

RNA Visualization

Andreas De Stefani*

Institute of Computer Graphics
Vienna University of Technology
Vienna / Austria

Abstract

Biological data of all kinds is proliferating at an incredible rate. If humans attempt to read such data in the form of numbers and letters, they will take in the information at a snail's pace. If the information is rendered graphically, however, human analysts can assimilate it and gain insight at a much faster rate.

RNA fundamentals, standard visualization methods (string, bracket dot, dot plot, mountain representation) and software tools which adopt these are presented in this paper. There is no RNA visualization showing all interesting information about RNA structures just as there is no software tool which provides all required RNA standard representations. Features and main drawbacks are discussed in this paper.

Keywords: RNA, Visualization, secondary structure, dot plot, mountain plot, circular representation

1 Introduction to Nucleic Acids

The DNA molecule, as the primary repository of genetic information in living systems, is constrained to be stable and predictably structured. RNA differs little from DNA in chemical terms, but by contrast contrives to exhibit remarkable conformational flexibility and functional versatility. The past few decades of intensive research have revealed, for example, that RNA physically conveys and interprets the genetic blueprint of every living cell; it performs essential structural roles in a number of molecular machines; its ability to form transient duplexes allows it to work as a switch; and it is likely to work as an essential catalyst in several biologically important reactions.

DNA is the basic hereditary material in all cells and contains all the information necessary to make proteins. DNA is a linear polymer that is made up of nucleotide units. The nucleotide unit consists of a base, a deoxyribose sugar, and a phosphate. There are four types of bases: adenine (A), thymine (T), guanine (G), and cytosine (C). Each base is connected to a sugar via a glycosyl linkage.

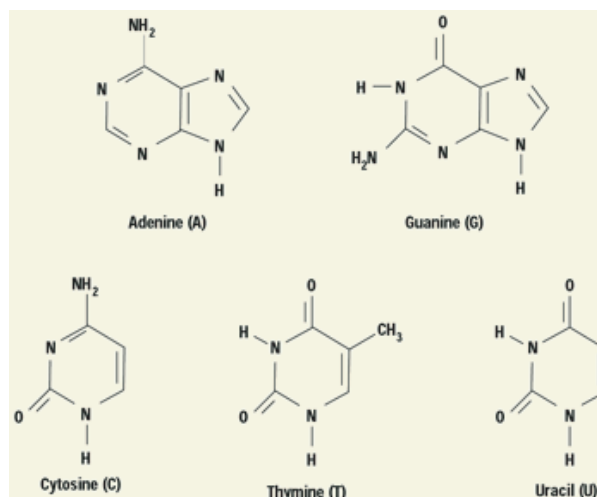


Figure 1: Bases that are found in DNA and RNA. Figures from [7].

In normal DNA, the bases form pairs: A to T and G to C. This is called complementarity. A duplex of DNA is formed by two complementary chains that are arranged in an anti-parallel manner.

DNA serves as the template for the synthesis of RNA. RNA is a polymer that contains ribose rather than deoxyribose sugars. The normal base composition is made up of guanine, adenine, cytosine, and uracil. See Figure 1.

RNA is the mobile form of genetic information, it is single stranded and can form complex and unusual shapes. The code is produced from one strand of the DNA by a process called "transcription". This produces mRNA which then is sent out of the nucleus where the message is translated into proteins.

There are several types of RNA, perhaps more than we are aware of. The synthesis of protein involves messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA), which are independent of each other and have different tasks to perform:

- The transfer RNA selects the amino acids.
- The messenger RNA dictates the order in which they are sequenced.

*andy@tbi.univie.ac.at

- The ribosome, which contains ribosomal RNA combined with protein, carries out the attaching process.

