Sequence-based identification of 3D structural modules in RNA with RMDetect

José Almeida Cruz & Eric Westhof

Structural RNA modules, sets of ordered non-Watson-Crick base pairs embedded between Watson-Crick pairs, have central roles as architectural organizers and sites of ligand binding in RNA molecules, and are recurrently observed in RNA families throughout the phylogeny. Here we describe a computational tool, RNA three-dimensional (3D) modules detection, or RMDetect, for identifying known 3D structural modules in single and multiple RNA sequences in the absence of any other information. Currently, four modules can be searched for: G-bulge loop, kink-turn, C-loop and tandem-GA loop. In control test sequences we found all of the known modules with a false discovery rate of 0.23. Scanning through 1,444 publicly available alignments, we identified 21 yet unreported modules and 141 known modules. RMDetect can be used to refine RNA 2D structure, assemble RNA 3D models, and search and annotate structured RNAs in genomic data.

Structured RNAs present hierarchical architectures in which double-stranded helices and single-stranded loops are organized in three-dimensional (3D) space by tertiary interactions. The helices are formed by stacks of Watson-Crick base pairs and the tertiary interactions consist mainly of non-Watson-Crick base pairs[1]. Tertiary interactions occur either between nucleotides in the same domain (for example, internal loops and junctions) or between nucleotides from distant domains (for example, loop-loop and loop-helix interactions, and pseudoknots). The tertiary interactions, by establishing local and specific contacts, build up 3D structural modules that are characterized by sets of non-Watson-Crick base pairs organized in a precise order. Modules occur recurrently in different RNAs stemming from any phylogenetic branch and display similar 3D shapes independently of the surrounding structural context[2]. They have important functional roles in RNA molecules as protein and RNA binding sites[3] and as local structural organizers in junctions or internal loops[4].

Structural RNA modules are often referred to as RNA motifs. We favor the term 'module' to distinguish between closely related, although distinct, concepts: sequence motifs, which are patterns of nucleotides; RNA motifs, which are sets of secondary structure elements (helices, single strands, hairpins, loops and others)[5,6]; and structural RNA modules, which are ensembles of stacked arrays of ordered non-Watson-Crick base pairs[3]. This distinction separates 'objects' that exist in tertiary structure from those that exist only in sequence.

The identification of a module in a RNA sequence can provide key information about the secondary structure and the resulting tertiary fold[7–9]. Therefore, the identification of a structural RNA module lends support to the identification of a transcript as a structured RNA[10], presents clues for the local function of the molecule[11,12] and explains chemical probing data because modules present defined chemical probing signatures and mutational data[4]. Recent tools for module searching in structures[13–16] illustrate the importance of module discovery. However, none of these tools have been designed to find modules in sequence alone.

Several RNA motif search tools are currently available. Some (RNAMotif[5] or MilPat[17]) rely on user-defined descriptors of sequence and secondary structure. Others (CMFinder[6]) infer assemblies of secondary-structure elements from homologous sequences. These tools search for specific secondary structure elements that can span up to hundreds of nucleotides with extensive helical regions and perform poorly when searching small sequence motifs with less than 20 nucleotides (**Supplementary Note 1**). The 3D structure prediction tools can, in theory, provide information about structural modules, but they require considerable amount of computer resources and expertise.

Here we present a computational tool for structural RNA module searching based solely on sequence information, which we called RNA 3D modules detection (RMDetect). To capture all the possible variations of the allowed tertiary interactions and base pairs, RMDetect relies on Bayesian network models, base-pair probability prediction and positional clustering of candidates. We tested the performance of RMDetect on 1,444 noncoding (nc)RNA alignments for finding four recurrent modules: G-bulge loop (referred to as G-bulge)[4], kink-turn[2,12], C-loop[2] and tandem GA/AG loop (referred to as tandem GA)[18]. From the 1,444 alignments, we identified 141 cases of known instances of the modules and 21 new candidates. RMDetect can be used on single sequences or on multiple sequence alignments and can be applied to any newly discovered module irrespective of the complexity or number of strands involved. The use of RMDetect with 2D structure algorithms can improve accuracy of predictions.

Architecture et Réactivité de l'ARN, Institut de Biologie Moléculaire et Cellulaire du Centre National de la Recherche Scientifique, Université de Strasbourg, Strasbourg, France. Correspondence should be addressed to E.W. (e.westhof@ibmc-cnrs.unistra.fr).

RECEIVED 8 NOVEMBER 2010; ACCEPTED 11 APRIL 2011; PUBLISHED ONLINE 8 MAY 2011; DOI:10.1038/NMETH.1603

Together with presently available modeling tools[7–9], RMDetect can be used to build relevant RNA models and also to search and annotate ncRNAs in genomic data.

Other modules not covered by the current implementation of RMDetect exist and new modules are likely yet to be discovered. Some structured RNA may not contain any of the modules discussed here. Therefore we also provide a tool to build Bayesian network models corresponding to new modules based simply on 3D coordinates of the new module and sequence alignments representative of the module, called RNA 3D modules builder or RMBuild.

RESULTS

Structural RNA modules and interaction networks

The most accurate way to characterize a module and its interaction network is to analyze crystal structures. The comparison of many instances of a given module conveys essential information about its structural regularity and variation. However, typically only a few of the possible sequences compatible with the given module are found in existing crystal structures. To obtain a larger sample of the range of possible sequence variation one must resort to carefully curated alignments of homologous sequences. Such alignments indicate the nucleotides that can occur at each position of a module.

Interaction networks represent both the sequential regularity and the variation present in structural modules without atomic details. They depict the nucleotide frequencies and base-base interactions for each module instance. After merging the interaction networks of all instances, one obtains an integrated interaction network that captures the full sequential regularity and variation of that module, irrespective to the specific molecule in which it is embedded and of the module location. We selected four known recurrent modules because they have key roles in many types of RNAs (**Fig. 1**).

Descriptions of modules

The G-bulge module is observed in the three rRNAs[19], in the lysine riboswitch[20], in the group I intron P7.1/P7.2 domain[21] and in the T-box leader[22]. G-bulge modules are formed by four stacked non-Watson-Crick base pairs (**Fig. 1a**) with a characteristic bulging G that participates in a triple interaction with the flanking base pair. The G-bulge module organizes internal loops and junctions, and often forms binding platforms for proteins[4,19].

