Current Bioinformatics, 2008, *3*, 000-000

Computational RNA Structure Prediction

Emidio Capriotti and Marc A. Marti-Renom*

Structural Genomics Unit, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain

Abstract: The view of RNA as simple information transfer molecule has been continuously challenged since the discovery of ribozymes, a class of RNA molecules with enzyme-like function. Moreover, the recent discovery of tiny RNA molecules such as µRNAs and small interfering RNA, is transforming our thinking about how gene expression is regulated. Thus, RNA molecules are now known to carry a large repertory of biological functions within cells including information transfer, enzymatic catalysis and regulation of cellular processes. Similar to proteins, functional RNA molecules fold into their native three-dimensional (3D) conformation, which is essential for performing their biological activity. Despite advances in understanding the folding and unfolding of RNA, our knowledge of the atomic mechanism by which RNA molecules adopt their biological active structure is still limited. In this review, we outline the general principles that govern RNA structure and describe the databases and algorithms for analyzing and predicting RNA secondary and tertiary structure. Finally, we assess the impact of the current coverage of the RNA structural space on comparative modeling RNA structures.

Keywords: RNA, secondary structure, tertiary structure, computational biology, structure prediction, comparative modeling.

INTRODUCTION

Recent discoveries have demonstrated the role of RNA as biological regulator as well as information-transfer molecule [1-3]. For example, RNA molecules have been associated with enzymatic functions [4], gene transcriptional regulation [4-6], and protein biosynthesis regulation [7]. The knowledge of its three-dimensional (3D) structure as well as its interactions with other biomolecules in the cell is essential for characterizing such functions. Initial descriptions of the molecular details about RNA secondary structure were already published by the end of the fifties [8, 9]. However, the first RNA structure (i.e., the yeast phenylalanine t-RNA) had to wait about 15 years to be experimentally determined in 1974 [10]. Only few years later, computational biologists started developing the first methods for RNA secondary structure prediction. In seminal works, Zuker [11, 12] and Nussinov [13, 14] provided the first computational algorithms to predict a list of RNA base-pairs from sequence. In 1990 Michel and Westhof derived a 3D model of a conserved core of group I introns [15]. Only recently, a significant number of RNA structures are being deposited in the Protein Data Bank (PDB) [16]. Since 1992, such depositions have been specifically collected and stored in the Nucleic Acid database (NDB) [17].

Classically, RNA structure determination has mostly been accomplished by X-Ray crystallography or Nuclear Magnetic Resonance (NMR) approaches and only a limited number of attempts have been carried for automatically predicting the 3D structure of a large RNA molecule [18, 19]. Despite that, the application of computational algorithms of RNA structure prediction has been one of the sources for characterizing the structural diversity in RNA molecules and its relationship to function. Most of the existing algorithms rely in the principle that RNA folding is a hierarchical process and that knowledge of its secondary structure (i.e., the determination of all base-pairing in a RNA sequence) may improve the prediction of its 3D conformation. Consequently, several *ab-initio* methods have been implemented in computational programs for predicting the base-pairs interactions in RNA from its sequence [20-22]. However, the growing number of available structural data for RNA molecules and the initial attempts for classifying their motifs [23] has opened the possibility for applying comparative approaches for RNA structure prediction. Comparative modeling is not a novel concept and has been applied to protein structure prediction for more than two decades. When an homologous structure is available, such approaches usually result in the most accurate protein structure models [24-26]. Although both RNA and proteins form compact and globular structures in solution, the driving forces in the RNA and protein folding are different. RNA folding is essentially driven by its base pair and its regular motifs [1] while the hydrophobic collapse of the protein core is the main force during protein folding [27]. In contrast to proteins, RNA sequence conservation within the same functional family is usually limited to very short fragments of nucleotides which still maintain a substantial conservation of their secondary structure [1]. Therefore, it seems fair to say that in general it will be more difficult to predict large RNA 3D structures than predicting protein structures.

We begin this review by describing the RNA structure and initial attempts for classifying of the RNA structural space. We continue by outlining recent developments and method for RNA sequence alignment, as well as for secondary and tertiary structure prediction from sequence. We then conclude by discussing possible implications of the use of comparative approaches to predict the 3D structure of RNA sequences based on existing known structures. The bibliography is not exhaustive, but an attempt was made to quote the latest papers or reviews in the relevant fields.

---

*Address correspondence to this author at the Structural Genomics Unit, Bioinformatics Department, Centro de Investigación Príncipe Felipe, Av. Autopista del Saler, 16, 46013 Valencia, Spain; Tel: +34 96 3289680; Fax: +34 96 3289701; E-mail: mmarti@cipf.es

1574-8936/08 $55.00+.00

RNA 3D-STRUCTURE

RNA Base Pairs

Over the last few years there has been a rapid grow in the number of RNA structures made available through the Protein Data Bank (PDB) [17] and compiled in the RNAbase database (Fig. **1**) [28]. This increment is mostly due to the recent structural determination of ribosome machineries [29-32]. Thus, the availability of such data has allowed the application of more robust classification of base-to-base (also referred as base-pair) interactions in RNA molecules. Although there are differences in the interaction of two RNA bases, a stable classification depending on the edges involved in the interaction (i.e., Watson-Crick (WC), Hoogsteen or Sugar edges) has already been proposed [33, 34]. In such classification, each base can form several non-bonded interactions that involve different types of atoms: (i) phosphate-phosphate interaction mediated by water molecules; (ii) phosphate-sugar interaction; (iii) sugar-sugar interaction; (iv) base-phosphate interaction; (v) base-sugar interaction and (vi) base-base interaction. Moreover, those six different interaction types can be formed in either a *cis* or *trans* states resulting in 12 possible different conformations (Fig. **2**) [35]. About 60% of the base-pairs in known RNA structures adopt the canonical WC-WC interaction in *cis* conformation (Fig. **3**). Moreover, when other WC-WC interactions in *trans* are also accounted as standard, the remaining non-WC base-pairs account only for the 24% of the 140,501 base pairs in the PDB database (i.e., as of December 2006 with 2,180 RNA chains extracted from 1,101 structures).

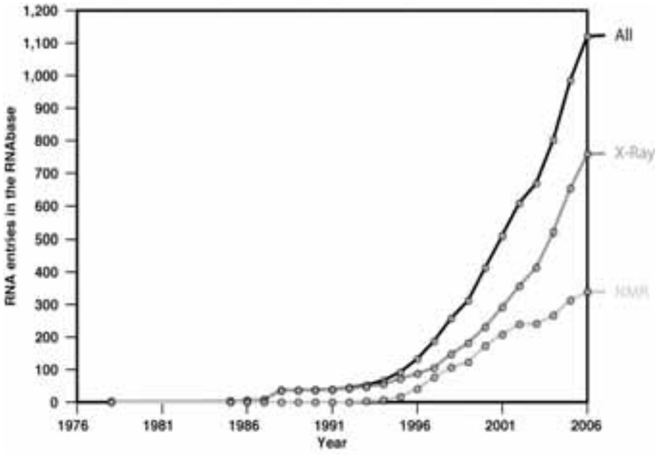

Fig. (1). Yearly growth of RNA structure deposition in the RNAbase database. Black curve shows all entries in the RNAbase, dark-grey curve shows only entries that were determined by X-Ray experiments and light-grey curve shows those determined by Nuclear Magnetic Resonance experiments.

RNA Backbone

Differently from proteins, RNA molecules are characterized by well-packed side-chains stabilized by hydrogen bonds and a flexible backbone. The RNA backbone is usually described by six continuous torsion angles between the phospohate (P), oxygen 5' (O5'), carbons 5' (C5'), 4' (C4'), 3' (C3'), and oxygen 3' (O3') atoms of a base (Fig. **4**). Only recently, Richardson and colleagues have analyzed a set of RNA structures with crystallographic resolutions higher than

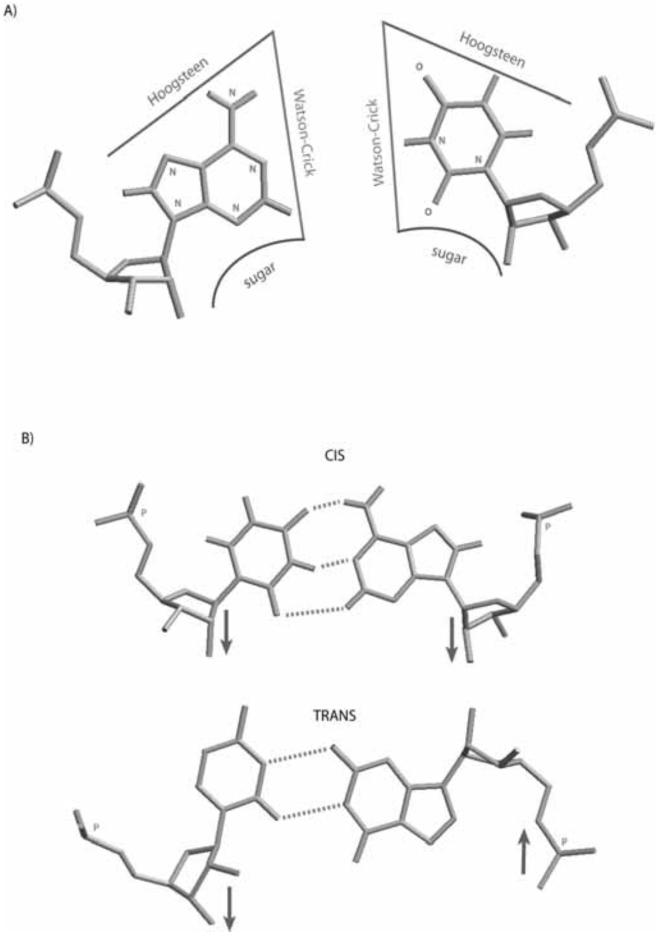

Fig. (2). Base-pair interactions. **A**) Watson-Crick (WC), Hoogsteen and sugar edges for a base-pair interaction. **B**) *cis* and *trans* states of a base-pair interaction.

