Chapter 9. Visualizing RNA sequence and structure

Contributors. Kornelia Aigner, Fabian Dressen, Valérie Fritsch, Tanja Gesell, Fabrice Jossinet, Yann Ponty, Gerhard Steger and Eric Westhof.

9.1 Introduction

Context. RNA molecules are biopolymers equipped with the dual ability to store genetic information and play functional roles in the cell. Up until recently, their biological function was seen as primarily dedicated to the synthesis of proteins, but it is now fully recognized that RNA molecules play key roles at all stages of cell life in Eukarya, Bacteria, Archaea and viruses. The advent of high-throughput experimental approaches has greatly increased the identification rate of new RNA entities. They can be either "simple" domains acting as *cis*-regulatory elements on the transcriptional and post-transcriptional regulation of gene expression or fully functional molecules acting as *trans*-regulatory elements. In the latter case, their sizes range from the tiny 18- to 25-nucleotides (nts) micro RNAs, up to 100-200 nts for the transcriptional regulators found in bacteria and to more than 10 000 nts for RNAs involved in gene silencing in higher eukaryotes. These non-coding RNAs (ncRNAs) are involved in various processes such as transcription, gene silencing, replication, RNA processing, RNA modification and RNA stability. However the biological function of a large majority of these RNAs still remains to be determined, motivating further functional studies.

[Figure 1 here]

RNA can establish stable structures, even with only a small number of nucleotides, giving opportunity for interaction. It then only takes about a dozen nucleotides for RNA to fold back on itself, forming a stable hairpin loop. Few more nucleotides will increase its potential complexity, allowing for a diversity of putative secondary structure motifs, like internal loops, three-way and four-way junctions. The first solved 3D structures of transfer RNAs highlighted the ability of RNA of around 70 nucleotides to adopt an intricate, compact and globular tertiary structure. Recent advances in molecular and structural biology made it feasible to experimentally determine the structure of larger RNAs ranging from sub-domains of ribozymes to full eukaryotic ribosomal RNAs, and lately to whole viruses genomes, as illustrated by Figure 1.

In RNA studies, a fruitful paradigm consists in using structure as a hint, if not a proxy, for function. Indeed it is extremely uncommon for RNA to act as an unstructured molecule all throughout the stages of its cellular life. For some RNA families, the mature and fully functional product is characterized by a highly structured RNA. For others, these structured states distinguish precursors that are recognized by molecular assemblies and maturated into smaller, less structured, RNAs. The folding of an RNA molecule is widely thought as a two-step process: First, canonical base-pairs form a stable **secondary structure**; Then the **tertiary structure** is stabilized by recurrent tertiary modules and long-range interactions. Energetically, the

secondary structure is the main component of RNA architecture. It produces the scaffold constraining the tertiary structure and its characterization constitutes an essential step in any effort to unravel its function. Its definition constitutes a first and essential step towards the comprehension of the folding abilities of RNA (Brion and Westhof 1997; Tinoco and Bustamante 1999).

[Figure 2 here]

Two strategies for RNA studies involving interactive visualization. Visualization and interactive manipulation of solved RNA structures have played an essential part in the identification of the first structural rules stabilizing RNA architectures. The accumulation of solved structures for RNAs, involved in a large range of biological functions, must be exploited to further our understanding of the common principles governing RNA structure. Such a structural perspective over RNA biology, termed the **top-down approach**, studies solved structures in order to discover fine structural principles. What could be more surprising to an external observer is that visualization and interactive manipulation are still the most efficient way to study RNA architecture. Indeed, even though the major recurrent secondary and tertiary motifs have been identified, we are only starting to decipher the rules constraining their evolution, limiting our ability to identify them automatically in new solved structures. The identification of the first evolution rules have been stimulated by the recent accumulation of genomic data and by the release of new tools interconnecting them with structural ones. Unfortunately, due to computational requirements, their usage in automated approaches is still nascent.

Besides the top-down approach, whose goals largely overlaps with that of structural biology, RNA biology routinely faces situations where only sequential data are available. Again in this case, structure can be used as a proxy for function, and can partly be inferred by computational methods. This is the **bottom-up approach**, which starts from the sequence of an RNA entity to the characterization of its 2D and 3D structures. Several bioinformatics approaches have been developed to automate this process. Due to many limitations, however, the results produced by these approaches have to be validated against user assumptions and results of low-resolution experimental approaches done in parallel. In general, it is an iterative refinement, alternating automated approaches with the interactive visualization and manual post-processing of the computed result, as illustrated by Figure 2. Moreover, although necessary, the characterization of an RNA structure is not sufficient. Indeed there is no one-to-one mapping between the presence of a structural motif and the ability to achieve a given biological function. Consequently, interactive visualization is required at all stages of the quest for the insertion points of RNA in biological networks.

Main objectives of RNA structural visualization. The visualization of RNA structure addresses a diversity of – possibly conflicting – objectives:

- *Assist in functional segmentation*: Structural motifs, features and functional domains should be easily spotted by the trained eye. To assist in this task, any graphical rendering method should present higher order elements and organizations in the most regular way. For instance, any interacting base pair and any bases that are consecutive in the backbone should be presented in such a way that helices naturally stand out, and

overlapping bases and motifs should be prohibited. By the same token, the layout of secondary structures should favor helix orientations that are multiple of 90 degrees, taking advantage of the viewer's propensity to partition the plane into cardinal directions.

- *Suggest higher-order organization:* The secondary structure of RNA can be thought of as a schematic 2D projection of a 3D molecule. Accordingly, the layout of RNA secondary structure can be designed to hint at its corresponding three-dimensional organization. For instance, the relative orientations of helices can suggest coaxial stacking or the three-dimensional conformation of a three-way junction.
- *Reveal similarity:* Structural and, ideally, functional similarity with other molecules should be evident from the inspection of the drawing. This induces a constraint of robustness, prescribing that minor changes in the structure do not impact too drastically the structure. For the same reason, established layouts and norms for extensively studied families of RNAs can sometimes be preferred to fully automated methods.
- *Provide context to position-specific annotations:* Structural or evolutionary data can be produced for each position in the sequence by a variety of experimental and computational methods. Mapping these data on a – possibly partially-determined – secondary structure can be useful to provide functional and structural insight, or reveal some inconsistencies between inferred models and experimental evidence.

9.2 The many facets of RNA data

9.2.1 RNA file formats

At a sequential level, RNA single sequences and sequence alignments are encoded using the ubiquitous, yet poorly structured, FASTA format. Since this format disallows structural annotations, the STOCKHOLM format is usually preferred for representing alignments, which are based on/used for the inference of a common structure.

Compared to proteins, a defining feature of RNA *in-silico* methods is a strong historical focus on the secondary structure. Both CONNECT (CT) files produced by Mfold/UnaFold (Zuker and Stiegler 1981; Markham and Zuker 2008), and BPSeq files arising from comparative modeling, offer the possibility to associate a partner to each position. Computational approaches usually disregard crossing interactions, typically giving rise to pseudoknots. The Vienna RNA dot-bracket notation takes explicitly advantage of this restriction to represent a secondary structure as a well-parenthesized string of characters, where any positions associated with matching brackets are paired. The PseudoBase (van Batenburg, Gultyaev et al. 2000) format extends this format to represent pseudoknots using multiple types of parentheses. Finally, the versatile RNAML format can also be used to represent RNA secondary structures, and is the only format to date that can represent base-triples and non-canonical base-pairs.

At the structural level, the PDB format is used to describe the atom coordinates associated with a 3D model. The feature-rich RNAML format is also able to represent the 3D coordinates of all-atoms models, along with the general elements of RNA architecture.

9.2.1.1 FASTA Format

```
>O.sativa.1 AJ489954.1/1-104
......UGGCUGUGACGACUAGGUGAAAUU.CAAGCUCAACAGACCAAAUCACAGGUCUC
..UCUCCAAGGCCUU.UGGAGAUGGGAUCUGUAUGCCGA.....GU..UUCCGCUC....
.AGCCG......
>O.sativa.2 AY013245.2/61987-62105
....GAUGGCAGUGACGACUUGGUAAUAUU.CAAGCUCAACAGACCAAAUCACAGGUCUU
CCUCUCUGGAUCCAC..UCCUCUGGGAUUGAUUUG..UAUGCCGAUUUUCCCGCUGAACC
GAGCCAUC....
>O.sativa.3 AJ307928.1/3-121
....GAUGGCAGUGACGACCUGGUAAUAUU.CAAGCUCAACAGACCAAAUCACAGGUCUU
..UCUCUCUGGAUCUACUCCUCAGGGAUUGAUUUG..UAUGCCGAUUUUUCCGCUGAACC
GAGCCAUC....
```

FASTA file for RFAM family RF00360/snoRNA Z107/R87 (cropped)

A FASTA file simply consists in a list of sequences, each preceded by a header. Each header line starts with a ">", usually followed by an accession number. The sequence starts on the next line, and must be broken into 60 characters-long pieces. After the end of the sequence, a new header may start another sequence. As shown in the above example, this format can be used to represent multiple sequence alignments through the introduction of a gap-character (A dot "." is the norm within RFAM, but a dash “-” is used by the Comparative RNA Web Site). This format is sometimes extended to include structural information, either as a consensus or as individual secondary structures.

9.2.1.2 ALN format

The ALN format was originally developed for ClustalW, one of the leading software for the automated multiple alignment of sequences. Today it is widely used within alignment tools.

```
CLUSTAL 2.1 multiple sequence alignment

M.musculus.1      UGGCCUCGUUCAAGUAAUCCAGGAUAGG--CU--GUG-CAGGUCCCAAGGGGCCUAUUCU 55
H.sapiens.2       UGGCCUCGUUCAAGUAAUCCAGGAUAGG--CU--GUG-CAGGUCCCAAU-GGCCUAU-CU 53
H.sapiens.3       GGACCCAGUUCAAGUAAUUCAGGAUAGGUUGU--GUG-CUGU--CCAG----CCUGUUCU 51
T.rubripes.1      CAACCGGGUUCAAGUAAUCCAGGAUAGGCUCU--GUAUCUGU--CUUGG---CCUAUGCU 53
H.sapiens.1       UGGCUGGAUUCAAGUAAUCCAGGAUAGGCUGUUUCCAUCUGU--G-AGG---CCUAUUCU 54
                   ..*   .********** *********   *    . * *      .    ***.* **

M.musculus.1      UGGUUACU---UGCACGGGGAC 74
H.sapiens.2       UGGUUACU---UGCACGGGGAC 72
H.sapiens.3       CCAUUACU--UGGCUCGGGGAC 71
T.rubripes.1      UGAUUACUUGCU-CUUGGAGG- 73
H.sapiens.1       UGAUUACUUGUUUCU-GGAGG- 74
                    .*****     *  **.*.
```

Example of a Clustal ALN file (*mir-26 microRNA precursor family*)

An example of a simple alignment can be seen in the above example. The file starts with a header line, identifying the data origin. The alignment is written in blocks of 60 residues

(including gaps). Each sequence block follows the general format "<*sequence-name*> <*aligned sequence*> <*residue-count*>". The ultimate line in a block indicates the conservation level within the column: "*" stands for perfect conservation; ":" stands for strong conservation, and "." stands for weakly conserved positions, while a whitespace " " indicates highly variable portions.

9.2.1.3 STOCKHOLM format

```
# STOCKHOLM 1.0
#=GF ID    mir-22
#=GF AC    RF00653
...
O.latipes.1             CGUUG.CCUCACAGUCGUUCUUCA.CUGGCU.AGCUUUAUGUCCCACG..
Gasterosteus_aculeat.1  GGCUG.ACCUACAGCAGUUCUUCA.CUGGCA.AGCUUUAUGUCCUCAUCU
R.esox.1                AGCUGAGCACA...CAGUUCUUCA.CUGGCA.GCCUUAAGGUUUCUGUAG
...
#=GC SS_cons            .<<<<.<<..<<<<<<<<<<<<<..<<<<..<<<<<<<.<<........
#=GC RF                 gGccg.acucaCagcaGuuCuuCa.cuGGCA.aGCuuuAuguccuuauaa

O.latipes.1             CCCCACCGUAAAGCU.GC.CAGUUGAAGAGCUGUUGUG..UGUAACC
Gasterosteus_aculeat.1  ACCAGC..UAAAGCU.GC.CAGCUGAAGAACUGUUGUG..GUCGGCA
R.esox.1                ACAGGC..UAAACCU.GC.CAGCUGAAGAACUGCUCUG..GCCAGCU
...
#=GC SS_cons            ....>>..>>>>>>>.>>.>>..>>>>>>>>>>>>>>...>>>>>>.
#=GC RF                 acaaac..UaaaGCu.GC.CaGuuGaaGaaCugcuGug..gucggCu
//
```

Stockholm formatted file fragment, excerpted from RFAM's seed alignment for miRNA mir-22. RFAM ID: RF00653.

STOCKHOLM format files are arguably the preferred representation for RNA multiple sequence alignments. As illustrated by the above example, a STOCKHOLM file breaks the global alignment of a set of sequences into portions of bounded width. Unique identifying prefixes (Organism name/Accession ID) are attached to regions within each portion to connect sequences in different portions. Additionally a set of mark-up lines, recognizable by their prefix "`#=`", specify various additional information, like accession ids, experimental parameters, bibliographical references…

Two mark-up line types that are especially relevant to RNA bioinformatics, are the sequence and secondary structure consensus, indicated by the "`#=GC RF`" and "`#=GC SS_cons`" respectively. The latter uses a parenthesis scheme to represent base-pairing positions using corresponding parentheses. Pseudoknots can be also represented using additional *alphabetical* systems of parentheses ("A" matches "a", "B"/"b"…).

9.2.1.4 Vienna RNA dot-bracket/Pseudobase notations

```
> Rat Alanine tRNA
GAGGAUUUAGCUUAAUUAAAGCAGUUGAUUUGCAUUUAACAGAUGUAAGAUAUAGUCUUACAGUCCUUA
((((((...((((.....)))).(((((.......)))))...(((((((...)))))))))))))).
```

Dot-bracket notation for the predicted structure (RNAFold) of the Rat Alanine tRNA.

In this format, the RNA sequence is coupled with a well-parenthesized expression denoting the base-pairs. This expression is well-parenthesized, meaning that any opening parenthesis can be unambiguously associated with a closing parenthesis, inducing a set of non-crossing base-pairs. For instance, in the above structure, the bases at first and ante-penultimate positions are base-paired.

```
       1590 1600 1610 1620 1630
#      |123456789|123456789|123456789|123456789|123456
$ 1590 AAAAAACUAAUAGAGGGGGGACUUAGCGCCCCCCAAACCGUAACCCC=1636
% 1590 ::::::::::::::[[[[[[:::::(((]]]]]]::::)))::::::
```

Pseudobase notation for the Gag/pro ribosomal frameshift site of Bovine Leukemia Virus
Source: Pseudobase, entry# PKB1.

Since matching parenthesis cannot induce crossing base pairs, pseudoknots cannot be represented strictly within this format. This format has therefore been extended by the PseudoBase to account for multiple parenthesis systems. In the above example, two helices are initiated by base-pairs at positions (1604,1623) and (1615,1630).

9.2.1.5 CONNECT (CT) format

```
80      dG = -33.48 [Initially -35.60]
1       U       0       2       80      1
2       G       1       3       79      2
3       G       2       4       78      3
4       G       3       5       77      4
5       A       4       6       76      5
6       U       5       7       0       6
7       G       6       8       75      7
...
75      U       74      76      7       75
76      U       75      77      5       76
77      C       76      78      4       77
78      C       77      79      3       78
79      U       78      80      2       79
80      A       79      0       1       80
```

CONNECT format for the Mfold 3.7 (Zuker and Stiegler 1981) predicted structure of the human let-7 pre-miRNA.

This format was introduced by Mfold (Zuker and Stiegler 1981), the historical tool for an ab-initio prediction of RNA secondary structure and is still used to date by several prediction tools. After an initial header consisting of the sequence length followed by a comment, each position is represented by its content and its neighboring bases through 6 fields, each encoded in a fixed-width (8 characters) column: 1) position; 2) IUPAC code for base; 3) position of previous base in the backbone (5'-3' order, 0 is used if first position); 4) position of next base in the backbone (5'-3' order, 0 is used if last position); 5) base-pairing partner position (0 if unpaired); 6) position (duplicated). Although such features cannot be predicted by Mfold, this format allows for a description of pseudoknots.

9.2.1.6 BPSEQ format

```
Filename: AM286415_b.bpseq
Organism: Yersinia enterocolitica subsp. enterocolitica 8081
Accession Numbers: AM286415
Citation and related information available at http://www.rna.ccbb.utexas.edu
1 U 0
...
117 U 0
118 U 236
119 G 235
120 C 234
121 C 233
122 U 232
123 G 231
124 G 230
...
230 C 124
231 C 123
232 A 122
233 G 121
234 G 120
235 C 119
236 A 118
...
```

Fragment of a BPSeq formatted 5s rRNA inferred by comparative modeling.
Source: Comparative RNA web site/Gutell Lab.

The BPSeq format is an alternative to the CT format introduced by the Comparative RNA Web site (CRW)/Gutell Lab. It essentially consists in a simplified version of the CT format. The file start with four self-explanatory lines identifying the data source and content, and is followed by the structure, specified through a sequence of a space-separated triplet of the form: 1) Position; 2) IUPAC code for base; 3) Base-pairing partner position (0 if unpaired).

9.2.1.7 PDB file format

The PDB (Berman, Westbrook et al. 2000) format is a comprehensive text-formatted representation of macromolecules used with the authoritative eponym repository of experimentally-derived 3D models. Originally introduced to represent protein structure, it has been enriched over the years to include fine details of the experimental protocol used for structure derivation.

