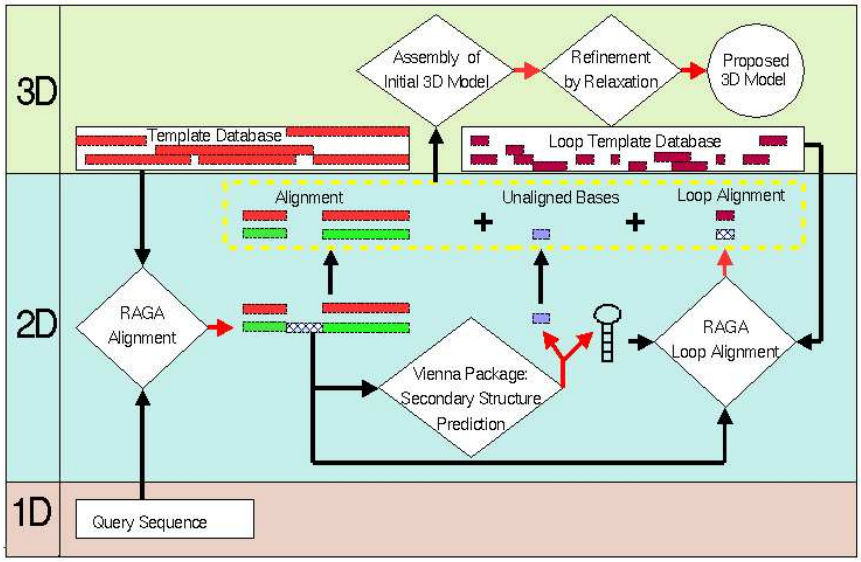


Figure 1: Illustration of the algorithmic network underlying the RNA threading method.

Preliminary Routines

Developing a prototype for an RNA Threading module, it is plausible to choose a template RNA class for which sufficient 3D structures have been elucidated that can serve as spatial scaffolds for the query sequence. The template database for the RNA Threading module is a tRNA database. A total of 28 unique 3D structures from the PDB[3] and Baltimore RNA[4] databases has been processed and gathered in an intern template database. Apart from the main, tRNA template database, a loop template database for the separate modeling of small regions that do not find any close match among the tRNA template molecules has been established. Currently, the loop motif selection for the RNA Threading module is limited to 100 tri- and tetraloops listed in the SCOR[5].

The RNA Threading module performs three basic functions on the template structures:

- Base pair detection,

[3]Protein Database, www.rcsb.or, 09/19/03
[4](http://www.rnabase.org, July 2003)
[5]http://scor.lbl.gov/scor.html, July 2003

- Base pair classification, and
- Loop decomposition.

Base Pair Detection. When reading in a template molecule of N residues from its coordinate file, the position vectors of those atom types which both pyrimidine (C, U) and purine (A, G) bases have in common are stored in separate arrays.

Let $B = b_1, b_2, \ldots, b_N$ be the RNA sequence of N residues and b_{ij} be a base pair between b_i and b_j, $i, j = 1, 2, \ldots, N; i < j$. The base pair detection routine identifies all possible base pair interactions between any two bases which fulfil the following four criteria:

1. The distance between the C_1^* atoms of two bases b_i, b_j must be ≥ 8.0Å. This excludes all cases in which the indices of the two residues do not fulfil the condition $j - i > 3$.

2. The distance between the two base hexagons must be ≤ 7.0Å. The center of reference $\vec{c}_i$ for the pyrimidine and purine rings of base b_i is defined as:
$$\vec{c}_i = 1/3 \left(\vec{N}_{1,i} + \vec{C}_{4,i} + \vec{C}_{5,i} \right)$$
This condition guarantees that the base rings are oriented in a plane with the edges of the hexagons facing each other.

3. The propeller twist (Figure 2) between the two base pair-planes must be $\mid \alpha \mid \leq 35\ ^\circ$.

   The upper limit of $\mid \alpha \mid \leq 35^\circ$ is set to ensure that, apart from the standard, canonical base pairs[6], all noncanonical base pairs are detected. Note that this does not imply that they are already classified as such in this step. Canonical base pairs in an A-form RNA helix have an avarage propeller twist of 16°, but noncanonical pairs, especially interactions between bases stacked inside hairpin loops, can be much more contorted [19]. Hence, the limit is set not too tightly in order to cover all potential base pair positions.

4. The normal projection of the connecting vector of the two N_1 atoms and one of the base planes is a measure for the distance between the base planes. The norm d_{ij} of this projection vector must be ≤ 2 Å (Figure 2). It is defined by the scalar product of either of the normals $\vec{n}_i$ or $\vec{n}_j$ and the interconnecting vector between $N_{1,i}$ and $N_{1,j}$:
$$d_{ij} = \parallel \left(\vec{n}_i \bullet \left(\vec{N}_{1,j} - \vec{N}_{1,i} \right) \right) \parallel$$

[6] A–U, U–A, C–G, G–C pairs with Watson-Crick//Watson-Crick geometry [18]

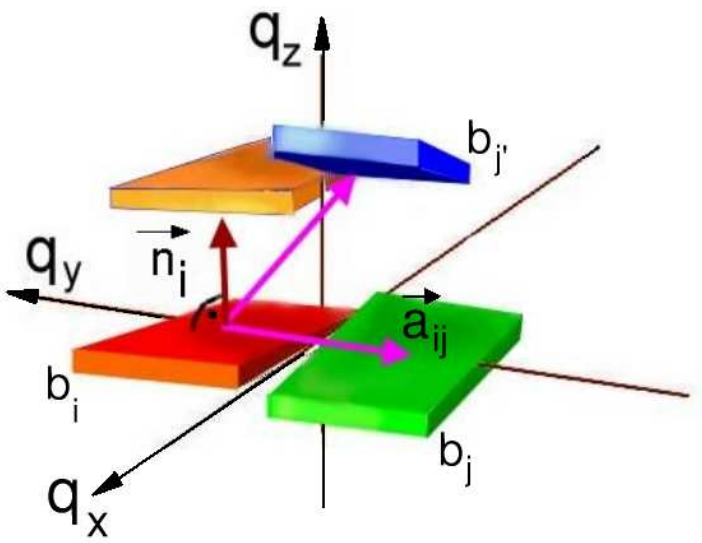


Figure 2: Schematic illustration of the distance between two base planes. The bases b_i and b_j are base pair partners. The vector between the N_1 atoms is represented by $\vec{a}_{ij}$. The normal projection of this vector must be ≤ 2Å. This way, the routine distinguishes the correct base pair partner b_j from bases like $b_{j'}$ which fulfil criteria 1-3, but are shifted along the direction of the normal.

Due to the propeller twist between the base planes, the norm d_{ij} would have a slightly different value if the normal $\vec{n}_j$ was used instead of $\vec{n}_i$.

Base Pair Classification. The Base Pair Classification routine tests the donor and acceptor atoms of each base pair if they meet the conditions for hydrogen bond interactions. A score is assigned to one of the interaction edges, if its distance from the center $\vec{c}_{ij} = 0.5(\vec{c}_j + \vec{c}_i)$ of the two base rings is within a cut-off e_{ij} defined by

$$e_{i,j} = \beta\ f; \quad f = \parallel 0.5(\vec{c}_j - \vec{c}_i) \parallel .$$

The cut-off is measured in Angstroem (Å). The value of the scalar β can be adjusted as a parameter; a reasonable value to start with is 1.3. When all hydrogen bonding sites of the base pair in question have been found, the routine scores 2 points for unambiguous donor/acceptor sites, and 1 point for sites that count among two different edges. This way, each base is associated with the edge that gained the highest score. The number of hydrogen bonds is determined by the number of corresponding donor/acceptor pairs whose atoms are within a distance of 3.45Å. Common values for distances between hydrogen bonds are [20]:
$-O-H\ldots \leftarrow 2.88\text{Å}\ldots \rightarrow O$

$-N-H\ldots \leftarrow 3.04\text{Å}\ldots \rightarrow N$.

