

FAST 1.0

USER TUTORIAL

Copyright Phillip S. Pang, 2011

FAST REQUIRES 4 FILES.

1) A control file, containing the peak areas and fragment lengths extracted from the trace for the RNA run only in the presence of DMSO (no electrophile).

Format: fragment length (tab) peak area EXAMPLE: 45.3 1294

NOTE: peak areas should be in points, not base pairs (see Peakscanner for details).

2) An experimental file, containing the peak areas and fragment lengths extracted from the trace for the RNA run in the presence of NMIA/1M7 (electrophile).

Format: fragment length (tab) peak area EXAMPLE: 45.3 1294

NOTE: peak areas should be in points, not base pairs (see Peakscanner for details).

3) The sequence file, containing the sequence of the RNA.

Format: FASTA format

EXAMPLE:

>sequence name

ATGTGTGTGACGCTGAGTGCGAG

NOTE: sequence should begin at the 5' end of the RNA of interest, and extend all the way to the 5' end of the primer. When working on large RNAs, the sequence file will be much longer for primers at the 3' end of the RNA. PURPOSE: this allows the program to assign NT positions relative to the entire RNA, rather than specific to the primer

4) A ladder file, containing the peak areas and fragment lengths extracted from the trace for the RNA run in the presence of ddGTP (or any other terminator).

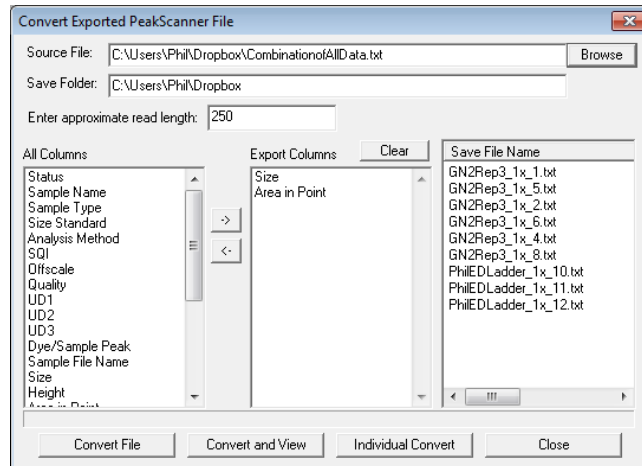
Format: fragment length (tab) peak area EXAMPLE: 45.3 1294

NOTE: peak areas should be in points, not base pairs (see Peakscanner for details).

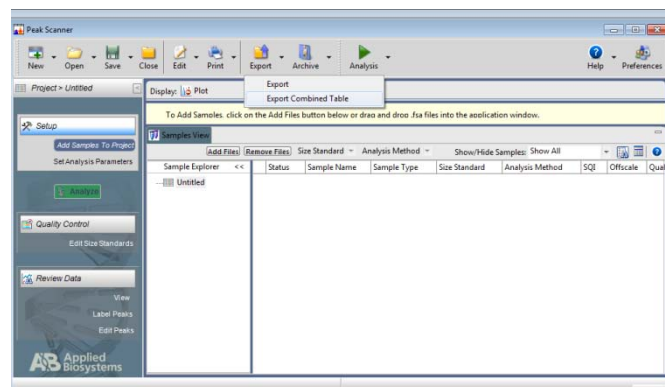
FAST GENERATES THESE FILES FOR YOU.

FAST>Convert File

The control file, experimental file and ladder file, can be generated using FAST>Convert File. This converts the output from PeakScanner into the proper format for FAST.



How do I get PeakScanner to output my data? PeakScanner>Export>Export Combined Table. FAST, like all computer programs, is limited by the principle of, "junk in, junk out". **Before exporting your data, it must have been properly analyzed and each trace should have been cursorily inspected for quality.**



What does "enter approximate read length:" mean? The probing reaction under single hit conditions has a maximum read length. The higher the concentration of the probing electrophile (e.g. NMIA), the shorter the read length, but the higher the signal for each peak. We have obtained read length between 200 to 400 nucleotides with 100mM NMIA. However, in the large majority of cases, read lengths of ~250 are more typical.

