

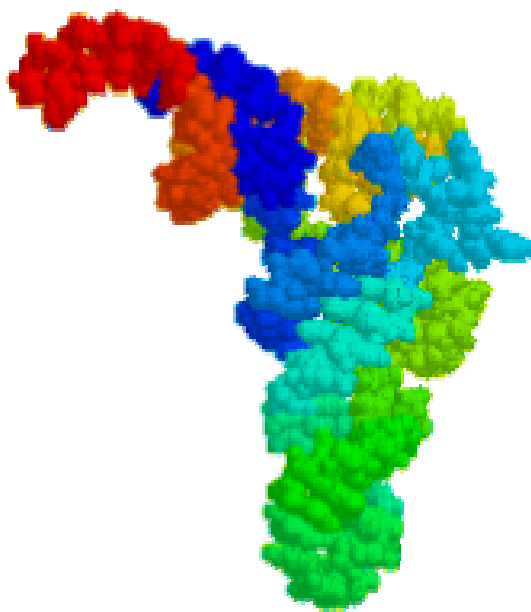
Introduction to Computational Biology

Team Project Report

RNA HEAT

RNA Helix Elimination and Acquisition Tool

Final Report



Team **MATRIX**
Gurushyam Hariharan
guru@ece.utexas.edu
James Zhang
jyzhang@mail.utexas.edu
Angie Li
alaysia@cs.utexas.edu

Contents

	Page
Biology Abstract	1
Introduction	2
Results	3
Conclusion	4
References	5
Appendix A: Tasks Listing	6
Appendix B: User Documentation	8

1. Abstract

A key problem in RNA folding is to understand, from an RNA molecule's primary structure, what properties allow for the formation of specific secondary structures, like those found in nature, in preference over other possible secondary structures. Study of secondary structure concentrates on observing which regions of an RNA molecule are base-paired to form helices, and which regions are free. The number of possible helices that can form increases exponentially with the length of the RNA, and the formation of non-Watson-Crick pairings among RNA bases further complicates this study. However, we believe that there exist certain fundamental rules for the formation of these helices, and with the identification of these rules and their order of precedence, we hope that we can achieve high accuracy in the prediction of RNA secondary structures from primary sequence data. Our application, RNA HEAT (RNA Helix Elimination and Acquisition Tool), can be used to graphically and interactively investigate the formation of possible RNA secondary structures in accordance to some of our predicted rules and compare them to the actual secondary structures of these RNA.

2. Introduction

RNA HEAT allows the user to apply various "filters", which add or eliminate helices based upon the rules implemented in these filters. For example, the current application offers filters based on helix length, simple distance, complex distance, nearby structures, and thermodynamic stability. RNA HEAT is a useful tool for visualizing the effect of applying these filters on an RNA sequence. Furthermore, it helps the user determine the similarity between the result of applying a sequence of filters to the RNA and the actual secondary structure of the RNA. Ultimately, the biological goal is to find rules, that when applied in the proper order, will result in a structure similar to how an RNA molecule folds in actuality.

A Java GUI (graphical user interface) was developed for the RNA HEAT application, where the user could select the filters to be applied, the parameters for each filter, and could choose the order in which to apply them. Our goals for the GUI included being able to view the status of each filter application, the result of the filter applications in a matrix form, the history of filters applied, and results on the screen that could be saved in various formats.

We strove for proper object-oriented design techniques in the development of this application, so that the program could be extensible, and others may add more filters in the future with relative ease.

3. Results

The RNA HEAT application that resulted from our efforts is a combination of its many implemented classes. For example, the class that is the necessary for all the filters is the RNA class. It contains the underlying data structures for representing the RNA sequences and helices. Each filter is a class that is implemented to make changes to an RNA object created by the RNA class. The GUI is another set of classes that communicates with the RNA and filters.

Our current application has a relatively stable implementation of the RNA class and its supportive data structures. The application has a reader that can read and parse BPSEQ files. We have implemented a basic filter that matches any possible combination of base pairings within an RNA molecule, and determines all the possible anti-parallel helices from those base pairings. From there, we have implemented the ability to filter helices that have length between two integers, that have E-loop structures at the end, that have AA or AG structures at either end, that have free energy between two values, and that have base-paired nucleotides that are positioned within a particular distance. The extensibility of the application has been achieved by making each filter apply its changes to the RNA in the same fashion.

