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

Knowing the structure of conserved structural RNAs is important to elucidate their function and mechanism of action. However, 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 three features. First, it uses significant covariation due to RNA structure and removes spurious covariation due to phylogeny. Second, it uses negative evolutionary 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 overlapping such as competing structures, base triplets, or other 3D non-antiparallel interactions. We present evidence that CaCoFold predictions are consistent with structures modeled from crystallography.

Author Summary

The availability of deeper comparative sequence alignments and recent advances in statistical analysis of RNA sequence covariation have made it possible to identify a reliable set of conserved base pairs, as well as a reliable set of non-basepairs (positions that vary without covarying). Predicting an overall consensus secondary structure consistent with a set of individual inferred pairs and non-pairs remains a problem. Current RNA structure prediction algorithms that predict nested secondary structures cannot use the full set of inferred covarying pairs, because covariation analysis also identifies important non-nested pairing interactions such as pseudoknots, base triples, and alternative structures. Moreover, although algorithms for incorporating negative constraints exist, negative information from covariation analysis (inferred non-pairs) has not been systematically exploited.

Here I introduce an efficient approximate RNA structure prediction algorithm that incorporates all inferred pairs and excludes all non-pairs. Using this, and an improved visualization tool, I show that the method correctly identifies many non-nested structures in agreement with known crystal structures, and improves many curated consensus secondary structure annotations in RNA sequence alignment databases.

Introduction

Having a reliable method to determine the structure of a conserved structural RNA would be an important tool to be able to elucidate important biological mechanisms, and will open the opportunity of discovering new ones. Structure and biological function can be closely related, as in the case of riboswitches where the structure dictates the biological function[1;2], or the bacterial CsrB RNA which acts as a sponge to sequester the CsrA protein[3], or the 6S RNA which mimics the structure of a DNA promoter bound to the RNA polymerase to regulate transcription[4].

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[5–10]. 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[11–16].

Several challenges remain in the determination of a conserved RNA structure using comparative analysis. There is ample evidence that pseudoknotted basepairs covary at similar levels as other basepairs, but 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 versus other sources. 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 previously introduced a method called R-scape (RNA Structural Covariation Above Phylogenetic Expectation)[17] that reports basepairs that significantly covary using a tree-based null model to estimate phylogenetic covariation from simulated alignments with similar base composition and number of mutations to the given one but where the structural signal has been shuffled. 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[18]. Where a pair of position shows no significant covariation, this method allows distinguishing between two different cases: a pair that has too little sequence variation and may still be a conserved basepair, versus a pair with adequate sequence variation but where the variation is inconsistent with a covarying basepair. This latter case should be rejected as basepairs. We call these pairs with variation but not covariation the negative basepairs.

Here we combine these two sources of information (positive in the form of significantly covarying basepairs, and negative in the form of pairs of positions unlikely to form basepairs) into a new RNA folding algorithm. The algorithm also introduces an iterative procedure that systematically incorporates all positive basepairs into the structure while remaining computationally efficient. The recursive algorithm is able to find pseudoknots, other non-nested interactions, alternative structures and triplet interactions provided that they are supported by covariation. The algorithm also predicts additional helices without covariation support but consistent with RNA structure. Helices with covariation-supported basepairs tend to be reliable. Additional helices lacking covariation support are less reliable and need to be taken as speculative.

We use the alignments provided by the databases of structural RNAs Rfam[19] and the Zasha Weinberg Database (ZWD)[20] to produce CaCoFold structure predictions. The number of positive pairs (that is, significantly covarying basepairs proposed by R-scape) is constant for a given alignment. We compare how many positive pairs are incorporated into CaCoFold structures versus annotated structures, comparing with structures derived by crystallography when possible.

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 basepairs. 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[17]. The set of all significantly covarying basepairs is called the positive set. We also calculate the covariation power for all possible pairs[18]. The set of all pairs that have variation but not covariation is called the negative set. Operationally, positive pairs have an E-value smaller than 0.05, and negative pairs are those with covariation power (the expected sensitivity of significantly covarying) larger than 0.95 and significance E-value larger than one. Non-significantly covarying pairs with an E-value between 0.05 and 1 are allowed (but not forced) to basepair regardless of power. All positive basepairs are included in the final structure, and all negative basepairs are forbidden to appear.

Fig. 1 illustrates the CaCoFold algorithm using a toy alignment (Fig. 1a) derived from the manA RNA, a structure located in the 5' UTRs of cyanobacterial genes involved in mannose metabolism[21]. After R-scape with default parameters identifies five positive basepairs (Fig. 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 (Fig. 1c). At each layer, it uses the Nussinov algorithm, one of the simplest RNA models[22]. 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. Fig. 2 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 all 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 (Fig. 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)[23]. The RBG model features the same basic elements as the nearest-neighbor thermodynamic model of RNA stability[24;25] such as basepair stacking, the length of the different loops, the length of the helices, the occurrence of multiloops, and others. RBG simplifies some details of loops in the models used in the standard thermodynamic packages, such as ViennaRNA[25], Mfold[26], or RNAStructure[27] resulting in fewer parameters, but it has comparable performance regarding folding accuracy[23]. Fig. 2 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 non-nested tertiary contacts, or base triplets. The S+ layers add alternative helices (complementary or not) to the main secondary structure, for that reason instead of a full loop model like RBG, the S+ layers use the simpler G6 RNA model[28;29] which mainly models the formation of helices of contiguous basepairs. Here we extend the G6 grammar to allow positive pairs that are parallel to each other in the RNA backbone, interactions that are not uncommon in RNA motifs. We name the modified grammar G6X (see Fig. 2 for a description).

The RBG and G6X model parameters are trained on a large and diverse set of known RNA structures and sequences as described[23]. At each layer, the corresponding probabilistic folding algorithm reports the structure with the highest probability using a CYK dynamic programming algorithm on a profile sequence that contains information on the proportion of each nucleic acid in each consensus column of 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 without covariation support.

We first break the S+ structures into individual alternative helices. A helix is operationally defined as a set of contiguous basepairs with at most two residues are unpaired (forming a one or two residue bulge or a 1x1 internal loop). Under this operational definition, a helix can consist of just one basepair, and each basepair belongs to one and only one helix. 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 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 (Fig. 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 green. CaCoFold reports and draws a consensus structure for the alignment. Conserved positions display the residue identity color coded by conservation (red >97%, black >90%, and gray >75%), otherwise a circle is displayed colored by column occupancy (red > 97%, black > 90%, gray >75%, white >50%).

We adapted the R2R software[30] to depict all non-nested pairs automatically (Fig. 1f). Alternative helices that do not overlap with the main nested structure are annotated as pseudoknots ("pk"). Alternative helices that overlap with the nested structure are annotated as triplets ("tr"). For 3D structures, non Watson-Crick basepairs (regardless of whether they overlap or not) are annotated as non-canonical ("nc").

If R-scape does not identify any positive basepair, one single layer is defined without positive pairs and constrained only by the negative pairs, 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 Fig. 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 finds pseudoknots, triplets and other long and short-range interactions

For a realistic example of how CaCoFold works, we present in Fig. 3 an analysis of transfer-messenger RNA (tmRNA). The tmRNA is a bacterial RNA responsible for freeing ribosomes stalled at mRNAs without a stop codon. The tmRNA molecule includes a tRNA-like structural domain, and a mRNA domain which ends with a stop codon. The tmRNA molecule is typically 230-400 nts[31], and its proposed structure includes a total of 12 helices forming four pseudoknots[32]. The core elements of the tmRNA structure are well understood, but the molecule has a lot of flexibility and is thought to undergo large conformational changes with the 4 pseudoknots forming a ring around a part of the small subunit of the ribosome[31].

We performed the analysis on the tmRNA seed alignment in Rfam (RF0023) which includes 477 sequences. The length of the consensus sequence is 354 nts, and the average pairwise percentage identity is 42% (Fig. 3a). In step one, the covariation analysis on the input alignment (ignoring the proposed consensus structure) results on 121 significantly covarying basepairs (Fig. 3b). This result is in agreement with the covariation power estimated for this alignment, which expects to find on average 109 significantly covarying basepairs[18]. In the next step, the maxCov algorithm requires 6 layers to explain all 121 positive basepairs (Fig. 3c). Next, the constrained folding of each of the 6 layers results on a total of 139 annotated pairwise interactions.

The covariation analysis also identifies 31,027 negative pairs (out of a total of 85,491 possible pairs for 414 columns analyzed), those are forbidden to form because they show variation but not covariation. In the final structure, 74 baseapairs are not reported do to the forbidden negative pairs (Fig. 3d). The final alternative helix filtering step reports: 5 pseudoknots, 3 triplets and 10 other covariations that are induced by coding constraints, which we describe in more detail below. All alternative helices have covariation support (Fig. 3e).

The CaCoFold structure for the tmRNA is given in Fig. 3f, and it includes the 12 helices and four pseudoknots[32]. It also proposes an additional helix (H13) with covariation support. We have not identified H13 in tmRNA crystal structures. Due to the amount of overlap between H13 and helix H2d, this could indicate the presence of two alternative competing structures.

In the helix H2d/helix H3 region, CaCoFold identifies three triplets, one of them (triplet 1) is confirmed by the structure derived from the tmRNA EM structure (13.6 Å) with PDB ID 3IZ4[33]. A different triplet for which we do not find covariation signal has been previously proposed in that same region (Fig. 3g). This is a complex region with many 3D contacts as helix H2d interacts both with the PK1 and PK4 pseudoknots[31].

CaCoFold identified 10 additional interactions associated to the mRNA domain. These tend to occur between contiguous residues. These interactions are not related to the RNA structure and arise from coding constraints (more details in Supplemental Fig. S6c). We observe this kind of covariation in other coding mRNA regions, not just in tmRNA. Finally, CaCoFold reports one covariation between the first and second position of the stop codon in the mRNA domain. The U residue in the first position of the stop codon is invariant, so a covariation involving this reside should not occur. This spurious covariation arises from a misalignment of the stop codon in the Rfam seed alignment. A small rearrangement of the alignment in that region results in a conserved stop codon.

We compared the tmRNA CaCoFold structure to the structure predicted for the same alignment by RNAalifold, a ViennaRNA program for predicting a consensus structure[34]. **??**(a) shows the output of RNAalifold. RNAalifold does not predict pseudoknots or any other non-nested structure, and it only identifies 6 of the 12 helices in the tmRNA structure (Fig. 3g). RNAalifold predicts 46 basepairs, but it does not assign confidence to the proposed basepairs. In **??**(b), the covariation analysis of the tmRNA alignment shows that 45 of the 46 RNAalifold basepairs covary. But it also indicates that there are 76 other covarying basepairs not present in the RNAalifold structure (Fig. 3b). CaCoFold brings together the basepair validation provided by the covariation analysis with a structure that incorporates all 121 inferred basepairs.

RNAs with structures improved by positive and negative signals

We have produced CaCoFold structures from the alignments provided by the databases of structural conserved RNAs Rfam[19] and ZWD[20]. Unlike the previous section where the proposed consensus structure was ignored, here we perform two independent covariation tests: one on the set of basepairs in the annotated consensus structure, another on the set of all other possible pairs (option "to improve an existing structure" in Methods). It is important to notice that because of this two-set analysis, CaCoFold builds on the knowledge provided by the alignments and the consensus structures of Rfam and ZWD. Using the positive and negative pairs obtained from the covariation test as constraints, the CaCoFold structure is then built anew.

One strength of the CaCoFold algorithm is in the association between covariation above phylogenetic expectation with RNA structure. For alignment with little or no significant covariation, CaCoFold behaves as the RBG model, which we have shown in benchmarks perform similarly to standard methods[23]. Because in the absence of covariation RNA structure prediction lack reliability and all methods perform comparably, we concentrate on the set of RNAs with covariation support.

Another strength of the CaCoFold algorithm is in incorporating all covariation signal present in the alignment into one structure. When the CaCoFold structure includes the same covarying pairs than the annotated structure, the differences between the two structures can only occur in regions not reliably predicted by either of the methods, thus we concentrate on the set of RNAs for which the CaCoFold structure has different covariation support than the annotated structure.

Because the set of positive pairs is constant and CaCoFold incorporates all of them, CaCoFold structures cannot have fewer positive pairs than the database consensus structures. Here we investigate the set of RNAs with CaCoFold structures with different (that is, more) covariation support than the annotated structures, and whether those differences are consistent with experimentally-determined 3D structures when available.

We identify 277 (out of 3,030) Rfam families and 105 (out of 415) 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 the CaCoFold structure has more covariation support than either the Rfam structure or the ZWD structure. Of the 314 RNAs, there are five for which the Rfam and ZWD alignments and structures are different (PhotoRC-II/RF01717, manA/RF01745, radC/RF01754, pemK/RF02913, Mu-gpT-DE/RF03012) and we include both versions in our analysis. In the end, we identify a total of 319 structural alignments for which the structure presented in the databases is missing positive basepairs, and CaCoFold proposes a different structure with more covariation support. In Table 1, we classify all structural differences into 15 types.

Modifications introduced by the extra covariations in the CaCoFold structure		# RNAs
Type 1	Helix extended by additional covariations	23
Type 2	New helix with covariation support	12
Type 3	One helix completely modified	7
Type 4	New pseudoknot with covariation support	16
Type 5	New junction/internal loop or coaxial stacking	17
Type 6	Internal loop/multiloop reshaped by coaxial stacking	12
Type 7	Hairpin/internal loop covariations (often nonWC)	19
Type 8	Non Watson-Crick (not within a loop) covariations	24
Type 9	Base triples	28
Type 10	(Cross,Side)-covariations (see text)	30
Type 11	Possible alternative structures	6
Type 12	Additional covariations in SSU and LSU rRNA	6
Type 13	Covariations not supporting a secondary structure	3
Type 14	Misalignment introducing spurious covariations	2
Type 15	Low power; inconclusive	114
CaCoFold structures with different (*i.e.* more) covariation support		319

21/319 RNAs with 3D structures

RNA	Rfam seed alignment	Types	Figure
RNase P RNA A-type[35]	RF00010	4,8	3a
SAM-I riboswitch[36]	RF00162	1,4,6	3b
U4 snRNA[37]	RF00015	2,5	3c
Cobalamin riboswitch[38]	RF00174	1,4,5	4a
tRNA[39;40]	RF00005	1,8,9	4b
U2 snRNA[41;42]	RF00004	11	4c
Bacterial SRP RNA[43]	RF00169	1	S2a
cyclic di-AMP riboswitch[44]	RF00379	1	S2b
YkoK leader[45]	RF00380	1	S2c
5S rRNA[46]	RF00001	3,5	S3a
FMN riboswitch[47]	RF00050	1,4	S3b
ZPM-ZTP riboswitch[48]	RF01750	4,9	S3c
Fluoride riboswitch[49]	RF01734	1,4	S4a
Glutamine riboswitch[50]	RF01739	4	S4b
Archaea SRP RNA[51]	RF01857	4	S4c
RNase P RNA B-type[52]	RF00011	5	S5a
group-II intron[53]	RF02001	5	S5b
U5 snRNA[54]	RF00020	5,7,10	S5c
Fungal U3 snoRNA[55]	RF01846	5	S6a
6S RNA[4]	RF00013	8	S6b
tmRNA[56]	RF00023	9,10,11,14	S6c

Table 1: **CaCoFold structures with different covariation support than the structures provided with the structural alignments.** CaCoFold structures with different covariation support can only have more positive basepairs. **(Left)** The 319 structural RNAs (from the Rfam and ZWD databases combined) for which the CaCoFold structure has more covariation support are manually classified into 15 categories. Each RNA is assigned to one main type, although they can belong to others as well. Examples of types 1-11 are presented in Fig. S7. A full description of all 319 RNAs is given in the supplemental table. **(Right)** Subset of 21/319 CaCoFold structures with more covariation support for which there is 3D structural information (not including the 6 rRNAs). We compare the 21 CaCoFold predicted structures to the 3D structures in Fig. 4, 5, and Supplemental Fig. S2-S6.

CaCoFold structures consistent with 3D structures

The set of 319 CaCoFold structures with more covariation support includes 27 RNAs that have 3D structures for representative sequences (out of a total of 97 families with 3D structures). We tested that for those RNAs (21 total, leaving aside 1 LSU and 5 SSU rRNA) the CaCoFold structure predictions are indeed supported by the 3D structures. Table 1 describes the 21 RNAs: 5S RNA, tRNA, 6S RNA, group-II intron, two bacterial RNase P RNAs (A-type and B-type), tmRNA, two SRP RNAs (bacterial and archaeal), four snRNAs (U2, U3, U4, and U5), and eight riboswitches (FMN, SAM-I, Cobalamin, Fluoride, Glutamine, cyclic di-AMP, and YkoK leader). The comparison of the CaCoFold structures for those 21 RNAs to 3D structures are presented in Fig. 4, 5 and supplemental Fig. S2-S6.

In Fig. 4a, we show the A-type RNase P RNA where CaCoFold identifies the two pseudoknots, one of which (P6) is not in the Rfam consensus structure. CaCoFold also identifies two long-range triplet interactions (tr_1 and tr_2) described in ref. 35, although for “tr_2” (between P8 and the P14 loop) there is a one-position discrepancy between the 3D structure and CaCoFold in the identity of the P14 residue. This could be due to a misalignment in the P14 loop, or an ambiguity in the correspondence between the consensus structure which accommodates many individual variants and the structure of one particular species, *Thermotoga maritima* in this case[35].

Fig. 4b shows the SAM-I riboswitch where CaCoFold identifies the reported pseudoknot[36]. Other RNAs for which CaCoFold identifies unannotated pseudoknots with covariation support confirmed by crystallography include five riboswitches: the Cobalamin riboswitch[38] (Fig. 5a), FMN riboswitch[47] (Fig. S3b), ZMP/ZTP riboswitch[48] (Fig. S3c), Fluoride riboswitch[49] (Fig. S4a), Glutamine riboswitch[50] (Fig. S4b), and the Archaeal SRP RNA[51] (Fig. S4c). Also in the SAM-I riboswitch, CaCoFold identifies an apparent lone Watson-Crick A-U pair in the junction of the four helices which in fact stacks with helix P1[36].

