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 miRNAs and small interfering RNA, is transforming our thinking about how gene expression is regulated. Thus, RNA molecules are now known to carry a large repertoire 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] have 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 methods 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.
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 phosphate (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.
It is possible that there might be errors in the backbone atoms, which may contain errors. Indeed, the authors of RNABase have analyzed and classified RNA torsion angles to conclude that an average one error can occur every two bases [28].

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 Structural 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, Murphy 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 the same 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.
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
RNA secondary structure is conserved between divergent sequences [62] and helps to confer 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 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, has 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 structure prediction problem [76]. The rationale behind such approach is the detection of a common base pair list by maximizing the sum of its basepair 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 [11-14]. 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 $O(N^2)$ time and $O(N^3)$ space, which makes it only usable for short sequences [98]. The partition function approach implemented by Dirks [99, 100] has an $O(N^4)$ 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 \(\eta(C_4'-i+1-P_i-C_4'-P_{i+1})\) and \(\theta(P_i-C_4'-i-P_{i+1}-C_4'-i+1)\) 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 \(\eta\) and \(\theta\) angles as well as the sugar pucker phase. The COMPADRES algorithm has been applied for identifying new RNA motifs such as \(p\)-turns, \(\Omega\)-turns, \(\alpha\)-loops, \(C2FA\) and \(\text{Hook} \) turns.

More recently, the ARTS [127, 128] and the DIAL [130] programs for structural comparison of RNA molecules 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 of 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 restraints 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


 Alignment and Consensus Secondary Structure Databases


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://rmp.uthct.edu/rmp/SRPDB/SRPDB.html.


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


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


RNA Alignment Methods by 2D Structural Information


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

Alignment Methods for RNA 2D Structures with Pseudoknots


RNA Searching Methods by Sequence/Structural Information


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


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


RNA Phylogenetic Analysis Tools

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


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/.


APPENDIX C – RNA SECONDARY STRUCTURE PREDICTION TOOLS

RNA Folding Software


RNA Single Sequence 2D Prediction with Pseudoknots


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/.


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


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/.


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**


**Simultaneous Alignment and Structure Prediction Methods**


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


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/.


**APPENDIX D – RNA TERTIARY STRUCTURAL TOOLS**


MANIP: a program that allows the rapid assembly of separate motifs into a complex three-dimensional architecture. http://www-ibmc.u-strasbg.fr/uptf9002/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/.


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


AKNOWLEDGEMENTS

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

REFERENCES