9.2.1.8 RNAML format

RNAML (Waugh, Gendron et al. 2002) is an XML format, introduced to address the dual need to unify data representations related to RNA, and to represent novel important features of its structure (e. g. non-canonical base-pairs and motifs). Although still challenged in its former goal by less structured domain-specific formats (arguably because of the intrinsic verbosity arising from its ambitious goals) RNAML has established itself as the format of choice in its latter goal, and is currently supported by most automated methods capturing non-canonical interactions and motifs.

9.2.2 RNA Databases

Due to its versatile nature, RNA data is organized in a number of databases, dedicated either to structural, sequential or functional family-specific data. Table 1 summarizes the authoritative sources of RNA data. The main drawback of the current situation is the present absence of unique identifier connecting the sequential, structural and contextual data within several databases. However this situation may only be temporary, as witnessed by recent structuring initiatives (Bateman, Agrawal et al. 2011; Birmingham, Clemente et al. 2011).

[Table 1 here]

9.3 From RNA tertiary structure to structural rules: the "top-down" approach

9.3.1 3D visualization of all-atoms RNA models

[Figure 3 here]

Most standard aspects of visualizing 3D structures of RNA can be performed adequately by molecular graphics tools designed for proteins, such as RasMol, Jmol, PyMOL and Swiss-PdbViewer. We refer to Chapter 10 for a detailed review of the general capacities of such tools, since most of these tools were introduced in the context of protein structure studies.

However, some specific representations are associated with the 3D visualization of RNA. For instance, the path from phosphorus atom to phosphorus atom is usually drawn larger and wider than the carbon alpha chain in proteins. Also, base pairs can be simplified by rectangles or rods linking the two sugar-phosphate backbones. An illustration of both these aspects can be found in the coarse-grain *cartoon* view of RNA implemented within the versatile PyMol. In combination with visual effects such as transparency, such a representation allows to emphasize in a clear manner some local structural feature of interest while retaining its context, as illustrated in Figure 3 and Figure 4.

9.3.2 Automated base-pair annotation of tertiary models

[Figure 4 here]

Given the ever-increasing complexity of solved tertiary structures, secondary structure can play the role of a 2D map to improve our understanding and handling of large RNA architectures. Novel algorithms have been made available in order to enable the automatic identification of all base-base interactions participating in the stability of the structure. The resulting **extended secondary structure** supplements the classical definition of helical and single-stranded with a description of non-canonical base-base interactions. Among several available classifications, the Leontis-Westhof nomenclature (see eponym display box) is the most recent and exhaustive one (Leontis and Westhof 2001), and has been universally adopted by existing implementations.

Starting from all-atom 3D models, MC-Annotate (Lemieux and Major 2006), RNAVIEW (Yang, Jossinet et al. 2003) and Fr3D (Sarver, Zirbel et al. 2008) typically produce a list of base-pairs annotations in the RNAML format. Following the Leontis-Westhof classification, each base-pair being associated interacting edges for both partners, and a relative orientation. Existing tools may also consider the local context of the base-pair, inferring stacking bases or stacking pairs, or the strand orientation, which cannot necessarily be deduced from the orientation. Integrated approaches such as S2S/Assemble typically rely on such tools for their own visualization, through calls to webservices.

9.3.3 Visualization and manipulation of RNA extended secondary structure

Using automated annotation software (see Section 9.3.2) or through manual inspection, an extended secondary structure can be derived directly from a solved tertiary structure. Some tools, like VARNA and S2S/Assemble, are then able to display tertiary interactions on top of the secondary structure, using Leontis-Westhof symbols, as illustrated by Figure 4 and Figure 7. The relative orientation of helices may either be derived from a projection of the helical axes in the tertiary model (Figure 4), or follow aesthetic principles (e.g. “equal angular increment”) prescribed by existing layout algorithms such as NAView (Bruccoleri and Heinrich 1988) (Figure 7).

9.3.4 Refining RNA alignments using 3D model data

By helping reveal an evolutionary pressure, comparative sequence analysis is one of the most fruitful approach for discovering new structural rules and motifs. For instance, the presence of compensatory mutations within an RNA family can be used to identify regular helices, and consequently the locations of single-stranded regions (hairpin and internal loops, bulges and multiple junctions). Unfortunately, automatic approaches to produce RNA structural alignments suffer from several limitations, leading to a necessity for interactive refinements, as discussed in Section 9.4.2.

When an all-atoms structural model is available for at least one of the aligned sequences, the additional knowledge of the base-pair orientation can be integrated to the interactive refinement of the alignment. The underlying rationale is that the opportunity for a conserved tertiary structure should be favored over alternatives. For instance, base-pair isostericity (Leontis, Stombaugh et al. 2002) with respect to the reference model should be preserved as much as possible within the alignment. The two leading software for such a scenario are described below.

9.3.4.1 S2S Software

[Figure 5 here]

S2S offers interesting graphical features adapted to RNA specificities. Starting from a solved 3D architecture described in a PDB file, S2S annotates it automatically and use it to produce a structural mask against which orthologous sequences can be aligned (Figure 5). Several graphical facilities have been developed to leverage this time-consuming task.

Firstly, since base-base interactions can be of long-range, S2S can produce several views of the same alignment that can be manipulated and edited independently. Between these views, the secondary and tertiary interactions are displayed only if their residues are present simultaneously in the alignment panel. Secondly, for each view, the graphical engine of S2S highlights the compatibility of orthologous sequence(s) and the local interaction schemes. More precisely, S2S computes, from a set of isostericity rules (Leontis, Stombaugh et al. 2002), the ability of the orthologous sequence to conserve the folding of the reference structure. Technically, two different colors can be chosen by the user to render the secondary and tertiary interactions. For each color, the darkest one indicates that the combination of the orthologous residues chosen by the user with the type of interaction observed in the reference structure will be isosteric with the reference base-pair. The lightest color indicates that the combination is geometrically possible but not necessarily isosteric with the reference base-pair, whereas the absence of color indicates that such combination has never been observed in solved structures so far. This structural mask enables identification of the *core* structure conserved in all of the orthologous RNA sequences in the alignment. Consequently, it also allows the identification of structural evolutionary pressure, and additional domains accumulated by these sequences during evolution.

9.3.4.2 BoulderAle Software

[Figure 6 here]

Boulder Ale (Stombaugh, Widmann et al. 2011) is a user-friendly web-based tool for editing structurally-informed RNA alignments. Not only does it show the primary sequence and the Watson-Crick base pairs but non-Watson-Crick base pairs and their isostericity within the alignment columns are also displayed. This editor allows for adjusting and evaluating an alignment manually with respect to additional structural information, represented as a base pair list. The base pair list can be calculated from a loaded consensus structure, which then is also displayed in the alignment in dot-bracket notation and contains the standard Watson-Crick and Wobble-base pairs. It can also be calculated from a loaded PDB-file. In the latter case all kinds of base pairs, including the non-Watson-Crick base pairs, are calculated from the given experimentally-determined structure via the program *FR3D* (Sarver, Zirbel et al. 2008). Manual editing of the base pair list is possible. Above the alignment window, several lines of annotations denote for each alignment column the kind of occurring base pairs, according to the base pair list. Triple base pairs are denoted as well.

For all sequences the isostericity of each base pair, compared to the corresponding base pair in the reference sequence, is shown in a color scheme, where green means isostericity, pink means non-isostericity and cyan means not allowed. Furthermore the alignment can be annotated with features or motifs and collapsed horizontally and vertically. The sequence order is changeable, gaps can be inserted and deleted, and sliding of nucleotides is possible.

If a reliable consensus structure and reference sequence or an appropriate PDB-structure is available, Boulder Ale is a comfortable tool to improve a structural RNA alignment, including non-Watson-Crick base pairs, to visualize the alignment itself, to calculate the base composition of its contained structural elements via Kings, and to visualize 2D structure plots of its sequences via an embedded version of the VARNA software (Darty, Denise et al. 2009).

9.3.5 Interactive 3D modeling of RNA architectures

Despite the fact that the number of available RNA tertiary structures has increased dramatically over the last few years (strengthening our knowledge of RNA structure and folding), a large majority of RNAs discovered so far are still to be crystallized. As an alternative, a three-dimensional architecture for an RNA molecule can be produced using a molecular modeling strategy. Such theoretical approaches have proved valuable in the past. A first method is based on a constraint satisfaction algorithm that searches the conformational space so that the models compatible with a given set of constraints are calculated, e.g. the MCFold/MCSym pipeline (Parisien and Major 2008). A second method uses a human iterative process to identify the structural constraints from a mixture of theoretical and experimental data allowing the construction of the model.

Among the different types of information, the availability of a solved tertiary structure for at least one homolog is perhaps the most powerful. Since molecular 3D architecture evolves much more slowly than sequences, structural data can be inferred for all the members of the RNA family with the same function. RNA 3D modeling by homology requires the alignment of the sequence to be modeled against some solved tertiary structure (a.k.a. "reference structure").

9.3.5.1 Interactive derivation of RNA all-atoms models from experimental data

Assemble also has the ability to render density maps to guide the user during the reconstruction of the RNA 3D model from experimental data. Assemble can load density maps described with the XPLOR or MRC file formats. COOT.

9.3.5.2 Interactive *ab-initio* 3D modeling

[Figure 7 here]

Starting from a single sequence, S2S (Jossinet and Westhof 2005) and Assemble (Jossinet, Ludwig et al. 2010) can delegate the computation of a first draft for the secondary structure to folding RNA algorithms made available as web services, such as RNAfold (Hofacker 2009) or Contrafold (Do, Woods et al. 2006). The usage of webservices offers consistent access to up-to-date prediction methods in a way that is transparent to the user (Curcin, Ghanem et al. 2005). Once retrieved and parsed, the result is displayed on an interactive graphical panel. S2S and Assemble also allow for an interactive edition of the resulting secondary structure. Base-base interactions and helices can then be easily added and removed, allowing for the design of structural features that cannot be easily computed/predicted automatically (like non-canonical interactions and pseudoknots).

In the absence of further 3D information, *ab-initio* modeling strategy are needed to extract a putative tertiary structure. To reduce this difficulty, Assemble has the ability to extract, from solved tertiary structures, local folding that can be applied to selections done in the 3D model. Using an embedded database made from a selection of annotated 3D architectures, the user can search for local RNA folds (a.k.a. RNA motifs (Leontis, Lescoute et al. 2006)) and store them in a local repository to apply their 3D fold to a selection within the 3D model.

9.3.5.3 Homology modeling from a preexisting 3D model

[Figure 8 here]

In the presence of a structural model is available, the availability of an alignment of good quality greatly helps modeling a sequence of interest. Such alignments are typically obtained through a refinement process using specialized editors (see Sections 9.3.4 and 9.4.2). Once done, for any orthologous sequence aligned to the reference structure, S2S can derive a secondary and a tertiary structure and export them to be used with Assemble to resume the modeling process.

When a user opens a tertiary structure stored in a PDB file, S2S and Assemble annotate it automatically using the RNAVIEW web service. The resulting extended secondary structure is then used for two different goals. Within Assemble, this secondary structure is displayed in a panel linked to a 3D scene rendering the original 3D architecture. Any selection done in the 2D panel highlights its 3D counterpart and centers the 3D scene on it (Figure 7). Within S2S, it defines the structural mask that will graphically guide the user during the construction of his structural alignment (Figure 5).

If the structural alignment is largely correct, the rest of the modeling process in Assemble should be limited to fine tuning through removal of steric clashes, and possibly completion of the local folding on inserted/deleted positions. Steric clashes can be fixed using a toolbox of Assemble allowing a modification of the torsion angles for any residue selected in the 3D scene. Residues not present in the reference structure and, consequently, not present in the 3D model derived by S2S, can be created *de novo* and exported in the 3D scene by selecting them in the 2D panel of Assemble.

Unfortunately, the differences between the sequence to be modeled and the reference structure can be more important. Orthologous sequences within a given RNA family can contain additional domains with sizes varying between about ten to more than hundred nucleotides (as observed for ribosomal RNAs (Yokoyama and Suzuki 2008) or RNaseP RNA (Kachouri, Stribinskis et al. 2005)). Due to the peculiarities of RNA architectures, such additional domains can form independently folded units that may be added to or removed from the common core with minimal alterations to the fold (Westhof and Massire 2004). The ability of RNA architectures to "accumulate" new modules has also for consequence to produce RNA alignments that become quickly unmanageable without adapted tools (Brown, Birmingham et al. 2009).

If present in the structural alignment, these additional domains are displayed as unfolded regions, and, consequently, as mere single-strands in the derived secondary structure displayed within Assemble. Such a partial model can then be completed by an ab-initio interactive approach such as that described in Section 9.3.5.2.

9.4 From RNA sequence to RNA function: the "bottom-up" approach

9.4.1 Visualizing RNA secondary structure

9.4.1.1 Main representations of RNA secondary structure

[Figure 9 here]

The diversity of applications and imperatives naturally leads to a variety of methods for drawing the secondary structure of RNA (see Figure 9), associated to different assets and disadvantages:

[Figure 10 here]

- **Linear arc-diagrams (Figure 9.3 and Figure 10):** In this representation, the sequence is drawn on the horizontal axis in the 5'-3' order, and base-pairing positions are linked together by arcs. Helices are easily identified as sets of consecutively nested arcs, and multiple loops naturally appear as the empty space delimited by a set of arcs/base-pairs. Complex features like pseudoknots can be also represented using linear diagrams, possibly using an alternative color to keep the focus on the secondary structure. However this representation suffers from a few limitations as the sequence length increases. Indeed, the *horizontal expansion* induced by its linearity may impede the identification of the sequential context of a structural motif, giving rise to sparse drawings where individual bases are no longer easily identified.

[Figure 11 here]

- **Mountain Plot (Figure 9.2 and Figure 11):** Here the sequence is again drawn linearly, but this representation also presents, at each position *i*, the number of base-pairs *nesting* the position, i.e. involving bases respectively before and after *i*. In this setting, helices give rise to *mountains* while terminal loops translate into *peaks*. This representation helps depict the hierarchical organization of RNA secondary structure, as nested helices translate into stacking mountains, easing the visual segmentation into domain.
  However pairing positions, at equal height on both sides of a mountain, can become hard to identify in this representation as the width of the mountain increases. For similar reasons, multiple-loops, represented by multiple mountains initiated from a common *plateau*, can be hard to distinguish from nested bulges, giving rise to plateaus at different height on their left and right side. Finally, this representation suffers from the same horizontal expansion issues as linear arc-diagrams.

[Figure 12 here]

- **(Outer-planar) graph (Figure 9.4 and Figure 12):** This popular – compact representation – draws a secondary structure as a graph with two types of edges (Backbone adjacency and base-pairs), while enforcing three major types of constraints:

  1. Helices should be drawn on a straight axis (ladder).
  2. Predefined distances should be respected between two connected bases.
  3. The resulting drawing should be non-overlapping.

  Since these constraints cannot always be simultaneously satisfied, and since the associated algorithmic problems are known to be intractable (NP-complete) (Auber 2006), existing software either produce a static picture, typically violating constraint **2.** whenever necessary, or produce an overlapping initial draft, providing editing facilities to *manually disentangle* the layout. Since the latter task may become demanding for large

RNAs, certain software may ease the user experience by proposing some template system, allowing reuse of an existing layout for homologous sequences (up to a certain level of structural dissimilarity).

The advantages of this representation are numerous. First it produces very compact drawings compared to the linear and circular representations. It also helps emphasize structural entities and domain (helices, interacting stems…), and may be tailored to faithfully reflect the tertiary organization (e.g. within S2S (Jossinet and Westhof 2005) and Assemble (Jossinet, Ludwig et al. 2010)). Finally the layout algorithm can be designed to be robust to small local structural changes, e.g. by giving a limited weight to unpaired bases in the general orientation of helices.

The main drawback of this widely used representation is the lack of universally accepted aesthetic principles guiding its layout. Indeed manually drawn graph representations have now been used for a couple of decades by structural biologists and *de facto* standards have been established for certain RNA families on a case-per-case basis. Since the principles underlying these *canonical* representations are not homogeneous and sometimes conflicting, they cannot be entirely captured by fully automated layout algorithms. A classic example, illustrated by Figure 12, is the tRNA cloverleaf-shaped secondary structure, typically expected as shown in Figure 12.B or Figure 12.C, but rendered as Figure 12.A or Figure 12.D by fully automated procedures. It follows that the need for interactive *a posteriori* manipulation cannot be entirely circumvented.

[Table 2 here]

[Figure 13 here]

9.4.1.2 Quick automated visualization of RNA secondary structure

In a number of situations, one is interested in getting a quick initial glance at a given secondary structure. For instance, one may focus on a given transcript, and may have gained access to putative secondary structures, either from a database or using some computational method. In such cases, a quick visualization of the secondary structure will help the educated user in a prior validation.

To this purpose, **VARNA** (Darty, Denise et al. 2009) and **RNAPLOT** (Gruber, Lorenz et al. 2008) are definitely the tools of choice, for interactive and static (command-line) visualization respectively. VARNA will conveniently accept most files formats and open/display them automatically through drag-and-drop gestures within a minimalist environment. RNAPLOT will take as input Vienna dot-bracket formatted files, and will produce compact (E)PS files that can be viewed or edited using vector graphics editing software. VARNA also offers similar command-line functionalities, but RNAPLOT may sometimes be more convenient because of its ability to draw, in a single run, multiple sequence/structures bundled within a single input file.

[Figure 14 here]

9.4.1.3 Visualizing complex structural features: Pseudoknots.

Secondary structures may feature pseudoknots, i.e. sets of base-pairs that are mutually crossing in a linear representation. Unfortunately, such motifs violate some of the assumptions

on which automated graph drawing algorithms are based. Consequently the automated layout of general pseudoknots remains an unsolved computational problem.

In spite of these limitations, the **Pseudoviewer** (Byun and Han 2009) implements an *ad hoc* algorithm that successfully draws most existing pseudoknots as planar graphs (Figure 14 **A.**). The resulting pictures are compact and aesthetically pleasing. A Windows-only graphical interface will allow for further editing/annotation of the produced layouts.

