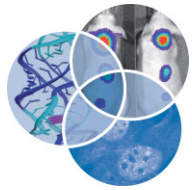
Advanced Review



RNA dot plots: an image representation for RNA secondary structure analysis and manipulations

Alexander Churkin and Danny Barash*

Dot plots were originally introduced in bioinformatics as dot-containing images used to compare biological sequences and identify regions of close similarity between them. In addition to similarity, dot plots were extended to possibly represent interactions between building blocks of biological sequences, where the dots can vary in size or color according to desired features. In this survey, we first review their use in representing an RNA secondary structure, which has mostly been applied for displaying the output secondary structures as a result of running RNA folding prediction algorithms. Such a result may often contain suboptimal solutions in addition to the optimal one, which can be easily incorporated in the dot plot. We then proceed from their passive use of providing RNA secondary structure snapshots to their active use of illustrating RNA secondary structure manipulations in beneficial ways. While comparison between RNA secondary structures can mostly be done efficiently using a string representation, there are notable advantages in using dot plots for analyzing the suboptimal solutions that convey important information about the structure of the RNA molecule. In addition, structure-based alignment of dot plots has been advanced considerably and the filtering of dot plots that considers chemical and enzymatic data from structure determination experiments has been suggested. We discuss these procedures and how they can be enhanced in the future by using an image representation to analyze RNA secondary structures and examine their manipulations.

How to cite this article:
WIREs RNA 2013, 4:205–216. doi: 10.1002/wrna.1154

INTRODUCTION

The analysis of biological sequences is considered a cornerstone in the field of bioinformatics. In particular, image representation has contributed substantially in this growing field to both the interpretation of biological data and the development of new methods. Dot plots, originally introduced in Refs 1 and 2 for comparing biological sequences, can also be used to represent interactions between building blocks of biological sequences. This can be done in the case of protein sequences and nucleic acid sequences.[3,4] In addition, the use of two-dimensional plots to investigate possible secondary structure elements in RNAs (at the time, these 2D plots were known as 'Tinoco plots') can be traced back to the seminal work by Tinoco et al.[5] Trifonov and Boshoi[6] used such 2D plots in their analysis to reveal common hairpins in 5S rRNA molecules. In a further advance forward on exploiting this type of representation, Jacobson and Zuker[7] showed how to use dot plots to predict well-determined regions in a viral genome, suggesting that the amount of cluttering in dot plots reflect the impossibility of accurate structure predictions. Since then, dot plots have been widely used when analyzing the results of RNA folding prediction software by energy

*Correspondence to: dbarash@cs.bgu.ac.il
Department of Computer Science, Ben-Gurion University, Beer-Sheva, Israel

minimization.[8–10] In this survey, we examine their benefits in representing RNA secondary structure in ways that are advantageous over other types of representation that received recent attention such as arc diagrams[11] and mountain plots.[12]

The secondary structure of an RNA molecule is a representation of the pattern, given an initial RNA sequence, of complementary base-pairings that are formed between the constituent nucleotides. The sequence, represented as a string of four letters, is a single strand consisting of the nucleotides A, C, G, and U, which are generally assumed to pair to form a secondary structure with minimum free energy. The computational prediction of the RNA secondary structure given a sequence is a challenging but practical problem: it can be directly tested in the laboratory with minimal experimental effort relative to, for example, RNA tertiary structure. Oftentimes, there is a well-known correspondence between the secondary structure of RNA and the molecule's ultimate function. RNA secondary structure can be represented in several ways: squiggle plot, graph representation, dot-bracket notation, circular plot, arc diagram, mountain plot, dot plot, and more. Squiggle plots are the most common representation of an RNA secondary structure. They have been ordinarily used since the beginning of investigations on RNA secondary structures[13] and their automatic display.[14,15] Bonds formed between paired bases are drawn as chords, usually short straight lines, while the nucleotide labels are placed along a curved line. Graph representations are varied as there are many ways to represent an RNA secondary structure with a simplified graph. Three major ways are the full graph representation where each nucleotide is a node,[16] a coarse grain representation where each motif is a node,[17] and a full tree leading to a homeomorphically irreducible tree.[18] Dot–bracket notation, which is basically a string of dots and brackets that represent the RNA secondary structure, was invented in Vienna and implemented in the Vienna RNA package.[19] It is most commonly used in the computer science community because it can easily be aligned next to an RNA sequence, where a dot represents an unpaired nucleotide and each base pair is symbolized by a pair of opening and closing brackets of the same type. Circular plots were first used in the seminal papers of Nussinov et al. on RNA folding prediction.[20,21] The backbone is represented by a circle, and the base pairs are symbolized by arcs in the interior of the circle. In arc diagrams, the RNA backbone is drawn as a straight line and the nucleotides of each base pair are connected by an arc (see, for example, their implementation in jViz.Rna[22] and R-chie.[11]). Mountain plots[23,24] are named so because they look similar to a mountain range. The nucleotide index number is plotted against the x-axis and the number of enclosing base pairs against the y-axis (a nucleotide k is enclosed by a base pair (i, j) if $i < k < j$). Along with dot plots, they have been used in applications such as the automatic detection of conserved base pairing patterns in RNA virus genomes.[25] Finally, dot plots are two-dimensional graphs in which the nucleotide sequence is written along both x- and y-axis, and a dot is placed in each position where a base pair occurs. Historically, apart from their meaningful application[7,26] in predicting well-determined regions mentioned above, they have been used extensively since then (e.g., as stem histograms in StructureLab,[27] in RNA folding prediction by energy minimization including long RNA sequences,[28,29] in detecting classes of RNA molecules,[30] and recently in RNA deleterious mutation prediction[31,32] as will be elaborated further). It should be mentioned that traditionally, mfold was used to display energy dot plots and the Vienna RNA package was used to display probability dot plots, but both software packages are now able to produce all combinations. A considerable advantage in using dot plots which will be discussed later is that they can contain more information about the RNA than only a single secondary structure. Dot plots can also be used to represent base pair probabilities, Watson-Crick base pairs, and dinucleotide information content in RNA sequence alignments related to RNA secondary structure prediction, which can be helpful for displaying information obtained from programs such as KnetFold[33] and CorreLogo.[34]

In the following sections we will first describe some basic themes in RNA folding prediction. These themes are instrumental for assessing the importance of dot plots in the RNA field. Subsequently, we will critically examine the use of dot plots for stem alignment and clustering, for RNA structure comparison, for illustrating suboptimal manipulations, and for analyzing structure with bilateral and morphological filters.

