Visualizing RNA base-pairing probabilities with RNAbow diagrams

DANIEL P. AALBERTS[1] and WILLIAM K. JANNEN
Department of Physics, Williams College, Williamstown, Massachusetts 01267, USA

ABSTRACT

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.

Keywords: chemical mapping; mutations; partition function; secondary structure; visualization

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 (Zuker 1989; Mathews et al. 2004; Markham and Zuker 2008), computing the partition function (McCaskill 1990; Hofacker et al. 1994; Hofacker 2003; Mathews et al. 2004; Markham and Zuker 2008), stochastically sampling the partition function (Ding and Lawrence 2003), enumerating states (Wuchty et al. 1999), kinetic approaches (Isambert and Siggia 2000; Hofacker et al. 2010), maximum-expected accuracy approaches (Do et al. 2006; Lu et al. 2009), comparative analysis (Cannone et al. 2002; Wiebe and Meyer 2010), and statistical methods (Andronescu et al. 2010; Gardner et al. 2011; Rivas et al. 2012). The accuracy of predictions has received scrutiny (Doshi et al. 2004; Dowell and Eddy 2004; Layton and Bundschuh 2005; Hajiaghayi et al. 2012). 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 1A. A classic "rainbow" diagram, see Figure 1B, 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 (Nussinov et al. 1978) the backbone is arranged in a circle with arcs again marking the pairs. Most compact is bracket notation (Hofacker et al. 1994), see Figure 1C, in which unpaired bases are periods and matching parentheses indicate paired bases. To represent non-nested 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 1D,F. Dots along diagonals indicate stems.

Because the eye naturally groups similar objects together (Metzger 2006), the dot plot representation in Figure 1D 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?

[1]**Corresponding author**
E-mail aalberts@williams.edu
Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.033365.112.

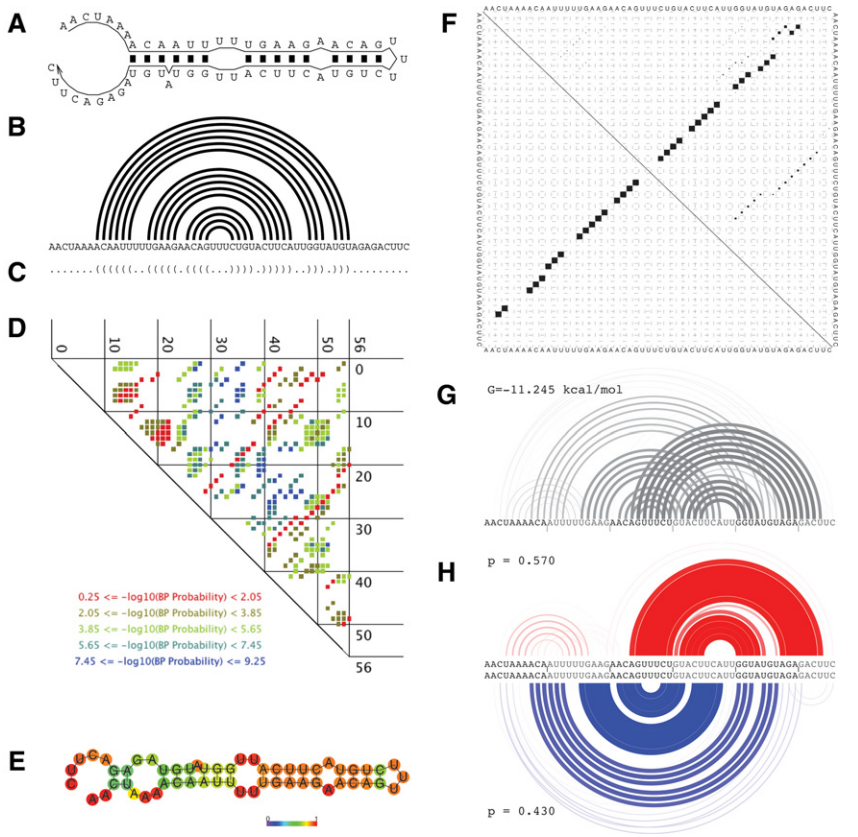

FIGURE 1. Depictions of secondary structures for the *L. collosoma* Spliced Leader sequence: (*A*) two-dimensional "airport terminal" diagram of the Minimum Free Energy (MFE) state; (*B*) classic "rainbow" diagram (MFE); (*C*) bracket notation with periods representing unpaired bases and parentheses 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 (*top* 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 (*B*) belongs to the less-probable blue cluster.

The Figure 1E hybrid approach adds to the MFE structure a color-coding of the bases according to their probability of pairing (De Rijk and De Wachter 1997). 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.

