ARTICLES
PUBLISHED ONLINE: 30 OCTOBER 2011 | DOI: 10.1038/NCHEM.1176
nature chemistry

A two-dimensional mutate-and-map strategy for non-coding RNA structure

Wipapat Kladwang[1], Christopher C. VanLang[2], Pablo Cordero[3] and Rhiju Das[1,3,4]*

Non-coding RNAs fold into precise base-pairing patterns to carry out critical roles in genetic regulation and protein synthesis, but determining RNA structure remains difficult. Here, we show that coupling systematic mutagenesis with high-throughput chemical mapping enables accurate base-pair inference of domains from ribosomal RNA, ribozymes and riboswitches. For a six-RNA benchmark that has challenged previous chemical/computational methods, this 'mutate-and-map' strategy gives secondary structures that are in agreement with crystallography (helix error rates, 2%), including a blind test on a double-glycine riboswitch. Through modelling of partially ordered states, the method enables the first test of an interdomain helix-swap hypothesis for ligand-binding cooperativity in a glycine riboswitch. Finally, the data report on tertiary contacts within non-coding RNAs, and coupling to the Rosetta/FARFAR algorithm gives nucleotide-resolution three-dimensional models (helix root-mean-squared deviation, 5.7 Å) of an adenine riboswitch. These results establish a promising two-dimensional chemical strategy for inferring the secondary and tertiary structures that underlie non-coding RNA behaviour.

The transcriptomes of living cells and viruses continue to reveal novel classes of non-coding RNA (ncRNA) with critical functions in gene regulation, metabolism and pathogenesis[1–3]. The functional behaviours of these molecules are intimately tied to specific base-pairing patterns, and these patterns are challenging to identify using existing strategies based on phylogenetic analysis[4,5], nuclear magnetic resonance (NMR)[6,7], crystallography[8–14], molecular rulers[15,16] or functional mutation/rescue experiments[17,18]. A more facile approach to the characterization of RNA structure involves high-throughput chemical mapping at single-nucleotide resolution. This method is applicable to RNAs as large as the ribosome as well as entire viruses, both *in vitro* and in their cellular milieu[19–21]. Measurements of the accessibility of every nucleotide to solution chemical modification can guide or filter structural hypotheses from computational models[22,23]. Nevertheless, approximations in computational models and in correlating structure to chemical accessibility limit the inherent accuracy of this approach[22–25].

This Article presents a strategy to expand the information content of chemical mapping by means of a two-dimensional 'mutate-and-map' methodology[26]. Here, sequence mutation acts as a second dimension in a manner analogous to initial perturbation steps in multidimensional NMR pulse sequences for structure determination[7] or pump–probe experiments in other spectroscopic fields[27]. Based on elegant precedents in group I intron studies[28,29], we reasoned that if one nucleotide involved in a base pair is mutated, its partner might become more exposed and thus be readily detectable by chemical mapping. In practice, some mutations might not lead to the desired 'release' of the pairing partners, and some mutations might produce larger perturbations, such as the unfolding of an entire helix. Nevertheless, if even a subset of the probed mutations leads to precise release of interacting nucleotides, the base-pairing pattern of the RNA could potentially be read out from this extensive data set. Indeed, our recent proof-of-concept studies have demonstrated systematic inference of Watson–Crick base pairs in a 20-base-pair DNA/RNA duplex[26] and a 35-nucleotide RNA hairpin[30]. However, these artificial systems were designed to include single long helices and thus may not adequately represent natural, functional non-coding RNAs with many shorter helices, extensive non-canonical interactions and multiple solution states.

We therefore sought to apply the mutate-and-map strategy to a diverse set of non-coding RNAs with available crystal structures for some states and unknown structures for other states. The benchmark, which includes ribozymes, riboswitches and ribosomal RNA domains (Supplementary Table S1), is challenging; an earlier (one-dimensional) chemical/computational approach missed and mispredicted ~20% of the benchmark's helices[24]. We can report that our mutate-and-map strategy achieves 98% accuracy in inferring Watson–Crick base-pairing patterns and gives useful confidence estimates through bootstrap analysis. Furthermore, the method permits the generation and falsification of structural hypotheses about partially ordered RNA states, as highlighted by new results on a glycine-sensing riboswitch. We focus predominantly on the basic but unsolved problem of RNA secondary structure inference[24,25,31] from biochemical data. Extensions of the method and advances in computational modelling may permit robust tertiary contact inference and three-dimensional models, and we present one such case as a proof-of-concept.

Results

Proof-of-concept for an adenine-binding riboswitch. We first established the information content and accuracy of the strategy for the 71-nucleotide adenine-sensing *add* riboswitch from *Vibrio vulnificus*, which has been studied extensively[6,17,32–35] and solved in the adenine-bound state using crystallography[9]. The *add* secondary structure is incorrectly modelled by the *RNAstructure* algorithm alone, but can be recovered through the inclusion of standard one-dimensional SHAPE (selective 2′ hydroxyl acylation with primer extension) data[24]; this RNA therefore serves as a well-characterized control. We prepared 71 variants of the RNA,

[1]Department of Biochemistry, Stanford University, Stanford, California 94305, USA, [2]Department of Chemical Engineering, Stanford University, Stanford, California 94305, USA, [3]Program in Biomedical Informatics, Stanford University, Stanford, California 94305, USA, [4]Department of Physics, Stanford University, Stanford, California 94305, USA. *e-mail: rhiju@stanford.edu

mutating each base to its complement using high-throughput polymerase chain reaction (PCR) assembly, *in vitro* transcription and magnetic bead purification methods in a 96-well format, as discussed previously[30]. SHAPE data for all mutants were collected in a single afternoon. For a detailed view, Fig. 1a compares electrophoretic traces for the starting sequence and the C18G variant in the presence of 5 mM adenine. For a global view, the electropherograms for all constructs are given in Fig. 1b. As in earlier studies[26,30], *Z*-scores (Fig. 2a, number of standard deviations from the mean accessibility; see Methods) highlight the most significant features of the data.

As with simpler model systems, the *add* mutate-and-map data demonstrate perturbations near mutation sites (marked at nucleotide 18 in Fig. 1a; distinct diagonal stripes (I) in Fig. 1b). More importantly, the data show numerous features corresponding to interacting pairs of sequence-separated nucleotides. For example, mutation C18G led to increased exposure of nucleotides 74–79, with the strongest effect at G78 (Fig. 1a; marked II in Fig. 1b). This observation

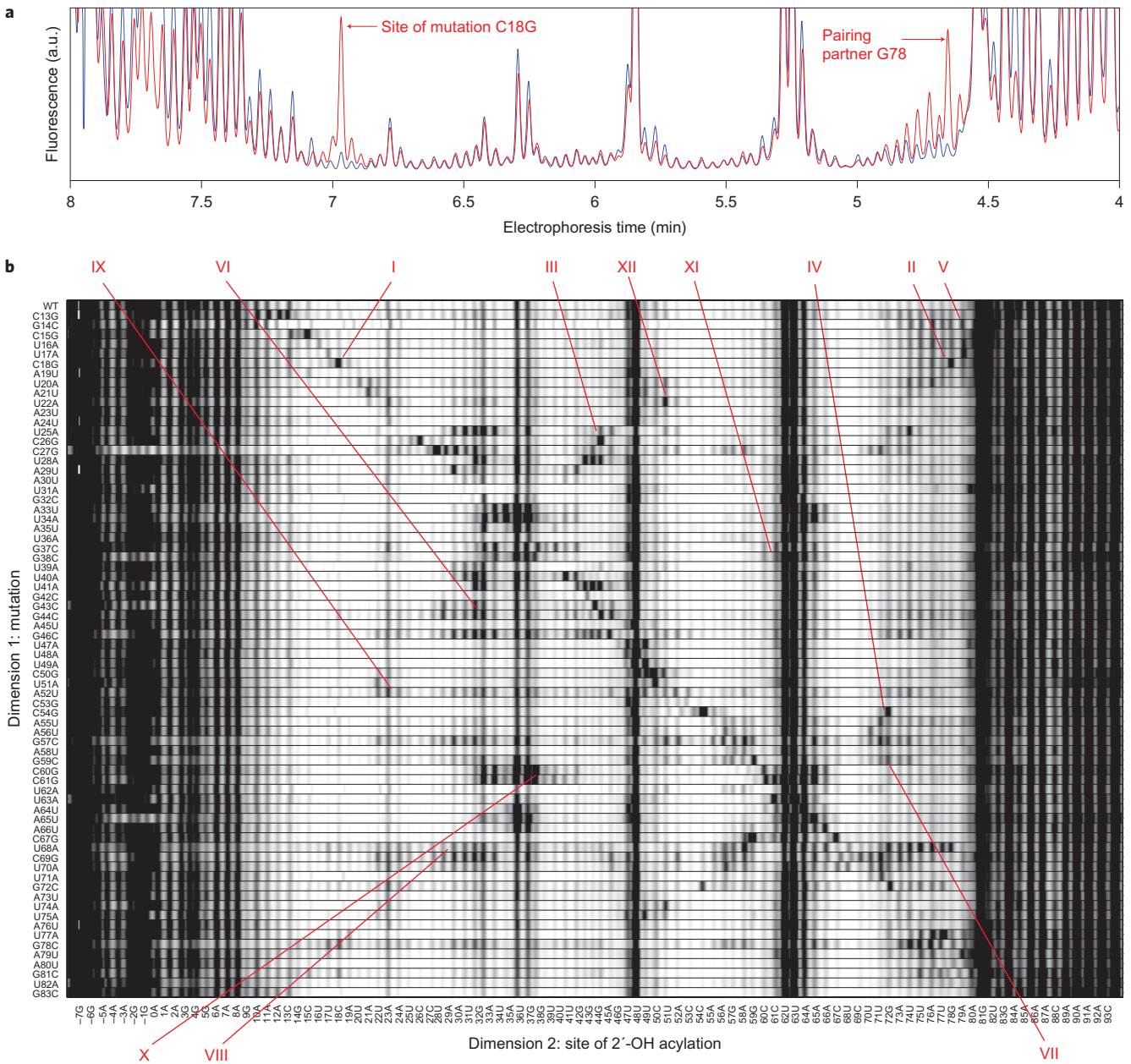


Figure 1 | The mutate-and-map method gives an information-rich picture of RNA structure. a, Mutating a nucleotide and mapping chemical accessibility reveals interactions in the three-dimensional structure of the RNA. The traces are for wild-type (blue) and C18G-mutated (red) variants of the adenine-binding domain of the *add* riboswitch. These 2′-OH acylation (SHAPE) data were read out by reverse transcription with fluorescently labelled primers and capillary electrophoresis; peaks (left to right) correspond to nucleotides from the 5′ to 3′ end of the RNA. Arrows mark exposure of the mutation site (C18) and of sequence-distant regions brought near this nucleotide by base-pairing (partner G78). **b**, Entire mutate-and-map data set across 71 single mutations, plotted in grey scale, revealing numerous elements of riboswitch structure. Dark features highlight: (I) the main diagonal stripe showing localized perturbations following C18G mutation; (II–IV) punctate features marking base pairs C18–G78, C26–G44 and C54–G72 in three different helices; (V–VII) more delocalized effects upon helix mutations G14C, G44C and G59C; (VIII) large-scale changes from C69G mutation due to secondary structure rearrangement; (IX) perturbations consistent with loss of adenine binding in A52U variant; (X) evidence for long-range tertiary contact between L2 and L3 upon mutation of C60 and C61 in L3; (XI) 'symmetric' mutations in L2 that affect L3; (XII) evidence for U22–A52 base pair in the adenine binding site.