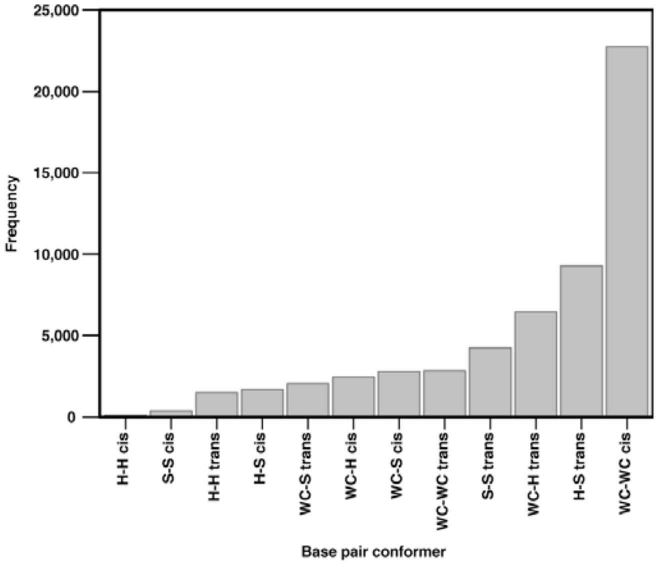

Fig. (3). Frequency of the 12 different conformations that a base-pair can adopt. Watson-Crick (WC), Hoogsteen (H), and sugar (S) base-pairs were used in the distribution obtained from 2,180 RNA chains from 1,178 structures in the PDB (December 06), which correspond to 140,501 base pairs.

3 Ångströms (Å) and no-atom clashes identifying 42 discrete RNA backbone conformers [36]. Other similar studies also concluded that RNA backbone is rotametic and can be classified into discrete conformers [37, 38]. These type of analyses have been possible because the quality and amount of determined RNA structures has considerably grown over the last years [28, 39, 40]. However, most large RNA structures can only be determined at resolutions lower than 2.5 Å. At this resolution, the phosphate and base plane can be accurately positioned but the sugar ring and the rest of the backbone atoms may contain errors. Indeed, the authors of RNABase database have analyzed and classified RNA torsion angles to conclude that on average one error can occur every two bases [28].

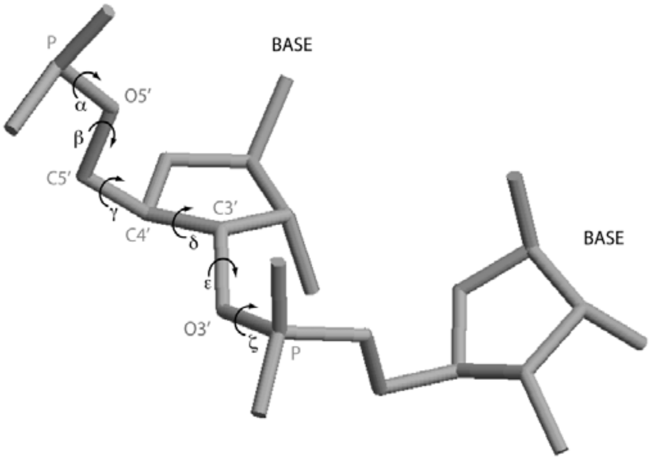

Fig. (4). RNA Backbone torsion angles.

RNA Motifs

RNA motifs correspond to recurrent RNA structural elements, which are subject to spatial constraints [23, 41]. This broad definition of RNA motifs already indicates the difficulty for uniquely describing or classifying them. RNA motifs are usually classified by its regular sequence patterns or its 3D conformation [23]. Here we will focus exclusively on the motifs that can be detected from structural information. RNA secondary structure, which can be reliably predicted from sequence [42], partially explains some of the known RNA motifs such as, bulges, hairpins, internal loops, and multi-helical motifs (Fig. **5**). However, the prediction of pseudo-knots is a more challenging task in secondary structure prediction programs because they contain two stem-loop motifs in which the first stem loop forms part of the second stem (Fig. **5B**). Structural data indicate that the final 3D RNA structure is mostly determined by its base-pair stacking (i.e., WC base pairs) and non-WC interactions. Thus, characterizing, analyzing, and ultimately predicting the stacking of those bases will help the goal of classifying complex RNA motifs.

RNA STRUCURAL DATABASES AND CLASSIFICATION

Since the seventies, when the first RNA structures became available [10], there has been an attempt to store, organize and classify the RNA structural space. Berman and coworkers developed the Nucleic Acid Database (NDB), which included known structures for DNA and RNA molecules [17]. NDB stores all molecules containing nucleic acid residues and complements them with additional information such as classification of nucleic acids and their interaction with proteins, backbone conformation angles, and base-pair classification. More recently, Murthy and Rose [28] have developed a RNA specialized database (RNABase), which collects and classifies RNA structures according to experimental properties and functional categories. For each entry in the RNABase, links to external databases, rasterized images and Ramachandran style maps of the backbone conformation angles are also provided. RNAbase identifies RNA structures with discrete conformational codes describing the multidimensional conformational space accessible to the structure. Thus, such codes can be used by their retrieval system for a fast search of structures occupying parts of the conformational space. RNAbase also provides a list of all 1,210 entries (August 2007) classified by their structural or functional categories (Table **1**). Thus, the information stored in the RNAbase allows the study of the relationship between sequence, structure and function of RNA molecules.

The Structural Classification of RNA (SCOR) database was developed in 2002 with the aim of organizing and classifying RNA structures for model building and engineering [40]. The SCOR database organizes RNA motifs in a hierarchical classification system similar to the SCOP database for protein domains [43]. However, the modularity of RNA at the sub-domain level makes the classification of RNA structures a more challenging task than for proteins. As a result, the classification in SCOR contains properties of directed acyclic graph architectures similar to that of the Gene Ontology database [44]. SCOR classifies RNA structures from three properties: first, the RNA structural classification describes RNA motifs according to the number of strands connecting double helices; second, the RNA functional classification divides each entry by the biological function of their molecule, motif and structural model; and third, the RNA tertiary interaction classification groups RNA molecules by their inter- and intra-molecular interactions differing from WC and non-WC base pairs. The SCOR database stores 8,270 structural motifs (October 2004), some of which are further classified into functional and RNA tertiary interaction classes (Table **2**). The SCOR database may prove very useful to identify hidden relationships between sequence, structure and function.

Although not explicitly using 3D structural information, the Rfam database [39] classifies non-coding RNA molecules into families of members that conserve sequence and secondary structure. It is known that, similar to proteins [45], the conservation of RNA secondary structure implies a degree of conservation of its function. The Rfam database uses this principle to detect sequence relationships by RNA secondary structure profiles derived from the so called ‘covariance models’ methods. Currently (February 2007), Rfam classifies ~33,000 RNA molecules into ~600 families providing family-based multiple alignments of consensus secondary structures. The use of Rfam stored data may prove useful for developing new methods for structural and functional RNA motif prediction.

There have been other attempts to store and classify known RNA structures using alignments and consensus secondary structures. Such databases are usually more suitable for focused research on particular classes of RNA molecules