RNAMovies (Figure 14 **B.**) and **VARNA** (Figure 14 **C.** and **D.**) both adopt another approach. They first consider a non-crossing subset of basepairs, then compute a layout using some standard algorithm and finally *complete* the drawing by drawing the remaining base-pairs. The result can then be post-processed either within a dedicated interface or using vector graphics software.

9.4.1.4 Visualizing the reliability of predictive methods

Following the emergence of ensemble-based approaches for the *ab-initio* computational prediction of RNA secondary structure, multiple criteria have been proposed to assess one's confidence in the prediction. For instance, assuming a Boltzmann distribution on the set of putative secondary structures, one can assign a probability to each base-pair, and Mathews (Mathews 2004) has shown that this probability correlates with the probability of experimentally observing this base-pair. In particular, the study showed that base-pairs associated with Boltzmann probabilities greater than 99% could be verified by experimental methods 91% of the time. Therefore such measures can be used as reliability indices, whose visualization mapped on a single structure can provide useful insight into the sequence structure relationship.

[Figure 15 here]

The **Vienna RNA webserver** (Gruber, Lorenz et al. 2008) offers a variety of representations for visualizing reliability information, summarized by Figure 15. In this illustrative example, the method is accurate in its predictions but predicts three extraneous base-pairs (blue ellipses) in the MFE structure compared to the RFAM consensus structure. The reliability plots **b.**, **c.** and **d.** tag this region as unstable, unreliable or highly entropic: The graph layout (Figure 15.b) colors theses bases in green, associated in the heat map with lower probabilities. In the mountain plot (Figure 15.b), the MFE structure plot departs from that of the average and centroid structures. Finally the dot-plot associates smaller squares to the corresponding base-pairs. The method also identifies a plausible – yet slightly energetically unfavorable – additional small hairpin (green ellipses), associated with smaller positional probabilities in the MFE graph layout (Figure 15.b), with an extra bump in the mountain plot of the average structure (Figure 15.c) and with a line of squares in the upper-right triangular matrix that is not seen in the lower-left part.

9.4.1.5 Interactive drawing of multiple homologous secondary structures

The dissemination of scientific results through publication is one of the motivations for the development of visualization tools and methods. Accordingly secondary structure diagrams have typically been used presented to illustrate functional mechanisms, map experimental and evolutionary evidences on a putative structure…

Since general-purpose software usually does not maintain universally desirable features of RNA layouts, such as backbone connectivity, the ability to support rich RNA-aware editing and annotation features is critical. Furthermore, support for output using vector formats will allow convenient, lossless, posterior editing, while keeping the file size extremely low (at infinite theoretical resolution). Among desirable features, one includes the ability to define **templating mechanisms** which, by separating the general layout from the specifics of a given RNA, allows to draw a set of homologous structures identically.

XRNA and **RNAViz** (De Rijk, Wuyts et al. 2003) are arguably the most mature tools for this application. Both offer rich editing and annotation features, as well as convenient gestures for disentangling the initial drawing but only support graphs representations. Compared to the leading tools, **VARNA** offers limited editing and annotation features, but can be used to produce other types of layouts.

[Figure 16 here]

R2R also produce aesthetic graph representations of the secondary structure, possibly annotated with additional information such as conservation levels (see Figure 16). Although the software does not offer a graphical user interface, it is highly customizable through a set of annotations which can be added to the input file, allowing the user to specify the relative orientations of helices, the layout policy for unpaired bases… Its learning curve may be a bit steep for non-technical savvy users, but it arguably produces the most aesthetic results and can accommodate for and save virtually any personal preferences.

9.4.2 RNA alignment tools including secondary structure

The perfect sequence/structure alignment, implemented by the Sankoff algorithm, is prohibitive in terms of computational costs and memory usage (see section **9.4.1.2.2**), and the simplifying heuristics proposed to work around these issues may produce misalignments. It is not uncommon that those mistakes can be detected by the educated eye. To allow for such manual corrections, RNA-alignment editors, displaying both sequence and structure simultaneously, can be a very valuable asset.

9.4.2.1 A workflow for the interactive refinement of RNA Alignments

[Figure 17 here]

Figure 17 describes a common workflow for the iterative refinement of the RNA structural alignments. Starting with an RNA sequence **[1]**, its secondary structure **[2]** can be calculated with an RNA folding program (eg *RNAfold* or programs included in the *Mfold/UNAfold* and *RNAstructure* packages). At this stage a secondary structure with minimum free energy, and thus stable, is predicted by using thermodynamic parameters. The result can be considerably improved by taking into account information from homologous sequences, which are homologous in terms of function and structure. A common way to get access to related sequences is to simply *BLAST* the query sequence **[3]**. The obtained sequences are relatively conserved in terms of sequence and may thus become the starting point of an iterative process (see later). A widely used program for aligning sequences (e.g. ClustalW) creates a multiple sequence alignment **[4]**. Since the common structure of a non-coding RNA is more conserved

than their sequences, this alignment is error-prone. An improvement of the alignment taking into account structural information is usually necessary **[5]**, therefore secondary structures of each sequence **[2]** are commonly used.

The use of a 3D-model **[6]** of a sequence, which is included in the alignment, is an alternative way to improve the alignment. If a fitting 3D-model is available (e.g. in the PDB), the alignment can be corrected in such a way that nucleotides of alignment columns can build basepairs that are similar or isosteric to that of the model's structure (*BoulderAle*, see section). While only few structures are solved, a vast amount of sequences are available; thus the usage of predicted structures is thus common and advantageous to evolve a multiple sequence/structure alignment from a pure sequence alignment **[from 4 to 5]**.

The iterative refinement of an alignment using structural information is an intricate manual operation and therefore needs a good alignment editor. The editor must be able to display both sequence and structure, allow for the sliding of nucleotides/gaps in the alignment and indicate the improvement. Several editors are currently available (see Table 3), differing with respect to their specific intention, their way of processing data, their algorithms, their features, their kind of input and output data, and their graphical user interface. Mostly a prealigned set of sequences can be loaded and additional structural information helps to improve the alignment.

A significant difference between the editors is the kind of structural data they process, which gives rise to their differing abilities to correct an RNA alignment. Naturally a 3D-model is a reliable basis for this procedure (*BoulderAle*, see section). A consensus structure ---as result of a multiple sequence-structure alignment--- might be used to align further sequences. A thermodynamic optimal structure for each sequence of the alignment allows, assuming a user-friendly graphical interface, for manually aligning the structures and thus improving the alignment. The structures are mostly displayed in Dot-Bracket-style (*4SALE, SARSE, RALEE*), while some editors provide clearly represented 2D-plots (*4SALE*).

Methods that take only the thermodynamically optimal structure into account neglect the possibility that an alternative structure is the biologically relevant structure. The partition function of all possible structures, as calculated by *RNAfold*, *UNAfold* and *RNAstructure*, or near-optimal structures, as calculated by *RNAfold*, *UNAfold*, *RNAstructure* and *CentroidFold*, for a sequence might be of higher importance for function than the single optimal structure.

Using base pairing probabilities derived from partition function calculation for each sequence is an evidentiary basis for creating good sequence and structure alignments. Pairing probabilities can be displayed in a dot plot, where pairing probabilities for each sequence are indicated by the size of dots in the dot plot, and possible helices build diagonals in the dot plot (see *ConStruct*). The alignment is refined by sliding nucleotides in the alignment window such that the structures in the dotplot lay upon each other. The more sequences build a basepair between two positions (columns) of the alignment, the higher is the probability of a consensus basepair at that positions. This probability is further increased if those positions show covariance. That means a mutation at one of these positions is accompanied by a specific mutation at the other position.

Covariation measurement in ncRNA is a basic principle for phylogenetic calculations **[8]** and an additional important method for construction of consensus structures. Covariance is evaluated

by some of the available RNA editors, e.g. *4SALE* which counts Compensatory Basepair Changes (CBC) or *ConStruct* which calculates either the Mutual Information index (MI) or optionally a covariation score as implemented in *RNAalifold*.

A multiple sequence/structure alignment allows the calculation of a consensus structure **[7].** The consensus structure **[7]** is a compacted source of information. The function of an RNA family is often detectable through its consensus structure since it contains structural (and sequential) peculiarities that recur within the family. Subtleties and specifics of single members of the family can be detected by comparison of its predicted secondary structures with the consensus structure **[10]**.

The whole process of finding homologous sequences, aligning them and finding a consensus structure can be refined through iteration. Once one has a consensus structure, more homologous sequences with less sequence conservation can be found through database search with a structural pattern **[9]** than by simple sequence search. This is due to the structural conservation of noncoding RNAs that often is accompanied with relatively low sequence conservation. The higher the sequential differences are, the higher is the information content of the alignment's columns, thus a predicted consensus structure gets even more reliable.

9.4.2.2 Re-aligning RNA sequences by taking base-pair probabilities of each sequence into account

Pure RNA sequence-alignments are often misaligned due to their low sequence conservation. Since RNA structure is far better conserved, the alignment can be refined by taking into account thermodynamically determined base-pair probabilities for each sequence. An RNA alignment editor helps to adjust the alignment in such a way that probable homologous base pairs are positioned in corresponding columns of the alignment.

For this approach the sequences are pre-aligned through a simple sequence alignment program (e.g. clustal) using either *FASTA*, *STOCKHOLM* or *VIENNA* format. Base pair probabilities are calculated through thermodynamic structure determination, e.g. *RNAfold*. To demonstrate the application, the curation of an alignment of Secis-element sequences is shown, using the RNA alignment-editor *ConStruct*.

[Figure 18 here]

The SECIS-element, an RNA structural motif that consists of a stem- loop structure, is relatively low conserved in sequence but highly conserved in structure. Accordingly, alignments created by standard sequence alignment programs are far from structurally correct. Structural information is crucial for producing a reliable alignment. Here, a pure sequence alignment of Secis-element sequences is displayed in an alignment window (see Figure 18: A, bottom).

Probabilities of base pairs for each sequence were calculated and are shown as green squares in a dot plot (see Figure 18: A top right dotplot; upper right and lower left triangle for thermodynamic and covariation pairing probabilities, respectively); base pairs of the selected sequence ---here the first sequence of the alignment--- are shown in blue.

Note the small yellow dots inside the blue squares, which indicates the mean pairing probability of all base pairs, and the "helix clustering" visible as a close accumulation of green diagonals in

the upper triangle of the dotplot. The low probability of consensus base pairs (small, yellowish and not larger, reddish dots) and the obvious non-superimposition of structures from different sequences both point to an incorrect multiple alignment. Here, five of the 14 sequences are already superimposed in their structure (note the colored nucleotides of identical sequence in the alignment window). Furthermore, from the dotplot it is already obvious that most other, not-superimposed structures can be aligned with those by mainly horizontal adjustment of base-pair positions. A major shift is necessary only for the two hdr_A sequences (see the off-diagonal helices in the dotplot). The user is guided during this adjustment process –i.e. regarding which of the sequences have to be selected, which nucleotides have to be moved in the alignment, etc.– by the direct interconnection between base pairs in the dotplot and corresponding nucleotides in the alignment editor. Additionally, the possibility of highlighting certain nucleotides or motifs in the alignment window by means of regular expressions (search pattern) might be of help during the manual refinement stage. In case of the SECIS elements this is the conserved GAA in the internal loop (see for example the orange colored motif in the alignment windows).

After the correction process from the sequence alignment in A to the corrected alignment shown in B, the alignment length is reduced by three nucleotides and all helices (green diagonals) except one are superimposed, thus building a consensus helix (red diagonal).

9.4.2.3 Calculating a consensus structure from a multiple alignment

A consensus structure of a set of RNA sequences represents the structural characteristics common to all or at least most of these sequences. It is calculated from an alignment of the sequences which is inspected columnwise. Each position of the consensus structure is the subsumption of the according alignment column. For this reason a correct alignment is crucial for achieving a correct consensus structure. There are two main approaches to calculating a consensus structure from an alignment, namely the thermodynamic and the covariation method.

The thermodynamic method considers each pair of columns in the alignment and computes the frequency of a base-pair within thermodynamically-determined structures. In the example from Figure 18, we illustrate the prediction of consensus structures not only with one structure per sequence, but with base pairing probabilities for the sequences. In this case, the probability of a consensus basepair is the sum of the base-pair probability over all sequences. Iteration through all pairs of positions results in a consensus structure with known probabilities for each position. In *ConStruct,* for all sequences, the probability of a basepair is indicated by the size of a square in a dot plot.

The second approach uses covariation of the sequences. Through evolutional mutation, sequences of RNA-families often show a relative low level of sequence conservation at which nevertheless the structure is conserved. Employing that instance, the probability of a consensus basepair can be calculated in measuring the amount of covarying positions. Covarying positions show joined nucleotide substitutions; that is, a mutation at the first position is compensated for by a specific mutation at the second position. A covariation score can either be calculated only for compensatory base pair changes of Watson-Crick and Wobble base pairs (*RNAalifold*), which abets helix detection, or in general for all possible pairs of bases which also detects tertiary interactions (Mutual Information Index used in *ConStruct*).

Covariation only occurs in sequences that are not too similar. Therefore the thermodynamic and the covariation method complement one another and a combination of both results in most reliable consensus structures. As shown in figure ConStruct the consensus structure calculated from the alignment of SECIS elements has increased consensus base pairing probability after alignment correction. The alignment has been manually corrected as described above and the consensus structure is calculated by summation of the thermodynamic and the covariation pairing probabilities. These probabilities are shown in the upper right (thermodynamic) and lower left (covariation) triangle of the dot plot before (see Figure 18: A) and after (see Figure 18: B) manual alignment correction.

9.4.2.4 Main Tools

In Section 9.3.4, we presented S2S and BoulderAle, two software that support tertiary annotations in addition to a secondary structure. Here we complement this list with major secondary structure-aware editors.

9.4.2.4.1 Construct: Comparing and aligning thermodynamic landscapes

[Figure 19 here]

ConStruct (Luck, Graf et al. 1999; Wilm, Linnenbrink et al. 2008) is an RNA alignment editor for improving RNA alignments while taking into account several structural information. The alignment is displayed in a standard way and all secondary structures are shown as overlaying dot plots of pairing probabilities in a separate window. The only required input is a standard alignment in *FASTA, Vienna* or *Stockholm* format. Structural information is obtained from thermodynamic base pairing probabilities (as calculated by the *RNAfold* software) and mutual information content. For each sequence a thermodynamic base pairing probability matrix is calculated; all matrices are superimposed, thus forming a consensus matrix, and are displayed in the top right triangle of the dot plot, where green and blue dots indicate base pairs and red dots indicate consensus base pairs. A score for covariation, either the mutual information index or the *RNAalifold* score, is calculated and shown in the bottom left triangle of the dot plot where tertiary interactions get visible. Editing the alignment is facilitated by corresponding movements in the dot plot (Seibel, Muller et al. 2006; Seibel, Muller et al. 2008). Superimposing of helices in the dot plot indicates improvement of the structural alignment.

Consensus structure prediction is based upon the weighted and filtered summation of the thermodynamic consensus dot plot and the covariation dot plot. Pairing probabilities of the consensus structure are indicated in its 2D-plot with colors from white to red. *ConStruct* runs on several Linux-distributions and Mac OSX.

[Figure 20 here]

9.4.2.4.2 4Sale

4Sale (Seibel et al., 2006; Seibel et al., 2008) is a user-friendly editor that allows for correcting an RNA alignment by simultaneously taking sequence and structure into account. It handles individual secondary structure information for each sequence, which can be manually edited. Rearranging or hiding of sequences and searching for sequence or structure patterns within the alignment facilitate handling large alignments.

Alignments or sequences are accepted as input in *FASTA* format. Sequences can be aligned via the program *ClustalW, DCA* or *DIALIGN.* Sequence and structure editing features are synchronized in both, the alignment and the 2D-structure viewer. The program runs on all systems.

9.4.2.4.3 SARSE

SARSE (Andersen, Lind-Thomsen et al. 2007) is an alignment editor that takes sequences in *FASTA* format and aligns them via the program *FoldalignM.* Structures are calculated via the program *Pfold* and displayed indirectly in terms of colored base pairs. Phylogenetic dependencies, calculated by taking compensatory base pair replacements of canonical basepairs into account, are made visible through clustering within the alignment. The program runs on Linux and Mac OSX.

9.4.2.4.4 RALEE

RALEE (Griffiths-Jones 2005) is a GNU-Emacs extension that takes Stockholm formatted sequences as input. An alignment and the consensus structure is calculated via the Vienna-RNA Package. A structure-based color scheme displays structures indirectly. Fetching sequences from Genbank is possible. The alignment window can be split horizontally and manually edited.

9.5 Perspectives

In this chapter we have concentrated on the “bottom-up” and ”top-down” approaches for RNA structures in order to demonstrate the importance of visualization tools including interactive manipulation for studying RNA architecture. Alternating automated approaches with the interactive visualization and manual post-processing of the computed results (Figure 2), we have described and illustrated this approach as an iterative refinement. We have noticed that the functional annotation of experimentally or computationally predicted structures remains an unsolved problem: Determination of RNA structure is necessary but not sufficient, and there is no clear definition of what constitutes functionality in terms of RNA structures. Additional data in computational analysis and experimental validations are necessary. On the one hand, ncRNAs are known to regulate gene expression at virtually every possible stage. On the other hand, a growing number of regulatory processes involving RNA in genetics and epigenetics are presently coming out thanks to advanced technology. As these additional data are increasing rapidly in volume and complexity, the characterization of RNA structures and the analysis of biochemical functions require a novel kind of theoretical biology as well as computational visualization methods and tools.