The kink-turn module, an asymmetric internal loop, leads to a sharp bend between two helical regions[12] (**Fig. 1b**). One of the helices contains exclusively Watson-Crick base pairs, and the three base pairs of second helical region, closest to the internal loop, usually form a GAA/GGA Hoogsteen-Sugar edge platform[2]. The kink-turn modules bind several ribosomal proteins. The *U4* small nuclear (sn)RNA and small nucleolar (sno)RNAs bind the 15.5 kDa protein in eukaryotes and the homologous archaeal protein L7 (ref. 23).

The C-loop module is an asymmetrical internal loop between two canonical helices. C-loops increase the helical twist between the helices[2] (**Fig. 1c**). C-loop modules have been observed in rRNAs and in a synthetase mRNA regulatory element[24].

The tandem-GA module is a small module formed by two consecutive Hoogsteen/sugar edge base pairs, A-G and G-A[18]. They are frequently observed within regular helices. We considered tandem GAs with four stacked base

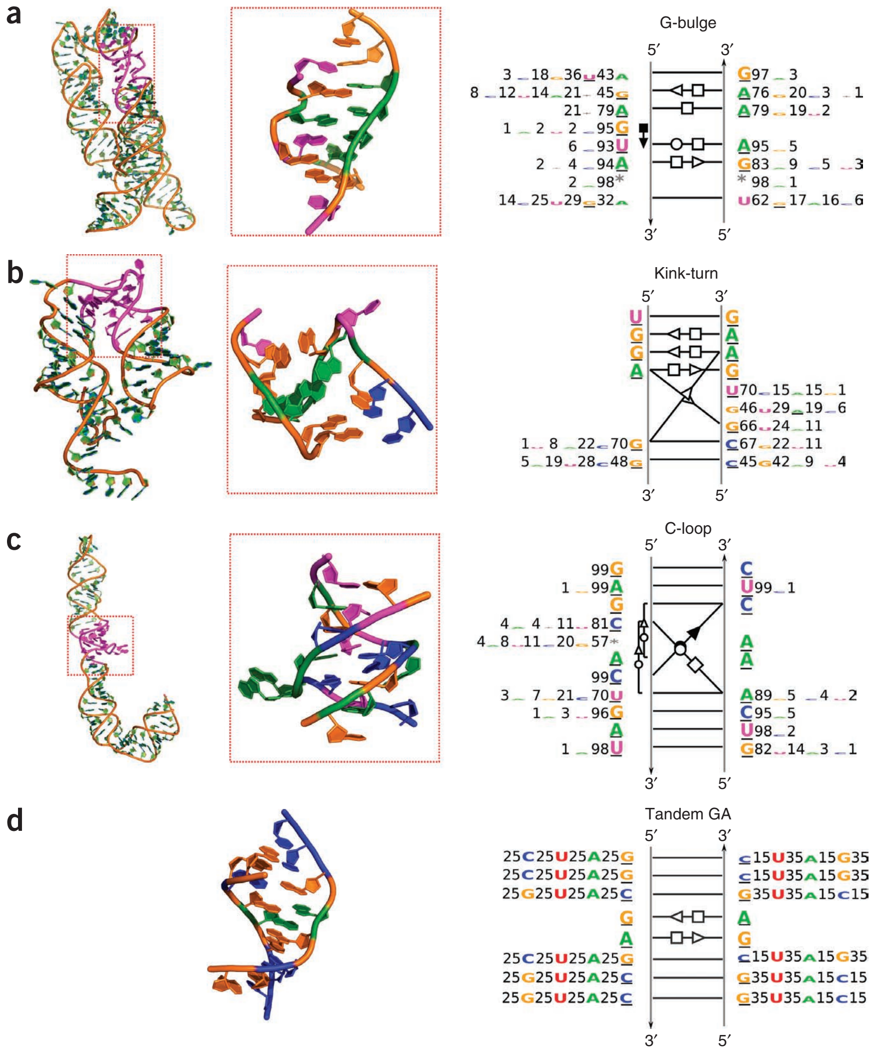

Figure 1 | Details of the analyzed RNA structural modules. (**a**) The G-bulge from the lysine riboswitch (Protein Data Bank (PDB) code 3DIG)[20]. (**b**) The kink-turn from the helix 46 of the bacterial large subunit (PDB code 2WRJ)[40]. (**c**) The C-loop from the helix 38 of the bacterial large subunit (PDB code 2WRJ)[40]. (**d**) A tandem GA from a synthetic RNA octamer (PDB code 1SA9)[41]. For each module, a detailed structure (center), the position in the original molecule (left) and the interaction network (right) are shown. The underlined bases in the interaction network correspond to the nucleotides present in the crystal structure. Numbers next to the bases indicate the observed percentages of each nucleotide in the alignment.

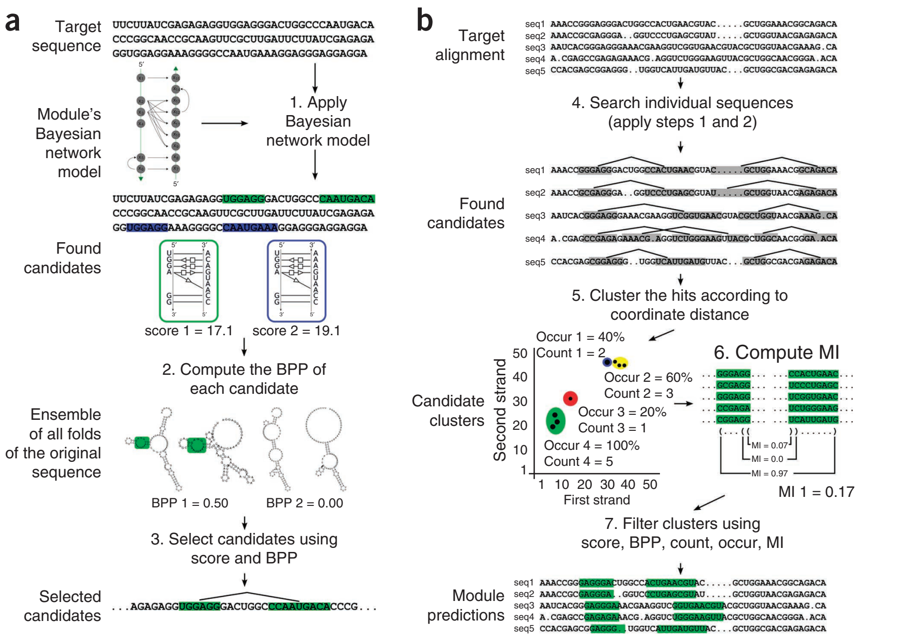
a
Target sequence
UUCUUAUCGAGAGAGGUGGAGGACUGGCCCAAUGACA
CCCGGCAACCGCAAGUUCGCUUGAUUCUUAUCGAGAGA
GGUGGAGGAAAGGGGCCAAUGAAAGGAGGGAGGAGGA
Module's Bayesian network model
1. Apply Bayesian network model
Found candidates
score 1 = 17.1
score 2 = 19.1
2. Compute the BPP of each candidate
Ensemble of all folds of the original sequence
BPP 1 = 0.50
BPP 2 = 0.00
3. Select candidates using score and BPP
Selected candidates
b
Target alignment
seq1
seq2
seq3
seq4
seq5
4. Search individual sequences (apply steps 1 and 2)
Found candidates
5. Cluster the hits according to coordinate distance
Candidate clusters
Occur 1 = 40%
Count 1 = 2
Occur 2 = 60%
Count 2 = 3
Occur 3 = 20%
Count 3 = 1
Occur 4 = 100%
Count 4 = 5
Second strand
First strand
50
40
30
20
10
1
1 10 20 30 40 50
6. Compute MI
MI = 0.07
MI = 0.0
MI = 0.97
MI 1 = 0.17
7. Filter clusters using score, BPP, count, occur, MI
Module predictions

