Shedding light on the dark matter of the biomolecular structural universe: Progress in RNA 3D structure prediction

Fabrizio Pucci[a], Alexander Schug[a,*]

[a] *John von Neumann Institute for Computing, Jülich Supercomputer Centre,*
Forschungszentrum Jülich, 52428 Jülich, Germany

Abstract

Structured RNA plays many functionally relevant roles in molecular life. Structural information, while required to understand the functional cycles in detail, is challenging to gather. Computational methods promise to complement experimental efforts by predicting three-dimensional RNA models. Here, we provide a concise view of the state of the art methodologies with a focus on the strengths and the weaknesses of the different approaches. Furthermore, we analyzed the recent developments regarding the use of coevolutionary information and how it can boost the prediction performances. We finally discuss some open perspectives and challenges for the near future in the RNA structural stability field.

Keywords: RNA structure prediction, Computational modeling, Direct Coupling Analysis, Coevolution

1. Introduction

In the last two decades growing attention has been dedicated to the understanding of RNA. As for proteins, RNA structure and function are closely tied and play a determining role in many biomolecular processes such as the splicing process, transcriptional and translational machineries, and RNA localization and decay [1]. Despite this importance, the number of available experimental RNA structures at an atomic level stored in public databases such as the Protein Data Bank (PDB) [2] or the Nucleic Acids Database (NDB) [3] remains limited due to challenging experimental problems related to the preparation and/or crystallization of RNAs that are usually more flexible and dynamic with respect to proteins [4]. Currently, more than 90% of structures stored in the PDB database [2] are proteins, while less than 5% of the human genome encodes for proteins. This discrepancy has stirred the curiosity of scientists and lead to the remaining 95% of the human genome sometimes being referred to as the dark matter of the genome [5, 6].

*corresponding author

Email addresses: f.pucci@fz-juelich.de (Fabrizio Pucci), al.schug@fz-juelich.de (Alexander Schug)

To overcome the lack of structurally resolved RNA, computational methods have complemented experimental efforts to get more insight into how RNA structure and dynamics determine its functions [7, 8, 9]. Significant efforts have been devoted to the construction of methods to predict the RNA secondary structure mainly employing thermodynamics-based models [10]. These methods have been recently achieved significant improvements by the incorporation of auxiliary structural information from high-throughput chemical probing technologies [11, 12].

However, even if the knowledge of the RNA secondary structure provides important information, it is not sufficient to fully explain RNA function or interactions with other biomolecules [13]. During the last years lot of attention has been focused on the construction of RNA 3D structure prediction tools of increasing accuracy and speed [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26].

In this review we provide a concise overview of these methodologies, present their strengths and limitations, and highlight the open challenges in RNA structure prediction. We will particularly underline the recent development related to the use of coevolutionary information to improve the accuracy of the RNA 3D structure prediction methods. The structural information remains, however, static and provided one piece to the puzzle of RNA function. Another important component is the dynamics of RNA, for example while undergoing large conformational rearrangements [27, 28], which is exhaustively covered in the excellent review [29].

2. From the RNA sequence to its 3D structure

The basic unit of RNA is the nucleotide that is formed by planar aromatic rings linked to a ribose unit that in turn is attached to a phospate group (see fig 1). The sequence of the different constituent nucleotides (adenine, guanine, cytosine, uracil) of a given RNA molecule is defined as its **primary structure**.

Nucleotides typically complement each other by forming the canonical base pairs A-U and C-G, which maximizes inter-nucleotide hydrogen bonding. This leads to short chains of nucleotides folding in anti parallel double helices. The nucleotides that do not form Watson-Crick base pairs can remain unpaired or establish less stable non-canonical base pairs forming internal and bulge loops, hairpins and junctions. The **secondary structure** is thus essentially defined as the set of base pairs occurring in the RNA molecules.

The **tertiary structure** is the complete set of three-dimensional coordinates of all atoms of the RNA structure. This includes formation of a plethora of tertiary motifs such as pseudoknots, stacking of helices, multiple base pairing, ribose zipper and loop-loop interactions that determine the molecular shape in the physical space.

An accurate computational prediction of the RNA tertiary structure starting from its sequence is particularly challenging as the 3D structure depends not only on the sequence but also on the environmental conditions such as the ion concentrations and temperature.