In our introduction, we proposed the following objectives as inherent to the visualization of RNA structures: to assist in functional segmentation, to suggest higher-order organization, to reveal similarity and, finally, provide a context to position-specific annotations. With respect to these goals, three of the main challenges for RNA visualization tools are currently as follows: (1) No automated tool exists for visualizing the larger RNAs that were recently made available by high throughput (large scale) structural mapping methods. The extent to which structural elements are required for the function of these RNAs is still an open question. While many different visual

representations exist for small RNAs, not much is done for these long RNAs: the challenge imposed by scale is evident, and macro and micro resolution may be taken into account interactively. A most effective balance between detailed visualization versus simplification is necessary, especially for these molecules, which can consist of several thousands of nucleotides. Generally, the diversity of scales for ncRNAs with lengths ranging from 20 to 10 000 nucleotides, does not only constitute a challenge to existing visualization techniques but also requires different representations depending on the respective sizes. (2) RNA often adopts multiple structures depending on experimental conditions, and none of the available tools currently supports these multiple structures to a satisfactory extent. Another example of multiple structures from a different perspective are RNA families: family members share functional and structural properties, which allow them to be studied as a whole, while facilitating both bioinformatics and experimental characterization. Side-by-side or multiple graphics or displays may be an option. However, in the case of RNA families, we might also ask: what is a visual unit for RNA families? (3) Even though RNA usually occurs in complexes with other RNAs or proteins, RNA-specific tools are not yet able to handle such complexes. RNA researchers could easily use standard molecular graphics tools to view such complexes, but this currently means losing RNA-specific features and representations, such as the Leontis-Westhof nomenclature diagrams.

The various needs for RNA-specific representations depend on the users' level of knowledge of structures and their respective tasks, i.e. an understanding of what is important in the case of a specific RNA, for example, depending on the size of the RNA, experimental conditions and insertion points of RNA in biological networks. This level of knowledge may vary considerably between users, e.g. between experimentalists and bioinformaticians. As graphics hardware is much faster today, options are far less limited and acceptable interactive performance is possible. However, there is also a need for new mathematical models. A bidirectional process is at work here, where visualization can lead to new methods and models and vice versa: a new mathematical model allows us to visualize a single structural aspect such as RNA complexes. Indeed, the question of what can be visualized depends on the definition of structure.

[Figure 21 here]

While alignments and structure visualization of RNA are being dealt with extensively, not much progress is being made with respect to the visualization of RNA phylogenies. This is of course an important issue in its own right (see Chapter 20), but phylogenetic visualization tools currently disregard RNA-specific aspects. However, various methods exist in combination with the visualization tools presented in this chapter. For example, in order to define a structure from a phylogenetic viewpoint a simulation programme was developed (Gesell and von Haeseler 2006) that evolves an RNA sequence along a phylogenetic tree and simulates an alignment. Although the software works on the command line, it can display the evolution of the RNA structure along a phylogenetic tree in combination with some of the visualization tools introduced in this chapter. Possible applications include visualizing the outcome of prediction methods along evolutionary time (see Figure 21), or visualizing the stability of RNAs and lineage specific structures in future research. In general, however, there is still a lack of interactive visual explorations that take sequence alignments, structures and phylogenetic trees into account. Also, from a phylogenetic point of view, the advanced visualization of substitution

matrices including phylogenetic trees may be another appropriate representation for underlining specific aspects of RNA structures. As phylogenetic trees are mathematical concepts (Chapter 20), a molecular image is an artificial representation of the molecule. It captures certain aspects of the structure such as physical properties or, more precisely, the definition of structure given a model under user-specific aspects.

While from a hardware point of view such high-end rendering methods nowadays allow for 3D photorealistic graphical techniques and animations, these methods do not always amount to the best choice, especially in the case of data analysis (Chapter 2). Following Ockham's parsimony principle, Edward Tufte asks the following question: What level of complexity is needed to get the message across? Simplification in molecular representations is important as long as no important information is lost. Biologists nevertheless often stick to "what RNA looks like". Independently of these "personal" opinions, the validation of different visualization tools shall demonstrate suitable methods for future research to RNA researchers. Indeed, the process of validating and evaluating the visualization tool is a part of the visualization process itself. What is true for experimental biology clearly also holds in the field of computational biology and visualization for data analysis. The value of even the most sophisticated algorithm and the most beautiful visualization remains unclear if the significance of the results cannot be assessed properly. For the validation of visualization tools we refer to Chapter 2. Here, as an example, experimental validation can prove the prediction of ncRNAs genes at a genome wide level and indirectly validate visualization tools. Comparative genome analysis is currently a widely used strategy for detecting and annotating ncRNAs. However, *de novo* detection of functional RNA structures, or even RNA genes, is still an ill-defined problem. Statistical analysis combined with visualization tools therefore constitutes a promising method for finding ncRNAs genes. Both novel evolutionary signals associated with ncRNAs such as mutation-rate asymmetries and intersections of predicted structures with transcriptomics data (Chapter 8), promoter/terminator signals or new histone-modification patterns (Chapter 6) add reliability to predictions. In addition to highlighting conserved sequence and structure motifs, visualization tools should highlight, sort and filter specific additional information to the eyes of experts. Macro and micro resolution as well as genome-structure displays will be necessary to fulfill these tasks aimed at improving comparative genomic ncRNAs screens.

In sum, while the importance of visualizations to the study of RNA architecture is recognized in the field and a number of tools already exist, there remains an immense potential and need for autonomous new RNA-specific tools to place the incoming amounts of data in context and nurture novel approaches in theoretical biology. To this purpose, connections need be established between RNA structure visualization and many other subfields of biological visualization, such as those presenting transcriptomics data (described in Chapters 7 and 8), genome data (Chapter 5) and the new field of epigenetics (Chapter 6), phylogenetic studies (Chapter 20) and, last but not least, other complexes such as protein structure (Chapter 10) and ligand binding sites (Chapter 11). As graphics hardware is much faster today, options are far less limited and acceptable interactive performance is possible. Moreover, designing effective visual encodings for our purpose requires a focus on scientific questions through the characterization of visualization systems and principles as described in Chapters 1 and 2. Further validation of both computational methods and visualization tools will be important for

future research. As discussions of the role played by RNA in the cell at present resemble a never-ending story, specific RNA sequence and structure visualization tools will be essential for this class of molecules – a class that has long been underestimated.

9.6 Acknowledgements

The authors wish to express their gratitude to Zasha Weinberg for his help with the data processing using R2R software. T.G. is funded by a fellowship of the Austrian genome research program GEN-AU and the GEN-AU project “Bioinformatics Integration Network III"

9.7 Tables

Name	Data type	Scope	Description	File formats	#Entries[1]	URL
PDB	All-atoms	General	RCSB Protein Data Bank – Global repository for 3D molecular models	PDB	~1,900 models	http://www.pdb.org
NDB	All-atoms, Secondary structures	General	Nucleic Acids Database – Nucleic acids models and structural annotations.	PDB, RNAML	~2,000 models	http://bit.ly/rna-ndb
RFAM	Alignments, Secondary structures[3]	General	RNA FAMilies – Multiple alignments of RNA as functional families. Features consensus secondary structures that are either predicted and/or manually curated.	STOCKHOLM, FASTA	~1,973 Alignments/consensus structures, 2,756,313 sequences	http://bit.ly/rfam-db
STRAND	Secondary structures	General	The RNA secondary STRucture and statistical ANalysis Database – Filtered merge of the secondary structures from 8 databases (incl.PDB, NDB, RFAM, RFAM, CRW…)	CT, BPSEQ, RNAML, FASTA, Vienna	4,666 structures	http://bit.ly/sstrand
PseudoBase(++)	Secondary structures	Pseudo knotted RNAs	PseudoBase – Secondary structure of known pseudonotted RNAs.	Extended Vienna RNA	359 structures	http://bit.ly/pkbase http://bit.ly/pkbaseplus
CRW	Sequence alignments, Secondary structures	Ribosomal RNAs, Introns	Comparative RNA Web Site – Manually curated alignments and statistics of ribosomal RNAs.	FASTA, ALN, BPSEQ	1,109 structures, 91,877 sequences	http://bit.ly/crw-rna
tRNAdb	Secondary structures	tRNAs	Transfer RNA database, featuring tRNA genes and sequences.	FASTA, Vienna RNA	12,554 structures	http://bit.ly/trnadb
miRBase	Sequences, Secondary structures[3]	miRNAs	Micro RNA database – published miRNA sequences and annotations	FASTA	16,772 structures	http://www.mirbase.org
RNase-P	Alignments, Secondary structures	RNAse-P	The RNAse P database – RNAse P sequences and structures established through comparative methods	CT, RNAML	521 structures	http://bit.ly/rnasepdb

IRESite	Secondary structures	IRES	The database of experimentally verified Internal Ribosome Entry Sites	FASTA	206 structures	http://iresite.org
snoRNA-LBME-db	Sequences	snRNAs	Small Nucleolar RNAs database.	FASTA	~400 sequences	http://bit.ly/snornas
tmRDB	Secondary structures	tmRNAs	tmRNA Database.	FASTA, CT	729 structures	http://bit.ly/tmrnadb

Table 1. Major RNA databases.
These sites are the main repositories of RNA sequence and structure data. In the absence of a single authoritative source for the secondary structure of RNA, multiple sources must be queried, depending on the functional family of interest.
[1] Statistics compiled in July 2011, only includes RNA –related data.
[2] PseudoBase++ is a user-friendly front-end for the HTML-formatted Pseudobase.
[3] Predicted structures
Citations: PDB (Berman, Westbrook et al. 2000); NDB (Berman, Olson et al. 1992); RFAM (Gardner, Daub et al. 2011); STRAND (Andronescu, Bereg et al. 2008); PseudoBase (van Batenburg, Gultyaev et al. 2001)/PseudoBase++ (Taufer, Licon et al. 2009); CRW (Cannone, Subramanian et al. 2002); tRNAdb (Juhling, Morl et al. 2009); miRBase (Kozomara and Griffiths-Jones 2011); RNAse-P (Brown 1999); IRESite (Mokrejs, Masek et al. 2010); snoRNA-LBME-db (Lestrade and Weber 2006); tmRDB (Zwieb, Gorodkin et al. 2003).

Name	Cost	Web	Win	Mac	Linux	Description	Ease of Use	Layouts	Interactivity	Output Formats	Pseudoknots	Non-canonical basepairs	Multiple structures	URL
PseudoViewer		•	•			A non-interactive web server/service especially fit for the visualization of pseudoknotted structures.	++	Graph	None	**EPS**, **SVG**, PNG, GIF	++			http://bit.ly/pseudoviewer
RNA2DMap		•	•	•	•	Flash application which maps structural and comparative data onto static drawings of ribosomal RNAs.	+	Graph	Annotations	**PDF**[2]	+	•		http://bit.ly/rna2dmap
RNAMovies		•	•	•	•	Java software that animates the transition between multiple secondary structures.	++	Graph	None	**SVG**, PNG, JPEG, GIF	+		•	http://bit.ly/rnamovies
RNAPLOT		•[1]	•	•	•	Simple command-line tool for drawing RNA secondary structure	+	Graph	None	**PS**, **SVG**, GML, XRNA				http://bit.ly/vienna-rna

RNAViz			•	•	•	Interactive software for the semi-automated production of publication-quality drawings.	+	Graph	Edition/Annotation	**PDF**[2]	+			http://bit.ly/rnaviz
R2R			•	•	•	Comprehensive command-line tool for the production of consensus diagrams.	-	Graph	None	**PDF**, **SVG**	+			http://bit.ly/r2r-soft
S2S/Assemble			•	•	•	Two suites of web services-based tools for the 3D modeling of RNA from sequence and 2^{ary} structure data.	+	Linear, Graph	Edition/Annotation	**SVG**	+	•		http://bit.ly/s2s-soft
VARNA		•	•	•	•	A visualization tool, initially designed as a Java webserver companion, coupled with a simple user-interface, and offering annotation and editing features.	++	Circular, Linear, Graph	Edition/Annotation	**EPS**, **SVG**, PNG, JPEG, GIF	+	•	•	http://varna.lri.fr
xRNA			•	•	•	Rich editing features	+	Graph	Edition/Annotation	EPS	+		•	http://bit.ly/xrna-soft

Table 2. RNA secondary structure visualization tools.
Such software can be used for the visualization and export of RNA secondary structure. Bold output formats indicate vector graphics, allowing for convenient post-processing (e.g. using ADOBE illustrator or Inkscape), and virtually unlimited resolution rendering (relevant to the production of publication-quality illustrations).

[1] Only available from the Vienna RNA webserver, available at http://rna.tbi.univie.ac.at/

[2] Export accessible through a *print* option, using custom drivers such as Adobe Distiller.

Citations : VARNA (Darty, Denise et al. 2009), Pseudoviewer (Byun and Han 2009), S2S/Assemble (Jossinet and Westhof 2005; Jossinet, Ludwig et al. 2010), RNAMovies (Kaiser, Kruger et al. 2007), RNAPLOT/Vienna package (Gruber, Lorenz et al. 2008; Hofacker 2009), RNAViz (De Rijk, Wuyts et al. 2003), R2R (Weinberg and Breaker 2011).

Name	**Cost**	**Web**	**Win**	**Mac**	**Linux**	**ease-of-use**	**Description**	**Input/ Output Formats**	**Displays secondary structure**	**Base Pairing Probabilities**	**3D model**	**Tertiary interactions/ Isostericity**	**Calculates covariation score**	**Calculates consensus structure**	**URL**
BoulderAle		•				++	Boulder ALignment Editor (ALE) is designed for editing and assessing alignments, based on isostericity of Watson-Crick and non-Watson-Crick base pairs.	FASTA, Stockholm			•	•			http://bit.ly/boulderale
ConStruct				•	•	+	Semi-automated tool for creating RNA alignments correct in terms of consensus sequence and consensus structure	FASTA, Vienna, Stockholm, Output Cons. Struct.: Rnaml, Stockholm	•	•		•	•	•	http://bit.ly/construct3
JALView		•	•	•	•	+	JALView aligns sequences using Web Services (Clustal, Muscle, MAFFT...), has sorting options (by name, tree order, percent identity, group). Calculates trees based on percent identity distances.	Fasta, PFAM, MSF, Clustal, BLC, PIR	•		•			•	http://www.jalview.org/

RALEE			•	•	•	+	The RALEE (RNA ALignment Editor in Emacs) tool provides a simple environment for alignment editing, including structure-specific color schemes.	Unblocked Stockholm	•					•	http://bit.ly/RNAralee
SARSE				•	•	+	Semi-automated RNA sequence editor (SARSE). Divides the sequences into subgroups with secondary structure differences.	Input: COL, txt Output: COL, txt, FASTA	One structure per sub group					One structure per sub group	http://sarse.ku.dk/
4SALE			•	•	•	++	4SALE is designed to handle sequence and secondary structure information of RNAs synchronously.	FASTA, Output Secondary structure: SVG	•				•		http://bit.ly/rna4sale

Table 3. Structure-aware RNA sequence alignment editors.

9.8 Figures

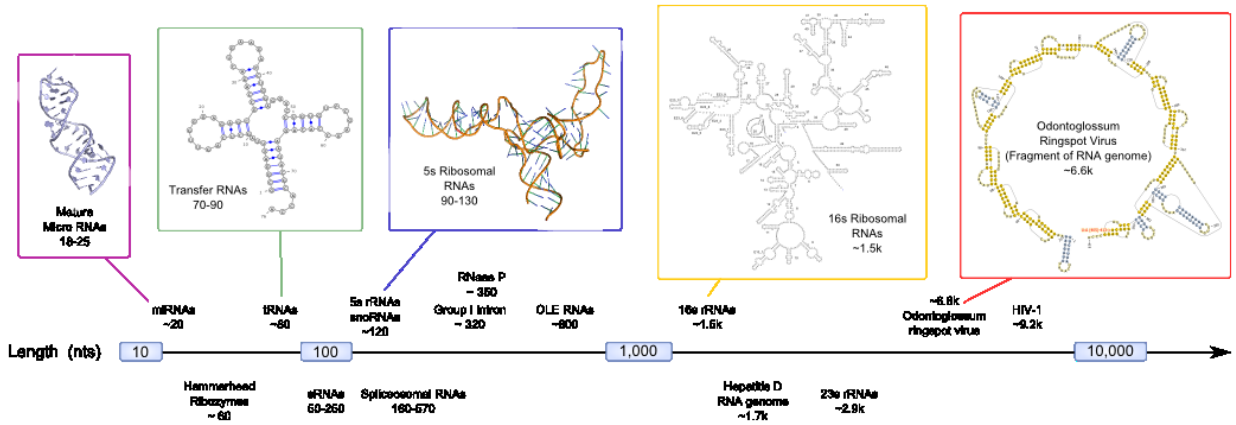

Figure 1. Diversity of scales for non protein-coding RNAs.
With lengths ranging from 20 to 10 000 nucleotides, structurally-resolved RNAs require a diversity of visualization techniques.

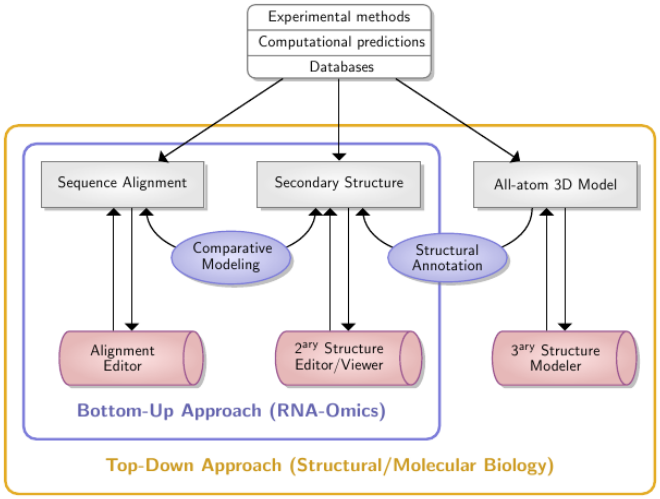

Figure 2. RNA annotation methods, relying on visualization for an iterative refinement of models and data.