Figure 2 | Steps of single- and multiple-sequence search algorithms. (**a**) In step 1 of the single-sequence search algorithm, first the Bayesian network model is applied to the target sequence to obtain potential candidates and their respective scores. Step 2 is to fold the target sequence and compute the proportion of the ensemble (set of all possible folds for that sequence) compatible with the candidates found in the previous step. This proportion is referred to as base-pair probability (BPP). In step 3, candidates are filtered using predefined score and BPP thresholds. (**b**) In the multiple sequence search algorithm used with multiple homologous sequences, step 4 is to apply the previous algorithm to each individual sequence of the target alignment to obtain the candidates for all sequences (seq1–seq5). Step 5 is to represent each candidate, in a matrix, using the starting column of the candidate strands in the alignment as coordinates. Cluster the candidates according to their location in the matrix and compute the frequency of each cluster in the alignment (occur). Overlapping candidates are discarded. Step 6 is to compute the average mutual information (MI) of each cluster as a measure of variation between positions. The MI of the cluster is the mean of the individual MI of expected Watson-Crick base pair positions, and it is normalized by the maximum possible MI (2 bits per base pair). In step 7, heuristic rules are used to filter candidates based on score, BPP, occur, count and MI values.

pairs: Watson-Crick, G-A, A-G and Watson-Crick. The module contains two sequences of four nucleotides, NGAN (in which N is any nucleotide), which can occur, by chance, once in each 16 random bases. This short sequence makes it difficult to distinguish tandem GAs from background. However, the conservation of the GA nucleotides across homologous sequences is usually distinguishable in sequence alignments (**Fig. 1d**).

Structural modules as Bayesian networks

The direct use of nucleotide distributions, observed in the interaction networks, to search for modules in sequences, presents the limitation of assuming statistical independence between the positions of the module. This independence is generally not verified. For example, Watson-Crick base pairs present a strong correlation between the bases. Sometimes, the same base pair can adopt more than one interaction type depending on the particular instance of a module, which imposes a dependency between bases even if, in some of the instances, one of the bases is fixed. Such a situation occurs in the kink-turn module in which the first base pair of the noncanonical stem usually adopts a Hoogsteen/Sugar edge interaction with an invariant A, but it can also adopt a Watson-Crick interaction, which imposes the corresponding isostericity constraints. Other dependencies can also occur between edge-interacting or stacking-interacting nucleotides. A way to overcome this limitation is to interpret an interaction network as a Bayesian network and explicitly model all the dependencies between the bases of the module observed in systematic structural alignments.

Bayesian networks are probabilistic models in which random variables and the dependency between them are represented as an acyclic directed graph. The nodes of the graph correspond to the random variables and the edges to the dependencies. Bayesian networks have been applied to sequence-analysis problems, for example, for detection of transcription factors[25]. For modeling RNA modules as Bayesian networks, the nodes represent individual bases occupying a defined structural position, and the edges represent the dependencies between them.

Single sequence search

When searching for structural modules in single sequences, RMDetect computes, for all subsequences, the log-likelihood score corresponding to the likelihood that the given subsequence was generated by the Bayesian network of the module. Owing to the small size of Bayesian networks and the short (four-letter) alphabet of nucleotides, this scan will normally produce a large number of medium to high score hits, many of them false positives. To reduce the number of false positives RMDetect uses the predicted joint base-pair probability of the module's Watson-Crick base pairs to select the subsequences for which a compatible secondary structure is likely to be observed (**Fig. 2a**).

To evaluate RMDetect for single sequence search we built 15 test cases, corresponding to the molecules in which the modules had been identified in crystal structures and for which we obtained reliable sequence alignments (**Supplementary Table 1** and Online Methods). We obtained Matthews correlation coefficient values of individual test cases[26], with fixed parameters, which varied between 0.93 for the kink-turn model and 0.13 for the tandem-GA model. We calculated the true positive rates to be above 0.5 for all but the tandem-GA module, indicating that RMDetect consistently found more than half of the positive candidates. However, false discovery rates higher than 0.5 for the three tandem-GAs confirmed the difficulty in discarding false positive candidates for small modules with few non-Watson-Crick interactions using single sequence information (**Table 1**). The diversity of the training set should be as complete as possible to obtain a representative model of a module. For example, a G-bulge model trained only

