Visualizing RNA base pairing probabilities with RNAbow diagrams

Daniel P. Aalberts and William K. Jannen
Department of Physics, Williams College,
33 Lab Campus Dr, Williamstown, MA 01267, USA

There are many effective ways to represent a minimum free energy RNA secondary structure that make it easy to locate its helices and loops. It is a greater challenge to visualize the thermal average probabilities of all folds in a partition function sum; dot plot representations are often puzzling. Therefore, we introduce the RNAbows visualization tool for RNA base pair probabilities. RNAbows represent base pair probabilities with line thickness and shading, yielding intuitive diagrams. RNAbows aid in disentangling incompatible structures, allow comparisons between clusters of folds, highlight differences between wild type and mutant folds, and are also rather beautiful.

I. INTRODUCTION

Graphical representations can profoundly influence our conception of physical reality or interpretation of data. For example, in conventional representations of RNA secondary structure the stems (regions of stacked base pairs) and loops (gaps between) are easily identified; however, showing only one set of base pairs makes invisible the prevalence of thermal fluctuations. In fact, the likelihood of being in even the most probable structure is exceedingly small and thermal fluctuations allow the molecule to explore many states. To characterize RNA structures in thermal equilibrium, better visualization methods are needed.

Much work has gone into developing computational methods to predict the secondary structure from the sequence, including: minimizing free energy [1–3], computing the partition function [2–6], stochastically sampling the partition function [7], enumerating states [8], kinetic approaches [9, 10], maximum-expected accuracy approaches [11, 12], comparative analysis [13, 14], and statistical methods [15–17]. The accuracy of predictions has received scrutiny [18–21]. Our particular interest is to visualize ensembles of structural states in thermal equilibrium as predicted by partition-function based methods.

A number of tools have also been developed to visualize RNA secondary structures. The minimum free energy (MFE) or other secondary structures can be depicted in two-dimensional "airport terminal" diagrams, in which the backbone defines the perimeter and lines or dots between bases denote the pairs, as in Figure 1(a). A classic "rainbow" diagram, see Figure 1(b), encodes the same information but instead of the backbone sequence forming the perimeter, it is stretched horizontally with the base pairs making long arcs. In circle diagrams [22], the backbone is arranged in a circle with arcs again marking the pairs. Most compact is bracket notation [6], see Figure 1(c), in which unpaired bases are periods and matching parentheses indicate paired bases. To represent nonnested pseudoknot structures, bracket notation requires additional delimiters, like [ ] or { }.

Partition-function based computational methods predict the thermal average probabilities P_{ij} of RNA base pairs rather than one single structure. The P_{ij} information is often represented in dot plots — a grid is made and the size or color of the dot at (i, j) indicates the probability of pairing base i with base j, as in Figure 1(d,f). Dots along diagonals indicate stems.

Because the eye naturally groups similar objects together [23], the dot plot representation in Figure 1(d) subliminally suggests that each color represents a unique structure. But closer examination reveals, for example, that base 41 along the horizontal axis forms red pairs with bases 4, 9, 13, and 35 along the vertical axis. So if there is not a single red structure, can one figure out which dots are consistent?

The Figure 1(e) hybrid approach adds to the MFE structure a color coding of the bases according to their probability of pairing [24]. This approach may leave the impression of a single static structure in which the predictions vary in certainty, rather than of a fluctuating molecule exploring many states and many local minima.

II. RESULTS

We introduce RNAbow diagrams as a more intuitive way to visualize RNA structures in thermal equilibrium. RNAbows are the partition function analog of rainbow diagrams. In RNAbow diagrams, we use the line thickness and shade of the arcs to represent the probability of a base pair. The single AllPairs RNAbow displays the entire partition function. In Figure 1(g) it is simple to see the two local minima structures because the eye naturally groups parallel lines. With RNAbows our perceptual inclinations help us, rather than hinder us.