RNA SECONDARY STRUCTURE

The RNA molecule is single stranded and folds upon itself to form base pairs. The secondary structure of RNA is the collection of base pairs, which occur in its three-dimensional structure. An RNA sequence can be represented as $R = r_1, r_2, r_3, \ldots, r_N$, where r_i is the ith (ribo)nucleotide, belonging to the set A, C, G, U. Referring to i as the ith base of the sequence and to j as the jth base of the sequence, a secondary structure on

R is a set S of ordered base pairs, written as (i, j) where $1 \leq i < j \leq n$ and satisfying the following constraints:

1. No sharp turns: If S contains i, j then $|j - i| \geq 4$.
2. No overlap of pairs: If S contains (i, j) then it cannot contain $(i, j)'$ (with $j' \neq j$) or (i', j) (with $i' \neq i$).
3. If (i, j) and (i', j') are two base pairs, assuming without loss of generality that $i \leq i'$, then either:

$i = i'$ and $j = j'$ (they are the same base pair),
$i < j < i' < j'$ ((i, j) precedes (i', j')), or
$i < i' < j' < j$ ((i, j) includes (i', j')).

The second constraint excludes triplets. It follows from this constraint that each r_i occurs either in exactly one pair or in no pairs, and r_i is described as paired or unpaired accordingly. The last constraint excludes pseudoknots, which occur when two base pairs (i, j) and (i', j') satisfy $i < i' < j < j'$. Pseudoknots are more complicated for RNA folding prediction methods by energy minimization to deal with and therefore they are oftenwise not taken into account to first-order approximation.

In addition to the aforementioned constraints, some common base pair definitions are:

1. Watson-Crick: If S contains (i, j) then (i, j) are either A and U, or U and A, or C and G, or G and C.
2. $G \cdot U$ wobble: If S contains (i, j) then (i, j) are either G and U, or U and G.

The $G \cdot U$ wobble base pair has comparable thermodynamic stability to Watson-Crick base pairs. These types of base pairs are known as canonical base pairs.

RNA FOLDING PREDICTION

Prediction of RNA secondary structure given its sequence is a problem of substantial interest in RNA biology. Computational methods that address this problem have been devised already in the 1970s, with the seminal works that introduced a dynamical programming formulation.[20,35,36] A major advance was to cast the problem as energy minimization and use thermodynamics and auxiliary information,[36,37] which resulted in Zuker's mfold[8] and the Vienna RNA package[9] that also uses the partition function algorithm in Ref 38.

Minimum free energy RNA folding prediction uses energy functions that are a superposition of the contribution of different energy components. Let $e(r_i, r_j)$ be the energy of a base pair between a base in position r_i and a base in position r_j. The energy of the entire structure S is given by:

$$E(S) = \sum_{i,j \in S} e(r_i, r_j). \quad (1)$$

A very simple model would choose the values of e at 37°C to assume −3, −2, and −1 kcal/mol for GC, AU, and GU canonical base pairs, respectively, taking into account the relative strengths of the most dominant base pairings.[5] However, this model is too simplified to consider stabilizing stacking energies for neighboring base pairs in double-helical regions and destabilizing energies for loops containing unpaired bases.[36] The extended thermodynamic parameter sets used by current programs such as mfold and the Vienna RNA package include large tables of sequence-specific loop energies for short loops, negative energy values for stacking interactions, and more.

Before demonstrating how energy minimization in current-day prediction packages is performed in principle using dynamic programming, the entity component called a 'loop' needs to be discussed since the total energy of a structure can be expressed as the sum over all loop energies. As shown in the section above, a secondary structure is a set S of ordered base pairs, basically represented as a list of noncrossing base pairs specified by indices assigned to positions in the sequence. One can distinguish two relations between base pairs: (1) (k, l) is said to be 'interior' to a base pair (i, j) if $i < k < l < j$, (2) (k, l) is said to be 'immediately interior' to (i, j) if there is no base pair (m, n) such that $i < m < k < l < n < j$. Every base pair in a structure closes exactly one loop. A loop closed by (i, j) consists of (i, j) itself, all base pairs which are immediately interior to it, and the unpaired regions between these base pairs. The formula used to calculate the energy for a loop depends on its type. Loops with no interior base pairs are called hairpin loops, while those with exactly one immediately interior base pair (k, l) are called 'interior'. Loops with more than one immediately interior base pair are called multibranched or multiloops. The interested reader is referred to more explanations in Ref 39 and to the derivation of the method in detail put forth in Ref 36.

It should be noted that following Ref 36, a k-loop decomposition was formulated in Ref 40 that defines the secondary structure motifs by k-cycles. In this type of decomposition, if (i, j) is a base pair in S and $i < k < j$, we say that k is accessible from

(i,j) if there is no $(i',j') \in S$ such that $i < i' < k < j' < j$. If both k and l are accessible then the base pair (k,l) is accessible. The set of $(k-1)$ base pairs and k' single-stranded bases accessible from (i,j) is called the k-loop closed by (i,j). The null k-loop, L_0 consists of those single- and double-stranded bases accessible from no base pair. This is referred to as the 'exterior loop', and its bases and base pairs are said to be 'free'. Given that m is the number of base pairs in the secondary structure, any secondary structure S partitions the sequence R uniquely into k-loops $L_0, L_1, L_2, \ldots, L_m$ where $m > 0$ iff $S \neq \emptyset$. Since each loop except for L_0 is uniquely determined by its closing base pair, $L(i,j)$ denotes the loop closed by the base pair (i,j). The sequence is then decomposed in the following way:

$$R = L_0 \cup (\underset{(i,j)\in S}{\cup} L(i,j)). \tag{2}$$

Note that the closing base pair is not contained in the k-loop (or k-cycle). With this type of decomposition,[40,41] one can develop a nomenclature for k-loops with the following cases and sub-cases:

1. $k = 1$: A 1-loop is called a hairpin loop.
2. $k = 2$: Let (i',j') be the base pair accessible from (i,j). Then the 2-loop is called:
   - (a) stacked pair if $i' - i = 1$ and $j - j' = 1$.
   - (b) bulge loop if $i' - i > 1$ or $j - j' > 1$, but not both.
   - (c) interior loop if both $i' - i > 1$ and $j - j' > 1$.
3. $k \geq 3$: These k-loops are called multibranched or multiloops.