Table 1 | RMDetect analysis of the single-sequence test set

Searched			Best MCC[c]					Fixed parameters[d]				
module	**Molecule[a]**	**Instances[b]**	**MCC**	**TPR[e]**	**FDR[f]**	**Score[g]**	**BPP[h]**	**MCC**	**TPR[e]**	**FDR[f]**	**Score[g]**	**BPP[h]**
G-bulge	16S rRNA bacteria	2 × 250	0.73	0.70	0.22	12.3	0.001	0.58	0.71	0.53	8.0	0.001
G-bulge	23S rRNA archaea	6 × 100	0.68	0.55	0.15	13.8	0.001	0.61	0.64	0.42		
G-bulge	23S rRNA bacteria	5 × 250	0.71	0.67	0.24	11.5	0.001	0.63	0.71	0.44		
G-bulge	Lysine riboswitch	1 × 150	0.66	0.48	0.09	8.3	0.010	0.64	0.51	0.20		
Kink-turn	16S rRNA bacteria	1 × 250	0.97	0.96	0.01	17.2	0.041	0.67	0.97	0.53	11.0	0.001
Kink-turn	23S rRNA archaea	5 × 100	0.70	0.61	0.20	14.5	0.001	0.64	0.67	0.40		
Kink-turn	23S rRNA bacteria	4 × 250	0.67	0.52	0.16	15.7	0.001	0.59	0.65	0.48		
Kink-turn	SAM riboswitch[i]	1 × 150	0.93	0.93	0.07	8.7	0.001	0.93	0.91	0.06		
Kink-turn	U4 snRNA	1 × 500	0.71	0.54	0.06	12.3	0.001	0.70	0.55	0.10		
C-loop	16S rRNA bacteria	1 × 250	0.84	0.85	0.16	18.5	0.011	0.80	0.91	0.29	16.0	0.010
C-loop	23S rRNA archaea	3 × 100	0.66	0.50	0.11	22.4	0.001	0.48	0.54	0.57		
C-loop	23S rRNA bacteria	3 × 250	0.62	0.56	0.32	15.9	0.021	0.60	0.58	0.38		
Tandem GA	16S rRNA bacteria	1 × 250	0.42	0.66	0.74	9.6	0.161	0.36	0.67	0.81	9.0	0.100
Tandem GA	23S rRNA archaea	1 × 100	0.41	0.21	0.19	9.9	0.990	0.13	0.33	0.95		
Tandem GA	23S rRNA bacteria	2 × 250	0.53	0.67	0.58	9.5	0.530	0.39	0.82	0.82		

[a]Sequence alignments searched. [b]Number of (module) instances present in the alignment: module instances present in each sequence times the number of sequences. [c]Sensitivity and specificity analysis for the parameter that maximize the Matthews correlation coefficient (MCC). [d]Sensitivity and specificity analysis for fixed score and bpp for all test sets of the same module. [e]True positive rate (TPR) = TP / (TP + FN). [f]False discovery rate (FDR) = FP / (TP + FP). [g]Threshold score. [h]Threshold BPP values used to discriminate candidates. [i]SAM, S-adenosylmethionine.

with 16S rRNA and 23S rRNA did not identify most of the lysine riboswitch G-bulge instances (**Supplementary Note 2**).

Multiple sequence alignment search

The increasing availability of databases of homologous, or related, sequences for many RNA molecules[27] and the existence of effective RNA sequence alignment tools for close sequences[28] provides powerful sources of information for module discovery. When searching for modules in aligned RNA sequences, even if the positions where the modules occur are misaligned, we expect that the true positive candidates would be located in columns relatively close to each other. When sequences are sufficiently divergent, which is the case of many RNA sequences, false positive candidates should be distributed across the alignment. Based on these assumptions, we devised a clustering strategy to exploit multiple sequence alignment information for module searching. We clustered candidates according to the distance on the column space of the alignment and selected the most represented clusters, with higher score candidates and covariation signals between bases of Watson-Crick base pairs, as potential hits (**Fig. 2b**).

To test RMDetect on multiple sequence alignments, we applied it to the same 15 datasets of the single sequence search. RMDetect correctly found all of the 37 known module instances (true positive rate of 1) with 11 false positive candidates (false discovery rate of 0.23), five of them falsely identified as tandem-GA modules. These results show that RMDetect is effectively improved by adding alignment information (**Table 2** and **Supplementary Data 1**).

Search in public databases

We applied RMDetect to multiple sequence alignments from the RFam database, the group I intron database[29] and new bacterial ncRNAs reported in references 30 and 31 (**Supplementary Data 2**). Using the same selection conditions as in previous tests, we selected 250 candidates. From those, 21 predictions correspond to presently unreported modules, 141 correspond to previously predicted or observed modules, and the remaining 88 were unconfirmed candidates (**Table 3** and **Supplementary Data 3**).

Rfam results

Searching 1,309 Rfam alignments resulted in 222 module candidates, 132 of which were known modules and 77 of which corresponded to unconfirmed candidates. Not surprisingly, 99 of the known candidates corresponded to kink-turns in the snoRNAs C/D or C/D′ boxes. We found 13 previously undetected modules, including one kink-turn, one G-bulge, three C-loop and eight tandem GA modules (**Supplementary Fig. 1**).

Table 2 | RMDetect analysis of the multiple-sequence test set

	G-bulge		Kink-turn		C-loop		Tandem GA	
Alignment searched	**TP (TPR)[a]**	**FP (FDR)[b]**	**TP (TPR)**	**FP (FDR)**	**TP (TPR)**	**FP (FDR)**	**TP (TPR)**	**FP (FDR)**
16S_P	2 (1.00)	0 (0.00)	1 (1.00)	2 (0.66)	1 (1.00)	0 (0%)	1 (1.00)	2 (0.66)
23S_A	6 (1.00)	0 (0.00)	5 (1.00)	1 (0.17)	3 (1.00)	3 (50%)	1 (1.00)	0 (0.00)
23S_P	5 (1.00)	0 (0.00)	4 (1.00)	0 (0.00)	3 (1.00)	0 (0%)	2 (1.00)	3 (0.60)
SamRS	–	–	1 (1.00)	0 (0.00)	–	–	–	–
LysRS	1 (1.00)	0 (0.00)	–	–	–	–	–	–
U4 snRNA	–	–	1 (1.00)	0 (0.00)	–	–	–	–
Total	14 (1.00)	0 (0.00)	12 (1.00)	3 (0.20)	7 (1.00)	3 (0.30)	4 (1.00)	5 (0.55)

–, not applicable.
[a]Number of true positives (TP) and true positive rate (TPR = TP / (TP + FN)). [b]Number of false positives (FP) and false discovery rate (FDR = FP / (TP + FP)).

Table 3 | RMDetect analysis of public database alignments

Database (aligments searched)		G-bulge	Kink-turn	C-loop	Tandem GA
Rfam (1,309 alignments)	Total selected candidates	13	119	22	68
	Known modules[a]	6	105 (99 snoRNAs)	0	21 (20 kink-turns)
	New candidates[b]	1	1	3	8
	Not confirmed[c]	6	13	19	39
Group I introns (14 alignments)	Total selected candidates	1	1	0	1
	Known modules[a]	1	0	0	0
	New candidates[b]	0	0	0	1
	Not confirmed[c]	0	1	0	0
Bacterial ncRNAs (121 alignments)	Total selected candidates	4	4	1	16
	Known modules[a]	3	0	0	5 (1 kink-turn)
	New candidates[b]	1	1	0	5
	Not confirmed[c]	0	3	1	6

[a]Number of selected candidates corresponding to known modules. [b]Number of selected candidates corresponding to new putative modules. [c]Number of false positive candidates or candidates for which no confirmation was possible.