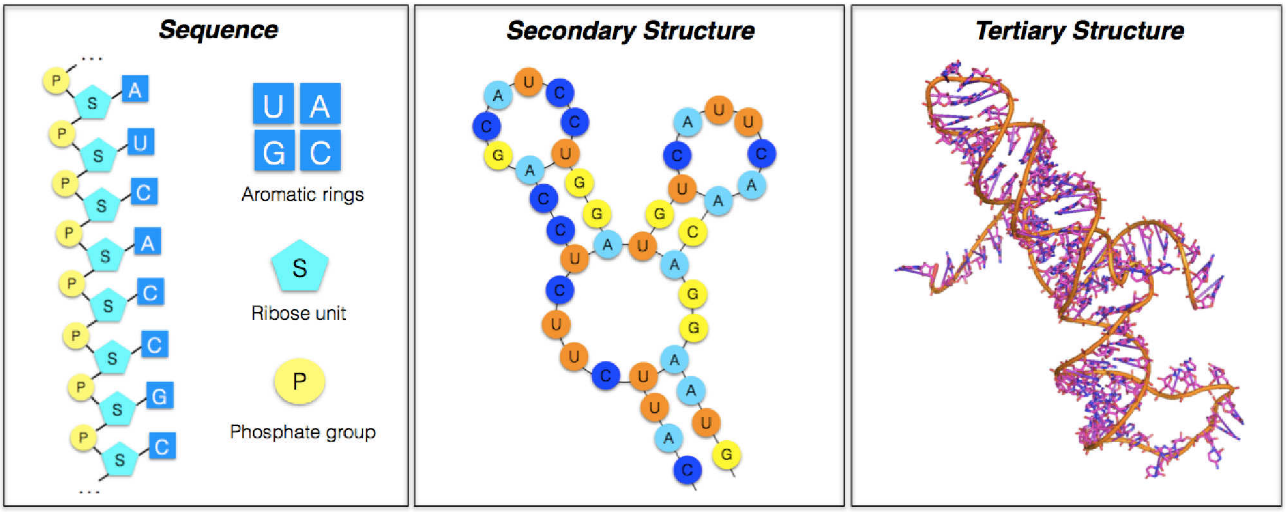


Figure 1: From primary (sequence), to secondary and to tertiary RNA structures.

3. Computation modeling of RNA 3D structure

Here, we review and compare some widely known methods for the prediction of the three-dimensional structure of RNA. The available approaches can be roughly divided in three different types: fragement-based, physics-based and comparative modeling. To compare the state of the art prediction methods and assess their performance, a blind experiment for the RNA 3D structure prediction has been established in the last years [30, 31, 32] with the last round focused on the challenging prediction of six RNA structures of riboswitches and ribozymes [32].

3.1. Fragment-based homology methods

The main idea behind this approach is to assemble the 3D prediction of target molecules using small fragments from libraries with similar sub-sequences. The theoretical justification of such a procedure comes from assuming that the distribution of the different conformations observed in known RNA structures for given fragment sequences is a good approximation for the conformation of similar or identical sub-sequences.

The basics steps of these methods consist first in the fragmentation of the secondary structure used as input. As a second step a search algorithm is employed to match these elements from fragment libraries constructed from databases of known RNA structures. Finally, all the elements are assembled together using different algorithms (see below) and, usually, a final refinement stage using atomic force field or coarse-grained potentials is performed.

One advantage of these methods is their computational efficiency as the fragmentation assembly drastically reduces the conformational search space. As the structural diversity of the fragment library directly limits the accuracy of the composed assembly, good results require a large and diverse library as well as a good scoring function. Here, we list methods belonging to this class and some of their characteristics.

- **RNAComposer** [15]: after the fragmentation step, the predicted secondary structure elements (stem, loops and single strands) constitute the input pattern for a search in the FRABASE 3D fragment data-set developed by the authors. From the matched elements a 3D structure is constructed by first superimposing and then merging them. Finally, an energy minimization is performed in the CHARMM force field [33].

- **Vfold3D** [18] uses a coarse-grained representation of the RNA. First, it utilizes VFold2D, a free energy-based model, to predict the secondary structure from which it extracts motifs (helices, hairpin loops, internal loops,...). From these motifs it searches the best template in the VFoldMTF database. After assembling the 3D structural motifs and the addition of all atoms to the coarse-grained structure (according to the template) it performs an all-atom structure refinement.

- **3dRNA** [21] uses a two-steps procedure where first the smallest secondary elements (SSEs) are assembled in hairpins and duplexes one by one following the 5' to 3' end direction. Then, these structures are further assembled into a complete tertiary structure by selecting the junction component from a junction database. Finally, to assure the chain connectivity, the assembled model is energy minimized in the AMBER 98 force field [34].

- **FARNA** [22] (Fragment Assembly of RNA) also uses a coarse-grained representation of the RNA structure and a fragment assembly strategy employing a Monte Carlo process that is guided by a low-resolution knowledge-based energy function. The authors developed knowledge-based base-pairing and base-stacking potentials to which they add several other terms such as the penalty for steric clashes. The structural model undergoes a second refining step in an all-atom potential to improve the accuracy and to better discriminate competing structural models. The two-step protocol is called **FARFAR** (Fragment Assembly of RNA with Full Atom Refinement) and is part of the ROSETTA package.

- **MC-Fold/MC-Sym** [25] this pipeline uses the combination of small motifs called nuclotide cyclic motifs (NCMs). The NCM-3D fragments are assigned to the given sequence by choosing the structure with higher probability of occurencies. Then the structural NCMs are concatenated using a Las Vegas probabilistic algorithm.

3.2. Physics-based methods

In contrast to the previous methods, the physics-based models do not use template structures in the assembly of RNA fragment/motifs but derive and parameterize energy functions depending on specific conformations, similar to approaches applied for proteins [35, 36]. These methods can be further separated in *ab-initio* approaches or knowledge-based approaches. In the latter methods, the energetic functions are derived using the inverse Boltzmann law from the probability of occurrences of certain sequence-structure elements in a dataset of known structures. In constrast, the *ab-initio* methods are based on force fields adopting usually harmonic potentials for bond lengths and angles, Lennard-Jones potentials for Van der Waals interactions, and electrostatic potentials that get reparameterized based on RNA structure and thermodynamics data.

Such energetic functions are then used in Molecular Dynamics (MD) simulations or Monte Carlo (MC) minimization often associated with enhanced sampling techniques such as temperature replica exchange or discrete molecular dynamics simulations in which the energy function is substituted with discrete step function potentials that drastically reduce the computational cost of the method.