The SAM-I riboswitch[36] CaCoFold structure also includes additional covariations that further extend existing helices P2a, P3 and P4. Other RNAs for which CaCoFold identifies additional covarying pairs in helices supported by 3D structures are given in Fig. S2: Bacterial SRP RNA[43] (Fig. S2a), cyclic di-AMP riboswitch[44] (Fig. S2b), and YkoK leader[45] (Fig. S2b)

In Fig. 4c, the U4 spliceosomal snRNA shows two covarying pairs identifying a new internal loop including a kink turn RNA motif[37]. Four other RNAs for which CaCoFold identifies key covarying residues are: RNase P RNA B-type[52] (Fig. S5a) where one covarying basepair identifies a new internal loop, the group-II intron[53] (Fig. S5b) where one covarying basepair defines a new three-way junction, U5 snRNA[54] (Fig. S5c) where a Y-Y covarying pair modifies a hairpin loop, and the Fungal U3 snoRNA (Fig. S6a) where a R-Y covarying pairs allows identifying the characteristic boxB/boxC boxes of the snoRNA[55].

In Fig. 5a, the CaCoFold structure for the Cobalamin riboswitch[38] includes a pseudoknot, a covarying pair identifying a multiloop with coaxial stacking, and additional covarying basepair in helices P1 and P2 all supported by the 3D structure[38]. CaCoFold also identifies other unreported covarying pairs in the internal loop between helices P7 and P8.

The tRNA CaCoFold structure (Fig. 5b) incorporates many long-range interactions, five of them are confirmed by the crystal structure with PDB ID 1EHZ, one of the higher resolution tRNA structures (1.93 Å). There is one more interaction identified by CaCoFold involving one anticodon residue and the discriminator residue in the acceptor stem. This anticodon/discriminator covariation results from the interaction of both residues with aminoacyl-tRNA synthetase[39]. CaCoFold identifies six additional covarying pairs not reported by RNAView on the 1EHZ tRNA crystal structure.

In Fig. 5c, the U2 spliceosomal snRNA describes a case of alternative structures. "Stem IIc" was originally proposed as possibly forming a pseudoknot with one side of Stem IIa, but was later discarded as non-essential for U2 function[41;57]. But later, a U2 conformational switch was identified indicating that Stem IIa and Stem IIc do not form a pseudoknot but are two competing helices promoting distinct splicing steps[42]. Both helices are important to the U2 function, and both have covariation support.

5S rRNA (Fig. S3a) shows the case of a region (the helix 4 and Loop E region) almost completely reshaped by the covariations found by CaCoFold, and in agreement with the 3D structure[46].

In addition to the coding mRNA signal in tmRNA (Fig. S6c), we have found another signal that produces non-phylogenetic covariations in the 6S RNA (Fig. S6b) which regulates transcription by direct binding to the RNA polymerase[4]. The 6S RNA structure mimics an open promoter and serves as a transcription template. Synthesis of a 13 nt product RNA from the 6S RNA results in a structural change that releases the RNA polymerase. We do not find any covariation evidence for the alternative helix of "isoform 2" in Ref. 4 (Fig. S6b), but we observe one covariation between the U initiating the RNA product and the previous position. We hypothesize an interaction of the two bases with the RNA polymerase.

Other CaCoFold structures with more covariation support

Based on what we learned from the 3D structures, we manually classified the 319 RNAs with modified structures into 15 categories (Table 1). In Supplemental Table S1, we report a full list of the RNA families and alignments with CaCoFold structures incorporating more positive covariation support, classified according to Table 1. In Fig. S7, we show representative examples of Types 1-12 amongst the RNAs with more covariation support but without 3D structures.

In **Type 1**, the extra positive basepairs incorporated by CaCoFold extend the length of an already annotated helix, as in the TwoAYGGAY RNA (RF01731) and drum RNA (RF02958) examples. **Type 2** includes cases in which several positive basepairs identify a new helix. We present the case of the Coronavirus 3'UTR pseudoknot, a pseudoknotted structure specific to coronaviruses, typically 54-62 nts in length found within the 3' UTR of the N gene. The alignment for this RNA in the Rfam 14.2 Coronavirus special release (RF00165) has a consensus sequence of 62 nts, and it annotates two helices forming a pseudoknot[58]. The CaCoFold structure includes one additional third helix with 2 positive pairs and compatible with the pseudoknot. The existing chemical modification data for the Coronavirus 3'UTR pseudoknot does not rule out the presence of this additional helix[58]. **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 original helix has no covariation support but 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 introducing new (three-way or higher) junctions or new 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[59], as in the case shown in Fig. S7, where an additional positive basepair identifies a tandem GA motif in the RtT RNA. For **Type 6**, we show another positive basepair in the DUF38000-IX RNA that highlights the coaxial stacking of two helices. Other more general non-Watson-Crick interactions are collected in **Type 8**, of which tRNA is an exceptional example in which almost all positions are involved in some covarying interaction. In Fig. S7 we show another example, Bacteroides-2, a candidate structured RNA[21]. **Type 9** are putative base triplets involved in more than one positive interaction. In general, one of the positive basepairs is part of an 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 base triplet that we name cross-covariation and side-covariations. A cross(side)-covariation appears when two covarying basepairs $i-j$ and $i'-j'$ that belong to the same helix are such that two of the four residues form another covarying interaction. If the extra covarying pair involves residues in one side of the helix ($i-i'$ or $j-j'$), we name it a side-covariation (annotated "sc" in the graphical representation). If the residues are in opposite sides of the helix ($i-j'$ or $j-i'$), it is a cross-covariation (annotated "xc"). We have observed side covariations in tmRNA (Fig. 3, and Fig. S6c) and other mRNA sequences. In Fig. S7, 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 competing 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 few and cannot provide confirmation of the proposed 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 (114/319) 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.

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 (*i.e.* supported by significant covariation). The strength and key features of the CaCoFold algorithm are in building RNA structures anchored both by all positive (significant covariation) and negative (variation in the absence of covariation) information provided by the alignment. 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.

CaCoFold is not the first method to use covariation information to infer RNA structures[11–16], but it is the first to our knowledge to distinguish structural covariation from that of phylogenetic nature, which is key to eliminate confounding covariation noise. CaCoFold is also the first method to our knowledge 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) to identify negative basepairs that are not allowed to form. (3) It uses a recursive algorithm that incorporates all positive basepairs even those that do not form nested structures, or involve positions already forming other basepairs. The CaCoFold algorithm uses different probabilistic folding algorithms at the different layers. (4) A visualization tool derived from R2R that incorporates all interactions and highlights the positive basepairs.

Overall, we have identified over two hundred RNAs for which CaCoFold finds new significantly covarying structural elements not present in curated databases of structural RNAs. For the 21 RNAs in that set with 3D information (leaving aside SSU and LSU rRNAs), we have shown that the new CaCoFold elements are generally supported by the crystal structures. Those new elements include new and re-shaped helices, basepairs involved in coaxial stacking, new pseudoknots, long-range contacts and base triplets. Reliable CaCoFold predictions could accelerate the discovery of still unknown biological mechanisms without having to wait for a crystal structure.

We have found interesting cases of significantly covarying pairs where the covariation is not due to RNA structure, the tRNA acceptor/discriminator covariation (Fig. 4) or the coding covariations associated to the messenger domain of tmRNA (Fig. 2, Fig. S6c) are examples. These covariations do not interfere with the determination of the RNA structure, which usually forms during the first layers of the algorithm, as they are added by higher layers on top of the RNA structure. The CaCoFold visual display of all layered interactions permits to identify the RNA structure and to asses its covariation support, and may help proposing hypotheses about the origin of other interactions of different nature.

Even for RNAs with a known crystal structure, because that experimental structure may have only captured one conformation, CaCoFold can provide a complementary analysis, as in the case of the U2 spliceosomal snRNA presented here (Fig. 5c). (Riboswitches also have alternative structures, but because Rfam alignments do not typically include riboswitch expression platform regions, we do not observe the alternatively structured regions of riboswitches in these data.)

CaCoFold improves the state of the art for accurate structural prediction for the many structural RNAs still lacking a crystal structure. This work provides a new tool for several lines of research such as: the study of significant covariation signatures of no phylogenetic origin present in messenger RNA, as those identified here in the tmRNA (Fig. 3, Fig. S6c); the study of the nature and origin of covariation in protein sequences; and the use of 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: `R-scape --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 structures in Fig. 1, 3 were obtained using this option.

- To improve a existing structure: `R-scape -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 CaCoFold structure is built anew using the positive and negative basepairs as constraints.

  The structures in Fig. 4, 5, and Supplemental Fig. S2-S7 were obtained using this option.

Extracting the RNA structure from a PDB file

The software is capable of obtaining the RNA structure from a PDB file for a sequence homolog to but not necessarily represented in the alignment, and transforms it to a consensus structure for the alignment.

For a given PDB[60] file, we use the software nhmmer[61] to evaluate whether the PDB sequence is homologous to the aligned sequences. If the PDB sequence is found to be a homolog of the sequences in the input alignment, we extract the RNA structure from the PDB file (Watson-Crick and also non-canonical basepairs and contacts) using the program RNAView[62]. An Infernal model is built using the PDB sequence and the PDB-derived RNAView structure[63]. All input sequences are then aligned to the Infernal PDB structural model. The new alignment includes the PDB sequence with the PDB structure as its consensus structure. We use the mapping of each sequence to the PDB sequence in this new alignment to transfer the PDB structure to the sequence as it appears in the input alignment. From all individual structures, we calculate a PDB-derived consensus structure for the input alignment. The R-scape software can then analyze the covariation associated with the PDB structure mapped to the input alignment.

For example, the PDB structure and covariation analysis in Fig. 5b for the tRNA (RF00005) was derived from the PDB file 1EHZ (chain A) using the options:

```
R-scape -s --pdb 1ehz.pdb --pdbchain A --onlypdb RF00005.seed.sto
```

The option `--pdbchain <chain_name>` forces to use only the chain of name `<chain_name>`. By default, all sequence chains in the PDB file are tested to find those with homology to the input alignment. The option `--onlypdb` ignores the alignment consensus structure. By default, the PDB structure and the alignment consensus structure (if one is provided) would be combined into one annotation.

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 1.5.2. The distribution of R-scape v1.5.2 includes external programs: FastTree version 2.1.10[64], Infernal 1.1.2[63], hmmer 3.3[65]. It also includes modified versions of the programs RNAView[62], and R2R version 1.0.6.1-49-g7bb81fb[30]. The R-scape git repository is at `https://github.com/EddyRivasLab/R-scape`.

For this manuscript, we used the databases Rfam version 14.1 (`http://rfam.xfam.org/`), the 10 new families and 4 revised families in Rfam 14.2, and ZWD (114e95ddbeb0) downloaded on February 11, 2019 (`https://bitbucket.org/zashaw/zashaweinbergdata/`). We used program RNAalifold from the ViennaRNA-2.4.12 software package[34].

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

Acknowledgments

I thank the Centro de Ciencias de Benasque Pedro Pascual in Benasque, Spain, where ideas for this manuscript were developed. Special thanks to Ioanna Kalvari and Anton Petrov for their assistance with the Rfam database, Zasha Weinberg for assistance with the database ZWD and the program R2R, and Sean R. Eddy for comments.

References

[1] A. S. Mironov, I. Gusarov, R. Rafikov, L. E. Lopez, K. Shatalin, R. A. Kreneva, D. A. Perumov, and E. Nudler, "Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria cell," *Cell*, vol. 5, pp. 747–756, 2002.

[2] W. C. Winkler, A. Nahvi, and R. R. Breaker, "Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression," *Nature*, vol. 419, pp. 952—956, 2002.

[3] P. Babitzke and T. Romeo, "CsrB sRNA family: sequestration of RNA-binding regulatory proteins," *Current Opinion in Microbiology*, vol. 10, no. 2, pp. 156–163, 2007.

[4] J. Chen, K. M. Wassarman, S. Feng, K. Leon, A. Feklistov, J. T. Winkelman, Z. Li, T. Walz, E. A. Campbell, and S. A. Darst, "6S RNA Mimics B-Form DNA to Regulate Escherichia coli RNA Polymerase," *Mol. Cell*, vol. 68, no. 2, pp. 388–397.e6, 2017.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] E. Bindewald, R. Hayes, Y. G. Yingling, W. Kasprzak, and B. A. S. BA, "RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign," *Nucl. Acids Res.*, vol. 36, pp. D392–D397, 2008.

[15] H. Kiryu, T. Kin, and K. Asai, "Rfold: an exact algorithm for computing local base pairing probabilities," *Bioinformatics*, vol. 24, pp. 367–373, 2008.

[16] 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.

[17] 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.

[18] E. Rivas, J. Clements, and S. R. Eddy, "Estimating the power of sequence covariation for detecting conserved RNA structure," *Bioinformatics*, 02 2020. btaa080.

[19] 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," *Nucl. Acids Res.*, vol. 46, pp. D335–D342, 2018.

[20] Z. Weinberg, "The Zasha Weinberg Database (ZWD)," 2018. Available: https://bitbucket.org/zashaw/zashaweinbergdata/. Accessed 11 February 2019.

[21] Z. Weinberg, J. X. Wang, J. Bogue, J. Yang, K. Corbino, R. H. Moy, and R. R. Breaker, "Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes," *Genome Biol.*, vol. 11(3):R31, 2010.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] Z. Z. Xu and D. H. Mathews, "Experiment-assisted secondary structure prediction with RNAstructure: Methods and Protocols," *Methods in Molecular Biology*, vol. 1490, pp. 163–176, 2016.

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

[29] 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.

[30] 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.

[31] F. Weis, P. Bron, E. Giudice, J. P. Rolland, D. Thomas, B. Felden, and R. Gillet, "tmRNA-SmpB: a journey to the centre of the bacterial ribosome.," *EMBO J*, vol. 29, pp. 3810–3818, 2010.

[32] S. T. Kelley, J. K. Harris, and N. R. Pace, "Evaluation and refinement of tmRNA structure using gene sequences from natural microbial communities," *RNA*, vol. 7, pp. 1310—1316, 2001.

[33] J. Fu, Y. Yaser Hashem, I. Wower, J. Lei, H. Y. Liao, C. Zwieb, J. Wower, and J. Frank, "Visualizing the transfer-messenger RNA as the ribosome resumes translation," *EMBO J*, vol. 29, pp. 3819—3825, 2010.

[34] R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "ViennaRNA package 2.0," *Algorithms Mol. Biol.*, vol. 6, p. 10, 2011.

[35] A. Torres-Larios, K. K. Swinger, A. S. Krasilnikov, T. Pan, and A. Mondragón, "Crystal structure of the RNA component of bacterial ribonuclease P," *Nature*, vol. 437, no. 7058, pp. 584–587, 2005.

[36] R. Montange and R. T. Batey, "Structure of the S-adenosylmethionine riboswitch regulatory mRNA element," *Nature*, vol. 441, pp. 1172–1175, 2006.

[37] R. Wan, C. Yan, R. Bai, L. Wang, M. Huang, C. C. L. Wong, and Y. Shi, "The 3.8 å structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis," *Science*, vol. 351, no. 6272, pp. 466–475, 2016.

[38] A. Peselis and A. S. Serganov, "Structural insights into ligand binding and gene expression control by an adenosylcobalamin riboswitch," *Nat Struct Mol Biol*, vol. 19, pp. 1182—1184, 2012.

[39] R. Giegé, M. Sissler, and C. Florentz, "Universal rules and idiosyncratic features in tRNA identity," *NAR*, vol. 26, pp. 5017—5035, 1998.

[40] H. Shi and P. B. Moore, "The crystal structure of yeast phenylalanine tRNA at 1.93 A resolution: a classic structure revisited," *RNA*, vol. 6, p. 1091–1105, 2000.

[41] M. Ares and A. H. Igel, "Mutations define essential and nonessential U2 RNA structures," *Mol. Biol. Rep.*, vol. 14, no. 2-3, pp. 131–132, 1990.

[42] R. J. Perriman and M. Ares, "Rearrangement of competing U2 RNA helices within the spliceosome promotes multiple steps in splicing," *Genes Dev.*, vol. 21, no. 7, pp. 811–820, 2007.

[43] S. F. Ataide, N. Schmitz, K. Shen, A. Ke, S. O. Shan, J. A. Doudna, and N. Ban, "The crystal structure of the signal recognition particle in complex with its receptor," *Science*, vol. 331, pp. 881–886, 2011.

[44] A. Gao and A. Serganov, "Structural insights into recognition of c-di-AMP by the ydaO riboswitch," *Proc. Natl. Acad. Sci. USA*, vol. 10, pp. 787–792, 2014.

[45] C. E. Dann, C. A. Wakeman, C. L. Sieling, B. S. C., I. Irnov, and W. C. Winkler, "Structure and mechanism of a metal-sensing regulatory RNA," *Cell*, vol. 130, pp. 878–892, 2007.

[46] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz, "The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution," *Science*, vol. 289, pp. 905–920, 2000.

[47] A. Serganov, L. Huang, and D. J. Patel, "Coenzyme recognition and gene regulation by a flavin mononucleotide riboswitch," *Nature*, vol. 458, pp. 233—237, 2009.

[48] C. P. Jones and A. R. Ferré-D'Amaré, "Recognition of the bacterial alarmone ZMP through long-distance association of two RNA subdomains," *Nat Struct Mol Biol*, vol. 22, no. 9, pp. 679–685, 2015.

[49] A. Ren, K. R. Rajashankar, and D. J. Patel, "Fluoride ion encapsulation by Mg2+ ions and phosphates in a fluoride riboswitch," *Nature*, vol. 486, pp. 85–89, 2012.

[50] A. Ren, Y. Xue, A. Peselis, A. Serganov, A.-H. H. M., and D. Patel, "Structural and dynamic basis for low-affinity, high-selectivity binding of L-Glutamine by the Glutamine riboswitch," *Cell Rep.*, vol. 13, pp. 1800–1813, 2015.

[51] C. Zwieb, R. W. V. Nues, M. A. Rosenblad, J. D. Brown, and T. Samuelsson, "A nomenclature for all signal recognition particle RNAs," *RNA*, vol. 11, no. 1, pp. 7–13, 2005.

[52] A. Kazantsev, A. A. Krivenko, D. J. Harrington, S. R. Holbrook, P. D. Adams, and N. R. Pace, "Crystal structure of a bacterial ribonuclease P RNA," *PNAS*, vol. 102, p. 13392–13397, 2005.

[53] N. Toor, K. S. Keating, S. D. Taylor, and A. M. Pyle, "Crystal Structure of a Self-Spliced Group II Intron," *Science*, vol. 320, no. 5872, pp. 77–82, 2008.

[54] C. Yan, J. Hang, R. Wan, M. Huang, C. C. Wong, and Y. Shi, "Structure of a yeast spliceosome at 3.6-angstrom resolution," *Science*, vol. 349, pp. 1182–1191, 2015.