Identifying the interacting edges is not sufficient for the association with one of the main 12 base pair classes [18]. At least, either the strand orientation or the orientation of the glycosyl bond (*cis* or *trans*) must also be known to distinctly categorize a base pair. For this purpose, a routine testing for parallel or anti-parallel strand orientation has been developed. The local strand orientation for a given base pair b_{ij} can be identified by the relative positions of the two ribose 2'–OH oxygens to a plane that compensates for the propeller twist between the two base planes. If the two 2'–OH oxygens are situated on the same side of this plane, the strand orientation is parallel; if not, it is anti-parallel. The calculation of the plane is done indirectly, using the covariance matrix $\mathbf{C_{ij}}$ of the atoms $N_{1,i},\ C_{4,i},\ C_{5,i},\ N_{1,j},\ C_{4,j}$ and $C_{5,j}$. The eigenvector $\vec{h}$ corresponding to the smallest eigenvalue of the covariance matrix points in the direction of least variance regarding the atom coordinates in relation to the center $\vec{c}_{ij}$ between the two bases. In other words, $\vec{h}$ would be the normal of the plane balancing out the propeller twist. The cosine of the angle α between $\vec{h}$ and either of the vectors pointing to the 2'-OH oxygens is defined by:

$$cos(\alpha_i) = \frac{\vec{O}^*_{2,i} \bullet \vec{h}}{\parallel \vec{O}^*_{2,i} \parallel \parallel \vec{h} \parallel}\ ,\ cos(\alpha_j) = \frac{\vec{O}^*_{2,j} \bullet \vec{h}}{\parallel \vec{O}^*_{2,j} \parallel \parallel \vec{h} \parallel}$$

If $cos(\alpha_i)$ and $cos(\alpha_j)$ have common algebraic signs, the strand orientation is parallel. Otherwise, the strand orientation is anti-parallel. With the combined information about the base pair geometry and the strand orientation, the base pairs can be classified according to the nomenclature proposed by E. Westhof and N. B. Leontis [18] whose classification scheme has been implemented in the RNA Threading module. Each base presents three sites for potential hydrogen bond interactions: the *Watson-Crick Edge*, the *Hoogsteen Edge* and the *Sugar Edge*. In the following, all atom types will be named according to the IUPAC[7] nomenclature. The index notation "*"refers to ribose atoms. The hydrogen bond involves the hydrogen at the *donor(Do)* atom and the free electron pair of the *acceptor(Ac)* atom. The donor, respectively acceptor atoms of the four bases comprise:

[7]International Union of Pure and Applied Chemistry
http://www.chem.qmul.ac.uk/iupac/class/nucle.html 02

	Adenine
Watson-Crick Edge	$N_6(Do), N1(Ac), C_2(Do)$
Hoogsteen Edge	$N_6(Do), N7(Ac)$
Sugar Edge	$C_2(Do), N3(Ac), O_2^*(Do)$
	Guanine
Watson-Crick Edge	$O_6(Ac), N1(Do), N_2(Do)$
Hoogsteen Edge	$O_6(Ac), N7(Ac)$
Sugar Edge	$N_2(Do), N3(Ac), O_2^*(Do)$
	Cytosine
Watson-Crick Edge	$O_2(Ac), N_4(Do), N_3(Ac)$
Hoogsteen Edge	$N_4(Do), C5(Do)$
Sugar Edge	$O_2(Ac), O_2^*(Do)$
	Uracil
Watson-Crick Edge	$O_4(Ac), N_3(Ac), O_2(Ac)$
Hoogsteen Edge	$O_4(Ac), C5(Do)$
Sugar Edge	$O_2(Ac), O_2^*(Do)$

The canonical G–C and A–U pairs are oriented in *cis* Watson-Crick//Watson-Crick geometry, meaning that both bases are oriented with their Watson-Crick Edges facing each other. The term *cis* refers to the mutual orientation of the glycosyl bonds which are situated between the C_1^* and N_9/N_1 atoms in pyrimidine/purine bases [18].

Loop Decomposition. After all base pairs of the template molecule with sequence $B = b_1, b_2, \ldots, b_N$ have been identified and characterized, the secondary structure S can be derived from this information. An RNA secondary structure is a set of base pairs $b_{ij},\ 1 \leq i \leq j \leq\ N$ organized in helices (*stems*) which are separated from each other by a variable number of single bases. Its energy can be described by the sum of the different loop energy contributions.
Based on these considerations, Zuker and Sankoff [16, 21, 22] developed a concept which allows the complete decomposition of any secondary structure S free of overlapping bases[8] into loop elements, including stacked base pairs that pass as loops of size zero (Figure 3). Each loop element consists of a loop-closing base pair b_{ij} and a variable number $a \geq 0$ of interior base pairs and single bases.
The implementation of the loop decomposition concept in the RNA Threading module starts with a list of the base pairs and unpaired bases in an index succession determined by the template sequence. The residue names (A, G,

[8]The condition of non-overlapping bases means that for two base pairs $b_{ij} \in S$ and $b_{i'j'} \in S$ it holds that if $i < i' \Rightarrow j' < j$.

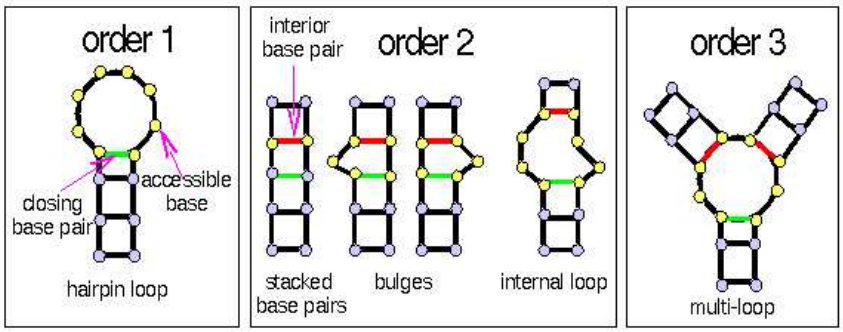
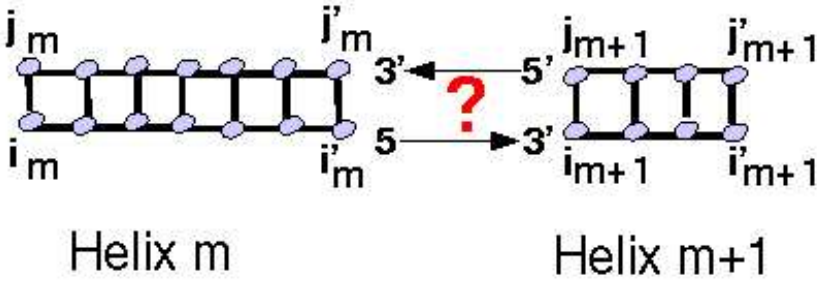


Figure 3: Illustration of the loop orders according to the loop decomposition concept. Note that two stacked base pairs can be seen as a loop of order 2 and size sero. The expression "accessible base" always refers to a specific closing base pair. Hence, hairpin loops comprise accessible bases, but no interior base pair. Stacked base pairs, bulges and internal loops have only one, multi-loops at least two interior base pairs.