The strength of the physics-based methods is that they are applicable to sequences with no known similar sequences or even sub-sequences. Their disadvantage is their need to explore a large conformational search space which increases computational demands and decreases their computational efficiency in comparison to fragment-based methods.

Here in the following the list of the computational tools that use this approaches.

- **iFoldRNA** [20] uses a simplified ”three-bead per nucleotide” representation of the RNA structure, and it is based on a replica-exchange discrete molecular dynamic (DMD) simulations protocol to span the conformational space. DMD incorporates base-pairing and base-stacking interactions into an energy function where in addition an entropic estimation of the loop formation is also considered.

- **NAST** [23] uses coarse-grained representation of RNA considering one quasi-atom per RNA nucleotide base. A simplified knowledge-based energy function, derived from the observed RNA geometries at the nucleotide level, is used to predict the target structure by global energy minimization. NAST necessitates as input the (known or predicted) secondary structure information and accepts also tertiary contacts to guide the folding.

- **SimRNA** [17] uses a coarse-grained representation of the RNA structures reducing the number of explicitly represented atoms per residue from about thirty to only five. It is based on dedicated RNA statistical potentials to compute the structure free energy and identify the native structure via Monte Carlo sampling.

3.3. Comparative homology-based modeling

Another type of methods uses homology modeling approaches by identifying structurally related template and geometrically aligning residues from the target onto corresponding residues in the template. Examples of this type of methods are **RNABuilder** [37] and **ModeRNA** [16]. The latter makes also extensively use of the evolutionary information by using multiple RNA sequence alignments to better reveal patterns of conservation that improve the accuracy of the prediction starting from the 3D template.

In order to improve the accuracy of homology-based methods it has been shown that the addition of multiple templates can be successfully employed. Another characteristic that is common to this type of methods is that they model (short) regions with no template by employing fragment-based insertion approaches. Finally the methods perform usually a geometry optimization using a force field in order to obtain physically reasonable conformations. The RNABuilder method for example uses a multi-resolution approach that handles at different level of resolution the forces, rigidifying certain bonds, residues or molecules part while keeping flexible the others.

The major drawback of this class of methods resides in the difficulty of having the template structure for the given sequence and an informative multiple sequence alignment. Indeed for the RNA structure templates there is the limitation of the number of structure deposited in the widely known database [3]. Regarding the alignments one can use those available for many RNA families in the Rfam database [38] or perform alignment via commonly used multiple RNA sequence alignment packages such as R-Coffe [39], Muscle [40] or Infernal [41].

The strength of these methods is their high accuracy with modest computational costs when good structural templates can be found. Their performance drops in the absence of such templates.

3.4. Performance assessment and RNA-Puzzles prediction

Most of the analyzed RNA structure prediction methods participated in the RNA-Puzzle competitions [30, 31, 32] in which a set of experimentally resolved RNA 3D structures had to be blindly predicted. To assess the performance of the predictors and rank the models, different metrics have been used such as the root mean square deviation (RMSD) between the predicted the experimental crystal structures that gives a more global information about the model's accuracy or the deformation index and the complete deformation profile matrix that instead capture the "local" accuracy at the nucleotide interaction level.

