RNA structure prediction using positive and negative evolutionary information

Elena Rivas
Department of Molecular and Cellular Biology,
Harvard University, Cambridge, Massachusetts 02138, USA

Keywords: structural RNAs, evolutionary covariation, variation, pseudoknots, triplets, alternative structures

rivaslab.org/R-scape

Abstract

Predicting a conserved RNA structure remains unreliable, even when using a combination of thermodynamic stability and evolutionary covariation information. Here we present a method to predict a conserved RNA structure that combines the following features: it uses significant covariation due to RNA structure and removes spurious covariation due to phylogeny. In addition to positive covariation information, it also uses negative information: basepairs that have variation but no significant covariation are prevented from occurring. Lastly, it uses a battery of probabilistic folding algorithms that incorporate all positive covariation into one structure. The method, named CaCoFold (Cascade variation/covariation Constrained Folding algorithm), predicts a nested structure guided by a maximal subset of positive basepairs, and recursively incorporates all remaining positive basepairs into alternative helices. The alternative helices can be compatible with the nested structure such as pseudoknots, or incompatible such as alternative structures, triplets, or other 3D non-antiparallel interactions. A comparison to curated databases of structural RNAs highlights multiple RNAs for which CaCoFold uncovers key structural elements such as coaxial-helix stacking, multi-way junctions, pseudoknots, and non Watson-Crick RNA motifs.

Introduction

The importance of comparative information to improve the prediction of a conserved RNA structure has been long recognized and applied to the determination of RNA structures[1–6]. Computational methods that exploit comparative information in the form of RNA compensatory mutations from multiple sequence alignments have been shown to increase the accuracy of RNA consensus structure prediction[7–12].

Still the determination of a conserved RNA structure using comparative analysis has challenges. There is ample evidence that pseudoknots covary at similar levels than other basepairs, however most comparative methods for RNA structure prediction can only deal with nested structures. Identifying pseudoknotted and other non-nested pairs that covary requires having a way of measuring significant covariation due to a conserved RNA structure. After all, all covariation scores are positive. In addition to using positive information in the form of basepairs observed to significantly covary, it would also be advantageous to use negative information in the form of basepairs that should be prevented from occurring because they show variation but not significant covariation.

To approach these challenges, we have introduced a method called R-scape (RNA Structural Covariation Above Phylogenetic Expectation)[13] that reports basepairs that significantly covary using a tree-based null model to estimate phylogenetic covariations from simulated alignments with similar base composition and number of mutations to the given one but where the structural signal has been perturbed. Significantly covarying pairs are reported with an associated E-value describing the expected number of non-structural pairs that could have a covariation score of that magnitude or larger in a null alignment of similar size and similarity. We call these significantly covarying basepairs for a given E-value cutoff (typically ≤ 0.05) the positive basepairs.

In addition to reporting positive basepairs, R-scape has recently introduced another method to estimate the covariation power of a pair based on the mutations observed in the corresponding aligned positions[14]. A pair without significant covariation could still be a basepair in which the two positions are very conserved. A pair without significant covariation but with covariation power should be rejected as a basepair. We call these the negative basepairs.

Here we combine these two sources of information (positive in the form of significantly covarying basepair, and negative in the form of pairs of positions unlikely to form a basepair) into a new RNA folding algorithm. The algorithm also introduces a new iterative procedure that systematically

incorporates all positive basepairs into the structure while remaining computationally efficient. The recursive algorithm is able to finds pseudoknots, other non-nested interactions, alternative structures and triplet interactions provided that they are supported by covariation.

Our method incorporates covariation-supported pairs into a set of helices, and may also predict additional helices that are consistent with that structure. The helices with covariation-supported pairs tend to be reliable, although their length may remain unclear. The additional helices lacking covariation support are less reliable and need to be taken as speculative. We compare the new structures informed by positive and negative evolutionary information to those proposed in the databases of structural RNAs Rfam[15] and the Zasha Weinberg Database (ZWD)[16], where our method highlights structural differences supported by positive basepairs that often uncover key structural RNA elements.

Results

The CaCoFold algorithm

The new RNA structure prediction algorithm presents three main innovations: the proposed structure is constrained both by sequence variation as well as covariation (the negative and positive basepairs respectively); the structure can present any knotted topology and include residues pairing to more than one residue; all positive basepairs are incorporated into a final RNA structure. Pseudoknots and other non-nested pairwise interactions, as well as alternative structures and tertiary interactions are all possible provided that they have covariation support.

The method is named Cascade covariation and variation Constrained Folding algorithm (CaCoFold). Despite exploring a 3D RNA structure beyond a set of nested Watson-Crick basepairs, the algorithm remains computationally tractable because it performs a cascade of probabilistic nested folding algorithms constrained such that at a given iteration, a maximal number of positive basepairs are forced into the fold, excluding all other positive basepairs as well as all negative basepairs. Each iteration of the algorithm is called a layer. The first layer, calculates a nested structure that includes a maximal subset of positive basepair. Subsequent layers of the algorithm incorporate the remaining positive basepairs arranged into alternative helices.

From an input alignment, the positive basepairs are calculated using the G-test covariation

measure with APC correction after removing covariation signal resulting from phylogeny, as implemented in the software R-scape[13]. The set of all significantly covarying basepairs is called the positive set. We also calculate the covariation power for all possible pair[14]. The set of all pairs that have variation but not covariation is called the negative set. All positive basepairs are included in the final structure, and all negative basepairs are forbidden to appear.

