Published online XXXX *Nucleic Acids Research, XXXX, Vol. XXXX, No. XXXX* **1–5**
doi:10.1093/nar/gkn000

RNA-MoIP: Prediction of RNA secondary structure and local 3D motifs from sequence data

Jason Yao [1], Vladimir Reinharz [2], François Major [3], Jérôme Waldispühl [1,*]

[1] School of Computer Science, McGill University, 3480 University Street, Montreal QC H3A 0E9, Canada; [2] Department of Computer Science, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel; [3] Institute for Research in Immunology and Cancer and Department of Computer Science and Operations Research, Université de Montréal, Montreal, QC H3C 3J7, Canada.

ABSTRACT

RNA structures are hierarchically organized. The secondary structure is articulated around sophisticated local 3D motifs shaping the full 3D architecture of the molecule. Recent contributions have identified and organized recurrent local 3D motifs, but applications of this knowledge for predictive purposes is still in its infancy.

We recently developed a computational framework, named **RNA-MoIP** (RNA Motifs over Integer Programming), to reconcile RNA secondary structure and local 3D motif information available in databases. In this paper, we introduce a web service using our software for predicting RNA hybrid 2D-3D structures from sequence data only. Optionally, it can be used for (i) local 3D motif prediction or (ii) the refinement of user-defined secondary structures. Importantly, our web-server automatically generates a script for the **MC-Sym** software, which can be immediately used to quickly predict all-atom RNA 3D models.

The web-server is available at **http://rnamoip.cs.mcgill.ca**.

INTRODUCTION

RNA folding is hierarchical. The secondary structures (Watson-Crick and Wobble base pairs) form rapidly, acting as a scaffold for the slower formation of 3D structures (1, 2). This observation has already been successfully used by previous software (3, 4, 5, 6, 7), which use secondary structure information to assist in tertiary structure prediction. Alternate approaches fully based on molecular dynamics solutions have also been proposed (8). In general, it is worth noting that the accuracy of all prediction methods significantly decreases as the length of the RNA sequence increases.

Because of their versatility and reliability, fragment assembly methods have become a popular strategy to predict RNA 3D structures (3, 4, 6, 9). These assemble diverse fragments of known 3D structures into complete composite 3D structures. Among them, the MC-Pipeline (4), which is composed of MC-Fold for predicting secondary structure from sequence and MC-Sym for assembling 3D structures from a sequence and structural constraints, has been successfully used in multiple contexts. These include the refinement of diffraction data and model fitting (10), and model screening in pharmaceutical applications (11). The present work aims to boost its performance and ease its utilization.

Our approach builds upon recent advances in RNA 3D structure analysis. By combining the hierarchal folding properties with the large public database of experimentally resolved 3D motifs from RNA3dMotif (12), we have developed a fast and accurate hybrid method, named RNA-MoIP,which uses an integer programming framework to predict RNA 2D-3D structures (i.e. secondary structures in which loop regions are annotated with local 3D structures) from sequence data alone (13). This predicted set of candidate 2D-3D structures can then be used as a template in the MC-Sym software to generate complete 3D structures.

To generate 2D-3D structures, RNA-MoIP uses RNA secondary structure templates (generated by default with RNAsubopt, from the ViennaRNA package (14)) in which it inserts candidate RNA 3D motifs. A distinctive aspect of RNA-MoIP lies in its ability to use a set of approximate candidate secondary structures instead of a single highly reliable template (13). It identifies the most promising secondary structures and eventually improve them by removing incorrectly predicted base pairs. It follows that RNA-MoIP can also be used to refine secondary structure predictions.