The above generalized decomposition into k-loops can assist further when developing algorithms for loop-dependent energy rules in more detail.[41]

A basic RNA folding algorithm that minimizes an energy function as described in Eq. (1) uses the following recursion:

$$W(i,j) = \min \begin{cases} V(i,j) \\ W(i+1,j) \\ W(i,j-1) \\ \min\limits_{i \leq k < j} \{W(i,k) + W(k+1,j)\} \end{cases}.$$

$$V(i,j) = min \begin{cases} \text{Hairpin}(i,j) \\ \min\limits_{i<k<l<j} \text{Interior}(i,j,k,l) + V(k,l) \\ \text{Multi}(i,j) + W(i+1,j-1) \end{cases} \tag{3}$$

There are two terms in the above formula for each subsequence i to j. V_{ij} is the mean free energy under the constraint that bases i and j form a base pair, while W_{ij} is the global mean free energy. The loop types 'Hairpin', 'Interior', and 'Multi' that were discussed above appear in the formula as functions that return the energy values of the corresponding loops. The recursive algorithm that was developed in Ref 36 and derived there in detail works by adding one nucleotide at a time to a sequence, and observing what the best structure is at each step. The last number to be computed is the minimum energy, which is the desired answer. The construction of the structure is achieved by a traceback through the matrices W and V.

It should be noted that the use of dot plots is not restricted to dynamic programming algorithms. For example, RNA folding prediction by energy minimization can also be performed using stochastic algorithms such as Monte Carlo-like methods or genetic algorithms.[10] In addition, information of known structures from biochemical experiments that are not computationally predicted can be incorporated into the dot plots.

SUBOPTIMAL SOLUTIONS

An important breakthrough in the field of folding prediction by energy minimization was the ability to calculate meaningful suboptimal solutions to the problem, which was first devised in Ref 42. Because of environmental considerations that cannot be modeled and imperfections of the model, the RNA may often be found in one of the suboptimal solutions as in the well-known case of tRNAs. The 'mfold style' suboptimal solutions[42] are computed along with a filtering step that ensures that suboptimal solutions differ to avoid redundancy. In mfold, the user can choose the percent of suboptimality, which is set by default to a value of $p = 5\%$. If this number is set to p, only foldings within $p\%$ from the minimum free energy will be computed. The energy dot plot generated by mfold contains the superposition of all possible foldings within $p\%$ of the minimum energy. A different way of calculating the suboptimal solutions was carried out in Ref 43 for the Vienna RNA package. It calculates all suboptimal solutions within an energy range above the minimum free energy without a pre-prescribed filtering step. The RNAsubopt program available in the Vienna RNA package outputs the suboptimal structures—sorted by mean free energy—in a dot–bracket notation, followed by the energy in kcal/mol. It is then possible to perform some filtering on the output, a feature that was also used in Ref 31 for predicting deleterious mutations and will be explained in the continuation.

The ability to calculate suboptimal solutions in RNA folding prediction is significant since the suboptimal structures convey important information on the system that can be represented in dot plots.

COARSE GRAIN REPRESENTATION

Having described various types of loops that can appear in an RNA secondary structure, the issue of coarse grain representation comes naturally as an important RNA motif called a 'stem' can be visualized in dot plots. A group of at least two consecutive base pairs is called a stem. A stem of k base pairs contain $k-1$ stacking interactions in addition to the hydrogen bonding of the base pairs. Since in an RNA structure dot plot a dot is placed in the ith row and jth column to represent the base pair (i,j), a dot plot consists of stems and missing dots between consecutive stems on the same diagonal is an indication of a loop.

Coarse graining is helpful in analyzing structural data that can be inferred, for example, from dot plots. The representation of RNA secondary structure as coarse grained tree-graphs was initially explored in Refs 17,44,45,19 and was used in Refs 46–48 to address the problem of deleterious mutation prediction in the secondary structure of RNAs. A coarse grained tree-graph T can be transformed to a Laplacian matrix $L(T)$ that is symmetric, with one row and column for each node on the tree. The Laplacian matrix is constructed according to $L(T) = D(T) - A(T)$, where $D(T)$ is the diagonal matrix of vertex degrees and $A(T)$ is the adjacency matrix. The rows and columns of $L(T)$ sum up to 0. The complete set of eigenvalues of the Laplacian matrix is called the spectrum of the graph and is independent of how graph vertices are labeled. More formally, let $T = (V, E)$ be a tree with vertex set $V = \{v_1, v_2, \ldots, v_n\}$ and edge set E. The degree of v is denoted by $d(v)$, where $v \in V$ is a vertex of T. The Laplacian matrix of T is $L(T) = (m_{ij})$, where

$$m_{ij} = \begin{cases} d(v_i), & \text{if } i = j, \\ -1, & \text{if } v_i, v_j \in E, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

$L(T)$ is a symmetric, positive semidefinite and singular matrix. The lowest eigenvalue of $L(T)$ is always 0, since all rows and columns sum up to 0. The eigenvalues of $L(T)$ are denoted by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n = 0$. The second smallest eigenvalue, λ_{n-1}, is called the algebraic connectivity[49] of T and labeled as $a(T)$.

The following properties that characterize the Laplacian matrix eigenvalues are useful for their application in analyzing RNA structure:

1. The eigenvalues of $L(t)$ are nonnegative and the first eigenvalue is 0.
2. The second smallest eigenvalue is the algebraic connectivity of T, denoted by $a(T)$.[49]
3. $0 \leq a(t) \leq 1$.
4. $a(T) = 0$ iff T is not connected. This can occur, for example, if separate fragments of an RNA library are to be analyzed. Since loops are connected through stems, $a(T) > 0$ for each RNA molecule.
5. $a(T) = 2(1 - cos(\pi/n)$ iff $T = P_n$ is a path on n vertices.[49]
   Example: RNA viruses tend to be linear in shape, corresponding to a path. For a potato spindle tuber viroid (PSTVd), for example, $a(T) = 2(1 - cos(\pi/24)) = 0.0171$ because it is a path on 24 vertices.
6. $a(T) = 1$ iff $T = K_{1,n-1}$ is a star on n vertices.[50,51]

Intuitively, the second eigenvalue of the Laplacian matrix is monotonically increasing from its lower value for a linear tree-graph structure to its highest value of 1.0 for a 'star shaped' tree-graph structure. As shown in Ref 48, where the interested reader is referred to for more details and illustration, the Laplacian matrix representation and its properties can assist in distinguishing mutations with interesting structural properties and filtering mutations that are less likely to be found deleterious. For the purpose of dot plots, coarse graining can serve as a complementary tool for analyzing structures though the standard dot plots representing base pairs are more informative in their content.