[55] Q. Sun, X. Zhu, J. Qi, W. An, P. Lan, D. Tan, R. Chen, B. Wang, S. Zheng, C. Zhang, X. Chen, W. Zhang, J. Chen, M.-Q. Dong, and K. Ye, “Molecular architecture of the 90S small subunit pre-ribosome,” *eLife*, vol. 6, p. e22086, 2017.

[56] D. Ramrath, H. Yamamoto, K. Rother, D. Wittek, M. Pech, T. Mielke, J. Justus Loerke, P. Scheerer, P. Ivanov, Y. Teraoka, O. Shpanchenko, K. H. Nierhaus, and S. C. M. T., “The complex of tmRNA–SmpB and EF-G on translocating ribosomes,” *Nature*, vol. 485, pp. 526––529, 2012.

[57] A. J. Zaug and T. R. Cech, “Analysis of the structure of Tetrahymena nuclear RNAs in vivo: telomerase RNA, the self-splicing rRNA intron, and U2 snRNA,” *RNA*, vol. 1, pp. 363–374, 1995.

[58] G. D. Williams, R.-Y. Chang, and D. A. Brian, “A phylogenetically conserved Hairpin-Type 3’ untranslated region pseudoknot functions in coronavirus RNA replication,” *Journal of Virology*, vol. 73, pp. 8349—8355, 1999.

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

[60] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The Protein Data Bank,” *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

[61] T. J. Wheeler and S. R. Eddy, “nhmmer: DNA homology search with profile HMMs,” *Bioinformatics*, vol. 29, pp. 2487–2489, 2013.

[62] H. Yang, F. Jossinet, N. Leontis, L. Chen, J. Westbrook, H. M. Berman, and E. Westhof, “Tools for the automatic identification and classification of RNA base pairs,” *Nucl. Acids Res.*, vol. 31.13, pp. 3450–3460, 2003.

[63] E. P. Nawrocki and S. R. Eddy, “Infernal 1.1: 100-fold faster RNA homology searches,” *Bioinformatics*, vol. 29, pp. 2933–2935, 2013.

[64] 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.

[65] S. R. Eddy, “Accelerated profile HMM searches,” *PLOS Comp. Biol.*, vol. 7, p. e1002195, 2011.

[66] P. Auffinger and E. Westhof, “Singly and bifurcated hydrogen-bonded base-pairs in tRNA anticodon hairpins and ribozymes,” *J. Mol. Biol.*, vol. 292, pp. 467–483, 1999.

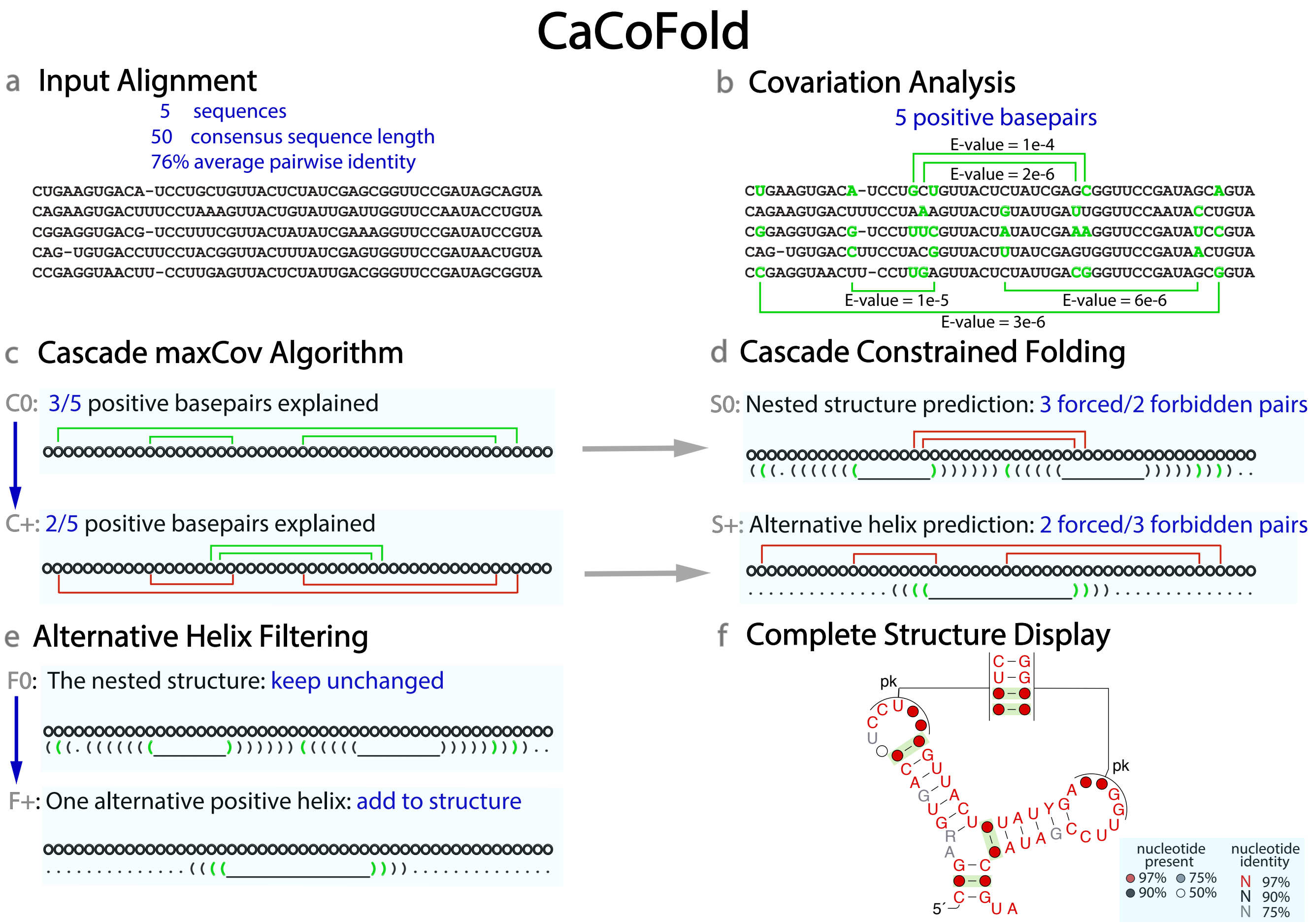
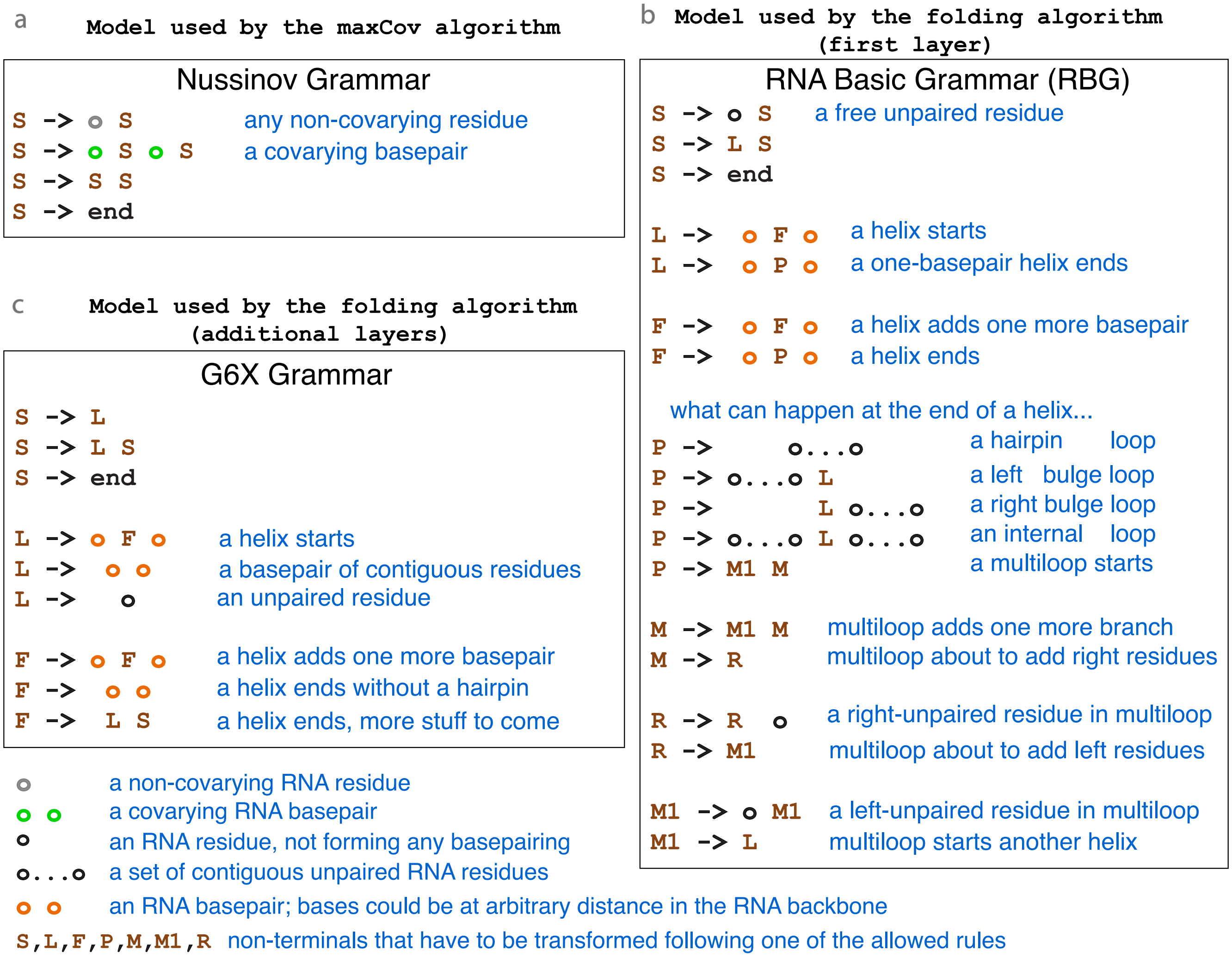

CaCoFold
a Input Alignment
5 sequences
50 consensus sequence length
76% average pairwise identity
CUGAAGUGACA-UCCUGCUGUUACUCUAUCGAGCGGUUCCGAUAGCAGUA
CAGAAGUGACUUUCCUAAAGUUACUGUAUUGAUUGGUUCCAAUACCUGUA
CGGAGGUGACG-UCCUUUCGUUACUAUAUCGAAAGGUUCCGAUAUCCGUA
CAG-UGUGACCUUCCUACGGUUACUUUAUCGAGUGGUUCCGAUAACUGUA
CCGAGGUAACUU-CCUUGAGUUACUCUAUUGACGGGUUCCGAUAGCGGUA
b Covariation Analysis
5 positive basepairs
E-value = 1e-4
E-value = 2e-6
E-value = 1e-5
E-value = 6e-6
E-value = 3e-6
c Cascade maxCov Algorithm
C0: 3/5 positive basepairs explained
C+: 2/5 positive basepairs explained
d Cascade Constrained Folding
S0: Nested structure prediction: 3 forced/2 forbidden pairs
S+: Alternative helix prediction: 2 forced/3 forbidden pairs
e Alternative Helix Filtering
F0: The nested structure: keep unchanged
F+: One alternative positive helix: add to structure
f Complete Structure Display
pk
pk
5´
nucleotide present
97%
75%
90%
50%
nucleotide identity
N 97%
N 90%
N 75%

Figure 1

Figure 1: **The CaCoFold algorithm. (a)** Toy alignment of five sequences. **(b)** The statistical analysis identifies five significantly covarying position pairs in the alignment (E-value < 0.05). Column pairs that significantly covary are marked with green arches, compensatory pairwise substitutions including G-U pairs (green) relative to consensus (black). **(c)** The maxCov algorithm requires two layers to explain all five covariations. In the first (C0) layer, three positive basepairs depicted in green are grouped together. In successive layers (C+), positive basepairs already taken into account (depicted in red) are excluded. **(d)** At each layer, a dynamic programming algorithm produces the most probable fold constrained by the assigned positive basepairs (green parentheses), to the exclusion of all negative basepairs and other positive basepairs (red arches). (This toy alignment does not include any negative basepairs.) Residues forming a red arch can pair to other bases. Basepairs that do not significantly covary are depicted by black parentheses. **(e)** The S+ alternative structures without positive basepairs that overlap in more that half of their residues with the S0 structure are removed. Alternative helices with positive basepairs are always kept. **(f)** The final consensus structure combining the nested S0 structure with the alternative filtered helices from all other layers is displayed automatically using a modified version of the program R2R. Positive basepairs are depicted in green.



a **Model used by the maxCov algorithm**

Nussinov Grammar

```
S -> o S           any non-covarying residue
S -> o S o S       a covarying basepair
S -> S S
S -> end
```

b **Model used by the folding algorithm (first layer)**

RNA Basic Grammar (RBG)

```
S -> o S           a free unpaired residue
S -> L S
S -> end

L ->  o F o        a helix starts
L ->  o P o        a one-basepair helix ends

F ->  o F o        a helix adds one more basepair
F ->  o P o        a helix ends

 what can happen at the end of a helix...
P ->     o...o        a hairpin     loop
P -> o...o L          a left   bulge loop
P ->       L o...o    a right bulge loop
P -> o...o L o...o    an internal   loop
P -> M1 M             a multiloop starts

M -> M1 M       multiloop adds one more branch
M -> R          multiloop about to add right residues

R -> R   o      a right-unpaired residue in multiloop
R -> M1         multiloop about to add left residues

M1 -> o M1      a left-unpaired residue in multiloop
M1 -> L         multiloop starts another helix
```

c **Model used by the folding algorithm (additional layers)**

G6X Grammar

```
S -> L
S -> L S
S -> end

L -> o F o      a helix starts
L ->  o o       a basepair of contiguous residues
L ->   o        an unpaired residue

F -> o F o      a helix adds one more basepair
F ->  o o       a helix ends without a hairpin
F ->  L S       a helix ends, more stuff to come
```

```
o        a non-covarying RNA residue
o o      a covarying RNA basepair
o        an RNA residue, not forming any basepairing
o...o    a set of contiguous unpaired RNA residues
o o      an RNA basepair; bases could be at arbitrary distance in the RNA backbone
S,L,F,P,M,M1,R  non-terminals that have to be transformed following one of the allowed rules
```

Figure 2

Figure 2: **RNA models used by the CaCoFold algorithm.** (**a**) The Nussinov grammar implemented by the maxCov algorithm uses the R-scape E-values of the significantly covarying pairs, and maximizes the sum of -log(E-value). (**b**) The RGB model used by the first layer of the folding algorithm. (**c**) The G6X model used by the rest of the layers completing the non-nested part of the RNA structure. 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.

tranfer-messenger RNA (tmRNA)

a Input Alignment

Rfam RF00023 seed alignment

477 sequences
354 consensus sequence length
357 average sequence length
42% average pairwise identity

b Covariation Analysis

All possible pairs analyzed equally

119 annotated basepairs in alignment
(not used in analysis)

414 columns analyzed:
121 positive basepairs (significantly covary)
109 positive basepairs expected by power
31,027 negative basepairs

c Cascade maxCov Algorithm

121 positive basepairs explained in 6 layers

layer 1: 69	layer 2: 41
layer 3: 5	layer 4: 3
layer 5: 2	layer 6: 1

d Cascade Constrained Folding

139	annotated pairwise interactions
121/139	positive basepairs
74	pairs not in final ss due to forbidden negative basepairs