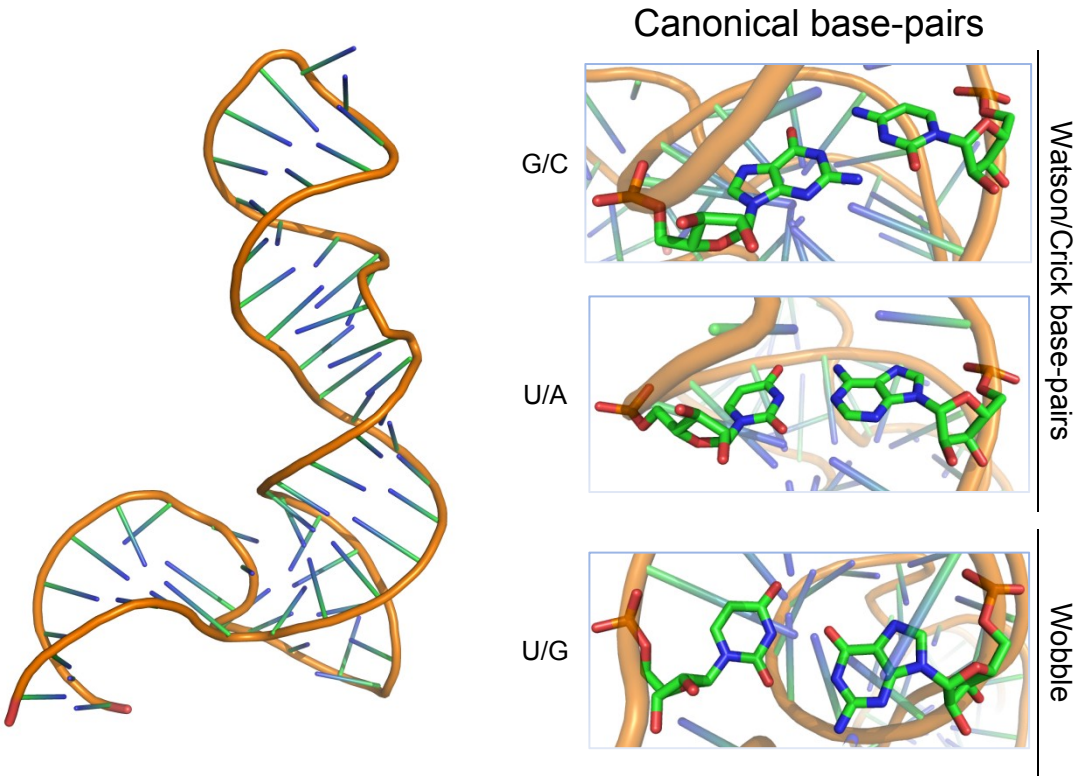

Figure 3. *Cartoon* view, using PyMol, of a 5s ribosomal RNA (PDBID: 1UN6, Left) and a focus on some exemplary canonical base-pairs (Right).
PyMol's *RNA-specific* implementation of the *cartoon* view, in combination with a transparency effect and a classic stick representation, enables a clear focus on local phenomena of interest, such as base-pairs or tertiary motifs.

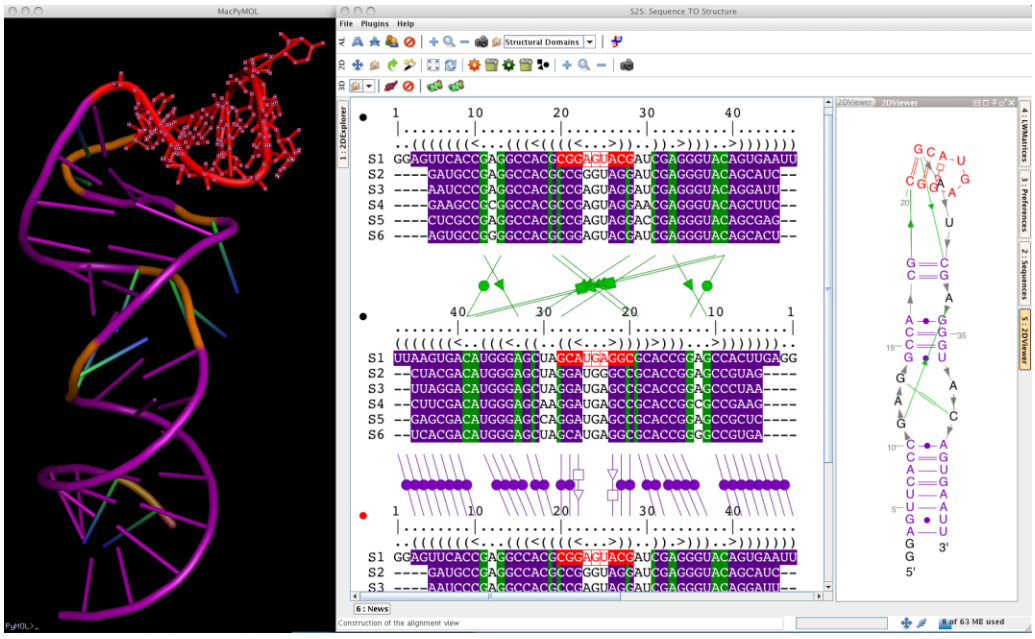

Figure 4. Tertiary annotation of an experimentally-derived 3D model using S2S and PyMol.

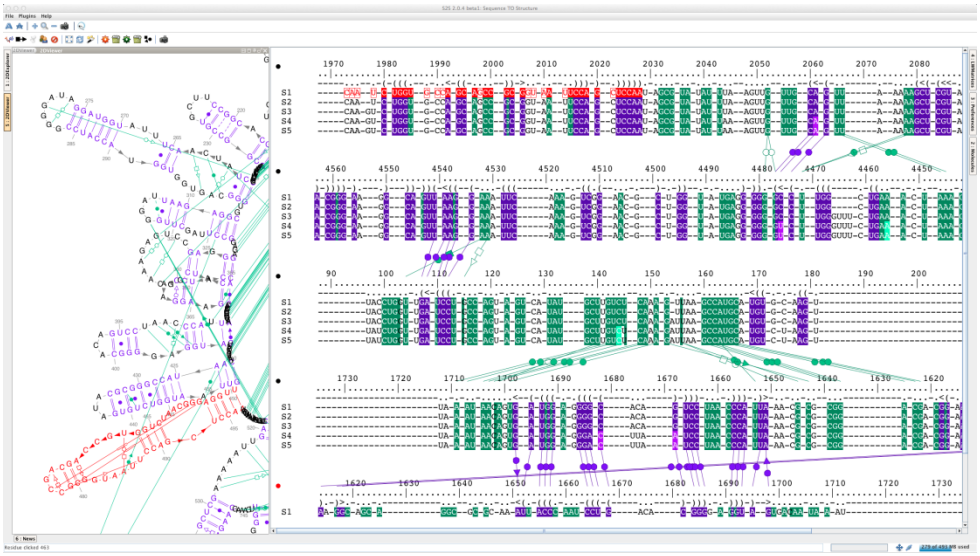

Figure 5. Screen capture of the S2S tool rendering a structural alignment of eukaryotic 18S ribosomal RNAs.

The structure of Triticum aestivum (PDBID : 3IZ7) has been used as the reference structure. It is displayed as the S1 sequence in the alignment panel and the 2D panel on the left renders its “extended secondary structure”. Several views of the same alignment have been produced. Between each view, secondary interactions (in purple) and tertiary interactions (in green) are displayed.

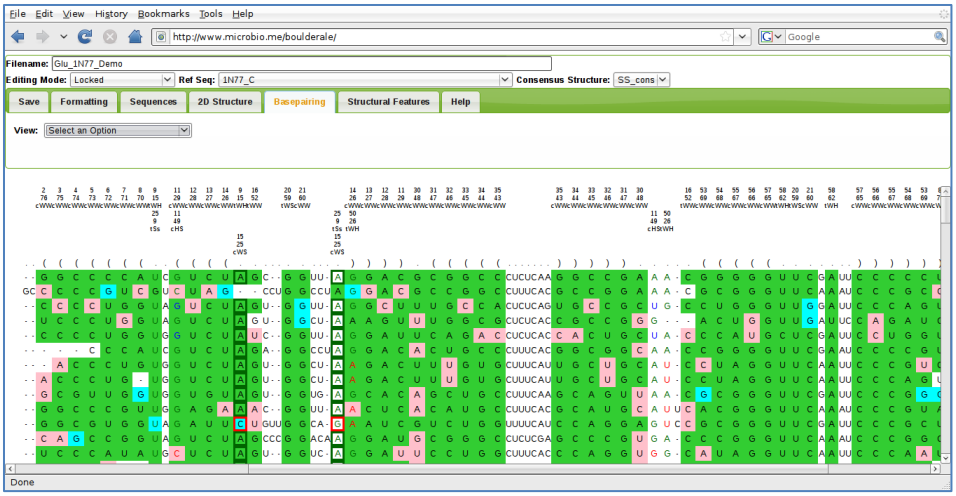

Figure 6. Isostericity-aware multiple alignment of RNA sequences using BoulderAle web tool.

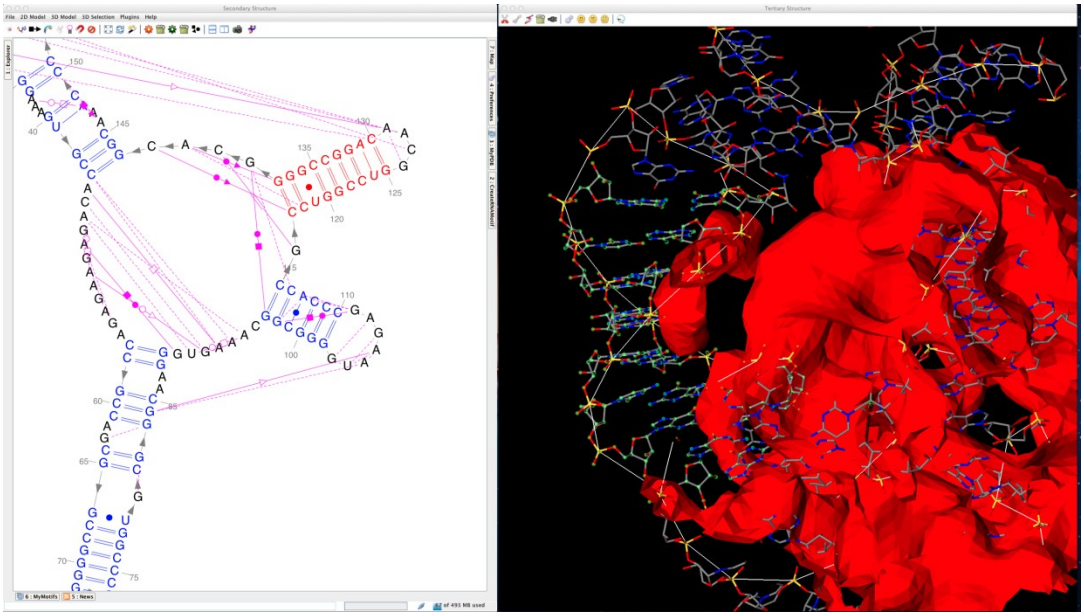

Figure 7. Screen capture of the Assemble tool rendering the solved structure of the ribonuclease P of the A-type (PDBID : 1U9S) along with a cryo-EM density map.
An « extended secondary structure » is automatically computed using the RNAVIEW webservice. Synchronized helix selection between the 2D and 3D panels helps cumulate the benefits of both representation levels.

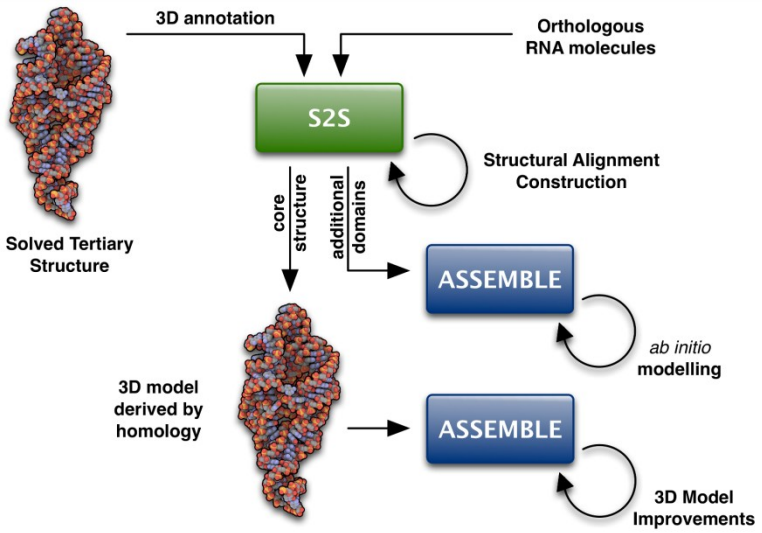

Figure 8. Description of the semi-automatic workflow to model RNA 3D architectures by combining S2S and Assemble.
Starting from a solved tertiary structure, S2S allows to align orthologous sequences against a reference structure. The core structure can then be derived and curated with Assemble (homology modeling). Any additional domain identified during the construction of the structural alignment can be modeled *de novo* using Assemble and its embedded structural database.

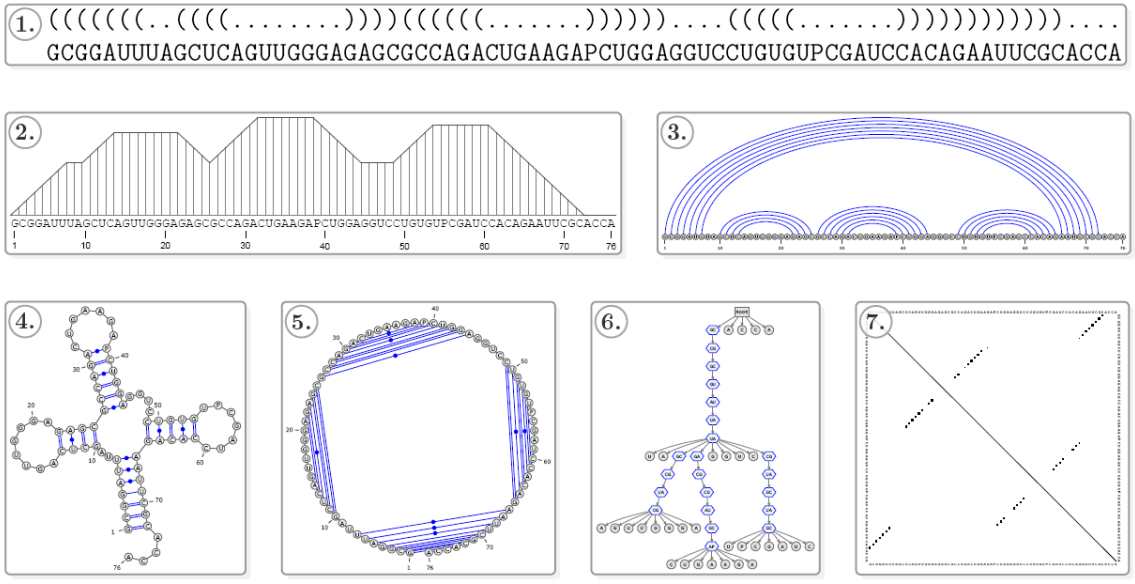

Figure 9. Main representations of RNA secondary structure.
The structure of a typical transfer RNA secondary structure, denoted by a well-parenthesized expression **(1)**, and drawn as: a mountain plot **(2)**, a linear arc-annotated sequence **(3)**, an outer-planar graph **(4)**, a circular Feynman diagram **(5)**, a tree **(6)** and a dot-plot **(7)**. **(3)**, **(4)** and **(5)** generated by VARNA (Darty, Denise et al. 2009), **(7)** generated by RNAFold (Gruber, Lorenz et al. 2008).

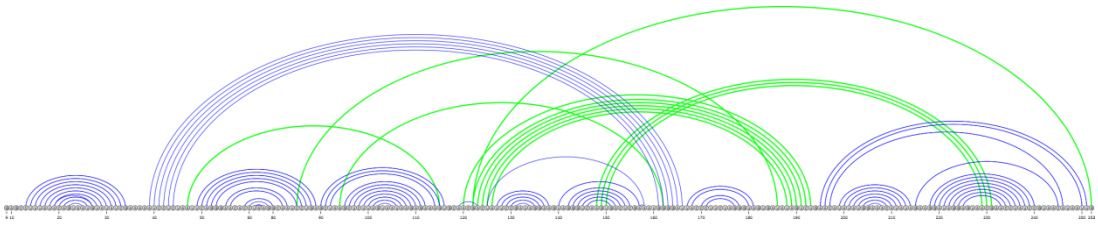

Figure 10. Linear arc-diagram of a pseudoknoted RNA secondary structure.
Secondary structure featuring pseudoknots (green arcs) of a ribozyme (Group I intron - Staphylococcus phage twort - PDBID 1Y0Q:A), inferred using RNAView (Yang, Jossinet et al. 2003) and drawn using VARNA (Darty, Denise et al. 2009).

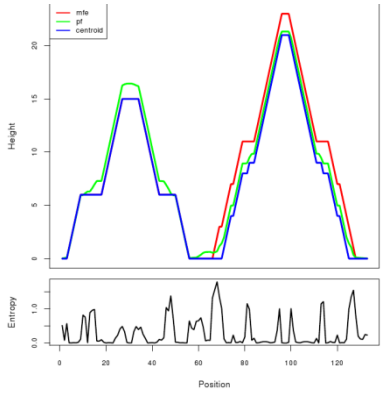

Figure 11. Mountain-view representation amplifies subtle local differences.
This representation helps in studying the divergence of a predicted structure from the average structure in the Boltzmann ensemble. Local divergence may have a cascading effects, increasing the height of local substructures, but their overall shape will remain unchanged, allowing for a quick identification. Image produced by the Vienna RNA websuite (Gruber, Lorenz et al. 2008) from an *E. caballus* snoRNA sequence (RFAM ID: RF00265).