Figure 1 illustrates the CaCoFold algorithm using a toy alignment (Figure 1a) derived from the manA RNA[17]. After R-scape with default parameters identifies five positive basepairs (Figure 1b), the CaCoFold algorithm calculates in four steps a structure including all five positive basepairs as follows.

(1) The cascade maxCov algorithm. The cascade maxCov algorithm groups all positive basepairs in nested subsets (Figure 1c). At each layer, it uses the Nussinov algorithm which is one of the simplest RNA models[18]. Here we use the Nussinov algorithm not to produce an RNA structure, but to group together a maximal subset of positive basepairs that are nested relative to each other. Each subset of nested positive basepairs will be later provided to a folding dynamic programming algorithm as constraints. Supplemental Figure S1 includes a detailed description of the Nussinov algorithm.

The first layer (C0) finds a maximal subset of compatible nested positive basepairs with the smallest cumulative E-value. After the first layer, if there are still positive basepairs that have not been explained because they did not fit into one nested set, a second layer (C1) of the maxCov algorithm is performed where only the still unexplained positive basepairs are considered. The cascade continues until all positive basepairs have been grouped into nested subsets.

The cascade maxCov algorithm determines the number of layers in the algorithm. For each layer, it identifies a maximal subset of positive basepairs forced to form, as well as a set of basepairs not allowed to form. The set of forbidden basepairs in a given layer is composed of all negative pairs plus those positive pairs not in the current layer.

The cascade maxCov algorithm provides the scaffold for the full structure, which is also obtained in a cascade fashion.

(2) The cascade folding algorithm. For each layer in the cascade with a set of nested positive basepairs, and another set of forbidden pairs, the CaCoFold algorithm proceeds to calculate the most probable constrained nested structure (Figure 1d).

Different layers use different folding algorithms. The first layer is meant to capture the main nested structure (S0) and uses the probabilistic RNA Basic Grammar (RBG)[19]. The RBG model features the same elements as the nearest-neighbor thermodynamic model of RNA stability[20;21] such as basepair stacking, the length of the different loops, the length of the helices, the occurrence of multiloops, and others. RBG is a simplification of the models used in the standard packages, such as ViennaRNA[21], Mfold[22], or RNAStructure[23], but it has comparable performance regarding folding accuracy[19]. Supplemental Figure S1 includes a description of the RBG algorithm.

The structures at the subsequent layers (S+ = {S1, S2,...}) are meant to capture any additional helices with covariation support that does not fit into the main secondary structure S0. We expect that the covariations in the subsequent layers will correspond to pseudoknots, and also to non Watson-Crick interactions, or even triplets. The S+ layers use the simpler G6 RNA model[24;25] which mainly models the formation of helices of contiguous basepairs. Here we extend the G6 grammar to allow positive pairs that can be next to each other in the RNA backbone, interactions that are not uncommon in RNA motifs. We name the modified grammar G6X (see Supplemental Figure S1 for a description).

The RBG and G6X models are trained on a large and diverse set of known RNA structures and sequences as described in the method TORNADO[19]. At each layer, the corresponding probabilistic folding algorithm reports the structure with the highest probability using a CYK dynamic programming algorithm on a consensus sequence calculated from the columns in the alignment.

Because the positive residues that are forced to pair at a given cascade layer could pair (but to different residues) at subsequent layers, the CaCoFold algorithm can also identify triplets or higher order interactions (a residue that pairs to more than one other residue) as well as alternative helices that may be incompatible and overlap with other helices.

(3) Filtering of alternative helices. In order to combine the structures found in each layer into a complete RNA structure, the S+ structural motifs are filtered to remove redundancies and elements without covariation support.

We first break the S+ structures into individual alternative helices. A helix is arbitrarily defined as a set of contiguous basepairs with at most two residues are not paired (forming a one or two residue bulge or a 1x1 internal loop). A helix is arbitrarily called positive if it includes at least one positive basepair. All positive alternative helices are reported. Alternative helices without any

covariation are reported only if they include at least 15 basepairs, and if they overlap in no more than 50% of the the bases with another helix already selected from previous layers. In our simple toy example, there is just one alternative helix. The alternative helix is positive, and it is added to the final structure. No helices are filtered out in this example (Figure 1d).

(4) Automatic display of the complete structure. The filtered alternative helices are reported together with the main nested structure as the final RNA structure. We use the program R2R to visualize the CaCoFold structure with all covarying basepairs annotated in a green hue. We adapted the R2R software to depict all non-nested pairs automatically (Figure 1f).

If R-scape does not identify any positive basepair, one single layer is defined without any pair constraints, and one nested structure is calculated. Lack of positive basepairs indicates lack of confidence that the conserved RNA is structural, and the proposed structure has no evolutionary support.

For the toy example in Figure 1, R-scape with default parameters identifies five positive basepairs. The CaCoFold algorithm requires two layers to complete. The first layer incorporates three nested positive basepairs. The second layer introduces the remaining two positive basepairs. The RBG fold with three constrained positive basepairs produces three helices. The G6X fold with two positive and three forbidden basepairs results in one alternative helix between the two hairpin loops of the main nested structure. In this small alignment there are no negative basepairs, and no alternative helices without covariation support have to be filtered out. The final structure is the joint set of the four helices, and includes one pseudoknot.