T, U)[9] are stored in a list of length N, so that the base b_i at sequence position i occupies the corresponding ith list entry. A second list of the same length contains the information whether the bases are paired or single bases. A "proof-reading" is performed on the secondary structure lists because the preceeding base pair detection routine might have codified pseudoknots or base triples in the secondary structure lists. A subroutine checks the indices of two successive base pairs, whether the conditions for a pseudoknot or a base triple apply and, if necessary, corrects the entries of the secondary structure lists. A pseudoknot will be cut out entirely and must be treated separately, whereas base triples are separated into a base pair b_{ij} satisfying condition $j - i > 3$ and a single base. Tertiary interactions between two hairpin loops as seen in tRNA are excluded from the secondary structure evaluation as well. The Secondary Structure Classification routine identifies such loop-to-loop interactions, if a base pair b_{ij} is surrounded by single bases which constitute a loop turn on either side of the pair.
It must be stressed that the justification for excluding pseudoknots, base triples and all tertiary contacts from the secondary structure classification is that they would interfere with the loop decomposition pattern. However, these substructures are a determining factor of the 3D fold and should find recognition in the future improvement of the algorithm (see 4).
After the proof-reading has been accomplished, the secondary structure lists are read in by the loop decomposition routine which scans them for helical regions. A helix is defined by a continuous succession of base pairs, or in more technical terms, successive loops of order 2 and size zero. Tracing the

[9] In this routine, modified bases are treated as the bases they are derived from

secondary structure lists, the routine sorts out four base indices i, i', j' and j that demarcate each helix:
$i < i' < j' < j;\ i' = i + x;\ j = j' + x,$
with $x + 1$ as the helix length. The four indices belong to the base pairs b_{ij} and $b_{i'j'}$ that are separated by $x - 1$ base pairs. This way, all helices of the structure are listed compactly and in consecutive order regarding their indices. Let i_m, i'_m, j_m and j'_m be the start and stop indices of the m^{th} helix. The relation of the base indices between helix m and the next helix $m + 1$ is $i_m < i'_m < i_{m+1} < i'_{m+1} < j'_{m+1} < j_{m+1} < j'_m < j_m$.



The algorithm allocates stems and single-based regions to certain categories of secondary structure elements. The different loop structures between helix m and helix $m + 1$ are identified as a

- a bulge between i'_m and i_{m+1}, if $i'_m + 1 \neq i_{m+1} \ \bigwedge\ j'_m - 1 = j_{m+1}$
- a bulge between j_{m+1} and j'_m, if $j'_m - 1 \neq j_{m+1} \ \bigwedge\ i'_m + 1 = i_{m+1}$
- a hairpin between i'_m and j'_m, if
  a.) both $i'_m + 1 \neq i_{m+1} \ \bigwedge\ j'_m - 1 \ \neq \ j_{m+1}$,
  b.) and all bases between index i'_m and index j'_m are single bases.
- an internal loop between i'_m and i_{m+1} on the one strand and j_{m+1} and j'_m on the other, if
  a.) both $i'_m + 1 \neq i_{m+1} \ \bigwedge\ j'_m - 1 \ \neq \ j_{m+1}$
  b.) and no other base pairs exists between the index positions j_{m+1} and j'_m.
- a multi-loop between i'_m and i_{m+1} on the one strand and j_{m+1} and j'_m on the other, if
  a.) both $i'_m + 1 \neq i_{m+1} \ \bigwedge\ j'_m - 1 \ \neq \ j_{m+1}$
  b.) and there is at least one base pair between the index positions j_{m+1} and j'_m. In other words, base pair $b_{ij,m+1}$ is not the only interior base pair accessible from the closing base pair $b_{ij,m}$.

The last condition for multi-branched loops comprises a range of possible substructures that may be located between the indices j_{m+1} and j'_m.

Sequence-To-Secondary Structure Alignments

The loop decomposition routine automatically generates a secondary structure file which contains the information about the stem numbers, helix indices and base pairs of a given 3D template structure. This file serves as input data for the alignment package RAGA.

For each tRNA query sequence, a secondary structure alignment using RAGA was conducted. The template of the highest scoring alignment was selected as being the most promising scaffold structure for the query sequence.

Regions of the query sequence that had remained unaligned in this first RAGA alignment with the tRNA templates had to be aligned seperately. According to the algorithmic network outlined in Figure 1, such bases constitute the non-common regions of query and template sequence. These gaps vary in length and may contain either only single bases or comprise both single bases and base pairs.

A second sequence-to-secondary structure alignment was conducted to extract structural information from the gap regions. The submitted regions covered five additional bases at the ends of each gap. This ensures that the structural information enclosed in the fringe bases could not get lost in case that the RAGA algorithm had opened the gap in between a secondary structure element.

Again, RAGA was employed for the alignment of the gap regions, but this time resorting to the structures in the loop motif database. The second RAGA run is split up into two phases:

- The gap region was treated like a new, shorter query sequence. The sequence stretch was aligned to the template structures of the loop database.

- Recall the algorithmic network (Figure 1) and the fact that RAGA processes a sequence and a secondary structure file. Dealing with short loop sequences, the alignment process can be reversed; if the gap region of the query sequence comprises a secondary structure element, it can be used as a template structure while the sequences of the loop motif database are treated as query sequences.

The secondary structure of short sequences can reliably be predicted with the program *RNAfold*. The structural information of the prediction was read into RAGA for the second part of the gap region alignment. Therefore the two RAGA runs for the gap regions yielded two alignments for each structure in the loop database. The alignment score differs between these two forms of alignments because the structure which is considered as the template contributes a partial secondary structure score to the the overall

aligment score. A suitable template structure for a specific gap region scores well in both RAGA runs.

3 Results

The application of the Base Pair Detection, Base Pair Classification and Loop Decomposition routines result in an improved RNA backbone visualization (hier fehlt noch der Verweis auf ein Bild einer Initialstruktur).
The following assembly routine creates an initial 3D structure embracing all atom positions for each of the query sequences.
The secondary structure alignments unambiguously associates most or even all bases of the query sequence with the bases of either the tRNA template or the loop motif template, if necessary. Apart from this one-to-one assignment, the RAGA alignments also provide information about common base pair positions between query and template structure. The coordinates for the aligned residues are derived from the template structure; copying residues whenever possible and approximating their positions, if the alignment position does not present a direct match.
When all atom coordinates have been assigned to the residues of the query sequence, the phosphodiester bonds are introduced. A subroutine scans residue b_1 to residue $b_N - 1$ of the proposed 3D structure and connects the 3'OH of each residues with the 5' P of the following residue through a single bond. Some of the phosphodiester bonds will remain deformed due to inaccuracies in the mapping of satellite to leader bases or at the interfaces of non-common loop regions. Correcting these bond lengths and angles is one of the challenges for the refinement routine. The residues of a specific query sequence divide up into five categories:

1. Base Pairs Conserved in both Position and Sequence. Both query and template sequence have the same base pair in the same alignment position. The threading copies the two base pair partners from the template and places them at the same coordinates which the template residues would occupy according to their PDB file.

2. Base Pairs Conserved only in Position. The base types differ between query sequence and template structure so that the base pair cannot simply be copied as before. Instead, a model base pair has to be loaded from a separate base pair file providing the atoms for the base pair of the query sequence. For a better understanding the template base pair at a given alignment position will henceforth be referred to as the *leader* base pair, whereas the base pair of the query sequence will be called the *satellite* base pair. The satellite base pair is mapped onto the corresponding template base pair. The file providing the satellite base pair can be generated from any RNA data file containing all possible canonical base pairs and a variety of

noncanonical pairs[10]. The atoms of the satellite base pair have to be assigned the correct coordinates in the new 3D model because the two residues are positioned in space relative to the origin of the coordinate system of the base pair file. In order to translate the satellite base pair to the correct position in the coordinate system of the threading model, a refererence point is needed. This point defines which atoms of the satellite base pair and the leader base pair should be mapped onto each other after the translation. As outlined in section 2, the center of reference is defined in the centroid of the hexagon of each satellite pyrimidine base b_{sat}:

$$\vec{c}_{sat} = 1/3 \left(\vec{N}_{1,sat} + \vec{C}_{4,sat} + \vec{C}_{5,sat} \right)$$