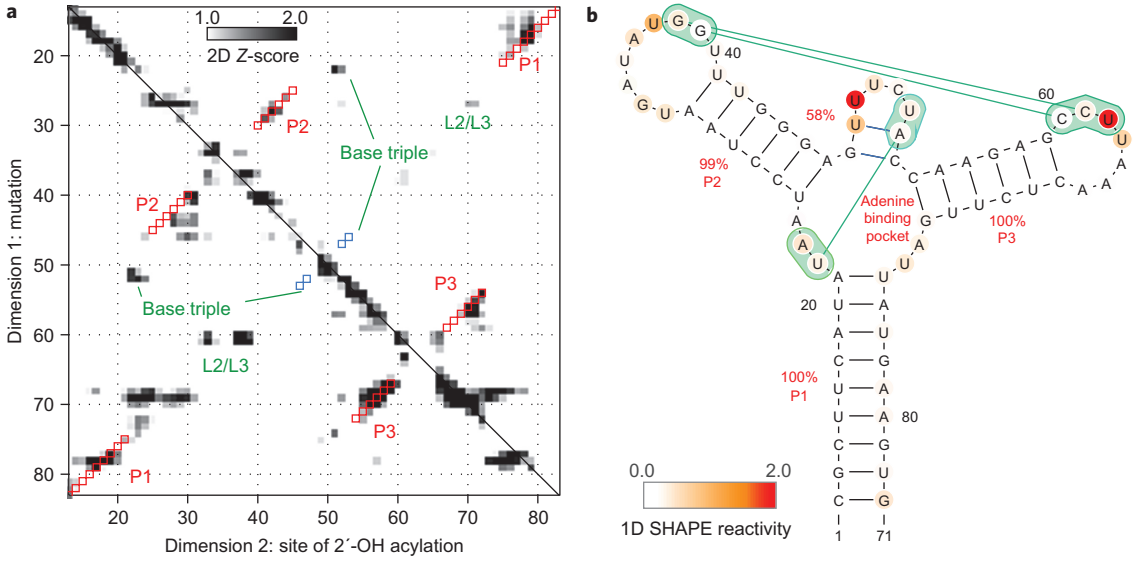


Figure 2 | Mutate-and-map data and secondary structure. a, Strong features of mutate-and-map data isolated by *Z*-score analysis (number of standard deviations from mean at each residue). Squares show secondary structure model guided by mutate-and-map data (red, match to crystallographic Watson–Crick stems; blue, match to non-Watson–Crick stem). **b**, Secondary structure derived from incorporating *Z*-scores into the *RNAstructure* modelling algorithm; bootstrap confidence estimates given as red percentage values. Additional tertiary contacts inferred from a separate clustering analysis are given in green. Nucleotides are coloured according to SHAPE reactivity.

strongly supports a C18–G78 base pair in the P1 helix. Such 'punctate' single-nucleotide-resolution base-pair features are also visible for the other helical stems of this RNA (for example, C26–G44 and C54–G72, marked III and IV in Fig. 1b). Several mutations led to more delocalized perturbations (V–VII, Fig. 1b; stems marked in Fig. 2a) due to disruption of multiple consecutive base pairs. Although not punctate, these features confirm interactions at helix-level resolution. Some single mutations produce larger-scale changes, reflecting shifts in the secondary structure (for example, C69G, VIII in Fig. 1b; see also U25A, G46C) or loss of adenine binding (for example, A52U, IX in Fig. 1b; see also Supplementary Figs S1 and S2). Interestingly, within each helix, some mutations gave strong signals, whereas others led to minimal perturbations (for example, the 3′ segment of P2 in Fig. 2a)[26,30], underscoring the need to survey all possible mutations.

To assess the predictive power of these data for structure modelling, we applied the measured *Z*-scores as energetic bonuses in the *RNAstructure* secondary structure prediction algorithm. We further estimated confidence values for all inferred helices by bootstrapping the mutate-and-map data and repeating the secondary structure calculation[36]. As expected from the visual analysis above, the crystallographic secondary structure was robustly recovered (Fig. 2a,b), with ≥99% bootstrap values for helices P1, P2 and P3. An 'extra' two-base-pair helix was also found, with a weak bootstrap value (58%); these nucleotides are in fact base-paired in the *add* riboswitch crystallographic model[9], but one pairing is a non-canonical Watson–Crick/Hoogsteen pair and part of a base triple. Together with additional mutate-and-map signals (X–XII, Fig. 1b), these data were sufficient for determining the global tertiary fold of the RNA, as described in the following.

A challenging benchmark of base pair inference. To complete our benchmark of the mutate-and-map strategy, we applied the method to RNAs for which base-pairing patterns have been more challenging to recover. The smallest of these, unmodified $tRNA^{phe}$ from *Escherichia coli*[10], offers a simple illustration of the information content of the new method (Fig. 3a). The *RNAstructure* algorithm mispredicted two of the four helices of the tRNA 'cloverleaf' (the D and anticodon helices; cf. Fig. 3b,c). Inclusion of one-dimensional SHAPE reactivities corrected these errors, but introduced an additional error, mispredicting the TψC helix (Fig. 3d); protection of the loop of this helix by tertiary contacts renders its modelling uncertain with one-dimensional data alone. The mutate-and-map SHAPE data for this tRNA (Fig. 3a) gave clear signals for all four helices. Applying the two-dimensional mutate-and-map data set to *RNAstructure* corrected the inherent inaccuracies of the algorithm and recovered the entire four-helix secondary structure (>99% bootstrap values; Fig. 3e). One additional edge base pair was predicted for the anticodon arm; this and other fine-scale errors are discussed in the following.

The remaining RNAs in our benchmark exceeded 100 nucleotides in length. As in the $tRNA^{phe}$ case, earlier chemical/computational methods assigned incorrect secondary structures to these sequences, but the mutate-and-map strategy led to accurate base-pairing patterns. The mutate-and-map data for a widely studied model RNA, the P4–P6 domain of the group I *Tetrahymena* ribozyme, gave visible features corresponding to all helices in the RNA[14] (Fig. 4a) and led to correct recovery of the secondary structure (Fig. 4b). One of the helices, P5c, was correctly modelled but with a weak bootstrap value (48%); this low score is consistent with conformational fluctuations in P5c identified in previous biochemical and NMR studies[37,38].

As a more stringent test of the mutate-and-map strategy, we applied the method to the *E. coli* 5S ribosomal RNA, a notable problem case for earlier chemical/computational approaches[22,39]. In particular, the segments around the non-canonical loop E motif have been mispredicted in all previous studies, including the most recent (one-dimensional) SHAPE-directed approach[24]. By providing pairwise information on interacting nucleotides (Fig. 4c), the mutate-and-map method recovered the entire secondary structure with high confidence (>90%; Fig. 4d). One extra helix (blue in Fig. 4d) corresponds to a segment that in fact forms non-canonical base pairs within the loop E motif.

The ligand-binding domain of the cyclic di-GMP riboswitch from *Vibrio cholerae* provided an additional challenge; helix P1 of this RNA was not found in the original phylogenetic analysis[18],

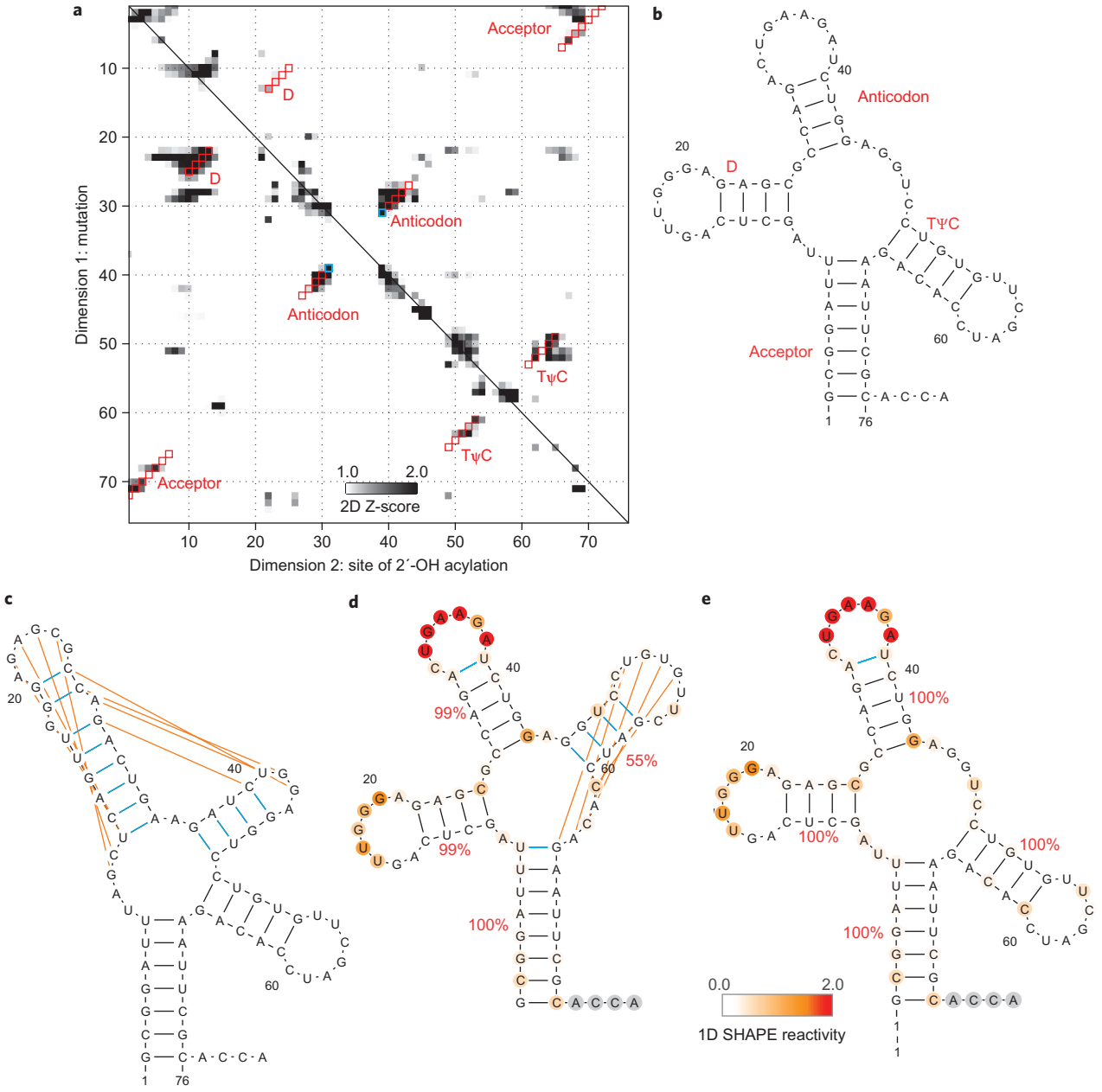


Figure 3 | Comparison of chemical/computational modelling approaches on tRNA^{phe}. **a**, Mutate-and-map *Z*-score data for tRNA^{phe} from *E. coli*. **b**–**e**, Secondary structure models of this RNA from crystallography (**b**), the *RNAstructure* algorithm without data (**c**), calculations guided by one-dimensional SHAPE data (**d**) and calculations guided by the two-dimensional mutate-and-map data (**e**). Red squares (**a**) give Watson–Crick base pairs from the mutate-and-map model that match the crystallographic secondary structure. Blue squares (**a**) or lines (**b**–**e**) give model Watson–Crick base pairs not present in the crystallographic secondary structure. Orange lines give crystallographic Watson–Crick base pairs missed in each model. Helix confidence estimates from bootstrapping one-dimensional (**d**) or two-dimensional (**e**) data are given as red percentage values; nucleotides are coloured according to SHAPE reactivity.

but was instead later revealed by crystallography. Based on measurements in the presence of 10 μM ligand, the mutate-and-map strategy (Fig. 4e) recovered nearly the entire secondary structure (7 of 8 helices), including P1 (Fig. 4f).

Blind prediction for the glycine riboswitch. As a final rigorous test, we acquired mutate-and-map data for an RNA for which a crystallographic model was not available at the time of modelling: the ligand-binding domain of the glycine-binding riboswitch from *Fusobacterium nucleatum*[40,41]. The mutate-and-map data in the presence of 10 mM glycine gave a secondary structure with nine helices (Fig. 5a); the model agreed with the nine helices that were identified by phylogeny. The secondary structure was confirmed by a crystallographic model released at the time of the submission of this Article[43].