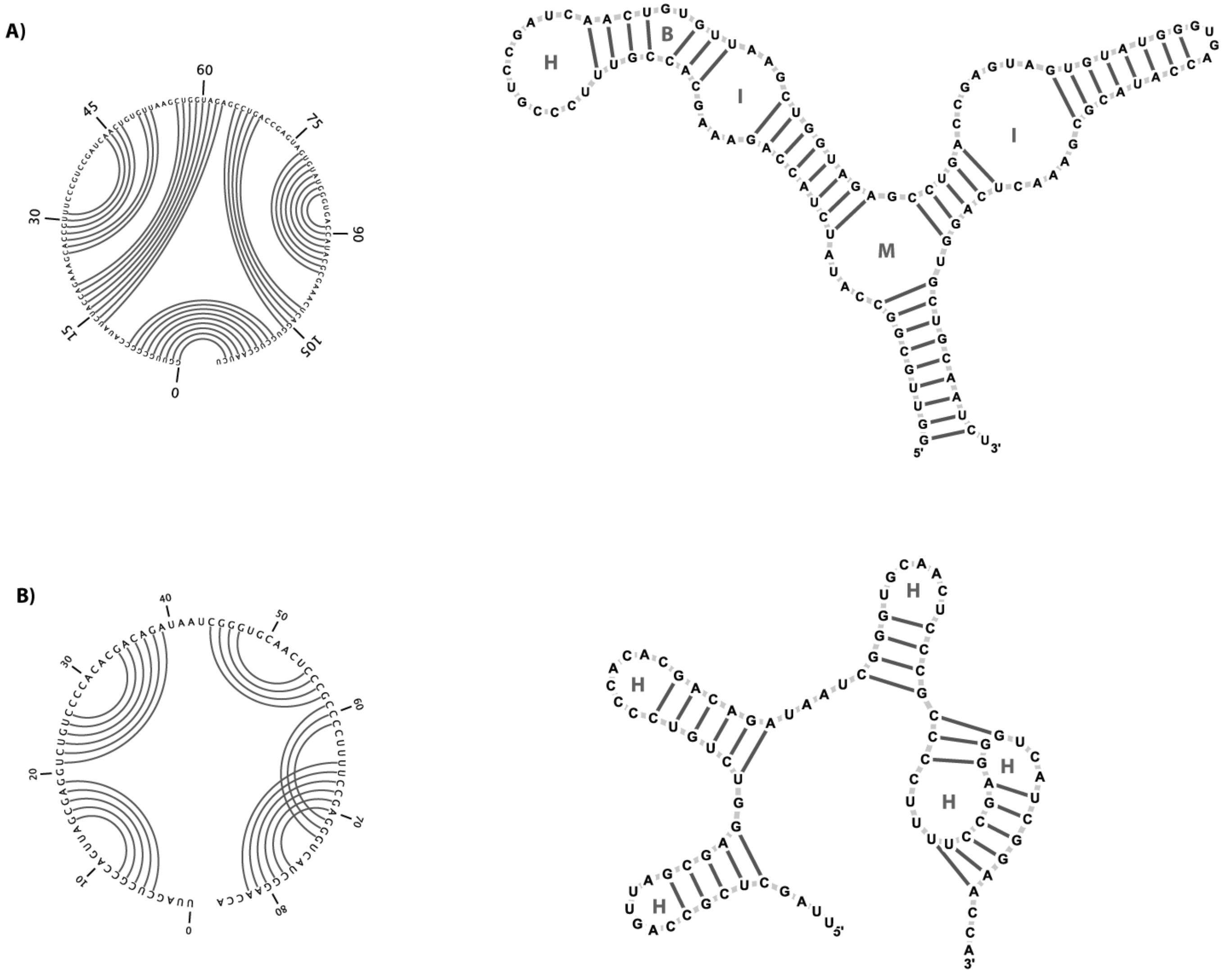

Fig. (5). RNA secondary structural motifs. Hairpin (H), internal (I), bulges (B) and multihelix (M) loops are shown in a circular Feynman (left) and secondary structure (rigth) representations. **A**) *S. cerevisiae* 5S ribosomal RNA (Genbank: X67579). **B**) tRNA-like molecule from the turnip yellow mosaic virus (Genbank: M58309). Images were produced with the jViz.RNA program [150].

or for evaluating consensus features of specific subsets of the RNA structural space. A list of such databases is included in the Appendix A.

RNA ALIGNMENT

RNA Sequence Alignment Methods

Similarly to protein sequences, RNA sequence alignment can be used for homology detection. Experimental evidences as well as computational analysis have shown that for protein sequences homology detection can be reliably done if the two sequences share more than 20-30% sequence identity [46, 47]. However, such a “twilight zone” [48] has not yet been determined for RNA homology detection. Methods such as BLAST and PSI-BLAST [49, 50], FASTA [51], CLUSTALW [52], MUSCLE [53, 54], or T-Coffee [55] have been developed or adapted to detect remote similarities between nucleic acid molecules as well as proteins. However, the detection of RNA homology for molecules with diverse sequence is not trivial and additional information such a predicted secondary structure may ensure a higher accuracy in both sequence homology detection and alignment quality.

RNA Secondary Structure Alignment Methods

One of the most challenging problems in modern computational RNA biology is the detection of an accurate secondary structure alignment between two or more RNA molecules. Several methods have been already developed [56-59]. For example, the RNAdistance program, which is available as part of the Vienna package, uses a tree-based model to coarsely represent and compare secondary structures based on edit distances [20]. The RNAforester program extends this simplified tree-model to the forest model, significantly improving both the time and space complexity of the searching algorithm [56]. The MARNA program uses multiple sequence and secondary structure information to generate more accurate multiple RNA alignments [58]. A new method based on pair stochastic tree adjoining grammars has been

Table1. RNAbase Classification

Category	Entries
Transfer RNAs	217
Ribosomal RNAs	283
Messenger RNAs	126
Transcription-related RNAs	86
Introns	26
Splicing-related RNAs	59
Signal recognition particle RNAs	22
Ribozymes	115
RNase P	21
Aptamers	30
Pseudoknots	31
Tetraloops	81
Bulges	69
DNA-RNA hybrids	115
PNA-RNA hybrids	1
Drug-RNA complexes	137
Viral & Phage RNAs	221

Number of RNA structure entries stored in the RNAbase classified by their functional categories.

developed to allow the inclusion of pseudo-knots during the alignment of two RNA molecules [60]. Finally, a new algorithm for aligning RNA pseudoknots has been designed for genome-wide scanning of novel non-coding RNA genes [61].

Table 2. SCOR Classification

Classification	Subclasses	RNA Motifs
Structural Classification	Internal Loops	5350
	Hairpin Loops	2920
Functional Cassification	Molecular Function	480
	Motif Function	179
	Structural Models	137
RNA Tertiary Interaction	Coaxial Helices	7
	Tetraloop-Receptor	1
	A-Minor Motif	240
	Kissing Hairpin Loops	32
	tRNA D-Loop:T-Loop	7
	Pseudoknots	17
	Ribose Zipper	657

Number of RNA structure motifs stored in the SCOR database classified by structural, functional and RNA tertiary interaction categories.

RNA secondary structure is conserved between divergent sequences [62] and helps to confers functional specificity [1]. Thus, it is expected that the knowledge of the RNA secondary structure may enhance RNA homology detection [63]. Several methods make use of known secondary structure by relying on: i) matching defined RNA sequence/structural patterns, ii) Hidden Markov models (HMM) or Stochastic Content Free Grammar (SCFG) methods, and iii) classical sequence alignment combined with the maximal pairing algorithms.

The first approach consists in defining an appropriate pattern derived from secondary structural information, which is then used in the search against a database of RNA sequences. The first application of such approach was used to identify possible homologous sequences of transport RNA (tRNA) molecules and the structural motif for the group I intron [64]. More recently, a declarative programming language has been designed to describe more complex RNA secondary structural elements [65, 66]. Such pattern matching methods have also previously been applied to detect homologous sequences to the Iron Responsive Element, the Histone stem-loop structure, and the Selenocysteine Insertion sequence [67].

The second approach uses trained HMM methods to search for homology in RNA sequence databases. RNA HMM usually takes as input the linear sequence and a tree representing its secondary structure. The output usually consists of an alignment for the secondary structure elements of the query sequence and the detected homolog sequences [68]. SCFG, a generalization of HMMs that allows for modeling pairwise interactions, are also used for detecting homology between a query sequence and a database of RNA sequences. SCFG has been previously used to model tRNA [69, 70] and small nucleolar RNA (snoRNA) [71]. The RSEARCH program [72] considers two different score matrices (i.e., 4 by 4 for single nucleotide alignment substitution matrix and a 16 by 16 substitution matrix) to score a set of aligned base pairs [73]. Using those matrices and the query sequence, the program first builds a tree-like structure encoding for the RNA sequence/structural features and then aligns the query sequence against each sequence in the database using a dynamic programming algorithm [74]. The INFERNAL program, a type of SCFG approach, have been used to build the Rfam database [75]. INFERNAL scoring function combines measures of sequence consensus and RNA secondary structure consensus, which allows the detection of RNA homologs that conserve their secondary structure more than their primary sequence [22].

Finally, the third class of approaches for RNA homology detection simultaneously explore possible solutions to the alignment and the secondary prediction problem [76]. The rationale behind such approach is the detection of a common base pair list by maximizing the sum of its base-pair weights. Thus, effectively merging the classical sequence alignment methods with the maximal pairing algorithm [13]. This type of approaches can be used to obtain both a sequence alignment and a consensus secondary structure. Available tools based on this procedure are discussed in the section dedicated to the methods for RNA secondary structure prediction.

RNA Phylogenetic Analysis

Probabilistic methods for phylogenetic analysis use substitution models defining the probability of a residue replacement. This kind of approach, often based on empirical models, has largely been used for protein sequence analysis [77, 78]. RNA phylogenetic analysis is facilitated by the compensating substitutions of paired bases [79], which depends on the thermodynamic stability of the intermediate folding state [80]. Several methods for phylogenetic analysis of RNA families use Bayesian and a Markov Chain Monte Carlo algorithms to find the most probable tree and posterior probabilities of clades [80, 81]. A recently developed method, MrBayes [82], performs Bayesian phylogenetic analysis by combining information from different data partitions or subsets evolving under different stochastic evolutionary models.

The list of the current available programs for RNA sequence alignment and phylogenetic analysis is reported in Appendix B.

RNA SECONDARY STRUCTURE PREDICTION

Single Sequence Free Energy Calculations

The RNA folding process is hierarchical [83], which means that local interactions occur first and are energetically stronger than tertiary interactions [84]. Therefore, RNA secondary structure provides a scaffold to its native 3D structure. This property already indicates that RNA secondary structure can be predicted without the knowledge of tertiary interactions. Unfortunately, and despite recent advances, RNA secondary structure prediction still constitutes a challenge in computational structural biology [84-87].

Protein folding studies by Anfinsen hypothesized that, at environmental conditions, the native structure is a unique, stable and kinetically accessible free energy minimum [88]. Thus, this general approach assumes that, at the equilibrium in physiological conditions, the native protein conformation is unique and determined by its sequence. The first algorithms for predicting the secondary structure of RNA molecules were developed assuming the same principles of the minimum free energy conformation search by dynamic programming [11, 12, 13, 14]. The scoring functions for such approaches were based on free energy parameters from physics, which were derived from empirical calorimetric experiments [89] or from known RNA structures deposited in the PDB [21]. Regardless of the scoring function used by such programs, most of them perform a complete evaluation of feasible features for a given RNA sequence to determine the minimal free energy conformation using a dynamic programming algorithm [87]. Unfortunately, and due to imperfect scoring functions, the minimum free energy approach (MFE) does not guarantee that the selected or predicted final structure will be the native structure and typically corresponds to a near-native conformation [87]. Other implementations of the MFE principle include the use of a heuristic search for suboptimal secondary structures [11, 89, 90], the computation of all suboptimal alignments near the optimal folding space [91], and the selection of suboptimal solutions based on RNA shape analysis [92].