CaCoFold increases the number of positive basepairs and helices

One key feature of a CaCoFold structure is that it incorporates all comparative evidence conveyed by the positive basepairs into one structure. Another key feature is the set of negative pairs not allowed to form because they would show variation but not covariation. In addition, CaCoFold provides a set of compatible basepairs obtained by constrained probabilistic folding. The set of compatible pairs is only indicative of a possible completion of the structure. They do not provide any additional evidence about the presence of a conserved structure, and some of them could be erroneous as it is easy to predict consistent RNA basepairs even from random sequences.

We compared the RNA structures presented in the databases Rfam and ZWD to those pro-

duced by CaCoFold on the same alignments. For the Rfam alignments, the CaCoFold structures incorporate 781 additional positive basepairs relative to those found in the given Rfam structures. For the ZWD database, the CaCoFold structures have 364 additional positive pairs. A total of 672 negative basepairs are removed from the Rfam CaCoFold structures, and 229 from the ZWD structures (Table 1).

The CaCoFold structure allows us to asses the reliability of the predicted helices. We define positive helices arbitrarily as those with at least one positive basepair. We separate positive helices found in the main nested structure from those found in the additional layers. Regarding positive nested helices, we identify 216 more positive nested helices for Rfam, and 76 more for the ZWD database. Regarding alternative helices is where we find the largest change as pseudoknots, triplets and incompatible helices tend to be less accurately annotated. By construction, the majority of alternative helices incorporated by CaCoFold are positive. We identify 374 more positive alternative helices for Rfam, and 215 for ZWD (Table 1).

In the next section, we classify all RNA families for which the CaCoFold structure include more positive basepairs and helices into different types depending on the kind of structural modification they introduce.

RNA structures improved by positive and negative signals

By comparing whole structures, we identify 276 Rfam and 105 ZWD RNA families for which the CaCoFold structure includes positive basepairs not present in the given structures. Because there is overlap between the two databases, in combination there is a total of 313 structural RNAs for which either the Rfam structure or the ZWD structure can be improved by including more significant covarying pairs. Of the 313 RNAs, there are five for which the Rfam and ZWD alignments and structures are quite different from each other (PhotoRC-II/RF01717, manA/RF01745, radC/RF01754, pemK/RF02913, Mu-gpT-DE/RF03012) so we include both versions in our analysis. In the end, we identify a total of 318 structural alignments for which the structure presented in the databases is missing positive basepairs, and CaCoFold proposes a modified structure.

In many cases, the change introduced by the CaCoFold structure is to include one additional positive basepair. But often, that extra positive basepair appear at critical positions indicating a key structural feature, such as an additional helix or pseudoknot, a new three-way junction, or a

	Rfam 1,242 RNA families			**ZWD** 377 RNA families		
	CaCoFold structure		Given structure	CaCoFold structure		Given structure
Basepairs						
Positive	9,349	+781	8,568	5,481	+364	5,117
Negative	672		n.a.	229		n.a.
Compatible	51,075		46,071	7,370		6,322
Positive helices (% of total helices)						
Nested	2,834 (40%)	+216	2,618 (38%)	1,352 (70%)	+76	1,276 (76%)
Alternative	431 (99%)	+347	84 (68%)	319 (99%)	+215	104 (83%)

Table 1: **Positive and negative signals found in CaCoFold structures.** Positive basepairs are those that significantly covary. Negative basepairs are pairs prevented to form because the two positions show variation but not covariation. Compatible pairs are all pairs in the proposed structure that are not positive. Positive helices have at least one positive basepair. In blue, the signed difference between the number of positives (basepairs or helices) in the CaCoFold versus given structures. Nested helices are those from the CaCoFold first layer nested structures, or annotated as nested in the consensus given structures. Alternative helices come from all other CaCoFold layers, or annotated as pseudoknots in the consensus given structures. For each database, we compile results for all families with a consensus given structure that has at least 10 basepairs and 10% power (that is, at least one expected covarying basepair). There is one alignment per family. Rfam uses seed alignments.

Modifications introduced by the extra covariations in the CaCoFold structure		Number of alignments
Type 1	Helix expanded by additional covariations	23
Type 2	One new helix with covariation support	11
Type 3	One helix completely modified	7
Type 4	New pseudoknot with covariation support	16
Type 5	New coaxial stacking	17
Type 6	Internal loop or multiloop reshaped by coaxial stacking	12
Type 7	Hairpin or internal loop covariations (often non Watson-Crick)	19
Type 8	Non Watson-Crick (not within a loop) covariations	24
Type 9	Triplets	28
Type 10	Cross-covariations (see text)	30
Type 11	Possibly alternative structures	6
Type 12	Additional covariations in SSU and LSU rRNA	6
Type 13	Multiple covariations not compatible with anything	3
Type 14	Misalignment introducing spurious covariations	2
Type 15	Low power; inconclusive	114
	Total number of structural RNA alignments	318

Table 2: **318 RNA structural alignments (from the Rfam and ZWD databases combined) for which the CaCoFold structure includes additional positive basepairs.** The RNA structural alignments are manually classified into 15 categories. Examples of types 1-11 are presented in Figure 2.

better definition of how different helices stack coaxially. We manually classified these 318 modified structures into 15 categories (Table 2). In Supplemental Table S1, we report a full list of the RNA families and alignments with better CaCoFold structure, classified according to Table 2.