e Alternative Helix Filtering

18 alternative helices →
5 pseudoknots
3 triplets
10 mRNA-induced covariations

f Complete structure display

g Structure comparison

Kelley *et al.*, RNA 2001, Fig 4

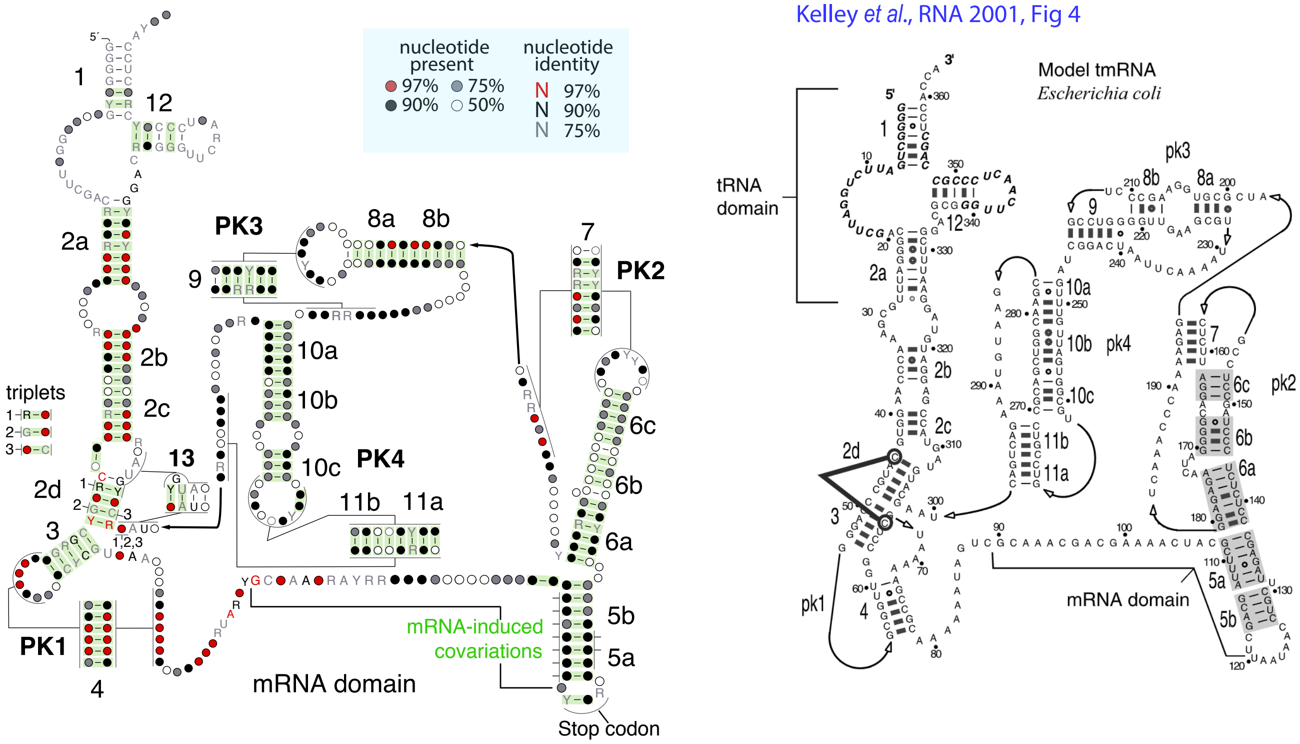
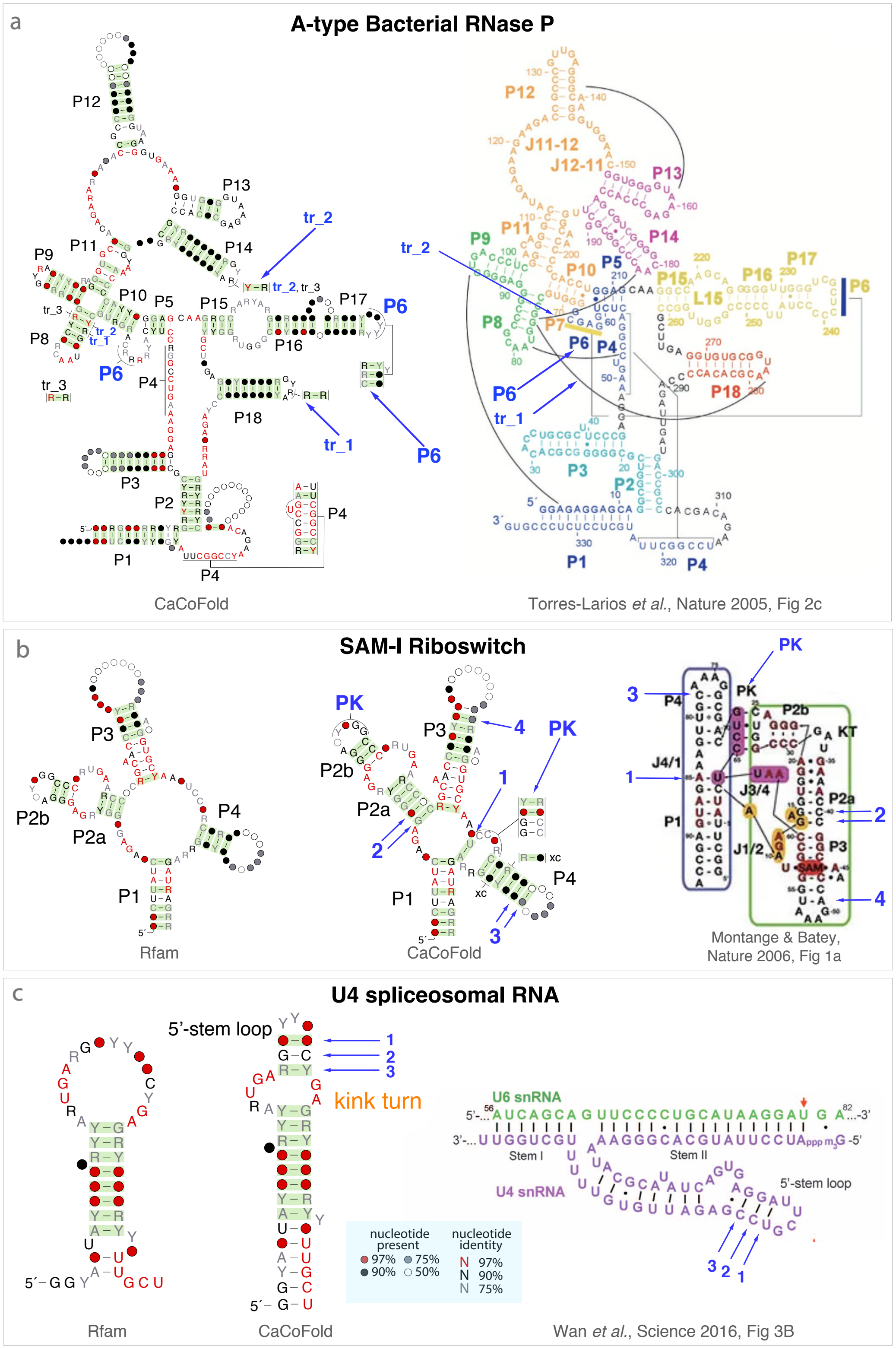


Figure 3

Figure 3: **The CaCoFold algorithm applied to the transfer-messenger RNA (tmRNA).** Steps (a) to (f) refer to the same methods as described in Fig. 1. Step (b) performs a statistical test that considers all possible pairs equally resulting in the assignment of 121 significantly covarying positive basepairs. The Rfam consensus structure in not used in the analysis. The whole analysis is performed using the single command `R-scape --fold` on the input alignment. The analysis takes 25 seconds (30s including drawing all the figures) on a 3.3 GHz Intel Core i7 MacBook Pro. The structural display in (f) has been modified by hand to match the standard depiction of the tmRNA secondary structure in (g). The thick line in (g) indicates the C-C triplet interaction proposed in Ref. 32. Details of the mRNA-induced covariations are given in Fig. S6c.


a A-type Bacterial RNase P

CaCoFold

Torres-Larios *et al*., Nature 2005, Fig 2c

b SAM-I Riboswitch

Rfam

CaCoFold

Montange & Batey, Nature 2006, Fig 1a

c U4 spliceosomal RNA

Rfam

CaCoFold

Wan *et al*., Science 2016, Fig 3B

Figure 4

Figure 4: **CaCoFold structures confirmed by known 3D structures (part 1/7).** Structural elements with covariation support introduced by CaCoFold relative to the Rfam annotation and corroborated by 3D structures are annotated in blue. **(a)** The A-type RNase P RNA CaCoFold structure includes one more helix (P6) and two long range interactions (tr_1 and tr_2) with covariation support relative to the Rfam structure (not shown). The blue arrows show their correspondence to the crystal structure[35]. The display of the CaCoFold structure has been modified by hand to match the standard depiction of the structure. **(b)** The SAM-I riboswitch CaCoFold structure shows relative to the Rfam structure one more helix forming a pseudoknot, and a A-U pair stacking on helix P1 both confirmed by the SAM-I riboswitch 2.9 Å resolution crystal structure of *T. tengcongensis*[36]. CaCoFold also identifies additional pairs with covariation support for helices P2a, P3 and P4. **(c)** The U4 snRNA CaCoFold structure identifies one more internal loop and one more helix than the Rfam structure confirmed by the 3D structure[37]. The new U4 internal loop flanked by covarying Watson-Crick basepairs includes a kink turn (UAG-AG). The non Watson-Crick pairs in a kink turn (A-G, G-A) are generally conserved (>97% in this alignment) and do not covary.

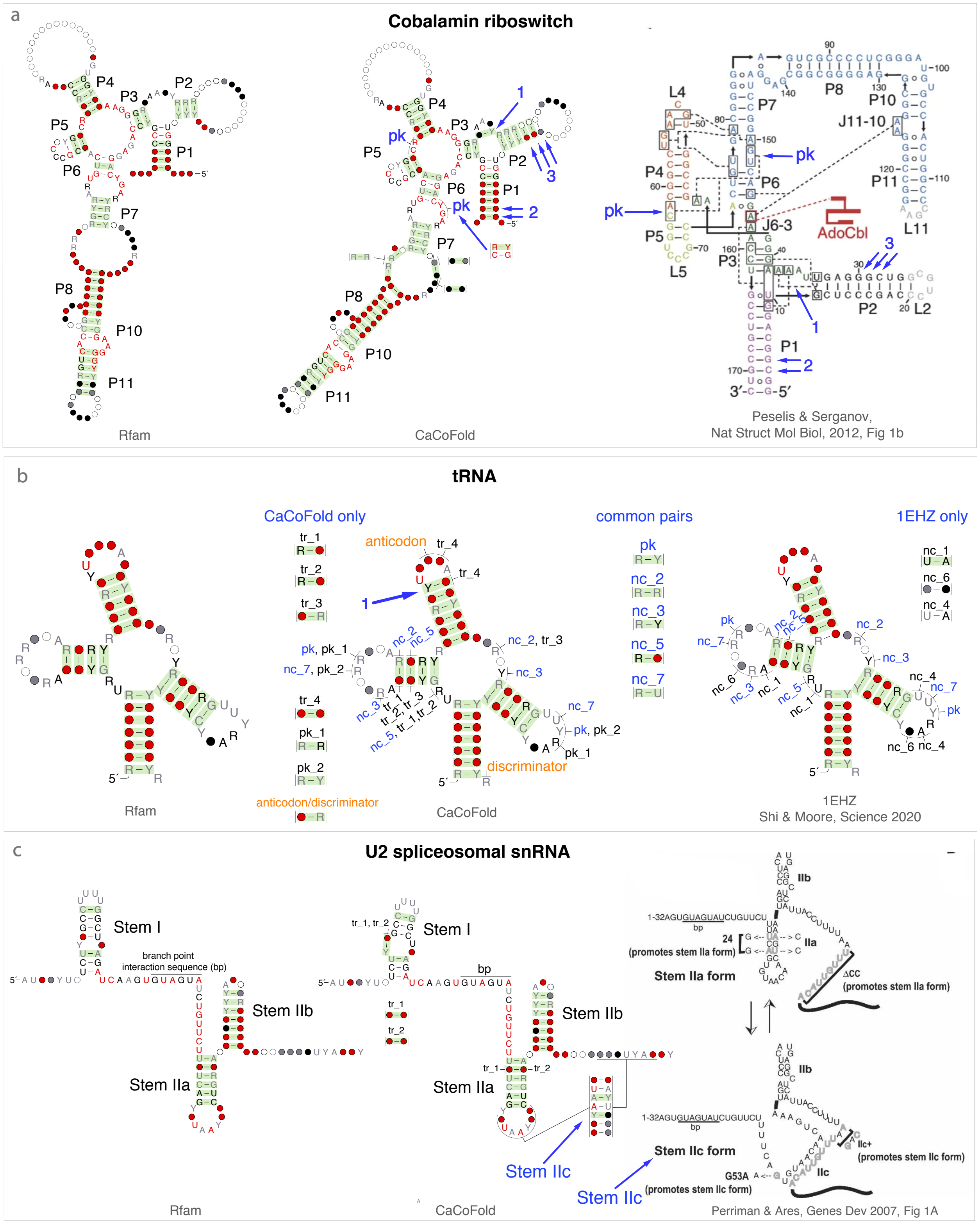

a
Cobalamin riboswitch
P1
P2
P3
P4
P5
P6
P7
P8
P10
P11
pk
1
2
3
Rfam
CaCoFold
L4
L5
L11
L2
J11-10
J6-3
AdoCbl
Peselis & Serganov,
Nat Struct Mol Biol, 2012, Fig 1b
b
tRNA
CaCoFold only
common pairs
1EHZ only
anticodon
discriminator
anticodon/discriminator
tr_1
tr_2
tr_3
tr_4
pk_1
pk_2
nc_1
nc_2
nc_3
nc_4
nc_5
nc_6
nc_7
Rfam
CaCoFold
1EHZ
Shi & Moore, Science 2020
c
U2 spliceosomal snRNA
Stem I
Stem IIa
Stem IIb
Stem IIc
branch point
interaction sequence (bp)
bp
Stem IIa form
Stem IIc form
(promotes stem IIa form)
(promotes stem IIc form)
ΔCC
IIc+
G53A
Rfam
CaCoFold
Perriman & Ares, Genes Dev 2007, Fig 1A

Figure 5

Figure 5: **CaCoFold structures confirmed by known 3D structures (part 2/7).** Structural elements with covariation support introduced by CaCoFold relative to the Rfam annotation and corroborated by 3D structures are annotated in blue. **(a)** Relative to the Rfam structure, the Cobalamin riboswitch CaCoFold structure adds one pseudoknot and one Watson-Crick basepair defining a four-way junction between helices P1, P2, and P3, both confirmed by the *S. thermophilum* crystal structure[38]. It also adds more covariation support for helices P1 and P2. **(b)** In CaCoFold structures, alternative helices that do not overlap with the nested structure are annotated as pseudoknots (pk), otherwise they are annotated as triplets (tr). For structures obtained from a crystal structure, non Watson-Crick basepairs are annotated as non-canonical (nc) regardless of whether they are overlapping or not with the nested structure. The tRNA CaCoFold structure has been re-annotated manually to match the labeling of the *S. cerevisiae* phenylalanine tRNA 1EHZ crystal structure (1.93 Å) for all common basepairs[40]. Four nc pairs and one pk pair with covariation support are found by CaCoFold and confirmed by the 1EHZ structure. Four base triplets (tr) and two pseudoknots (pk) have covariation support but have not been assigned to any basepair type by RNAView. The additional positive basepair (marked “1”) in the anticodon hairpin is a non-canonical basepairs that has also been confirmed[66]. **(c)** In the U2 spliceosomal RNA, both Stem IIa and Stem IIc have covariation support and compete to promote different splicing steps[42].

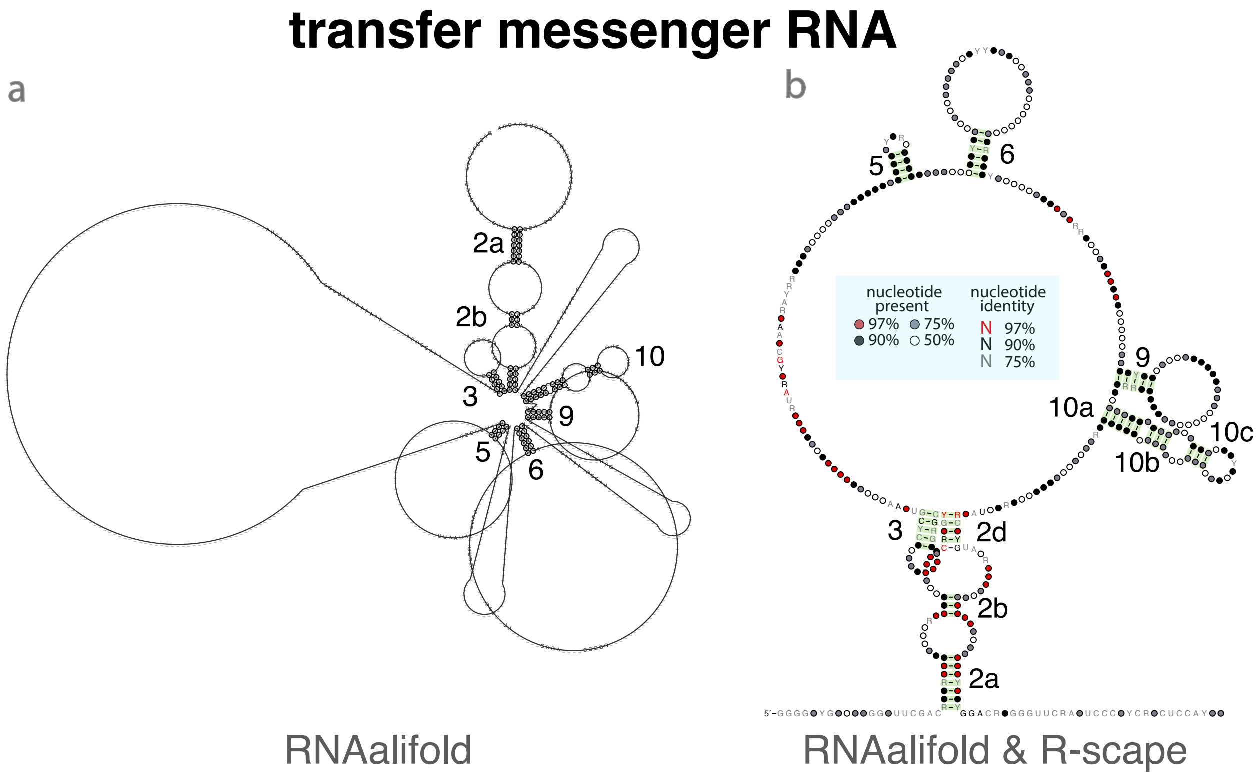


Figure S1. **tmRNA structure predicted by RNAalifold and covariation analysis.** **(a)** The RNAalifold predicted consensus structure output for the tmRNA Rfam seed alignment (RF00023) obtained using default parameters. The RNAalifold structure consists of 46 basepairs, and it annotates (at least partially) 6 of the 12 helices in the structure[32]: 2 (a,b,d), 3, 5, 6, 9, and 10 (a,b,c), see Fig. 3g. **(b)** The covariation analysis of the RNAalifold structure indicates that 45 of the 46 RNAalifold basepairs have covariation support (shown in green). It also identifies 76 other basepairs with covariation support not in the proposed RNAalifold structure (not shown in figure). The display of all 121 positive pairs can be seen in Fig. 3f. (Columns with more than 75% gaps have been removed from the display.)