Overall accuracy of secondary structure modelling. Overall, the mutate-and-map method demonstrated high accuracy in secondary structure inference for a benchmark of six diverse RNAs including 661 nucleotides in 42 helices (Table 1). As a baseline, an earlier method, using *RNAstructure* directed by one-dimensional SHAPE data, gave a false negative rate and false discovery rate of 17% and 21%, respectively, on this benchmark[24]. The mutate-and-map method recovered 41 of 42 helices, giving a sensitivity of 98% and a false negative rate of 2%, nearly an order of magnitude less than the previous method. The only missing helix was a two-base-pair

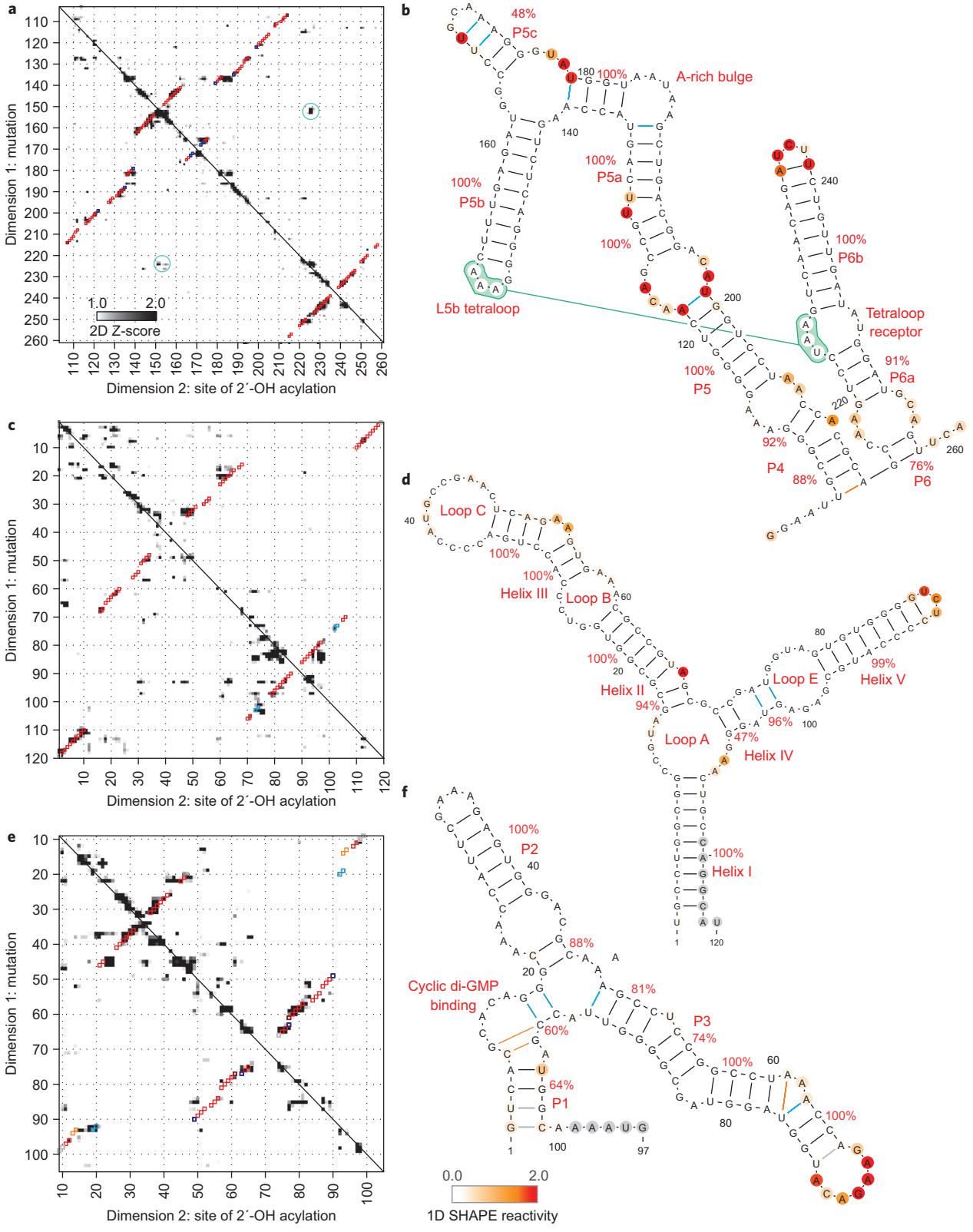


Figure 4 | Accurate secondary structure models for non-coding RNAs. a–f, Mutate-and-map *Z*-score data and resulting secondary structure models for the P4–P6 domain of the *Tetrahymena* group I ribozyme (**a**,**b**), the 5S ribosomal RNA from *E. coli* (**c**,**d**) and the domain that binds cyclic di-guanosine monophosphate from the *V. cholerae* VC1722 riboswitch (in the presence of 10 μM ligand; **e**,**f**). Colouring of squares (**a**,**c**,**e**) and lines and nucleotides (**b**,**d**,**f**) are as in Fig. 3.

helix in the cyclic diGMP riboswitch (see below). At a finer resolution, a small number (<6%) of base pairs in mutate-and-map calculated helices were either missed or added relative to the crystallographic secondary structures (1 and 11 of 197 base pairs, respectively; Supplementary Table S2). All these errors were either G–U or A–U pairs at the edges of otherwise correct helices (Figs 2– 5 and Table S2). Variation of the assumed coefficient of the two-dimensional *Z*-scores in the *RNAstructure* energy bonus

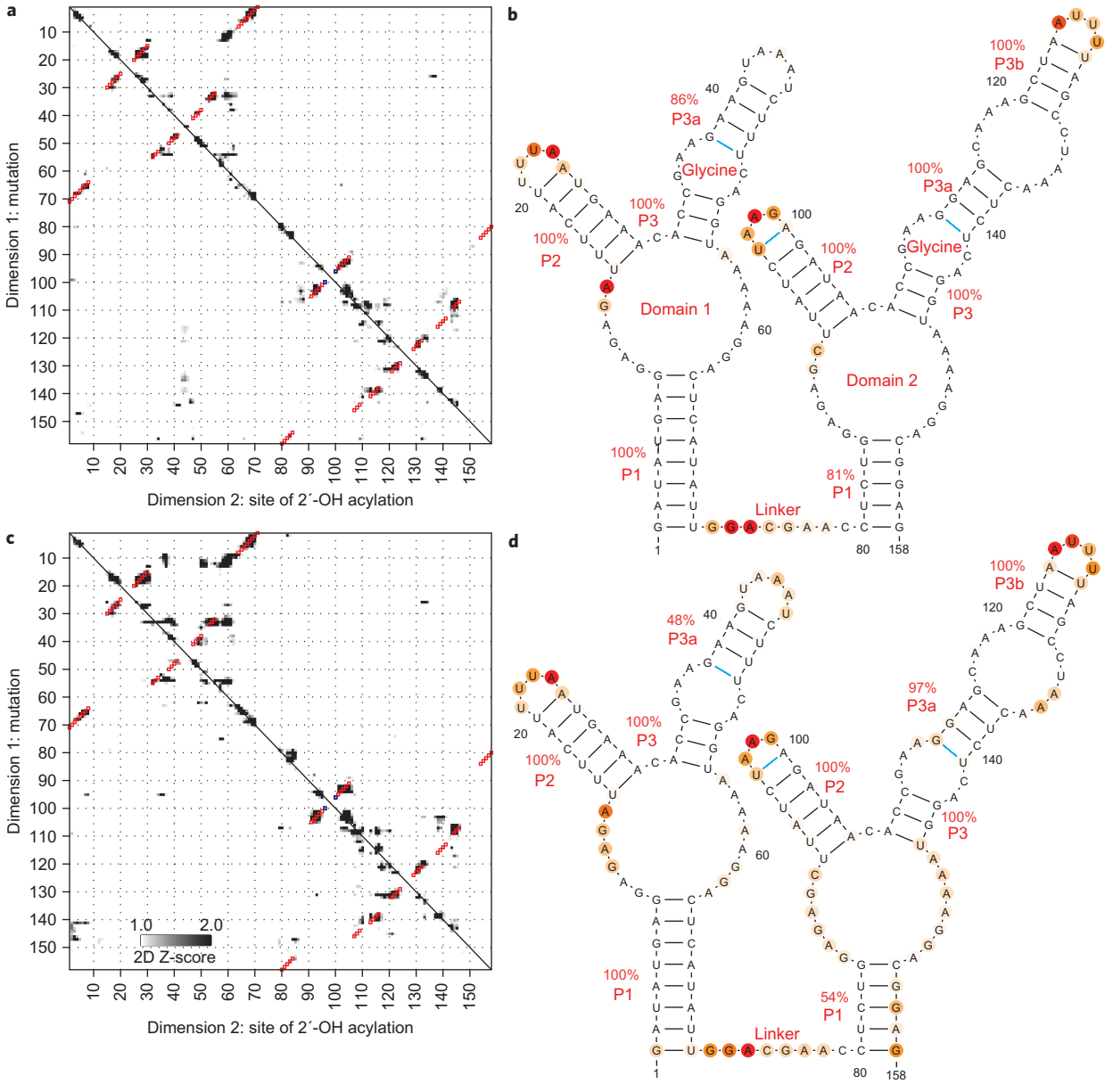


Figure 5 | Two states of a glycine-binding riboswitch. a–d, Mutate-and-map *Z*-score data and resulting secondary structure models for the double-ligand-binding domain of the *F. nucleatum* glycine riboswitch with 10 mM glycine (**a**,**b**) and without glycine (**c**,**d**), indicating no inter-domain helix swap upon glycine binding. Colouring of squares (**a**,**c**) and lines and nucleotides (**b**,**d**) are as in Fig. 3.

or addition/subtraction of an offset did not improve the recovery at the level of base pairs or helices (data not shown).

In terms of the false discovery rate, the mutate-and-map method gave only three extra helices, all of which were the smallest possible length (2 bp). As discussed above, two of these extra helices in fact correspond to non-canonical stems observed in crystallographic models. The remaining false helix gave a weak bootstrap value (60%) and may correspond to a stem sampled in the ligand-free conformation of the cyclic diGMP riboswitch (see below). The overall positive predictive value was 93–98% depending on whether the non-canonical helices are counted as correct. The false discovery rate was 2–7%, nearly an order of magnitude less than the earlier one-dimensional SHAPE-directed method (21%). Somewhat surprisingly, using both the one-dimensional SHAPE data and two-dimensional mutate-and-map data gave slightly worse accuracy than using the two-dimensional data alone (false negative rate of 7% compared with 2%); this result may reflect inaccuracies in interpreting absolute SHAPE reactivity, as opposed to *Z*-score changes in reactivity induced by mutations. We conclude that secondary structures derived from the mutate-and-map method are accurate (~2% error rates) for structured non-coding RNAs.

Testing an 'inter-domain helix swap' hypothesis for glycine riboswitch cooperativity. Beyond recovering known information about non-coding RNA secondary structure, we sought to generate or falsify novel hypotheses that would be difficult to explore using standard structural methods. The three riboswitch ligand-binding domains for adenine, cyclic di-GMP and glycine provide interesting test cases because their ligand-free states will generally be partially ordered and thus difficult to crystallize. First, application of the mutate-and-map strategy indicated that the secondary structure of the *add* riboswitch ligand-binding domain remains the same in adenine-free and adenine-bound states

Table 1 | Accuracy of RNA secondary structure models.