In Figure 2, we show representative examples of the most important types of structural changes (Types 1-12). In **Type 1**, the extra positive basepairs incorporated by CaCoFold extend the length of an already annotated helix, as in the RF0169 bacterial small SRP example. **Type 2** represented by the RF01794 sok RNA includes cases in which several positive basepairs identify a new helix in the structure. **Type 3** includes seven cases in which a helix without positive basepairs in the given structure gets refolded by CaCoFold into a different helix that includes several positive basepairs. For the RF03068 RT-3 RNA example, the refolded helix has 8 positive basepairs. **Type 4** describes cases in which positive basepairs reveal a new helix forming a pseudoknot. There are 16 of these cases, of which chrB RNA is an example. **Type 5** and **Type 6** are cases in which the additional positive basepairs refine the secondary structure, either by defining new (three-way or higher) junctions or internal loops, (**Type 5**) or by adding positive basepairs at critical positions at the end of helices that help identify coaxial stacking (**Type 6**). **Type 7** describes cases in which the extra positive basepairs are in loops (hairpin or internal). Types 5, 6, and 7 often identify recurrent RNA motifs[26], as in the case shown in Figure 2, where and additional positive basepair identifies a tandem GA motif in the RtT RNA, and another positive basepair in a U4 RNA loop highlights a kink turn motif. Other more general non-Watson-Crick interactions are collected in **Type 8**, of which tRNA is a exceptional example in which almost all positions are involved in some covarying interaction. **Type 9** include cases of positions involved in more than one positive basepair. In general, one of the positive basepairs is part of a extended helix, but the other is in general not nested and involves only one or two contiguous pairs. **Type 10** includes a particular type of triplets that we name cross-covariations. A cross-covariation is defined as a positive basepair such that each position forms another positive Watson-Crick basepair with a different position, such that, the two other positive Watson-Crick basepairs are part of the same helix. In Figure 2, we show an example of a helix with four cross-covariations. As an extreme example, the bacterial LOOT RNA with approximately 43 basepairs in six helices includes 28 cases of cross-covariations. **Type 11** includes a few cases in which an alternative positive helix is incompatible with another positive helix. These cases are candidates for possible mutually exclusive

structures. The SSU and LSU ribosomal RNA alignments are collected in **Type 12**. These are large structures with deep alignments in which about one third of the basepairs are positive. For the LSU rRNA, CaCoFold finds between 8 (Eukarya) to 22 (bacteria) additional positive basepairs. **Type 13** include just three cases for which the positive basepairs are scattered throughout, and do not confirm an structure. **Type 14** identifies two cases in which the Rfam and ZWD alignments report different sets of positive basepairs. These suggest the possibility of a misalignment resulting in spurious covariations. Finally, **Type 15** collects about a third (114/318) of the alignments for which CaCoFold identifies only one or two positive basepairs while the original structure has none. None of these alignments has enough covariation to support any particular structure. These alignments also have low power of covariation to decide whether there is a conserved RNA structure in the first place.

In the supplemental materials, we provide the original Rfam and ZWD alignments annotated with the CaCoFold structure, as well as R2R depictions of the original and CaCoFold structures as in Figure 2.

The R-scape covariation analysis and CaCoFold structure prediction including pseudoknots for all 3,016 seed alignments in Rfam 14.1 (which includes four SSU and three LSU rRNA alignments; ranging in size from SSU rRNA Archaea with 1,958 positions to LSU rRNA Eukarya with 8,395 positions) takes a total of 724 minutes performed serially on a 3.3 GHz Intel Core i7 MacBook Pro.

Discussion

The CaCoFold folding algorithm provides a comprehensive description and visualization of all the significantly covarying pairs (even if not nested or overlapping) in the context of the most likely complete RNA structure compatible with all of them. This allows an at-a-glance direct way of assessing which parts of the RNA structure are well determined and which should to be taken just as suggested compatible solutions. The strength of the CaCoFold algorithm is in building RNA structures anchored by all positive (significant covariation) and negative (variation in the absence of covariation) information provided by the alignment.

The purpose of the CaCoFold algorithm is not to be just another folding algorithm, but to integrate in a structure all the information that covariation (positive basepairs) and covariation power

(negative basepairs) combined provide. The helices proposed by CaCoFold that do not include any positive basepair are just to be taken as plausible completions of the structure. Additional data in the form of crystallographic information or alignments with more power would allow to resolve the rest of the structure.

CaCoFold is not the first method to use covariation information to infer RNA structures[7–12], but it is the first to distinguish structural covariation from that of phylogenetic nature, which is key to eliminate confounding covariation noise. And I believe R-scape is also the first method to use negative evolutionary information to discard unlikely basepairs. CaCoFold differs from previous approaches in four main respects: (1) It uses the structural covariation information provided by R-scape which removes phylogenetic confounding. The specificity of R-scape is controlled by an E-value cutoff. (2) It uses the variation information (covariation power) provided by R-scape to identify negative basepairs that should not be allowed to form. (3) It uses a recursive algorithm that incorporates all positive basepairs even those that do not form nested structures, and those that involve positions already forming other basepairs. The CaCoFold algorithm uses different folding algorithms at the different layers. (4) A visualization tool that incorporates all interactions and highlights the positive basepairs.