How accurate do I need to be? Not very accurate at all. This is a guesstimate, and if you are off by 50 to 100 nucleotides it does not matter. However, because the program often uses the median statistic, the inclusion of many non-quality data points will

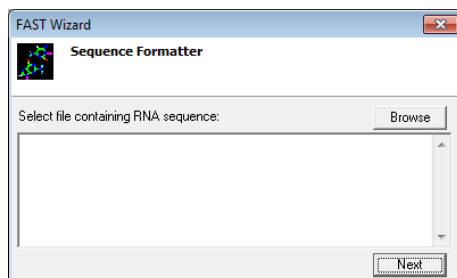
eventually cause the algorithm to require more user input when later generating a ladder assignment table.

How do I know what my read length is? The easiest way is to simply use 250 if you're using an NMIA concentration of 100 mM. The next fastest, most simple way is to "eyeball" your trace, and see where the signal goes from high quality to low quality. A more quantifiable way is to perform a posthoc analysis: having selected a long read length, and completed the FAST wizard, the program will generate a file called "reactivity.txt". This is the experimental data after it has been normalized to the control data, and the control data subtracted from it. A plot of fragment length versus peak area will indicate the NT region where the points are significantly scattered around $X=0$.

What if I have files with different read lengths? If all the traces have been generated using the same concentration of probing electrophile, then minor differences are not concerning, and the same read length can be used. This number merely demarcates the outer limit of what you consider to be quality data in the trace (by fragment length). At the end of the wizard, you will be able to refine this number further for each specific trace.

FAST>Sequence Formatter

The sequence file can be generated using FAST>Sequence Formatter



Select a FASTA formatted RNA sequence, that contains the sequence for the entire RNA. If for example, you are determining the structure of the first 1000 nucleotides of HCV, using six different primers spaced 200 NTs apart, then the 1000 NT RNA sequence should be selected.

The program then asks you for a primer file. This primer file can be formatted either with commas or returns separating your primers. Continuing on with the example, given six primers for the above 1000 nucleotide HCV RNA, create a simple text file of the primer sequences that will be used in the shape analysis, separated by commas or returns.

EXAMPLE: GCGTGCGAGTCGT, CGTGAGTGTGCCGAAAA, TTGTGTGAGA

The program finds the complementary region in the RNA sequence, and outputs a FASTA formatted file for each primer, starting at the 5' end of the RNA and continuing to the 5' end of the given primer.

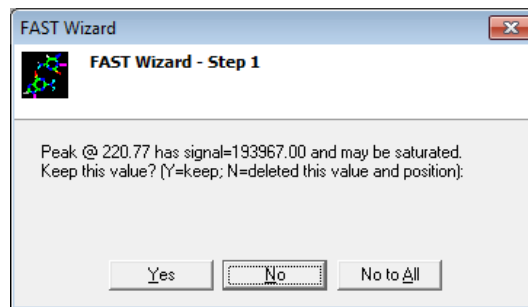
YOU ARE NOW READY TO START THE FAST WIZARD: FAST>WIZARD

The fast wizard contains many dialog boxes that give you a number that has already been entered. The user simply needs to click next. These dialog boxes are advanced features, allowing an expert user or programmer to test the effect of altering the default parameters. They should remain unchanged by the majority of users.

If multiple primers have been used, each primer has its own set of four files: a control (DMSO) file, an experimental (NMIA) file, a ladder (ddGTP) file, and a sequence file.