To facilitate comparisons at a glance we introduce the difference RNAbow diagram, such as Figures 1(h), 2, and 3. Two folds, top and bottom, are juxtaposed. Color highlights the differences between folds. When $P_{ij}^{\rm top} > P_{ij}^{\rm bot}$, the top arc's color is set proportional to the relative probability excess $X_{ij}^{\rm top} = (P_{ij}^{\rm top} - P_{ij}^{\rm bot})$, otherwise $X_{ij}^{\rm top} = 0$. We then either use the (hue, saturation,

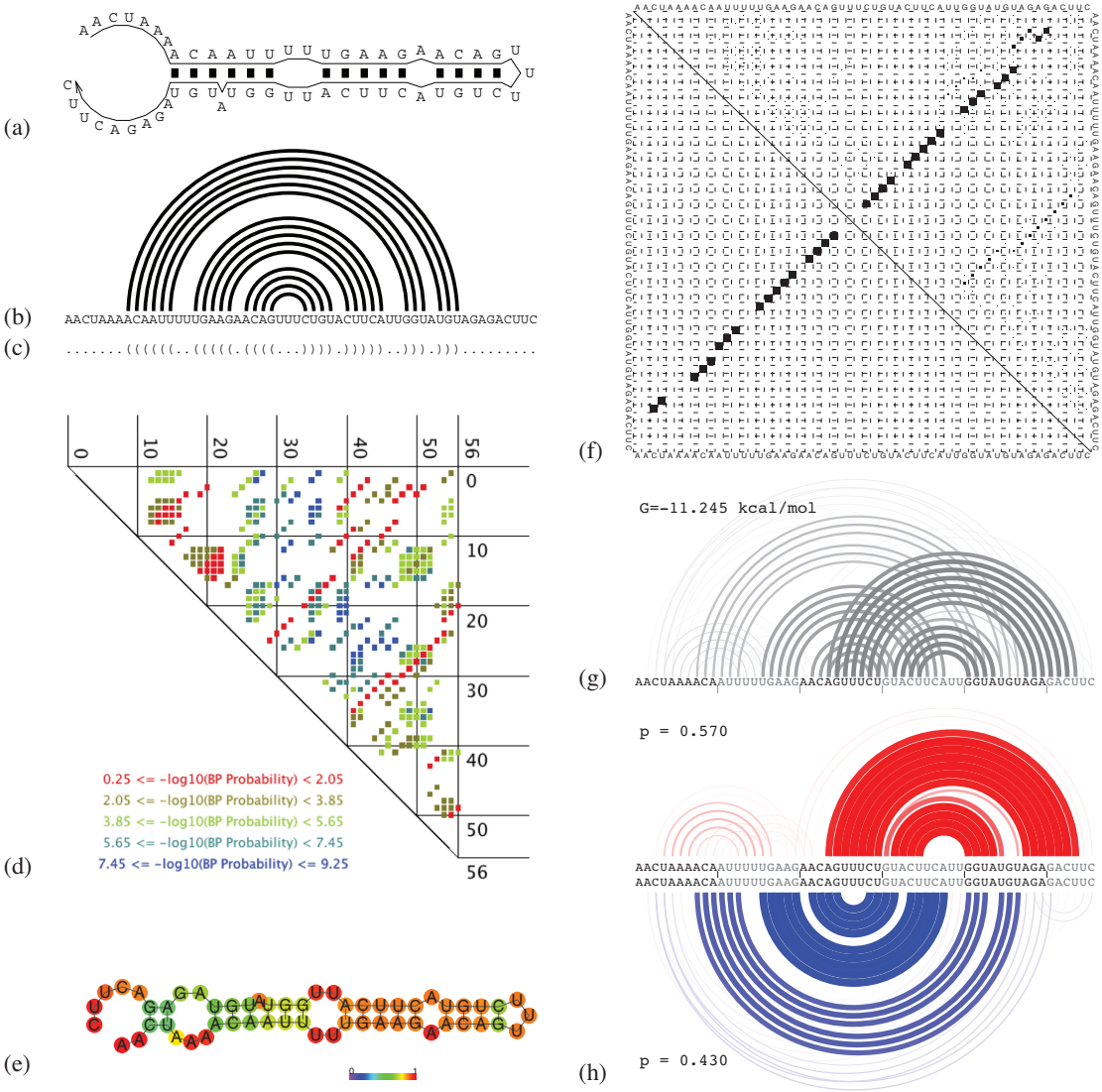

FIG. 1: Depictions secondary structures for the *L. collosoma* Spliced Leader sequence: (a) two-dimensional "airport terminal" diagram of the Minimum Free Energy (MFE) state, (b) classic "rainbow" diagram (MFE), (c) bracket notation with periods representing unpaired bases and parentheses indicating paired bases (MFE). (d) A dot plot with partition function probabilities P_{ij} with base i vertical and base j horizontal. Color is assigned on the basis of the logs of probabilities. [Graphics adapted from RNAStructure.] (e) ViennaRNA's prediction (using slightly different free energy rules) with bases color coded according to their partition function probabilities. [Graphics adapted from ViennaRNA.] (f) A Dot Plot available from ViennaRNA uses box size proportional to probability (upper triangle), but the grid obscures low probability pairs. (g) An AllPairs RNAbow diagram with the line width and darkness proportional to the probability of the base pairs. (h) A Clusters RNAbow diagram after resolving into the two dominant clusters, with probability 0.57 (red) and 0.43 (blue); Note that the MFE state [Fig. 1(b)] belongs to the less probable blue cluster.

value) or RGB color models, with

$$(H, S, V) = (\text{red},\ X_{ij}^{\text{top}},\ X_{ij}^{\text{top}} - P_{ij}^{\text{top}} + 1),$$
$$(R, G, B) = (255 \cdot X_{ij}^{\text{top}}/P_{ij}^{\text{top}},\ 0,\ 0),$$

for pair (i, j) on the top. Formulas for bottom arcs are analogous. Pairs with similar weight are colored black, extra weight drives top pairs toward red and bottom pairs toward blue.

In Figure 1(g) we see two dominant structural classes in the total partition function. To visualize each local minima we first have to partition the partition function; we use our *PF* method, which is described fully in the Supplemental Information. The idea is to identify the base pair (i, j) which is most incompatible with other base pairs. We then split the partition function into two, one with the (i, j) pair *Prohibited* and one with the (i, j) pair *Forced* to exist. The resulting P and F clusters