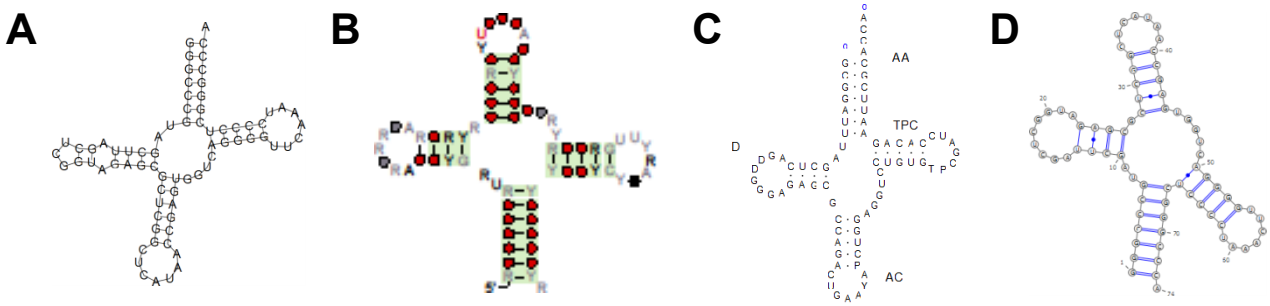

Figure 12. Various graph representations for tRNA secondary structure.
Produced by RNAPlot (**A**, Default setting) (Gruber, Lorenz et al. 2008), R2R (**B**, Template) (Weinberg and Breaker 2011), RNAViz (**C**, Template) (De Rijk, Wuyts et al. 2003), and VARNA (**D**, Radiate algorithm) (Darty, Denise et al. 2009). Source: RFAM seed alignment for the tRNA family (**A**, **B**, **D**, RFAM ID: RF00005) and Yeast Phenylalanine tRNA (**C**).

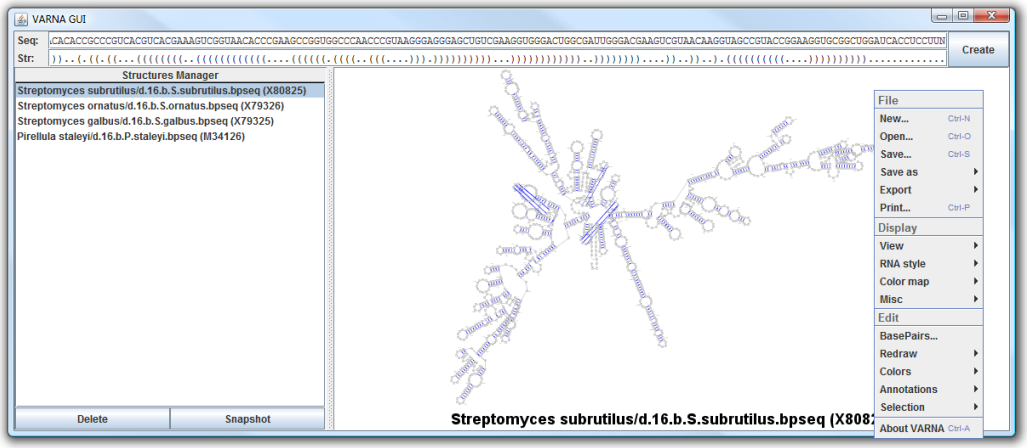

Figure 13. VARNA's minimal graphical user interface.
Numerous functionalities can be accessed through a pop-up menu (Right). Four 16s ribosomal RNAs excerpted from the CRW database (Cannone, Subramanian et al. 2002) are quickly loaded through drag-and-drop gestures, and drawn using the NAView algorithm (Bruccoleri and Heinrich 1988).

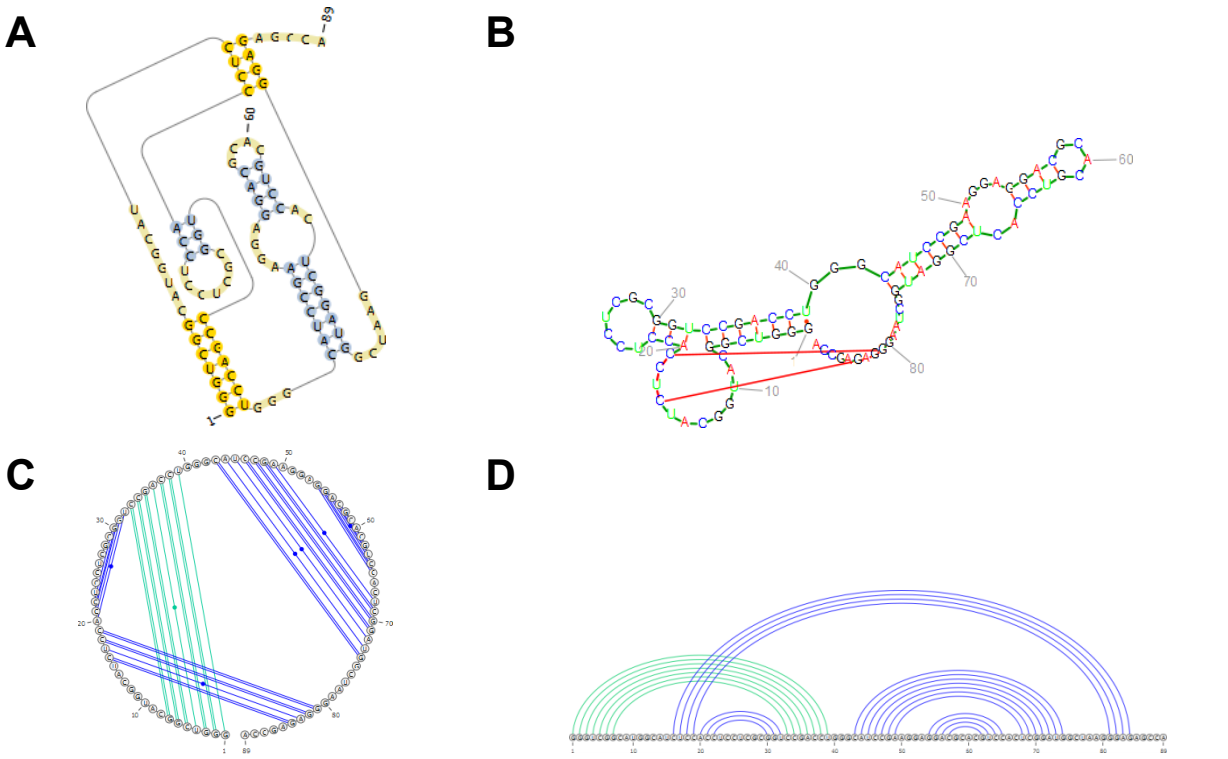

Figure 14. Automated drawing of an H-type pseudoknot.
(A) PseudoViewer 3 (Byun and Han 2009) proposes an elegant planar graph layout for pseudoknots. **(B) RNAMovies** supports pseudoknots by first extracting a maximal non-crossing subset of base-pairs, and adding pseudoknoted base-pairs afterward. **(C)** & **(D)** The circular and linear representations (created here by **VARNA**) remain largely unaffected by the presence of

pseudoknots. Source: Hepatitis delta virus (*Italy* variant), PseudoBase ID: PKB76 (van Batenburg, Gultyaev et al. 2001).

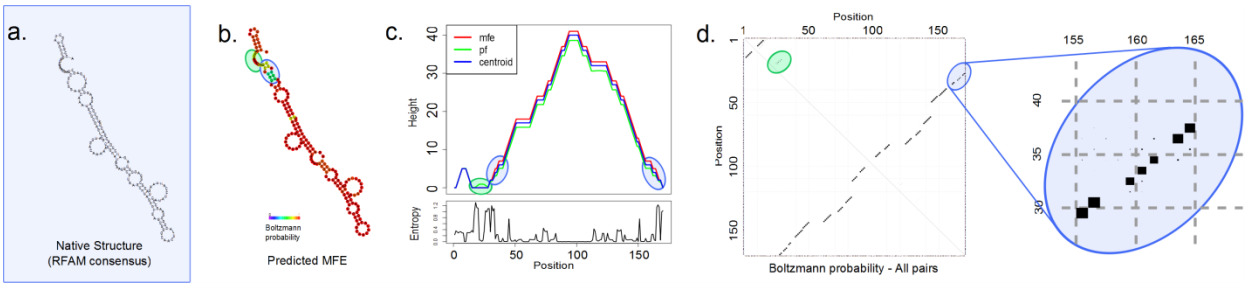

Figure 15. Visualizing the reliability of predictive methods.
RFAM-derived native secondary structure (**a.**, RFAM ID: RF02001) and reliability-annotated Minimum Free-Energy (MFE) prediction (**b.**, **c.** and **d.**), using RNAFold (Gruber, Lorenz et al. 2008), for the D1-D4 domain of a group II catalytic intron in *A. capsulatum*. **b.** Predicted secondary structure, drawn as a graph and color-annotated with the Boltzmann probability of the predicted base-pairing status. **c.** Top: Joint mountain plot of the MFE structure (red line), average structure in the Boltzmann ensemble (green line) and centroid structure (blue line) obtained through statistical sampling (Ding, Chan et al. 2005). Bottom: Positional entropy, or variability of base-pairing in the Boltzmann equilibrium. **d.** Base-pairing matrix (dot-plot), showing the contacts induced by the MFE structure (Lower-left triangular part) against the base-pairing probabilities (Upper-right triangular part).

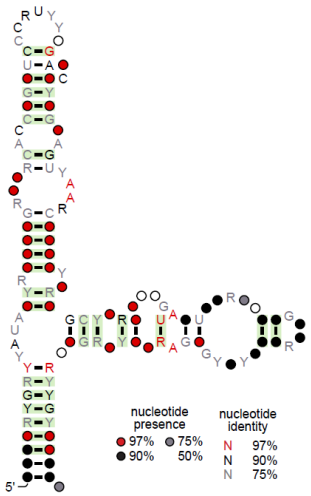

Figure 16. Semi-automated production of publication-quality consensus diagrams using R2R.
Source: RFAM 5s ribosomal RNA seed alignment (RFAM ID: RF00001).

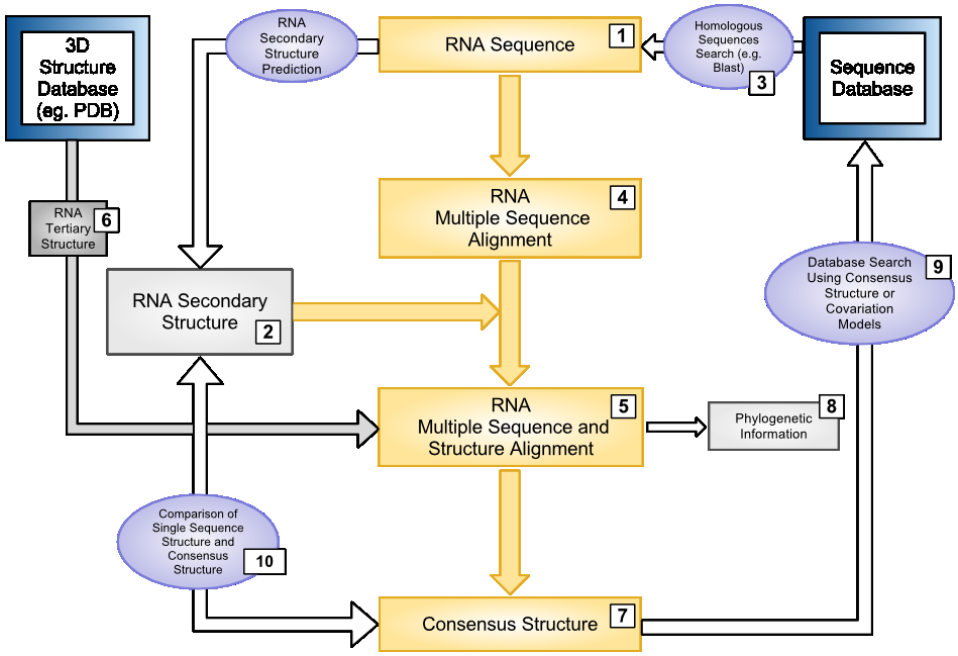

Figure 17. Iterative refinement of RNA structural alignment.

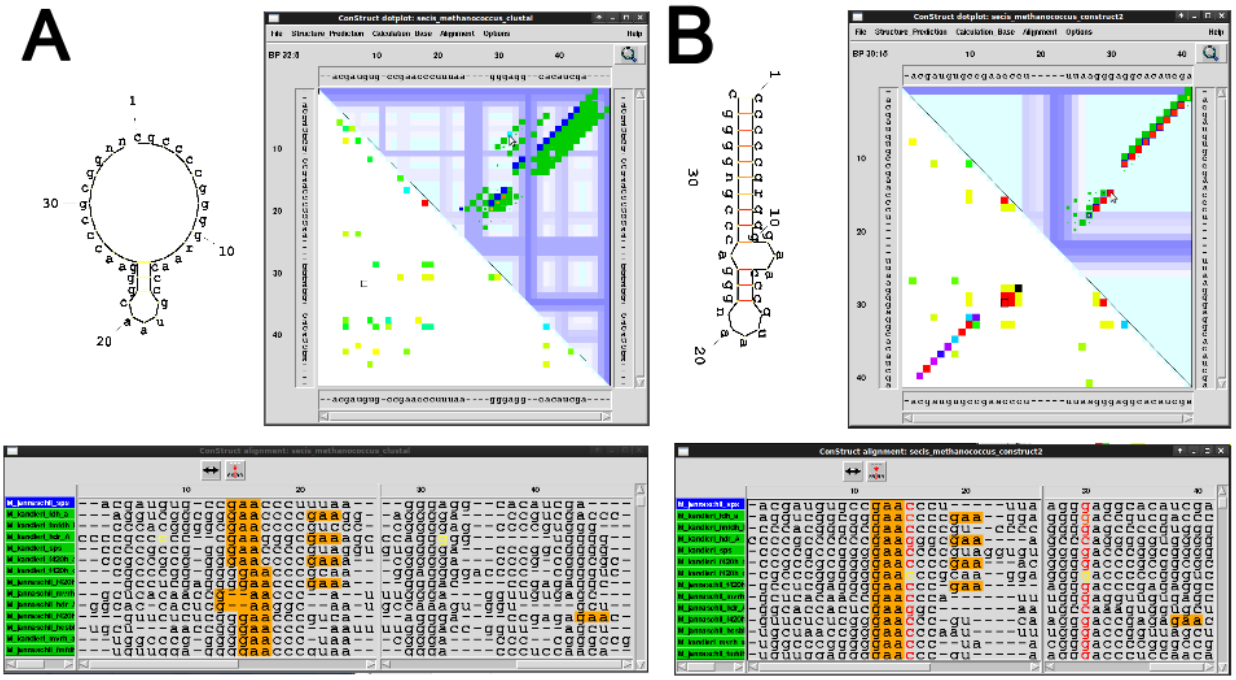

Figure 18. Visualization of alignments by ConStruct.
An alignment of SECIS elements created by CLUSTALW (A) and after manual optimization/correction using CONSTRUCT (B). In both cases predicted consensus structures

and CONSTRUCT's GUI are shown. Top left: Corresponding drawings of consensus structures (annotated with the consensus sequence) generated by CONSTRUCT; consensus base pairing probability is color-coded from white to red. Top right: Corresponding dotplots: the base pairing probability of individual sequences (dark blue for the selected sequence M_janaschii_sps and green for others) is shown top-right in CONSTRUCT's main window; yellow to red dots show the consensus pairing probability; white to light blue bars denote gaps. The lower-left triangle shows the MI (Mutual Information index) in rainbow-colors from yellow to red. Bottom: Corresponding alignment windows. Nucleotides participating in a base pair to which the cursor points in the dotplot are automatically highlighted [colored by pairing probability from p = 0 (black) to p = 1 (red)]. The motif GAA (orange background), which is conserved in the internal loop, has been highlighted using the built-in regular expression search.

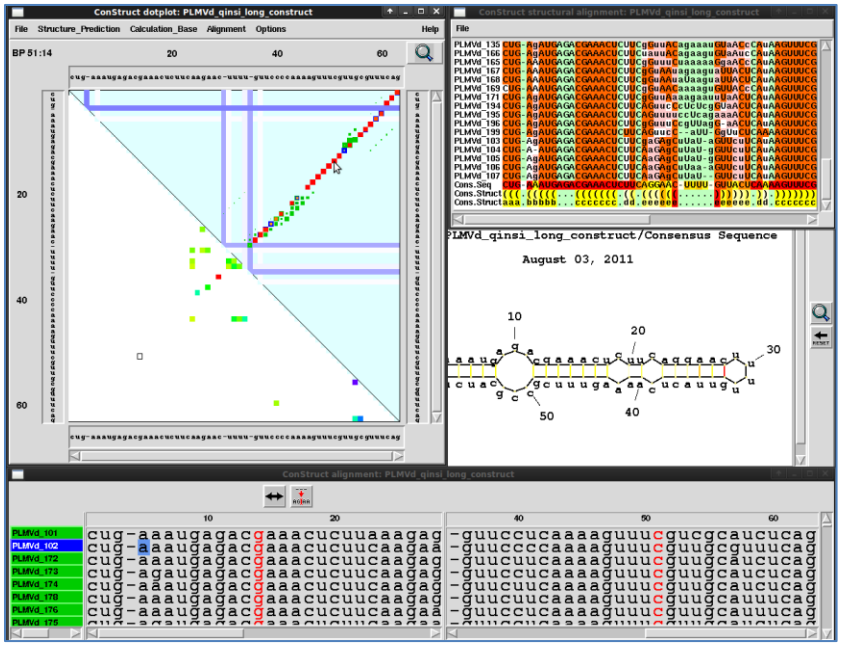

Figure 19. Synchronized view of secondary structure, base pair probabilities and sequence alignment using ConStruct.