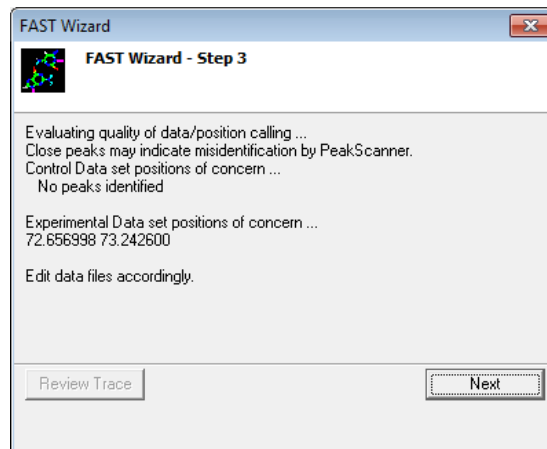
Step 1: select the control file generated for the primer

If saturated peaks exist in the data, you will be asked whether or not to include these. The default is to remove all saturated peaks. This is an advanced feature, and "no to all" should be selected.



Step 2: select the experimental file generated for the primer

Step 3: assessing the quality of the data sets.



Peakscanner algorithmically detects peaks and integrates their areas. While accurate 99% of the time, especially when run with the parameters described below, mistakes are still possible. In this example, the program is indicated that the peaks at 72.65 and

73.24 are close together. Peaks that are close together often indicate either noise in the trace, a mis-called peak, or some other artifact in the trace. Sometimes, however, close peaks are simply too real peaks close together.

Fast solution: delete all peaks of concern. (MAKE SURE YOU SAVE the file after you make any changes. Make sure you are making changes to the correct file. The program opens the experimental data set in front of the control data set.) FAST contains algorithms that can adjust for missing peaks, and the nucleotides corresponding to these peaks are assigned a SHAPE reactivity value of -999, which is read by the RNAstructure program as no data.

Advanced solution: Using PeakScanner, inspect the traces of the control, experimental, and ladder data, simultaneously, in the indicated area. Given the spacing of peaks before and after the area of concern, and the shape of the peaks themselves, it should be obvious what are not the peaks represent 2 actual separate peaks, should be combined, or one or both of the peaks deleted. Note: failure to properly combine or delete a peak in an area of concern may generate a flag at the very end of the program indicating that the shape data has been limited. In this instance you will see a shape file with many -999 (no data) positions.

PEAK SCANNER PARAMETERS (Updated 3-26-2011):

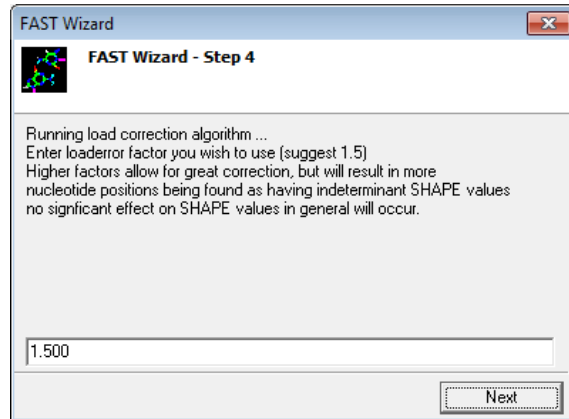
Analysis Method: Start with: Sizing Default - NPP.

- 1) Peak Detection>Range>Sizing>Limit Range 50 to 490 (if using the ROX 500 size standard); The local southern method requires two internal standard standards below and above, for accurate sizing.
- 2) Peak Characteristics>Blue (assuming you are using a 6FAM labeled primer)>Polynomial Degree = 2; Peak Window Size=31.
- 3) Minimum Peak heights>Blue=30; RED=50.

Sizing Standard: (assuming you are using a ROX500 size standard from ABI)

Remove the 250 and 340 size standards, as these are thought to be sensitive to minor instrumental temperature fluctuations.

Step 4: Loaderror Factor: This is an advanced feature, and the default value of 1.5 should be used.