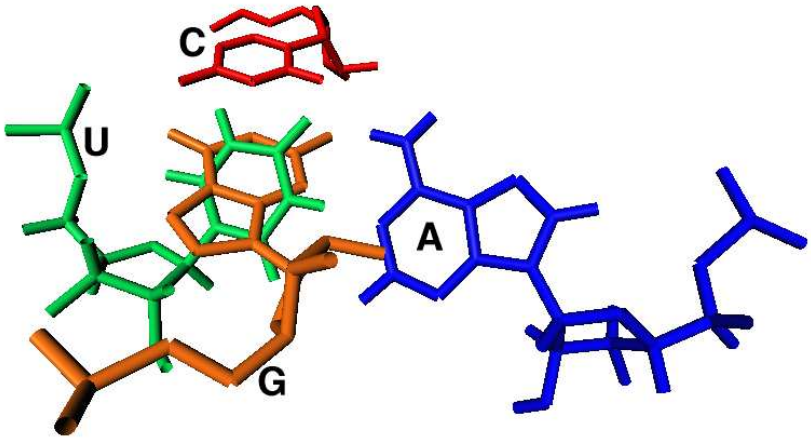


Figure 4: First step of mapping a satellite A–U base pair to its leader G–C base pair. The satellite cytosine is translated, so that its centroid is aligned with the centroid of the corresponding leader adenine. Color code: leader base pair: C/red, G/orange; satellite base pair: U/green, A/blue.

Analogously, the centroid $\vec{c}_{leader}$ of the leader base is computed in the same way. The leader base can either be a pyrimidine or a purine, depending on the RAGA alignment. There is no compelling reason for choosing the satellite pyrimidine base instead of the satellite purine base as the first reference point. The satellite base pair is both translated and rotated so that the hexagons of satellite and leader bases become as congruent as possible. Since the relative orientation of the two planes of the base pair partners is the decisive aspect in the formation of hydrogen bonds, the hexagons of the two base pairs are mapped onto each other in order to maintain the base

[10]The threading module accesses the PDB data of the *E.coli* 5S rRNA E-loop (1A51) for the satellite base pair atom coordinates

plane orientation of the leader base pair in the satellite base pair.
The translation vector $\vec{t}$ is

$$\vec{t} = \vec{c}_{leader} - \vec{c}_{sat}.$$

Adding this vector to all atoms of the two satellite residues assigns them to the new position with the satellite pyrimidine centroid aligned onto the corresponding leader centroid (Figure 4).
In a second step, the entire satellite base pair has to be rotated to approximate the orientation of the leader base pair. It should be pointed out that the following rotations do not equalize the difference in propeller twist between leader and satellite base pair. On the contrary, the satellite base pair should be seen as an entity whose propeller twist must be preserved. The rotations of the satellite base pair rather serve to approximate the position of the leader base pair so that the helix curvature is preserved.
The pyrimidine base plane of the satellite base pair is adjusted firstly. In the second step, the deviation between the purine base plane and its corresponding leader base is successively decreased until the constellation of the least possible deviation on both sides of the base pair is achieved.
In practice, the angle α between the base planes of the satellite pyrimidine base and its leader base is computed. The rotation center of all following rotations is situated in the overlapping centroids of the satellite pyrimidine and its leader base $\vec{c}_{leader,sat} = \vec{c}_{leader} = \vec{c}_{sat}$. The whole satellite base pair is rotated α degrees around an axis $\vec{r}$ defined by the vector product of the two normals $\vec{n}_{leader}$ and $\vec{n}_{sat}$.

$$\vec{r} = \vec{n}_{leader} \times \vec{n}_{sat}.$$

After this rotation, the satellite pyrimidine base plane and the corresponding leader base plane have the same normals. Still, there is a difference on the other side of the two base pairs, namely between the base planes of the satellite purine base and its corresponding leader base. Thus, a similar rotation also adjusted the base planes on this side (Figure 5).

However, it is not possible to exactly align both centroids of the satellite base pair to both centroids of the leader base pair while bringing the two sets of normals in line at the same time. Though canonical pairs are isosterical, each pair type has a slightly different propeller twist. Therefore, the orientation of the leader base pair is approximated by iteratively adjusting the normals of the satellite purine base and its leader base. The angle between the two normals is reduced during the rotation by one degree as long as this procedure improves the overall situation. Since the entire satellite base pair is rotated without altering its propeller twist, the decreasing difference between the normals on the purine side of the base pair causes the difference between the normals on the pyrimidine

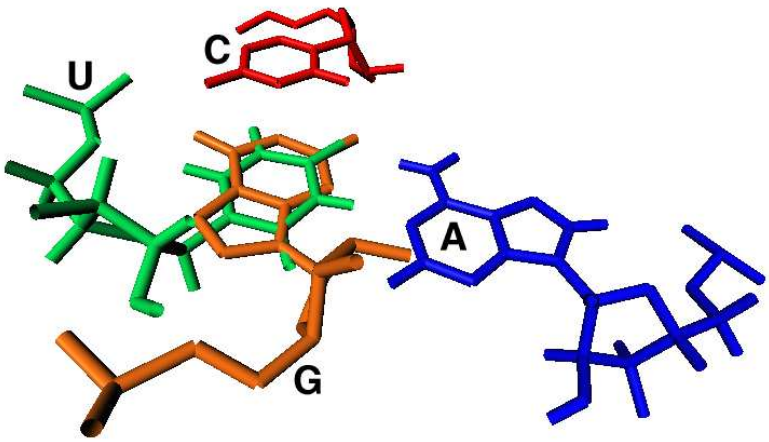


Figure 5: After the normals of the base planes have been adjusted, the hexagons of both leader and satellite base pair approximately lie in the same plane. Color code as in Figure 4.

side to increase again. As soon as the two differences exceeds the point of equality, the iteration is stopped. This way, an approximation of the least deviation on both sides of the base pair is achieved. The method successfully superimposes the planes of satellite and leader base pair, if the difference in propeller twist between satellite and leader is small. Problems that may occur when noncanonical base pairs are involved, are discussed in section 4.

In the next step the strand orientation is checked as described in section 2. However, this time the task is to transfer the strand orientation of the leader base pair to the satellite base pair. Therefore, the orientations of the two 2'-OH oxygens of the satellite pyrimidine base on the one hand and of the corresponding leader base on the other hand are compared. If they haved different orientations, the entire satellite base pair is flipped by 180 degrees around an axis connecting the C_1^* atoms to adjust the strand orientation between satellite and leader base pair.

If the satellite pyrimidine base is aligned with a purine leader base, the ribose part of the satellite base pair will be shifted a few degrees sideways. Due to this effect, the phosphodiester bond length are incorrect afterwards. Before correcting the phosphodiester bonds, a further rotation adjusts the distance between the centroids of the satellite purine base and its leader base. The triangular constellation between

- the superimposed centroids of the satellite pyrimidine base and its leader base,

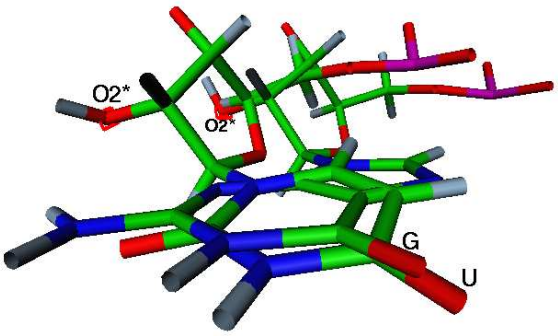


Figure 6: To preserve the strand orientation between satellite and leader base pair, the two 2'OH oxygens must be situated on the same side of the plane represented by the base hexagons.