RNA molecules form a structure of helical regions interspersed with single stranded areas. This structure is important in the function of these molecules, and its understanding has already contributed to the understanding of processes such as the splicing of group I and group II introns, the functional role of rRNA in protein synthesis and the function of RNase P [2]. In the case of rRNA, secondary structure features can be helpful to fine-tune the alignment of sequences for phylogenetic studies. The secondary structure of RNA molecules can be studied using experimental, thermodynamical and comparative methods. Programs that calculate the most thermodynamically favorable structure such as mfold [6] produce connection data: a list of bases and of numbers indicating secondary structure interactions. In DCSE [13] the structural information is incorporated in the alignment by interspersing the sequence with special symbols denoting the start and end of structural features. A special "helix numbering line" contains the names for the helix strands, and indicates which are complementary. Although these forms of structural information are very useful, they cannot be used for publications as they are difficult to evaluate. Since the classical 2D drawing of the secondary structure is easier to grasp and more aesthetically pleasing, it is the preferred visualization for publications.

2 RNA and Standard-Representation Fundamentals

RNA is transcribed (or synthesized) in cells as single strands of (ribose) nucleic acids. However, these sequences are not simply long strands of nucleotides. Rather, intra-strand base pairing will produce structure motives. See Figure 2.

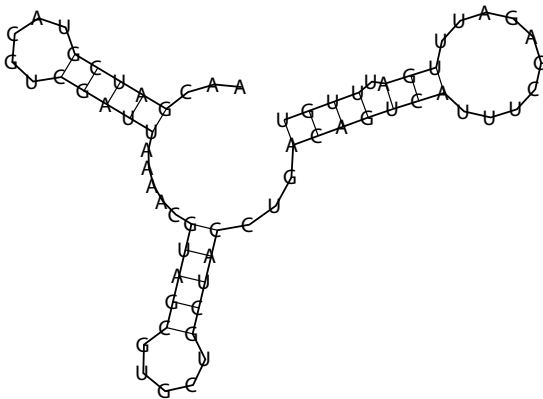


Figure 2: RNA sequence secondary representation. Created using the Vienna RNA Web Server based on ViennaRNA-1.5 [4].

The structure formation process of RNA can conceptually be partitioned into two consecutive stages. First, the specific sequence (the string of bases) or primary structure, is transformed into a pattern of complementary base pairings called the secondary structure. Second the secondary structure distorts, to form a three dimensional spatial structure or tertiary structure. It is hard to solve the structure prediction problem for RNA structures since the number of degrees of freedom of the RNA chain is very high.

2.1 String Representation

This is the simplest representation, it gives only information about the sequence of adenine (A), uracil (U), guanine (G), and cytosine (C). Adjacent letters mean that there is a connection between the two bases. No further information about base pairs or folding is provided, it shows the RNA sequence in its initial form, as a single strand.

Sequence in string representation:

```
AACGAUCGUAACGUCGAUUAAAACGUAGCGUGUCGUACCUACAGUCAUUUCGAGAUUUGAUUUUGU
```

2.2 Bracket Dot Representation

Bracket Representation: is a string of parenthesis and dots of length n. A base-pair between nucleotides i and j is indicated by an open bracket '(' at position i and a close bracket ')' at position j. Unpaired nucleotides are indicated with a dot '.'.

Sequence and structure in bracket dot notation:

```
AACGAUCGUAACGUCGAUUAAAACGUAGCGUGUCGUACCU
...((( ((( (.....) )))).....((( ((( (.....) )))))).
```

2.3 Classic Secondary Structure Representation

The secondary structure of RNA is formed by aggregation of base pairs of purine and pyrimidine bases. G and C, respectively A and U are complementary bases which can form strong hydrogen bonds, a weaker base pair, often referenced as "wobble" base pair, is also possible between G and U.

A secondary structure S is formally defined as the set of all base pairs (i, j) with i < j such that for any two base pairs (i, j) and (k, l) with i <= k the two following conditions hold :

1. i = k if and only if j = l.
2. There are no knots or pseudo knots allowed. For any two base pairs (i, j) and (k, l) the condition i < k < l < j or k < i < j < l must be satisfied.

The first condition simply means that each nucleotide can take part in at most one base pair. Several examples of tertiary interactions breaking this condition are known, including base triplets, G-quartets and A-platforms.