RNA	Length*	Number of helices†								
		Cryst.	No data		1D		1D + 2D		2D	
			TP	FP	TP	FP	TP	FP	TP	FP
Adenine riboswitch‡	71	3	2	3	3	0 (1)	3	0 (1)	3	0 (1)
$tRNA^{phe}$	76	4	2	3	3	1	4	0	4	0
P4–P6 RNA	158	11	10	1	9	2	9	2	11	0
5S rRNA	118	7	1	9	6	3	7	0 (1)	7	0 (1)
c-di-GMP riboswitch‡	80	8	6	2	6	2	7	1	7	1
Glycine riboswitch‡	158	9	5	3	8	1	9	0	9	0
Total	661	42	26	21	35	9 (10)	39	3 (5)	41	1 (3)
False negative rate§			38.1%		16.7%		7.1%		2.4%	
False discovery rate‖			44.7%		20.4 (22.2)%		7.1 (11.4)%		2.3 (6.8)%	

*Length of RNA in nucleotides. †Cryst, number of helices in crystallographic model; TP, true positive helices; FP, false positive helices; 1D, models using one-dimensional SHAPE chemical mapping data; 2D, models using mutate-and-map data. For FP, a helix was considered incorrect if its base pairs did not match the majority of base pairs in a crystallographic helix. Numbers in parentheses required that the matching crystallographic base pairs have Watson–Crick geometry. ‡Ligand-binding riboswitches were probed in the presence of small-molecule partners (5 mM adenine, 10 μM cyclic di-guanosine-monophosphate or 10 mM glycine). All experiments were carried out with 10 mM $MgCl_2$, 50 mM Na-HEPES, pH 8.0. §False negative rate = (Cryst–TP)/TP. ‖False discovery rate = FP/(FP + TP). Numbers in parentheses count matches of model base pairs to non-Watson–Crick crystallographic base pairs as false discoveries.

(Supplementary Fig. S3), consistent with biophysical data from other approaches[6,35]. In contrast, mutate-and-map data indicate that the cyclic di-GMP riboswitch shifts its secondary structure near P1 on ligand binding (Supplementary Fig. S4). This shift is potentially involved in the mechanism of the riboswitch[12,13,18] and may account for the weak phylogenetic signature of the P1 helix.

Among these 'non-crystallographic' targets, we were most interested in the glycine-binding riboswitch, which exhibits cooperative binding of two glycines to separate domains and is under intense investigation by several groups[40–44]. Analogous to the tense/relaxed equilibrium in the Monod–Wyman–Changeaux model for haemoglobin[45], we considered that cooperativity might stem from an inter-domain helix swap. In this model, an alternative ('tense') secondary structure involving non-native interactions between the two domains would be rearranged upon glycine binding. Although one-dimensional mapping experiments (Fig. 5 and refs 40–42) show changes in the RNA following glycine binding, these data are consistent with a range of secondary structures and do not provide stringent tests of the inter-domain swap model. Furthermore, the model is not easily testable by crystallography[43,44], which, if successful, is biased towards more structured conformations.

Application of the mutate-and-map strategy (Fig. 5b,d) gave a strong test of the hypothesis; the data in the absence of glycine gave the same domain-separated secondary structure as under conditions with glycine bound (Fig. 5a,c). Any changes in secondary structure for these constructs are thus either at edge base pairs or are negligible. We note that additional 5′ and 3′ flanking elements are likely to play critical roles in the modes of genetic regulation of these RNAs[2]; these longer segments are now under investigation.

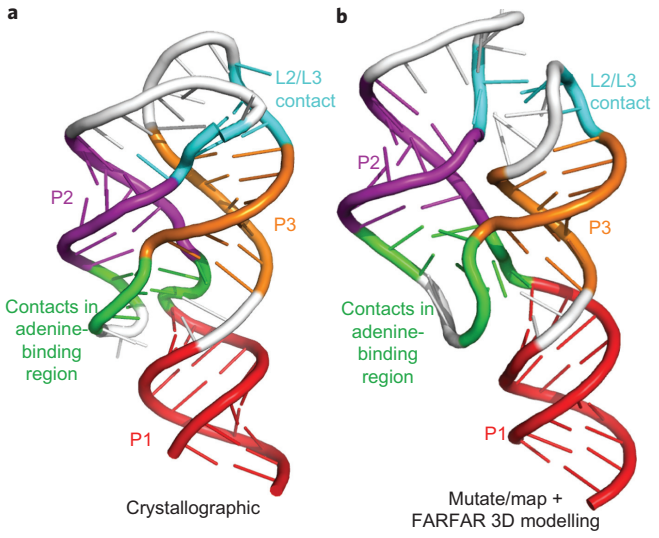


Figure 6 | Three-dimensional modelling from mutate-and-map data. **a**,**b**, Models of the (ligand-bound) adenine riboswitch derived from X-ray crystallography (**a**) and from the chemical/computational protocol introduced here (**b**). *De novo* modelling (using the Rosetta FARFAR algorithm) was carried out with secondary structure (P1, P2, P3) and tertiary contacts (L2/L3 and two contacts in adenine-binding region) inferred from solution mutate-and-map data. The mutate-and-map model agrees with the crystallographic model at nucleotide resolution (helix root-mean-squared deviation (RMSD) of 5.7 Å; overall RMSD of 7.7 Å).

Tertiary structure and cooperative fluctuations. The analysis described above focused on the first level of RNA structure, the Watson–Crick base-pairing pattern. Nevertheless, many non-coding RNAs use tertiary contacts and ordered junctions to position Watson–Crick helices into intricate three-dimensional structures. Qualitatively, we found evidence for numerous such tertiary interactions in the mutate-and-map data of these RNAs. For example, the *add* riboswitch is stabilized by tertiary interactions between the loops L2 (nucleotides 32–38) and L3 (nucleotides 60–66). In the presence of 5 mM adenine, mutations at G37 and G38 resulted in exposure of their partners C61 and C60 (X, Fig. 1b), and vice versa (XI, Fig. 1b; L2/L3 pseudoknot marked in Fig. 1b). Nevertheless, other mutations led to longer range effects (VIII, IX in Fig. 1b) due to cooperative unfolding of subdomains of tertiary structure or loss of adenine binding. For example, mutation of nucleotide A52 (VIII) gave chemical accessibilities that were different from the adenine-bound wild type RNA throughout the sequence, but consistent with the adenine-unbound state (Supplementary Fig. S1).

To extract tertiary base-pairing information, we could not use the *RNAstructure* method above, as it focuses on Watson–Crick base pairs. We therefore implemented filters enforcing strong, punctate signals and symmetry but not A–U, G–C or G–U pairing (see ref. 30, Methods and Supplementary Fig. S5). This analysis, independent of any computational models of RNA structure, recovered the majority of Watson–Crick helices in this benchmark. The analysis also recovered three tertiary contacts: the L2/L3 interaction of the *add* riboswitch (X and XI in Fig. 1b) and a U22–A52 base pair in the adenine binding pocket (XII, Fig. 1b); and a tetraloop/receptor interaction in the P4–P6 RNA (Fig. 4a,b). These features are accurate, but, in most test cases, their number is significantly less than the number of helices, precluding effective three-dimensional modelling. For the one case in which multiple tertiary contacts could be determined, the *add* adenine-sensing

riboswitch, we carried out three-dimensional modelling using the FARFAR *de novo* assembly method. The algorithm gave a structural ensemble with helix RMSD of 5.7 Å and overall RMSD of 7.7 Å to the crystallographic model[9] (Fig. 6a,b). This resolution is comparable to the average distance between nearest nucleotides (5.9 Å) and significantly better than model accuracy without mutate-and-map data (helix RMSD 8.9 Å; overall RMSD 16.9 Å) or values expected by chance ($P < 1 \times 10^{-3}$ for modelling a 71-nt RNA with secondary structure information[46]). These results on a favourable case suggest that rapid inference of three-dimensional structure for general RNAs might be achievable with other chemical probes that discriminate non-canonical interactions (for example, dimethyl sulfate[30,47] for A-minor interactions) or more sophisticated methods for mining tertiary information or ligand-binding sites from mutate-and-map data. We note also that features corresponding to cooperative changes in chemical accessibility, while not reporting on specific tertiary contacts, can reveal 'excited' states in the folding landscapes of the RNA that may be functional[48,49]. We are making the information-rich data sets acquired for this Article publicly available in the Stanford RNA Mapping Database (http://rmdb.stanford.edu) to encourage the development of novel analysis methods to explore tertiary contact extraction and landscapes.

Discussion

We have demonstrated that a mutate-and-map strategy permits the high-throughput inference of non-coding RNA base-pairing patterns. With error rates of ~2% and confidence estimates via bootstrapping, the method determines the secondary structures of riboswitch, ribosomal and ribozyme domains for which earlier chemical/computational approaches gave incorrect models. In addition to recovering known structures, the mutate-and-map data permit the rapid generation and falsification of hypotheses for structural rearrangements in three ligand-binding RNAs in partially ordered ligand-free states, including a cooperative glycine riboswitch with a poorly understood mechanism. Finally, the data yield rapid information on tertiary contacts of ncRNAs. Although not sufficient to yield crystallographic-quality structure models, in an adenine-sensing riboswitch, the data permit the modelling of the three-dimensional helix arrangement of the RNA at nucleotide resolution (5.7 Å). Further insights will come from detailed biophysical modelling of secondary and tertiary structure fluctuations induced by mutations; the public availability of these information-rich data sets should promote such analyses.

The mutate-and-map method only requires commercially available reagents, widely accessible capillary electrophoresis sequencers and freely available software. Further, each data set was acquired and analysed in a week or less. Therefore, for non-coding RNA domains up to ~300 nucleotides in length, the technology should be applicable as a front-line structural tool. The combined expense of the mutagenesis and mapping grows as the square of the RNA length. Thus, characterization of transcripts with thousands of nucleotides is presently challenging but may be facilitated by next-generation sequencing strategies[50].

Expanding experimental technologies from one to multiple dimensions has transformed fields ranging from NMR to infrared spectroscopy. We propose that the mutate-and-map strategy will be analogously enabling for chemical mapping approaches, permitting the confident secondary structure determination and tertiary contact characterization of non-coding RNAs that are difficult or intractable for previous experimental methods. Applications to full-length RNA messages *in vitro* or in extract, to complex ribonucleoprotein systems, and even to full viral RNA genomes appear feasible and are exciting frontiers for this high-throughput approach.

Methods

Mutate-and-map experimental protocol and data processing. Preparation of DNA templates, *in vitro* transcription of RNAs, SHAPE chemical mapping, and capillary electrophoresis were carried out in a 96-well format, accelerated through the use of magnetic bead purification steps, as has been described previously[26,30,51]. Data were analysed with the HiTRACE[52] software package, and *Z*-scores were computed in MATLAB. A complete protocol is given in the Supplementary Methods. Code for analysing mutate-and-map data is being made available as part of HiTRACE. *Z*-scores were used for secondary structure inference and sequence-independent feature analysis by single-linkage clustering, as described in the Supplementary Methods.

Secondary structure inference. The *Fold* executable of the *RNAstructure* package (v5.3) was used to infer secondary structures. The entire RNA sequences (Supplementary Table S1), including added flanking sequences, were used for all calculations. The flag '–T 297.15' set the temperature to match our experimental conditions (24 °C). The flags '–sh' and '–x' were used to input (one-dimensional) SHAPE data files and (two-dimensional) base-pair energy bonuses (equal to -1 kcal mol^{-1} times the *Z*-scores), respectively. In the *RNAstructure* implementation, the pseudoenergies were applied to each nucleotide forming an edge base pair, and doubly applied to each nucleotide forming an internal base pair[23]. Additional flags '–xs' and '–xo' permitted scaling and offset of the *Z*-score bonuses, but default values of 1.0 kcal mol^{-1} and 0.0 kcal mol^{-1}, respectively, were found to be optimal. For bootstrap analyses, mock SHAPE data replicates were generated by randomly choosing mutants with replacement[36]. The analysis is being made available as a server at http://rmdb.stanford.edu/structureserver. Secondary structure images were prepared in VARNA[53].

Assessment of secondary structure accuracy. A crystallographic helix was considered correctly recovered if more than 50% of its base pairs were observed in a helix by the computational model. (In practice, 40 of 41 such helices in models based on mutate-and-map data retained all crystallographic base pairs.) Helix slips of ± 1 were not considered correct (that is, the pairing (i,j) was not allowed to match the pairings $(i,j-1)$ or $(i,j+1)$).