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 1G 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 1H, 2, and 3. Two folds, top and bottom, are juxtaposed. Color highlights the differences between folds. When $P_{ij}^{\text{top}} > P_{ij}^{\text{bot}}$, the top arc's color is set proportional to the relative probability excess $X_{ij}^{\text{top}} = (P_{ij}^{\text{top}} - P_{ij}^{\text{bot}})$, otherwise $X_{ij}^{\text{top}} = 0$. We then either use the (hue, saturation, value) or RGB color models, with

$$(H, S, V) = (\text{red}, X_{ij}^{\text{top}}, X_{ij}^{\text{top}} - P_{ij}^{\text{top}} + 1),$$

$$(R, G, B) = (255\, X_{ij}^{\text{top}}/P_{ij}^{\text{top}}, 0, 0),$$

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 1G 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) that 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 describe two local free-energy minima, including fluctuations. These are visualized with a Clusters RNAbow in Figure 1H.

In the more probable red cluster of Figure 1H, one can see the thermal equilibrium between states in which G31 pairs to either U45 or U47. And, one can also see a possible UAAA/UUUG hairpin duplex early in the sequence that 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 1H, one sees gradations in the stability of the hairpin's stem that are not seen in the MFE structure (Fig. 1B) because the MFE bonds either exist or not. Notice also that the MFE structure is one of the states in the least probable of the clusters.

If desired, the *PF* procedure could be repeated again on each cluster to further disentangle structures. Notice also in Figure 1H that the maximum probabilities P_{ij} within each

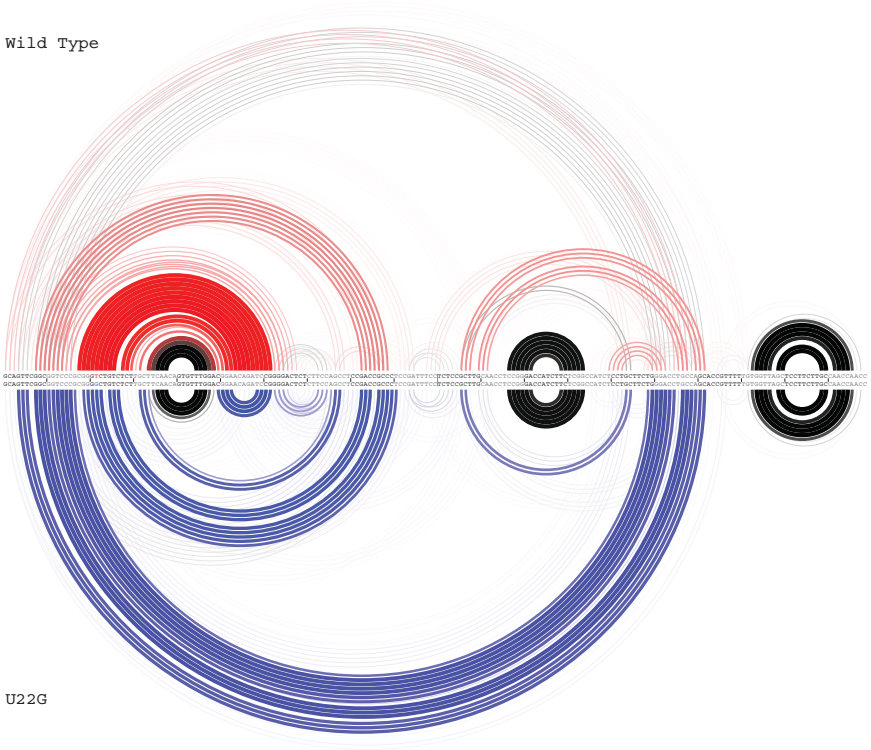

FIGURE 2. The partition functions of wild-type (red) and the U22G mutant (blue) 5′ UTR of ferritin light-chain mRNA are depicted with Difference RNAbows. The colors are set proportional to the difference between the clusters such that common elements are black, while the distinct elements are either 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.

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 that 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 (Halvorsen et al. 2010) 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 that shows how information about which bases are unpaired obtained from chemical mapping experiments can be incorporated in partition function calculations.

ACCESS

From http://rna.williams.edu/ users can create their own RNAbows with a choice of ViennaRNA, RNAstructure, or UNAFold (Hofacker 2003; Mathews et al. 2004; Markham and Zuker 2008) to compute the partition functions. Three RNAbows tools are available:

AllPairs to visualize the entire partition function (Fig. 1G) 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 (Fig. 1H) using the *PF* method described in the Supplemental Information, and

Difference RNAbows to highlight the differences between the partition functions of two sequences (Figs. 1H, 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 precomputed using other algorithms. Source code for RNAbows is also available by request for incorporation into other applications.