2.4 Linked Graph Representation

Secondary structure graphs as defined above can be drawn by placing the bases of a sequence equidistant to one another on a line. Pairing bases are connected by arcs, see Figure 3.

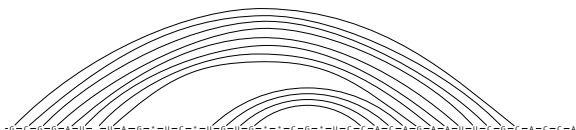


Figure 3: The secondary structure of tRNA^{Phe} in linked graph representation. Figure as seen in [3]

2.5 Circular Representation

A particularly easy way to draw secondary structure graphs was suggested by Ruth Nussinov [11]. The bases of the sequence are placed equidistant to one another on a circle and for each base pair a chord is drawn between the two bonded bases. Since the structures are un-knotted by definition, no two chords will intersect. See Figure 4 for a circular representation of tRNA^{Phe}.

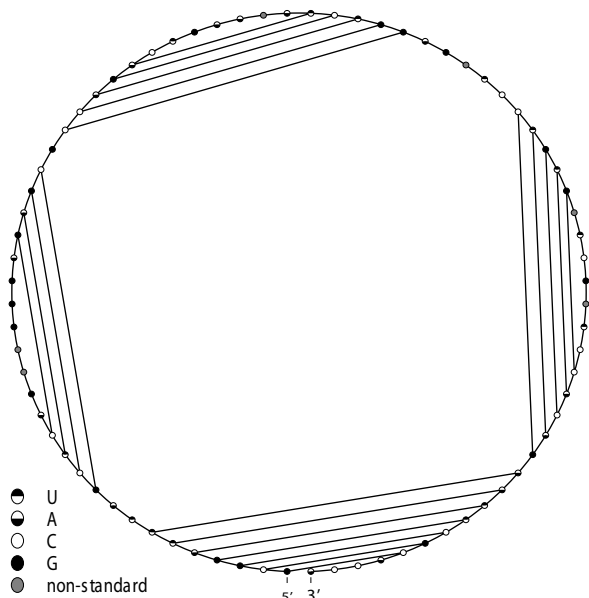


Figure 4: The secondary structure of tRNA^{Phe} in Circular representation. Figure from [3].

There is a very similar circular representation where the nucleotides are stretched out uniformly along the circumference of a circle in the same way, but the base

pairs are represented by circular arcs that link paired bases and meet the circle at right angles.

2.6 Dot Plot Representation

A dot plot shows the base pairs which can be built by the RNA, each possible base pair is represented by a square. The size and color (optional) of the dots represent additional information, the size gives information on the frequency at which the base are paired at the given point. The color, if available, shows how many different base pairs appear at this point. The upper right triangle shows the pairing possibilities of the entire sequence including sub-optimal structures, the lower left shows a single structure which returns the minimum free energy.

A dot plot is a two-dimensional graph in which the size of the dot at position (i, j) within the graph represents the probability P_{ij} of the base pair. Figure 5 shows the tRNA^{Phe} as an example. The plot is divided into two triangles. The upper right triangle contains the base pairing probability matrix (P_{ij}) ; the size of the squares is proportional to the pairing probability. The lower-left triangle displays the minimum free energy structure for comparison.