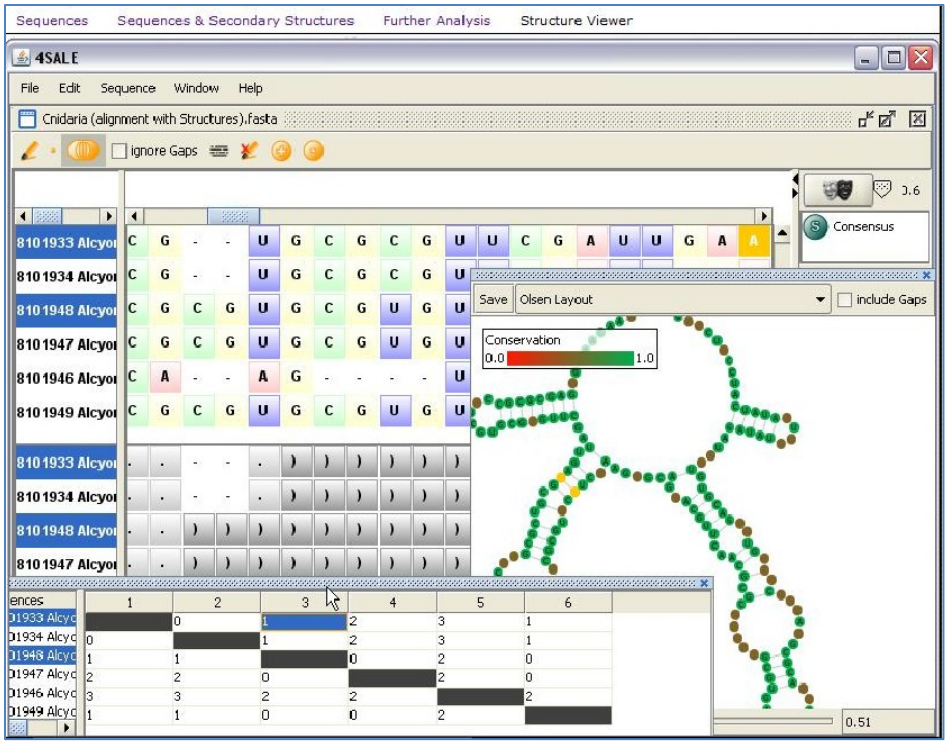

Figure 20. Realignment of structure/sequences using 4Sale.

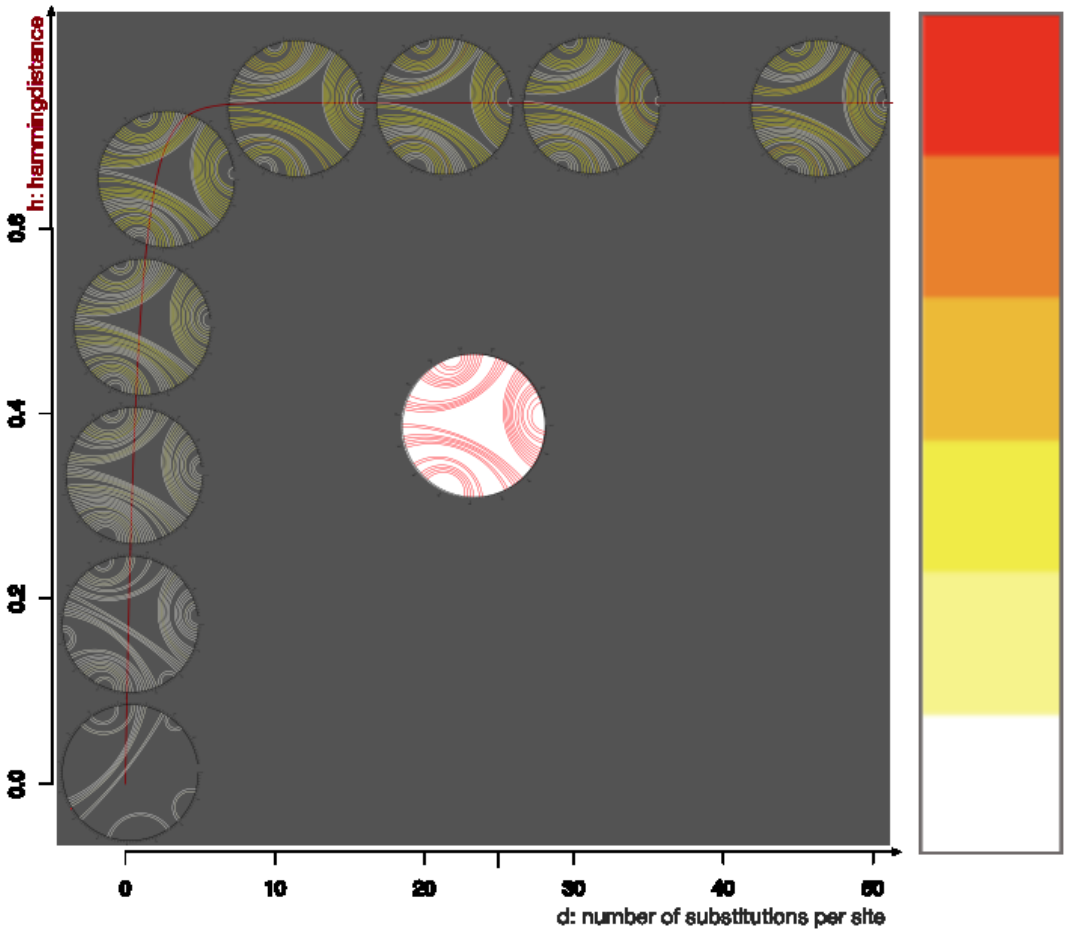

Figure 21. Visualizing simulated alignments along star trees under the constraints of archebacteria 5s RNA using the command line tool SISSI and the visualization tool ConStruct.

Color code is given by ConStruct for the base pair probability from white to yellow to red with the highest probability (right). The red line describes the Hamming distance *h* (y-axis) as a function of the genetic distance *d*, measured in number of substitutions per site (x-axis). Using structure prediction with mutual information content and starting with very little information the prediction becomes increasingly similar to the underlying 5s RNA constraint in the middle of the figure.

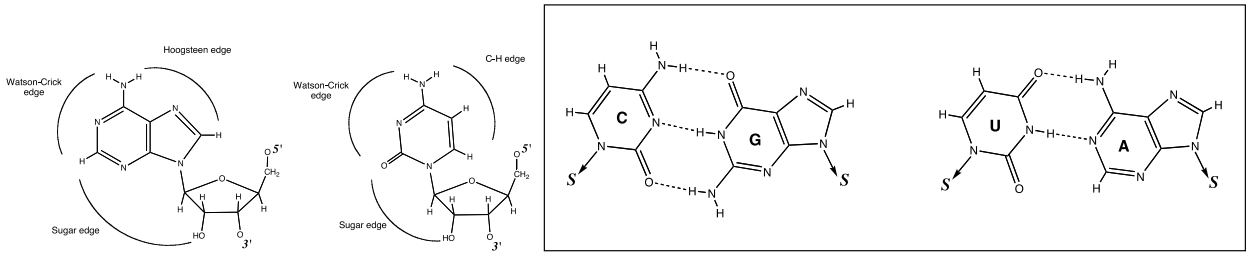

Figure 22. The three edges of nucleic acid bases involved in edge-to-edge interactions mediated by hydrogen bonding (Left), and the canonical Watson-Crick C=G and U-A base pairs (Right).

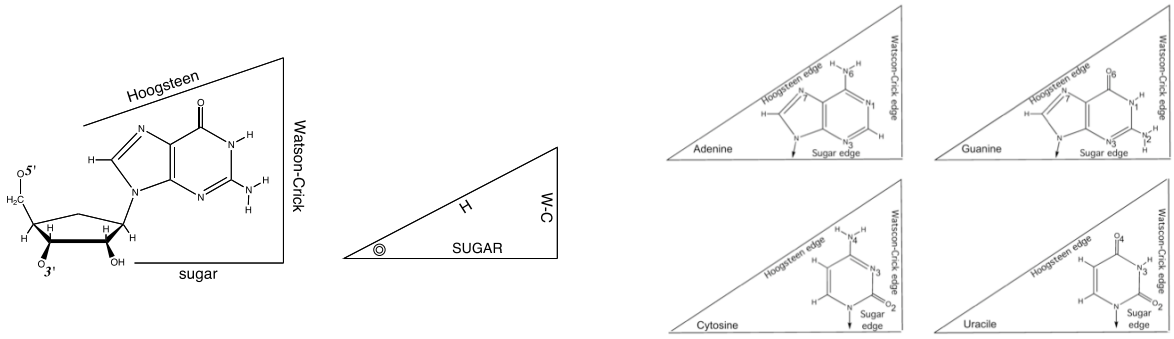

Figure 23. Abstraction of nucleic acid bases as triangles.

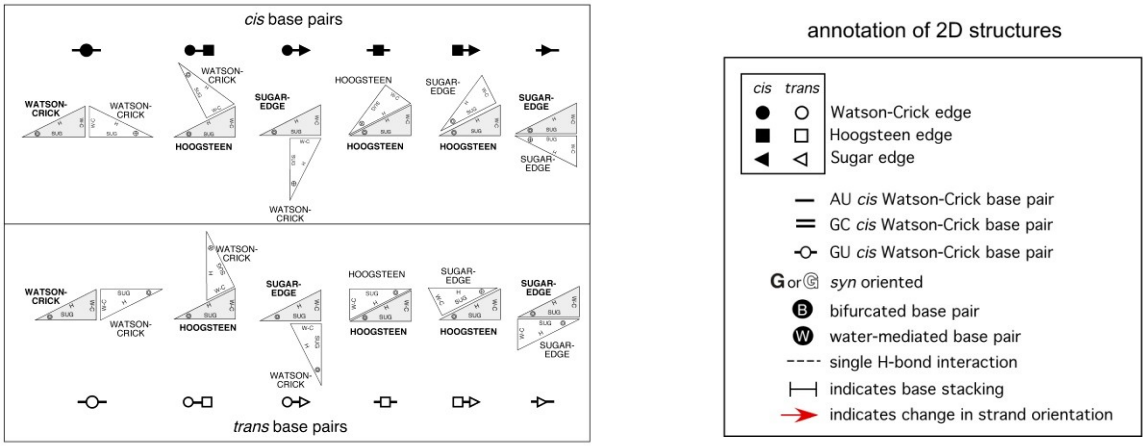

Figure 24. The twelve possible base-pairing geometries (Left), and symbols for three-dimensional structural features in 2D representations of RNA structures (Right).

9.9 Display boxes

Display Box: RNA folding: A primer

Generalities. RNA is a macromolecule analogous to a chain of nucleotides, each supporting one of the **(nucleo)bases**: **A**denine, **U**racil, **G**uanine and **C**ytosine. Like DNA, the sequential nature of RNA allows for an abstraction of each molecule as a **sequence** over the alphabet A, C, G, and U (possibly extended to include non-conventional or modified bases). However, unlike DNA, RNA is single stranded and **folds** on itself, allowing for the formation of one or several complex three-dimensional structures which are stabilized by the interaction of some of its nucleotides.

One of the driving force of RNA folding is the process of **base pairing**, the establishment of hydrogen bonds between the atoms of bases at arbitrary sequential distance. Such a **base pair** involves, on both partners, one out of three possible edges and may be subject to diverse relative orientations, as represented schematically in Figure 22. Due to their stability and ubiquity, A-U, G-C and G-U base pairs involving Watson-Crick edges in *cis* orientation (see Figure 24) have drawn the attention of earlier studies of RNA studies, and are usually referred to as **canonical** base pairs. However the importance of non-canonical base pairs is increasingly acknowledged and their incorporation in predictive models has been a source of substantial improvement (Parisien and Major 2008).

RNA structure(s). The **secondary structure** of an RNA structure consists in a (sub)set of its base pairs. The precise definition of the secondary structure is, however, quite flexible and may either refer to canonical base pairs only, or preclude complex topological features called pseudoknots.

Besides the secondary structure, which only provides partial structural information, **all-atoms three-dimensional models** can be obtained. Such models typically specify the relative positions of every atom, as derived from experimental methods, possibly in conjunction with experimental parameters and measures like the model resolution. Tertiary motifs are modular elements in RNA in which base-pair patterns are associated with definite tertiary organization, and therefore constitute a bridge between the secondary and tertiary structure.

Experimental techniques. All atoms models of RNA molecules can be produced using either **NRM spectroscopy** or **X-ray crystallography** techniques. Typical resolutions of less than 3 Å are currently reported for more than 60% of the structures deposited in the PDB. Quite remarkably, advances in multidimensional NMR have recent made possible a *real-time* study of RNA kinetics (Lee, Gal et al. 2010).

Much information about the fold of a RNA sequence can also be gathered using **chemical and enzymatic probing**. In chemical probing techniques, the accessibility and reactivity of some atomic positions (e.g. the N7 of adenine or the N3 of cytosine) can be assessed in various conditions (*in vitro* or *in vivo*). In enzymatic probing techniques, accessible phosphodiester bonds either in single-stranded or double-stranded regions are cleaved using specific enzymes. This information can be used, possibly in combination with computational methods, to perform high-throughput secondary determination possible at a genomic scale (Watts, Dang et al. 2009) (e.g. HIV-1).

Display Box: Computational methods for RNA structure prediction

Computational methods for the automated prediction of the native structure(s) of RNA typically fall in two categories, depending on the available data. If a single sequence is available, *ab-initio* methods postulate some form of energy model to either recover a minimal free-energy (MFE) structure, or an ensemble of representative methods. When homologous sequences are available, one can assume structure conservation throughout evolution, and use a comparative approach for refined predictions.

Ab-initio single fold methods. Building on nearest neighbor energy rules (Tinoco, Borer et al. 1973), coupled with an universal decomposition of RNA secondary structures into loops, efficient algorithmic solutions were proposed for the problem of predicting the minimal free-energy secondary structure (Zuker and Stiegler 1981).

These approaches are based on an algorithmic technique called Dynamic Programming (DP), which implicitly traverses the ensemble of *all* possible conformations in polynomial time on the length (typically taking a few seconds on a laptop computer for RNAs of length 300). From this exploration, existing software either retrieves the most stable structure, a set of suboptimal or reports some statistics of the ensemble. Ensemble methods are additionally able to associate Boltzmann probabilities to individual base pairs, which can either be used to assess the confidence in the prediction or as indicating the structuration (or lack thereof) of transcripts.

The main tools for the *ab initio* prediction are MFold/Unafold (Zuker and Stiegler 1981; Markham and Zuker 2008) and RNAfold (Hofacker 2009), both based on a thermodynamic model. Typical computational ab-initio methods recover about 73% of the base-pairs (Mathews, Disney et al. 2004). The Contrafold software (Do, Woods et al. 2006) is based on an alternative probabilistic model and reports slightly better performances, while losing the mechanistic quality of an energy-based prediction. Recent contributions have also extended existing energy models and algorithms to incorporate non-canonical base-pairs and motifs. MC-Fold (Parisien and Major 2008) uses statistically-derived pseudo-potentials to compute a secondary structure that may possibly include non-canonical base pairs. The resulting set of base pairs can then be used as a scaffold for a three-dimensional modeling, allowing for an *ab-initio* prediction of 3D models, falling within a few Angstroms RMSD of experimental models for RNAs of moderate length (~50nts) RNAs.

Ensemble methods. Given the susceptibility of MFE prediction methods to small variations of the experimentally-derived energy model, earlier studies have considered subsets of suboptimal structures (Zuker 1989; Wuchty, Fontana et al. 1999). The seminal work of McCaskill (McCaskill 1990) furthered these studies by postulating a Boltzmann equilibrium, in which each possible conformation is assigned a probability within a Boltzmann probability distribution. This allowed for an exact computation of base-pairing probabilities, later confirmed as a good indicator of confidence in predicted base-pairs (Mathews 2004), as base-pairs associated with Boltzmann probabilities greater than 0.99 were verified experimentally in 91% of occurrences (Mathews 2004). Clustering was also used to estimate a representative set of structures, improving the specificity of predictions (Ding 2006).

Computational methods for RNA structure prediction (continued)

Comparative methods. The accuracy of *ab-initio* prediction tools is limited by several factors. Simplifications are made in the underlying model: tertiary interactions and kinetic factors are ignored, and the stability of multiple loops is approximated for computational reasons. Furthermore the accuracy of thermodynamic models, whose parameters are partially extrapolated from experimental observation, is intrinsically limited. Prediction accuracy can, however, be increased by taking into account additional information gained from aligned homologous sequences. Namely the structural conservation, at both the secondary and tertiary structure levels, of an ncRNA can be much higher than its sequence conservation. This is due to the fact that the structure of certain ncRNAs is much more constrained by their biological function than their sequence. For instance, RNase P has a sequence identity of only 35 – 55% while secondary structures are very similar (Brown and Pace 1992). Under the hypothesis of a selective pressure towards a given structure, a mutation in a base-paired region will typically be compensated by a further mutation that reestablishes the original pairing scheme, restoring the original shape of the molecule. Such an evolutionary device, called compensatory mutations, can be witnessed from the inspection of multiple RNA sequence alignments.

Consensus structure prediction. A consensus structure represents a common structure for a set of homologous RNAs. A set of RNAs can have more than one consensus structure when its biological function is based on more than one structural conformation (e.g. riboswitches). To find an underlying consensus structure for a set of ncRNAs, the best approach accuracy-wise would be to simultaneously align the sequences and structures of a set of RNAs. This problem was addressed in the infancy of the field (Sankoff 1985), leading to an exact algorithm whose runtime scales like $O(n^{3m})$ for m sequences of length n.

Computational costs are usually specified using the O or Θ notations, which describe the general behaviour, either exactly (Θ) or as an upper-bound (O), for the evolution of time or memory as a function of the sequence length. In these notations, all constant terms are neglected.

For instance, the (true) multiple alignment of m sequences of length n requires an exact effort of $2^m * n^m$, using a classic generalization of the Smith-Waterman algorithm. Hence the computational cost for a multiple alignment will be denoted as either $\Theta(n^m)$, or $O(n^m)$, the former being more precise (exact equivalent) than the latter (upper-bound). If we assume that every step of calculation needs 1 µs, the following computing times and memory usage arise for a (true) multiple alignment of sequences of length 100.