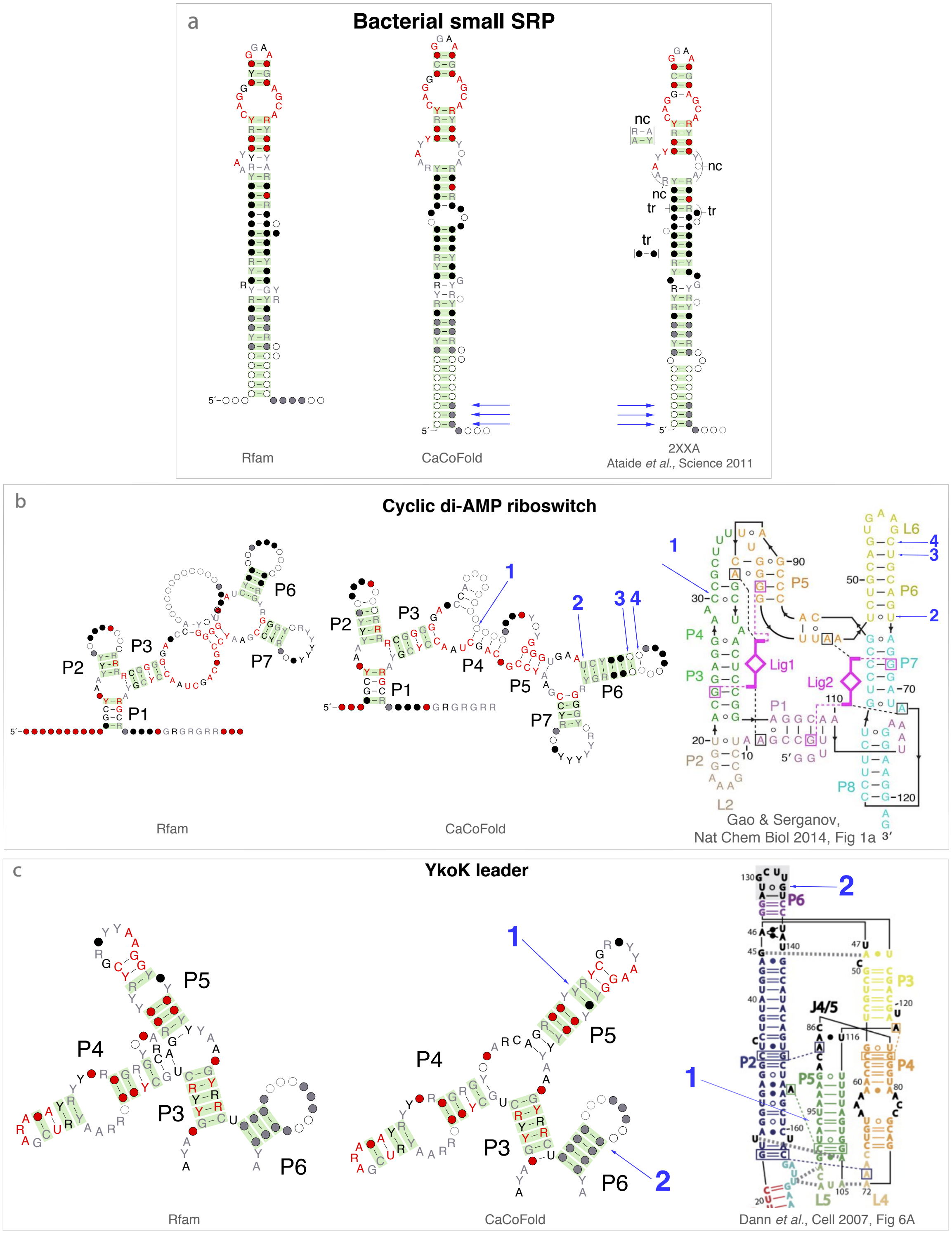

a
Bacterial small SRP
nc
tr
Rfam
CaCoFold
2XXA
Ataide et al., Science 2011
b
Cyclic di-AMP riboswitch
P1
P2
P3
P4
P5
P6
P7
P8
L2
L6
Lig1
Lig2
1
2
3
4
Rfam
CaCoFold
Gao & Serganov,
Nat Chem Biol 2014, Fig 1a
c
YkoK leader
P2
P3
P4
P5
P6
J4/5
L4
L5
1
2
Rfam
CaCoFold
Dann et al., Cell 2007, Fig 6A

Figure S2.

Figure S2. **CaCoFold structures confirmed by known 3D structures (part 3/7).** Structural elements with covariation support introduced by CaCoFold relative to the Rfam annotation and corroborated by 3D structures are annotated in blue. All three cases are examples of CaCoFold structures with more covariation support in the form of more positive basepairs to helices already present in the consensus Rfam structures. **(a)** The SRP complex 2XXA PDB X-ray diffraction structure has 3.94 Å resolution[43]. The PDB-derived consensus structure was obtained as described in Methods. **(b)** For the cyclic di-AMP riboswitch, the region around helix P4 is highly variable in the Rfam alignment, and none of the proposed structures has covariation support. The displayed CaCoFold structure showing helix P4 was obtained using a consensus reference sequence (instead of the default profile sequence). The rest of the structure has covariation support and remains invariant.

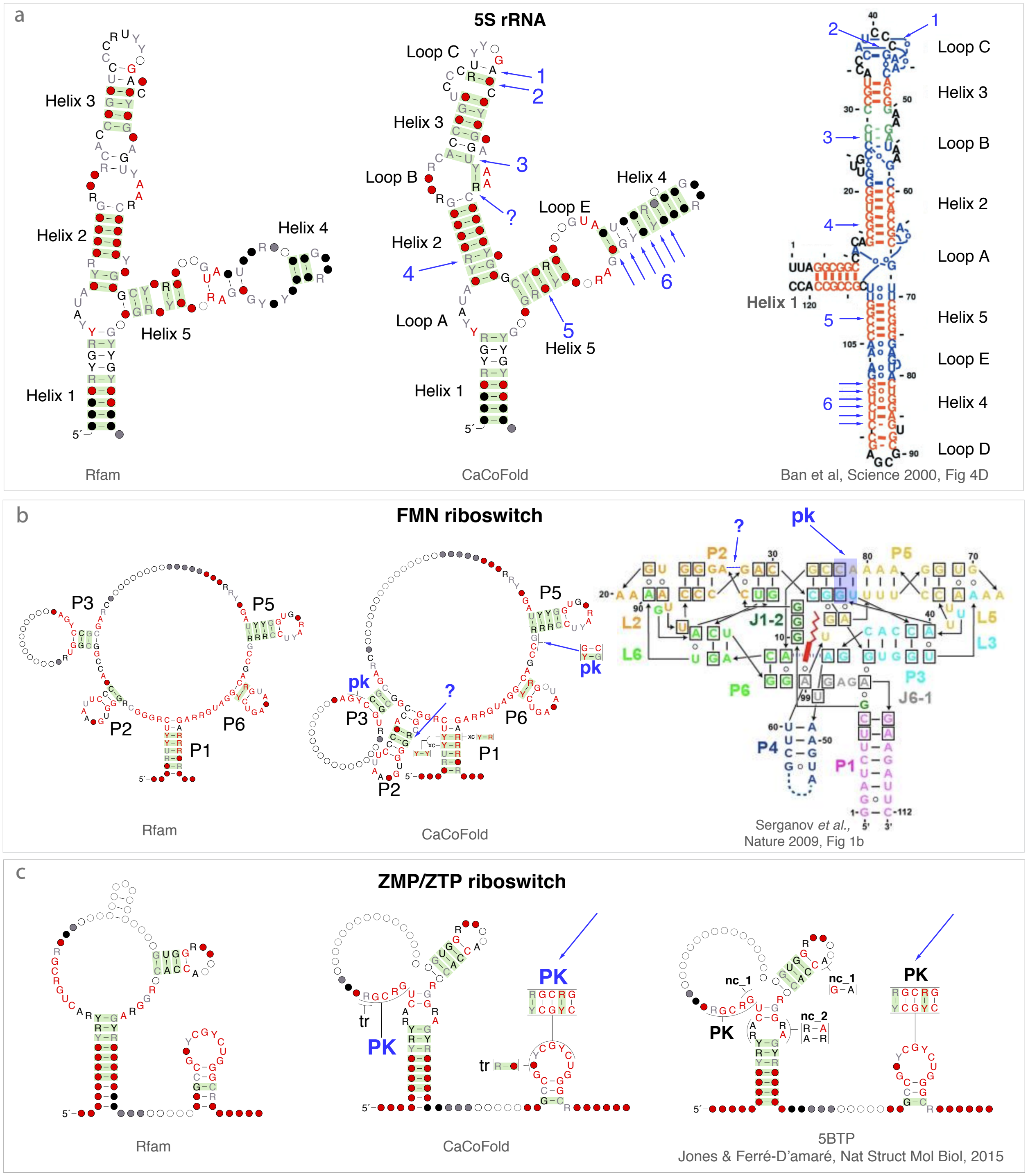

a
5S rRNA
Helix 1
Helix 2
Helix 3
Helix 4
Helix 5
Loop A
Loop B
Loop C
Loop D
Loop E
1
2
3
4
5
6
?
Rfam
CaCoFold
Ban et al, Science 2000, Fig 4D
b
FMN riboswitch
P1
P2
P3
P4
P5
P6
L2
L3
L5
L6
J1-2
J6-1
pk
?
Rfam
CaCoFold
Serganov et al.,
Nature 2009, Fig 1b
c
ZMP/ZTP riboswitch
PK
tr
nc_1
nc_2
Rfam
CaCoFold
5BTP
Jones & Ferré-D'amaré, Nat Struct Mol Biol, 2015

Figure S3.

Figure S3. **CaCoFold structures confirmed by known 3D structures (part 4/7).** Structural elements with covariation support introduced by CaCoFold relative to the Rfam annotation and corroborated by 3D structures are annotated in blue. **(a)** The 5S rRNA CaCoFold structure remodels Helix 4 (six basepairs) and Loop C (two basepairs) in agreement with the crystal structure[46]. A Y-R covarying basepair in Loop B is not described in the 3D structure. **(b)** The FMN riboswitch CaCoFold structure identifies a confirmed 2-basepair pseudoknotted helix, and one covarying pair in helix P2 that is different than in the 3D structure[47]. **(c)** The covarying pseudoknot identified by CaCoFold in the ZPM-ZTP riboswitch is confirmed by the *Fusobacterium ulcerans* X-ray diffraction structure (2.82 Å)[48].

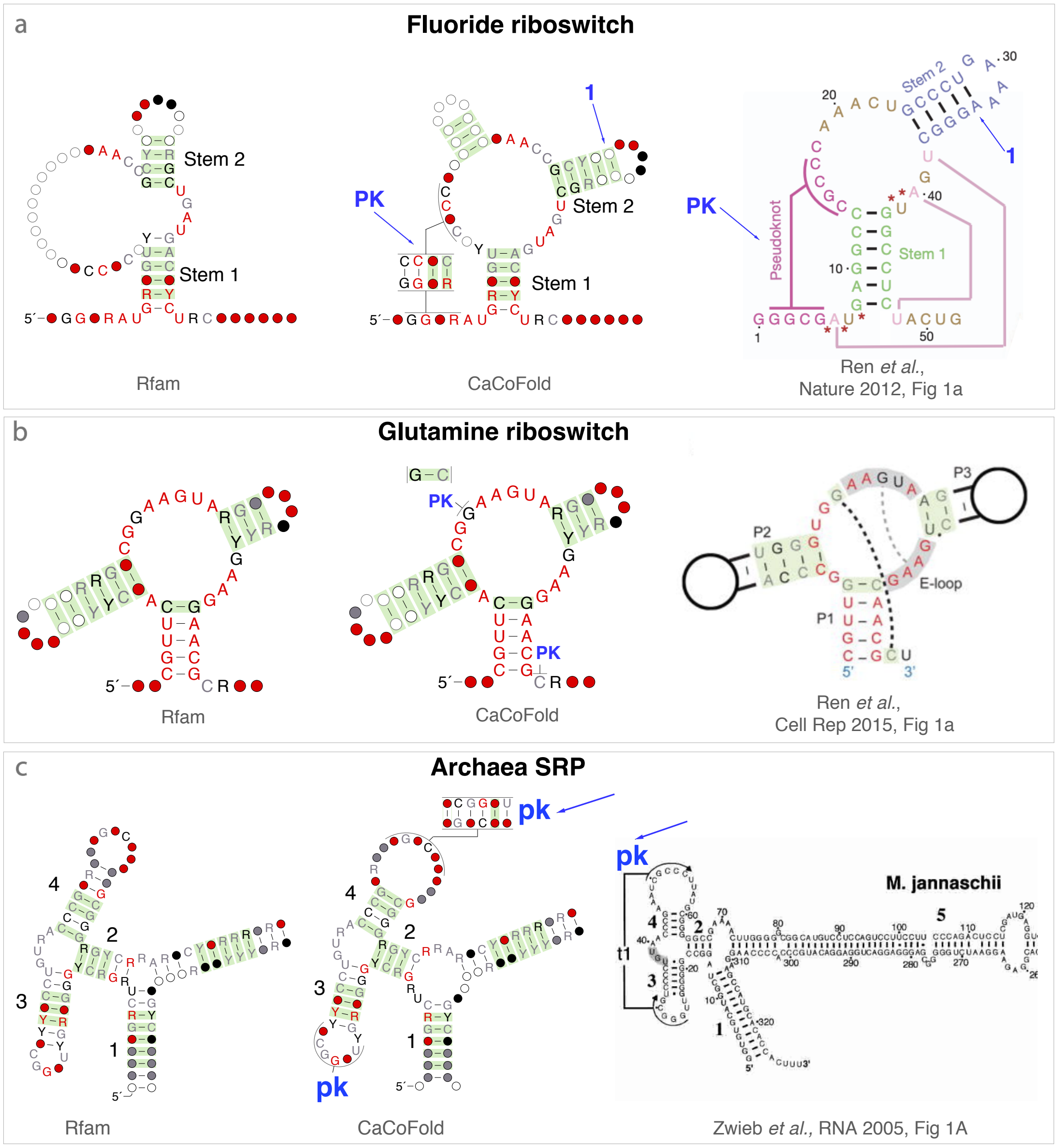


Figure S4. **CaCoFold structures confirmed by known 3D structures (part 5/7).** Structural elements with covariation support introduced by CaCoFold relative to the Rfam annotation and corroborated by 3D structures are annotated in blue. All three cases are examples of CaCoFold structures with more covariation support in the form of a new helix forming a pseudoknot all confirmed by the 3D structures.

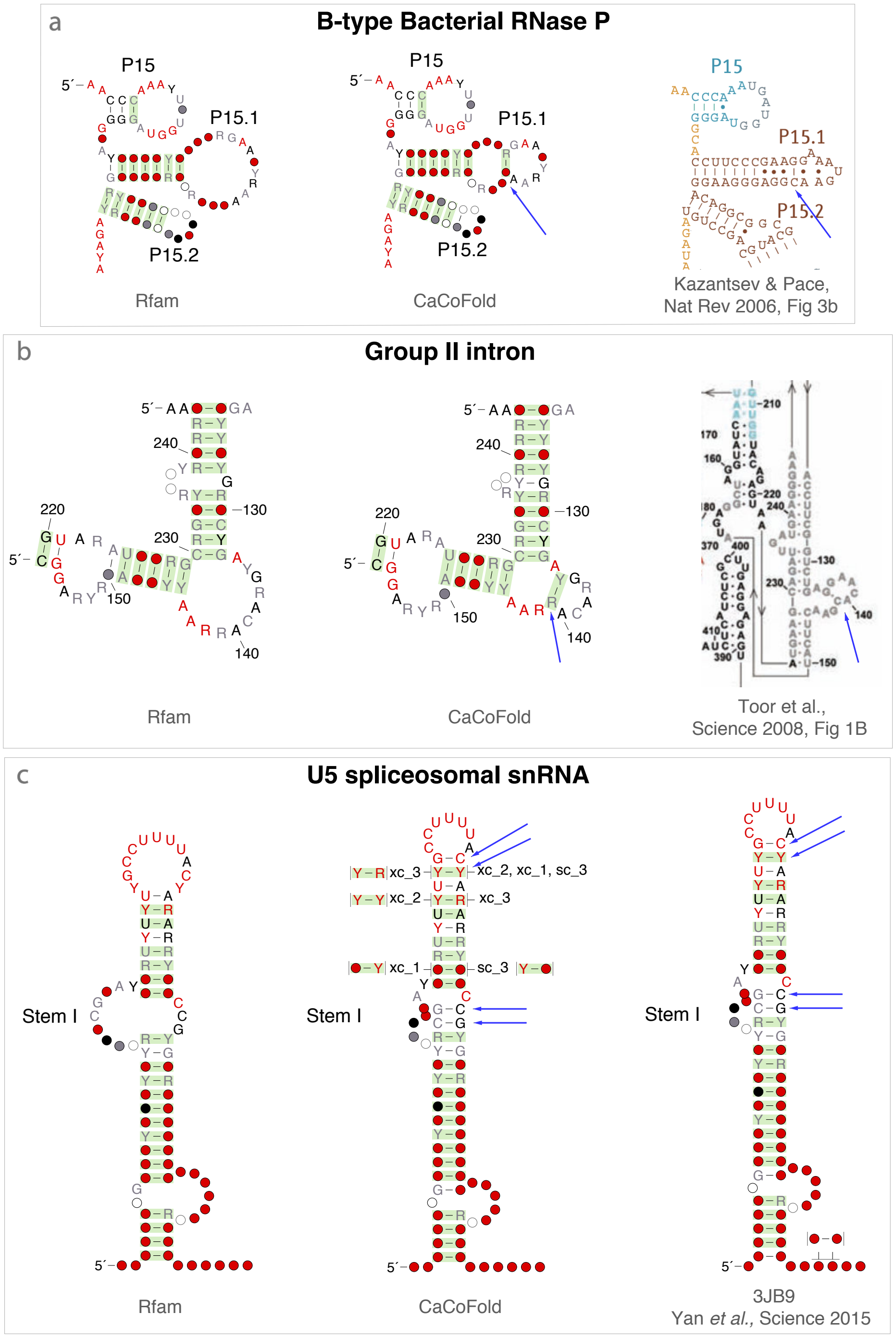


Figure S5. **CaCoFold structures confirmed by known 3D structures (part 6/7).** Structural elements with covariation support introduced by CaCoFold relative to the Rfam annotation and corroborated by 3D structures are annotated in blue. **(a)** An additional covarying pair introduces a new internal loop in the B-type RNase P RNA confirmed by Ref. 52, Fig. 3b. **(b)** An additional covarying pair introduces a new three-way junction an the group-II intron[53]. **(c)** In the U5 snRNA, an additional Y-Y covarying pair that modifies a hairpin loop is confirmed by the *S. pombe* spliceosomal RNA cryo-EM structure 3JB9 (3.60 Å)[54].