We have identified over two hundred RNAs in the best curated databases of structural RNAs for which the CaCoFold structure finds new structural elements not present in the originally provided structure. Amongst those new elements we find new helices, basepairs involved in coaxial stacking, new pseudoknots, and triplets.

This work has not addressed the following lines that are also interesting. Analysis of the significant covariation in non Watson-Crick basepairs. The study of significant covariation signatures that do not have a phylogenetic origin in protein-coding mRNA and protein sequences. To use variation and covariation information to improve the quality of RNA structural alignments.

Methods

Implementation

The CaCoFold algorithm has been implemented as part of the R-scape software package. For a given input alignment, there are two main modes to predict a CaCoFold structure using R-scape

covariation analysis as follows,

- To predict a new structure: `--fold`

  All possible pairs are analyzed equally in one single covariation test. This option is most appropriate for obtaining a new consensus structure prediction based on covariation analysis in the absence of a proposed structure.

  The structure in Figure 1 was obtained using this option.

- To improve a existing structure: `-s --fold`

  This option requires that the input alignment has a proposed consensus structure annotation. Two independent covariation tests are performed, one on the set of proposed base pairs, the other on all other possible pairs.

  The structures in Figure 2 were obtained using this option.

Availability

A R-scape web server is available from rivaslab.org/R-scape. The source code can be downloaded from a link on that page. A link to a preprint version of this manuscript with all supplemental information and the R-scape code is also available from that page.

This work uses R-scape version 1.4.0. The distribution of R-scape v1.4.0 includes the external programs: FastTree version 2.1.10[27], and R2R version 1.0.6.1-49-g7bb81fb[28]. The R-scape git repository can be found at https://github.com/EddyRivasLab/R-scape.

For this manuscript, we used the databases Rfam version 14.1 (http://rfam.xfam.org/), and ZWD (114e95ddbeb0) downloaded on February 11, 2019 (https://bitbucket.org/zashaw/zashaweinbergdata/).

All alignment used in the manuscript are provided in the Supplemental Materials.

Acknowledgments

We thank the Centro de Ciencias de Benasque Pedro Pascual in Benasque, Spain, where ideas for this manuscript were developed. We thank I. Kalvari and A. Petrov for their assistance with the

Rfam database, and Z. Weinberg for assistance with the database ZWD and the program R2R. We thank Sean R. Eddy for comments.

References

[1] R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, and A. Zamir, "Structure of a ribonucleic acid," *Science*, vol. 14, pp. 1462–1465, 1965.

[2] H. F. Noller, J. Kop, V. Wheaton, J. Brosius, R. R. Gutell, A. M. Kopylov, F. Dohme, W. Herr, D. A. Stahl, R. Gupta, and C. R. Woese, "Secondary structure model for 23S ribosomal RNA," *Nucl. Acids Res.*, vol. 9, pp. 6167–6189, 1981.

[3] R. R. Gutell, B. Weiser, C. R. Woese, and H. F. Noller, "Comparative anatomy of 16S-like ribosomal RNA," *Prog. Nucl. Acids Res. Mol. Biol.*, vol. 32, pp. 155–216, 1985.

[4] N. R. Pace, D. K. Smith, G. J. Olsen, and B. D. James, "Phylogenetic comparative analysis and the secondary structure of Ribonuclease P RNA – a review," *Gene*, vol. 82, pp. 65–75, 1989.

[5] K. P. Williams and D. P. Bartel, "Phylogenetic analysis of tmRNA secondary structure.," *RNA*, vol. 2, pp. 1306–1310, 1996.

[6] F. Michel, M. Costa, C. Massire, and E. Westhof, "Modeling RNA tertiary structure from patterns of sequence variation.," *Meth. Enzymol.*, vol. 317, pp. 491–510, 2000.

[7] R. R. Gutell, A. Power, G. Z. Hertz, E. J. Putz, and G. D. Stormo, "Identifying constraints on the higher-order structure of RNA: Continued development and application of comparative sequence analysis methods," *Nucl. Acids Res.*, vol. 20, pp. 5785–5795, 1992.

[8] V. R. Akmaev, S. T. Kelley, and G. D. Stormo, "Phylogenetically enhanced statistical tools for RNA structure prediction," *Bioinformatics*, vol. 16, pp. 501–512, 2000.

[9] B. Knudsen and J. Hein, "Pfold: RNA secondary structure prediction using stochastic context-free grammars," *Nucl. Acids Res.*, vol. 31, pp. 3423–3428, 2003.

[10] E. Bindewald and B. A. Shapiro, "RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers," *RNA*, vol. 12, pp. 342–352, 2006.

[11] H. Kiryu, T. Kin, and K. Asai, "Robust prediction of consensus secondary structures using averaged base pairing probability matrices," *Bioinformatics*, vol. 23, pp. 434–441, 2007.

[12] S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler, "RNAalifold: improved consensus structure prediction for RNA alignments," *BMC Bioinformatics*, vol. 9, p. 474, 2008.

[13] E. Rivas, J. Clements, and S. R. Eddy, "A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs," *Nature Methods*, vol. 14, pp. 45–48, 2017.