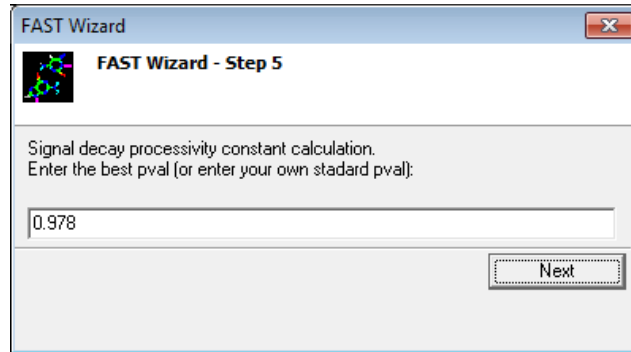


Advanced information: FAST experimental to control data sets in the following manner: A) each control datapoint is subtracted from its corresponding experimental data point, for all data points. B) the median of these differences is calculated. C) the experimental data set is multiplied by a factor ranging from near zero to as high as four, and a step size of .001. For each step, the median of the differences is again calculated. D) the algorithm attempts to determine the multiplication factor that minimizes this median difference.

Because the distribution of points and the experimental data should be skewed relative to the distribution of points and the control data, this has the potential to potentially result in overcorrection, especially if the signal strength for the experimental and control data are significantly different. Practically, this is not a significant issue, as it will merely result in more positions not being assigned a shape reactivity values. The load error factor allows the user to eliminate sets of control and experimental data points from the median calculation, when the experimental data point has very high reactivity. Increasing this value tells the program to make no adjustments for the skew in the experimental data set. Decreasing this value tells the program to be very aggressive about eliminating skew, but at the risk of using many fewer points to normalize the experimental to the control data.

Step 5(a): Click Next, agreeing to the load correction factor (not to be confused with loaderror factor) calculated by FAST.

Step 5(b): signal decay process the constant calculation. See the manuscript by Pang et al. for detailed description of this parameter.



This parameter corrects for any potential general trend in the data, where shorter fragments have higher intensities than longer fragments - a phenomenon known as signal decay. Click Next.

Step 6: enter the sequence file **for the primer** generated previously using sequence formatter.

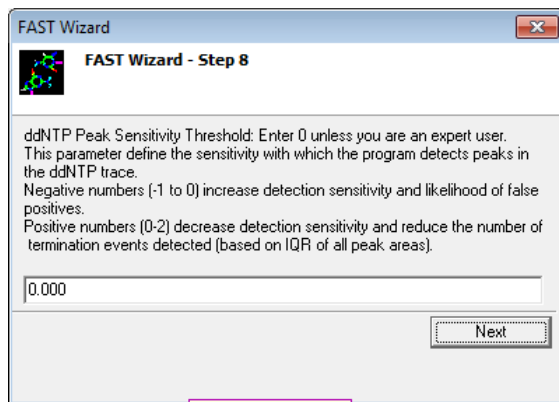
DO NOT select the sequence file for the entire RNA (if using multiple primers.)

Step 7: using a ladder file to generate a ladder_{position} file

The input for this step is the ddGTP ladder file. A ladder_{position}.txt file -- the name under which it is saved (and also referred to as the ladder assignment table) is a table containing a list of fragment lengths and their corresponding nucleotide positions. This file is the calibration file that will subsequently be used to assign nucleotide positions to the experimental and control data, given their fragment lengths. A ladder file will need to be generated for each section (as defined by the read length for each primer) of the RNA of interest. Once generated it can be reused for all subsequent analyses of that same RNA region, IF AND ONLY IF the actual RNA molecule is the same. (In other words, if you make a mutation to your RNA, you must generate a new ladder file.)

This step, along with step 3, are the only semi-automated steps in FAST. The majority of user time occurs at this step. When the ddGTP ladder file is of excellent quality, this step may take less than a minute. When the ladder file is of poor quality, or contains many saturated peaks due to degradation or reverse transcriptase pausing, this step may take 15 to 30.

Step 8(a): peak sensitivity threshold. This is an advanced feature, and users should click NEXT.