Currently, RNA-MoIP exists as a command-line script that accepts as minimum input a primary sequence and a pool of candidate secondary structures generated by RNAsubopt. However, this implementation requires users to have supporting software installed on their machine. RNAsubopt, which generates the candidate secondary structures, and the Gurobi Optimizer, which contains a Python interpreter and integer programming solver for which RNA-MoIP is written, are needed. It also requires the manual transfer of results between RNAsubopt, RNA-MoIP, and MC-Sym. Limitations of the motif database also complicate the process of assembling the motifs suggested by RNA-MoIP into an MC-Sym readable format. Missing structural data

[*] To whom correspondence should be addressed. Tel: +1 514 398 5018; Email: jerome.waldispuhl@mcgill.ca

(attributable to experimental factors such as crystallography resolution) and gaps in motif coverage make the preparation of scripts for MC−Sym a time consuming process. This limited the utility of RNA−MoIP for multi-sequence processes and made the utility inaccessible to many users.

The goal of this server was to develop a computational pipeline to allow users to automatically generate MC−Sym input scripts with the optimizations provided by RNA−MoIP. We developed an end-to-end solution to automate the process such that a user can directly generate a pool of predicted RNA tertiary structures (output by MC−Sym) from any single primary sequence input alone. Additionally, we implemented tools to visualize the motif insertions and ease interpretation and selection from the candidate solutions. Their locations in the secondary structure are directly shown in an annotated schema and an additional interface allows users to visualize the 3D atomic model of each instance. In the following sections, the RNA−MoIP web-server and its methods are described.

WEB SERVER

The RNA−MoIP web-server is available at http://rnamoip.cs.mcgill.ca/ and runs on an Ubuntu Server, on a Dell PE T610 2x Intel Quad core X5570 Xeon Processor, 2.93 GHz 8M Cache, 64 GB Memory (8 x 8 GB), 1333 MHz Dual Ranked RDIMMs for 15 Processors.

The back-end is written in Python 2.7.3 and bash. For each task one processor is allocated and the Gurobi optimizer v.6.5.0 (7) API for Python is used to solve the integer programming equations. Each job is assigned a unique directory on the server and all data is preserved for up to two weeks.

The front-end is designed using the Bootstrap 3.3.7 css framework. Webpages are generated using Python 2.7.3 standard library cgi module and utilize Javascript.

Input

The minimal input to perform a query is an RNA sequence. Up to 5 sequences in the FASTA format can be provided simultaneously and all results will be presented in the same directory.

An ensemble of additional constraints can be provided by the users. Those are available under the collapsible *Advanced Options* menu. The options are (i) the maximal fraction of base pairs in the secondary structure that can be removed to accommodate the insertion of motifs (default 30%), (ii) the largest number of components—or strands—in a motif that can be inserted (default 3, or three way junctions, only up to four way junctions is available now), and (iii) the maximum number of solutions to output (default 1). This is the number of optimal solutions per secondary structure that are returned. If different secondary structures achieve solutions with the same score all of them will be displayed. Additionally, by default, candidate secondary structures are computed with RNAsubopt, sampling all structures within 3 KCal/mol from the minimal free energy structure. A stochastic sampling can also be selected, or a set of secondary structures in the dot bracket notation can be provided by the user.

A name can be assigned to a new job, or a random ID will be provided, that can be used to jump directly to the working directory. In the top navigation bar, there is a search menu that users can input either their job name, or the job ID that is randomly assigned and associated with the created directory. This information is provided to the user immediately after the job has been submitted. In addition, a general help page is available with explanations for each of the options, as well as an about page with references and author contact information.