Three-dimensional modelling with Rosetta. Three-dimensional models were acquired using the Fragment Assembly of RNA with Full Atom Refinement (FARFAR) methodology[51] in the Rosetta framework. Briefly, ideal A-form helices were created for each helix greater than two base pairs in length in the modelled secondary structure. Then, remaining nucleotides were modelled by FARFAR as separate motifs interconnecting these ideal helices, generating up to 4,000 potential structures. Finally, these motif conformations were assembled in a Monte Carlo procedure, optimizing the FARNA low-resolution potential and tertiary constraint potentials defined by the sequence-independent clustering analysis of mutate-and-map data. Runs without mutate-and-map data used the one-dimensional SHAPE-directed secondary structure (which agrees with crystallography for the *add* riboswitch) and constraints only for the two-base-pair non-canonical helix (G47–C54,U48–A53). Explicit command lines and example files are given in the Supplementary Information. The code, as well as a Python job-setup script *setup_rna_assembly_jobs.py* and documentation, are being incorporated into Rosetta release 3.4, which is freely available to academic users at http://www.rosettacommons.org. Before release, the code is available on request from the authors. The *P*-value for the *add* riboswitch was estimated by comparing the all-atom RMSD (7.7 Å) to the range expected by chance (13.5 ± 1.8 Å), as described in ref. 46.

Received 18 April 2011; accepted 15 September 2011; published online 30 October 2011

References

1. Yanofsky, C. The different roles of tryptophan transfer RNA in regulating trp operon expression in *E. coli* versus *B. subtilis*. *Trends Genet.* **20,** 367–374 (2004).
2. Winkler, W. C. & Breaker, R. R. Genetic control by metabolite-binding riboswitches. *Chembiochem* **4,** 1024–1032 (2003).
3. Zaratiegui, M., Irvine, D. V. & Martienssen, R. A. Noncoding RNAs and gene silencing. *Cell* **128,** 763–776 (2007).
4. Levitt, M. Detailed molecular model for transfer ribonucleic acid. *Nature* **224,** 759–763 (1969).
5. Lehnert, V., Jaeger, L., Michel, F. & Westhof, E. New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the *Tetrahymena thermophila* ribozyme. *Chem. Biol.* **3,** 993–1009 (1996).
6. Lee, M. K., Gal, M., Frydman, L. & Varani, G. Real-time multidimensional NMR follows RNA folding with second resolution. *Proc. Natl Acad. Sci. USA* **107,** 9192–9197 (2010).
7. Wuthrich, K. NMR studies of structure and function of biological macromolecules (Nobel lecture). *Angew. Chem. Int. Ed.* **42,** 3340–3363 (2003).
8. Cruz, J. A. & Westhof, E. The dynamic landscapes of RNA architecture. *Cell* **136,** 604–609 (2009).
9. Serganov, A. *et al.* Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem. Biol.* **11,** 1729–1741 (2004).
10. Byrne, R. T., Konevega, A. L., Rodnina, M. V. & Antson, A. A. The crystal structure of unmodified tRNAPhe from *Escherichia coli*. *Nucleic Acids Res.* **38,** 4154–4162 (2010).

11. Correll, C. C., Freeborn, B., Moore, P. B. & Steitz, T. A. Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell* **91,** 705–712 (1997).
12. Smith, K. D., Lipchock, S. V., Livingston, A. L., Shanahan, C. A. & Strobel, S. A. Structural and biochemical determinants of ligand binding by the c-di-GMP riboswitch. *Biochemistry* **49,** 7351–7359 (2010).
13. Kulshina, N., Baird, N. J. & Ferre-D'Amare, A. R. Recognition of the bacterial second messenger cyclic diguanylate by its cognate riboswitch. *Nature Struct. Mol. Biol.* **16,** 1212–1217 (2009).
14. Cate, J. H. *et al.* Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* **273,** 1678–1685. (1996).
15. Lemay, J. F., Penedo, J. C., Mulhbacher, J. & Lafontaine, D. A. Molecular basis of RNA-mediated gene regulation on the adenine riboswitch by single-molecule approaches. *Methods Mol. Biol.* **540,** 65–76 (2009).
16. Das, R. *et al.* Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proc. Natl Acad. Sci. USA* **105,** 4144–4149 (2008).
17. Mandal, M. & Breaker, R. R. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nature Struct. Mol. Biol.* **11,** 29–35 (2004).
18. Sudarsan, N. *et al.* Riboswitches in eubacteria sense the second messenger cyclic di-GMP. *Science* **321,** 411–413 (2008).
19. Culver, G. M. & Noller, H. F. *In vitro* reconstitution of 30S ribosomal subunits using complete set of recombinant proteins. *Methods Enzymol.* **318,** 446–460 (2000).
20. Adilakshmi, T., Lease, R. A. & Woodson, S. A. Hydroxyl radical footprinting *in vivo*: mapping macromolecular structures with synchrotron radiation. *Nucleic Acids Res.* **34,** e64 (2006).
21. Wilkinson, K. A. *et al.* High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol.* **6,** e96 (2008).
22. Mathews, D. H. *et al.* Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA* **101,** 7287–7292 (2004).
23. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA structure determination. *Proc. Natl Acad. Sci. USA* **106,** 97–102 (2009).
24. Kladwang, W., VanLang, C. C., Cordero, P. & Das, R. Understanding the errors of SHAPE-directed RNA modeling. *Biochemistry* **50,** 8049–8056 (2011).
25. Quarrier, S., Martin, J. S., Davis-Neulander, L., Beauregard, A. & Laederach, A. Evaluation of the information content of RNA structure mapping data for secondary structure prediction. *RNA* **16,** 1108–1117 (2010).
26. Kladwang, W. & Das, R. A mutate-and-map strategy for inferring base pairs in structured nucleic acids: proof of concept on a DNA/RNA helix. *Biochemistry* **49,** 7414–7416 (2010).
27. Cho, M. Coherent two-dimensional optical spectroscopy. *Chem Rev* **108,** 1331–1418 (2008).
28. Pyle, A. M., Murphy, F. L. & Cech, T. R. RNA substrate binding site in the catalytic core of the Tetrahymena ribozyme. *Nature* **358,** 123–128 (1992).
29. Duncan, C. D. & Weeks, K. M. SHAPE analysis of long-range interactions reveals extensive and thermodynamically preferred misfolding in a fragile group I intron RNA. *Biochemistry* **47,** 8504–8513 (2008).
30. Kladwang, W., Cordero, P. & Das, R. A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. *RNA* **17,** 522–534 (2011).
31. Shapiro, B. A., Yingling, Y. G., Kasprzak, W. & Bindewald, E. Bridging the gap in RNA structure prediction. *Curr. Opin. Struct. Biol.* **17,** 157–165 (2007).
32. Lemay, J. F., Penedo, J. C., Tremblay, R., Lilley, D. M. & Lafontaine, D. A. Folding of the adenine riboswitch. *Chem. Biol.* **13,** 857–868 (2006).
33. Rieder, R., Lang, K., Graber, D. & Micura, R. Ligand-induced folding of the adenosine deaminase A-riboswitch and implications on riboswitch translational control. *Chembiochem* **8,** 896–902 (2007).
34. Lemay, J. F. *et al.* Comparative study between transcriptionally- and translationally-acting adenine riboswitches reveals key differences in riboswitch regulatory mechanisms. *PLoS Genet.* **7,** e1001278 (2011).
35. Noeske, J. *et al.* An intermolecular base triple as the basis of ligand specificity and affinity in the guanine- and adenine-sensing riboswitch RNAs. *Proc. Natl Acad. Sci. USA* **102,** 1372–1377 (2005).
36. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall, 1998).
37. Wu, M. & Tinoco, I. Jr. RNA folding causes secondary structure rearrangement. *Proc. Natl Acad. Sci. USA* **95,** 11555–11560 (1998).
38. Vicens, Q., Gooding, A. R., Laederach, A. & Cech, T. R. Local RNA structural changes induced by crystallization are revealed by SHAPE. *RNA* **13,** 536–548 (2007).
39. Leontis, N. B. & Westhof, E. The 5S rRNA loop E: chemical probing and phylogenetic data versus crystal structure. *RNA* **4,** 1134–1153 (1998).
40. Mandal, M. *et al.* A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science* **306,** 275–279 (2004).
41. Lipfert, J. *et al.* Structural transitions and thermodynamics of a glycine-dependent riboswitch from *Vibrio cholerae*. *J. Mol. Biol.* **365,** 1393–1406 (2007).
42. Kwon, M. & Strobel, S. A. Chemical basis of glycine riboswitch cooperativity. *RNA* **14,** 25–34 (2008).
43. Butler, E. B., Xiong, Y., Wang, J. & Strobel, S. A. Structural basis of cooperative ligand binding by the glycine riboswitch. *Chem. Biol.* **18,** 293–298 (2011).
44. Huang, L., Serganov, A. & Patel, D. J. Structural insights into ligand recognition by a sensing domain of the cooperative glycine riboswitch. *Mol. Cell* **40,** 774–786 (2010).
45. Monod, J., Wyman, J. & Changeux, J. P. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **12,** 88–118 (1965).
46. Hajdin, C. E., Ding, F., Dokholyan, N. V. & Weeks, K. M. On the significance of an RNA tertiary structure prediction. *RNA* **16,** 1340–1349 (2010).
47. Tijerina, P., Mohr, S. & Russell, R. DMS footprinting of structured RNAs and RNA–protein complexes. *Nature Protoc.* **2,** 2608–2623 (2007).
48. Nikolova, E. N. *et al.* Transient Hoogsteen base pairs in canonical duplex DNA. *Nature* **470,** 498–502 (2011).
49. Korzhnev, D. M., Religa, T. L., Banachewicz, W., Fersht, A. R. & Kay, L. E. A transient and low-populated protein-folding intermediate at atomic resolution. *Science* **329,** 1312–1316 (2010).
50. Lucks, J. B. *et al.* Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl Acad. Sci. USA* **108,** 11063–11068 (2011).
51. Das, R., Karanicolas, J. & Baker, D. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Methods* **7,** 291–294 (2010).
52. Yoon, S. *et al.* HiTRACE: high-throughput robust analysis for capillary electrophoresis. *Bioinformatics* **27,** 1798–1805 (2011).
53. Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25,** 1974–1975 (2009).

Acknowledgements

The authors thank A. Laederach and J. Lucks for comments on the manuscript and the authors of *RNAstructure* for making their source code freely available. This work was supported by the Burroughs-Wellcome Foundation (CASI to R.D.), the National Institutes of Health (T32 HG000044 to C.C.V.) and a Stanford Graduate Fellowship (to P.C.).

Author contributions

R.D. conceived and designed the experiments. W.K., C.C.V. and R.D. performed the experiments. C.C.V., P.C. and R.D. analysed the data. R.D. wrote the paper. All authors discussed the results and commented on the manuscript.

Additional information

The authors declare no competing financial interests. Supplementary information accompanies this paper at www.nature.com/naturechemistry. Reprints and permission information is available online at http://www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.D.

nature chemistry

SUPPLEMENTARY INFORMATION

DOI: 10.1038/NCHEM.1176

A two-dimensional mutate-and-map strategy for non-coding RNA structure

Wipapat Kladwang[1], Christopher C. VanLang[2], Pablo Cordero[3], and Rhiju Das[1,3,4]*

Departments of Biochemistry[1] & Chemical Engineering[2], Program in Biomedical Informatics[3], and Department of Physics[4], Stanford University, Stanford CA 94305

** To whom correspondence should be addressed: rhiju@stanford.edu. Phone: (650) 723-5976. Fax: (650) 723-6783.*

Index

Supporting Methods

Preparation of model RNAs

The DNA templates for each RNA (Table S1) consisted of the 20-nucleotide T7 RNA polymerase promoter sequence (TTCTAATACGACTCACTATA) followed by the desired sequence. Double-stranded templates were prepared by PCR assembly of DNA oligomers up to 60 nucleotides in length (IDT, Integrated DNA Technologies, IA) with Phusion DNA polymerase (Finnzymes, MA). For each mutant, an automated MATLAB script was used to determine which primers required single-nucleotide changes and to generate 96-well plate spreadsheets for ordering and guiding pipetting for PCR assembly reactions.

Assembled DNA templates were purified in 96-well Greiner microplates with AMPure magnetic beads (Agencourt, Beckman Coulter, CA) following manufacturer's instructions. Sample concentrations were measured based on UV absorbance at 260 nm measured on Nanodrop 100 or 8000 spectrophotometers. Verification of template length was accomplished by electrophoresis of all samples and 10-bp and 20-bp ladder length standards (Fermentas, MD) in 4% agarose gels (containing 0.5 mg/mL ethidium bromide) and 1x TBE (100 mM Tris, 83 mM boric acid, 1 mM disodium EDTA).