DOT PLOTS FOR STEM ALIGNMENT AND CLUSTERING

Dot plots representing RNA secondary structure, as illustrated in Figure 1, depict stems. They also convey information about loops of various types by observing where there are no dots displayed, but each diagonal line of dots represents a stem. Therefore, a most obvious use of dot plots is to illustrate stem alignment, which can be viewed in Figure 1 for the corresponding RNA structures. For actually computing multiple alignments, as was proposed in the past for sequence alignments of general dot plots[52] (not necessarily RNAs), a method for performing structure-based multiple alignment of RNA probability dot plots was developed in Ref 53. The method is based on a simplified variant of Sankoff's algorithm for simultaneous folding and alignment.[54] It is available

in the Vienna RNA package in programs named PMcomp and PMmulti.

As a consequence of the above, it is tempting to think of a dot plot application in the context of clustering structures and to detect common structural features in each class of related RNAs. The latter objective has some roots in Ref 56 and was further explored in 30, where a trial was done to superimpose many dot plots on each other that belong to the same RNA class in order to filter out random noise and remain with significant common structural features. More recently, a genome-scale structure-based clustering of RNAs called LocARNA (local alignment of RNA) that has resemblance to the alignment method of probability dot plots mentioned above,[53] similarly being based on a simplified variant of Sankoff's algorithm, was put forth in Ref 57 along with biologically motivated clustering analyses. In LocARNA,57,58 dot plots are used for comparing known and predicted structures in various clusters. By default, LocARNA creates a dot plot that is a base pair probability matrix for each sequence in the multiple-sequence input, where the structures are predicted by Vienna's RNAfold unless a fixed structure is specified. This facilitates structure comparison. It can therefore be concluded that illustrating stem alignment is something that dot plots are found to be useful for in a straightforward way, and moreover the alignment and clustering of dot plots can be worked out for some detailed structure-based analyses of RNAs.

DOT PLOTS FOR RNA STRUCTURE COMPARISON

When dealing with dot plots, it is plausible to think of a distance measure between dot plots that takes advantage of this type of representation. This was tried in Ref 59 with an improvement in efficiency described in Ref 60. Simple subtraction of dot plot images will not work because a small shift of stems in the dot plot will result in a large computed distance, which is a nondesirable result. Initially, ideas that were examined included dynamic time warping and Fast Fourier Transform (FFT) registration, after which histogram correlation that has been shown useful in computer vision and computational geometry applications[61–63] was also found in this case to be suitable. The histogram correlation method works as follows. Let A and B be two dot plot diagrams representing secondary structures. The distance measure denoted by D_{HC} (for histogram correlation) is calculated by:

$$D_{HC}(A,B) = \frac{\mathrm{Dist}(A,B)}{\mathrm{Corr}(A,B)}, \tag{5}$$

where $\mathrm{Dist}(A,B)$ is taken as root mean square or Hausdorff distance for the groups of points of both dot plot diagrams and $\mathrm{Corr}(A,B)$ stands for correlation, which in our case is a four-dimensional histogram correlation:

$$\mathrm{Corr}(A,B) = \sqrt{X_C(A,B) \times Y_C(A,B) \times D_C(A,B) \times I_C(A,B)}, \tag{6}$$

where $X_C(A,B)$ is the correlation between the vectors that sum all the point on each X column of the matrix, $Y_C(A,B)$ is correspondingly for the vectors that sum all the points on each Y row of the matrix, $D_C(A,B)$ is correspondingly for the vectors that sum all the points on each diagonal $SW-NE$, and $I_C(A,B)$ is the correlation between the vectors that sum all the points on each inverse diagonal $SE-NW$. These correlations are computed by calculating normalized cross-correlations between two one-dimensional vectors. The way to generate a one-dimensional series vector from the two-dimensional matrix that represents a dot plot diagram is by traversing the diagram, each time on a specific axis, and summing all the values on that axis. When considering two vectors, $X(i)$ and $Y(i)$, $i = 0, 1, \ldots, N-1$, the cross-correlation Corrat delay d is defined as:

$$\mathrm{Corr}(d) = \frac{\sum_i [(X(i) - MX) \times (Y(i-d) - MY)]}{\sqrt{\sum (X(i) - MX)^2 \times \sum (Y(i-d) - MY)^2}}, \tag{7}$$

where MX and MY are the means of the corresponding series, and $d = 0, 1, 2, \ldots, N-1$ represents all the possible delays. The cross-correlation between vectors X and Y is then:

$$\mathrm{Cross\_Correlate}(X,Y) = \mathrm{Max}_d(\mathrm{Corr}(d)), \tag{8}$$

where $\mathrm{Corr}(d)$ is as defined above. The cross-correlation will be maximal when the two compared vectors are identical, or contain identical areas. More details and explanation of the formulas is available in Ref 59. In Ref 60, a simplification was proposed to rotate the image in 45° angle, thereby avoiding the need to compute the cross-correlation from four different directions. In this proposed modification, histogram correlation is still used for RNA structure comparison. While histogram correlation seems to be a robust way to compare dot plots, it remains questionable whether for RNA structure comparison the dot plots comparison can offer any advantage over the