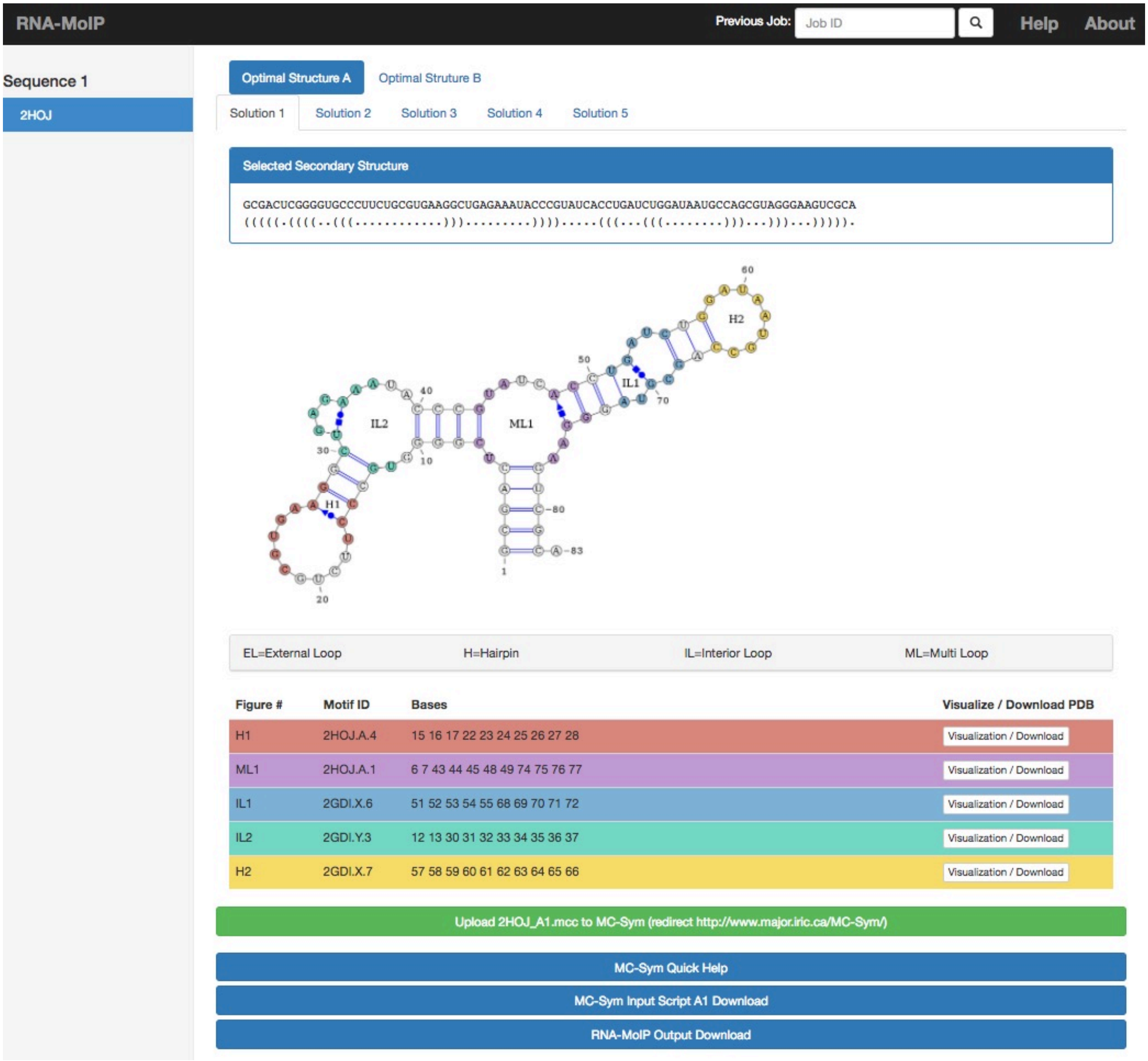
Output

The main window of the output is shown in Fig 1. Upon job submission, a working directory is generated that can be accessed by Job ID or job name (if provided). Each working directory is maintained on the server for a minimum of 2 weeks. The directory consists of four main elements: tabs to view the different optimal structures and their solutions, links to the assembled MC−Sym script(s) with an option to automatically submit, a diagram of the secondary structure selected by RNA−MoIP indicating the position of the inserted motifs, and a table with links to view interactive 3D models of those inserted motifs. It is important to note that many competing structures can have the same optimal value in the IP framework (13), in which case each of them will be displayed.