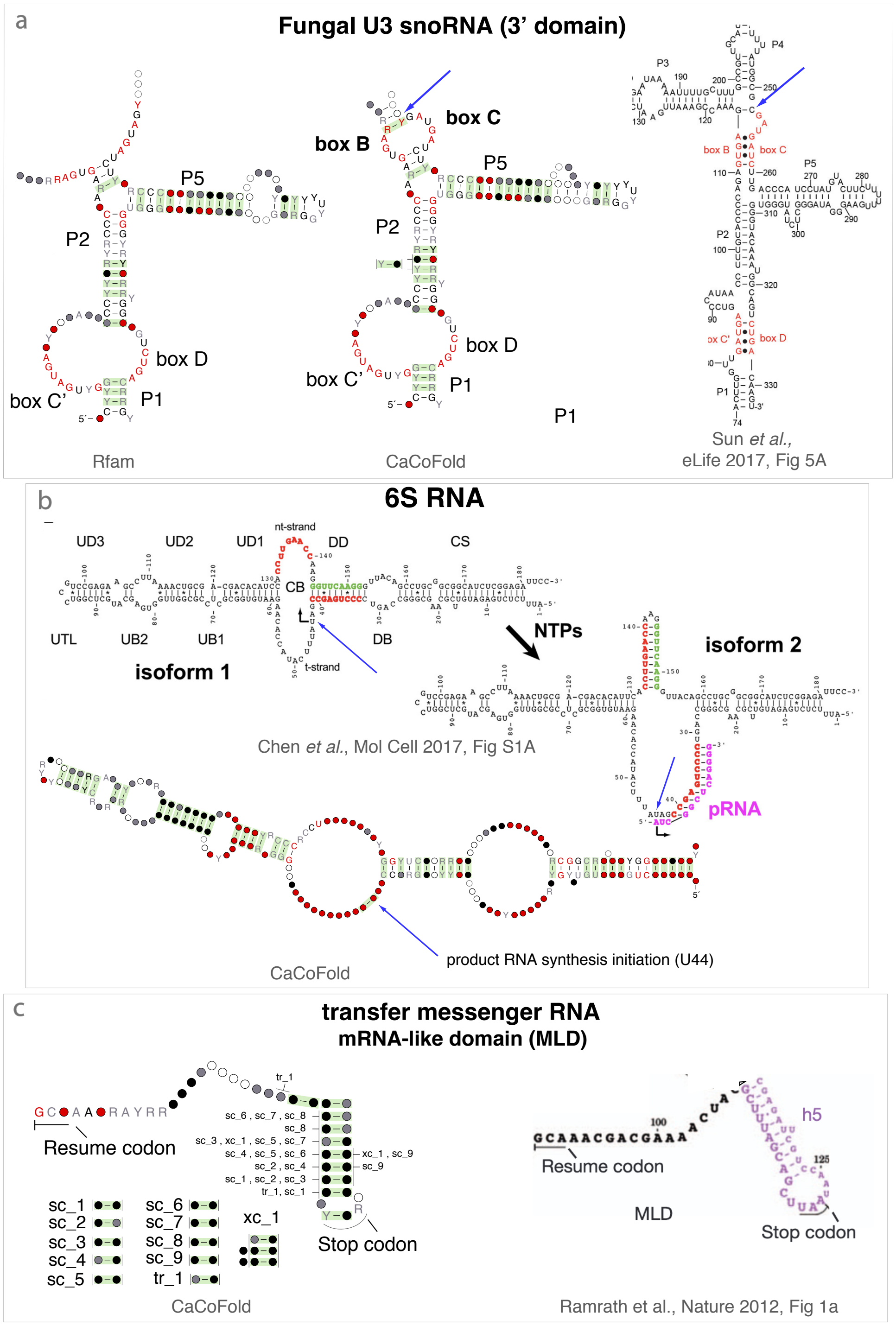


Figure S6. **CaCoFold structures confirmed by known 3D structures (part 7/7).** Structural elements with covariation support introduced by CaCoFold relative to the Rfam annotation and corroborated by 3D structures are annotated in blue. **(a)** The U3 snoRNA CaCoFold structure adds a covarying pair closing the boxB/boxC of the snoRNA[54]. **(b)** 6S RNA covarying pair at the RNA synthesis initiation site not associated to RNA structure[4]. **(c)** Side-covariation in the mRNA-like domain of tmRNA not due to RNA structure.

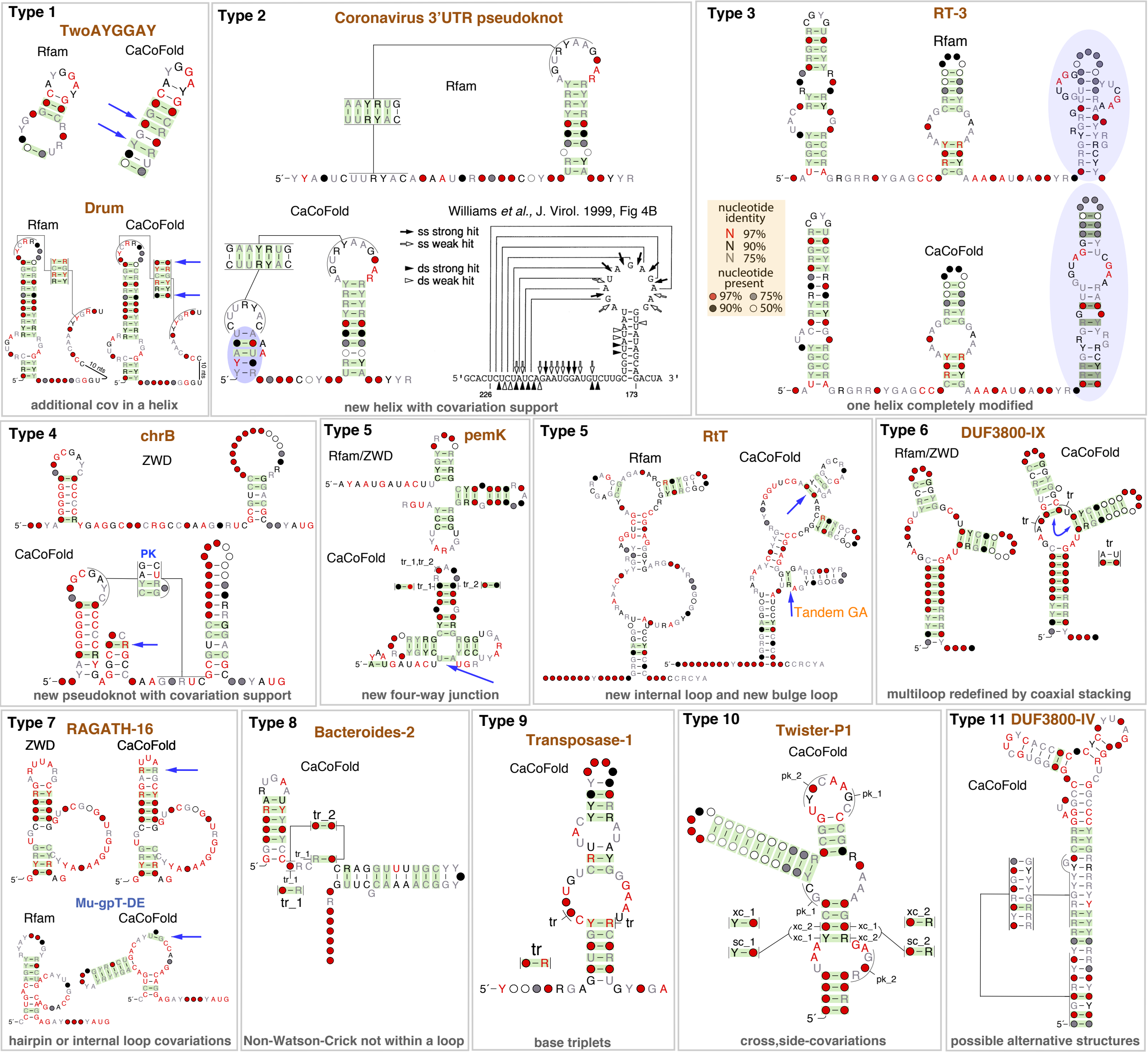
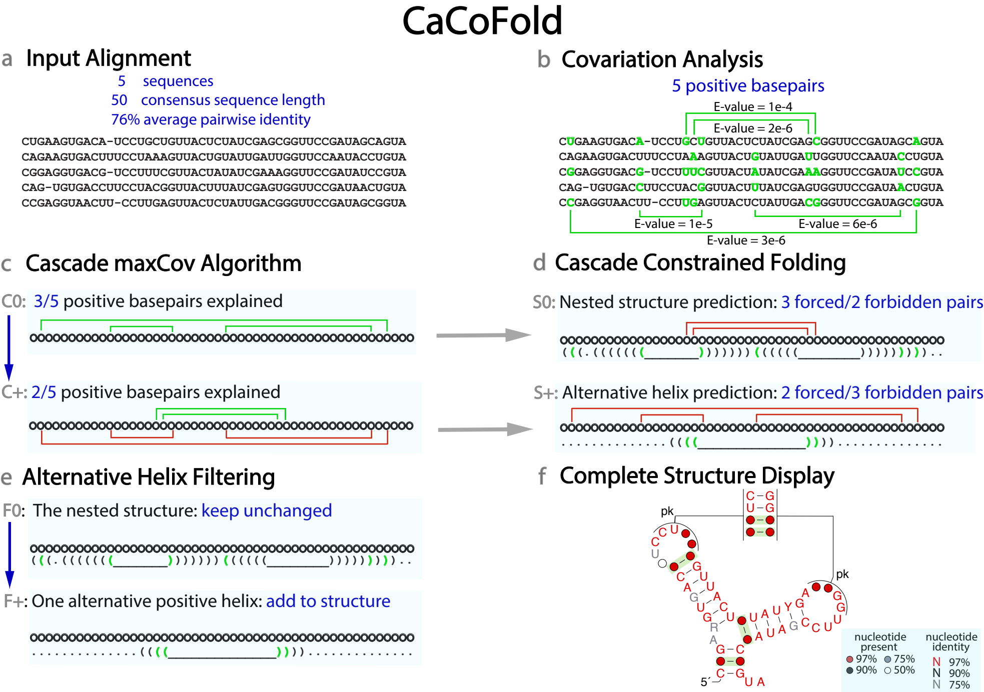


Figure S7. **Examples of RNAs without a 3D structure for which the CaCoFold structure has more positive basepairs (green shading) than the structure given by the corresponding database.** We provide examples of differences corresponding to Types 1 to 11. A description of all different types is given in Table 1.


CaCoFold
a Input Alignment
5 sequences
50 consensus sequence length
76% average pairwise identity
CUGAAGUGACA-UCCUGCUGUUACUCUAUCGAGCGGUUCCGAUAGCAGUA
CAGAAGUGACUUUCCUAAAGUUACUGUAUUGAUUGGUUCCAAUACCUGUA
CGGAGGUGACG-UCCUUUCGUUACUAUAUCGAAAGGUUCCGAUAUCCGUA
CAG-UGUGACCUUCCUACGGUUACUUUAUCGAGUGGUUCCGAUAACUGUA
CCGAGGUAACUU-CCUUGAGUUACUCUAUUGACGGGUUCCGAUAGCGGUA
b Covariation Analysis
5 positive basepairs
E-value = 1e-4
E-value = 2e-6
CUGAAGUGACA-UCCUGCUGUUACUCUAUCGAGCGGUUCCGAUAGCAGUA
CAGAAGUGACUUUCCUAAAGUUACUGUAUUGAUUGGUUCCAAUACCUGUA
CGGAGGUGACG-UCCUUUCGUUACUAUAUCGAAAGGUUCCGAUAUCCGUA
CAG-UGUGACCUUCCUACGGUUACUUUAUCGAGUGGUUCCGAUAACUGUA
CCGAGGUAACUU-CCUUGAGUUACUCUAUUGACGGGUUCCGAUAGCGGUA
E-value = 1e-5
E-value = 6e-6
E-value = 3e-6
c Cascade maxCov Algorithm
C0: 3/5 positive basepairs explained
C+: 2/5 positive basepairs explained
d Cascade Constrained Folding
S0: Nested structure prediction: 3 forced/2 forbidden pairs
(((.(((((((________)))))))(((((((________)))))))))..
S+: Alternative helix prediction: 2 forced/3 forbidden pairs
..............((((________________))))..............
e Alternative Helix Filtering
F0: The nested structure: keep unchanged
(((.(((((((________)))))))(((((((________)))))))))..
F+: One alternative positive helix: add to structure
..............((((________________))))..............
f Complete Structure Display
pk
C – G
U – G
pk
C C U
U
G C U A U G A U C G U G R U A
A G – C C – G U A
U A U Y G A G G U U C G A U A C C
5´
nucleotide present
97% 75%
90% 50%
nucleotide identity
N 97%
N 90%
N 75%

a **Model used by the maxCov algorithm**

Nussinov Grammar

```
S -> o S          any non-covarying residue
S -> o S o S      a covarying basepair
S -> S S
S -> end
```

b **Model used by the folding algorithm (first layer)**

RNA Basic Grammar (RBG)

```
S -> o S        a free unpaired residue
S -> L S
S -> end

L ->  o F o     a helix starts
L ->  o P o     a one-basepair helix ends

F ->  o F o     a helix adds one more basepair
F ->  o P o     a helix ends

  what can happen at the end of a helix...
P ->     o...o              a hairpin      loop
P -> o...o L                a left   bulge loop
P ->       L o...o          a right bulge loop
P -> o...o L o...o          an internal    loop
P -> M1 M                   a multiloop starts

M -> M1 M     multiloop adds one more branch
M -> R        multiloop about to add right residues

R -> R  o     a right-unpaired residue in multiloop
R -> M1       multiloop about to add left residues

M1 -> o M1    a left-unpaired residue in multiloop
M1 -> L       multiloop starts another helix
```

c **Model used by the folding algorithm (additional layers)**

G6X Grammar

```
S -> L
S -> L S
S -> end

L -> o F o      a helix starts
L ->  o o       a basepair of contiguous residues
L ->   o        an unpaired residue

F -> o F o      a helix adds one more basepair
F ->  o o       a helix ends without a hairpin
F ->  L S       a helix ends, more stuff to come
```

```
o              a non-covarying RNA residue
o  o           a covarying RNA basepair
o              an RNA residue, not forming any basepairing
o...o          a set of contiguous unpaired RNA residues
o  o           an RNA basepair; bases could be at arbitrary distance in the RNA backbone
S,L,F,P,M,M1,R non-terminals that have to be transformed following one of the allowed rules
```

tranfer-messenger RNA (tmRNA)

a Input Alignment

Rfam RF00023 seed alignment

477 sequences
354 consensus sequence length
357 average sequence length
42% average pairwise identity

b Covariation Analysis

All possible pairs analyzed equally

119 annotated basepairs in alignment
(not used in analysis)

414 columns analyzed:
121 positive basepairs (significantly covary)
109 positive basepairs expected by power
31,027 negative basepairs

c Cascade maxCov Algorithm

121 positive basepairs explained in 6 layers

layer 1: 69	layer 2: 41
layer 3: 5	layer 4: 3
layer 5: 2	layer 6: 1

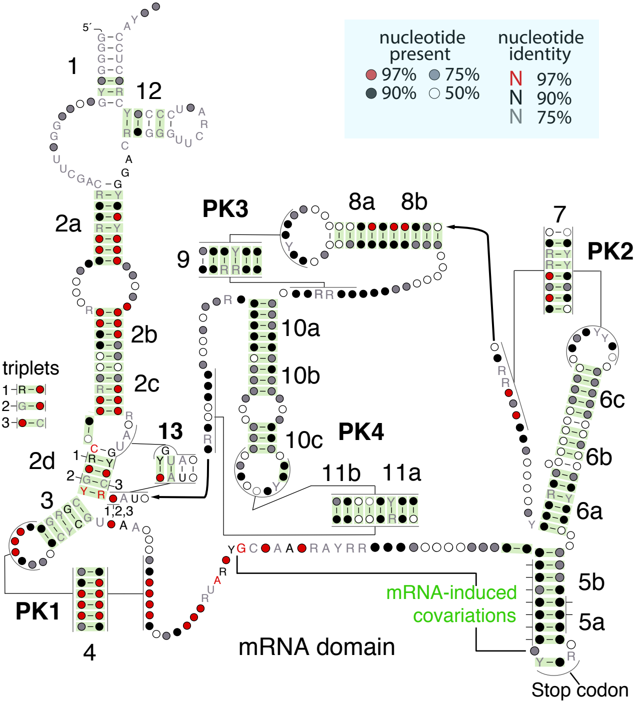
d Cascade Constrained Folding

139	annotated pairwise interactions
121/139	positive basepairs
74	pairs not in final ss due to forbidden negative basepairs

e Alternative Helix Filtering

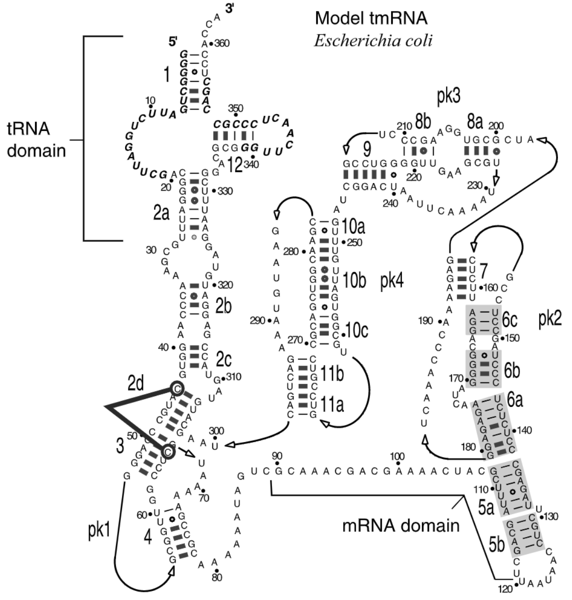
18 alternative helices →
- 5 pseudoknots
- 3 triplets
- 10 mRNA-induced covariations