We detected the newly predicted kink-turn in 353 (16%) sequences in the variable region of the cobalamin riboswitch alignment[32]. With realignment, using the predicted kink-turn as an anchor, we established the full conservation of the tandem-GA sequences as well as the perfect pairing of at least two Watson-Crick base pairs in both helical stems with strong covariation (**Fig. 3a**). Another strong candidate was a G-bulge found in 109 (11%) sequences of the Hepatitis C virus stem-loop SL-VII[33]. Unlike the cobalamin riboswitch candidate, this G-bulge is conserved, correctly aligned and stands out in the secondary structure derived from the full alignment (**Fig. 3b**). Although alternative folding is possible, in which the G-bulge region participates in a helix interrupted by two bulged adenines, the conservation of the AGUA-GA sequences and the covariation of the base pairs in the hairpin support the prediction of a G-bulge. We detected three potential C-loops in the *c-mic* internal ribosome entry site (IRES)[34] (**Fig. 3c**), enterovirus *cis*-acting replication element (*CRE*)[35] and *QUAD* bacterial ncRNA[36] in 37 (45%), 112 (54%) and 174 (49%) sequences respectively. In the first case, we found the candidate in a region flanking a pseudoknot in the structure of the IRES. The covariation of the helices and the conservation of the characteristic 'CAC' motif support the prediction. In the cases of the enterovirus *CRE* and *QUAD* RNA the candidates stand out from the originally proposed secondary structure with no rearrangement needed.

We detected a tandem GA in 157 sequences (40%) of the *rtT* alignment, a bacterial ncRNA observed as a transcription product of the tyrT operon of *Escherichia coli*[37]. The detected module suggests a rearrangement of one internal loop of the originally proposed structure. It is possible that the module is not present in all sequences. We detected a second tandem GA in 16 sequences (47%) of the 5′ untranslated region of the voltage-gated potassium channel mRNA where the proposed secondary structure suggests the detected module. A tandem GA, predicted in 20 (71%) sequences in the *purD* alignment[38], stands at an internal loop compatible with a rare type of kink-turn with four nucleotides in the bulge (**Fig. 3d**). One can rearrange the secondary structure to obtain the minimal tandem GA maintaining the covariation, but we cannot discard the possibility of a more complex module. Notably, 21 of the identified tandem GAs correspond to kink-turns. This is not surprising because kink-turns contain a tandem GA and that the first base of the bulge can often be predicted as forming a base pair with the base in the opposite strand.

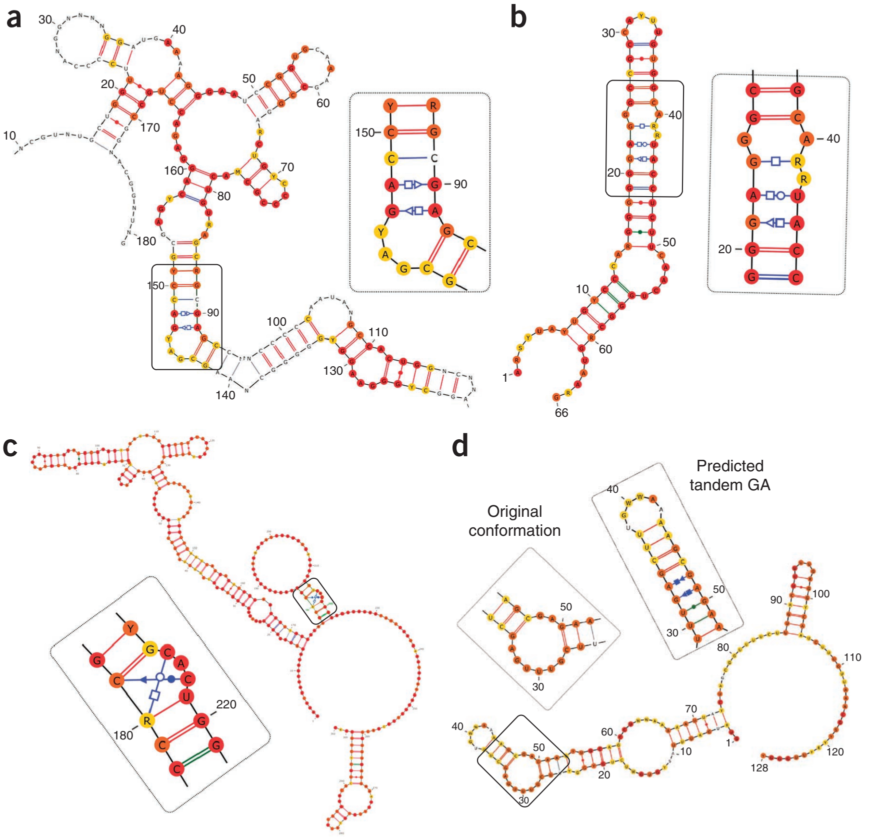

Figure 3 | Examples of the newly predicted modules. (**a**) Kink-turn in the Cobalamin riboswitch. (**b**) G-bulge in the SL VII domain of the Hepatitis C virus. (**c**) C-loop in the internal ribosomal entry site from the C-myc mRNA. (**d**) Tandem GA in the *purD* bacterial RNA (original conformation as in ref. 38).

RMDetect allowed the correct detection of the three modules (two G-bulges and one kink-turn) in the T-box riboswitch[22], the two modules (one G-bulge and one kink-turn) in the lysine riboswitch[20] and the G-bulge module in the IRES of the Hepatitis C virus[39].

Group I intron results

Searching the 14 alignments of the group I intron database, we detected the known G-bulge module present in the P7 domain of type IA2 introns, confirmed by the crystal structure of the phage Twort intron[21]. Additionally we detected a tandem GA in 12 (38%) sequences of type IC2 intron. This candidate was predicted in P5d domain (**Supplementary Fig. 2**).