The working directory is composed of a tabbed main navigation element and a sidebar linking to other directories for batch (.fasta) uploaded jobs. In the main navigation window, a link allows the script to be automatically submitted to the MC−Sym server, where the user is then directed to the MC−Sym control panel for the submitted job. Additionally, users can view details about the selected secondary structure and motifs inserted. If a maximum of more than one solution was requested, the solutions can be compared using the tabbed navigation. For each solution, a secondary structure diagram is created dynamically using VARNA, with non-canonical interactions annotated in the inserted motifs (15). This allows for an easy graphical comparison between candidates in the set of solutions. Motifs are inserted with sequence based constraints; thus a set of valid 3D structures can be assigned to to the same location. Below, users can view details of the motifs selected by RNA−MoIP, including the bases it spans, and a link to visualize or download the PDB file of each valid instance. For each inserted motif, a specific candidate atomic structure can be selected from a list and visualized in the browser using JSmol. A direct link to download the PDB file of each instance of the inserted motif is also available.

MATERIALS AND METHODS

The RNA−MoIP web-server consists mainly of the RNA−MoIP program with a set of new tools easing its use, analysis of the output, and interface with MC−Sym for an all-atoms prediction. Here we describe its main components: the secondary structure generation, RNA−MoIP, the automatic generation of MC−Sym scripts, and the visualization system. The automatic script generation is a novel approach to systematize the strategy in (16) which can be quite involved even for small RNAs with few motifs. This new procedure allows a seamless transition from the RNA−MoIP2D-3D structure to to the full-atom predictions of MC−Sym.



Sequence 1

2HOJ

Optimal Structure A | Optimal Struture B

Solution 1 | Solution 2 | Solution 3 | Solution 4 | Solution 5

Selected Secondary Structure

```
GCGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCUGAUCUGGAUAAUGCCAGCGUAGGGAAGUCGCA
(((((.((((..(((............))).........)))).....(((...(((........)))...)))...))))).
```

EL=External Loop   H=Hairpin   IL=Interior Loop   ML=Multi Loop

Figure #	Motif ID	Bases	Visualize / Download PDB
H1	2HOJ.A.4	15 16 17 22 23 24 25 26 27 28	Visualization / Download
ML1	2HOJ.A.1	6 7 43 44 45 48 49 74 75 76 77	Visualization / Download
IL1	2GDI.X.6	51 52 53 54 55 68 69 70 71 72	Visualization / Download
IL2	2GDI.Y.3	12 13 30 31 32 33 34 35 36 37	Visualization / Download
H2	2GDI.X.7	57 58 59 60 61 62 63 64 65 66	Visualization / Download

Upload 2HOJ_A1.mcc to MC-Sym (redirect http://www.major.iric.ca/MC-Sym/)

MC-Sym Quick Help

MC-Sym Input Script A1 Download

RNA-MoIP Output Download

Figure 1. The results screen of the RNA−MoIP web-server for a fasta input. Shown is the output for the *E. coli thi-box riboswitch* 2HOJ. In the main dashboard, the corrected secondary structure is shown with a 2D secondary structure visualization generated from VARNA. Multiple solutions for each sequence can be toggled using the tabbed navigation. In the side navigation bar users can access other sequences if submitted in the same *.fasta* file.

Secondary Structures Generation

For each sequence, if no secondary structure is provided, the web-server provides two ways to generate a pool of secondary structures. In both cases RNAsubopt is used. RNAsubopt can either list all structures at an energy range from the minimal free energy structure, or use a stochastic sampling approach. By default, structures are sampled using the energy range model at 3 KCal/mol. Under this setting, at least 10 or more RNAs are typically sampled, although the sample size may drastically increase for longer sequences. The range can be modified by the user. In the stochastic model, secondary structures are sampled with probability equal to their Boltzmann weight in the ensemble. The number of sampled structures is 25 by default and can be modified by the user.

The size and makeup of the pool of candidates can significantly impact predictions made by the server. We have previously found that low ranking structures can result in models with the highest accuracy, measured by RMSD, after refinements by RNA−MoIP (13). This stresses the importance of optimizing efficiency versus accuracy of the model at the secondary structure level. A large sample is necessary to provide RNA−MoIP greater flexibility in model selection.