- the centroid of the satellite purine base and
- the centroid of the leader base associated with the satellite purine base

is illustrated in Figure 5. The centroids of the satellite purine base and its leader base are brought together in a rotation so that the angle between the three centroids disappeares (Figure 7). The last rotation corrects the majority of the deformed phosphodiester bond lengths. The glycosyl bond is set free for stretching and bending, whereas the base rings are frozen in their current position. Then each of the two satellite ribose parts including the phosphor atoms is rotated so that the satellite riboses are separately mapped onto the leader riboses (Figure 8). The rotation adjusts the normals of the "ribose planes" described by the plane between the atoms C_1^*, C_2^*, and O_4^*.
The improvement of the backbone trace comes at the price of inaccurate glycosyl bond angles and lengths. This problem is solved by the refinement routine.

3. Aligned Single Bases. This category refers to single bases that are aligned either with the tRNA template (common loop regions) or with the loop template (non-common loop regions). Similar to the preceeding category, each single base is copied from the base pair file and then translated in the direction of its leader base, so that the two centroids are mapped onto each other. Subsequently, the twist between the base planes of satellite and leader base is balanced out and the strand orientation is adjusted as described above. A a result, leader and satellite base planes share the same centroid and the same normals. The satellite base is then rotated around the centroid so that the distance between the two C_1^* atoms described by

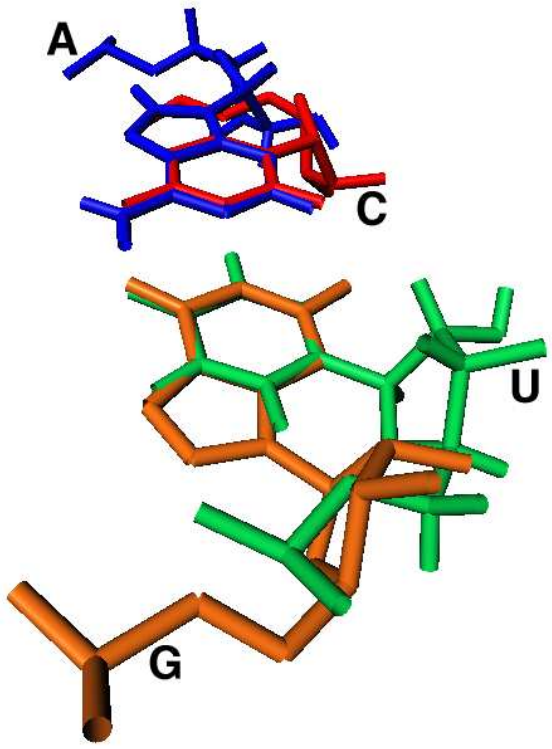


Figure 7: The satellite base pair is rotated so that the hexagons on either side of leader and satellite base pair become superimposed.

the angle between the centroid and the two C_1^* atoms dissolves.

4. Bases Aligned to a Loop Template. There is no generally applicable algorithm how to fit a separately aligned loop motif into the entire 3D model. After all other base pairs and single bases have been placed in their correct positions, a gap remains in the new structure, indicating the vacant space where the loop region should be positioned. Using the amiraMol® environment, the complete loop must be interactively fit into the entire 3D model structure. The distance between the variable loop and the two residues flanking the loop in the main model structure was adjusted so that they could form phophosphodiester bonds with the first and the last loop residue.

5. Unaligned Single Bases. The fifth and last category considers unaligned single bases that did not find a leader base at all, neither in the alignments with the tRNA templates, nor in the gap region alignments. Fortunately this "worst case scenario" concerns only a few bases at the fringes of gaps inserted in the query sequence during the alignment with the tRNA template.

For the time being, there is no algorithmic principle how to model such un-

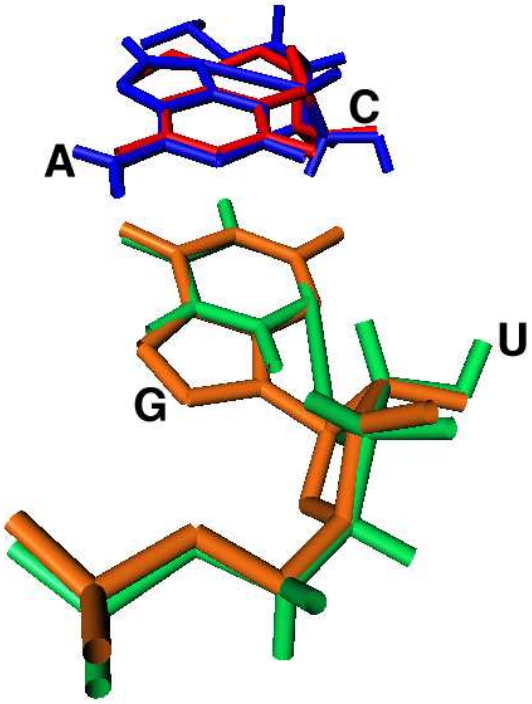


Figure 8: In the last step the ribose parts of the bases are re-arranged separately in order to maintain the helix curvature. This improvement is bought by stretch-bending the glycosyl bonds.

aligned residues. Nevertheless, the threading module opens up interactive modeling possibilities to cope with such problematic residues and finding a distinctive solution in each case. For example, single residues representing an insertion in the query sequence can be placed in between the two residues of the template structure which are flanking the insertion site. Furthermore, longer regions of the query sequence that do not find any leader bases in the best scoring alignment can often be replaced using a second alignment with a minor scoring template.

Refinement by Relaxation An MD relaxation including water molecules and ions was performed on the initial structures.

Starting off from each of the initial structures, the conserved base pair positions were preserved while single-based loop regions and all backbone phosphodiester bonds were allowed to arrange freely in space. A complete rearrangement of the initial structure is neither possible nor desirable. The maintenance of structurally essential features of the template in the threading model implies that the base pairs of the initial structure must not be destroyed during the MD relaxation. Therefore, each base pair, excluding

the ribose parts, was treated as an entity that obeys rigid body dynamics. The parameters for the MD run can be gathered from the Appendix (section 6).

Template Selection The template database reflects a comprehensive selection of currently available, complete tRNA 3D structures. The database was designed to support the development of a prototype algorithm. tRNAs represent both functionally and structurally well-investigated molecules standing out for their suitability as template molecules. A general criterion is that they adopt a key position in the cellular machinery throughout all organisms. Even more decisive is that they incorporate a variety of important RNA loop structures (several hairpin motifs centered around a multi-loop) without being infeasibly complex. Being a medium-sized RNA species of comparable length (70-90 nucleotides) they show interestingly diverse primary sequences and a remarkably high level of structural conservation. These facts underline a close structure-to-function relationship which is the foundation of threading approaches. In principle, the proposed method is applicable to other RNA systems as well.

The amount of time required for the entire threading process strongly depends on the individual query sequence. In cases that do not require interactive/separate modeling of loop regions, the computational effort for the secondary structure alignments and the relaxation (minutes to several hours) exceeds the time spent on the assembly (seconds) described in 3 by far. Interactive modeling may take only minutes if just a single unaligned residue occurs, but it may also extend to several hours if a suitable loop template must be selected from the loop database.

Query Sequence Selection The query sequences have been selected deliberately to test the functionality of the RNA threading algorithm. Four entire tRNA query sequences have been modeled with the prototype module (see below). Two of these models, namely *Saccharomyces cerevisae* $\mathrm{tRNA}^{\mathrm{ser}}$ and *Thermus thermophilus* $\mathrm{tRNA}^{\mathrm{pro}}$ can be compared to 3D structures deposited in the PDB. In contrast to that the models of *Prochlorococcus spec.* $\mathrm{tRNA}^{\mathrm{glu}}$ and *Prochlorococcus spec.* $\mathrm{tRNA}^{\mathrm{pro}}$ can only be compared to their template structures because they are based on recently elucidated sequence data[11].

Moreover, the ability of the algorithm to identify common loop motifs has been tested on an HIV-1 purine-rich hairpin loop sequence bearing structural similarity to the tRNA anticodon loop.