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

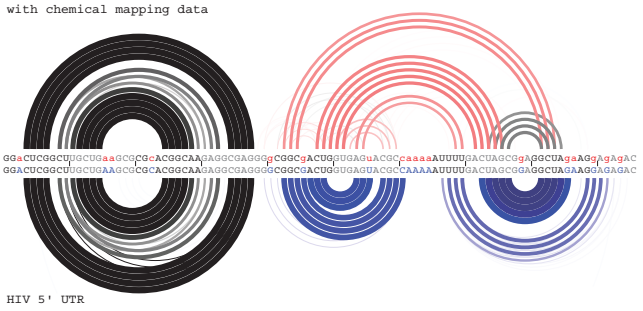

FIGURE 3. Chemical mapping experiments indicate which bases are unpaired, and this information can also be visualized with Difference RNAbows. Here, for the 5′-UTR region of HIV-1, we compare the constrained (Schroeder et al. 2011) partition function where lowercase bases are forced to be unpaired (*top*, red) with the unconstrained partition function (*bottom*, blue).

R-chie package (Lai et al. 2012) also uses double arcs to visualize helices predicted from multiple alignments with the Transat program (Wiebe and Meyer 2010). 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.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank Alain Laederach, Dave Mathews, Duane Bailey for discussions, and Julian Hess and Chris Warren for programming assistance. This work was supported by the National Institutes of Health (grant no. GM080690 to D.P.A.) and the National Science Foundation (grant no. MCB-0641995 to D.P.A.).

Received March 20, 2012; accepted December 21, 2012.

REFERENCES

Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP. 2010. Computational approaches for RNA energy parameter estimation. *RNA* **16:** 2304–2318.

Cannone JJ, Subramanian S, Schnare MN, Coollet JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Müller KM, 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).

De Rijk P, De Wachter R. 1997. RnaViz, a program for the visualization of RNA secondary structure. *Nucleic Acids Res* **25:** 4679–4684.

Ding Y, Lawrence CE. 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res* **31:** 7208–7301.

Do CB, Woods DA, Batzoglou S. 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22:** e90–e98.

Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR. 2004. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics* **5:** 105.

Dowell RD, Eddy SR. 2004. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics* **5:** 71.

Gardner DP, Ren PY, Ozer S, Gutell RR. 2011. Statistical potentials for hairpin and internal loops improve the accuracy of the predicted RNA structure. *J Mol Biol* **413:** 473–483.

Hajiaghayi M, Condon A, Hoos HH. 2012. Analysis of energy-based algorithms for RNA secondary structure prediction. *BMC Bioinformatics* **13:** 22.

Halvorsen M, Martin JS, Broadaway S, Laederach A. 2010. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet* **6:** e1001074.

Hofacker IL. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res* **31:** 3429–3431.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte Fur Chemie* **125:** 167–188.

Hofacker IL, Flamm C, Heine C, Wolfinger MT, Scheuerman G, Stadler PF. 2010. BarMap: RNA folding on dynamic energy landscapes. *RNA* **16:** 1308–1316.

Isambert H, Siggia ED. 2000. Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme. *Proc Natl Acad Sci* **97:** 6515.

Lai D, Proctor JR, Zhu JYA, Meyer IM. 2012. R-chie: A web server and R package for visualizing RNA secondary structures. *Nucleic Acids Res* **40:** e95.

Layton DM, Bundschuh R. 2005. A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. *Nucleic Acids Res* **33:** 519–524.

Lu ZJ, Gloor JW, Mathews DH. 2009. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA* **5:** 1805–1813.

Markham NR, Zuker M. 2008. UNAFold: software for nucleic acid folding and hybriziation. In *Bioinformatics, Volume II. Structure, Functions and Applications, number 453 in Methods in Molecular Biology*, chapter 1 (ed. JM Keith), pp. 3–31. Humana Press, Totowa, NJ.

Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci* **101:** 7287–7292.

McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29:** 1105–1119.

Metzger W. 2006. *Laws of Seeing*. MIT Press, Cambridge, MA.

Nussinov R, Pieczenik G, Griggs JR, Kleitman DJ. 1978. Algorithms for loop matchings. *SIAM J Appl Math* **35:** 68–82.

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

Schroeder SJ, Stone JW, Bleckley S, Gibbons T, Mathews DM. 2011. Ensemble of secondary structures for encapsidated satellite tobacco mosaic virus RNA consistent with chemical probing and crystallography constraints. *Biophys J* **101:** 167–175.

Wiebe NJP, Meyer IM. 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.

Wuchty S, Fontana W, Hofacker IL, Schuster P. 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49:** 145–165.

Zuker M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* **244:** 48–52.