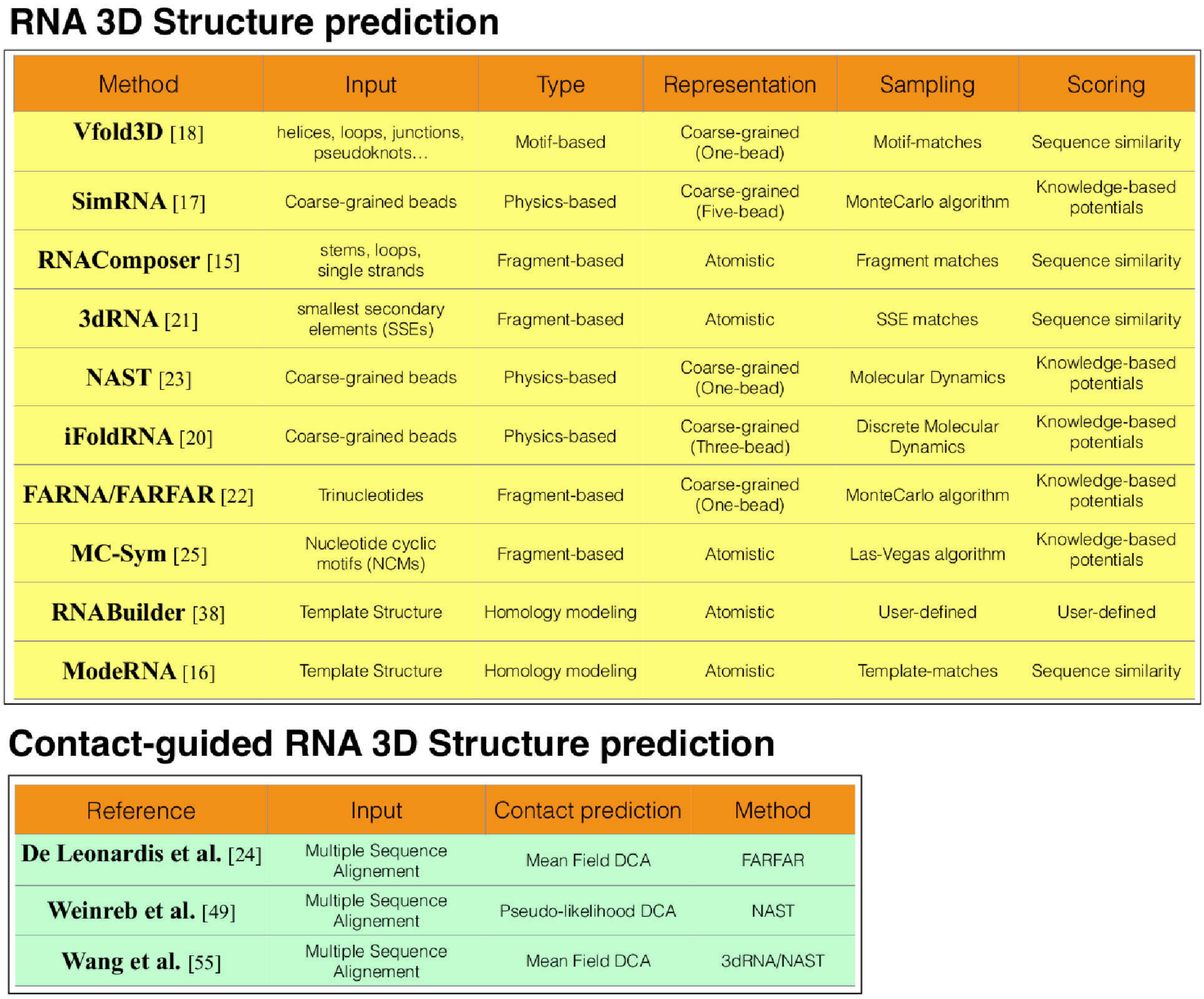
In the first RNA-Puzzle round [30] in addition to two simple small targets that were relatively well predicted, the more challenging riboswitch structure were not accurately reproduced with a mean RMSD accuracy of about 15 Å. Moreover while most methods achieve good performance on Watson-Crick base pairs, non-Watson-Crick interactions remain difficult to predict and clash score remains generally quite high.

In the second RNA-Puzzle round [31] the best RMSDs for a long nucleotide sequence range between 6.8 and 11.7 Å indicating a global improvement of methods' performance. A substantial amelioration for non-Watson-Crick interactions prediction is also observed.

Finally in the last RNA-Puzzle competition [32], the predictions achieved a consistently high level of accuracy especially when a high-homology template can be identified. For example in the case of the SAM-I riboswitch aptamer prediction that has as template (PDB code 3QIR), the average RMSD over all predicted models is about 4.3 Å, with a standard deviation of less then 2 Å.

Unfortunately, when the homology with the template is not high enough, the accuracy of methods is still not satisfactory and depends on the length of the RNA sequence. Small RNA sequences can be predicted with good accuracy as exemplified in the case of the ZTP riboswitch predicted with an averaged RMSD of about 6 Å and as also shown in the previous RNA-Puzzle round. For long sequences such as the *ydaO* riboswitch no method is capable of reliably predicting the native three dimensional conformation with an average RMSD of about 16 Å.

In order to improve the structure prediction of these challenging targets, there is a need for new and more performing algorithms. In the next section we will thus present recent progress in this direction and more in detail we will show how the coevolutionary information can been used to improve significantly the methods' accuracy.



RNA 3D Structure prediction

Method	Input	Type	Representation	Sampling	Scoring
Vfold3D [18]	helices, loops, junctions, pseudoknots...	Motif-based	Coarse-grained (One-bead)	Motif-matches	Sequence similarity
SimRNA [17]	Coarse-grained beads	Physics-based	Coarse-grained (Five-bead)	MonteCarlo algorithm	Knowledge-based potentials
RNAComposer [15]	stems, loops, single strands	Fragment-based	Atomistic	Fragment matches	Sequence similarity
3dRNA [21]	smallest secondary elements (SSEs)	Fragment-based	Atomistic	SSE matches	Sequence similarity
NAST [23]	Coarse-grained beads	Physics-based	Coarse-grained (One-bead)	Molecular Dynamics	Knowledge-based potentials
iFoldRNA [20]	Coarse-grained beads	Physics-based	Coarse-grained (Three-bead)	Discrete Molecular Dynamics	Knowledge-based potentials
FARNA/FARFAR [22]	Trinucleotides	Fragment-based	Coarse-grained (One-bead)	MonteCarlo algorithm	Knowledge-based potentials
MC-Sym [25]	Nucleotide cyclic motifs (NCMs)	Fragment-based	Atomistic	Las-Vegas algorithm	Knowledge-based potentials
RNABuilder [38]	Template Structure	Homology modeling	Atomistic	User-defined	User-defined
ModeRNA [16]	Template Structure	Homology modeling	Atomistic	Template-matches	Sequence similarity

Contact-guided RNA 3D Structure prediction

Reference	Input	Contact prediction	Method
De Leonardis et al. [24]	Multiple Sequence Alignement	Mean Field DCA	FARFAR
Weinreb et al. [49]	Multiple Sequence Alignement	Pseudo-likelihood DCA	NAST
Wang et al. [55]	Multiple Sequence Alignement	Mean Field DCA	3dRNA/NAST

Figure 2: 3D RNA structure prediction methods and their principal characteristics

4. Including evolutionary information to improve 3D structure prediction

4.1. Residue co-evolution and contact prediction

A significant amount of data obtained from high-throughput sequencing technologies provides us an invaluable source of evolutionary information that can be used in order to improve the protein [44, 45, 42, 50, 46] and RNA structure prediction [24, 49]. The basic idea behind these approaches is tracing co-variation of amino acid or nucleic acid pairs in proteins and RNA belonging to homologous families. Such co-variation indicates structural proximity of the involved residues and is hence related to biomolecular structural and stability properties. Compensatory mutations occur when a mutation with a detrimental effect at a given site, interact with a secondary mutations at another site to restore the molecular fitness [47] thus indicating the tendency of co-evolving residues to represent physical interactions that are important for the stability and function of biomolecules.