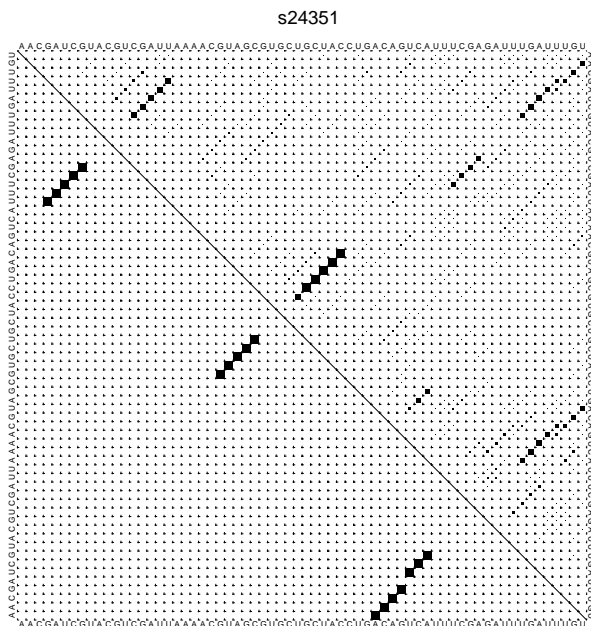


Figure 5: RNA Dot Plot. The lower left triangle shows the base pair probability, and the upper right the minimum free energy. Figure created using the Vienna RNA Web Server based on ViennaRNA-1.5 [4].

2.7 Mountain Plot Representation

Paulien Hogeweg, B. Hesper and Danielle Konings conceived a related graphical method for the comparison of RNA secondary structures called *mountain representation* [5, 9, 8] by identifying ' (, ') ', and ' . ' , with "up",

“down”, and “horizontal”, respectively. See Figure 6 for a mountain representation.

- *Peaks* correspond to hairpins. The symmetric slopes represent the stems enclosing the unpaired bases in the hairpin loop, which appear as a plateau.
- *Plateaus* represent unpaired bases. When interrupting sloped regions they indicate bulges or interior loops, depending on whether they occur alone or paired with another plateau on the other side of the mountain at the same height respectively.
- *Valleys* indicate the unpaired regions between the branches of a multi-stem loop or, when their height is zero, they indicate unpaired regions separating the components of secondary structures.

The height of the mountain at sequence position k is simply the number of base pairs that enclose position k ; *i.e.*, the number of all base pairs (i, j) for which $i < k$ and $j > k$. The mountain representation allows for straightforward comparison of secondary structures.

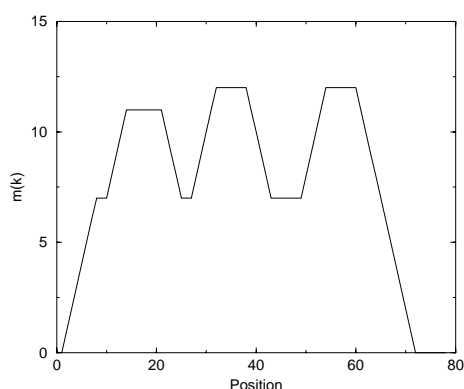


Figure 6: The secondary structure of tRNA^{Phe} in *mountain representation*. Figure from [3]

3 Special Requirements Visualizing RNA Sequences

Graphic representation of information containing RNA sequences can help the human analyst to find interesting and biologically relevant patterns.

Usually RNA sequences are about 10^2 - 10^4 bases long and have a complex structure, which requires visualizations allowing to compare structures and showing how conserved different structures are.

Classical representation for instance is not very useful if you are interested in how conserved (stable) the structure is, identifying pairing probabilities or for chronologic enumeration. Sometimes color is used to define how conserved (how stable) the nucleotides are at this position:

- *red* this base pair pairs always exactly with the same bases.
- *green* there is always a base pair, but not perforce always with the same base. This means that there could be a AU or a GC base pair, but not an unpaired base.
- *light green* There are two types of base pairs, but there is exactly one possible structure in which the base doesn't pair at this position.

4 Overview of Existing Tools

Although several programs exist that produce 2D structure drawings, they share some of the following problems: Most are too tightly coupled to an energy minimization prediction program to be of general use. Furthermore, the user cannot easily change the produced layout: Much effort has been put into automatically producing a layout where none of the helices overlap, but this often does not properly emphasize similarities in structure because of insertions or deletions in less conserved areas. Other common problems are limitations to the size of a molecule that can be displayed, and the inability to handle complex structural elements such as pseudo-knots.