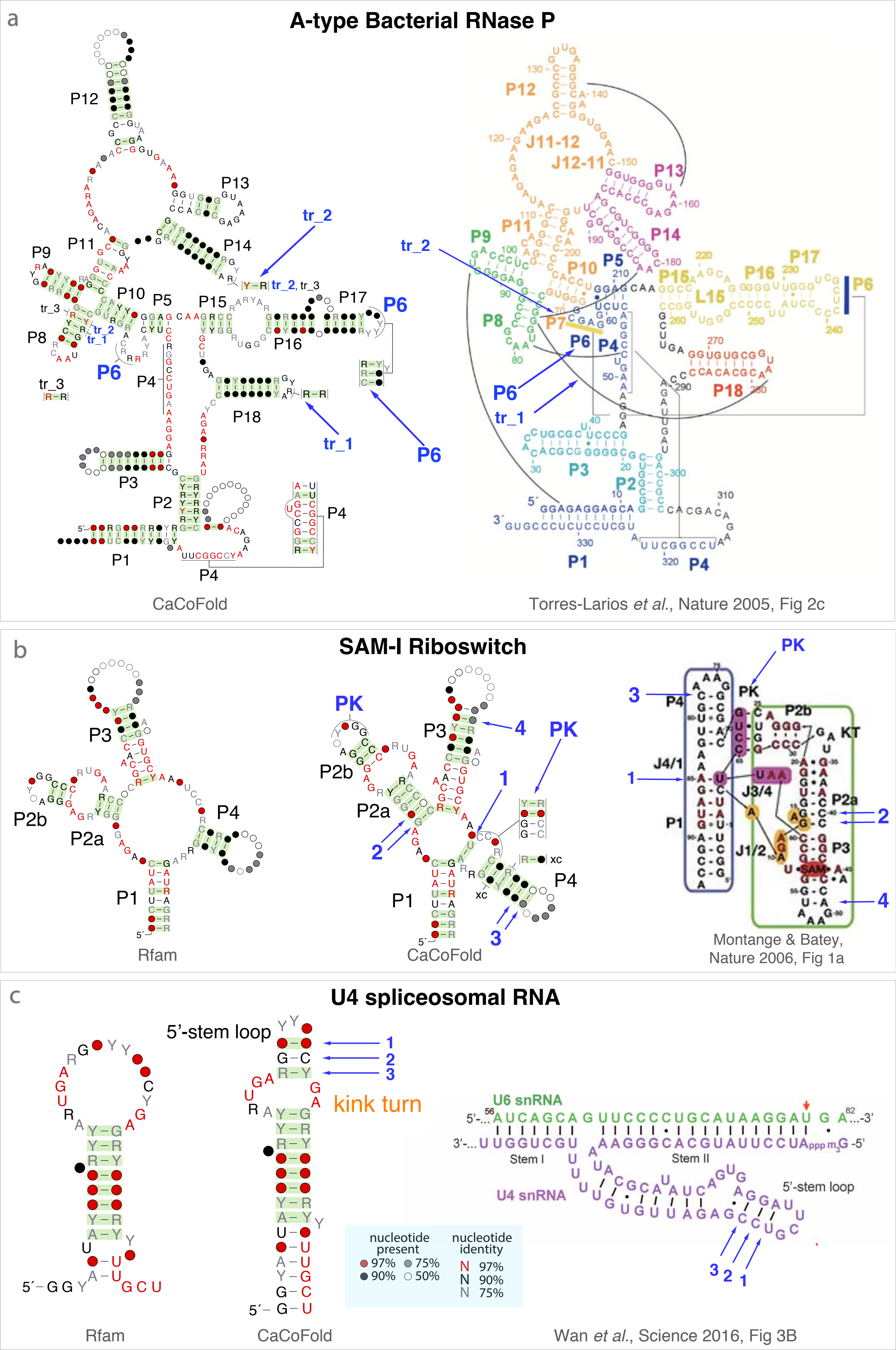
f Complete structure display



g Structure comparison

Kelley *et al*., RNA 2001, Fig 4





a **A-type Bacterial RNase P**

CaCoFold

Torres-Larios *et al.*, Nature 2005, Fig 2c

b **SAM-I Riboswitch**

Rfam

CaCoFold

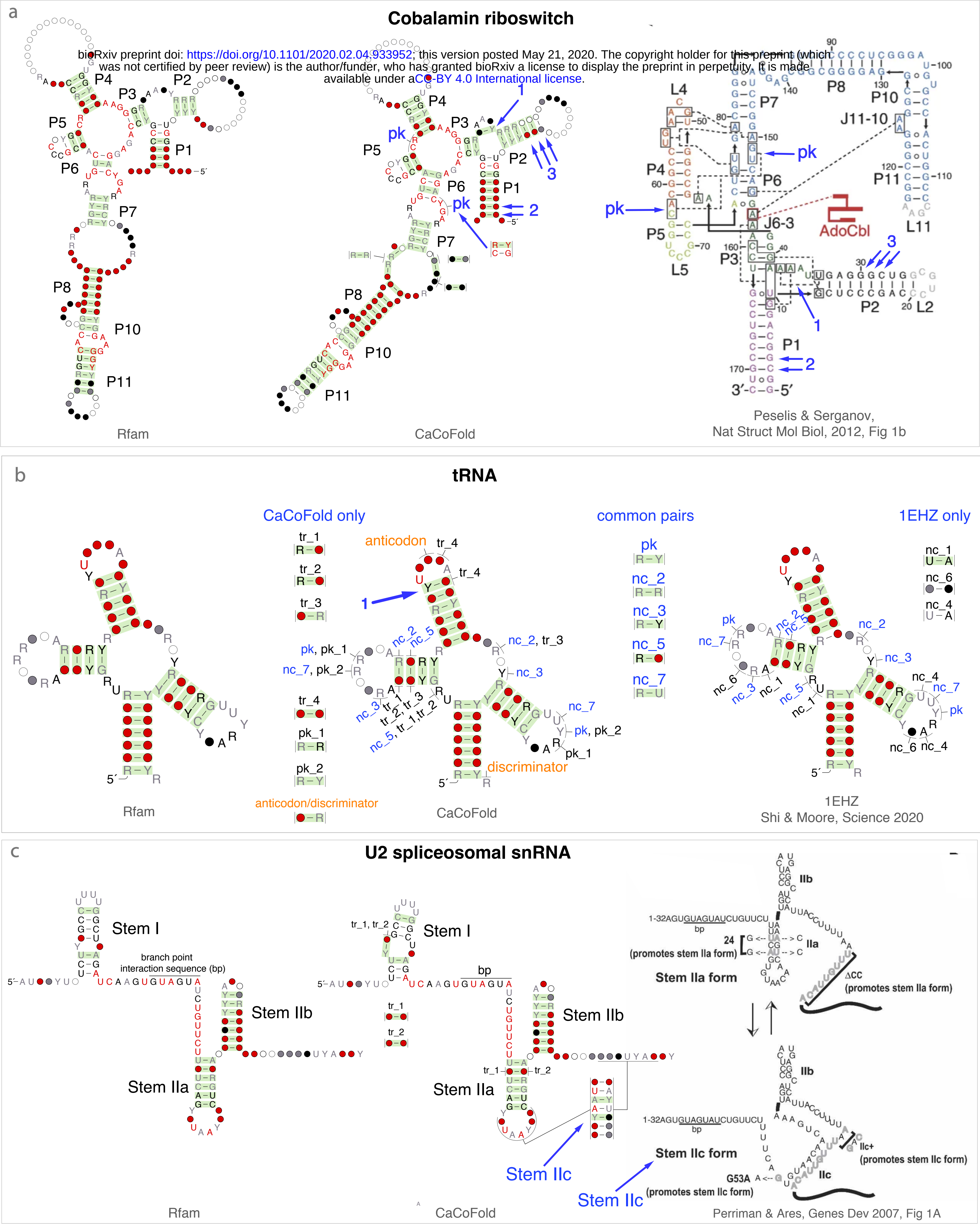
Montange & Batey, Nature 2006, Fig 1a

c **U4 spliceosomal RNA**

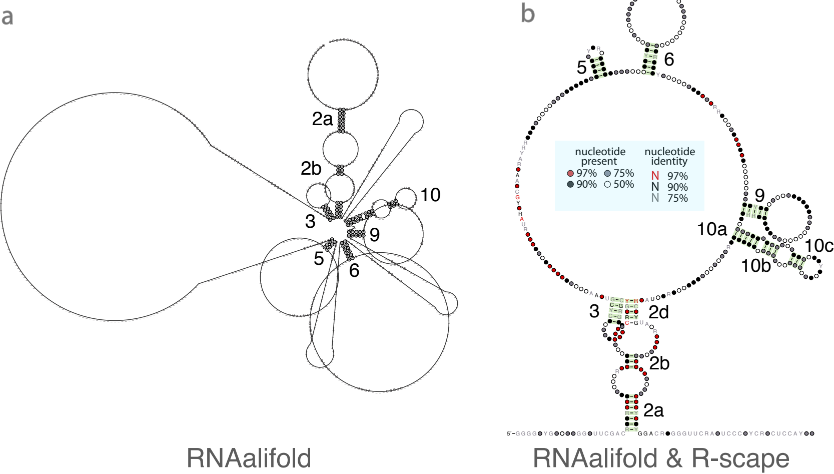
Rfam

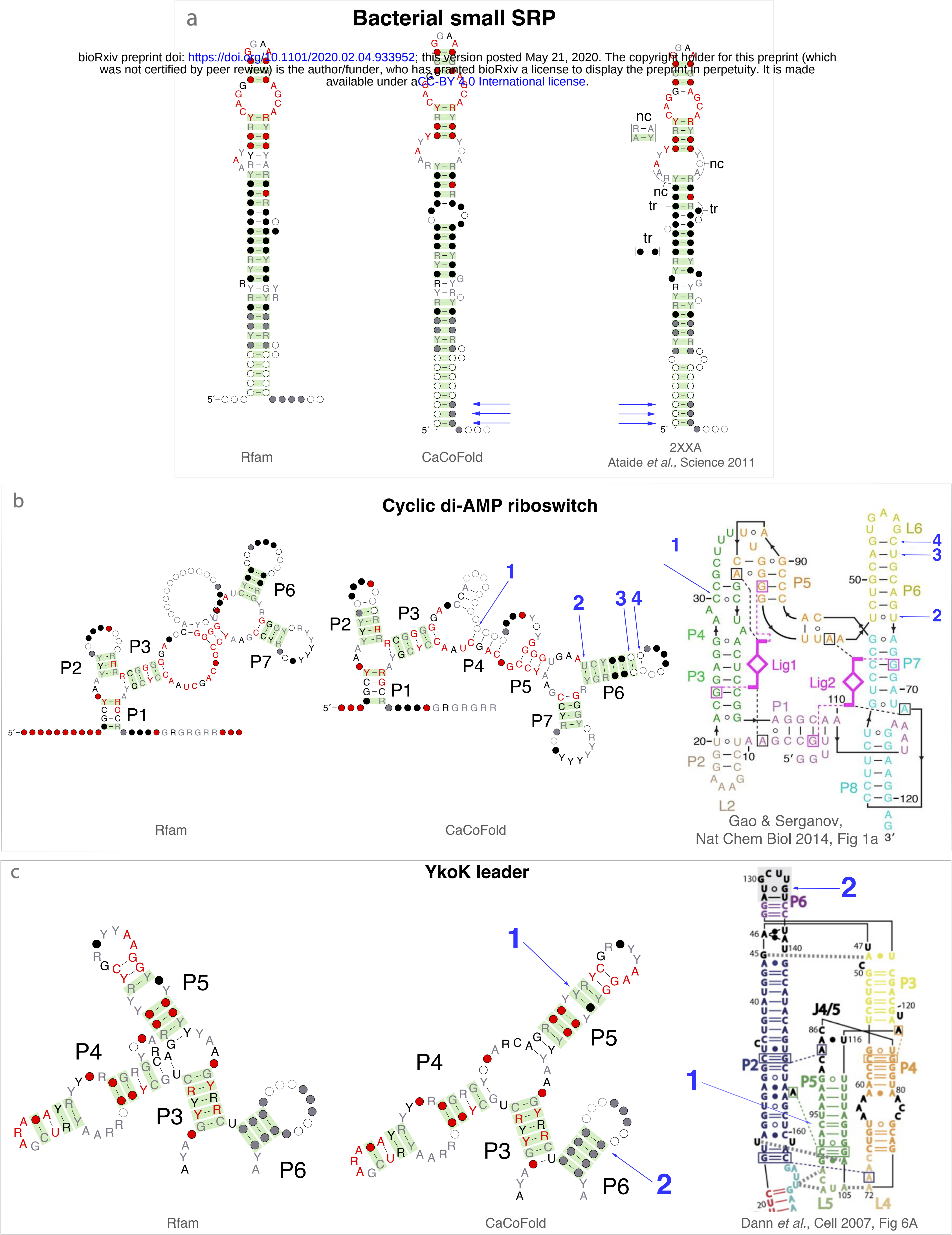
CaCoFold

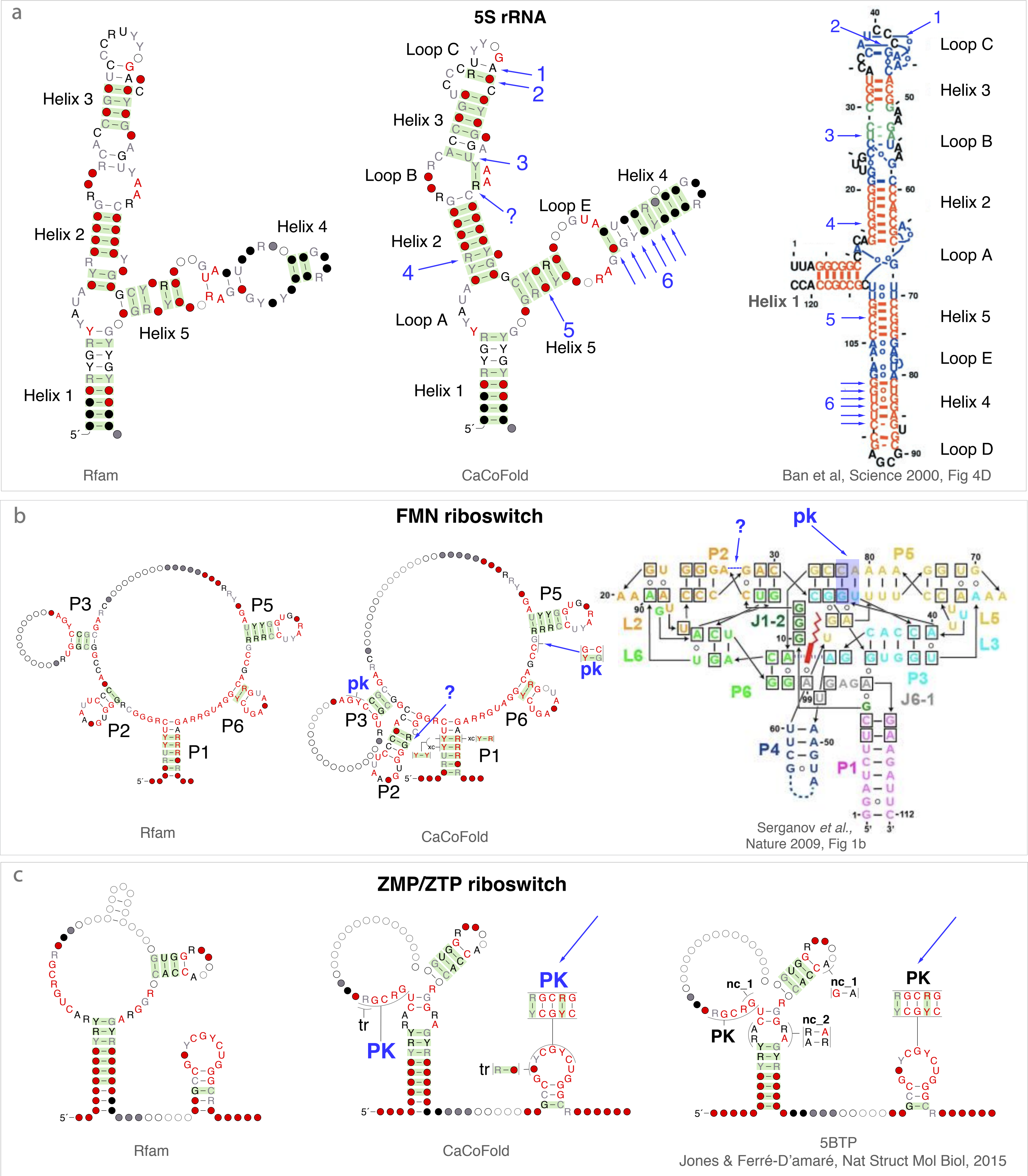
Wan *et al.*, Science 2016, Fig 3B

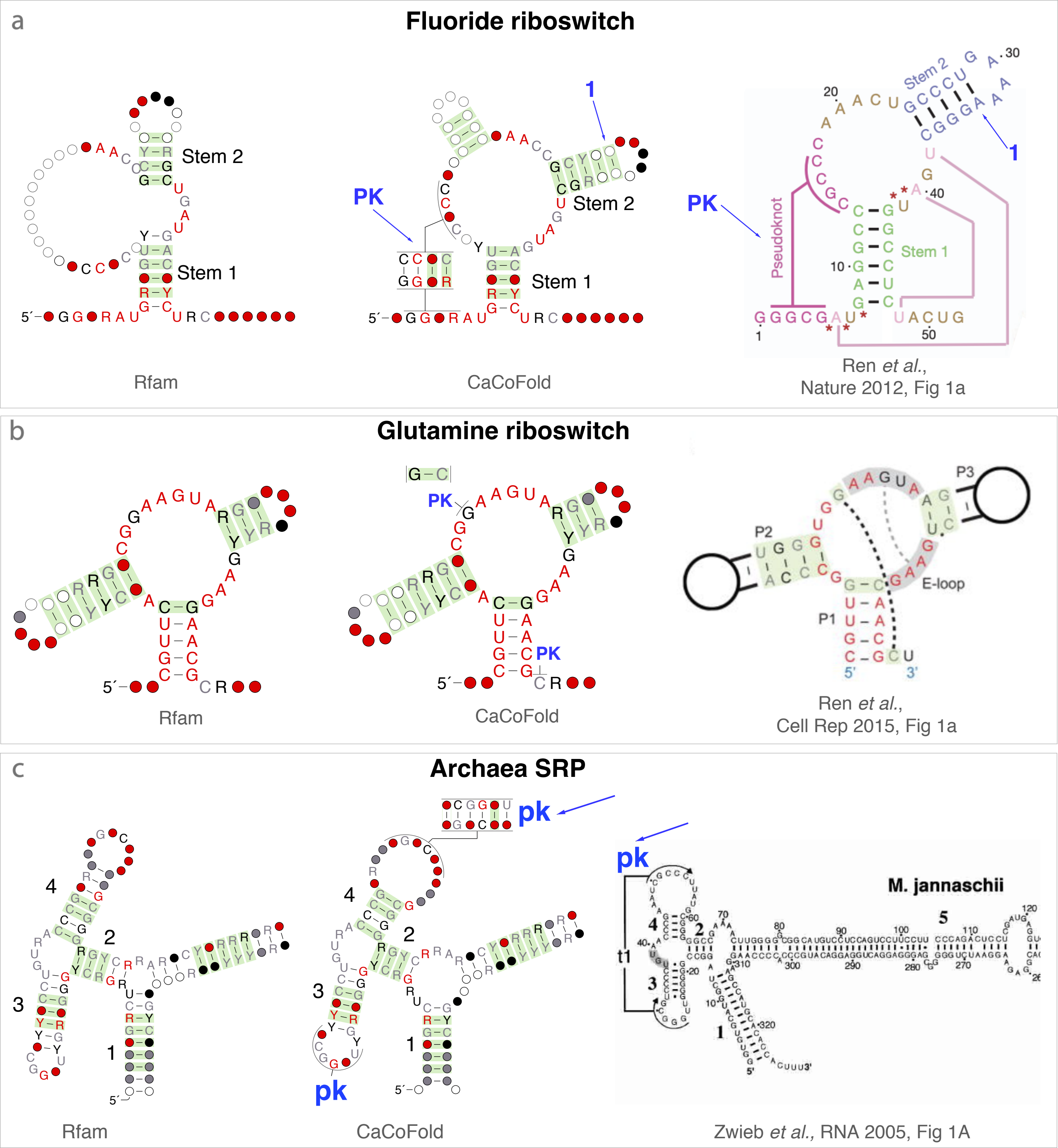

a
Cobalamin riboswitch
P1
P2
P3
P4
P5
P6
P7
P8
P10
P11
Rfam
pk
1
2
3
CaCoFold
L4
L5
L11
J11-10
J6-3
AdoCbl
Peselis & Serganov,
Nat Struct Mol Biol, 2012, Fig 1b
b
tRNA
CaCoFold only
tr_1
tr_2
tr_3
tr_4
pk_1
pk_2
anticodon/discriminator
anticodon
discriminator
common pairs
pk
nc_2
nc_3
nc_5
nc_7
1EHZ only
nc_1
nc_6
nc_4
Rfam
CaCoFold
1EHZ
Shi & Moore, Science 2020
c
U2 spliceosomal snRNA
Stem I
branch point
interaction sequence (bp)
Stem IIb
Stem IIa
Stem IIc
Rfam
CaCoFold
IIb
IIa
IIc
24
(promotes stem IIa form)
ΔCC
Stem IIa form
Stem IIc form
IIc+
(promotes stem IIc form)
G53A
1-32AGUGUAGUAUCUGUUCU
bp
Perriman & Ares, Genes Dev 2007, Fig 1A