In the last decade many statistical methods have been developed to identify co-evolving residue pairs in a multiple sequence alignment (MSA) [48]. One can assume that such correlation occurs due to the spatial proximity of the two residues even if it can also arise from indirect effects related to the transitivity of the interaction between pairs and tertiary residues.

The use of the statistical methods such as maximum entropy models (MEM) or direct coupling analysis [24, 44, 46, 50] allows to unravel the transitive effects in the network of constrained residue-residue interactions and thus they give more efficient and robust contact-prediction. Using these statistical-based approaches one can detect long-range tertiary contacts from sequence covariation whose prediction difficulty has been one of the main limitation to the advancement of the computational RNA 3D structure prediction methods.

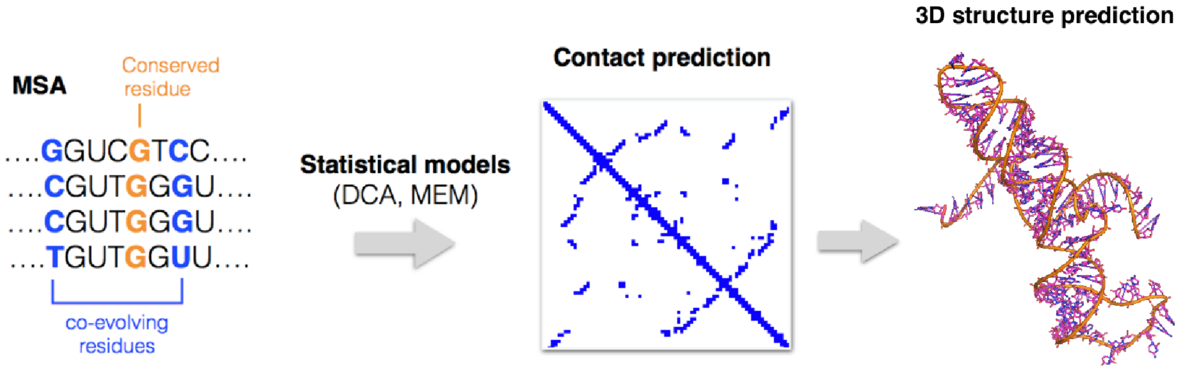


Figure 3: Statistical-based contact prediction from coevolutionary data improved the 3D RNA structure prediction

4.2. Direct coupling analysis (DCA)

The basic assumption of this method is to associate the probability of observation $P(\sigma)$ of a given sequence $\sigma = (a_1,\, a_2\, ...a_L)$ of length L in a MSA to the Hamiltonian energetic function $H(\sigma)$ using the Boltzmann law

$$P(\sigma) = \frac{1}{\mathcal{Z}} e^{-\beta H(\sigma)} \tag{1}$$

where β is the temperature and $\mathcal{Z}$ is the partition function of the system, and where the Hamiltonian is assumed to have the following simplified form

$$H(\sigma) = -\sum_{i}^{L} h_i(a_i) - \sum_{i}^{L-1} \sum_{j=i+1}^{L} J_{ij}(a_i, a_j) \tag{2}$$

consisting only of single site terms, *i.e.* $h_i(a_i)$, and residue pair interactions $J_{ij}(a_i, a_j)$. These parameters can be inferred from the MSA using a plethora of different approaches. For example in [49] a pseudo-maximum likelihood (pmlDCA) approximation have been employed, while a computational intensive message-passing algorithm (mpDCA) is used in [44] and a more efficient mean field algorithm (mfDCA) in [50]. The list of other type of popular algorithm used in the inverse inference step can be found in [24].

There are also pitfalls. Frequently, some species are over-represented in the MSA, e.g. because of their medical importance or the ease of handling them experimentally. Thus, these sequences need to be re-weighted. In addition, the quality of the MSA such as the proper placement of gap regions influences contact prediction accuracy. This loss of contact prediction precision directly leads to a decreased quality of 3D prediction. Another drawback is that the DCA prediction of the tertiary contacts is far from being perfect with only a modest overall true positive (TP) predicted contacts; it should be noted, however, that only relatively few (O(10)) higher ranking pairs that show higher TP rate are already sufficient to boost the performance of the structural modeling. Still, these methods significantly boost performance without too much computational effort.

4.3. Contact guided 3D RNA-structure prediction

While the use of coevolutionary data has been already fruitfully applied to protein structure determination during the last decade [42, 43, 44, 45, 46, 50, 51, 52, 53], the contact guided prediction of the three-dimensional RNA structure is relatively new. Indeed, the previous mutual information (local) approach to the extraction of coevolution signals from MSA was not sufficiently accurate [56] to provide reliable tertiary contact predictions.