RNA-MoIP

RNA-MoIP is an integer programming framework that refines secondary structure predictions to accommodate the insertion of RNA 3D motifs by removing base pairs. Motifs are structural units shared across many different RNA molecules including hairpins, internal loops, and k-way junctions. Between different RNAs, they are expected to share a similar local 3D structure based on folding interactions at the primary and secondary structure level. In the RNA-MoIP framework, motifs are inserted based only on sequence compatibility. The motifs database is populated by RNA3dMotif (12), which extracts them from the structures in the Protein Data Bank. Currently, 4695 motifs are indexed as part of the RNA-MoIP database. We have previously shown that by incorporating publicly accessible 3D motif data into structural predictions, we can improve the accuracy and running time of our models compared to MC-Sym alone.

Given a sequence and secondary structure, RNA-MoIP performs two functions in parallel: i) refinement of secondary structures by removing base pairs to accommodate the insertion of database motifs and ii) insertion of motifs from the database.

A fraction of base pairs from the given secondary structure can be removed to accommodate the insertion of motifs. This has a two-fold advantage of enhancing the flexibility of the secondary structure to accept motifs, and improving the accuracy of secondary structure by removing incorrectly predicted base pairs. The IP framework aims to balance the cost of removing predicted canonical-base pairs with the insertions of motifs. It additionally prioritizes the insertion of large motifs instead multiple small ones, aiming to maximize coverage while maintaining the general structure of the original secondary structure prediction. The IP framework assigns linear penalties for base-pair deletions and scores motifs based on the square of the component length.

Formally, RNA-MoIP minimizes the following function, given an ensemble of base pairs D, an ensemble of motifs Mot, the total length of a motif M^x, and all its occurrences C. A score is given as follows:

$$10 * \sum_{(u,v)\in B} D_{u,v} - \sum_{x\in \text{Mot}^j} \left((|M^x|)^2 \cdot \sum_{(x,k,l)\in \text{Seq}_1^j} C_{k,l}^{x,1} \right) \quad \textbf{(1)}$$

Each pair of sequence structures gets a score. RNA-MoIP selects the structures with the minimal score as the most probable solution. The output provides the secondary structure(s) it has selected from the given input pool, a list of motifs it has inserted, and a list of bases that have been removed.

MC-Sym and Input Script Assembly

To generate full-atom predictions, we offer a direct submission to the MC-Sym web-server. MC-Sym requires the instructions to perform the prediction in a specific format, *.mcc*, which is automatically generated and provided to the user. We describe here how the script for MC-Sym is produced.

The script consists of the following main sections:
Sequence: the RNA sequence
Library: list of fragments used to assembly the RNA secondary structure. Includes motifs, stems and links
Backtrack: the order in which fragments are joined. Fragments must be merged contiguously, overlapping with a previously placed fragment that serves as an “anchor”
Relation: defines dangling at the 5’ and 3’ ends separately from the rest of the model

MC-Sym assembles RNA fragments. Each must be individually defined in the library section, and are assembled in an overlapping fashion. In other words newly placed fragments must overlap with at least one residue from a previously placed fragment. The only exceptions are the dangling ends which are defined in the *Relation* section and are treated independently by MC-Sym. The quality and feasibility of the full-atom prediction is greatly influenced by the order in which the fragments are merged.

There are three main type of fragments. First *stacked canonical base pairs* are used to build the stems. Second *links*, which indicate that a certain nucleotide follows another, are used to define nucleotides with less constraints. Finally *motifs*, which were previously inserted by RNA-MoIP.

To limit the size of the conformational space early on and to improve the accuracy of the model, the largest fragment is positioned first. This has the two-fold advantage of significantly constraining the search space of most nucleotides from the beginning, and building around a central portion of the molecule containing a three-way or four-way junction. This is beneficial as these have greater stability than external regions due to their highly constrained structures. The rest of the structure is assembled around it using a depth-first prioritization to constrain free-ends and limit long-range stem interactions.