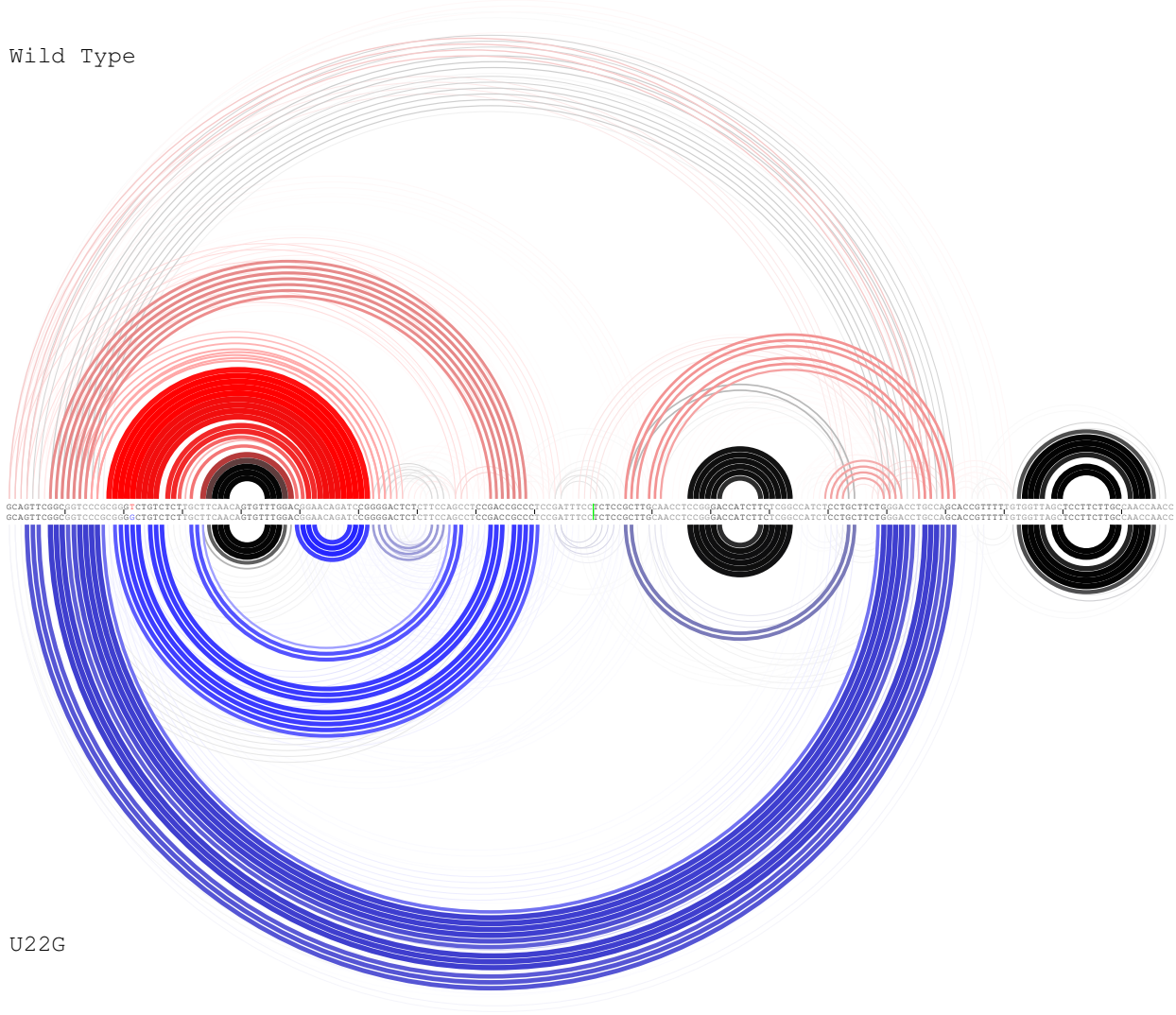

FIG. 2: The partition functions of Wild Type (red) and the U22G mutant (blue) the 5′UTR of ferritin light chain mRNA are depicted. The colors are set proportional to the difference between the clusters such that common elements are black, while the distinct elements are either red or blue. The dramatic effect of this single nucleotide polymorphism on the secondary structure is evident. Other base changes within loop regions have less influence.

describe two local free energy minima, including fluctuations. These are visualized with a Clusters RNAbow in Figure 1(h).

In the more probable red cluster of Figure 1(h), one can see the thermal equilibrium between states in which G_{31} pairs to either U_{45} or U_{47}. And one can also see a possible UAAA/UUUG hairpin duplex early in the sequence which has no topological barrier with the later strong hairpin; it is formed only about one-quarter of the time in this cluster.

In the blue cluster of Figure 1(h), one sees gradations in the stability of the hairpin's stem which are not seen in the MFE structure [Figure 1(b)] because the MFE bonds either exist or not. Notice also that the MFE structure is one of the states in the less probable of the clusters.