Bacterial ncRNA results

Several modules had been originally identified on 121 alignments of structured ncRNAs from recently published metagenomic data[30,31]. We applied our algorithm to all of these alignments. We found a new kink-turn in the GEMM-II alignment, a new G-bulge in group-II-D1D4-1 molecule and five new tandem-GA modules (**Supplementary Fig. 2**). The twoAYGGAY motif[31] bacterial RNA alignment is an interesting case: we detected two tandem GAs in the same hairpin stem, a distal one (10 base pairs (bp) from the loop) in 80 sequences (39%), and a proximal one (2 bp from the loop) in 28 sequences (14%). All combinations of one and both tandem GAs can be found in the alignment (**Supplementary Fig. 3**). Homologous sequences that do not contain the module instead have Watson-Crick base pairs at the positions corresponding to the module. This alignment raises interesting questions about how structural modules evolve and interchange and how this will affect the final 3D geometry of the molecule. However, RMDetect missed four of seven previously reported G-bulge in the dataset. One was that of GOLLD that differs slightly from the defined G-bulge model, putting it outside the scope of the Bayesian network model. Two others correspond to Dictyoglomi-1 G-bulges that spanned more than the sliding window length (150 nucleotides). We detected the missing modules by reapplying the algorithm with no window length limitation. We discarded the final Dictyoglomi-1 G-bulge despite a high score (17.0) and occurrence (75%) owing to the small alignment (4 sequences) and total conservation in the module region (mutual information score of 0.0). This observation highlights the fact that the RMDetect parameters, although necessary owing to the large number of searched alignments, will not guarantee the exhaustive search of sequence space. When searching a small number of alignments, different window lengths and steps together with more relaxed selection criteria should be applied.

DISCUSSION

In the single-sequence test set we detected more than half of the searched modules in molecules as complex as the ribosome. In multiple sequence alignment test sets, we identified all known modules with an overall false discovery rate of 0.23. We extended the search to 1,444 publicly available alignments used without realignment. We found most of the known modules in all major classes of structured ncRNAs and identified 21 new candidates. With the RMBuild tool (**Supplementary Note 3**), our approach can be extended to additional modules and newly discovered ones. The Bayesian network models can be further improved with new instances of known modules.

RMDetect is available as **Supplementary Software** and at http://sourceforge.net/projects/rmdetect/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank R. Backofen for useful suggestions. J.A.C. is supported by the Ph.D. Program in Computational Biology of the Instituto Gulbenkian de Ciência, Portugal (sponsored by Fundação Calouste Gulbenkian, Siemens SA and Fundação para a Ciência e Tecnologia; SFRH/BD/33528/2008).

AUTHOR CONTRIBUTIONS

J.A.C. conceived the algorithms, performed the computations and wrote the manuscript. E.W. conceived the research and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Leontis, N.B., Stombaugh, J. & Westhof, E. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.* **30**, 3497–3531 (2002).
2. Lescoute, A. *et al.* Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Res.* **33**, 2395–2409 (2005).
3. Leontis, N.B. & Westhof, E. Analysis of RNA motifs. *Curr. Opin. Struct. Biol.* **13**, 300–308 (2003).
4. Leontis, N.B. & Westhof, E. A common motif organizes the structure of multi-helix loops in 16 S and 23 S ribosomal RNAs. *J. Mol. Biol.* **283**, 571–583 (1998).
5. Macke, T.J. *et al.* RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* **29**, 4724–4735 (2001).
6. Yao, Z., Weinberg, Z. & Ruzzo, W.L. CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics* **22**, 445–452 (2006).
7. Parisien, M. & Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**, 51–55 (2008).
8. Jossinet, F., Ludwig, T.E. & Westhof, E. Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics* **26**, 2057–2059 (2010).
9. Das, R., Karanicolas, J. & Baker, D. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods* **7**, 291–294 (2010).
10. Westhof, E. The amazing world of bacterial structured RNAs. *Genome Biol.* **11**, 108 (2010).
11. Moore, P.B. Structural Motifs in RNA. *Annu. Rev. Biochem.* **68**, 287–300 (1999).
12. Klein, D.J. *et al.* The kink-turn: a new RNA secondary structure motif. *EMBO J.* **20**, 4214–4221 (2001).
13. Djelloul, M. & Denise, A. Automated motif extraction and classification in RNA tertiary structures. *RNA* **14**, 2489–2497 (2008).
14. Sarver, M. *et al.* FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.* **56**, 215–252 (2008).
15. Apostolico, A. *et al.* Finding 3D motifs in ribosomal RNA structures. *Nucleic Acids Res.* **37**, e29 (2009).
16. Zhong, C., Tang, H. & Zhang, S. RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic Acids Res.* **38**, e176 (2010).
17. Thébault, P. *et al.* Searching RNA motifs and their intermolecular contacts with constraint networks. *Bioinformatics* **22**, 2074–2080 (2006).
18. Gautheret, D., Konings, D. & Gutell, R.R. A major family of motifs involving G? A mismatches in ribosomal RNA. *J. Mol. Biol.* **242**, 1–8 (1994).
19. Leontis, N.B., Stombaugh, J. & Westhof, E. Motif prediction in ribosomal RNAs Lessons and prospects for automated motif prediction in homologous RNA molecules. *Biochimie* **84**, 961–973 (2002).

20. Serganov, A., Huang, L. & Patel, D.J. Structural insights into amino acid binding and gene control by a lysine riboswitch. *Nature* **455**, 1263–1267 (2008).
21. Golden, B.L., Kim, H. & Chase, E. Crystal structure of a phage Twort group I ribozyme-product complex. *Nat. Struct. Mol. Biol.* **12**, 82–89 (2005).
22. Wang, J., Henkin, T.M. & Nikonowicz, E.P. NMR structure and dynamics of the specifier loop domain from the *Bacillus subtilis* tyrS T box leader RNA. *Nucleic Acids Res.* **38**, 3388–3398 (2010).
23. Kuhn, J.F., Tran, E.J. & Maxwell, E.S. Archaeal ribosomal protein L7 is a functional homolog of the eukaryotic 15.5kD/Snu13p snoRNP core protein. *Nucleic Acids Res.* **30**, 931–941 (2002).
24. Sankaranarayanan, R. *et al.* The structure of threonyl-tRNA synthetase-tRNA(Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site. *Cell* **97**, 371–381 (1999).
25. Ben-Gal, I. *et al.* Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics* **21**, 2657–2666 (2005).
26. Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).
27. Gardner, P.P. *et al.* Rfam: updates to the RNA families database. *Nucleic Acids Res.* **37**, D136–D140 (2009).
28. Wilm, A., Mainz, I. & Steger, G. An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol. Biol.* **1**, 19 (2006).
29. Zhou, Y. *et al.* GISSD: Group I Intron Sequence and Structure Database. *Nucleic Acids Res.* **36**, D31–D37 (2008).
30. Weinberg, Z. *et al.* Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. *Nature* **462**, 656–659 (2009).
31. Weinberg, Z. *et al.* Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.* **11**, R31 (2010).
32. Vitreschak, A.G. Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. *RNA* **9**, 1084–1097 (2003).
33. Lee, H. *et al.* *cis*-acting RNA signals in the NS5B C-terminal coding sequence of the Hepatitis C virus genome. *J. Virol.* **78**, 10865–10877 (2004).
34. Le Quesne, J.P. *et al.* Derivation of a structural model for the c-myc IRES. *J. Mol. Biol.* **310**, 111–126 (2001).
35. Paul, AV. *et al.* Identification of an RNA hairpin in poliovirus RNA that serves as the primary template in the *in vitro* uridylylation of VPg. *J. Virology* **74**, 10359–10370 (2000).
36. Wassarman, K.M. *et al.* Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.* 1637–1651 (2001).
37. Bösl, M. & Kersten, H. A novel RNA product of the tyrT operon of *Escherichia coli*. *Nucleic Acids Res.* **19**, 5863–5870 (1991).
38. Weinberg, Z. *et al.* Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res.* **35**, 4809–4819 (2007).
39. Klinck, R. *et al.* A potential RNA drug target in the hepatitis C virus internal ribosomal entry site. *RNA* **6**, 1423–1431 (2000).
40. Gao, Y.-G. *et al.* The structure of the ribosome with elongation factor G trapped in the posttranslocational state. *Science* **326**, 694–699 (2009).
41. Jang, S.B. *et al.* Structures of two RNA octamers containing tandem G.A base pairs. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 829–835 (2004).