#Sequences	2	4	8
Computation Time	$2^2 * 100^2$ µs = 40 ms	$2^4 * 100^4$ = 1600 s	$2^8 * 100^8 \approx 3 * 10^{12}$ s ≈ 83 years
Memory Usage	≈ 10 kB	10^8 ≈ 95 MB	10^{16} · 2 byte ≈ 18 TB

The prohibitive amount of time and memory required for the alignments of only 8 sequences of length 100 is the main motivation for resorting to approximate algorithmic schemes, such as heuristics.

Because of this prohibitive computational cost, unsuitable for real-world datasets, many practical approaches have been implemented, through a relaxation of the optimization scheme, in acceptable time and memory. However, due to their heuristic nature, these cannot perform equally as well as Sankoff's algorithm. These heuristic approaches can be categorized as follows (Gardner and Giegerich 2004):

- First align sequences then predict common structure. An alignment is computed by either pure sequence or by sequence-structure alignment methods. The initial alignment can be refined using one of the available RNA sequence-structure editors (see Sections 9.3.4 and 9.4.2) (Bindewald and Shapiro 2006; Bernhart, Hofacker et al. 2008; Wilm, Higgins et al. 2008).
- First predict structures then align them (Hochsmann, Voss et al. 2004; Dalli, Wilm et al. 2006; Moretti, Wilm et al. 2008).
- Align and predict structures simultaneously. Despite bearing a prohibitive computational cost, the Sankoff algorithm can still be used to perform pairwise alignments when combined with practical (yet error-prone)

optimizations. (Perriquet, Touzet et al. 2003; Hofacker, Bernhart et al. 2004; Holmes 2005; Yao, Weinberg et al. 2006; Bauer, Klau et al. 2007; Harmanci, Sharma et al. 2007; Kiryu, Tabei et al. 2007; Lindgreen, Gardner et al. 2007; Torarinsson, Havgaard et al. 2007; Will, Reiche et al. 2007; Harmanci, Sharma et al. 2008)

Application domains and performances. While all methods for predicting consensus structures outperform *ab-initio* methods, the third approach is the most accurate. In fact, for a set of sequences with an average pairwise sequence identity (APSI) below 55%, it is the only approach which gives reasonable results, but it is also the most demanding in terms of computing resources. The above classification into categories is, however, not quite distinct. Some heuristics to the Sankoff algorithm (Sankoff 1985), for instance, restrict their search space taking into account a primary sequence alignment. Typical sensitivity/specificity tradeoffs of 80%/80% can be achieved from only two homologous sequences (Gardner and Giegerich 2004) by – computationally intensive – automated comparative methods.

Display box: Leontis-Westhof classification of tertiary interactions and associated schematics.

All-atom models are extremely rich information-wise, and can be immensely valuable for a detailed structural study of smaller molecules. However detailed 3D contents can quickly become overwhelming when studying larger molecules or assemblies. Furthermore, the task of establishing, either visually or computationally, the homology of molecules can become time-consuming in the absence of simplified representations. One therefore needs representations for the secondary structure diagrams at various levels of abstraction.

Specificities and difficulties arise when results from the three-dimensional structures are projected onto a planar diagram representing a mixture of the secondary structure and the three-dimensional structure. Three-dimensional structures are extremely complex and rich in interatomic contacts and only a fraction of those can be adequately represented in simplified diagrams. The three-dimensional structure obtained either from crystallography or from nuclear magnetic resonance, yields directly the secondary structure, i.e. the set of non-Watson-Crick base pairs, together with the tertiary structure, i.e. the set of non-Watson-Crick base pairs (see Figure 22), of co-axial stacking of helices, and more generally of the stacking of bases. A widely adopted schematic representation for RNA tertiary structure is the **Leontis-Westhof nomenclature**, which offers a satisfactory tradeoff between simplicity and expressivity.

Within this nomenclature, the supporting plane around each type of base is partitioned in three regions, or **edges,** as shown in Figure 23. In addition to the **Watson-Crick** edge, which is involved in stable *canonical* base-pairs, the nomenclature distinguishes the **Hoogsteen edge**, which creates the opportunity for base-triplets. Although Hoogsteen edges are usually used for purines only, the same name is used to refer to the C-H edge of pyrimidines. The **sugar edge**, named as such because it includes 2’ hydroxyl group of the ribose, and is sometimes referred to as the shallow-groove edge. Accordingly, each base can be represented by a triangle whose relative dimensions symbolizes the area covered by the backbone and base. The symbol in the Sugar/Hoogsteen corner indicates the orientation of the sugar-phosphate backbone relative to the plane of the pairing: a circle means that the 5’ $\rightarrow$ 3’ direction comes from back to front, as exemplified by Figure 23, while a cross means that it goes from front to back.

Using this simplified representation, base-pairs mediated by hydrogen bonds can be visualized as assemblies of annotated triangles involving two edges of the triangle. The local orientation can match or differ, and two strands can therefore be **parallel** (the 5’ $\rightarrow$ 3’ directions on both strand points towards the same direction) or **antiparallel** (in case of opposite directions). Each contact can then be designated unambiguously stating the interacting edges of the two bases (Watson-Crick, Hoogsteen, or Sugar edge) and the relative glycosidic bond orientation, ***cis*** or ***trans***. The resulting 12 types of base-pairs can further abstracted by introducing dedicated symbols, as illustrated by Figure 24, and enriched with symbols for other key architectural elements, such as **base stacking**. Such interactions and additional features constitute a complete and expressive data model for *in silico* structural approaches such as S2S/Assemble, and greatly ease the visualization of complex structural motifs, as illustrated by Figure 4, Figure 7 and Figure 5.

9.10 References

Andersen, E. S., A. Lind-Thomsen, et al. (2007). "Semiautomated improvement of RNA alignments." RNA **13**(11): 1850-1859.

Andronescu, M., V. Bereg, et al. (2008). "RNA STRAND: the RNA secondary structure and statistical analysis database." BMC Bioinformatics **9**: 340.

Auber, D. D., Maylis; Domenger,Jean-Philippe; Dulucq, Serge (2006). "Efficient drawing of RNA secondary structure." Journal of Graph Algorithms and Applications **10**(2).

Bateman, A., S. Agrawal, et al. (2011). "RNAcentral: A vision for an international database of RNA sequences." RNA **17**(11): 1941-1946.
Bauer, M., G. W. Klau, et al. (2007). "Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization." BMC Bioinformatics **8**: 271.
Berman, H. M., W. K. Olson, et al. (1992). "The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids." Biophys J **63**(3): 751-759.
Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-242.
Bernhart, S. H., I. L. Hofacker, et al. (2008). "RNAalifold: improved consensus structure prediction for RNA alignments." BMC Bioinformatics **9**: 474.
Bindewald, E. and B. A. Shapiro (2006). "RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers." RNA **12**(3): 342-352.
Birmingham, A., J. C. Clemente, et al. (2011). "Meeting report of the RNA Ontology Consortium January 8-9, 2011." Stand Genomic Sci **4**(2): 252-256.
Brion, P. and E. Westhof (1997). "Hierarchy and dynamics of RNA folding." Annu Rev Biophys Biomol Struct **26**: 113-137.
Brown, J. W. (1999). "The Ribonuclease P Database." Nucleic Acids Res **27**(1): 314.
Brown, J. W., A. Birmingham, et al. (2009). "The RNA structure alignment ontology." RNA **15**(9): 1623-1631.
Brown, J. W. and N. R. Pace (1992). "Ribonuclease P RNA and protein subunits from bacteria." Nucleic Acids Res **20**(7): 1451-1456.
Bruccoleri, R. E. and G. Heinrich (1988). "An improved algorithm for nucleic acid secondary structure display." Comput Appl Biosci **4**(1): 167-173.
Byun, Y. and K. Han (2009). "PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots." Bioinformatics **25**(11): 1435-1437.
Cannone, J. J., S. Subramanian, et al. (2002). "The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs." BMC Bioinformatics **3**: 2.
Curcin, V., M. Ghanem, et al. (2005). "Web services in the life sciences." Drug Discov Today **10**(12): 865-871.
Dalli, D., A. Wilm, et al. (2006). "STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time." Bioinformatics **22**(13): 1593-1599.
Darty, K., A. Denise, et al. (2009). "VARNA: Interactive drawing and editing of the RNA secondary structure." Bioinformatics **25**(15): 1974-1975.
De Rijk, P., J. Wuyts, et al. (2003). "RnaViz 2: an improved representation of RNA secondary structure." Bioinformatics **19**(2): 299-300.
Ding, Y. (2006). "Statistical and Bayesian approaches to RNA secondary structure prediction." RNA **12**(3): 323-331.
Ding, Y., C. Y. Chan, et al. (2005). "RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble." RNA **11**(8): 1157-1166.
Do, C. B., D. A. Woods, et al. (2006). "CONTRAfold: RNA secondary structure prediction without physics-based models." Bioinformatics **22**(14): e90-98.
Gardner, P. P., J. Daub, et al. (2011). "Rfam: Wikipedia, clans and the "decimal" release." Nucleic Acids Res **39**(Database issue): D141-145.
Gardner, P. P. and R. Giegerich (2004). "A comprehensive comparison of comparative RNA structure prediction approaches." BMC Bioinformatics **5**: 140.
Gesell, T. and A. von Haeseler (2006). "In silico sequence evolution with site-specific interactions along phylogenetic trees." Bioinformatics **22**(6): 716-722.
Griffiths-Jones, S. (2005). "RALEE--RNA ALignment editor in Emacs." Bioinformatics **21**(2): 257-259.

Gruber, A. R., R. Lorenz, et al. (2008). "The Vienna RNA websuite." Nucleic Acids Res **36**(Web Server issue): W70-74.
Harmanci, A. O., G. Sharma, et al. (2007). "Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign." BMC Bioinformatics **8**: 130.
Harmanci, A. O., G. Sharma, et al. (2008). "PARTS: probabilistic alignment for RNA joinT secondary structure prediction." Nucleic Acids Res **36**(7): 2406-2417.
Hochsmann, M., B. Voss, et al. (2004). "Pure multiple RNA secondary structure alignments: a progressive profile approach." IEEE/ACM Trans Comput Biol Bioinform **1**(1): 53-62.
Hofacker, I. L. (2009). "RNA secondary structure analysis using the Vienna RNA package." Curr Protoc Bioinformatics **Chapter 12**: Unit12 12.
Hofacker, I. L., S. H. Bernhart, et al. (2004). "Alignment of RNA base pairing probability matrices." Bioinformatics **20**(14): 2222-2227.
Holmes, I. (2005). "Accelerated probabilistic inference of RNA structure evolution." BMC Bioinformatics **6**: 73.
Jossinet, F., T. E. Ludwig, et al. (2010). "Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels." Bioinformatics **26**(16): 2057-2059.
Jossinet, F. and E. Westhof (2005). "Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure." Bioinformatics **21**(15): 3320-3321.
Juhling, F., M. Morl, et al. (2009). "tRNAdb 2009: compilation of tRNA sequences and tRNA genes." Nucleic Acids Res **37**(Database issue): D159-162.
Kachouri, R., V. Stribinskis, et al. (2005). "A surprisingly large RNase P RNA in Candida glabrata." RNA **11**(7): 1064-1072.
Kaiser, A., J. Kruger, et al. (2007). "RNA Movies 2: sequential animation of RNA secondary structures." Nucleic Acids Res **35**(Web Server issue): W330-334.
Kiryu, H., Y. Tabei, et al. (2007). "Murlet: a practical multiple alignment tool for structural RNA sequences." Bioinformatics **23**(13): 1588-1598.
Kozomara, A. and S. Griffiths-Jones (2011). "miRBase: integrating microRNA annotation and deep-sequencing data." Nucleic Acids Res **39**(Database issue): D152-157.
Lee, M. K., M. Gal, et al. (2010). "Real-time multidimensional NMR follows RNA folding with second resolution." Proc Natl Acad Sci U S A **107**(20): 9192-9197.
Lemieux, S. and F. Major (2006). "Automated extraction and classification of RNA tertiary structure cyclic motifs." Nucleic Acids Res **34**(8): 2340-2346.
Leontis, N. B., A. Lescoute, et al. (2006). "The building blocks and motifs of RNA architecture." Curr Opin Struct Biol **16**(3): 279-287.
Leontis, N. B., J. Stombaugh, et al. (2002). "The non-Watson-Crick base pairs and their associated isostericity matrices." Nucleic Acids Res **30**(16): 3497-3531.
Leontis, N. B. and E. Westhof (2001). "Geometric nomenclature and classification of RNA base pairs." RNA **7**(4): 499-512.
Lestrade, L. and M. J. Weber (2006). "snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs." Nucleic Acids Res **34**(Database issue): D158-162.
Lindgreen, S., P. P. Gardner, et al. (2007). "MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing." Bioinformatics **23**(24): 3304-3311.
Luck, R., S. Graf, et al. (1999). "ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure." Nucleic Acids Res **27**(21): 4208-4217.
Markham, N. R. and M. Zuker (2008). "UNAFold: software for nucleic acid folding and hybridization." Methods Mol Biol **453**: 3-31.
Mathews, D. H. (2004). "Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization." RNA **10**(8): 1178-1190.
Mathews, D. H., M. D. Disney, et al. (2004). "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure." Proc Natl Acad Sci U S A **101**(19): 7287-7292.

McCaskill, J. S. (1990). "The equilibrium partition function and base pair binding probabilities for RNA secondary structure." Biopolymers **29**(6-7): 1105-1119.
Mokrejs, M., T. Masek, et al. (2010). "IRESite--a tool for the examination of viral and cellular internal ribosome entry sites." Nucleic Acids Res **38**(Database issue): D131-136.
Moretti, S., A. Wilm, et al. (2008). "R-Coffee: a web server for accurately aligning noncoding RNA sequences." Nucleic Acids Res **36**(Web Server issue): W10-13.
Parisien, M. and F. Major (2008). "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data." Nature **452**(7183): 51-55.
Perriquet, O., H. Touzet, et al. (2003). "Finding the common structure shared by two homologous RNAs." Bioinformatics **19**(1): 108-116.
Sankoff, D. (1985). Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems, SIAM.
Sarver, M., C. L. Zirbel, et al. (2008). "FR3D: finding local and composite recurrent structural motifs in RNA 3D structures." J Math Biol **56**(1-2): 215-252.
Seibel, P. N., T. Muller, et al. (2006). "4SALE--a tool for synchronous RNA sequence and secondary structure alignment and editing." BMC Bioinformatics **7**: 498.
Seibel, P. N., T. Muller, et al. (2008). "Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE." BMC Res Notes **1**: 91.
Stombaugh, J., J. Widmann, et al. (2011). "Boulder ALignment Editor (ALE): a web-based RNA alignment tool." Bioinformatics **27**(12): 1706-1707.
Taufer, M., A. Licon, et al. (2009). "PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots." Nucleic Acids Res **37**(Database issue): D127-135.
Tinoco, I., Jr., P. N. Borer, et al. (1973). "Improved estimation of secondary structure in ribonucleic acids." Nat New Biol **246**(150): 40-41.
Tinoco, I., Jr. and C. Bustamante (1999). "How RNA folds." J Mol Biol **293**(2): 271-281.
Torarinsson, E., J. H. Havgaard, et al. (2007). "Multiple structural alignment and clustering of RNA sequences." Bioinformatics **23**(8): 926-932.
van Batenburg, F. H., A. P. Gultyaev, et al. (2001). "PseudoBase: structural information on RNA pseudoknots." Nucleic Acids Res **29**(1): 194-195.
van Batenburg, F. H., A. P. Gultyaev, et al. (2000). "PseudoBase: a database with RNA pseudoknots." Nucleic Acids Res **28**(1): 201-204.
Watts, J. M., K. K. Dang, et al. (2009). "Architecture and secondary structure of an entire HIV-1 RNA genome." Nature **460**(7256): 711-716.
Waugh, A., P. Gendron, et al. (2002). "RNAML: a standard syntax for exchanging RNA information." RNA **8**(6): 707-717.
Weinberg, Z. and R. R. Breaker (2011). "R2R--software to speed the depiction of aesthetic consensus RNA secondary structures." BMC Bioinformatics **12**: 3.
Westhof, E. and C. Massire (2004). "Structural biology. Evolution of RNA architecture." Science **306**(5693): 62-63.
Will, S., K. Reiche, et al. (2007). "Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering." PLoS Comput Biol **3**(4): e65.
Wilm, A., D. G. Higgins, et al. (2008). "R-Coffee: a method for multiple alignment of non-coding RNA." Nucleic Acids Res **36**(9): e52.
Wilm, A., K. Linnenbrink, et al. (2008). "ConStruct: Improved construction of RNA consensus structures." BMC Bioinformatics **9**: 219.
Wuchty, S., W. Fontana, et al. (1999). "Complete suboptimal folding of RNA and the stability of secondary structures." Biopolymers **49**(2): 145-165.
Yang, H., F. Jossinet, et al. (2003). "Tools for the automatic identification and classification of RNA base pairs." Nucleic Acids Res **31**(13): 3450-3460.

Yao, Z., Z. Weinberg, et al. (2006). "CMfinder--a covariance model based RNA motif finding algorithm." Bioinformatics **22**(4): 445-452.
Yokoyama, T. and T. Suzuki (2008). "Ribosomal RNAs are tolerant toward genetic insertions: evolutionary origin of the expansion segments." Nucleic Acids Res **36**(11): 3539-3551.
Zuker, M. (1989). "On finding all suboptimal foldings of an RNA molecule." Science **244**(4900): 48-52.
Zuker, M. and P. Stiegler (1981). "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information." Nucleic Acids Res **9**(1): 133-148.
Zwieb, C., J. Gorodkin, et al. (2003). "tmRDB (tmRNA database)." Nucleic Acids Res **31**(1): 446-447.