Evaluation of the Models The quality of the final models cannot be measured by an energetical criterion because the MD simulation does not serve the purpose of investigating the thermodynamics of the model, but rather to ease the tensions caused by the remaining deformed bonds. This implies that the energy of the refined structure does not represent a resalistic

[11] with permission of Hanspeter Herzel, Institute fot Theoretical Biology, HU Berlin

value. It is an indicator energy which decreases significantly if the simulation has successfully relaxed the unnatural bond length and angles of the initial structure.

Instead of resorting to an energic criterion, the *Root-Mean-Square Deviation (RMSD)* is used to quantify the difference between the 3D models and their reference structures disposed in the PDB. To compare the maintenance of the backbone curvature, rather than the overlap of all atoms in the base rings, the RMSD is calculated on the basis of the backbone atoms ($P^*, O5^*, O3^*, C5^*, C3^*, C4^*$). The RMSD and PDB access codes for all query sequences are listed in table 1 on page 27.

HIV-1 A-Rich Hairpin Loop Model. The reference structure has PDB access code 1BVJ. A hairpin loop from template structure *Haloarcula marismortui* 23S rRNA gained the highest alignment score with the HIV-1 purine-rich hairpin because it contains the same platform motif according to the SCOR classification. The only difference is that the canonical U2–A9 closing base pair in 1BVJ is a G–C pair in the template and the adenine at position A7 has been replaced by a cytosine in the template. All essential features of the reference structure are preserved in the model of 1BVJ. The RMSD between model and reference structure is only slightly higher than the RMSD between template and reference structure.

Prochlorococcus spec. tRNA Models. The 3D models of *Prochlorococcus spec.* tRNA^{glu} and *Prochlorococcus spec.* tRNA^{pro} cannot be compared to reference structures because there is no PDB data for these structures available so far. Figures 9 and 10 show the refined models and their template structures (1G59 and 1H4Q, respectively) both in a wireframe and backbone representation. Note that the acceptor stem of the *Prochlorococcus spec.* tRNA^{pro} model has been added using the minor scoring alignment with *E. coli* tRNA^{asp} (2TRA).

Saccharomyces cerevisae tRNA^{ser} Model. The model of *Saccharomyces cerevisae* tRNA^{ser} is particularly difficult to evaluate because the reference structure 5TRA is a theoretical model as well and has not been verified yet. The template, *Saccharomyces cerevisae* tRNA^{arg} (1F7U), lacks the variable loop almost completely. Figures 11 show that the shape of the model bears a greater similarity to the reference structure than to the template. The high RMSD values between model and reference structure can be attributed to backbone trace deviations in the acceptor stem, anticodon and variable loop regions[12]. These single-based parts of the structures can adopt variable orientations. Therefore, the RMSD, when measured between model and reference structure, is calculated both including and excluding the single-based regions. The variable loop region has been modeled using a second loop template, namely a hairpin loop from *T.thermophilus* 16S

[12] variable loop region: res. 44-56, anticodon loop region: res. 30-40, 3' end of acceptor stem: res. 82-85

rRNA (1J5E). Thus, the RMSD between the model and the reference structure 5TRA includes the variable loop region, whereas it is excluded from the RMSD calculation between the template 1F7U and either the model or the reference structure.
Obviously, the RMSD increases significantly if the 3D model is composed of more than a single template structure.
Thermus thermophilus $tRNA^{pro}$ Model.
The relatively high RMSD between model and reference structure can be explained by the different orientations of the anticodon bases in the template structure *E. coli* $tRNA^{f\text{-}met}$ (2FMT) on the one hand and in the reference structure (1HQ4) on the other hand. Since the query sequence has been mapped onto the template, the anticodon structure is copied to the 3D model which accounts for the RMSD between the backbones of the model and its reference structure. Excluding the anticodon region from the calculation yields a considerably improved RMSD.

4 Problem Discussion

Databases. Similar to the tRNA template database, a separate template database for each RNA type should be established. The categories could be chosen according to the RNAbase classification.
Moreover, the loop template database which comprises only tri- and tetraloops so far, must be extended to all other exterior and interior loop motifs according to the SCOR database.
Base pair Classification. During the Base Pair Classification routine, two adjacent edges of the base in question sometimes gain the same score due to the ambiguity of some of the sites. In this case, a decision either between Watson-Crick and Sugar Edge, or Watson-Crick and Hoogsteen Edge must be made; ambiguity between Hoogsteen and Sugar Edge does not occur because these two edges are separated by the Hoogsteen Edge. The routine prefers the Watson-Crick Edge as default.
Problems occur if chemically modified bases lack one or several of the described donor/acceptor sites because the currently implemented routine treats them as normal bases so far. In the further development of the module a comprehensive description of the hydrogen bonding sites of modified bases should be added.
Secondary Structure Classification. The template folds are constrained to structures comprising only a single multi-loop so far. The algorithm detects a single multi-loop by recognizing several interior base pairs following a loop closing base pair and expects to find only further stem loops (but not another multi-loop) in the succession of each interior base pair. It fails to allocate the correct loop category if a multi-loop branches out into another multi-loop. A recursive algorithm focussed on

the detection of recurring structural motifs, could substantially improve the classification routine.

Tertiary interactions. Tertiary interactions are identified and sorted out by the base pair detection routine, but the information could also be used as a constraint in the subsequent modeling process. A comprehensive RNA-profile comprising structural information on all three organization levels could be processed either in the secondary structure alignment (see the next point: *Sequence-to-Secondary Structure Alignment*), or in the refinement routine (see *Refinement by Relaxation*).

Asssembly of the Initial Structures. It should be mentioned that problems occur if the difference of the propeller twist is large between leader and satellite base pair, e.g. if a noncanonical satellite base pair with a large propeller twist is aligned on a regular canonical base pair or vice versa. Then the plane normals of the satellite base pair cannot be completely adjusted to the leader base pair under the constraint that the relative orientation of the satellite base planes is preserved. In this case, a noticable difference in the propeller twist between leader and satellite base pair will remain.

A similar problem occurs at all base pair positions, where a canonical base pair in the template has been replaced by a noncanonical base pair in the query sequence. There is no generally applicable rule for which geometry class the noncanonical base pair belongs to according to the classification scheme [18]. The geometry of noncanonical base pairs cannot be predicted automatically yet.

Refinement Procedures. Primarily, Molecular Dynamics (MD) simulations are employed to describe the dynamics and internal energies of real molecules. However, the MD applied to the RNA initial models is used as a tool for the relaxation of unnatural bond length and angles, thereby adjusting them to their equilibrium values.

5 Conclusion

Achieved Objectives. The RNA Threading algorithm described in this article produces full-atom 3D model structures that can be compared to the experimentally verified structures deposited in the PDB. Algorithms have been developed that decompose 3D template structures into their secondary structure elements. This information has been deposited in two specific template databases. An algorithmic network has been established that integrates public algorithms for a sequence-to-structure alignment (RAGA) and secondary structure prediction (Vienna package). The resulting initial 3D models have been refined with a MD relaxation technique developed at the ZIB for Conformation Dynamics. The RNA 3D model structures provide

a good starting point for further refinement processes because they do not only retrace the backbone of the template structure, but represent complete new structures at the atomic resolution level.

6 Appendix

6.1 MD Relaxation Parameters

The RNA structure was placed within a virtual box of water molecules and sodium ions compensating for the negative charges of the backbone phosphor atoms. Surrounding the box, periodic boundary conditions are assumed. The electrostatic interactions were calculated according to the reaction field approach proposed by I. G. Tironi, R. Sperb, P. E. Smith, and W. F. van Gunsteren [28]. The integrator performs $t_n = 20,000 - 30,000$ iterations at a resolution of $\Delta t = 0.001 \cdot 10^{-12} s$. During the equilibration phase the model structure is frozen in its position, while the water molecules and the ions are set free to arrange themselves around the RNA molecule. The resulting model is the state with the lowest energy that is reached within this time span.
The time step must be short because the initial structures comprise bond deformations that result in large forces $\vec{F}(\vec{q})$ acting on the point masses. The momenta are drawn from a Maxwell distribution of velocities at 1000K. Every 100 integration steps new momenta are chosen to prevent the system from heating up or cooling down.
The high temperature should facilitate conformational changes over high energy barriers.
For all MD relaxations, a cut-off for the non-bonded interactions in the MMFF94 force field is set to 12Å. Moreover, all bases, excluding the ribose part, and base pairs are handled as rigid bodies. Therefore, the integrator scheme had to be adjusted to deal with base pairs obeying rigid body dynamics according to [29].