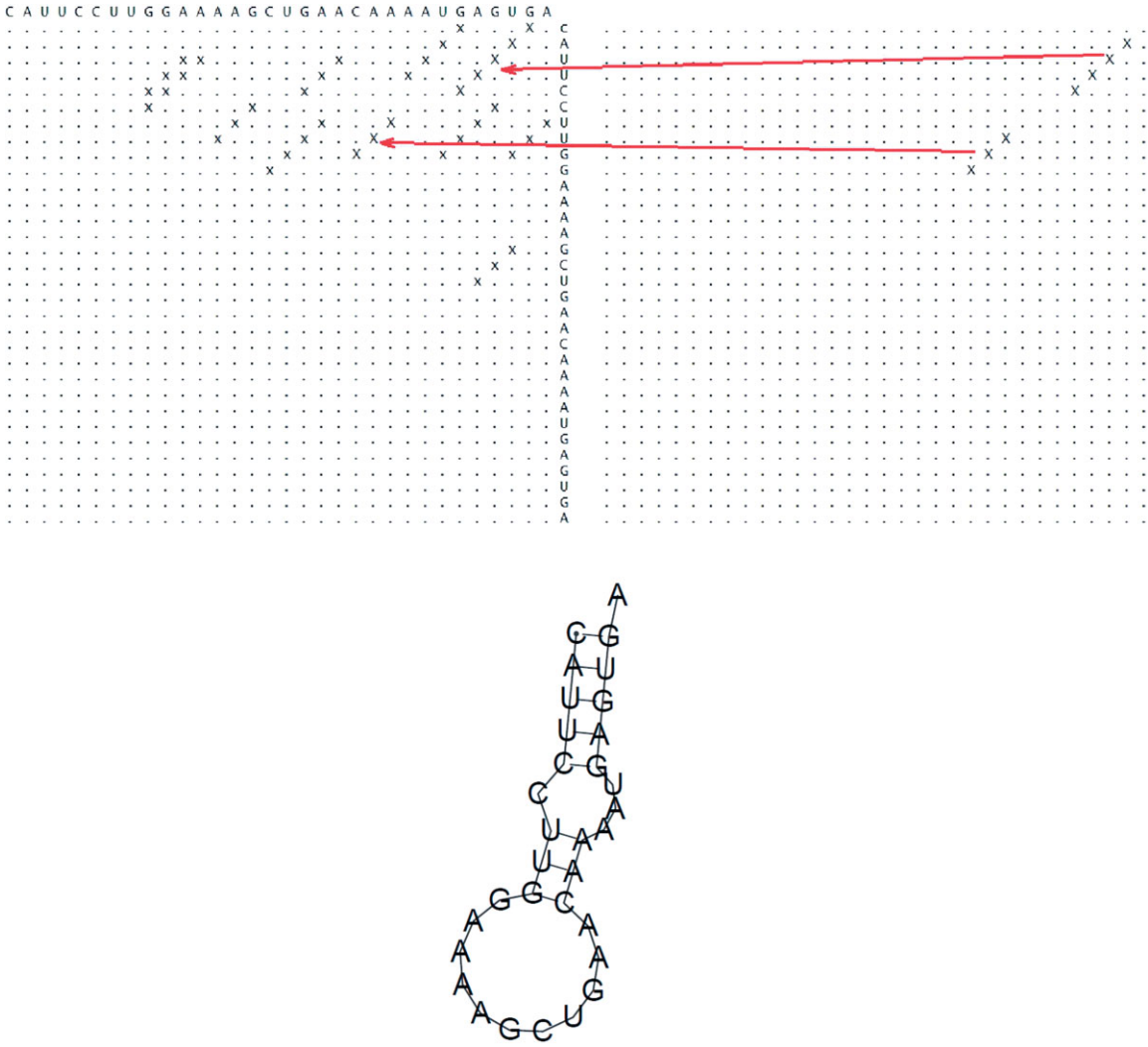


FIGURE 1 | An illustration of stem alignment using dot plots. The folding prediction result of an RNA segment taken from the small nuclear RNA SNORD73,[55] using a dot plot representation (top left) that includes suboptimal solutions and a secondary structure drawing of the optimal solution (bottom). Panel on the right depicts a dot plot containing artificially drawn stems of interest, with arrows indicating a desired alignment of these stems with the structure contained in the dot plot of panel on the left.

traditional strings comparison used in mfold[64] or the Vienna RNA package.[19] A widely used string comparison is available, for example, with the RNAdistance program in the Vienna RNA package.[19] In Vienna, the dot-bracket notation was devised to represent an RNA secondary structure, which is basically a string of dots and brackets. A dot represents an unpaired nucleotide and each base pair is symbolized by a pair of opening and closing brackets of the same type. It is then computationally. It is computationally quite affordable to calculate distances between dot–bracket strings using the so-called 'base-pair distance' in $O(N)$, where N is the length of the sequences. A recent comparison on metric methods performed in Ref 65 showed that the aforementioned RNA structure comparison using dot–bracket strings works sufficiently well for practical purposes. The traditional base-pair distance was recently challenged by the 'relaxed base-pair score[66]', which is intended to give a more biologically realistic measure of the difference between RNA structures, but it does not employ dot plots and is therefore outside the scope of the methods described here.

DOT PLOTS FOR ILLUSTRATING SUBOPTIMAL MANIPULATIONS

As mentioned in a previous section, suboptimal solutions to the RNA folding prediction problem are highly important for inferring about the structure

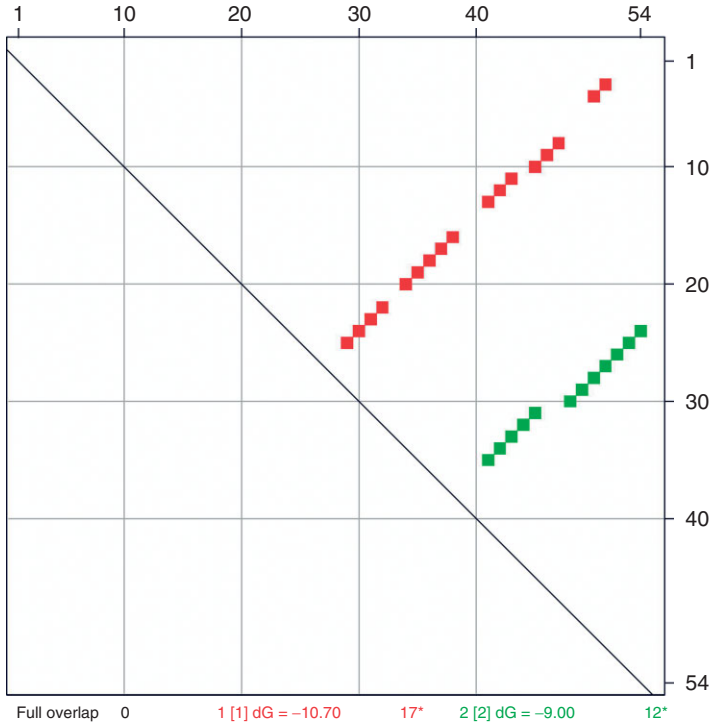


FIGURE 2 | Dot plot of a bistable RNA. A dot plot example using UNAFold[68] of a bistable RNA from *l*eptomonas collosoma spliced leader[67] that assumes two different stable conformations. Optimal structure is depicted in red dots and suboptimal structure in green dots, respectively.

of the RNA molecule. Dot plots are suitable and convenient for visually examining these suboptimal solutions. This can be observed from a simple example of an RNA switch, taken from the *leptomonas collosoma* spliced leader RNA[67]. In this example, the structure of a short RNA sequence that is known from biological experiments to be bistable (alternating between two highly stable states) is well predicted by energy minimization. After this sequence was provided as input and the percent of suboptimality was increased to 20% in order to capture the two stable structures, a dot plot containing these predicted structures was generated using UNAFold.[68] The dot plot is depicted in Figure 2. It is possible to view the two stable conformations, one is optimal and the other is suboptimal, in the same dot plot. A point mutation in the *leptomonas collosoma* that was observed experimentally[67] and can be predicted computationally[48] can actually manipulate the suboptimal solution to become optimal, after which the suboptimal structure becomes dominant and takes charge of the system. The dot plot is useful in representing such potentially important suboptimal solutions along with the optimal one.