In the 1990, McCaskill first implemented a method based on equilibrium partition function for secondary structure and associated probabilities of various substructures [93]. Such method allowed the statistical characterization of the equilibrium ensemble of RNA secondary structures. It has been noticed that higher base-pair probabilities, computed by the partition function approach, correspond to higher predictive reliability when considering structures determined by comparative sequence analysis [90].

More recently, new computational approaches based on statistical samplings of known RNA secondary structures [21] or genetic algorithms [94-96] have also been implemented for secondary structure prediction. However, most of the methods described so far are based on the recursive approach, which is not suitable for predicting RNA pseudoknots. It has been demonstrated that the prediction of secondary structure motifs with pseudoknots is a NP-complete problem making it computational intractable [97]. To address this problem, modified dynamic programming [98-100] and stochastic context-free grammar algorithms [101] have been recently introduced. For example, the PKNOTS program implements thermodynamic folding in a rather large subclass of pseudoknots on $O(N^4)$ and $O(N^6)$ time space, which makes it only usable for short sequences [98]. The partition function approach implemented by Dirks [99, 100] has an $O(N^5)$ complexity. Despite this computational complexity, the accuracy for pseudoknots prediction has significantly increase by using an innovative dynamic partner sequence stacking algorithm [102].

Appendix C lists some available methods for secondary structure prediction including those for pseudoknot prediction.

Multiple Sequence Comparison

RNA secondary structure prediction from single-sequence somehow neglects the evolutionary forces acting upon RNA sequence variation. Therefore, the inclusion of multiple sequences for predicting the RNA secondary structure allows the incorporation of constraints based on the commonalities of the compared sequences [103]. Evolution tends to conserve RNA secondary structure more than sequence [62]. This observation is widely used by different types of approaches for secondary structure prediction from multiple sequences. Such algorithms can be coarsely grouped in *align plus fold*, *simultaneous align and fold*, and *fold plus align* types of methods. A comprehensive review and benchmark of RNA secondary structure prediction methods shows that the first type of approaches on average reach the best level of accuracy, while the latter approaches result in less accurate predictions [104].

Align plus fold. It has been observed that a mutation in a RNA molecule is usually compensated by a second mutation in the paired base [105, 106]. Several methods for secondary structure prediction use this principle by attempting to detect such covariance between different positions in the multiple sequence alignment. An initial implementation of such approach used mutual information theory to extract the covariance between bases [107, 108]. However, those approaches resulted in limited accuracy [109] and have been replaced by more recent implementations such as the RNAalifold program [110], which scores possible solutions by combining a free-energy term with a covariance term, the Pfold program [111], which uses a evolutionary SCFG approach, or the

ILM program [112, 113], which combines thermodynamic and mutual information in a single score.

Simultaneous align and fold. This class of approaches for RNA secondary structure prediction simultaneously explore possible solutions to the alignment and its secondary structure. In 1985, Sankoff proposed the first rigorous mathematical treatments of this problem [76]. His approach used a dynamic programming algorithm to search the structural conformational space, which made the method computationally expensive. Thus, current tools implement restricted versions of original Sankoff algorithm. The Foldalign program [114, 115] heuristically considers local sequence alignments and maximum number of base pairs at the same time. The Dyalign program [116] is a pairwise alignment method that searches for common low energy structures between two sequences. The algorithm complexity is reduced by considering a maximum value of sequence distance between two aligned residues and by limiting the size of any internal loop. Finally, the Carnac program [117, 118], which is not a strict implementation of a simultaneous align and fold approach, relies on a thermodynamic model with energy minimization by combining information from locally conserved elements and mutual information between sequences.

Fold plus align. This approach first folds the RNA sequences using single sequence secondary structure prediction methods and then aligns the resulting structures using tree-based methods [119]. The RNAforester [120] and MARNA [58] programs can be classified under this approach and have been already introduced in this review.

RNA 3D STRUCTURE PREDICTION

RNA Structure Comparison

The increase over the last decade of the number of available structures deposited in the PDB, including X-ray and NMR models, has stimulated the structural biology community to develop computational tools for analyzing the RNA structural space [35, 121-130]. Next, we outline some of those methods.

The NASSAM program [131] was designed for identifying common sub-structural motifs between two RNA structures. The NASSAM program implements a simplified vector representation of each nucleic acid base with respect its position in the structure. Then the vectors and their edges are transformed in a graph connecting the bases and compared using the Ullman subgraph isomorphism algorithm.

The PRIMOS program [132], similarly to the Ramachandran's approach used to investigate the conformational space of the protein backbone, describes a RNA structure with pseudo torsion angles η ($C4'_{i-1}$-P_i-$C4'_i$-P_{i+1}) and θ (P_i -$C4'_i$ -P_{i+1}-$C4'_{i+1}$) obtained with the AMIGOS program [121]. Then the search comparison is done over the simplified version of the RNA structural representation allowing the identification of common small motifs between two RNA structures or a RNA structural motif and a database of RNA structures.

Both, PRIMOS and NASSAM have been successfully used to identify 3D motifs in RNA structural databases but they are unable to identify unknown motifs. The COMPADRES program [133] was developed to overcome such limitation. COMPADRES searches for consecutive RNA fragments with five or more nucleotides described by specific η and θ angles as well as the sugar pucker phase. The COMPADRES algorithm has been applied for identifying new RNA motifs such as p-turns, Ω-turns, α-loops, C2FA and Hook turns.

More recently, the ARTS [127, 128] and the DIAL [130] programs for structural comparison of RNA molecues have been developed to overcome the limitation of sequence continuity. The ARTS program describes RNA structures by a set of contiguous *quadrats* (i.e., four phosphate atoms located in two successive base pairs). The program then identifies very similar *quadrats* between two RNA structures and uses them as seeds for the final alignment. Two *quadrats* are considered similar by ARTS if their rigid superimposition is within a given RMSD threshold. Finally, the algorithm finds the maximal matching in a bipartite graph between the two structures by extending the structure alignment that maximizes the number of aligned bases and base pairs. The DIAL program uses a dynamic programming algorithm to align two RNA structures based on a scoring function that combines a base, a dihedral angle, and a base-paring similarity measure. DIAL can be run as a web server and provides the user with the option of producing global (Needleman-Wunsch), local (Smith-Waterman), or global-semiglobal (motif search) alignments.

Algorithms for RNA Structure Prediction

Predicting the 3D structure of an RNA molecule is not an easy task and usually requires of an important human intervention [134]. Compared to the current status of protein structure prediction, not a fully automated approach is able to reliably predict a RNA 3D structure from its sequence. However, over the last years, a plethora of methods have been developed that aid the manual or semi-automatic prediction of RNA structures. Next we outline some of such programs:

The ERNA-3D program [18] automatically generates a RNA 3D structure starting for its secondary structure. ERNA-3D, which has successfully been used to model the structure of transfer-messenger RNA molecules [19], is able to model RNA motifs by using high-resolution structural information from the SCOR database.

The MANIP program [135] builds complete RNA structural models based on the assembly of fragments from a library o RNA motifs. The final refinement protocol combines canonical as well as non-canonical base pairing constraints with restraints imposed by covalent geometry, stereochemistry, and van der Waals contacts.

The S2S framework [136] allows the end-user to easily display, manipulate and interconnect heterogeneous RNA data, such as multiple sequence alignments, secondary and tertiary structures.

The Nucleic Acid Builder program (NAB) [137], which can also be used for modelling proteins and small molecules, was developed to build helical and non-helical nucleic acid molecules. The program is based in the AMBER forcefield [138] to optimize by molecular dynamic simulations a set of restrains derived from known 3D structures.

The MC-Sym program [139] builds 3D RNA structures using the coordinates and relations between bases from known RNA structures. Additional constraints can be applied to the model during the building procedure to ensure the conservation of particular structural features. The program implements a symbolic language that is used to describe RNA structure properties and constraints executed by its interpreter. Like for NAB, Mc-Sym uses molecular dynamic simulations to minimize the energy of the predicted structure.

The RNA2D3D program [140], builds RNA structural models by first spacing the atoms of a nucleotide along a fixed backbone and then predicting the final structure of the model by an helical winding procedure. The model is further refined by interactively moving groups of nucleotides to better-fit known structural information or by minimizing it using molecular dynamics simulations.

Finally, a new approach [141], inspired by the Rosetta low-resolution protein structure prediction method [142], has been applied to predict the 3D structure of 20 RNA sequences of ~30 nucleotides. The authors report that their method is able to correctly predict the native conformation for ~90% of WC and about one-third of non-WC base pairs. Their results also suggest that improvements in the energy function together with the use of predictions from phylogenetic approaches are necessary for an accurate structure prediction of more complex RNA molecules.

The Appendix D lists some available methods for RNA tertiary structure analysis and prediction.

PERSPECTIVES

RNA structural determination, either by X-Ray crystallography or NMR, has recently significantly increased the number of known RNA structures in the PDB. This growth is mostly due to the determination of several structures of the ribosome machinery, which include very large and complex RNA structures [29-32]. Moreover, the recent advances in chemical synthesis of RNAs will likely result in even a faster increase in the number and diversity of determined RNA structures [32]. The available structural data on RNA molecules already shows the existence of regular and recurrent RNA motifs. Thus, the next logical step for structural biologists would be to detect, store, analyze and classify such structural motifs to aid in *ab-initio* or knowledge-based structural prediction of whole RNA sequences [143]. Even when secondary structure prediction methods are reaching a good level of accuracy, our ability to reliably predict a RNA structure from its sequence is nowadays limited [134]. However, the knowledge of a large number of determined or predicted RNA structures could help in the biologically relevant goal of detecting non-coding RNA molecules from genomic sequences. We believe that the use of comparative approaches may soon result in large-scale predictions of both secondary and tertiary structure prediction of RNA molecules.