Contiguous assembly is a challenge due to the fact that some motifs may have bases that are undefined in their 3D structures, as explained in (16). Therefore undefined nucleotides are subsequently merged using links that overlap the already placed motifs and stems. Because links are unconstrained, they have a very large conformational search space. The server therefore attempts to limit this type of fragment by placing them only after all adjacent stems and motifs have been placed. The stems connecting the motifs are built from chains of stacked canonical base pairs. The assembly process proceeds until all residues in the sequence have been defined by the script.

MC-Sym jobs are run on the IRIC web-server. MC-Sym offers tools to optimize, analyze and retrieve results. Energy minimization is recommended prior to analyzing RNA-MoIP output. Clustering using the k-means algorithm can also identify significant structures out of the complete pool of results.

Motif Insertion Visualization

The RNA-MoIP web-server provides tools for the visualization of motif insertion locations and 3D structures of inserted motifs. An annotated secondary structure diagram is dynamically generated by the VARNA visualization applet (15). Using the motif-to-sequence base mapping defined during processing, the secondary structure is annotated and coloured based on insertion location. Additionally, non-canonical binding interactions are annotated using the Leontis-Westhof symbolic notation. Motifs can contain sophisticated non-canonical interactions that help form the backbone of the tertiary structure. Therefore, representing these interactions can allow us to better understand the folding configuration of the models, and make comparisons amongst `RNA-MoIP` candidate outputs.

The web-server also provides a tool for users to view 3D models of the RNA motifs. Visualization is performed using JSmol, a JavaScript framework that allows users to interactively view 3D molecular structures. `MC-Sym` parses a number of candidate structures for each motif selected. Using `RNA-MoIP`, users can visualize and download the PDB files for each of these candidates directly from their browsers. This allows for an easy way to visualize and analyze the structural components used to build the model.

CONCLUSIONS

There have been many attempts to computationally solve the RNA tertiary structure prediction problem. Previous contributions have used molecular dynamics and short fragment (1-3 nucleotide) assemblies to predict RNA tertiary structures (8, 9). The `RNA-MoIP` pipeline offers a fast and accurate approach to solving the prediction problem using IP. By combining the available database with the flexibility of the IP model, our solution improves accuracy of alternative prediction software for RNA sequences of all lengths (13). Compared to `MC-Sym` alone, applying `RNA-MoIP` constraints enables us to quickly produce secondary structures that would otherwise run for an unspecified period of time.

Improvements planned for the `RNA-MoIP` server include upgrades to the motif database. New RNA models are constantly being added to the PDB database from which our motifs are extracted. We hope to build in a way of automatically updating our motif database as new listings are added, to provide constantly up-to-date results. We later also hope to allow users to filter the motifs used in their search, and provide them with the option of adding their own custom motif database to search on the server. For the `RNA-MoIP` framework, an open problem remains optimizing the size of the pool of suboptimal structures used as input. As mentioned, large pools significantly increase the runtime of `RNA-MoIP`, though prior results have shown that low ranking candidates can offer significant improvements to base-pair accuracy (for molecule 2DU3, `RNA-MoIP` selects the 163rd candidate with a base-pair accuracy of 91% vs 43% on average). We hope to solve this by clustering similar `RNAsubopt` structures to reduce the size of the search pool and look through only meaningful structures.

The `RNA-MoIP` server was designed with the goal of making the unique IP based approach to tertiary structure prediction available to all users. The server-side processing and optimizations streamline the assembly pipeline to seamlessly go from any primary sequence directly to the 3D structure predictions of `MC-Sym`, and allow for the exploration of the intermediate solutions. The web-server was designed with ease-of-use in mind: its minimal design and visualization tools make the server accessible for users of any background. We believe that it will serve as a valuable tool for applications requiring both fast and accurate 3D structure predictions, such as RNA structure, function, and interaction prediction.