[14] E. Rivas, J. Clements, and S. R. Eddy, "Estimating the power of sequence covariation for detecting conserved RNA structure," 2019.

[15] I. Kalvari, J. Argasinska, N. Quinones-Olvera, E. P. Nawrocki, E. Rivas, S. R. Eddy, A. Bateman, R. D. Finn, and A. I. Petrov, "Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families," *NAR*, vol. 46, no. D1, pp. D335–D342, 2018.

[16] Z. Weinberg, "Zasha Weinberg Database." [https://bitbucket.org/zashaw/zashaweinbergdata/], 2018.

[17] Z. Weinberg, J. Perreault, M. M. Meyer, and R. R. Breaker, "Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis," *Nature*, vol. 462, pp. 656–659, 2009.

[18] G. M. Landau, U. Vishkin, and R. Nussinov, "An efficient string matching algorithm with k differences for nucleotide and amino acid sequences," *Nucl. Acids Res.*, vol. 14, no. 1, pp. 31–46, 1986.

[19] E. Rivas, R. Lang, and S. R. Eddy, "A range of complex probabilistic models for RNA secondary structure prediction that include the nearest neighbor model and more," *RNA*, vol. 18, pp. 193–212, 2012.

[20] D. H. Mathews and D. H. Turner, "Prediction of RNA secondary structure by free energy minimization," *Curr Opin Struct Biol*, vol. 16, pp. 270–278, 2006.

[21] R. Lorenz, S. H. Bernhart, C. H. Z. Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "ViennaRNA Package 2.0," *Algorithms Mol Biol*, vol. 6, pp. 1748–7188, 2011.

[22] N. R. Markham and M. Zuker, "UNAFold: software for nucleic acid folding and hybridization," *Methods Mol. Biol.*, vol. 453, pp. 3–31, 2008.

[23] Z. Z. Xu and D. H. Mathews, "Secondary structure prediction of single sequences using RNAstructure," *Meth. Mol. Biol.*, vol. 1490, pp. 15–34, 2016.

[24] B. Knudsen and J. Hein, "RNA secondary structure prediction using stochastic context-free grammars and evolutionary history," *Bioinformatics*, vol. 15, pp. 446–454, 1999.

[25] R. D. Dowell and S. R. Eddy, "Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction," *BMC Bioinformatics*, vol. 5, p. 71, 2004.

[26] N. B. Leontis and E. Westhof, "Analysis of RNA motifs," *Curr Opin Struct Biol*, vol. 13, pp. 300–308, 2003.

[27] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree 2 - approximately maximum-likelihood trees for large alignments," *PLOS ONE*, vol. 5, p. e9490, 2010.

[28] Z. Weinberg and R. R. Breaker, "R2R – software to speed the depiction of aesthetic consensus RNA secondary structures," *BMC Bioinformatics*, vol. 12, p. 3, 2011.