ONLINE METHODS

Data sources. All crystal structures were obtained from the Protein Data Bank (PDB)[42]. The alignments of the bacterial ribosome (both subunits) are from ref. 2; archaeal ribosomal large subunit data were obtained both from ref. 2 and the comprehensive ribosomal RNA databases (Silva) version 102 (ref. 43), the later was manually corrected in the regions corresponding to the studied modules; Rfam alignments correspond to version Rfam 9.1 and were downloaded from http://rfam.sanger.ac.uk/[27], group I intron alignments were downloaded from group I intron sequence and structure database (GISSD)[29]; and the new bacterial RNA alignments are from the supplementary information in references 30,31. All 2D diagrams were produced using the visualization applet for RNA secondary structure (VARNA)[44].

Design of the Bayesian networks. The design of a Bayesian network is a two-step process. First, the network topology is established, that is, the set of dependencies between all variables of the model is determined. Second, the parameters describing the probability distribution of each node based on the observed data and specified dependencies are computed.

In the present case the Bayesian network topology closely follows the established interaction networks. All Watson-Crick (WC) base pairs, most of the non-WC base pairs and some base stacking interactions will map to edges of the Bayesian network. Interactions involving a fully conserved nucleotide were not included because they would not add any information to the Bayesian network. Some additional edges were included that connect structurally important but less conserved bases to bulged bases (**Supplementary Fig. 4**).

A multinomial distribution, corresponding to the occurrence probability of the four nucleotides and a gap, is associated to each node of the Bayesian network. In the case of dependent nodes, the local distribution is conditioned by the parent nodes distributions. The parameters were estimated by maximum likelihood using the sequence alignments of each module as the observations (**Supplementary Fig. 5**). Given the high number of observations (5,735 observations for G-bulge, 7,677 observations for kink-turns and 3,545 observations for C-loops) the parameters correspond to the relative frequency of each nucleotide in the full sample[45]. As the different alignments have different sequence frequencies, counts were normalized so that all alignments would contribute equally to the final count. The tandem-GA module was an exception to the above method as the parameters were not computed from sequence alignment data but were defined based on an ideal tandem GA. For this module, the first base of each WC base pair has a distribution identical to the nucleotide content of the sequence, and the second base had a conditional probability of $P(U|A) = P(G|C) = 1.0$; $P(A|U) = P(C|G) = 0.6$; and $P(G|U) = P(U|G) = 0.4$. The two non-WC base pairs were invariant with probabilities $P1(A) = P2(G) = 1.0$.

Finally, each WC base pair was classified as mandatory or optional. This information is used to compute the base pair probabilities and mutual information when filtering candidates.

Interaction networks analysis for parameter estimation. We analyzed 11 G-bulge modules. Computed nucleotide frequencies confirmed previous predictions by isostericity analysis[4]. Four base pairs were invariant in all occurrences of the module: the bulged G-U *cis* Hoogsteen/sugar edge, the A-G *trans* Hoogsteen/sugar edge, the U-A *trans* WC/Hoogsteen and the A-A *trans* Hoogsteen/Hoogsteen. In the remaining positions some variation was allowed (**Supplementary Fig. 6**).

We analyzed 14 kink-turn instances and grouped them into families based on their interaction network similarities (**Supplementary Fig. 7**). To obtain a consensus interaction network, we excluded the instances KT-16S-11-P, the interaction network of which is unique and too divergent from all other families; and KT-23S-15-A and KT-23S-58-A, which present atypical nucleotide insertions in the short strand (in the abbreviations, KT is kink turn; 23S or 16S indicate large or small ribosomal subunit; 11, 15 or 58 are helix 11, 15 or 58; and A or P indicate archaeal or bacterial rRNA alignment).

Seven analyzed C-loops revealed an invariant core formed by two crossing, noncanonical interactions pairing the first and last bases of the loop with the bases of the flanking base pairs in the opposite strand. Despite this interaction regularity the C-loop presents big sequence variation, except for the first base of the loop, invariably a C, and the third base of the loop, either a A or a C with the same frequency (**Supplementary Fig. 8**).

Single sequence search algorithm. A formal definition of the single sequence search algorithm used by RMDetect (**Fig. 2**) can be stated as: let M be a structural RNA module; S be a nucleotide sequence to be searched for M; M_{BN} a Bayesian network model of M; M_{GC} a null model in which all the bases are independent and have the same nucleotide distribution of S; $sp_{ij} = \{seq_i, seq_j\}$ a pair of non-overlapping subsequences of S starting from positions i and j, corresponding to the strands of the module; and WC_{M} be the set of all WC base pairs from M. For simplicity, we will describe only modules formed by pairs of subsequences, that is, modules with two strands. The extension to modules formed by more than two strands, as in n-way junctions[2], would simply require redefinition of sp as a tuple $sp_{i1,\ldots,in} = (seq_{i1}, \ldots, seq_{in})$.