Automatic protein structure prediction methods can reliably predict at least one domain for about one third of the known sequences [144, 145]. Such large-scale applications are usually only available for comparative approaches, which require the knowledge of homologous structures to the query sequence [25, 146, 147]. The current amount and diversity of known structures of RNA molecules may allow the development of similar approaches for RNA structure prediction. However, it is difficult to predict whether such methods will be readily applicable to RNA and, more importantly, will result in similar reliable models. Thus, an exhaustive analysis is necessary to determine which methods, previously developed for protein structural prediction, can be easily adapted and used to predict the RNA structure. For example, in a recent article Das and Baker have used the Rosetta program, initially developed for protein structure prediction, to predict the structure of 20 small RNA sequences [141]. Such approaches will require a complete classification of the RNA sequence and structure space, which will provide the exact relationship between sequence and structure for RNA molecules. Once the relationship is well characterized, the computational community will have the basis for developing methods for comparative RNA 3D structure prediction allowing an accurate and automatic prediction of the structure and function of RNA molecules. Moreover, the large amount of RNA structural information may prove useful in increasing the accuracy of methods for predicting non-coding RNA genes. Compared to proteins, RNA secondary structure more strongly determines its tertiary structure. Thus, small changes in sequence may result in a different base paring, which in turn changes its 3D conformation. For this reason, a better description and classification of RNA motifs can have a direct impact in RNA structure prediction and non-coding RNA homology search methods [57, 62, 148, 149]. We believe that in the near future we will see an increasing number of methods being developed for RNA structure prediction. Such methods will likely result in a more accurate description of the role of RNA molecules in biological processes.

APPENDIX A - DATABASES

Tertiary Structure Databases

NDB: the nucleic acid database: a repository of three-dimensional structural data from nucleic acids. http://ndbserver.rutgers.edu/.

RNAbase: an annotated database of all publicly available RNA structures. http://www.rnabase.org/.

SCOR: a comprehensive perspective and understanding of RNA motif structure, function, tertiary interactions and their relationships. http://scor.lbl.gov/index.html.

Alignment and Consensus Secondary Structure Databases

Rfam: a large collection of multiple sequence alignments and covariance models covering common non-coding RNA families. http://www.sanger.ac.uk/Software/Rfam/.

Ribosomal Database Project-II: The Ribosomal Database Project (RDP) provides ribosome related data and services to the scientific community. http://rdp.cme.msu.edu/index.jsp.

European rRNA database: a complete or nearly complete SSU (small subunit) and LSU (large subunit) ribosomal RNA sequences database. http://bioinformatics.psb.ugent.be/webtools/rRNA/.

CRW Site: alignments, structure models and phylogenetic analyses of 5S, 16S and 23S rRNA, Group I and II introns and tRNA. http://www.rna.ccbb.utexas.edu/.

SRPDB: the Signal Recognition Particle Database contains aligned, annotated and phylogenetically ordered sequences related to structure and function of SRP. http://rnp.uthct.edu/rnp/SRPDB/SRPDB.html.

The tmRNA Website Research: a collection of tmRNA sequences, alignments, secondary structures and other information. http://www.indiana.edu/~tmrna/.

tmRDB: tmRDB (tmRNA Database) provides aligned, annotated and phylogenetically ordered sequences related to structure and function of tmRNA. http://rnp.uthct.edu/rnp/tmRDB/tmRDB.html.

The RNase P Database: a compilation of RNase P sequences, sequence alignments, secondary structures, three-dimensional models, and accessory information. http://jwbrown.mbio.ncsu.edu/RNaseP/home.html.

Viral RNA Structure Database: Viral structures from TBI. http://rna.tbi.univie.ac.at/cgi-bin/virusdb.cgi.

APPENDIX B – ALIGNMENT AND PHYLOGENETIC TOOLS

General Sequence Alignment Methods

BLAST: The Basic Local Alignment Search Tool (BLAST) compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. http://www.ncbi.nlm.nih.gov/BLAST/.

CLUSTALW: a general purpose multiple sequence alignment program. http://www.ebi.ac.uk/clustalw/.

FASTA: a searcher for local alignments against proteins and nucleotides databases. http://www.ebi.ac.uk/fasta/.

MUSCLE: MUltiple Sequence Comparison by Log-Expectation. http://www.drive5.com/muscle/index.htm.

T-COFFEE: a collection of tools for computing, evaluating and manipulating multiple alignments of DNA, protein sequences and structures. http://www.tcoffee.org/.

RNA Alignment Methods by 2D Structural Information

LGSFAligner: a Local Gapped Subforest Aligner for pairs of RNA secondary structures. http://www.comp.nus.edu.sg/~bioinfo/LGSFAligner/.

MARNA: a program that considers both primary sequence and the secondary structure to align RNA sequences. http://biwww2.informatik.uni-freiburg.de/Software/MARNA/index.html.

MiGaL: a program for comparing RNA secondary structures and building phylogenetic trees. http://www-igm.univ-mlv.fr/~allali/migal/index.php.

RNA_align: aligns two RNA structures using an edit distance model. http://www.csd.uwo.ca/faculty/bma/rna_align/.

RNAdistance: a program for calculating measures for secondary structure dissimilarity. http://www.tbi.univie.ac.at/~ivo/RNA/.

RNAforester: a program for comparing and aligning RNA secondary structures *via* the "forest alignment" approach. http://bibiserv.techfak.uni-bielefeld.de/rnaforester/.

RNAshapes: a program that uses the "consensus shapes" method to predict an abstract shape common to RNA sequences. http://bibiserv.techfak.uni-bielefeld.de/rnashapes/submission.html.

Alignment Methods for RNA 2D Structures with Pseudoknots

PSTAG Pair stochastic tree adjoining grammars (PSTAGs) for aligning and predicting RNA secondary structures. http://phmmts.dna.bio.keio.ac.jp/pstag/.

RNA Searching Methods by Sequence/Structural Information

ERPIN: an RNA motif search program. http://tagc.univ-mrs.fr/erpin/.

HomoStRscan: a program for discovering homologous RNAs in complete genomes. http://protein3d.ncifcrf.gov/shuyun/homostrscan.html.

Infernal: a program to construct a RNA profile based upon an alignment and consensus structure. http://infernal.janelia.org/.

Palingol: a descriptive programming language to describe nucleic acid's secondary structure and scan sequence databases. http://wwwabi.snv.jussieu.fr/public/Palingol/.

PatSearch: a program for finding a defined pattern against a sequence(s). http://www.ba.itb.cnr.it/BIG/PatSearch/.

PHMMTS: an implementation of pair hidden Markov models on tree structures. http://phmmts.dna.bio.keio.ac.jp/.

RaveNnA: a software package for fast covariance modeling of RNA sequences. http://bliss.biology.yale.edu/~zasha/ravenna/.

RNACAD: a stochastic context-free grammar (SCFG) RNA modeling package. http://www.cse.ucsc.edu/~mpbrown/rnacad/index.html.

RNAmotif: a program for searching a database for RNA sequences that match a "motif" of secondary structure interactions. http://www.scripps.edu/mb/case/.

RSEARCH: a program for aligning an RNA query to target sequences, using SCFG algorithms. http://selab.wustl.edu/cgi-bin/selab.pl?mode=software#rsearch.

RSmatch: a RNA Secondary Structure Matcher. http://exon.umdnj.edu/software/RSmatch/.

RNA Phylogenetic Analysis Tools

CBCanalyzer: a program for inferring phylogenies based on compensatory base changes. http://cbcanalyzer.bioapps.biozentrum.uni-wuerzburg.de/.

jRNA: a program for exploring insect (and other less interesting) phylogenies using RNA secondary structure. http://hymenoptera.tamu.edu/rna/index.php.

MrBayes: a program for the Bayesian estimation of phylogeny. http://mrbayes.csit.fsu.edu/.

PHASE: a program designed for use with RNA sequences that have a conserved secondary structure, e.g., rRNA and tRNA. http://www.bioinf.manchester.ac.uk/resources/phase/.

RRNADIST, RRNAML and RNAML: a program for modeling empirical substitutions for Ribosomal RNA. http://www.uhnresearch.ca/labs/tillier/rRNA/rna.html.

SISSI: a software tool for generating data of related sequences along a given phylogeny. http://www.cibiv.at/ software/sissi/.

APPENDIX C – RNA SECONDARY STRUCTURE PREDICTION TOOLS

RNA Folding Software

Afold: a program for analyzing internal loops within RNA secondary structures. ftp://ftp.ncbi.nlm.nih.gov/pub/ogurtsov/Afold/.

CONTRAfold: a secondary structure prediction method based on conditional log-linear models (CLLMs). http:// contra.stanford.edu/contrafold/.

Kinfold: a program to simulate the stochastic folding kinetics of RNA sequences into secondary structures. http://www.tbi.univie.ac.at/~xtof/RNA/Kinfold/.

Mfold: an MFE RNA structure prediction algorithm. http://www.bioinfo.rpi.edu/applications/mfold/.

Rdfolder: a RNA folding program that uses energy weighted Monte Carlo simulations. http://rna.cbi.pku.edu.cn/.

RNAfold: an MFE RNA structure prediction algorithm. http://www.tbi.univie.ac.at/~ivo/RNA/.

RNA Kinetics: a program to simulate the dynamics of RNA secondary structure by the means of kinetic analysis of folding transitions of a growing RNA molecule. http://www.ig-msk.ru/RNA/kinetics/.

RNAstructure: a Windows implementation of the Zuker algorithm for RNA secondary structure prediction based on free energy minimization. http://rna.urmc.rochester.edu/rnastructure.html.

Sfold: a statistical sampling of all possible RNA secondary structures. http://www.bioinfo.rpi.edu/applications/sfold/srna.pl.