If desired, the *PF* procedure could be repeated again on each cluster to further disentangle structures. Notice also in Figure 1(h) that the maximum probabilities P_{ij} within each daughter cluster approach 1 while the most probable pair in the parent cluster was 0.57, roughly the weight p_P. It is easy to imagine applications to visualizing riboswitches which exhibit a conformational change between two folds.

In Figure 2, we present a Difference RNAbow comparison of the 5′ UTR of Ferritin Light Chain wild type to the U22G mutant [25] associated with Hyperferritinemia cataract syndrome. This single nucleotide polymorphism dramatically changes the folding pattern. In particular, the loss of the Iron Response Element, the brightest red hairpin in Figure 2, disrupts binding by an iron-response

protein.

In Figure 3 we present a Difference RNAbow which shows how information about which bases are unpaired obtained from chemical mapping experiments can be incorporated in partition function calculations.

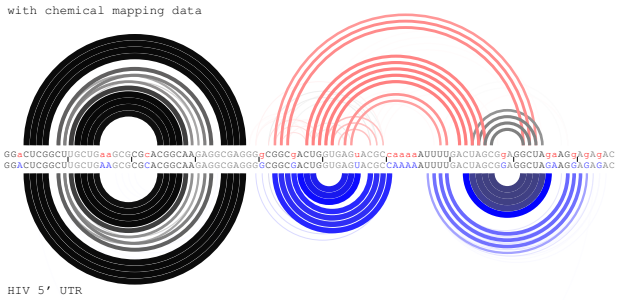

FIG. 3: Chemical mapping experiments indicate which bases are unpaired, and this information can also be visualized with Difference RNAbows. Here, for the 5′-UTR region of HIV-1, we compare the constrained [26] partition function where lower-case bases are forced to be unpaired (top, red) with the unconstrained partition function (bottom, blue).