RNADraw and RNAViz (described in section 4.1 and 4.2) are some of the most advanced free RNA structure drawing tools. RNAViz is focused on structure drawing and layout, RNADraw allows calculation and visualization of different representations such as optimal structure, basepair-probability matrix and heat curve.

4.1 RNADraw

4.1.1 General Description

RNADraw is a program for RNA secondary structure calculation and analysis. RNADraw is Freeware. It offers RNA optimal structure / basepair-probability matrix / heat curve calculation on Intel x86 compatible computers, providing a consistent user interface with many possibilities to view, print, import/export and edit calculation results. Although RNADraw has old windows (win 3.11) look and feel it has a user-friendly user interface supporting drag and drop, right-button menus containing data manipulating functions.

Via a toolbar button or main menu option it is possible to modify/save/load all energy parameters used by the calculational algorithms in RNADraw.[10]

4.1.2 System Requirements

To install and run RNADraw on your system you will need the following:

- An Intel x86 (or compatible) computer (386+)
- Windows 95, Windows NT or Window 3.1/3.11 with Win32s version 1.2 or later

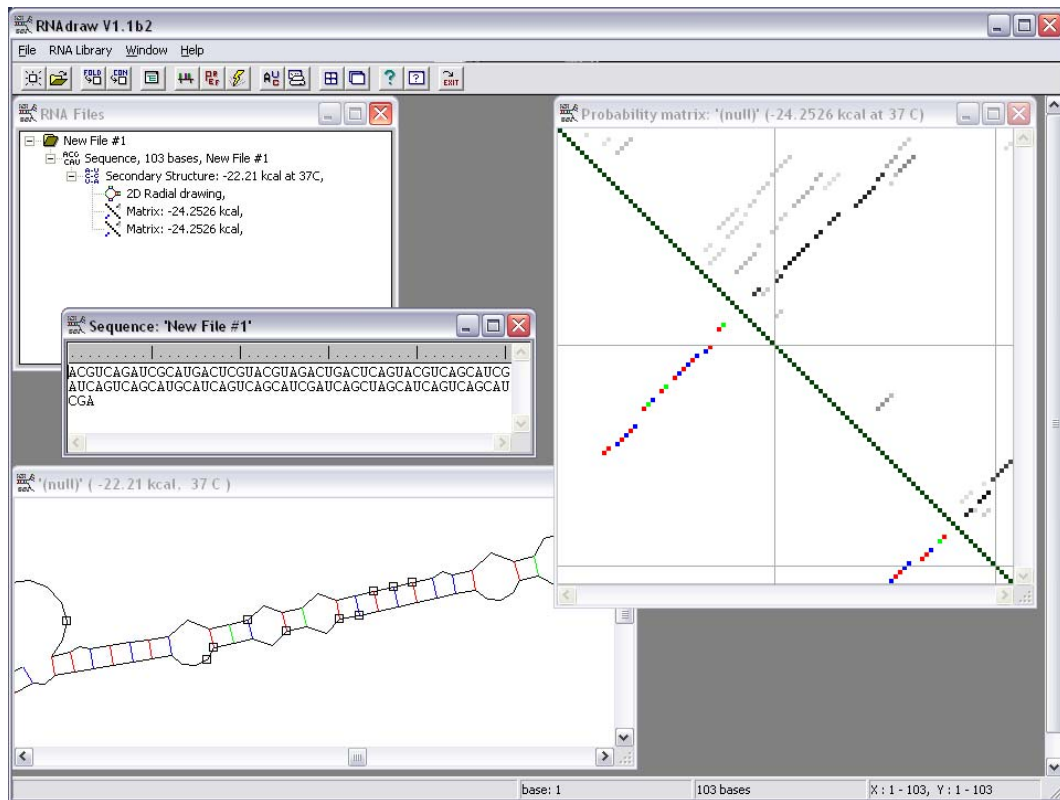


Figure 7: RNA Draw V1.1b2 Screenshot under Windows XP.