ACKNOWLEDGEMENTS

Conflict of interest statement. None declared.

REFERENCES

1. Ignacio Tinoco Jr and Carlos Bustamante. How RNA folds. *Journal of Molecular Biology*, 293(2):271 – 281, 1999. [doi:10.1006/jmbi.1999.3001].
2. Philippe Brion and Eric Westhof. Hierarchy and dynamics of RNA folding. *Annual review of biophysics and biomolecular structure*, 26(1):113–137, 1997. [doi:10.1146/annurev.biophys.26.1.113].
3. Zhiyong Wang and Jinbo Xu. A conditional random fields method for RNA sequence-structure relationship modeling and conformation sampling. *Bioinformatics*, 27(13):i102–10, Jul 2011. [doi:10.1093/bioinformatics/btr232].
4. Marc Parisien and François Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183):51–5, Mar 2008. [doi:10.1038/nature06684].
5. Matthew G Seetin and David H Mathews. Automated RNA tertiary structure prediction from secondary structure and low-resolution restraints. *J Comput Chem*, 32(10):2232–44, Jul 2011. [doi:http://dx.doi.org/10.1002/jcc.21806].
6. Mariusz Popenda, Marta Szachniuk, Maciej Antczak, Katarzyna J Purzycka, Piotr Lukasiak, Natalia Bartol, Jacek Blazewicz, and Ryszard W Adamiak. Automated 3D structure composition for large RNAs. *Nucleic Acids Res*, 40(14):e112, Aug 2012. [doi:http://dx.doi.org/10.1093/nar/gks339].
7. Michal J Boniecki, Grzegorz Lach, Wayne K Dawson, Konrad Tomala, Pawel Lukasz, Tomasz Soltysinski, Kristian M Rother, and Janusz M Bujnicki. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res*, 44(7):e63, Apr 2016. [doi:http://dx.doi.org/10.1093/nar/gkv1479].
8. Shantanu Sharma, Feng Ding, and Nikolay V Dokholyan. iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, 24(17):1951–2, Sep 2008. [doi:10.1093/bioinformatics/btn328].
9. Rhiju Das and David Baker. Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci U S A*, 104(37):14664–9, Sep 2007. [doi:http://dx.doi.org/10.1073/pnas.0703836104].
10. Jeremiah J. Trausch, Joan G. Marcano-Velázquez, Michal M. Matyjasik, and Robert T. Batey. Metal ion-mediated nucleobase recognition by the ZTP riboswitch. *Chemistry & Biology*, 22(7):829 – 837, 2015. [doi:10.1016/j.chembiol.2015.06.007].
11. James Palacino, Susanne E Swalley, Cheng Song, Atwood K Cheung, Lei Shu, Xiaolu Zhang, Mailin Van Hoosear, Youngah Shin, Donovan N Chin, Caroline Gubser Keller, et al. SMN2 splice modulators enhance U1-pre-mRNA association and rescue SMA mice. *Nature chemical biology*, 11(7):511–517, 2015. [doi:10.1016/j.chembiol.2015.06.007].
12. Mahassine Djelloul and Alain Denise. Automated motif extraction and classification in RNA tertiary structures. *RNA*, 14(12):2489–97, Dec 2008. [doi:10.1261/rna.1061108].
13. Vladimir Reinharz, François Major, and Jérôme Waldispühl. Towards 3D structure prediction of large RNA molecules: an integer programming framework to insert local 3D motifs in RNA secondary structure. *Bioinformatics*, 28(12):i207, 2012. [doi:10.1093/bioinformatics/bts226].
14. Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011. [doi:10.1186/1748-7188-6-26].
15. Kévin Darty, Alain Denise, and Yann Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974–5, August 2009. [doi:10.1093/bioinformatics/btp250].
16. Jérôme Waldispühl and Vladimir Reinharz. Modeling and predicting RNA three-dimensional structures. *RNA Bioinformatics*, pages 101–121, 2015. [doi:10.1007/978-1-4939-2291-8_6].