transfer messenger RNA




a
Bacterial small SRP
Rfam
CaCoFold
2XXA
Ataide et al., Science 2011
nc
tr
b
Cyclic di-AMP riboswitch
P1
P2
P3
P4
P5
P6
P7
P8
Lig1
Lig2
L2
L6
1
2
3
4
Rfam
CaCoFold
Gao & Serganov,
Nat Chem Biol 2014, Fig 1a
c
YkoK leader
P2
P3
P4
P5
P6
J4/5
L4
L5
1
2
Rfam
CaCoFold
Dann et al., Cell 2007, Fig 6A


a
5S rRNA
Helix 1
Helix 2
Helix 3
Helix 4
Helix 5
Loop A
Loop B
Loop C
Loop D
Loop E
1
2
3
4
5
6
?
Rfam
CaCoFold
Ban et al, Science 2000, Fig 4D
b
FMN riboswitch
P1
P2
P3
P5
P6
pk
?
L2
L3
L5
L6
J1-2
J6-1
P4
Rfam
CaCoFold
Serganov et al., Nature 2009, Fig 1b
c
ZMP/ZTP riboswitch
PK
tr
nc_1
nc_2
Rfam
CaCoFold
5BTP
Jones & Ferré-D'amaré, Nat Struct Mol Biol, 2015



a

Fluoride riboswitch

Rfam

CaCoFold

Ren *et al.*, Nature 2012, Fig 1a

b

Glutamine riboswitch

Rfam

CaCoFold

Ren *et al.*, Cell Rep 2015, Fig 1a

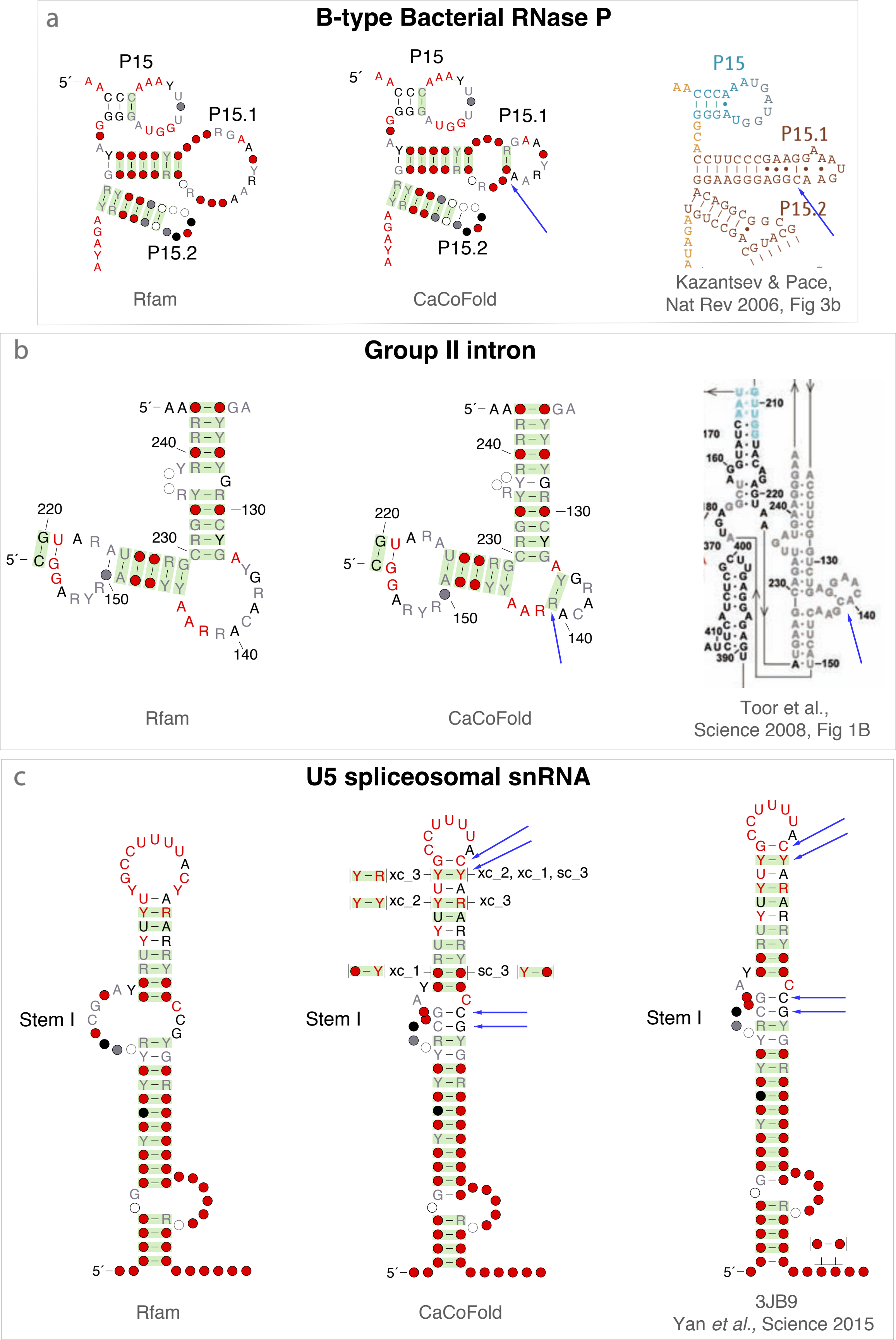
c

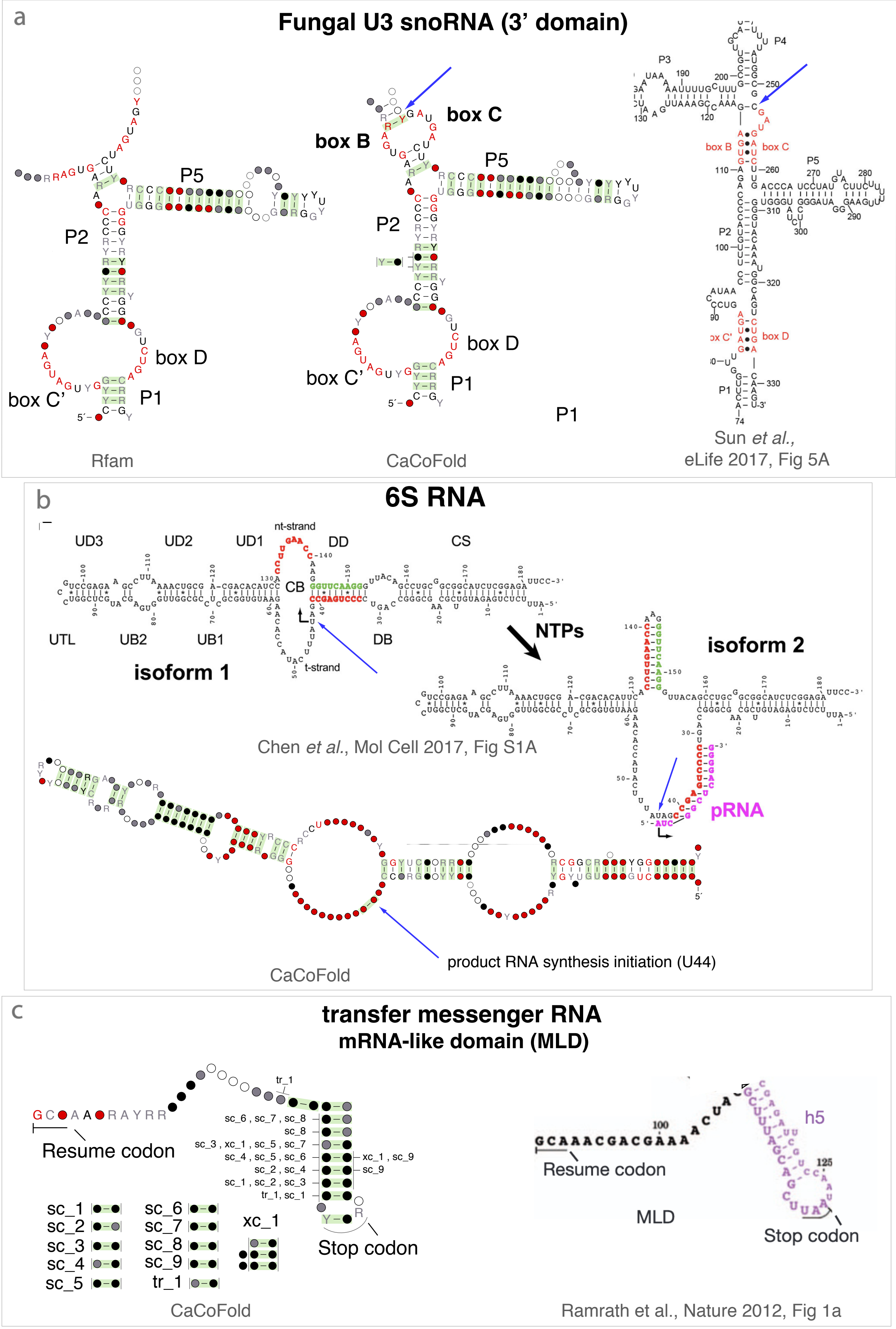
Archaea SRP

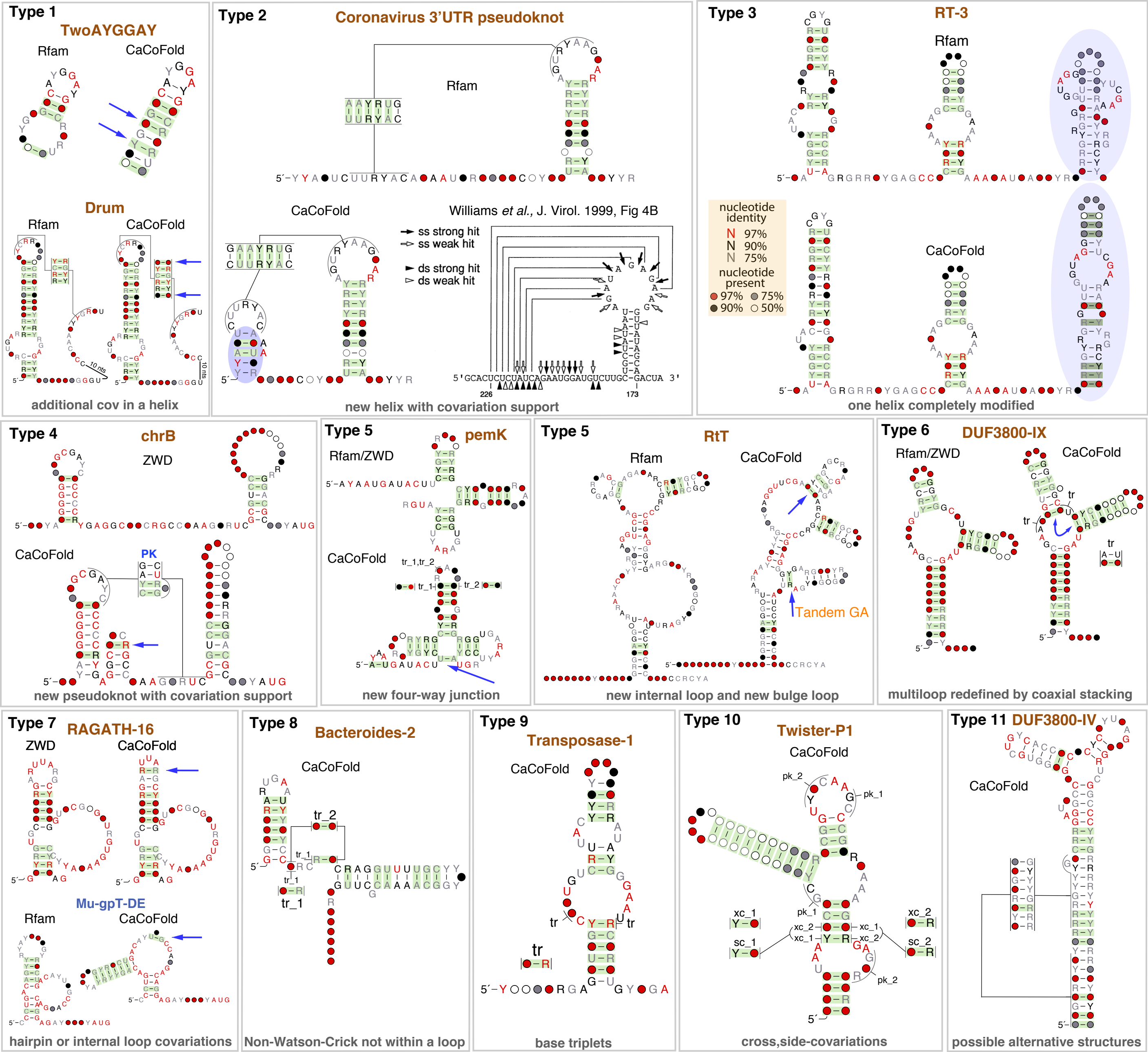
Rfam

CaCoFold

Zwieb *et al.*, RNA 2005, Fig 1A


a
B-type Bacterial RNase P
P15
P15.1
P15.2
5´
Rfam
CaCoFold
P15
P15.1
P15.2
Kazantsev & Pace,
Nat Rev 2006, Fig 3b
b
Group II intron
5´
240
220
230
130
150
140
Rfam
CaCoFold
210
170
160
220
240
180
370
400
130
230
140
410
390
150
Toor et al.,
Science 2008, Fig 1B
c
U5 spliceosomal snRNA
Stem I
5´
Rfam
xc_3
xc_2, xc_1, sc_3
xc_2
xc_3
xc_1
sc_3
Stem I
CaCoFold
Stem I
3JB9
Yan et al., Science 2015


a
Fungal U3 snoRNA (3' domain)
P5
P2
box D
box C'
P1
5´–
Rfam
box B
box C
P5
P2
box D
box C'
P1
CaCoFold
P4
P3
190
200
250
130
120
box B
box C
110
260
P5
270
280
310
290
300
P2
100
320
90
box C'
box D
P1
330
74
3'
P1
Sun et al.,
eLife 2017, Fig 5A
b
6S RNA
nt-strand
UD3
UD2
UD1
DD
CS
CB
UTL
UB2
UB1
DB
t-strand
isoform 1
NTPs
isoform 2
pRNA
Chen et al., Mol Cell 2017, Fig S1A
5´
product RNA synthesis initiation (U44)
CaCoFold
c
transfer messenger RNA
mRNA-like domain (MLD)
Resume codon
tr_1
sc_6 , sc_7 , sc_8
sc_8
sc_3 , xc_1 , sc_5 , sc_7
sc_4 , sc_5 , sc_6
xc_1 , sc_9
sc_2 , sc_4
sc_9
sc_1 , sc_2 , sc_3
tr_1, sc_1
Stop codon
sc_1
sc_2
sc_3
sc_4
sc_5
sc_6
sc_7
sc_8
sc_9
tr_1
xc_1
CaCoFold
GCAAACGACGAAA
100
Resume codon
h5
125
MLD
Stop codon
Ramrath et al., Nature 2012, Fig 1a


Type 1
TwoAYGGAY
Rfam
CaCoFold
Drum
Rfam
CaCoFold
10 nts
10 nts
additional cov in a helix
Type 2
Coronavirus 3'UTR pseudoknot
Rfam
CaCoFold
Williams et al., J. Virol. 1999, Fig 4B
ss strong hit
ss weak hit
ds strong hit
ds weak hit
5'GCACUCUCUAUCAGAAUGGAUGUCUUGC-GACUA 3'
226
173
new helix with covariation support
Type 3
RT-3
Rfam
CaCoFold
nucleotide identity
N 97%
N 90%
N 75%
nucleotide present
97%
75%
90%
50%
one helix completely modified
Type 4
chrB
ZWD
CaCoFold
PK
new pseudoknot with covariation support
Type 5
pemK
Rfam/ZWD
CaCoFold
tr_1,tr_2
tr_1
tr_2
new four-way junction
Type 5
RtT
Rfam
CaCoFold
Tandem GA
new internal loop and new bulge loop
Type 6
DUF3800-IX
Rfam/ZWD
CaCoFold
tr
tr
tr
multiloop redefined by coaxial stacking
Type 7
RAGATH-16
ZWD
CaCoFold
Mu-gpT-DE
Rfam
CaCoFold
hairpin or internal loop covariations
Type 8
Bacteroides-2
CaCoFold
tr_2
tr_1
tr_1
Non-Watson-Crick not within a loop
Type 9
Transposase-1
CaCoFold
tr
tr
tr
base triplets
Type 10
Twister-P1
CaCoFold
pk_2
pk_1
pk_1
pk_2
xc_1
xc_2
xc_1
xc_2
xc_2
xc_1
sc_1
sc_2
cross,side-covariations
Type 11
DUF3800-IV
CaCoFold
possible alternative structures