A similar concept was used in Ref 31 to solve the problem of multiple-point deleterious mutation prediction in an efficient manner, illustrated by dot plots. As can be found in Ref 31 and the additional example demonstrated in Figure 3, by computing all suboptimal solutions in advance,[43] it is possible to choose as candidates for deleterious mutations only those that enhance a suboptimal solution and suppress the optimal one. In the dot plot in Figure 3, the dots corresponding to the optimal solution are labeled with 'o' and the dots corresponding to the suboptimal solution are labeled with 's'. As illustrated, the mutation at position (4, 18) was picked because it enhances the suboptimal solution, which in turn becomes the optimal solution. Although the position of the mutation within the sequence and its corresponding structure is calculated with a string representation most efficiently, the method can be best understood and illustrated with dot plots. The advantage of a dot plot is that in the same image, it can contain both the optimal solution and the suboptimal ones in a visually convenient way. As discussed in the relevant section, suboptimal solutions to the folding prediction problem by energy minimization convey important information about the system. Thus, manipulations of suboptimal solutions as in the prediction of deleterious mutations mentioned above are an example of a successful use of dot plots.

DOT PLOTS ANALYSIS WITH MORPHOLOGICAL FILTERS

For general dot plots in bioinformatics, an image-processing approach to pursue the analysis of dot plots derived from sequence and structural data was

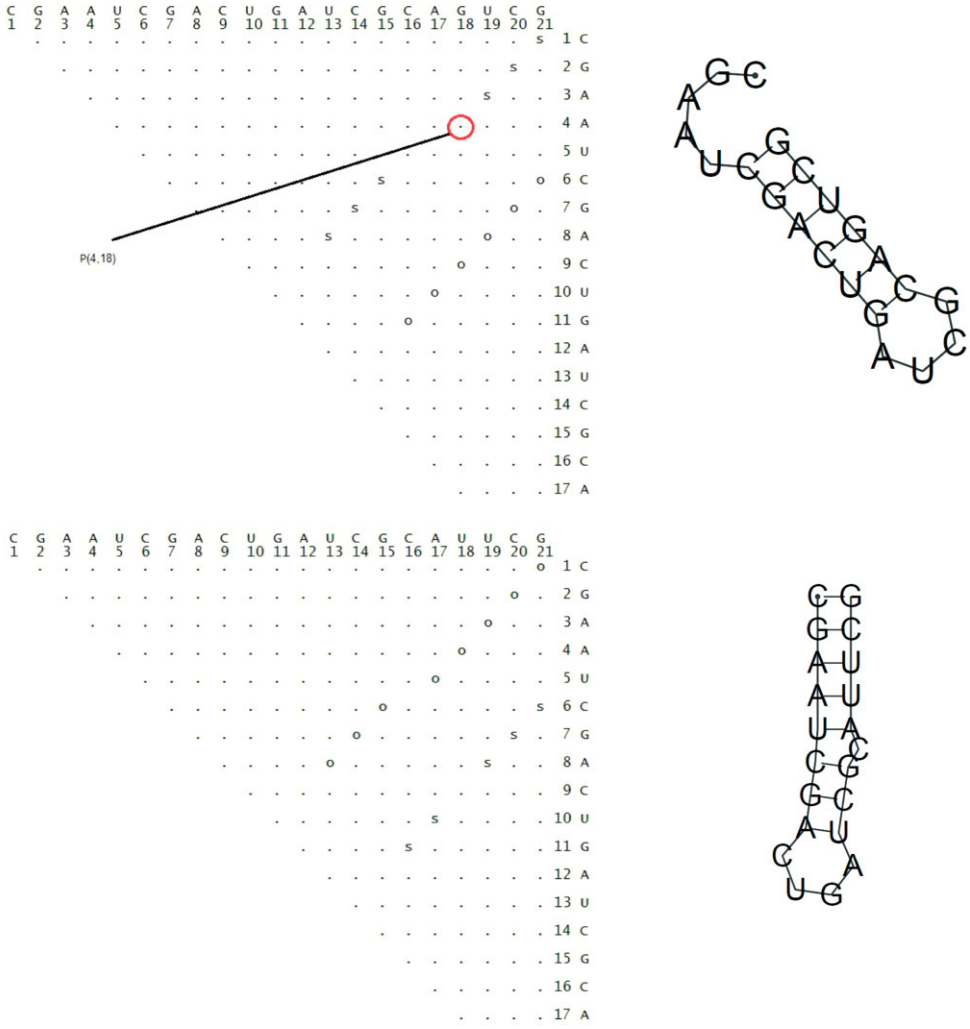


FIGURE 3 | An example of suboptimal manipulations using dot plots. The folding predicted solution of an artificial RNA is drawn using a dot plot representation (top left), which contains both the optimal solution labeled 'o' and a suboptimal solution labeled 's'. The mutation at position (4,18) is circled and pointed to with a line because it can potentially enhance a suboptimal solution. The RNA secondary structure drawing of the optimal solution is shown (top right). The folding predicted solution of the mutated artificial RNA is drawn using a dot plot (bottom left), showing that the suboptimal solution of the wild-type became the optimal solution of the mutated sequence. The RNA secondary structure drawing of the new optimal solution is shown (bottom right).

described in Ref 69. Examples were given from 'PAM' and 'Eisenberg' dot plots for their beneficial use in protein research.

For RNA secondary structure dot plots, simple filtering of the dots to enable structural analysis was carried out in Ref 6 to identify common base-pairing regions in related sequences of ribosomal RNAs. In Ref 70, further discussing the filtering of dot plots in the context of RNA secondary structure, the idea that chemical and enzymatic data from structure determination experiments can be incorporated to the filtering of dot plots was put forth. Nowadays, with more advanced ways to perform chemical and enzymatic probing, such an approach could be found very useful for structure analysis alongside with methods for the direct consideration of chemical probing data in RNA folding prediction.[71,72] Because of the manner in which RNA dot plots are generated, with some valued dots forming diagonals that represent stems, especially appealing in such a case would be the idea to perform structural analysis using morphological filters.[73] The morphological element could then be a diagonal of a given length of points that correspond to a stem of interest, leading to a filtering approach that distinguishes between desired stems and stems that are less valuable by some criteria. A bilateral, edge-preserving filter that was also conceptually studied in Ref 74 could then be used

to preserve large deviations in energy or probability of forming a base pair while filtering is performed in order to extract important information contained in the dot plot, which is basically an image with base pairing strengths taken as intensities.