Vsfold4: a program that folds single RNA sequences using an extended energy model. http://www.rna.it-chiba.ac.jp/~vsfold/vsfold4/.

RNA Single Sequence 2D Prediction with Pseudoknots

HotKnots: a heuristic algorithm for the prediction of RNA secondary structures including pseudoknots. http://www.cs.ubc.ca/labs/beta/Software/HotKnots/.

Hpknotter: a heuristic approach for detecting RNA H-type pseudoknots. http://bioalgorithm.life.nctu.edu.tw/HPKNOTTER/.

KineFold: a folding kinetics program of RNA sequences including pseudoknots. http://kinefold.curie.fr/.

NUPACK: a dynamic programming algorithm based on the partition function for the prediction of a restricted class of RNA pseudoknots. http://www.acm.caltech.edu/~niles/software.html.

Pknots-RG: a dynamic programming algorithm for the prediction of a restricted class of RNA pseudoknots. http://bibiserv.techfak.uni-bielefeld.de/pknotsrg/.

Pknots: a dynamic programming algorithm for optimal RNA pseudoknot prediction using the nearest neighbour energy model. http://selab.wustl.edu/cgi-bin/selab.pl?mode=software#pk.

PLMM-DPSS: a high sensitivity method for RNA pseudoknot prediction using "Pseudoknot Local Motif Model and Dynamic Partner Sequence Stacking". http://bioinformatics.ist.unomaha.edu:8080/x/PLMM_DPSS.html.

Algorithm for 2D with Suboptimal Predictions

Barriers: a program to compute local minima and energy barriers of the RNA folding landscape. http://www.tbi.univie.ac.at/~ivo/RNA/Barriers/.

RNAshapes: a program to select unique suboptimal structures (shapes) based on an abstract representation of RNA secondary structure. http://bibiserv.techfak.uni-bielefeld.de/rnashapes/submission.html.

RNALOSS: a program for local and optimal secondary structure computation. http://clavius.bc.edu/clotelab/RNALOSS/.

RNAsubopt: a program to calculate all suboptimal secondary structures within a user defined energy range above the minimum free energy (mfe). http://www.tbi.univie.ac.at/~ivo/RNA/.

Algorithm for 2D with Suboptimal Predictions with Pseudoknots

MPGAfold: A massively parallel genetic algorithm for predicting RNA secondary structures with pseudoknots. http://binkley.ncifcrf.gov/~bshapiro/mpgaFold/mpgaFold.html.

RNA 2D Prediction by Alignments

BayesFold: a program to find, rank, and draw the likeliest structures for a sequence alignment. http://jaynes.colorado.edu/Bayes/.

ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. http://www.biophys.uni-duesseldorf.de/construct3/.

Garna: a program for predicting secondary structures of RNAs using a genetic algorithm. http://wwwmgs.bionet.nsc.ru/mgs/programs/2dstructrna/.

GPRM: program for finding common secondary structure elements in large families of unaligned RNA sequences. http://bioinfo.cis.nctu.edu.tw/service/gprm/.

Genebee: a RNA alignment folding program that uses a combination of free-energy and mutual information. http://www.genebee.msu.su/services/rna2_reduced.html.

Pfold: a program for folding alignments using a SCFG trained on rRNA alignments. http://www.daimi.au.dk/~compbio/rnafold/.

RNAalifold: a program for folding alignments using a combination of free-energy and a co-variation measures. http://www.tbi.univie.ac.at/~ivo/RNA/.

RNAlishapes: a tool for RNA structure analysis based on aligned RNA sequences. http://rna.cyanolab.de/.

RNA-Decoder and CORSmodel: programs for comparative prediction of RNA secondary structure. http://www.ebi.ac.uk/~meyer/rnadecoder/.

RNAGA: a program for predicting common RNA secondary structures using a genetic algorithm. http://bioweb.pasteur.fr/seqanal/interfaces/rnaga.html.

X2s: an X windows program for analyzing and editing an alignment of RNA sequences as well as predicting their RNA secondary structure. http://www.binf.ku.dk/~pgardner/bralibase/x2s.tar.gz.

RNA 2D Prediction by Alignments with Pseudoknots

Circles: an experimental Windows 95/98/NT program for inferring RNA secondary structure using a comparative method. http://taxonomy.zoology.gla.ac.uk/rod/circles/.

HXMATCH: a program for computing a consensus structure including pseudoknots based on an alignment of a few RNA sequences. http://www.tbi.univie.ac.at/papers/ SUPPLEMENTS/HXMATCH/.

ILM: an iterated loop matching program that evaluates stems in an alignment using a combination of free-energy and mutual information. http://www.cs.wustl.edu/~zhang/projects/rna/ilm/.

KnetFold: a program that computes a consensus RNA secondary structure from an RNA sequence alignment based on machine learning approaches. http://knetfold.abcc.ncifcrf.gov/.

Mifold: a matlab package for investigating mutual information content of RNA alignments. http://www.lcb.uu.se/~evaf/MIfold/.

Simultaneous Alignment and Structure Prediction Methods

CaRNAc: a program for comparative analysis combined with MFE folding. http://bioinfo.lifl.fr/RNA/carnac/.

Consan: a program to produce pairwise RNA secondary structural alignments. http://selab.wustl.edu/people/robin/consan/.

Cmfinder: an RNA motif prediction tool. http://bio.cs.washington.edu/yzizhen/CMfinder/.

COVE: an implementation of stochastic context free grammar methods for RNA sequence/structure analysis. http://selab.wustl.edu/cgi-bin/selab.pl?mode=software#cove.

Dynalign: a "full energy model" and comparative information to align and fold 2 two RNA sequences. http://rna.urmc.rochester.edu/dynalign.html.

Pmmatch: a program of the Vienna package, implements a variant of the Sankoff algorithm. http://rna.tbi.univie.ac.at/cgi-bin/pmcgi.pl.

Foldalign1: a program that predicts conserved local sequence and hair-pin structures using CONSENSUS and CLUSTAL-like heuristics. http://foldalign.kvl.dk/1.0/.

Foldalign2: a program to structurally align two sequences using a light-weight energy model in combination with RIBOSUM like score matrices. http://foldalign.ku.dk/.

RNAcast: a RNA consensus abstract shapes technique for multiple RNA folding. http://bibiserv.techfak.uni-bielefeld.de/rnacast/.

RNAmine: is a software tool for extracting the structural motifs from a set of RNA sequences. http://www.ncrna.org/RNAMINE/.

ScaRNA: a Stem Candidate Aligner for RNA that aligns two RNA sequences and calculates similarities, based upon estimated common secondary structures. http://www.scarna.org/scarna/.

SEED: a program that uses suffix arrays to enumerate complementary regions, possibly containing interior loops, as well for matching RNA secondary structure expressions. http://bio.site.uottawa.ca/software/seed/.

Slash: a tool that combines the programs COVE and Foldalign1. http://foldalign.kvl.dk/server/index.html.

Stemloc: a comparative RNA-structure finder that uses accelerated pairwise stochastic context-free grammars. http://biowiki.org/dart.

T-LARA: a program that produces a global fold and alignment of ncRNA families using integer linear programming and Lagrangian relaxation. http://www.planet-lisa.net/.

Simultaneous Alignment and Structure Prediction Methods with Pseudoknots

comRNA: a program for predicting common RNA secondary structure motifs in a group of related sequences. http://ural.wustl.edu/~yji/comRNA/.

APPENDIX D – RNA TERTIARY STRUCTURAL TOOLS

AMIGOS/PRIMOS/Qnifft: a series of tools for working with RNA/DNA structure files. http://www.pylelab.org/software/index.html.

ARTS: Alignment of RNA Tertiary Structures based on quadrats RNA representation. http://bioinfo3d.cs.tau.ac.il/ARTS/.

COMPADRES: a program to automatically find without prior knowledge, recurrent RNA motifs in three-dimensional structures. http://www.pylelab.org.

ERNA-3D: a program to create and manipulate RNA 3D coordinates. http://www.erna-3d.de/.

FR3D: a program to find small RNA motifs of 2 to 20 nucleotides from a PDB file. http://rna.bgsu.edu/FR3D/.

MANIP: a program that allows the rapid assembly of separate motifs into a complex three-dimensional architecture. http://www-ibmc.u-strasbg.fr/upr9002/westhof/download.html.

MC-Sym: a software that builds RNA 3-D structures using coordinates and relations between residues extracted from X-ray crystallography and NMR. http://www-lbit.iro.umontreal.ca/mcsym/.

NAB: a program to construct models for non-helical nucleic acids. http://www.scripps.edu/mb/case/.

NASSAM: a program that searches for motifs and formations of nucleic acid bases in 3D space within nucleic acid PDB formatted structures. http://202.185.55.243/grafss/ nassam/.

Ribostral an RNA 3D alignment analyzer and viewer based on base-pair isostericities. http://rna.bgsu.edu/ ribostral/.

RNA2D3D: a program that visualizes and compares RNA 3D structures available in the package StructureLab. http://www-lmmb.ncifcrf.gov/~bshapiro/software.html.

ROSETTA: a software suite for predicting and designing protein structures, protein folding mechanisms, and protein-protein interactions. http://www.bakerlab.org.

S2S: a program to display, manipulate and interconnect RNA sequence and structure data. http://bioinformatics.org/ S2S/.

YAMMP: a package for molecular simulations of reduced representation models. http://rumour.biology.gatech. edu/Programs/YammpWeb/.

AKNOWLEDGEMENTS

MAMR and EC acknowledge support from a Marie Curie International Reintegration Grant (FP6-039722) and Generalitat Valenciana (GV/2007/065).

REFERENCES

[1] Schuster P, Stadler PF, and Renner A. RNA structures and folding: from conventional to new issues in structure predictions. *Curr Opin Struct Biol* **1997**; 7: 229-35.