Acknowledgement The work of F. Cordes has been supported by the German Federal Ministry of education and research (grant no. 031U109A/031U209A, Berlin Center for Genome Based Bioinformatics). All visualizations in this work have been rendered using amiraMol® a visualization software package being developed at the Zuse Institut Berlin (ZIB). The authors would like to thank D. Baum, J. Schmidt-Ehrenberg and H. Herzel for helpful discussions.

References

[1] G. Storz: *an expanding universe of noncoding RNAs*, Science (2001), 296, 1260–1263

[2] S. R. Eddy: *Non-coding RNA genes and the modern RNA world*, Nature Review Genetics (2001), 2, 919–929

[3] R. H. Symons: *Small Catalytic RNAs*, Annual Review of Biochemistry (1992), 61, 641–671

[4] Lai EC: *microRNAs: runts of the genome assert themselves*, Current Biology (2003), 13, R925–36

[5] P. Burgstaller, A. Jenne, M. Blind: *Aptamers and aptazymes: accelerating small molecule drug discovery*, Curr. Opin. Drug Discov. Devel. (2002), 5, 690–700

[6] G. J. Hannon: *RNA interference*, Nature (2002), 418, 244–251

[7] A. Thakur: *RNA interference Revolution*, Electronic J. of Biotech. (2003), 6,39–49

[8] D. Shortle: *Prediction of protein structure*, Current Biology (2000), 10, R49–R51

[9] D. Shortle: *Structure prediction: The state of the art*, Current Biology (1999), 9, R205–R209

[10] A. Sali, E. Shakhnovich and M. Karplus: *How does a protein fold?*, Nature (1994), 369:248–251

[11] B. Rost, R. Schneider, C. Sander: *Protein Fold Recognition by Prediction-based Threading*, J. Mol. Biol. (1997), 270, 471–480

[12] N. Alexandrov, R.Nussinov, R.M. Zimmer: *Fast protein fold recognition via sequence to structure alignment and contact capacity potentials*, Biocomputing: Proceedings of the 1996 Pacific Symposium, edited by Lawrence Hunter and Teri Klein, World Scientific Publishing Co., Singapore (1996)

[13] I. Tinoco Jr., C. Bustamante: *How RNA Folds*, J. Mol. Biol. (1999), 293, 271–281

[14] P. Brion, E. Westhof: *Hierarchy and Dynamics of RNA folding*, review article in Annual Review of Biophysics and Biomolecular Structure (1997), 26, 113–137

[15] S. A. Woodson: *Recent insights on RNA folding mechanisms from catalytic RNA*, Cell. Mol. Life Sci. (2000), 57, 796–808

[16] M. Zuker: *Prediction of RNA Secondary Structure by Energy Minimization*, in Computer Analysis of Sequence Data (1994), A.M. Griffin and H.G. Griffin eds., Methods in Molecular Biology, Humana Press Inc., 267–294

[17] J. Schmidt-Ehrenberg, D. Baum, H.-C.Hege: *Visualizing Dynamic Molecular Conformations*, R. J. Moorhead, M. Gross, K. I. Joy (eds.), Proceedings of IEEE Visualization 2002, Boston MA, USA, IEEE Computer Society, IEEE Computer Society Press Oct./Nov. (2002), 235–242

[18] N. B. Leontis, E. Westhof: *Geometric nomenclature and classification of RNA base pairs*, RNA (2001), 7, 499–512

[19] Ke Shi, M. Wahl, M. Sundaralingam: *Crystal structure of an RNA duplex r(GGGCGCUCC)2 with non-adjacent GU base pairs*, Nucleic Acids Research (1999), 27, 2196–2201

[20] L. Stryer: *Biochemistry*, 4th Edition (1999), chapter 1, 7

[21] M. Zuker, D. H. Matthews, D. H. Turner: *Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide*, in RNA Biochemistry and Biotechnology (1999), J. Barciszewski & B.F.C. Clark eds., NATO ASI Series, Kluwer, Academic Publishers, Dordrecht, NL, 11–43

[22] D. Sankoff, J. B. Kruksal, S, Mainville, R. J. Cedergren: *Fast algorithms to determine RNA secondary structures containing multiple loops*, in Time Wars, string edits and macromolecules: the theory and practice of sequence comparison (1983), D. Sankoff, J. B. Kruksal eds., Addisaon-Wesley, Reading, MA, chapter 3, 93–120

[23] C. Notredame, E. A. O'Brien, D. G. Higgins: *RAGA: RNA sequence alignment by genetic algorithm*, Nucleic Acids Research (1997), 25, 4570–4580

[24] C. Notredame, D. G. Higgins: *SAGA: Sequence alignment by genetic algorithm*, Nucleic Acids Research (1996), 24, 1515–1524

[25] C. Notredame: *Utilisation des algorithmes genetiques pour l'analyse de sequences biologiques*, Dissertation (1998), University Paul Sabatier, France

[26] D. E. Goldberg: *Genetic algorithms in search, optimization and machine learning*, Addison-Wesley (1989), New York

[27] T. A. Halgren: *Merck Molecular Force Field. I. Basis, Form, Scope, Parametrization, and Performance of MMFF94*, J. Comp. Chem. (1996), 17, 490–519

[28] I. G. Tironi, R. Sperb, P. E. Smith, W. F. van Gunsteren: *A generalized reacton field method for molecular dynamics simulations*, J. Chem. Phys. (1995), 102, 5441–5459

[29] A. Dullweber, B. Leimkuhler, R. McLachlan: *Symplectic splitting methods for rigid body molecular dynamics*, J. Chem. Phys. (1997), 107,5840–5851

Query: HIV-1 A-rich hairpin Template 1EIY, res. 29–40 Reference 1BVJ, res. 7-18		Query: HIV-1 A-rich hairpin Template 1BN0, res. 7-18 Reference 1BVJ, res. 5-16	
a.)	0.43	a.)	1.18
b.)	0.61	b.)	1.17
c.)	1.86	c.)	3.23
d.)	1.80	d.)	3.10
e.)	1.90	e.)	3.00
Query: HIV-1 A-rich hairpin Template 1NEM, res. 4–20 Reference 1BVJ, res. 4-20		Query: HIV-1 A-rich hairpin Template 1JJ2, res. 1195–1206 Reference 1BVJ, res. 7-18	
a.)	0.70	a.)	0.83
b.)	1.01	b.)	0.85
c.)	4.50	c.)	2.18
d.)	4.70	d.)	2.13
e.)	4.65	e.)	1.95
Query: *P. spec.* $tRNA^{glu}$ Template: 1G59, res. 1-72 No reference structure available		Query: *P. spec.* $tRNA^{pro}$ Main Template: 1H4Q, res. 1-67 No reference structure available	
a.)	0.43	a.)	0.65*
b.)	0.51	b.)	0.80*
$tRNA^{ser}$ Main Template 1F7U, res. 1-76 Reference 5TRA, res. 1-85		Query: *T. thermophilus* $tRNA^{pro}$ Template 2FMT, res. 1-77 Reference 1H4Q, res. 1-67	
a.)	2.10	a.)	0.74
b.)	2.16	b.)	2.23
c.)	8.16 (2.54)**	c.)	2.55 (1.82)***
d.)	7.80 (2.44)**	d.)	2.64 (1.94)***
e.)	2.17	e.)	2.57

Table 1: Root-Mean-Square Deviations (RMSD) between 3D models, templates and reference structures.
a.) template vs. initial structure
b.) template vs. refined structure
c.) reference vs. initial structure
d.) reference vs. refined structure
e.) template vs. reference structure
(*) excluding acceptor stem; res. 1-3 and 71-74
(**) excluding variable loop; res. 44-56, anticodon loop: res. 30-40 and 3' end of the acceptor stem; res. 82-85
(***) excluding anticodon loop: res. 31-35

Figure 9: 3D model of *Prochlorococcus spec.* $tRNA^{glu}$ based on template structure 1G59. Note that there is no reference structure for this model in the PDB.
Upper left: The template structure; *Thermus thermophilus* $tRNA^{glu}$ (1G59).
Upper right: Backbone representation of 1G59.
Lower left: The refined 3D model of *Prochlorococcus spec.* $tRNA^{glu}$.
Lower right: Backbone representation of the model.
Color code lefts: A/violet, C/blue, G/green, U/red.
Color code rights: dangling/violet, hairpin/blue, helical/green, multi-loop/yellow.