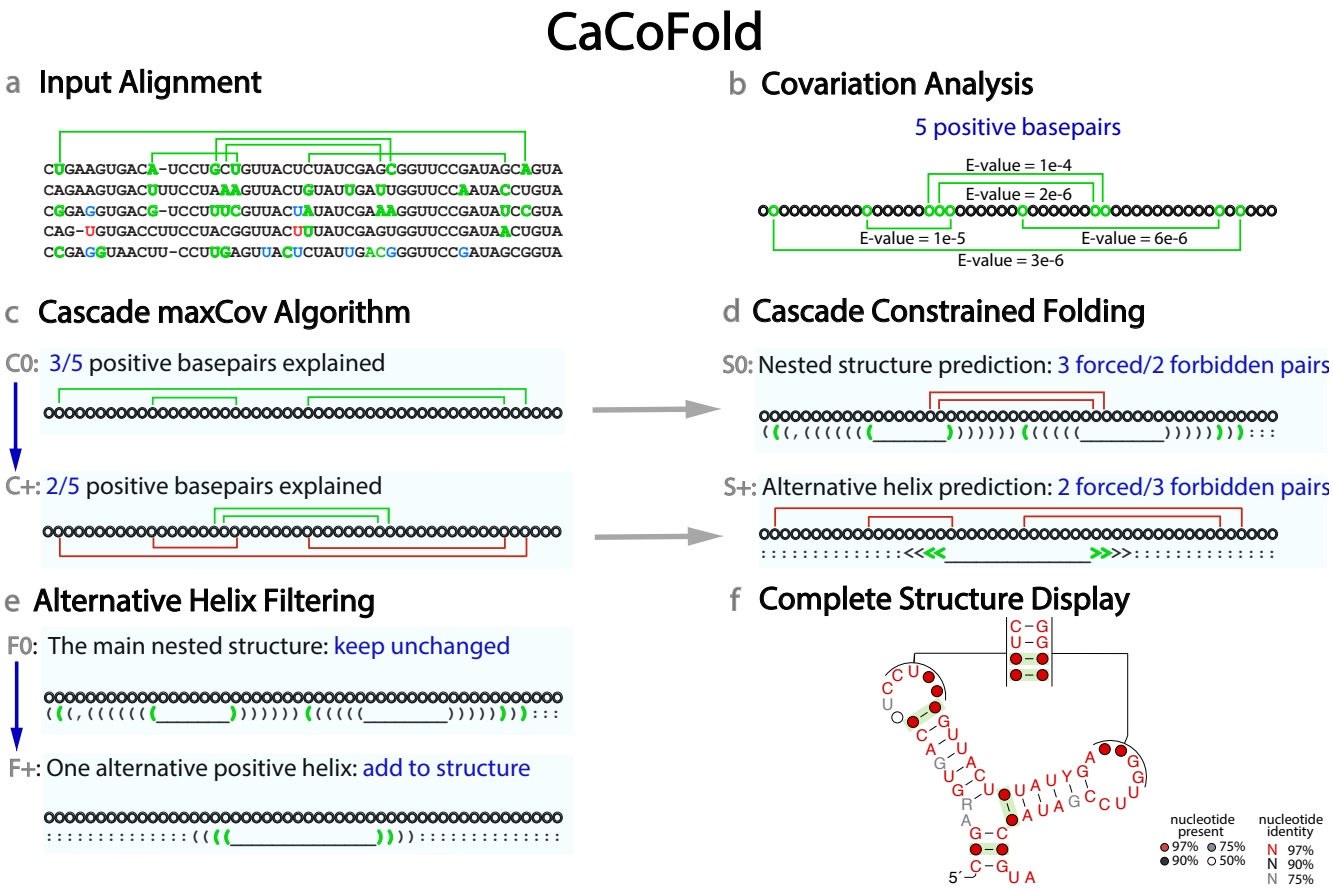

Figure 1: **The CaCoFold algorithm.** **(a)** Toy alignment of five sequences. Residue pairs that significantly covary (positive basepairs) are highlighted in green **(b)**. **(c)** The maxCov algorithm uses a cascade of dynamic programming algorithms to group all positive basepairs in maximal nested subsets that can be explained together. Positive basepairs that are not explained by the first (C0) layer in the recursion, are incorporated in successive layers (C+) where the already explained positive basepairs are excluded. The maxCov algorithm determines the number of layers in the cascade. This example requires two layers. **(d)** At each layer, a dynamic programming algorithm produces the best fold constrained by the positive basepairs assigned by the maxCov algorithm to that layer (green parentheses), to the exclusion of all other positive basepairs (red arches), although the involved residues could form other basepairs. In addition, pairs of residues that have power but do not covary (negative basepairs) are prevented from forming, but each residue can still pair with other residues. (This toy alignment does not include any negative basepairs.) **(e)** The S+ alternative structures are filtered such that alternative helices with positive basepairs (positive helices) are added to the structure. Alternative helices without positive basepairs that overlap in more that half of the residues with the S0 structure are removed. **(f)** The final structure combining the nested S0 structure with the alternative filtered helices from all other layers is displayed automatically adapting the program R2R. Positive basepairs are depicted in green.