Advanced information: ddGTP letter generation by reverse transcription is not as clean a process as DNA sequencing. Aberrant peaks are not uncommon. Therefore, this parameter is set to use only the top 25% of peaks (by area) in the ddGTP trace --because specificity is critical, and sensitivity less important. The number entered refers to multiples of the interquartile range (IQR, as defined by subtracting the first quartile from the third quartile). Thus, entering the -1, would put threshold such that the top 75% of peaks (by area) would be included in the analysis. Entering 1 would put the threshold such that only peak areas greater than the third quartile plus one IQR. Last sensitivity is unlikely to be required. Greater sensitivity will increase the number of ddGTP termination events detected, but at the risk of introducing aberrant peaks that may confuse the FAST algorithm, resulting in a ladderposition file containing many "REVIEW THIS REGION" Flags.

Step 8(b): Reviewing the ladderposition file.

Areas of concern are demarcated in the fourth column, with "REVIEW THIS REGION". This indicates regions where the algorithm was not able to confidently assign a nucleotide position to the fragment length. Column 3 and four must contain numerical data in order for the algorithm to proceed; thus, all "REVIEW THIS REGION" flags must be addressed by either deleting the row, or clicking "approve" -- which then changes this text value to a numerical value.

For detailed description of how to "review this region", please see the video tutorial. Briefly, the trace for the ddGTP letter is viewed in the region of concern. Knowing that, for example, fragment 181.94 has been assigned a nucleotide position of 183, we would then determine whether it is reasonable that the peak at 190.81 should be assigned the nucleotide position of 193; especially as position 195.11 has been assigned a nucleotide position of 196. By simply counting the intervening number of peaks between these assignments, it should be obvious whether or not the assignment at position 190.81 is correct. If it is difficult to identify the intervening peaks between assignments, call up the trace for the experimental arm beneath the trace for the ddGTP ladder; this often helps to count the intervening peaks. To aid in the counting process, you can click on the button "sequence" which calls up the reverse complement and highlights in red all the G's of interest.

No	Fragment Size	NT Position	Delta	R\Angle Distance
1	51.20	54	2.799	0.514
2	62.75	66	3.248	1.564
3	63.93	67	3.073	1.087
4	68.84	72	3.161	1.318
5	71.59	75	3.412	2.091
6	80.40	84	3.597	2.774
7	81.64	85	3.358	1.908
8	89.35	93	3.650	2.990
9	90.64	94	3.358	1.909
10	91.96	95	3.041	1.010
11	94.27	97	2.729	0.401
12	95.43	98	2.571	0.199
13	97.68	100	2.322	0.023
14	98.84	101	2.162	0.002
15	99.98	102	2.023	0.042
16	101.04	103	1.956	0.081
17	106.70	109	2.297	0.014
18	108.84	111	2.161	0.002
19	109.93	112	2.071	0.022
20	116.46	119	2.542	0.169
21	131.15	134	2.848	0.602
22	136.17	139	2.825	0.560
23	137.32	140	2.678	0.329
24	141.65	144	2.350	0.033
25	145.03	147	1.970	0.072
26	154.64	156	1.358	0.993
27	159.94	161	1.064	1.813
28	163.73	165	1.271	1.210
29	166.85	168	1.148	1.552
30	176.48	178	1.518	0.651
31	177.61	179	1.385	0.929
32	181.94	183	1.059	1.827
33	190.81	193	2.190	REVIEW THIS REGION
34	195.11	196	0.889	2.415
35	197.07	198	0.929	2.270

Take-home point: if there are only a few rows that require review, and no significant gaps between assign nucleotide positions (i.e a gap of over 20 nucleotides between subsequent assignments), these rows can simply be deleted without any effect.

While in general, rows which need to be reviewed are few, it has occurred that the vast majority of rows require review. This is likely due to one of the two following situations. First, the sequence contains a large region likely over 30 nucleotides in length, without any G's. This makes it challenging for the algorithm to properly assign the next G after this large region. Second, the trace contains a region of saturated peaks that are the result of degradation or reverse transcriptase pausing -- such a region of 5 to 10 peaks may also confuse the algorithm.