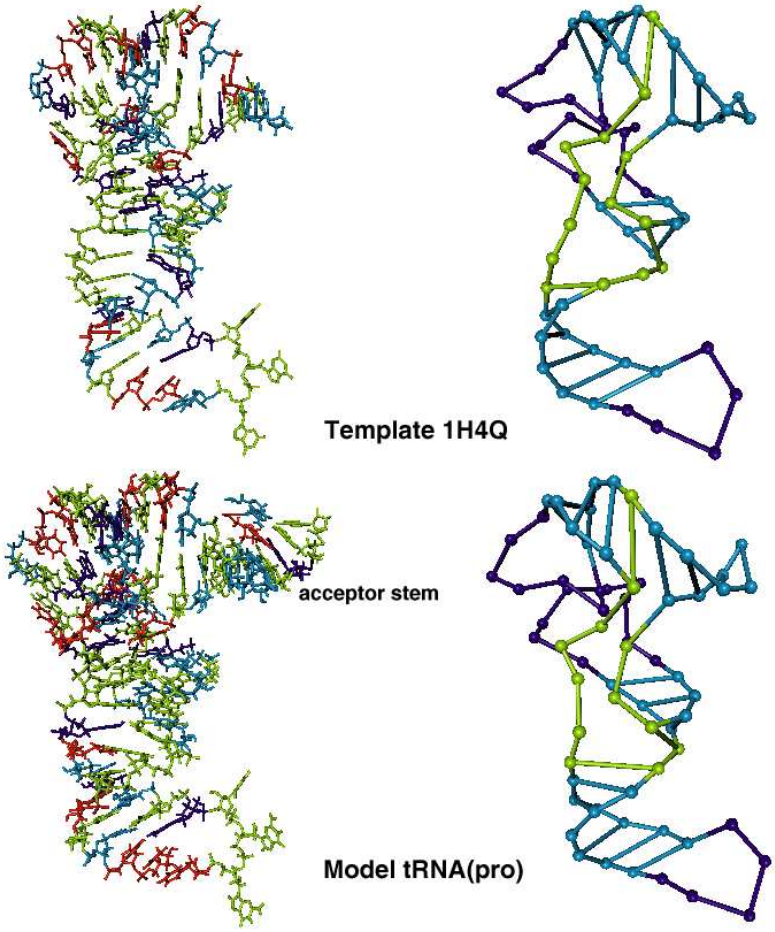


Figure 10: 3D model of *Prochlorococcus spec.* $tRNA^{pro}$ based on template structure 1H4Q. Note that there is no reference structure for this model in the PDB.
Upper left: The template structure; *Thermus thermophilus* $tRNA^{pro}$ (1H4Q).
Upper right: Backbone representation of 1H4Q.
Lower left: The refined 3D model of *Prochlorococcus spec.* $tRNA^{pro}$. Note that the acceptor stem is taken from the model based on the minor scoring template *E. coli* $tRNA^{asp}$ (2TRA). Therefore, it does not appear in the backbone representation.
Lower right: Backbone representation of the model.
Color code lefts: A/violet, C/blue, G/green, U/red.
Color code rights: hairpin/violet, helical/blue, multi-loop/green.

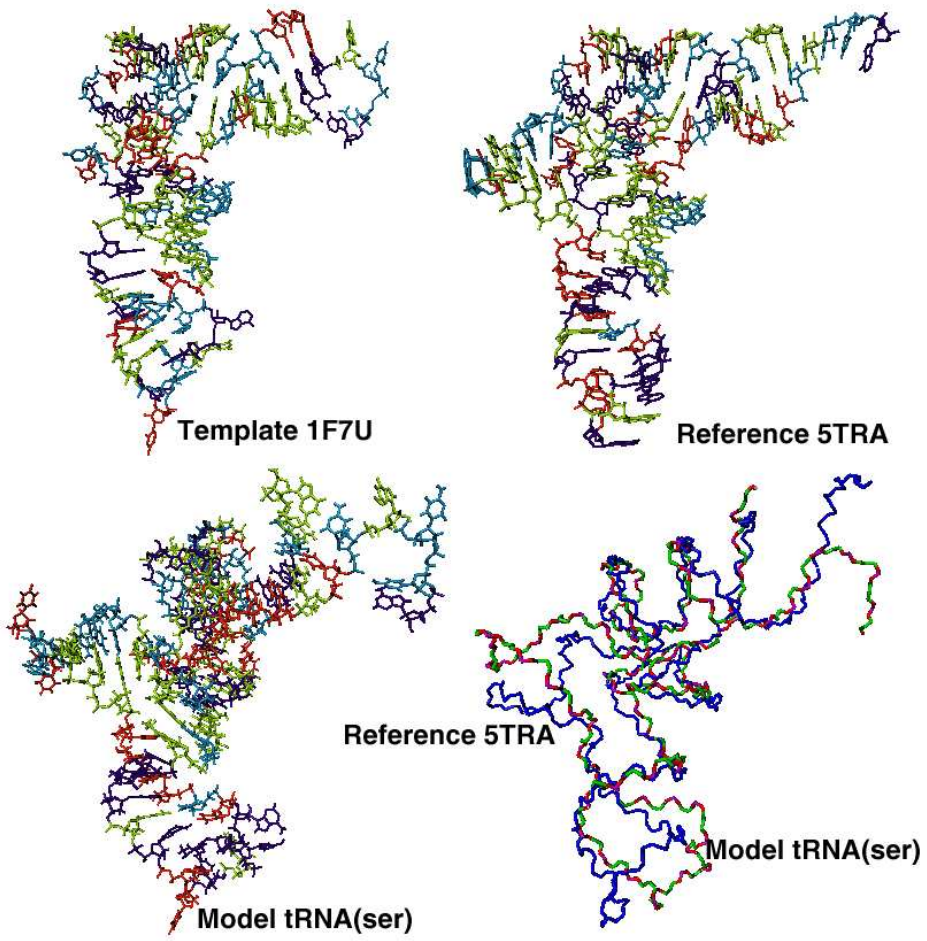


Figure 11: 3D model of *Saccharomyces cerevisae* $tRNA^{ser}$ based on the template structures 1F7U and 1J5E.
Upper left: The main template structure; *Saccharomyces cerevisae* $tRNA^{asp}$ (1F7U).
Upper right: The reference structure: theoretical model of *Saccharomyces cerevisae* $tRNA^{ser}$. (5TRA)
Lower left: The refined 3D model of *Saccharomyces cerevisae* $tRNA^{ser}$.
Lower right: Superposition of the backbone traces of model and reference structure.
Color code: A/violet, C/blue, G/green, U/red.
Color code: reference structure 1F7U/blue line, model backbone/ball-and-stick representation.