- at least 8Mb of RAM, the more the better
- 500 Kb free disk space

4.1.3 Features and Visualizations

Structure Calculation

In addition to data visualization, RNADraw offers calculation using the minimum energy structure prediction algorithm which was ported from the Rnafold program included in the Vienna RNA package.[4]

Multiple temperatures/temperature ranges for structure calculation can be set, base-pairing constraints added and other calculation flags can be easily set. The coordinates for the returned minimum energy structure are calculated with another ported program, i.e. Naviem by R. Brucoleri.[1]

Structure View

RNADraw visualizes the sequence in classical secondary structure representation (Hairpin), described in Section 2.3.

You can adjust which information will be shown, such as bases, labels, basepairs etc. You can easily zoom, pan and rotate the viewed structure by pressing the mouse buttons in different parts of the window. The mouse cursor changes appropriately.

The associated right-button menu includes several manipulation functions. For example, "Mark Bases" lets

you search for and mark sequence tokens in numerous ways. "Adjust Basepairs" allows the pairing / unpairing of individual basepairs to let you inspect structure / energy changes in different conformations.

It is possible to apply basepair probabilities to the structure view, resulting in thicker lines for higher probabilities.

Matrix Calculation and view

The Matrix view window can be opened from the "Open Matrix" menu option, which is available on the structure view / structure entry right-button menu. The top right triangle shows the probability matrix, the bottom left triangle shows the minimum energy structure. The probabilities are grayscale-coded (higher probability = darker gray). Is possible to zoom up and pan with the scrollbars, show alignment lines etc.. Moving with the mouse over the matrix displays the base-pairing probabilities of the underlying base-pair.

You can manually pair/unpair basepairs in the structure by clicking with the left mouse button on a possible base-pair in the lower left triangle. If you have the corresponding structure window open, you can directly monitor structure changes resulting from your basepair edits.

From the right-button menu of the matrix window you can open the matrix probability histogram window. Here, you can see the probability frequency distribution of the matrix.

It is possible to "extract" structures from the probability matrix, applying the current cutoff frequency. This allows to view structures of certain "probability levels", and to compare them with thermodynamically optimal structures.

Both above windows can be exported and printed in the same manner as with structure views .

Heat calculation and view

The specific heat calculation algorithm in RNAdraw was ported from the Rnaheat program included in the Vienna RNA package. [4]

The heat view can be opened from the "Show Heat" option, which is available on the sequence entry right-button menu. The scroll bar at the bottom of the window lets move the above blue marker back and forth to see individual heat at specific temperatures in the status bar.

Import/Export functions

- load Win/UNIX text file into sequence editor
- paste from the clipboard
- GENbank data files can be imported.
- Rnafold output files can be used for input
- export sequence as win text file
- all structure/matrix/heat windows can be printed and exported as bmp files

4.1.4 Strengths

- easy installation
- sequence window containing all open files and calculated structures in a tree view.
- easy and fast input of sequences
- many inport functions
- colored Matrix plot
- multiple windows

4.1.5 Weaknesses

- no toolbar for editing
- poor interactivity
- no ps or svg output
- no alternative structure representation
- not cross platform

4.2 RNAViz

4.2.1 General Description

RNAViz is a user-friendly, portable, windows-type program for producing publication-quality secondary structure drawings of RNA molecules.[12] Drawings can be created and the layout of a structure can be changed easily. Display of special structural elements such as pseudoknots or unformatted areas is possible. Sequences can be automatically numbered, and several other types of labels can be used to annotate particular bases or areas. Although the program does not try to produce an initially non-overlapping drawing, the layout of a properly positioned structure drawing can be applied to newly created drawings using skeleton files. In this way a range of similar structures can be drawn with a minimum of effort. Skeletons for several types of RNA molecule are included with the program.

4.2.2 System Requirements

RNAViz needs a modified version of Tcl and several extensions. Binary distributions of the modified Tcl and the RnaViz package are available for Linux and MS Windows 95 on the rRNA server at URL <http://rrna.uia.ac.be>.