CONCLUSIONS

RNA secondary structure dot plots provide a compact image representation that is available in major RNA folding prediction software and could be used more often for structure analysis in highly beneficial ways. Many RNA secondary structures can be superimposed on a single dot plot, and this fact makes them very useful for comparative analyses.[28]

Although there are a few purposes for which other representations are found more useful, there are many cases where dot plots are indispensable compared to other representations. Admittedly, the simple two-dimensional drawings (squiggle plots) are the most common for displaying single RNA secondary structures. Similarly, dot–bracket strings are a compact representation in the text writing of multiple lines of structures and are also the most efficient representation for the purpose of a fast RNA secondary structure comparison. However, dot plots are more convenient than any other representation method for the display and analysis of multiple structures in a single plot. Dot plots can be viewed as digital images that basically contain information on RNA structural motifs called stems since each dot represents a base pair. Therefore, an obvious use of dot plots is in the illustration of stem alignments. An efficient method for computing the alignment of RNA probability dot plots was developed in Ref 53. Moreover, methods that attempt to manipulate *suboptimal solutions* of the RNA folding prediction problem can be conveniently understood and illustrated using dot plots. In addition, ongoing advances in chemical and enzymatic probing experiments for structure determination offer high prospects to revive ideas that were put forth in modeling RNA secondary structure by the filtering of dot plots using experimental data.[70] Exploiting this approach to fullest could then be done in improved ways relative to Ref 70 by the aid of more advanced filtering approaches on dot plots. For example, one could think of using bilateral filtering as was conceptually explained in Ref 74 along with morphological filters,[73] in which the morphological element is a diagonal that reflects the basic properties of this type of image representation.

REFERENCES

1. Fitch W. Locating gaps in amino acid sequences to optimize the homology between two proteins. *Biochem Genet* 1696, 3:99–108.
2. Gibbs AJ, McIntyre GA. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur J Biochem* 1970, 16:1–11.
3. Maizel JV, Lenk RP. Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc Natl Acad Sci USA* 1981, 78:7665–7669.
4. Staden R. An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. *Nucl Acids Res* 1982, 10:2951–2961.
5. Tinoco I, Uhlenbeck OC, Levine MD. Estimation of secondary structure in ribonucleic acids. *Nature* 1971, 230:363–367.
6. Trifonov EN, Bolshoi G. Open and closed 5S ribosomal RNA, the only two universal structures encoded in the nucleotide sequences. *J Mol Biol* 1983, 169:1–13.
7. Jacobson AB, Zuker M. Structural analysis by energy dot plot of a large mRNA. *J Mol Biol* 1993, 233:261–269.
8. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucl Acids Res* 2003, 31:3406–3415.
9. Hofacker IL. Vienna RNA secondary structure server. *Nucl Acids Res* 2003, 31:3429–3431.
10. Shapiro BA, Kasprzak W, Grunewald C, Aman J. Graphical exploratory data analysis of rna secondary structure dynamics predicted by the massively parallel genetic algorithm. *J Mol Graph Model* 2006, 25:514–531.
11. Lai D, Proctor JR, Zhu JY, Meyer IM. R-chie: a web server and R package for visualizing RNA secondary structures. *Nucl Acids Res* 2012, 40:e95.
12. Bo X, Wang S. TargetFinder: a software for antisense oligonucleotide target site selection based on MAST and secondary structures of target mRNA. *Bioinformatics*, 2005, 21:1401–1402.
13. Fresco JR, Alberts BM, Doty P. Some molecular details of the secondary structure of ribonucleic acid. *Nature* 1960, 188:98–101.
14. Muller G, Gaspin Ch, Etienne A, Westhof E. Automatic DISPLAY of RNA secondary structures. *Comput Appl Biosci* 1993, 9:551–561.
15. Shapiro BA, Lipkin LE, Maizel JV. An interactive technique for the display of nucleic acid secondary structure. *Nucl Acids Res* 1982, 10:7041–7052.