Recent investigations [24, 49, 55] instead show that the use of a global approach to extract the top-ranked site-pairs with stronger co-evolutionary signals can be efficiently employed as distance constraints in modeling tools.

In [24] the authors show that in the prediction of the structure of six representative riboswitches with the Rosetta-based method FARFAR, the use of predicted tertiary contacts by mfDCA improves the RMSD in average by about 30% with respect to the case in which only secondary structure information (SSI) is provided. In figure 4 we report this explicit comparison for all the six structure considered.

These results have been confirmed in [42] where the authors show a significant improvement of prediction quality when the evolutionary based contact prediction computed via the pmlDCA approach. In this work, contacts are used as spatial constraints in the NAST coarse-grained structure prediction method. A further confirmation in [55] highlights that prediction RMSDs for the same structures as analyzed in [24, 42] are lowered by about 30% when using the tertiary contacts predicted via mfDCA in the 3dRNA method [21] compared to not using such tertiary contact constraints. [42] and [57] also demonstrated how

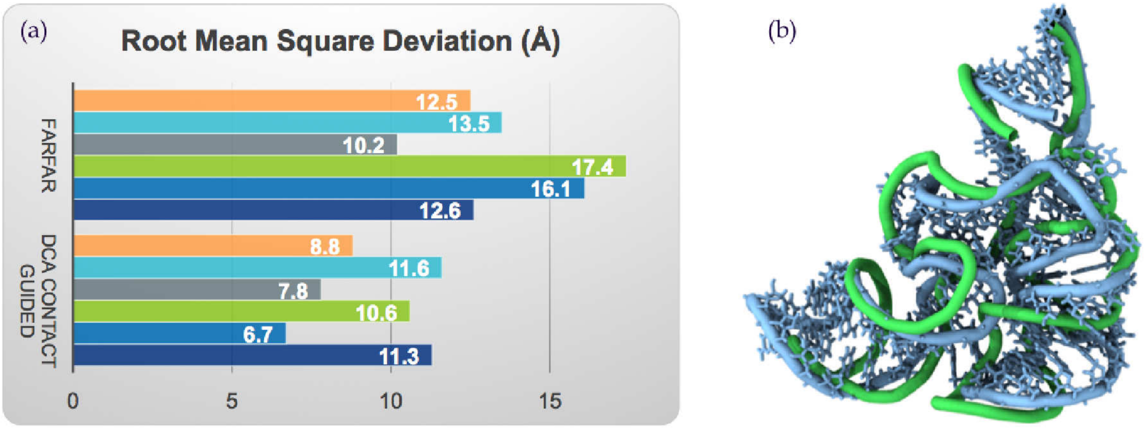


Figure 4: (a) DCA contact-guided RNA structure prediction improvement with respect to the state of the art (Rosetta-based) method for the six riboswitches from [24]. (b) Overlay of the DCA-contact guided predicted (blue) and the experimental structure (green) for the thiamine pyrophosphate-specific (TPP) riboswitch (PDB code 2gdi). In the prediction, the first 100 top contacts as computed via mean field DCA from the MSA of the RF00059 family have been used as constraints in the FARFAR method.

DCA-based methods show good accuracy in the prediction of intermolecular RNA-protein contacts.

5. Future challenges and outlook

Even if in the last decade tremendous advances has been achieved in RNA structure prediction, its accuracy still is not as high as for protein structure prediction. Moreover there are open and intriguing challenges in the field that will hopefully be tackled in the close future:

- The role played by the environmental conditions such as ions that strongly influence the RNA structure has to be fully investigated and clarified [58, 59, 60]. Since *in vivo* RNA can adapt different conformations with respect to *in vitro* ones, this will be also important to understand such differences and give important information for RNA biology.

- In the next years, thanks to the advancement of the next generation sequencing technologies, the amount of sequence information will continue to increase exponentially. Currently coevolutionary methods focus on the prediction of two-site interactions (contacts), but this increased amount of information promises to also allow to predict higher order correlations that could further boost structure prediction methods.

- Further improvements of RNA force fields will continue to increase the accuracy of predictions. These can help to better understand the role of the different RNA conformations, their stability and to gain new insights about the RNA structural dynamics.

- Combining structure prediction methods or simulations with experimental data such as Selective 2-hydroxyl acylation analyzed by primer extension (SHAPE) [61], Fluorescence Resonance Energy Transfer (FRET) [62] or small angle X-Ray scattering (SAXS) [63, 64, 65] will allow to probe RNA structures where a single method fails [66].
- Inter-molecular protein interactions and contacts can be predicted via DCA and related methods [67]. This could be transferred to RNA.
- Finally, it becomes more and more clear that base modifications such as the methylation or deamination play an important role in RNA biology by modifying the structure as well as the function of RNA. It could be thus of great interest in the next future to address and investigate these (epi)transcriptomics data to better understand all biological processes in which the RNA is involved.

References

[1] Y. Wan, M. Kertesz, R.C. Spitale, E. Segal, H. Chang, Understanding the transcriptome through RNA structure, Nat. Rev. Genet. 12, 10.1038 (2011).

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, Shindyalov I.N., Bourne P.E.. The Protein Data Bank H.M., Nucl. Ac. Res., 28: 235-242 (2002).

[3] B.C. Narayanan at al., The Nucleic Acid Database: new features and capabilities, Nucl. Ac. Res. 42, D114-D122 (2013).