The sources are also available there for people who want to port the code to other systems.

4.2.3 Features and Visualizations

User Interface

The graphical user interface has mostly native look and feel. Some of the features are zoom, undo/redo, printing, copy/paste ...

Each base is a canvas object that can be individually addressed, tags on bases keep track of which part of the structure the base belongs to. RNAViz allows to draw multiple structures on one page. Depending on the selection mode, individual bases, segments, helices, trees or structures can be manipulated.

Arranging the layout of a structure

RNAViz draws secondary structures in classical representation (hairpin), it doesn't provide additional representations.

Each structure on the page consists of a number of individual objects, such as bases, base pair connections or helix names. An object can be selected by clicking on it using the first mouse button. The current structure can be moved as a whole by clicking outside the structure and dragging. Extra objects can be added to or removed from the selection by clicking on them with the "Control" key pressed.

A structure on the page can be rearranged quickly by clicking on a base or on the selection and dragging it to a different position. When the selection is released, the bases connecting the selection to the rest of the structure

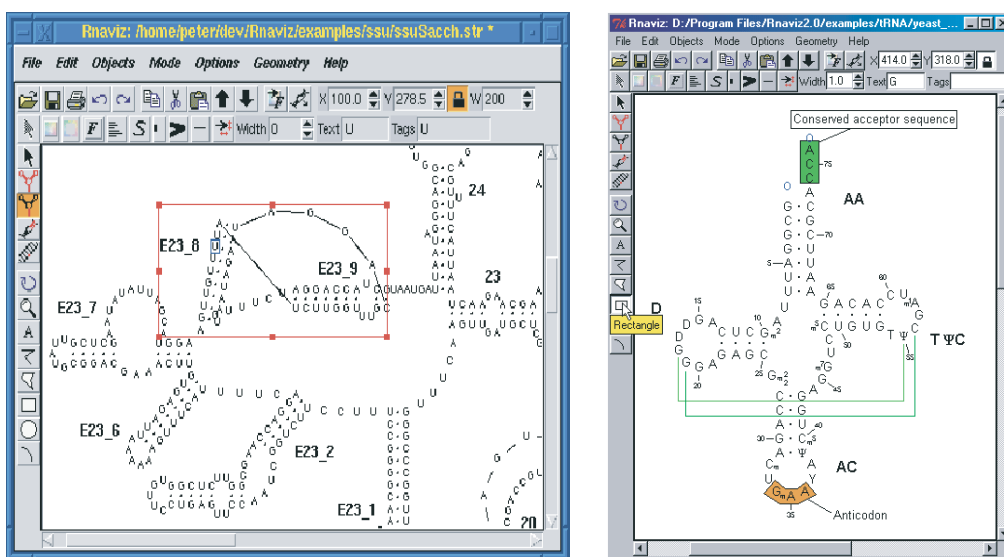


Figure 8: RNAviz2 Screenshot under Windows and Linux.

will be rearranged so as to maintain a correct structure drawing.

There are different selection modes allowing either selecting the entire helix or the single base simply by clicking on a base. This makes possible every special arrangement of objects. Other selection modes are the "select tree" and "select sub-element" modes, which automatically select a tree or a set of segments of a helix.

Flipping helices

It is often interesting to draw the figure or some helices of the figure counter-clockwise. RNAviz allows the user to flip the entire drawing or any helix to make this possible. This is often necessary in order to nicely arrange pseudo knots.

Labeling a structure

All helices are automatically labeled with their helix name. This label will move together with the helix. Another type of label is the base numbering. Base numbers can be added automatically at a specified interval starting from a specific base. Base numbers can also be individually added or removed. RNAviz also contains a limited drawing component. Several types of objects such as texts, rectangles, ovals, lines and polygons can be created and edited. Any of these objects can be used as a label by linking them to a certain base.