For each sp_{ij} compute the corresponding score$_{ij}$:

$$\mathrm{score}_{ij} = \log_2\left(\frac{P\left(sp_{ij} \mid M_{\mathrm{BN}}\right)}{P\left(sp_{ij} \mid M_{GC}\right)}\right);$$

For each sp_{ij} compute bpp$_{ij}$, the corresponding joint base pair probability of all WC base pairs:

$$\mathrm{BPP}_{ij} = \frac{e^{-\frac{\mathrm{Ens.\ FE}_{ij}}{kT}}}{e^{-\frac{\mathrm{Ens.\ FE}_{\mathrm{all}}}{kT}}},$$

in which Ens. FE stands for the free energy of a folding ensemble, Ens. FE$_{\mathrm{all}}$ corresponds to the folding of the unconstrained original sequence, and Ens. FE$_{\mathrm{ij}}$ corresponds to the folding of the original sequence constrained by the base pairs of (WC_M) in the positions determined by sp_{ij}.

Select all sp_{ij} with score$_{ij}$ and bpp$_{ij}$ higher than a given threshold. These will be considered the candidates for the module considered.

The single sequence search was performed with a window length of 150 nt and a window step of 75 nt. All candidates scoring less than the specified score and BPP values (**Table 1**) were discarded, and the remaining ones were retained as candidates. We considered a candidate as true positive (TP) if it occurred in the same sequence positions as the known module instances (plus or minus two positions to account for unexpected gaps and alignment errors) all the other candidates are considered false positives (FP). The free energies of the ensembles were computed with 'RNAfold -p'[46] (-p to calculate the partition function) to obtain Ens. FE_{all} (the unconstrained FE) and RNAfold -p –C' (–C to calculate structures subject to constraints) to obtain Ens. FE_{motif} (the constrained FE). The parameters used to compute the joint base pair probabilities where $T = 274.5K$ and $k = 1.98717 \times 10^{-3}$ kcal mol^{-1} (from Vienna package source code). The algorithm performance scaled linearly with sequence length (for a fixed window length) and scaled quadratically with window length (**Supplementary Notes 4** and **5** and **Supplementary Figs. 9** and **10**).

Multiple sequence search algorithm. As seen in the single sequence search algorithm, each module candidate can be defined by an ordered pair of alignment coordinates $cand_{ij} = (seq_i, seq_j)$. A hierarchical clustering algorithm was applied to group the candidates of the different sequences according to their distance. The algorithm is as follows. (i) Remove all overlapping candidates on the same sequence retaining only the one with the higher score. (ii) Each candidate, $cand_{ij}$, will be assigned to the cluster, $cluster_{ij}$, centered at the position (i,j). (iii) Merge all pairs of clusters for which dist $(cluster_{ij}, cluster_{kl}) <$ DLIMIT, where dist $(cluster_{ij}, cluster_{kl}) = \max(|i - k|, |j - l|)$. Notice that i, j, k and l are columns of the alignment and DLIMIT is the maximum tolerated column distance between two candidates so that they can be considered to belong to the same cluster. (iv) Recompute the center of each cluster as the most represented position (i,j). (v) Repeat from (iii) until no more clusters are merged.

At the end, each cluster will correspond to a module candidate characterized by five measures: (i) absolute number of the aligned sequences in which the candidate occurs (sequence count); (ii) percentage of aligned sequences in which the candidate occurs (occurrence); (iii) mean score of all candidates; (iv) mean BPP of all candidates; and (v) mutual information (MI) between the bases of each WC base pair from WC_M, measured as along all candidates[47] (**Supplementary Note 6**). Thus, for a cluster to be considered it must be sufficiently represented in the alignment, must have a score and BPP higher than the defined threshold and should have covariance between WC base pairs, supporting the evolutionary pressure on conservation of the secondary structure of the module (**Fig. 2b**).

The multiple sequence search algorithm, described above, produced a set of clusters that was filtered according to the following conditions: (i) (sequence count > 2) and (occurrence ≥ 10%); (ii) MI > 0 or (occurrence > 33% and sequence_count > 10); (iii) score ≥ limit_score; and (iii) BPP ≥ limit_BPP.

Both limit_score and limit_BPP vary across the models. Limit_score was 8.0 for G-bulge, 11.0 for kink-turn, 16.0 for C-loop and 9.0 for tandem-GA. limit_bpp is 0.1 for the tandem GA, 0.01 for the C-loop and 0.001 for all other models. These values were chosen as they allowed the detection of at least half of the modules in all but one single sequence search test case (**Table 2**). The DLIMIT distance, discussed above, was set to five columns.

At the end of this process each selected cluster corresponds to a module prediction that was manually validated according to the compatibility with published structure, sequence alignment or co-variation information obtained from the alignment.

Test cases for known modules. Fifteen test cases were generated each corresponding to one module and one alignment (**Supplementary Table 1**). For each test case, the original alignment was randomly split in one training set and one test set. The training set was used to compute model parameters, and the test set was used for the search. The training set was then augmented with sequences from the other alignments containing the searched module. As a negative control, each sequence of the test set was duplicated and shuffled to preserve the nucleotide composition of the sequence. Single sequence and multiple sequence search algorithms were performed in each test set as described above. For example, when searching for the G-bulge module in the 16S rRNA sequences, the training set was composed by 523 randomly selected sequences of the 16S rRNA alignment plus all 6,956 sequences from 23S bacterial rRNA, 23S archaea rRNA and lysine-riboswitch alignments. The test set included the remaining 250 sequences of the 16S rRNA alignment plus 250 shuffled sequences. Both algorithms were applied as described above.

Search in database alignments. We systematically searched 1,309 RFam families, 14 group I intron alignments and 121 alignments of structured ncRNAs from meta-genomic data[30,31]. The Rfam alignments with more than 7,000 sequences were reduced to shorter versions containing 500 randomly selected sequences from the original alignment. The group I intron alignments were converted to Stockholm format. All alignments were searched 'as-is' with no realignment or manual adjustments. The following alignments were excluded from the search: the U4 snRNA, all small and large subunit rRNAs (4 families) and the SAM riboswitch that were used for training; the group I intron alignment that were searched in specific databases; the tRNA familiy; And all the families with less than five sequences (56 families).

Implementation and software availability. The described algorithms were implemented as a set of python scripts publicly available as open source from: http://sourceforge.net/projects/rmdetect/. A user guide is provided (**Supplementary Note 3**).

42. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
43. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**, 7188–7196 (2007).
44. Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–1975 (2009).
45. Durbin, R. *et al. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, 1998).
46. Hofacker, I.L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie* **125**, 167–188 (1994).
47. Lindgreen, S., Gardner, P.P. & Krogh, A. Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics* **22**, 2988–2995 (2006).

doi:10.1038/nmeth.1603