16. Waterman MS. Secondary structure of single stranded nucleic acids. *Adv Math Suppl Stud* 1978, I:167–212.
17. Shapiro BA. An algorithm for comparing multiple RNA secondary structures. *Comput Appl Biosci* 1988, 4:387–393.
18. Fontana W, Konings DAM, Stadler PF, Schuster P. Statistics of RNA secondary structures. *Biopolymers* 1993, 33:1389–1404.
19. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. *Monatsh Chem* 1994, 125:167–188.
20. Nussinov R, Pieczenik G, Grigg JR, Kleitman DJ. Algorithms for Loop Matchings. *SIAM J Appl Math* 1978, 35:68–82.
21. Nussinov R, Jacobson AB. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci USA* 1980, 77:6309–6313.
22. Wiese KC, Glen E, Vasudevan A. JViz-Rna, a java tool for RNA secondary structure visualization. *IEEE Trans Nanobio* 2005, 4:212–218.
23. Hogeweg P, Hesper B. Energy directed folding of RNA sequences. *Nucl Acids Res* 1984, 12:67–74.
24. Konings DAM, Hogeweg P. Pattern analysis of RNA secondary structure similarity and consensus of minimal-energy folding. *J Mol Biol* 1989, 207:597–614.
25. Hofacker IL, Stadler PF. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comput Chem* 1999, 23:401–414.
26. Zuker M, Jacobson AB. ''Well Determined'' regions in RNA secondary structure predictions. Application to small and large subunit rRNA. *Nucl Acids Res* 1995, 23:2791–2798.
27. Shapiro BA, Kasprzak W. STRUCTURELAB: a heterogeneous bioinformatics systems for RNA structure analysis. *J Mol Graph* 1996, 14:194–205.
28. Fekete M, Hofacker IL, Stadler PF. Predicting RNA base pairing probabilities on massively parallel computers. *J Comp Biol* 2000, 7:171–182.
29. Zuker M, Mathews DH, Turner DH. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In Barciszewski J, Clark FC, eds. *RNA biochemistry and biotechnology. NATO ASI Series*. Dordrecht, NL: Kluwer Academic Publishers; 1999, 11–43.
30. Horesh Y, Amir A, Michaeli S, Unger R. RNAMAT: an efficient method to detect classes of RNA molecules and their structural features. In Proceedings of the 26th International Conference of the IEEE Engineering in Medicine and Biology (EMBS), vol 4. San Francisco, CA; 2004, 2869–2872.
31. Churkin A, Barash D. An efficient method for the prediction of deleterious multiple-point mutations in the secondary structure of RNAs using suboptimal folding solutions. *BMC Bioinformatics* 2008, 9:222.
32. Waldispühl J, Devadas S, Berger B, Clote P. Efficient algorithms for probing the RNA mutation landscape. *PLOS Comput Biol* 2008, 4:e1000124.
33. Bindewald E, Shapiro BA. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA* 2006, 12:342–352.
34. Bindewald E, Schneider TD, Shapiro BA. CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments. *Nucl Acids Res* 2006, 34:W405–W411.
35. Waterman MS, Smith TF. RNA secondary structure: a complete mathematical analysis. *Math Biosci* 1978, 42:257–226.
36. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl Acids Res* 1981, 9:133–148.
37. Turner DH, Mathews DH, Sabina J, Zuker M. Expanded sequence dependence of thermodynamics parameters improves prediction of RNA secondary structure. *J Mol Biol* 1999, 288:911–940.
38. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 1990, 29:1105–1119.
39. Dimitrieva S, Bucher P. Practicality and time complexity of a sparsified RNA folding algorithm. *J Bioinf Comput Biol* 2012, 10:1241007.
40. Zuker M, Sankoff D. RNA secondary structures and their prediction. *Bull Math Biol* 1984, 46:591–621.
41. Zuker M. RNA folding prediction: the continued need for interaction between biologists and mathematicians. *Lect Math Life Sci* 1986, 17:86–123.
42. Zuker M. On finding all suboptimal foldings of an RNA molecule. *Science* 1989, 244:48–52.
43. Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 1999, 49:145–165.
44. Shapiro BA, Zhang KZ. Comparing multiple RNA secondary structures using tree comparisons. *Comput Appl Biosci* 1990, 6:309–318.
45. Le SY, Nussinov R, Maizel J. Tree graphs of RNA secondary structures and their comparisons. *Comput Biomed Res* 1989, 22:461–473.
46. Margalit H, Shapiro BA, Oppenheim BA, Maizel JV. Detection of common motifs in RNA secondary structures. *Nucl Acids Res* 1989, 17:4829–4845.
47. Barash D. Deleterious mutation prediction in the secondary structure of RNAs. *Nucl Acids Res* 2003, 31:6578–6584.
48. Barash D. Second eigenvalue of the Laplacian matrix for predicting RNA conformational switch by mutation. *Bioinformatics* 2004, 20:1861–1869.
49. Fiedler M. Algebraic connectivity of graphs. *Czechoslovak Math J* 1973, 23:298–305.
50. Merris R. Characteristic vertices of trees. *Lin Multi Alg* 1987, 22:115–131.

51. Grone R, Merris R. Algebraic connectivity of trees. *Czechoslovak Math J* 1987, 37:660–670.
52. Vingron M, Argos P. Motif recognition and alignment for many sequences by comparison of dot-matrices. *J Mol Biol* 1991, 218:33–43.
53. Hofacker IL, Bernhart SHF, Stadler PF. Alignment of RNA base pairing probability matrices. *Bioinformatics* 2004, 20:2222–2227.
54. Sankoff D. Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. *SIAM J Appl Math*. 1985, 45:810–825.
55. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. *Nucl Acids Res* 2003, 31:439–441.
56. Unger R, Harel D, Sussman JL. DNAMAT: an efficient graphic matrix sequence homology algorithm and its application to structural analysis. *Comput Appl Biosci* 1986, 2:283–289.
57. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLOS Comput Biol* 2007, 3:e65.
58. Will S, Joshi T, Hofacker IL, Stadler PF, Backofen R. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA* 2012, 18:900–914.
59. Ivry T, Michal S, Avihoo A, Sapiro G, Barash D. An image processing approach to computing distances between RNA secondary structure dot plots. *Alg Mol Biol* 2009, 4:4.
60. Tsang HH, Jacob C. RNADPCompare: an algorithm for comparing RNA secondary structures based on image processing techniques. In Proceedings of the IEEE Congress on Evolutionary Computation (CEC). New Orleans, LA; 2011, 1288–1295.
61. Boutin M, Kemper G. Which point configuration are determined by the distribution of their pairwise distances. *Int J Comput Geom Appl* 2007, 17:31–43.
62. Funkhouser T, Kazhdan M, Min P, Shilane P. Shape-based retrieval and analysis of 3D models. *Comm ACM* 2005, 48:58–64.
63. Lowe DG. Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 2004, 60:91–110.
64. Zuker M, Jaeger J, Turner DH. A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structure determined by phylogenetic comparison. *Nucl Acids Res* 1991, 19:2707–2714.
65. Gruber A, Bernhart SHF, Hofacker IL, Washietl S. Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics* 2008, 9:122.
66. Agius P, Bennett KP, Zuker M. Comparing RNA secondary structure using a relaxed base-pair score. *RNA* 2010, 16:865–878.
67. LeCuyer KA, Crothers DM. The Leptomonas Collosoma spliced leader RNA can switch between two alternate structural forms. *Biochemistry* 1993, 32:5301–5311.
68. Markham NR, Zuker M. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol* 2008, 453:3–31.
69. Trelles-Salazar O, Zapata EL, Dopazo J, Coulson AFW, Carazo J. An image-processing approach to dotplots: An X-window-based program for interactive analysis of dotplots derived from sequence and structural data. *Comput Appl Biosci* 1995, 11:301–308.
70. Quigley GJ, Gehrke L, Roth DA, Auron PE. Computer-aided nucleic acid secondary structure modeling incorporating enzymatic digestion data. *Nucl Acids Res* 1984, 12:347–366.
71. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci USA* 2004, 101:7287–7292.
72. Washietl S, Hofacker IL, Stadler PF, Kellis M. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucl Acids Res* 2012, 40:4261–4272.
73. Serra J. *Image analysis and mathematical morphology*. London: Academic Press; 1982.
74. Barash D. A fundamental relationship between bilateral filtering, adaptive smoothing and the nonlinear diffusion equation. *IEEE Trans Pattern Anal Mach Intell* 2002, 24:844–847.