Configuring objects

Each object has properties such as font, text, color, line width and position. The "Configure Objects" dialog offers a versatile interface to change these properties for any object or groups of objects. Several parameters can be set that limit property changes to objects fulfilling certain cri-

teria. This way the properties of either the currently selected objects, the objects of the current drawing or all objects can be changed. In addition, the changes can be limited to a specific type of objects such as bases, base pairings, base numbers, helix names or labels.

4.2.4 Strengths

- publication-quality secondary structure drawings of RNA secondary structures
- recognises CT, RNAML and DCSE alignment formats
- multiple structures on page
- simple but powerful WYSIWYG editing using different selection modes
- allows display of pseudoknots
- free choice of fonts, colors, linewidths for any object
- graphical objects for annotation (rectangle, oval, lines, text)
- linking of graphics, text to certain bases
- independent scaling of structure drawings

4.2.5 Weaknesses

- no alternative views
- requires TCL (in older versions a patched TCL version)
- difficult to install

5 Conclusions

In this paper, we presented different visualization methods for RNA sequences, i.e classic, linked graph, circular, dot plot and mountain plot representations, and described their main functionality. There are many software tools which allow to calculate and display RNA sequences using the discussed visualization techniques, two of them were analyzed in more detail.

There is no known software tool which uses all of the representations discussed in this paper, the classical representation is mostly used and many tools provide a dot plot representation as well. Only few tools are cross platform, the few which are cross platform are mostly written in C and have to be recompiled, which is not always a simple task.

All of the tools have several drawbacks and there is no tool which provides all RNA representations. If you looking for a tool which allows calculation and you only need classic representation or maybe a matrix (dot) plot, RNADraw should work for you.

If you are interested in a tool which allows to change the layout of the sequence manually, RNAViz would be the better choice. The tool has a nice interactive interface which allows drawing and arranging of a structure, annotation and labeling, but provides only visualization in the classic representation. No other representations like mountain or dot-plot are available. Another drawback is that installation of RNAViz under windows requires additional libraries and is not trivial.

References

- [1] R. Brucoleri and G. Heinrich. An improved algorithm for nucleic acid secondary structure display. *Computer Applications in the Biosciences* 4, pages 167–173, 1988.
- [2] A.E. Dahlberg. The functional role of ribosomal rna in protein synthesis. *Cell* 57, pages 525–529, 1989.
- [3] Martin Fekete. Prediction of rna secondary structure using parallel computers. pages 10 – 14, 1997.
- [4] I. L. Hofacker, W. Fontana, P. F. Stadler L. S. Bonhoeffer, M. Tacker, and P. Schuster. Vienna RNA Package. <http://www.tbi.univie.ac.at>, 2003. (Free Software).
- [5] Pauline Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nucl. Acid. Res.*, 12:67–74, 1984.
- [6] A.B. Jacobson and M Zuker. Structural analysis by energy dot plot of a large mrna.
- [7] Kevin D. Jones. Membrane immobilization of nucleic acids, part 2: Probe attachment techniques. <http://www.devicelink.com/ivdt/archive/01/09/002.html>, 2001.
- [8] D.A.M. Konings. Pattern analysis of RNA secondary structures. *Proefschrift, Rijksuniversiteit te Utrecht*, 1989.
- [9] D.A.M. Konings and P. Hogeweg. Pattern analysis of RNA secondary structure, similarity and consensus of minimal-energy folding. *J. Mol. Biol.*, 207:597–614, 1989.
- [10] Ole Matzura and Anders Wennborg. Rndraw: an integrated program for rna secondary structure calculation and analysis under 32-bit microsoft windows, 1996.
- [11] Ruth Nussinov, George Piecznik, Jerrold R. Griggs, and Daniel J. Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35(1):68–82, 1978.
- [12] Jan Wuyts Peter De Rijk and Rupert De Wachter. Rnaviz2: an improved representation of rna secondary structure. *Bioinformatics* 19, pages 299–300, 2003.
- [13] Peter De Rijk and Rupert De Wachter. Dcse, an interactive tool for sequence alignment and secondary structure research. *Computer Applications in the Biosciences* 9, pages 735–740, 1993.