III. ACCESS

From `http://rna.williams.edu/` users can create their own RNAbows with a choice of UNAfold, ViennaRNA, or RNAstructure [2, 3, 5] to compute the partition functions. Three RNAbows tools are available:

AllPairs to visualize the entire partition function [Figure 1(g)] with base pairs denoted by arcs whose width and shading is proportional to the probability of the pair,

Clusters to split that partition function into two clusters [Figure 1(h)] using the *PF* method described in Supplemental Information, and

Difference RNAbows to highlight the *differences* between the partition functions of two sequences [Figures 1(h), 2, 3].

The RNAbow graphics are rendered in EPS, PDF, or SVG formats for easy import into other applications. All graphics are vector based, so there is no image degradation at any scale.

Advanced users can also import P_{ij} data they have pre-computed using other algorithms. Source code for RNAbows is also available by request for incorporation into other applications.

IV. CONCLUSION

Visualization tools can offer insights into problems beyond mere representation of data. With RNAbows, a partition function's base pair probability information becomes easier to use and more powerful. Our instinctive pattern-matching ability allows us to quickly compare clusters of structures. Incompatible clusters can be disentangled by eye, or by the *PF* procedure described.

The basic RNAbows tools are extensible to represent any data set with varying coupling strengths and then to highlight differences between conditions. The difference RNAbow juxtaposes arcs and highlights differences with color. The recent R-chie package [27] also uses double arcs to visualize helices predicted from multiple alignments with the Transat program [14]. Making side-by-side comparisons of multiple folds is less straightforward with dot plots.

Furthermore, the mental misconceptions that come from looking at a static MFE, PDB, or consensus structure — forgetting that thermal fluctuations open and close pairs, or forgetting that not all pairs are equally stable — are challenged subliminally by the gradations of the shadings in RNAbow pairs.

Acknowledgments

The authors thank Alain Laederach, Dave Mathews, Duane Bailey for discussions, Julian Hess and Chris Warren for programming assistance. This work was supported by the National Institutes of Health [grant number GM080690 to D.P.A.] and the National Science Foundation [grant number MCB-0641995 to D.P.A.].

[1] Zuker, M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science* **244**, 48-52.

[2] Mathews, D.H.; Disney, M.D.; Childs, J.L.; Schroeder, S.J.; Zuker, M.; and Turner, D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA* **101**, 7287–7292.

[3] Markham, N.R. and Zuker, M. (2008) UNAFold: software for nucleic acid folding and hybriziation. In Keith, J. M., editor, *Bioinformatics, Volume II. Structure, Functions and Applications*, number 453 in *Methods in Molecular Biology*, chapter 1, pages 3–31. Humana Press, Totowa, NJ.

[4] McCaskill, J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers.* **29**, 1105–1119.

[5] Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.

[6] Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M. and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte Fur Chemie*, **125**, 167-188.

[7] Ding, Y. and Lawrence, C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.* **31**, 7208–7301.

[8] Wuchty, S., Fontana, W., Hofacker, I.L., and Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49**: 145–165.

[9] Isambert, H. and Siggia, E.D. (2000) Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme *Proc. Natl. Acad. Sci. USA*, 97 (12), 6515.

[10] Hofacker, I.L, *et al.* (2010) BarMap: RNA folding on dynamic energy landscapes. *RNA* **16**: 1308–1316.

[11] Do, C.B., Woods, D.A., and Batzoglou, S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models *Bioinformatics* **22**: e90–e98.

[12] Lu, Z.J., Gloor, J.W., and Mathews, D.H. (2009). Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*. **5**, 1805–1813.

[13] Cannone J.J., *et al.* (2002). The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and Other RNAs. *BMC Bioinformatics*, **3**:2 (and Erratum 3:15).

[14] Wiebe, N.J.P. and Meyer, I.M. (2010) TRANSAT-A Method for Detecting the Conserved Helices of Functional RNA Structures, Including Transient, Pseudo-Knotted and Alternative Structures. *PLoS Comp. Biol.*, **6**: e1000823..

[15] Andronescu, M., Condon, A., Hoos, H.H., Mathews, D.H. and Murphy, K.P. (2010) Computational approaches for RNA energy parameter estimation. *RNA*, **16**, 2304–2318.

[16] Gardner, D.P., Ren, P.Y., Ozer, S. and Gutell, R.R. (2011) Statistical Potentials for Hairpin and Internal Loops Improve the Accuracy of the Predicted RNA Structure. *J. Mol. Biol.*, **413**, 473–483.