[2] Guerrier-Takada C, Gardiner K, Marsh T, Pace N, and Altman S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **1983**; 35: 849-57.

[3] Doudna JA and Cech TR. The chemical repertoire of natural ribozymes. *Nature* **2002**; 418: 222-8.

[4] Staple DW and Butcher SE. Pseudoknots: RNA structures with diverse functions. *PLoS Biol* **2005**; 3: e213.

[5] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **2004**; 116: 281-97.

[6] Dorsett Y and Tuschl T. siRNAs: applications in functional genomics and potential as therapeutics. *Nat Rev Drug Discov* **2004**; 3: 318-29.

[7] Doudna JA. Structural genomics of RNA. *Nat Struct Biol* **2000**; 7 Suppl: 954-956.

[8] Doty P, Boedtker H, Fresco JR, Haselkorn R, and Litt M. Secondary Structure In Ribonucleic Acids. *Proc Natl Acad Sci U S A* **1959**; 45: 482-99.

[9] Fresco JR, Alberts BM, and Doty P. Some molecular details of the secondary structure of ribonucleic acid. *Nature* **1960**; 188: 98-101.

[10] Kim SH, Suddath FL, Quigley GJ, *et al*. Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science* **1974**; 185: 435-40.

[11] Zuker M and Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* **1981**; 9: 133-48.

[12] Zuker M and Sankoff D. RNA secondary structure and their prediction. *Bulletin of Mathematical Biology* **1984**; 46: 591-621.

[13] Nussinov R, Piecznik G, Grigg JR, and Kleitman DJ. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics* **1978**; 35: 68–82.

[14] Nussinov R and Jacobson AB. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A* **1980**; 77: 6309-13.

[15] Michel F and Westhof E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J Mol Biol* **1990**; 216: 585-610.

[16] Berman HM, Battistuz T, Bhat TN, *et al*. The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* **2002**; 58: 899-907.

[17] Berman HM, Olson WK, Beveridge DL, *et al*. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys J* **1992**; 63: 751-9.

[18] Zwieb C and Muller F. Three-dimensional comparative modeling of RNA. *Nucleic Acids Symp Ser* **1997**: 69-71.

[19] Burks J, Zwieb C, Muller F, Wower I, and Wower J. Comparative 3-D modeling of tmRNA. *BMC Mol Biol* **2005**; 6: 14.

[20] Hofacker IL. Vienna RNA secondary structure server. *Nucleic Acids Res* **2003**; 31: 3429-31.

[21] Do CB, Woods DA, and Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **2006**; 22: e90-8.

[22] Dowell RD and Eddy SR. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics* **2006**; 7: 400.

[23] Leontis NB, Lescoute A, and Westhof E. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol* **2006**; 16: 279-87.

[24] Baker D and Sali A. Protein structure prediction and structural genomics. *Science* **2001**; 294: 93-6.

[25] Marti-Renom MA, Stuart AC, Fiser A, *et al*. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **2000**; 29: 291-325.

[26] Sali A and Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **1993**; 234: 779-815.

[27] Dill KA. Dominant forces in protein folding. *Biochemistry* **1990**; 29: 7133-55.

[28] Murthy VL and Rose GD. RNABase: an annotated database of RNA structures. *Nucleic Acids Res* **2003**; 31: 502-4.

[29] Schuwirth B, Borovinskaya M, Hau C, *et al*. Structures of the Bacterial Ribosome at 3.5 A Resolution. *Science* **2005**; 310: 827-834.

[30] Gao H, Ayub MJ, Levin MJ, and Frank J. The structure of the 80S ribosome from Trypanosoma cruzi reveals unique rRNA components. *Proc Natl Acad Sci U S A* **2005**; 102: 10206-11.

[31] Noller HF. RNA structure: reading the ribosome. *Science* **2005**; 309: 1508-14.

[32] Felden B. RNA structure: experimental analysis. *Curr Opin Microbiol* **2007**.

[33] Leontis NB and Westhof E. Geometric nomenclature and classification of RNA base pairs. *Rna* **2001**; 7: 499-512.

[34] Leontis NB, Stombaugh J, and Westhof E. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res* **2002**; 30: 3497-531.

[35] Yang H, Jossinet F, Leontis N, *et al*. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res* **2003**; 31: 3450-60.

[36] Murray LJ, Arendall WB, 3rd, Richardson DC, and Richardson JS. RNA backbone is rotameric. *Proc Natl Acad Sci U S A* **2003**; 100: 13904-9.

[37] Schneider B, Moravek Z, and Berman HM. RNA conformational classes. *Nucleic Acids Res* **2004**; 32: 1666-77.

[38] Hershkovitz E, Sapiro G, Tannenbaum A, and Williams LD. Statistical analysis of RNA backbone. *IEEE/ACM Trans Comput Biol Bioinform* **2006**; 3: 33-46.

[39] Griffiths-Jones S, Moxon S, Marshall M, *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **2005**; 33: D121-4.

[40] Klosterman PS, Tamura M, Holbrook SR, and Brenner SE. SCOR: a Structural Classification of RNA database. *Nucleic Acids Res* **2002**; 30: 392-4.

[41] Leontis NB and Westhof E. Analysis of RNA motifs. *Curr Opin Struct Biol* **2003**; 13: 300-8.

[42] Jossinet F, Ludwig TE, and Westhof E. RNA structure: bioinformatic analysis. *Curr Opin Microbiol* **2007**.

[43] Andreeva A, Howorth D, Brenner SE, *et al*. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **2004**; 32 Database issue: D226-9.

[44] Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **2000**; 25: 25-9.

[45] Bateman A, Coin L, Durbin R, *et al*. The Pfam protein families database. *Nucleic Acids Res* **2004**; 32 Database issue: D138-41.

[46] Liu J and Rost B. Domains, motifs and clusters in the protein universe. *Curr Opin Chem Biol* **2003**; 7: 5-11.

[47] Tian W and Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* **2003**; 333: 863-82.

[48] Rost B. Twilight zone of protein sequence alignments. *Protein Eng* **1999**; 12: 85-94.

[49] Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic local alignment search tool. *J Mol Biol* **1990**; 215: 403-10.

[50] Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997**; 25: 3389-402.

[51] Pearson WR and Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **1988**; 85: 2444-8.

[52] Thompson JD, Higgins DG, and Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **1994**; 22: 4673-80.

[53] Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **2004**; 5: 113.

[54] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **2004**; 32: 1792-7.

[55] Notredame C, Higgins DG, and Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **2000**; 302: 205-17.

[56] Hochsmann M, Voss B, and Giegerich R. Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE/ACM Trans Comput Biol Bioinform* **2004**; 1: 53-62.

[57] Liu J, Wang JT, Hu J, and Tian B. A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC Bioinformatics* **2005**; 6: 89.

[58] Siebert S and Backofen R. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics* **2005**; 21: 3352-9.

[59] Jansson J, Hieu NT, and Sung WK. Local gapped subforest alignment and its application in finding RNA structural motifs. *J Comput Biol* **2006**; 13: 702-18.

[60] Matsui H, Sato K, and Sakakibara Y. Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures. *Bioinformatics* **2005**; 21: 2611-7.

[61] Dost B, Han B, Zhang S, and Bafna V. Structural Alignment of Pseudoknotted RNA. *Proceedings of RECOMB 2006.* **2006**: 143-158.

[62] Freyhult EK, Bollback JP, and Gardner PP. Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res* **2007**; 17: 117-25.

[63] Eddy SR. Computational genomics of noncoding RNA genes. *Cell* **2002**; 109: 137-40.

[64] Gautheret D, Major F, and Cedergren R. Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. *Comput Appl Biosci* **1990**; 6: 325-31.

[65] Macke TJ, Ecker DJ, Gutell RR, *et al*. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res* **2001**; 29: 4724-35.

[66] Billoud B, Kontic M, and Viari A. Palingol: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. *Nucleic Acids Res* **1996**; 24: 1395-403.

[67] Pesole G, Liuni S, and D'Souza M. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics* **2000**; 16: 439-50.

[68] Sakakibara Y. Pair hidden Markov models on tree structures. *Bioinformatics* **2003**; 19 Suppl 1: i232-40.

[69] Sakakibara Y, Brown M, Hughey R, *et al*. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res* **1994**; 22: 5112-20.

[70] Eddy SR and Durbin R. RNA sequence analysis using covariance models. *Nucleic Acids Res* **1994**; 22: 2079-88.

[71] Lowe TM and Eddy SR. A computational screen for methylation guide snoRNAs in yeast. *Science* **1999**; 283: 1168-71.

[72] Klein RJ and Eddy SR. RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics* **2003**; 4: 44.

[73] Van de Peer Y, Van den Broeck I, De Rijk P, and De Wachter R. Database on the structure of small ribosomal subunit RNA. *Nucleic Acids Res* **1994**; 22: 3488-94.

[74] Eddy SR. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* **2002**; 3: 18.

[75] Griffiths-Jones S, Bateman A, Marshall M, Khanna A, and Eddy SR. Rfam: an RNA family database. *Nucleic Acids Res* **2003**; 31: 439-41.

[76] Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics* **1985**; 45: 810-825.

[77] Dayhoff MO. Atlas of Protein Sequence and Structure. *Natl. Biomed. Res. Found.* **1978**; 5: 345-352.

[78] Henikoff S and Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **1992**; 89: 10915-9.

[79] Smith AD, Lui TW, and Tillier ER. Empirical models for substitution in ribosomal RNA. *Mol Biol Evol* **2004**; 21: 419-27.

[80] Hudelot C, Gowri-Shankar V, Jow H, Rattray M, and Higgs PG. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. *Mol Phylogenet Evol* **2003**; 28: 241-52.

[81] Huelsenbeck JP and Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **2001**; 17: 754-5.