[4] R. Schroeder, A. Barta, K. Semrad, Strategies for RNA folding and assembly. Nat. Rev. Mol. Cell. Biol. 5, 908-919 (2004).

[5] P. Kapranov, G. St Laurent, Dark Matter RNA: Existence, Function, and Controversy. Front. Genet. 23, 10.3389/fgene.2012.00060 (2012).

[6] K.R. Chi, The dark side of the human genome. Nature 538, 10.1038/538275a (2016).

[7] M. Zuker, P.O. Stiegler, Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, Nucleic Acids Res. 9, 133-48 (1981).

[8] K. Aigner, F. Dressen, G. Stege, Methods for Predicting RNA Secondary Structure, RNA 3D Structure Analysis and Prediction, 19-41, Springer (2012).

[9] S.A. Mortimer, M.A. Kidwell, J.A. Doudna, Insights into RNA structure and function from genome-wide studies, Nat Rev Genet. 15, 469-79 (2014).

[10] J.S. Reuter and D.H. Mathew, RNAstructure: software for RNA secondary structure prediction and analysis, BMC Bioinformatics 11, 129 (2010).

[11] P. Ge, S. Zhang, Computational analysis of RNA structures with chemical probing data, Methods. 79-80,60-6 (2015).

[12] R. Lorenz, M.T. Wolfinger, A. Tanzer, I.L. Hofacker, Predicting RNA Secondary Structures from Sequence and Probing Data, Methods 103, 86-98 (2016).

[13] C. Laing, T.J. Schlick, Computational approaches to 3D modeling of RNA, Phys Condens Matter. 22, 283101 (2010).

[14] M. Rother, K. Rother, T. Puton, J.M. Bujnicki, ModeRNA: a tool for comparative modeling of RNA 3D structure, Nucl. Ac. Res. 39, 4007-22 (2011).

[15] M. Popenda et al., Automated 3D structure composition for large RNAs, Nucl. Ac. Res. 40, e112 (2012). M. Biesiada, K.J. Purzycka, M. Szachniuk, J. Blazewicz, R.W. Adamiak, Automated RNA 3D Structure Prediction with RNAComposer, Methods Mol. Biol. 1490, 199-215 (2016).

[16] M. Rother, K. Rother, T. Puton, J.M. Bujnicki, ModeRNA: a tool for comparative modeling of RNA 3D structure, Nucl. Ac. Res. 39, 4007-22 (2011).

[17] M.J. Boniecki et al., SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction, Nucl. Ac. Res. 20, 44 (2016).

[18] C. Zhao, X. Xu, S.J. Chen, Predicting RNA Structure with Vfold, Methods Mol. Biol. 1654, 3-15 (2017).

[19] C.Y. Cheng, F.C. Chou, R. Das, Modeling complex RNA tertiary folds with Rosetta, Methods Enzym. 553, 35-64 (2015).

[20] F. Ding, S. Sharma, P. Chalasani, V.V. Demidov, N.E. Broude, N.V. Dokholyan, Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms, RNA 14, 1164-73 (2008).

A. Krokhotin, K. Houlihan, N.V. Dokholyan, iFoldRNA v2: folding RNA with constraints, Bioinformatics 31, 2891-3 (2015).

[21] Y. Zhao, Y. Huang, Z. Gong, Y. Wang, J. Man, Y. Xiao, Automated and fast building of three-dimensional RNA structures, Sci Rep. 2,734 (2012).

[22] R. Das, D. Baker, Automated de novo prediction of native-like RNA tertiary structures, Proc. Natl. Acad. Sci. U. S. A. 11, 104 (2007).

[23] C. Flores, R.B. Altman, Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters, RNA 15, 1769-1778 (2010).

[24] E. De Leonardis et al., Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction, Nucl. Ac. Res. 43, 10444-10455.

[25] M. Parisien, F. Major, The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data, Nature 452, 5155 (2008).

[26] T. Cragnolini, P Derreumaux and S. Pasquali, Ab initio RNA folding, J. Phys.: Condens. Matter 27, 233102 (2017).

[27] P.C. Whitford, et al., Nonlocal helix formation is key to understanding S-adenosylmethionine-1 riboswitch function, Biophys. J. 96:2, 2-9 (2009).

[28] B. Lutz, et al., Differences between cotranscriptional and free riboswitch folding, Nucl. Ac. Res. 42, 26872696 (2014).

[29] J. Sponer et al., RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview, Chem. Rev., 118, 4177-4338 (2018).

[30] J.A. Cruz et al., RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction, RNA 18, 610-625 (2012).

[31] Z. Miao et al., RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures, RNA 21, 1066-1084 (2015).

[32] Z. Miao et al.,RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme, RNA. 23, 655-672 (2017).

[33] B.R. Brooks et al., CHARMM: The Biomolecular Simulation Program, J Comput Chem. 30, 1545-1614 (2009).

[34] D.A. Case et al.,AMBER 2018, University of California, San Francisco (2018).

[35] A. Schug, T. Herges, W. Wenzel, Reproducible protein folding with the stochastic tunneling method, Phys. Rev. Lett. 91(15), 158102 (2003).

[36] A. Schug, et al., Comparison of stochastic optimization methods for allatom folding of the trpcage protein, ChemPhysChem 6(12), 2640-2646 (2005).