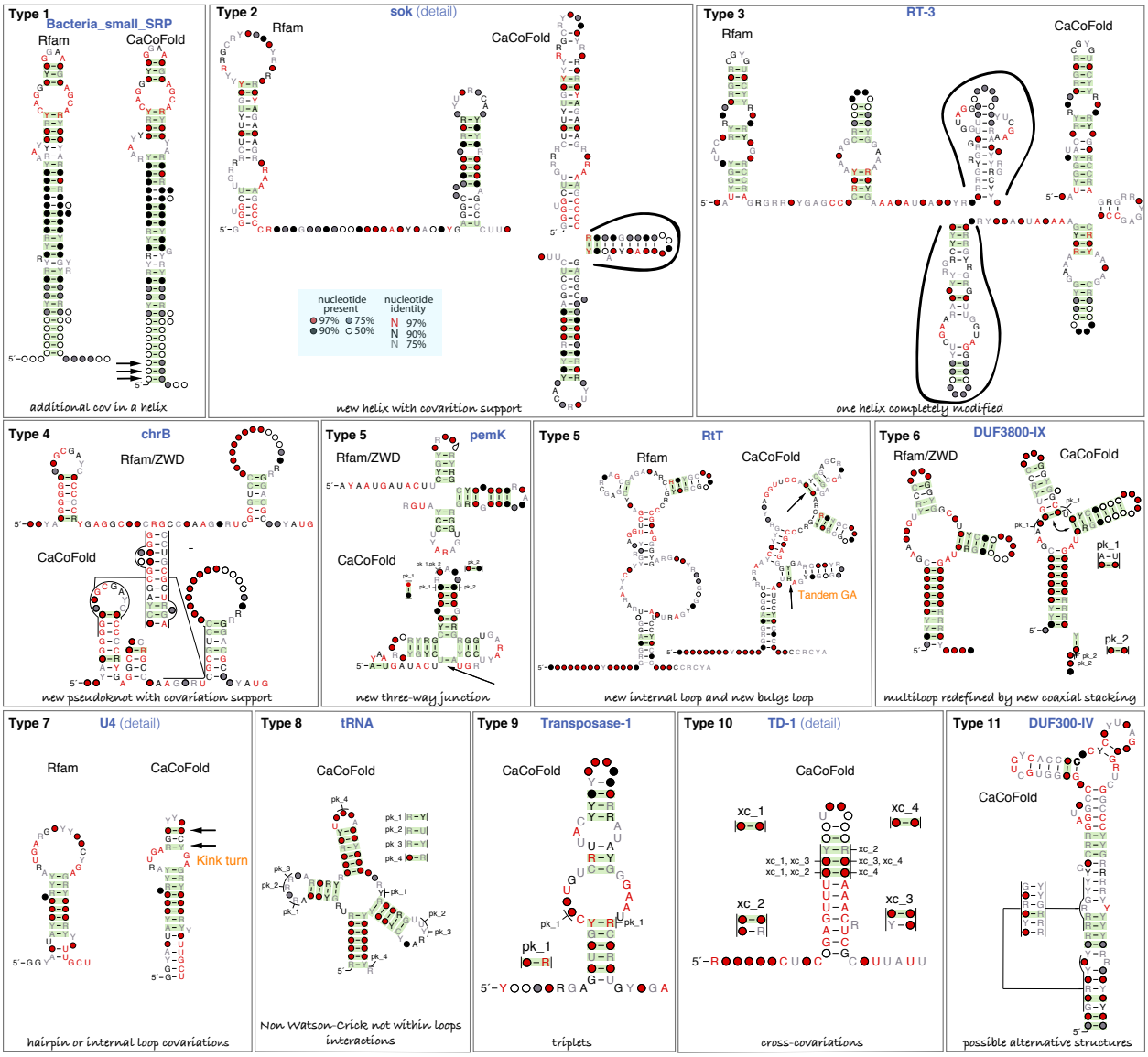

Figure 2: **Examples of RNAs for which the CaCoFold structure has more positive basepairs than the structure given by the corresponding database.** We provide examples of improvements corresponding to Types 1 to 11. A description of all different types is given in Table 2.

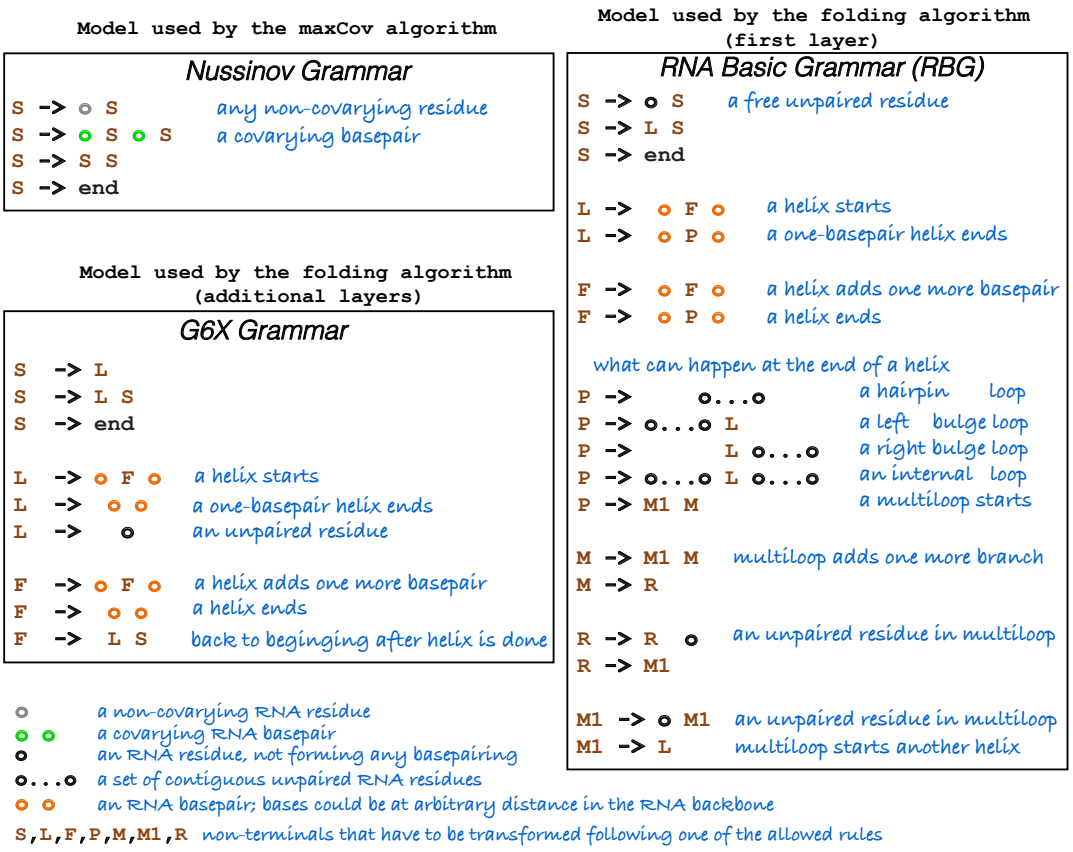

Figure S1. **RNA models used by the CaCoFold algorithm.** The maxCov algorithm implements the Nussinov grammar using as scores the R-scape E-values of the significantly covarying pairs. The maxCov algorithm determines the number of layers necessary to include all covarying basepairs. For the first layer, the CaCoFold folding algorithm uses a probabilistic version of the RBG grammar. The rest of the layers completing the non-nested part of the RNA structure use the G6X grammar. For the RGB and G6X models, the F nonterminal is a shorthand for 16 different non-terminals that represent stacked basepairs. The three models are unambiguous, that is, given any nested structure, there is always one possible and unique way in which the structure can be formulated by following the rules of the grammar.