[82] Ronquist F and Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **2003**; 19: 1572-4.

[83] Tinoco I, Jr. and Bustamante C. How RNA folds. *J Mol Biol* **1999**; 293: 271-81.

[84] Mathews DH and Turner DH. Prediction of RNA secondary structure by free energy minimization. *Curr Opin Struct Biol* **2006**; 16: 270-8.

[85] Ding Y. Statistical and Bayesian approaches to RNA secondary structure prediction. *Rna* **2006**; 12: 323-31.

[86] Mathews DH. Revolutions in RNA secondary structure prediction. *J Mol Biol* **2006**; 359: 526-32.

[87] Reeder J, Hochsmann M, Rehmsmeier M, Voss B, and Giegerich R. Beyond Mfold: recent advances in RNA bioinformatics. *J Biotechnol* **2006**; 124: 41-55.

[88] Anfinsen CB. Principles that govern the folding of protein chains. *Science* **1973**; 181: 223-30.

[89] Mathews DH, Sabina J, Zuker M, and Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **1999**; 288: 911-40.

[90] Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *Rna* **2004**; 10: 1178-90.

[91] Wuchty S, Fontana W, Hofacker IL, and Schuster P. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **1999**; 49: 145-65.

[92] Giegerich R, Voss B, and Rehmsmeier M. Abstract shapes of RNA. *Nucleic Acids Res* **2004**; 32: 4843-51.

[93] McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **1990**; 29: 1105-19.

[94] van Batenburg FH, Gultyaev AP, and Pleij CW. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J Theor Biol* **1995**; 174: 269-80.

[95] Gultyaev AP, van Batenburg FH, and Pleij CW. The computer simulation of RNA folding pathways using a genetic algorithm. *J Mol Biol* **1995**; 250: 37-51.

[96] Shapiro BA, Wu JC, Bengali D, and Potts MJ. The massively parallel genetic algorithm for RNA folding: MIMD implementation and population variation. *Bioinformatics* **2001**; 17: 137-48.

[97] Lyngso RB and Pedersen CN. RNA pseudoknot prediction in energy-based models. *J Comput Biol* **2000**; 7: 409-27.

[98] Rivas E and Eddy SR. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol* **1999**; 285: 2053-68.

[99] Dirks RM and Pierce NA. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J Comput Chem* **2003**; 24: 1664-77.

[100] Dirks RM and Pierce NA. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *J Comput Chem* **2004**; 25: 1295-304.

[101] Cai L, Malmberg RL, and Wu Y. Stochastic modeling of RNA pseudoknotted structures: a grammatical approach. *Bioinformatics* **2003**; 19 Suppl 1: i66-73.

[102] Huang X and Ali H. High sensitivity RNA pseudoknot prediction. *Nucleic Acids Res* **2007**; 35: 656-63.

[103] Gaspin C and Westhof E. An interactive framework for RNA secondary structure prediction with a dynamical treatment of constraints. *J Mol Biol* **1995**; 254: 163-74.

[104] Gardner PP and Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* **2004**; 5: 140.

[105] Rousset F, Pelandakis M, and Solignac M. Evolution of compensatory substitutions through G.U intermediate state in Drosophila rRNA. *Proc Natl Acad Sci U S A* **1991**; 88: 10032-6.

[106] Kirby DA, Muse SV, and Stephan W. Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc Natl Acad Sci U S A* **1995**; 92: 9047-51.

[107] Chiu DK and Kolodziejczak T. Inferring consensus structure from nucleic acid sequences. *Comput Appl Biosci* **1991**; 7: 347-52.

[108] Gutell RR, Power A, Hertz GZ, Putz EJ, and Stormo GD. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res* **1992**; 20: 5785-95.

[109] Lindgreen S, Gardner PP, and Krogh A. Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics* **2006**; 22: 2988-95.

[110] Hofacker IL, Fekete M, and Stadler PF. Secondary structure prediction for aligned RNA sequences. *J Mol Biol* **2002**; 319: 1059-66.

[111] Knudsen B and Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res* **2003**; 31: 3423-8.

[112] Ruan J, Stormo GD, and Zhang W. ILM: a web server for predicting RNA secondary structures with pseudoknots. *Nucleic Acids Res* **2004**; 32: W146-9.

[113] Ruan J, Stormo GD, and Zhang W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics* **2004**; 20: 58-66.

[114] Havgaard JH, Lyngso RB, and Gorodkin J. The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucleic Acids Res* **2005**; 33: W650-3.

[115] Torarinsson E, Havgaard JH, and Gorodkin J. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics* **2007**.

[116] Mathews DH and Turner DH. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol* **2002**; 317: 191-203.

[117] Perriquet O, Touzet H, and Dauchet M. Finding the common structure shared by two homologous RNAs. *Bioinformatics* **2003**; 19: 108-16.

[118] Touzet H and Perriquet O. CARNAC: folding families of related RNAs. *Nucleic Acids Res* **2004**; 32: W142-5.

[119] Shapiro BA and Zhang KZ. Comparing multiple RNA secondary structures using tree comparisons. *Comput Appl Biosci* **1990**; 6: 309-18.

[120] Hochsmann M, Toller T, Giegerich R, and Kurtz S. Local similarity in RNA secondary structures. *Proc IEEE Comput Soc Bioinform Conf* **2003**; 2: 159-68.

[121] Duarte CM and Pyle AM. Stepping through an RNA structure: A novel approach to conformational analysis. *J Mol Biol* **1998**; 284: 1465-78.

[122] Sarver M, Zirbel CL, Stombaugh J, Mokdad A, and Leontis NB. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J Math Biol* **2007**.

[123] Lu XJ and Olson WK. Resolving the discrepancies among nucleic acid conformational analyses. *J Mol Biol* **1999**; 285: 1563-75.

[124] Lu XJ, Babcock MS, and Olson WK. Overview of nucleic acid analysis programs. *J Biomol Struct Dyn* **1999**; 16: 833-43.

[125] Gendron P, Lemieux S, and Major F. Quantitative analysis of nucleic acid three-dimensional structures. *J Mol Biol* **2001**; 308: 919-36.

[126] Lu XJ and Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res* **2003**; 31: 5108-21.

[127] Dror O, Nussinov R, and Wolfson H. ARTS: alignment of RNA tertiary structures. *Bioinformatics* **2005**; 21 Suppl 2: ii47-ii53.

[128] Dror O, Nussinov R, and Wolfson HJ. The ARTS web server for aligning RNA tertiary structures. *Nucleic Acids Res* **2006**; 34: W412-5.

[129] Freyhult E, Moulton V, and Clote P. RNAbor: a web server for RNA structural neighbors. *Nucleic Acids Res* **2007**.

[130] Ferre F, Lorenz WA, Ponty Y, and Clote P. DIAL: A web server for RNA structural alignment using dihedral angles. *Nucleic Acids Research* **2007**; in press.

[131] Harrison AM, South DR, Willett P, and Artymiuk PJ. Representation, searching and discovery of patterns of bases in complex RNA structures. *J Comput Aided Mol Des* **2003**; 17: 537-49.

[132] Duarte CM, Wadley LM, and Pyle AM. RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res* **2003**; 31: 4755-61.

[133] Wadley LM and Pyle AM. The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic Acids Res* **2004**; 32: 6650-9.

[134] Shapiro BA, Yingling YG, Kasprzak W, and Bindewald E. Bridging the gap in RNA structure prediction. *Curr Opin Struct Biol* **2007**.

[135] Massire C and Westhof E. MANIP: an interactive tool for modelling RNA. *J Mol Graph Model* **1998**; 16: 197-205, 255-7.

[136] Jossinet F and Westhof E. Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics* **2005**; 21: 3320-1.

[137] Macke TJ and Case D, Modeling unusual nucleic acid structures., In: Molecular Modeling of Nucleic Acids, Leontis N and SantaLucia J, Editors. American Chemical Society: Washington, DC. 1998; 379-393.

[138] Cheatham TE, 3rd and Young MA. Molecular dynamics simulation of nucleic acids: successes, limitations, and promise. *Biopolymers* **2000**; 56: 232-56.

[139] Major F. Building three-dimensional ribonucleic acid structures. *Computing in Science & Engineering* **2003**; 5: 44-53.

[140] Yingling YG and Shapiro BA. The prediction of the wild-type telomerase RNA pseudoknot structure and the pivotal role of the bulge in its formation. *J Mol Graph Model* **2006**; 25: 261-74.

[141] Das R and Baker D. Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci U S A* **2007**.

[142] Bradley P, Malmstrom L, Qian B, *et al.* Free modeling with Rosetta in CASP6. *Proteins* **2005**.

[143] Leontis NB, Altman RB, Berman HM, *et al.* The RNA Ontology Consortium: an open invitation to the RNA community. *Rna* **2006**; 12: 533-41.

[144] Pieper U, Eswar N, Davis FP, *et al.* MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* **2006**; 34: D291-5.

[145] Kopp J and Schwede T. The SWISS-MODEL Repository: new features and functionalities. *Nucleic Acids Res* **2006**; 34: D315-8.

[146] Eswar N, John B, Mirkovic N, *et al.* Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res* **2003**; 31: 3375-80.

[147] Schwede T, Kopp J, Guex N, and Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* **2003**; 31: 3381-5.

[148] Pedersen JS, Bejerano G, Siepel A, *et al*. Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Comput Biol* **2006**; 2.

[149] Washietl S, Hofacker IL, and Stadler PF. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* **2005**; 102: 2454-2459.

[150] Wiese KC, Glen E, and Vasudevan A. JViz.Rna--a Java tool for RNA secondary structure visualization. *IEEE Trans Nanobioscience* **2005**; 4: 212-8.

Received: February 22, 2007 Revised: May 30, 2007 Accepted: May 30, 2007