In these scenarios, while time-consuming, it is still possible to review each region as indicated above, and either correct the assign nucleotide position or the fragment length for that nucleotide position.

Note: Because the Local Southern method is used to assign NT positions to peaks, there must exist at least two assigned ddGTP peaks below and two assigned ddGTP peaks above any experimental peak being assigned a NT position. Thus, SHAPE files are automatically truncated to meet these criteria.

Advanced information: At this juncture the ladderposition file is merely a calibration or assignment file -- the location of the G's in the RNA sequence correlated to their

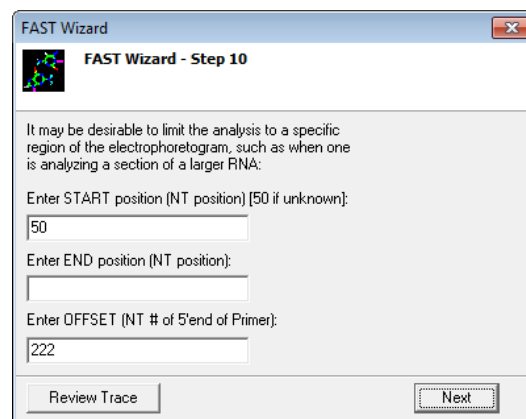
fragment lengths. This calibration file is then used as size standards in a Local Southern, to assign nucleotide positions to peaks in the experimental and control data sets. As the Local Southern method is indifferent to how the calibration file is generated, there is no reason to not include fragment length:nucleotide position rows that correspond to non-G nucleotides. In other words, when correcting row 33 in this example, I could either change the nucleotide position 193 such that it matches the correct nucleotide position for fragment length 190.81; or conversely, assigned position 193 the proper fragment length. In this example, the correct nucleotide position for fragment length 190.81 is 192. Although 192 is not a G, this does not matter.

Step 7(repeated): Having already generated a ladderposition file, click No.

Advanced information: this step merely provides flexibility for advanced users who wish to regenerate a ladderfile or alter the name of the ladderfile.

Step 9: Select the ladderposition file you created in step 8.

Step 10: generate the final SHAPE file



The numbers here refer to the nucleotide numbers counting from the 3' end of the RNA (counting from the 5' end of the primer). This dialogue offers the user one more opportunity to truncate the users data, based on the quality of the electrophoretic trace. At the start of this tutorial we used "Convert File" and "Approximate read length" to perform a similiar function. The point of this dialog boxes to allow the user to now refine for each individual trace, a specific read length.

When should this number be different than the number entered for approximate read length? Imagine that you have a trace that contains a series of highly saturated peaks around nucleotide 150, despite your read length being theoretically 250. Looking at the trace, it is obvious that all peaks after 150 are tiny, ragged, and of poor quality. This step would allow you the opportunity to enter 150 rather than 250, to avoid having poor quality data in the SHAPE file.

"OFFSET" is calculated for you by the program, and is based on the length of the sequence file for the specific primer -- when using multiple primers on a large RNA, this ensures that shape reactivities are assigned nucleotide positions relative to the entire large RNA.

NOW WHAT?

Combing SHAPE Data

If the user has generated multiple SHAPE files for different regions of the same RNA, then these SHAPE files can be combined using excel or any simple text editor.

Regarding overlap: overlap in SHAPE data, from primers that are near one another is advantageous. I have found that usually the first 10-20 NTs of SHAPE data (nearest the primer) is less accurate than the SHAPE data generated from an upstream primer. Conversely, I have found that where the quality of the SHAPE trace obviously degenerates, this is where I truncate the SHAPE data for that primer.