In vitro RNA transcription reactions were carried out in 40 μL volumes with 10 pmols of DNA template; 20 units T7 RNA polymerase (New England Biolabs, MA); 40 mM Tris-HCl (pH 8.1); 25 mM $MgCl_2$; 2 mM spermidine; 1 mM each ATP, CTP, GTP, and UTP; 4% polyethylene glycol 1200; and 0.01% Triton-X-100. Reactions were incubated at 37 °C for 4 hours and monitored by electrophoresis of all samples along with 100–1000 nucleotide RNA length standards (RiboRuler, Fermentas, MD) in 4% denaturing agarose gels (1.1% formaldehyde; run in 1x TAE, 40 mM Tris, 20 mM acetic acid, 1 mM disodium EDTA), stained with SYBR Green II RNA gel stain (Invitrogen, CA) following manufacturer instructions. RNA samples were purified with MagMax magnetic beads (Ambion, TX), following manufacturer's instructions; and concentrations

were measured by absorbance at 260 nm on Nanodrop 100 or 8000 spectrophotometers.

SHAPE measurements

Chemical modification reactions consisted of 1.2 pmols RNA in 20 μL with 50 mM Na-HEPES, pH 8.0, and 10 mM $MgCl_2$ and/or ligand at the desired concentration (see Table S1); and 5 μL of SHAPE modification reagent. The modification reagent was 24 mg/ml N-methyl isatoic anhydride freshly dissolved in anhydrous DMSO. The reactions were incubated at 24 °C for 15 to 60 minutes, with lower modification times for the longer RNAs to maintain overall modification rates less than 30%. In control reactions (for background measurements), 5 μL of deionized water was added instead of modification reagent, and incubated for the same time. Reactions were quenched with a premixed solution of 5 μL 0.5 M Na-MES, pH 6.0; 3 μL of 5 M NaCl, 1.5 μL of oligo-dT beads (poly(A) purist, Ambion, TX), and 0.25 μL of 0.5 mM 5´-rhodamine-green labeled primer (AAAAAAAAAAAAAAAAAAAGTTGTTGTTGTTTCTTT) complementary to the 3´ end of the MedLoop RNA [also used in our previous studies (*1, 2*)], and 0.05 μL of a 0.5 mM Alexa-555-labeled oligonucleotide (used to verify normalization). The reactions were purified by magnetic separation, rinsed with 40 μL of 70% ethanol twice, and allowed to air-dry for 10 minutes while remaining on a 96-post magnetic stand. The magnetic-bead mixtures were resuspended in 2.5 μL of deionized water.

The resulting mixtures of modified RNAs and primers bound to magnetic beads were reverse transcribed by the addition of a pre-mixed solution containing 0.2 μL of SuperScript III (Invitrogen, CA), 1.0 μL of 5x SuperScript First Strand buffer (Invitrogen, CA), 0.4 μL of 10 mM each dnTPs [dATP, dCTP, dTTP, and dITP (*3*)], 0.25 μL of 0.1 M DTT, and 0.65 μL water. The reactions (5 μL total) were incubated at 42 °C for 30 minutes. RNA was degraded by the addition of

5 μL of 0.4 M NaOH and incubation at 90 °C for 3 minutes. The solutions were neutralized by the addition of 5 μL of an acid quench (2 volumes 5 M NaCl, 2 volumes 2 M HCl, and 3 volumes of 3 M Na-acetate). Fluorescent DNA products were purified by magnetic bead separation, rinsed with 40 μL of 70% ethanol, and air dried for 5 minutes. The reverse transcription products, along with magnetic beads, were resuspended in 10 μL of a solution containing 0.125 mM Na-EDTA (pH 8.0) and a Texas-Red-labeled reference ladder (whose fluorescence is spectrally separated from the rhodamine-green-labeled products). The products were separated by capillary electrophoresis on an ABI 3100 or ABI 3700 DNA sequencer. Reference ladders for wild type RNAs were created using an analogous protocol without chemical modification and the addition of, e.g., 2´-3´-dideoxy-TTP in an amount equimolar to dTTP in the reverse transcriptase reaction.

Data processing

The HiTRACE software(*4*) was used to analyze the electropherograms. Briefly, traces were aligned by automatically shifting and scaling the time coordinate, based on cross correlation of the Texas Red reference ladder co-loaded with all samples. Sequence assignments to bands, verified by comparison to sequencing ladders, permitted the automated peak-fitting of the traces to Gaussians. Data were normalized so that, within each mutant, the mean band intensity was unity for all nucleotides except the 20 nucleotides closest to the 5´ and 3´ ends. Individual replicate data sets, including aligned electropherograms and quantified band intensities, are being made publically available in the Stanford RNA Mapping Database (http://rmdb.stanford.edu).

For each data set, Z-scores were calculated as follows. Let the observed band intensities be s_{ij} with $i = 1, 2, \ldots N$ indexing the band numbers, and $j = 1, 2, \ldots M$ indexing the nucleotides that were mutated. Then, the mean band intensities μ_i and standard deviations σ_i were computed using their standard definitions:

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} s_{ij}$$
$$\sigma_i = \left[\frac{1}{N}\sum_{j=1}^{N}(s_{ij}-\mu_i)^2\right]^{1/2} \quad (1)$$

And the Z-scores were computed as:

$$Z_{ij} = [s_{ij} - \mu_i]/\sigma_i \quad (2)$$

Only data with $Z_{ij} > 0.0$ and associated with bands with mean intensity μ_i less than a cutoff $\mu_i^{MAX} = 0.8$ were kept, since the mutate-and-map approach seeks to identify site-specific release of nucleotides that are protected in the starting sequence and most variants. [Varying μ_i^{MAX} from 0.5 to 1.0 gave indistinguishable results for models.] To avoid introducing additional noise, we did not correct for attenuation of longer reverse transcription products; because this effect should be similar for all mutants (and was observed to be such in the data), it scales s_{ij}, μ_i, and σ_{ij} identically (at a given nucleotide *i*) and did not affect the final Z_{ij} scores in (2). Further, to again avoid introducing unnecessary noise, we did not explicitly subtract background measurements, as they should also subtract out of the Z-score expression (2). Nevertheless, control measurements for all RNAs without SHAPE modification were carried out. For a small number (<0.1%) of nucleotides in specific mutants, weak mutant-specific background bands were observed (likely due to sequence-specific reverse transcriptase stops). An analogous Z-score was carried out for these control background measurements; nucleotides with "background Z-scores" greater than 6.0 were identified as anomalous and set to zero in analyzing Z_{ij} for mutate/map measurements. For data sets with more than one independent replicate (the *add* adenine-sensing riboswitch, the P4-P6 domain, and the *F. nucleatum* glycine-

sensing riboswitch), Z_{ij} values were averaged across the replicates. The analysis is available as a single MATLAB script *output_Zscore_from_rdat.m* within the HiTRACE package.

Inference of contacts through sequence-independent clustering

The Z-scores Z_{ij} [see above, eq. (2)] define possible long-range contacts in each RNA. As in prior work (*1, 2*), mutate/map pairs with statistically strong signals ($Z_{ij} > Z_{min}$; Z_{min}= 1.0) were considered. The pair (*i*,*j*) was defined as neighboring any strong signals at (*i*–1, *j*), (*i*+1, *j*), (*i*, *j*–1), (*i*, *j*+1), or the symmetry partner (*j*,*i*); strong signals were then grouped by single-linkage clustering. Final selection of clusters used simple but stringent filters. Clusters with at least 8 pairs, involving at least three independent mutations, and including at least one pair of symmetry partners were taken as defining long-range interactions with strong support. For this selection, mutations involved in more than 5 such clusters were omitted as potentially being associated with cooperative, extended structural effects beyond the disruption of a single base pair, helix, or tertiary contact. The analysis is available as a single MATLAB script *cluster_z_scores.m* within the HiTRACE package.

Three-dimensional modeling with Rosetta: command-lines and example files

Three-dimensional models were acquired using the Fragment Assembly of RNA with Full Atom Refinement (FARFAR) methodology(*5*) in the Rosetta framework.

First, ideal A-form helices were created with the command line:

```
rna_helix.exe -database <path to database> -nstruct 1 -fasta
stem2_add.fasta -out:file:silent stem2_add.out
```

where the file `stem2_add.fasta` contains the sequence of the P2 helix, as determined by the mutate-and-map data:

```
>stem2_add.fasta
uccuaauuggga
```

Then, for each RNA loop or junction motif that interconnects these helices, 4,000 models were created with FARFAR. For example, in the adenine riboswitch, two loops (L2 & L3) and the adenine-binding junction are the non-helical motif portions. The command line for building L2 onto the P2 helix is:

```
rna_denovo.<exe> -database  <path to database> -native
motif2_1y26_RNA.pdb -fasta motif2_add.fasta -params_file
motif2_add.params -nstruct 100 -out:file:silent motif2_add.out -cycles
5000 -mute all -close_loops -close_loops_after_each_move -minimize_rna
-close_loops -in:file:silent_struct_type rna -in:file:silent
stem2_add.out -chunk_res  1-6 16-21
```

Here, the optional "-native" flag, inputting the crystallographic structure for the motif, permits rmsd calculations but is not used in building the model. The file `motif2_add.params` defines the P2 stem within this motif-building run:

```
STEM    PAIR 6 16 W W A PAIR 5 17 W W A PAIR 4 18 W W A PAIR 3 19 W W A
PAIR 2 20 W W A PAIR 1 21 W W A
OBLIGATE PAIR 1 21 W W A
```

Finally, the models of separately built motifs and helices are assembled through the FARNA Monte Carlo procedure:

```
rna_denovo.<exe> -database  <path to database> -native 1y26_RNA.pdb -
fasta add.fasta -in:file:silent_struct_type binary_rna  -cycles 10000 -
nstruct 200 -out:file:silent add_assemble.out -params_file
add_assemble.params -cst_file
add_mutate_map_threetertiarycontacts_assemble.cst -close_loops  -
in:file:silent  stem1_add.out stem2_add.out stem3_add.out
motif1_add.out motif2_add.out motif3_add.out -chunk_res   1-9 63-71 13-
18 28-33 42-47 55-60 1-18 28-47 55-71 13-33 42-60
```

In the above command-line, the helix and loop definitions are given by `add_assemble.params`:

```
CUTPOINT_CLOSED  9 18 47
STEM  PAIR 1 71 W W A  PAIR 2 70 W W A  PAIR 3 69 W W A  PAIR 4 68 W W
A  PAIR 5 67 W W A  PAIR 6 66 W W A  PAIR 7 65 W W A  PAIR 8 64 W W A
PAIR 9 63 W W A
OBLIGATE PAIR 9 63 W W A

STEM  PAIR 13 33 W W A  PAIR 14 32 W W A  PAIR 15 31 W W A  PAIR 16 30
W W A  PAIR 17 29 W W A  PAIR 18 28 W W A
OBLIGATE PAIR 18 28 W W A
```

```
STEM  PAIR 42 60 W W A  PAIR 43 59 W W A  PAIR 44 58 W W A  PAIR 45 57
W W A  PAIR 46 56 W W A  PAIR 47 55 W W A
OBLIGATE PAIR 47 55 W W A
```

The constraint file `add_mutate_map_threetertiarycontacts_assemble.cst` encodes regions in tertiary contact (here including the short two-base-pair helix) inferred from the mutate-and-map data:

```
[ atompairs ]
N3 10 N3 39 FADE -100 10 2 -20.0
N3 10 N1 40 FADE -100 10 2 -20.0
N1 11 N3 39 FADE -100 10 2 -20.0
N1 11 N1 40 FADE -100 10 2 -20.0
N1 12 N1 40 FADE -100 10 2 -20.0
N3 10 N3 39 FADE -100 10 2 -20.0
N1 25 N3 49 FADE -100 10 2 -20.0
N1 25 N3 50 FADE -100 10 2 -20.0
N1 26 N3 48 FADE -100 10 2 -20.0
N1 26 N3 49 FADE -100 10 2 -20.0
N1 26 N3 50 FADE -100 10 2 -20.0
N3 27 N3 49 FADE -100 10 2 -20.0
N3 27 N3 50 FADE -100 10 2 -20.0
N3 35 N1 40 FADE -100 10 2 -40.0
N1 34 N3 41 FADE -100 10 2 -40.0
```

These constraints give a bonus of –20.0 kcal/mol if the specified atom pairs are within 8 Å; the function interpolates up to zero for distances by a cubic spline beyond 10.0 Å. Note that the Rosetta numbering here starts with 1 for the first nucleotide of the 71-nucleotide adenine binding domain, and is offset by 12 from the numbering in the crystal structure 1Y26. 5000 models of the full-length RNA were generated, and the lowest-energy conformation was taken as the final model. (For the adenine riboswitch, the next nine lowest energy models were within 2 Å RMSD of this model, indicating convergence.) Example files for carrying out the calculation are being distributed with Rosetta release 3.4 in rosetta_demos/RNA_Assembly/, including setup script *setup_rna_assembly_jobs.py*.