[37] S.C. Flores, Y. Wan, R. Russel, S.B. Altman, Predicting RNA structure by multiple template homology modeling, Pac. Symp. Biocomput. 216-227 (2010).

[38] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, S.R. Eddy, Rfam: an RNA family database, Nucl. Ac. Res. 31, 439-441 (2003).

[39] P. Di Tommaso et al., T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension, Nucl. Ac. Res. 39, W13-W17 (2011).

[40] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucl. Ac. Res. 32, 1792-1797 (2004).

[41] E. P. Nawrocki and S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches, Bioinformatics

29, 2933-2935 (2013)

[42] D.S. Marks et al., Protein 3D structure computed from evolutionary sequence variation, PLoS ONE 6 (2011).

[43] T.A. Hopf, L.J. Colwell, R. SheridaN, B. Rost, C. Sander, D.S. Marks, Three-dimensional structures of membrane proteins from genomic sequencing Cell. 149, 1607-1621 (2012).

[44] M. Weigt, R.A. White, H. Szurmant, J.A. Hoch, T. Hwa Identification of direct residue contacts in protein-protein interaction by message passing. Proc. Natl. Acad. Sci. USA. 106, 67-72 (2009).

[45] A. Schug, M. Weigt, J.N. Onuchic, T. Hwa, H. Szurmant, High-resolution protein complexes from integrating genomic information with molecular simulation, Proc. Natl. Acad. Sci USA 106(52): 22124-22129 (2009).

[46] A.E. Dago et al., Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis, Proc. Natl. Acad. Sci. USA 109, E1733-E1742 (2012).

[47] D.N. Ivankov, A.V. Finkelstein, F.A. Kondrashov, A structural perspective of compensatory evolution, Curr. Opin. Struct. Biol. 26, 104-112 (2014).

[48] D. de Juan, F. Pazos, A. Valencia, Emerging methods in protein co-evolution, Nat Rev Genet. 14, 249-61 (2013).

[49] C. Weinreb, A. Riesselman, J.B. Ingraham, T. Gross, C. Sander, D.S. Marks, 3D RNA and functional interactions from evolutionary couplings, Cell. 165, 963-975 (2016).

[50] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. USA 108, E1293-E1301 (2011).

[51] S. Ovchinnikov et al., Protein structure determination using metagenome sequence data, Science 355, 294-298 (2017).

[52] J.I. Sulkowska, F. Morcos, M. Weigt, T. Hwa, J.N. Onuchic, Genomics-aided structure prediction Proc. Natl. Acad. Sci USA 109, 10340-10345 (2012).

[53] D.T. Jones, T. Singh, T. Kosciolek, S. Tetchner, MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31, 9991006 (2015).

[54] Zerihun MB, Schug A, Biomolecular coevolution and its applications: Going from structure prediction toward signaling, epistasis, and function, Biochem Soc. Trans. 45, 1253-1261 (2017).

[55] J. Wang, K. Mao, Y. Zhao, C. Zeng, J. Xiang, Y. Zhang, and Y. Xiao, Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotidenucleotide interactions from direct coupling analysis, Nucl. Ac. Res. 45, 6299-6309 (2017).

[56] J.Y. Dutheil , F. Jossinet, E. Westhof, Base pairing constraints drive structural epistasis in ribosomal RNA sequences. Mol. Biol. Evol. 27, 18681876 (2010).

[57] Q. Zhou et al.,Global pairwise RNA interaction landscapes reveal core features of protein recognition, Nat. Commun. 9, 2511 (2018).

[58] D.E. Draper, A guide to ions and RNA structure, RNA 10, 335-43 (2004).

[59] Y. Zhu, Z. He, S.-J. Chen, TBI Server: A Web Server for Predicting Ion Effects in RNA Folding, PLoS ONE 10, e0119705 (2015).

[60] S. Roy, S.P. Hennelly, H. Lammert, J. N. Onuchic, K.Y. Sanbonmatsu, Magnesium controls aptamer-expression platform switching in the SAM-I riboswitch, Nucleic Acids Research, gky1311 (2019).

[61] S. Kirmizialtin, S.P. Hennelly, A. Schug, J.N. Onuchic, K.Y. Sanbonmatsu, Integrating molecular dynamics simulations with chemical probing experiments using SHAPE-FIT, Meth. Enzym. 553: 215-234 (2015).

[62] I. Reinartz et al., Simulation of FRET dyes allows quantitative comparison against experimental data, J. Chem. Phys. 148(12): 123321-6 (2018).

[63] Z. Xia, D.R. Bell, Y. Shi, P. Ren, RNA 3D Structure Prediction by Using a Coarse-Grained Model and Experimental Data, J. Phys. Chem. B 117, 31353144 (2013).

[64] Weiel M, Reinartz M., and Schug, A., Rapid interpretation of small-angle X-ray scattering data, PloS Comp. Biol., 3, e1006900 (2019).

[65] A. Gupta et al., Formation of a Secretion-Competent Protein Complex by a Dynamic Wrap-around Binding Mechanism, J. Mol. Bio., 430:3157-3169 (2018).

[66] A. Ponce-Salvatierra et al., Computational modeling of RNA 3D structure based on experimental data, Biosci. Rep. 39, BSR20180430 (2019).

[67] H. Szurmant and M. Weigt, Inter-residue, inter-protein and inter-family coevolution: bridging the scales, Curr. Op. Struct. Biol. 50, 26-32 (2018).