[17] Rivas, E., Lang, R. and Eddy, S.R. (2012) A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA*, **18**, 193–212.

[18] Dowell, R.D. and Eddy, S.R. (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, **5**: 71.

[19] Doshi, K.J., Cannone, J.J., Cobaugh, C.W. and Gutell, R.R. (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, **5**: 105.

[20] D. M. Layton and R. Bundschuh (2005) A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation *Nucl. Acids Res.* **33**: 519–524.

[21] Hajiaghayi, M., Condon, A. and Hoos, H.H. (2012) Analysis of energy-based algorithms for RNA secondary structure prediction. *BMC Bioinformatics*, **13**: 22.

[22] Nussinov, R., Pieczenik, G., Griggs, J.R. and Kleitman, D.J. (1978) Algorithms for loop matchings. *Siam J. Appl. Math.*, **35**, 68–82.

[23] Metzger, W. *Laws of Seeing* MIT Press, Cambridge, MA. (2006)

[24] De Rijk,P. and De Wachter,R. (1997) RnaViz, a program for the visualisation of RNA secondary structure. *Nucleic Acids Res.*, **25**, 4679–4684.

[25] Halvorsen, M., Martin, J.S., Broadaway, S. and Laederach, A. (2010) Disease-Associated Mutations That Alter the RNA Structural Ensemble. *PLoS Genet.*, **6**: e1001074.

[26] Schroeder, S.J., Stone, J.W., Bleckley, S., Gibbons, T. and Mathews, D.M. (2011) Ensemble of Secondary Structures for Encapsidated Satellite Tobacco Mosaic Virus RNA Consistent with Chemical Probing and Crystallography Constraints. *Biophys. J.*, **101**, 167-175.

[27] Lai, D., Proctor, J.R., Zhu, J.Y.A. and Meyer, I.M. (2012) R-CHIE: a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Res.*, **40**, e95.

Supplemental Information

Here we will describe our approach to disentangle incompatible secondary structures, resulting in two clusters of compatible structures, as seen in Figure 1(h).

Most folding algorithms [1–6] use the set of *nested* structures to build their partition functions. In the partition function, it is possible to have two probable pairs P_{ij} and P_{kl} which are non-nested, see Figure 1(g), but non-nested pairs do not co-occur in any of the individual states in the partition function sum.

Non-nestedness of pairs is therefore a hallmark of the incompatibility of the structures.

In our simple and iterative *PF* clustering method, we sum over non-nested (k, l) pairs to find the (i, j) pair that produces the biggest non-nestedness score,

$$\psi_{ij} = P_{ij} \sum_{k<i<l<j} P_{kl} + P_{ij} \sum_{i<k<j<l} P_{kl}. \quad (1)$$

Once the maximum ψ_{ij} has been identified, we then perform a new partition function computation with bases (i, j) forced to pair to create the *Forced F* cluster.

The F cluster is the portion of the original partition function that contains the (i, j) pair. The F cluster's free energy

$$G_F = -RT \log \sum_{s \in F} e^{G_s},$$

and base pair probabilities P_{ij}^F are output. The fraction of the original partition function is

$$p_F = \exp\{-(G_F - G_0)/kT\},$$

where G_0 is the original free energy. The original partition function minus the forced partition function leaves all states in which the (i, j) pair is prohibited, the P cluster. The values for the prohibited cluster P can be obtained from the conservation of probability:

$$p_P = p_0 - p_F$$

and

$$p_P P_{ij}^P = p_0 P_{ij}^0 - p_F P_{ij}^F,$$

or by computing the partition function with a prohibit (i, j) constraint.

It is clear from Figure 1(h) that the effect that forcing even one base pair can have in dividing the ensemble of structures is remarkable.

Our *PF* procedure can also be repeated on any previously defined cluster as well at the root; each time separating further incompatibilities. The overall run time is $O(KN^3)$, where K is the number of clusters. [The sums of Eq. (1) appear to require $O(N^4)$ operations if one sums over the indices; however, by iterating over a list of the $O(N)$ base pairs which exceed a probability threshold, $P_{ij} > \theta \sim 10^{-6}$, Eq. (1) requires only $O(N^2)$ operations.] A detailed description of our related NESTOR algorithm, which instead creates clusters from stochastically sampled states, is also forthcoming.