Table S1. Benchmark for the mutate-and-map strategy for noncoding RNA base pair inference.

RNA, source	Solution conditions[a]	Off–set[b]	PDB[c]	Sequence & Secondary Structure[d]
Adenine riboswitch, *V. vulnificus* (*add*)	Standard + 5 mM adenine	–8	**1Y26** 1Y27 2G9C 3GO2 …	ggaaaggaaagggaaagaaaCGCUUCAUAUAAUCCUAAUGAUAUGGUUUGGG AGUUUCUACCAAGAGCCUUAAACUCUUGAUUAUGAAGUGaaaacaaaacaaa gaaacaacaacaacaac .....(((((((((...((((((.........))))))........((((((.......))))))..)))))))))..... .....
tRNA^{phe}, *E. coli*	Standard	–15	**1L0U** 1EHZ	ggaacaaacaaaacaGCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAG AUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCAaaaccaaagaaac aacaacaacaac .............(((((((..((((........)))).((((..... ..))))).....(((((.......))))))))))..... .....
P4–P6 domain, *Tetrahymena* ribozyme	Standard, 30% methyl-pentanediol[e]	89	**1GID** 1L8V 1HR2 2R8S	ggccaaaacaacgGAAUUGCGGGAAAGGGGUCAACAGCCGUUCAGUACCAAG UCUCAGGGGAAACUUUGAGAUGGCCUUGCAAAGGGUAUGGUAAUAAGCUGAC GGACAUGGUCCUAACCACGCAGCCAAGUCCUAAGUCAACAGAUCUUCUGUUG AUAUGGAUGCAGUUCAaaaccaaaccaaagaaacaacaacaacaac .............(((((...((((((.....(((.((((.(((..((((((((....)))))))))..((.......)).....))).....)))))))....))))))..))).)))))(((...(((...(((((((((...)))))))))..))))...))..... .....
5S rRNA, *E. coli*	Standard	–20	**3OFC** 3OAS 3ORB 2WWQ …	ggaaaggaaagggaaagaaaUGCCUGGCGGCCGUAGCGCGGUGGUCCCACCU GACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGG GGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAUaaaacaaaacaaagaa acaacaacaacaac .....(((((((((.....(((((((((....(((((((.............))))..)))...)))))).))..((.......(((((((...)))))))).......))...))))))))..... .....
Cyclic di-GMP riboswitch, *V. cholerae* (*VC1722*)	Standard + 10 μM cyclic diguanosine monophosphate	0	**3MXH** 3IWN 3MUV 3MUT …	ggaaaaauGUCACGCACAGGGCAAACCAUUCGAAAGAGUGGGACGCAAAGCC UCCGGCCUAAACCAGAAGACAUGGUAGGUAGCGGGGUUACCGAUGGCAAAAU Gcauacaaaccaaagaaacaacaacaacaac .........((((......((...(((((((....)))))))...)).((.(((((((((..((.........)))))))..))))))...)).)) .....
Glycine riboswitch, *F. nucleatum*	Standard + 10 mM glycine	–10	**3P49**	ggacagagagGAUAUGAGGAGAGAUUUCAUUUUAAUGAAACACCGAAGAAGU AAAUCUUUCAGGUAAAAAGGACUCAUAUUGGACGAACCUCUGGAGAGCUUAU CUAAGAGAUAACACCGAAGGAGCAAAGCUAAUUUUAGCCUAAACUCUCAGGU AAAAGGACGGAGaaaacacaacaaagaaacaacaacaacaac .........(((((((.....((((((....)))))).(((...((((. ....)))).)..)))........)))))))..........((((.....(((((.....)))))..(((...((((....((((....)))).....)))))..))) .....))))..... .....

[a] Standard conditions are: 10 mM $MgCl_2$, 50 mM Na–HEPES, pH 8.0 at 24 °C.

[b] Number added to sequence index to yield numbering used in previous biophysical studies, and in Figs. 1-5 of the main text.

[c] The first listed PDB ID was the source of the assumed crystallographic secondary structure; other listed IDs contain sequence variants, different complexes, or different crystallographic space groups and confirm this structure.

[d] In the sequence, lowercase symbols denote 5´and 3´ buffer sequences, including primer binding site (last 20 nucleotides). In all cases, designs were checked in RNAstructure and ViennaRNA to give negligible base pairing between added sequences and target domain. Structure is given in dot–bracket notation, and here denotes base pairs for which there is crystallographic evidence. A long-range two-base-pair helix [25–50, 26–49] in the adenine riboswitch is involved in an extensive non-canonical loop-loop interaction and is not included.

[e] 2-methyl-2,4,-pentanediol (MPD) was included due to reports that its presence in crystallization buffer can change SHAPE reactivity of the P4-P6 RNA (Vicens et al. (2007) RNA 13, 536–48). Mutate/map measurements without MPD gave different reactivities in the P5abc region and ambiguous modeling results in the P5c region; the crystallographic helix or a helix with single-nucleotide register shift was observed in models from different bootstrap replicates.

Table S2. Base-pair-resolution analysis of the helices recovered by the mutate-and-map method.

RNA	Crystallographic	Correctly recovered	Missed[a]			Extra base pairs[a]		
			A-U	G-U	G-C	A-U	G-U	G-C
Adenine riboswitch[b]	21	21	0	0	0	0	0	0
$tRNA^{phe}$	20	20	0	0	0	1	0	0
P4-P6 RNA	48	47	1	0	0	4	1	0
5S rRNA	34	34	0	0	0	0	0	0
c-di-GMP riboswitch[b]	25	23	0	0	0	2	0	0
Glycine riboswitch[b]	40	40	0	0	0	1	2	0
Total	188	185	1	0	0	8	3	0

[a] Number of missed or extra base-pairs within helices that match crystallographic helices. (The only crystallographic helix not recovered in the mutate-and-map models is a short stem with two G-C base pairs in the cyclic di-GMP riboswitch.)

[b] Ligand-binding riboswitches were probed in the presence of small-molecule partners (5 mM adenine, 10 μM cyclic di-guanosine-monophosphate, or 10 mM glycine). All experiments were carried out with 10 mM $MgCl_2$, 50 mM Na-HEPES, pH 8.0.

Figure S1. Example of large-scale change in RNA structure induced by single mutation: loss of ligand binding. In the mutate-and-map data set of main text Fig. 1, the mutation A52U leads to perturbations throughout the adenine riboswitch ligand-binding domain. (A) Quantified, background-subtracted areas for A52U compared to the wild type sequence ('WT') in 10 mM $MgCl_2$, 50 mM Na-HEPES, pH 8.0, and 5 mM adenine. Site of mutation is marked with a red arrow. (B) Perturbations from the A52U mutation are similar to differences between the adenine-free and adenine-bound state of the riboswitch. Data shown are 'gold-standard' measurements (*6*) averaged over five independent replicates for the wild type riboswitch without (red) and with (blue) 5 mM adenine.

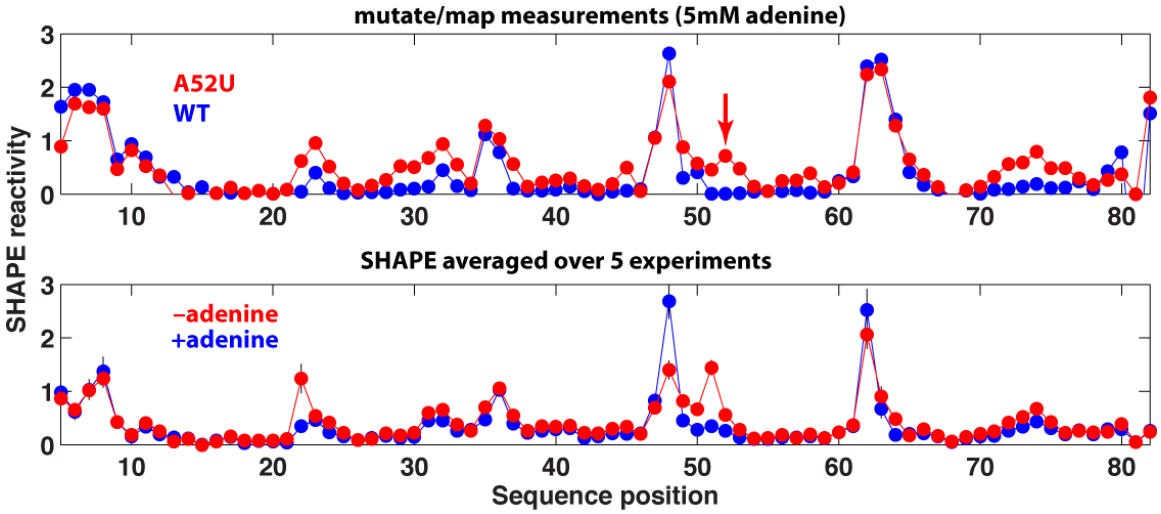


Figure S2. Example of large-scale change in RNA structure induced by single mutation: change in secondary structure. Background-subtracted SHAPE reactivities for the adenine riboswitch ligand-binding domain with C69G mutation are shown as coloring from white to red. The mutation site 69 normally lies in the P3 helix of the wild type RNA; the mutation leads to a large-scale change in SHAPE reactivity. The new reactivity is explained well by a change in secondary structure from the adenine-binding structure (cf. Fig. 1d in main text, and also SI Fig S3). The new model was estimated from *RNAstructure* guided by these 1D data (*7*); bootstrap values (*6*) given as percentages.

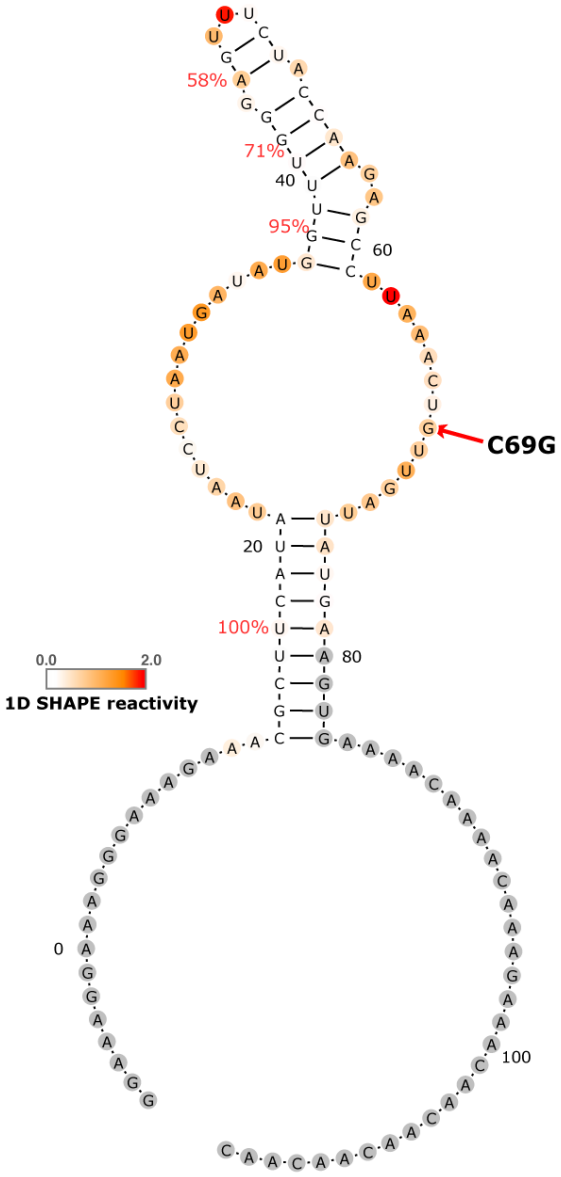


Figure S3. Mutate-and-map analysis of a partially ordered state of the adenine riboswitch. (a) Mutate-and-map data (Z-scores) are given in gray-scale for the adenine-binding domain from the *add* riboswitch, *V. vulnificus*, without adenine present. Red squares mark crystallographic secondary structure of the RNA in its adenine-bound form. (b) Dominant secondary structure for the ligand-free adenine riboswitch, inferred from mutate-and-map data, is not distinguishable from the ligand-bound form (see Main Text Fig. 1c-d). Cyan lines mark an 'extra' helix that is also seen in the ligand-bound state; the helix corresponds to neighboring Watson-Crick base pair and non-Watson-Crick base pair seen in the crystallographic ligand-bound model. Bootstrap confidence estimates for each helix are given in red.