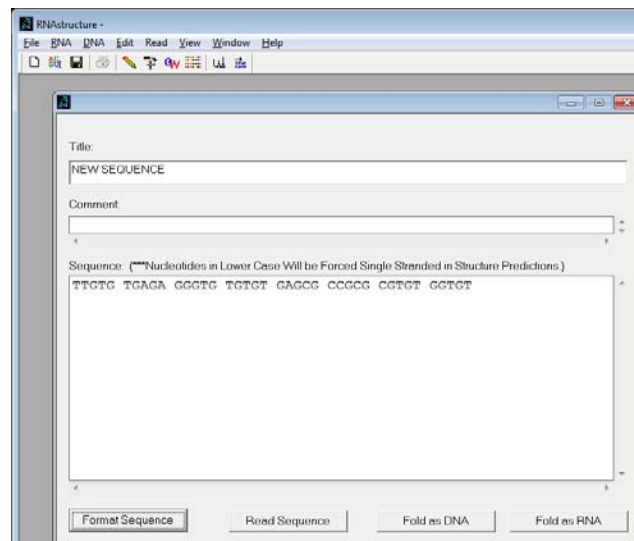
Using RNAstructure

The art of using RNAstructure with SHAPE constraints is beyond the scope of this tutorial. RNAstructure contains an extensive manual, and descriptions of how to incorporate SHAPE information.

Briefly:

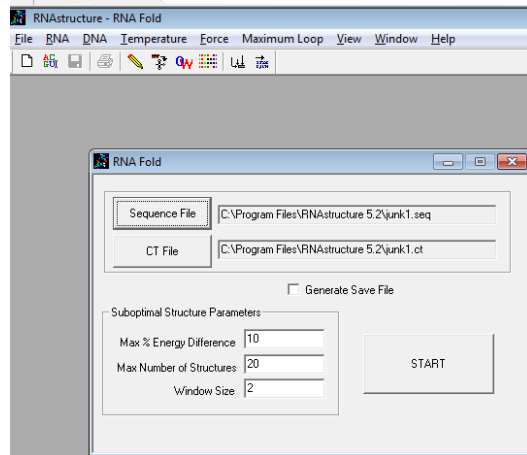
1) FILE> NEW SEQUENCE

Enter the sequence for the entire RNA molecule (NOT JUST THE SEQUENCE FOR THE SPECIFIC PRIMER)

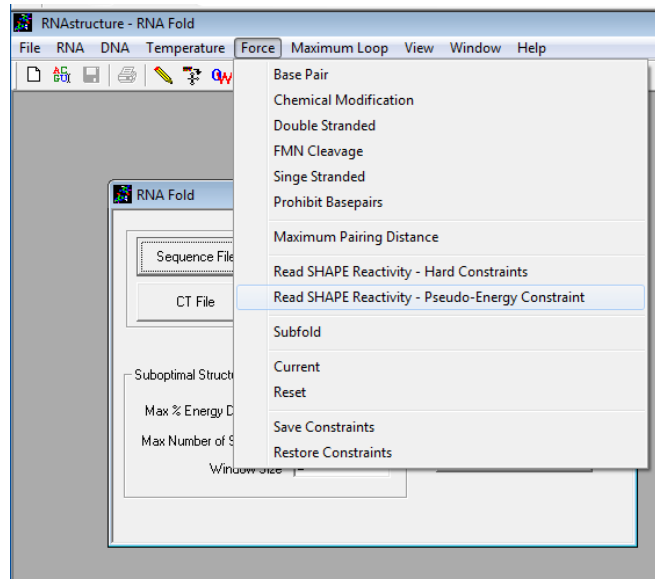


2) Click "Fold as RNA"

3) Select the Sequence you just saved by clicking "Sequence File"



4) THEN select "FORCE>READ SHAPE REACTIVITY - pseudo-energy constraints"



5) THEN click "START"

I have found that RNAview is a great program to quickly view RNA structures; I have found RNAviz to be excellent for making publication quality images. The ADD tags feature allows you to color code your structure according to SHAPE reactivity.

>END