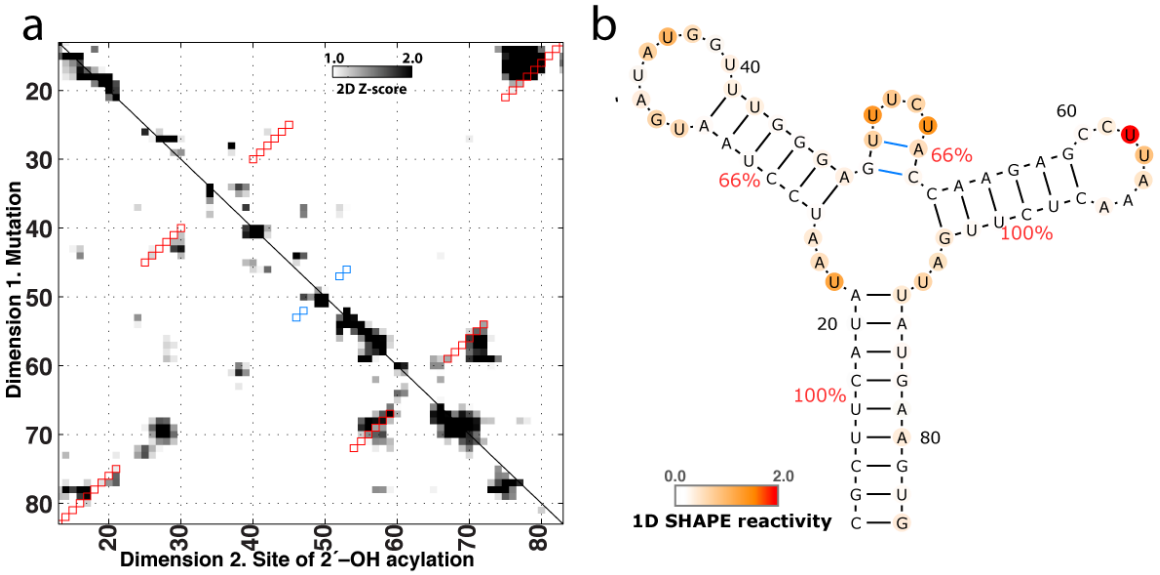


Figure S4. Mutate-and-map analysis of a partially ordered state of a cyclic di-guanosine monophosphate (c-di-GMP) riboswitch. (a) Mutate-and-map data (Z-scores) are given in gray-scale for the c-di-GMP-binding domain from the *VC1722* riboswitch, *V. cholerae*, without c-di-GMP present. Red squares mark crystallographic secondary structure of the RNA in its c-di-GMP-bound form. (b) Secondary structure models for this ligand-free c-di-GMP riboswitch, inferred from mutate-and-map data, is different from models for the ligand-bound state near the P1 stem and c-di-GMP binding region. Cyan squares (a) and lines (b) are mutate-and-map base pairs not present in (ligand-bound) crystal structure; orange annotations are crystallographic base pairs not present in the mutate-and-map model. Bootstrap confidence estimates for each helix are given in red.

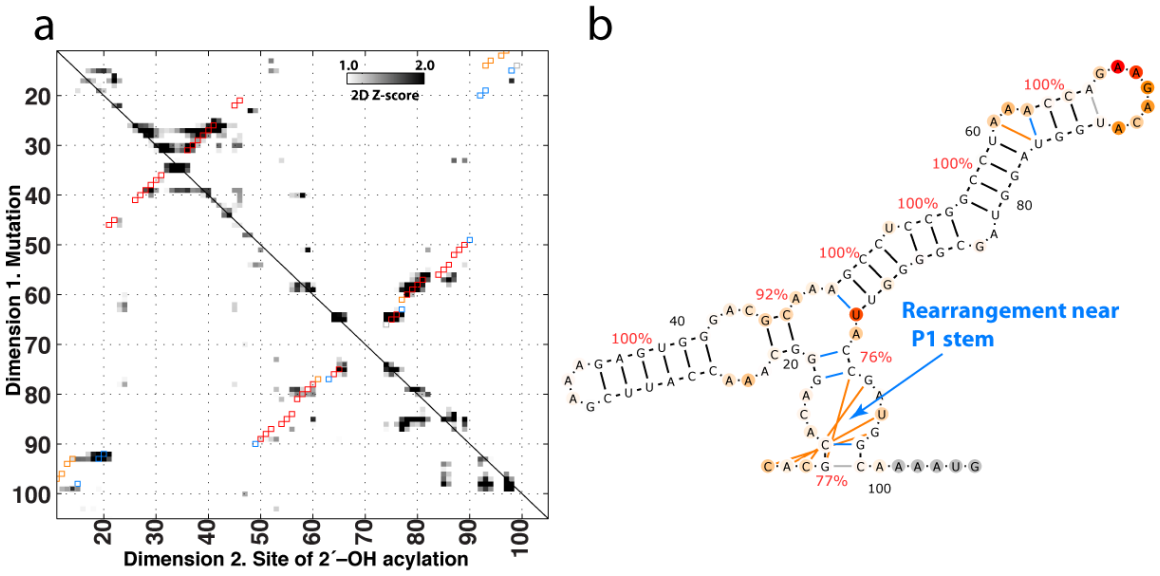
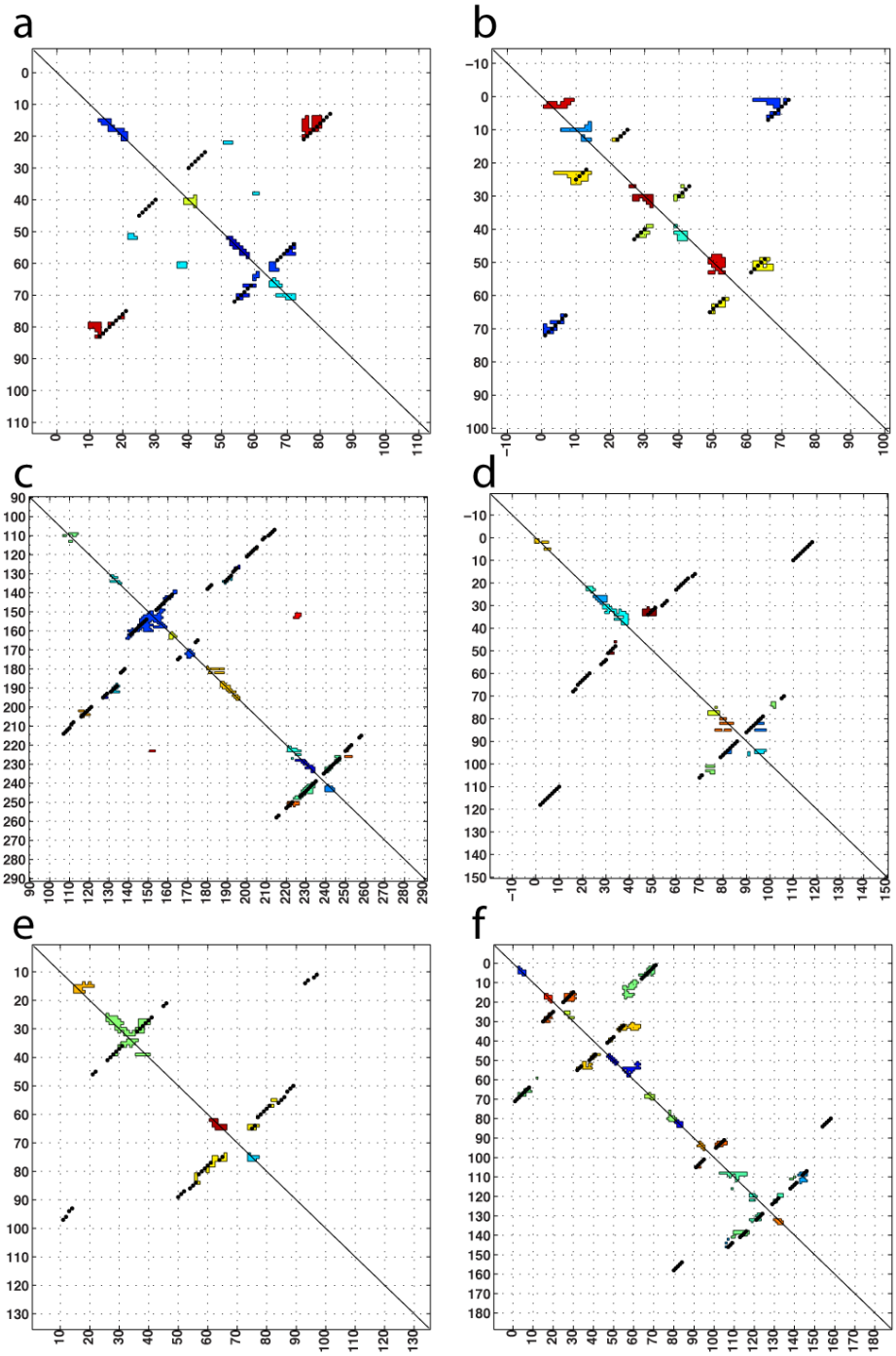


Figure S5. Accurate inference of contacting regions in structured non-coding RNAs through sequence-independent analysis of mutate-and-map data. Cluster analysis of *Z*-scores, using filters for signal strength, number of independent mutants, and symmetry of features (see Methods); final clusters are shown in different, randomly chosen colors. Base pairs from crystallographic secondary structure are marked as black symbols. RNAs are (a) adenine riboswitch, (b) tRNA(phe), (c) P4-P6 RNA, (d) 5S rRNA, (e) c-di-GMP riboswitch, and (f) glycine riboswitch. Riboswitch data were collected with ligands present.



References for Supporting Information

1. Kladwang, W., and Das, R. (2010) A mutate-and-map strategy for inferring base pairs in structured nucleic acids: proof of concept on a DNA/RNA helix, *Biochemistry 49*, 7414-7416.
2. Kladwang, W., Cordero, P., and Das, R. (2011) A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA, *RNA 17*, 522-534.
3. Mills, D. R., and Kramer, F. R. (1979) Structure-independent nucleotide sequence analysis, *Proc Natl Acad Sci U S A 76*, 2232-2235.
4. Yoon, S., Kim, J., Hum, J., Kim, H., Park, S., Kladwang, W., and Das, R. (2011) HiTRACE: high-throughput robust analysis for capillary electrophoresis, *Bioinformatics 27*, 1798-1805.
5. Das, R., Karanicolas, J., and Baker, D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure, *Nat Methods 7*, 291-294.
6. Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011) Understanding the errors of SHAPE-directed RNA modeling, *Biochemistry*, in press.
7. Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. (2009) Accurate SHAPE-directed RNA structure determination, *Proc Natl Acad Sci U S A 106*, 97-102.