As for the GUI, we have been able to implement the most of the features that had been proposed. The user is able to open sequence files for filtering. The GUI is able to render a 2-D matrix view of the RNA helices. The view can be zoomed in or out. Helix information can be seen by clicking on any helix in the 2-D view. Each filter application's parameters can be viewed with the filter history. The user can undo filters and apply them in any order, and is able to save the 2-D view either as a JPEG or PNG graphic file. A help section has also been implemented as a user reference.

For more detail on the features were proposed and those that have been implemented in RNA HEAT, see *Appendix A: Tasks Listings*. For more information on the GUI, as well as on how the filters function, see the help section user documentation in *Appendix B: User Documentation*, or access user documentation online at: <http://www.rna.icmb.utexas.edu/CLASS/2003-BIO384K/projects/MatriX/documentation.html>.

4. Conclusion

Within one semester, we have been able to create the current version of RNA HEAT, a professionally presented, basic functional application. We have been able to deliver a working application that reduces the number of RNA helices out of all possible helices by the application of rules as filters. While implementing the various filters, we have seen that RNA HEAT has the potential to visualize evidence of whether or not predicted rules are really likely patterns in RNA folding. Possible improvements of the application include improvements in speed and memory usage, as well as the implementation of additional filters.

5. References

- P. Chan. *The Java(TM) Developers Almanac 1.4, Volume 1* (2002) Addison Wesley Professional.
- J. Connone, S Subramanian, et. al. *The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intro and other RNAs.* (2002) BMC Bioinformatics.
- Robert Eckstein, Marc Loy and Dave Wood. *Java Swing.* O'Reilly and Associates. 1998.
- T. Elgavish, et. al. *AA.AG@Helix.Ends: A:A and A:G Base-pairs at the Ends of 16 S and 23 S rRNA Helices.* (2001) Journal of Molecular Biology.
- D.S. Fields and R.R. Gutell. *An analysis of large rRNA sequences folded by a thermodynamic method.* 1996. Folding and Design 1:419-430.
- R.R. Gutell *Chapter 6 - Comparative Sequence Analysis and the Structure of 16S and 23S rRNA.* 1996. From Book *Ribosomal RNA: structure, evolution, processing and function in protein biosynthesis.* Telford Press.
- R.R. Gutell, et. al. *A Story: Unpaired Adenosine Bases in Ribosomal RNAs.* (2000) Journal of Molecular Biology.
- L.H. Rodrigues *Building Imaging Applications with Java Technology* (2001) Addison Wesley.
- Discussions with Dr. Robin Gutell.
- Internal Gutell lab notes regarding RNA folding.

Appendix A: Tasks Listing

<u>Tasks/Objectives</u>	<u>Finished</u>	<u>In Progress</u>	<u>Planned</u>
Implementation of underlying data structure for representing RNA and helices.	95%	5%	0%
Implement reader class to read and parse BPSEQ file and generate RNA objects.	100%	0%	0%
Agreed on the interface for all filters	95%	5%	0%
Agree on a generic mechanism to pass arguments to the filters	80%	20%	0%
Generic Base-pair filter	100%	0%	0%
Ability to filter helices of size greater than N	100%	0%	0%
Ability to filter helices of size less than N	100%	0%	0%
Ability to filter helices of size between N and M	100%	0%	0%
Ability to filter E-Loop helices	100%	0%	0%
Ability to filter helices whose base-paired nucleotides are positioned within a particular distance (range)	90%	0%	10%
Ability to filter AA&AG helices	100%	0%	0%
Ability to filter helices according to Energy levels	100%	0%	0%
Ability to filter helices according to frequency of certain base-pairs	0%	0%	100%
Support for consecutive application of helix-filters	90% (working)	0%	10% (optimization)
Implementation of Scores for Filters	10%	0%	80%
Add the ability to undo filters	100%	0%	0%
Implement the filter history list	100%	0%	0%
Implemented GUI frontend for corresponding functionalities of the backend.	100%	0%	0%
Add printing capabilities	0%	0%	100%
Add support to save images in jpeg, gif, or ps format.	80%	0%	20%
Ability to render the 2-dimensional view of the RNA helices.	100%	10%	0%

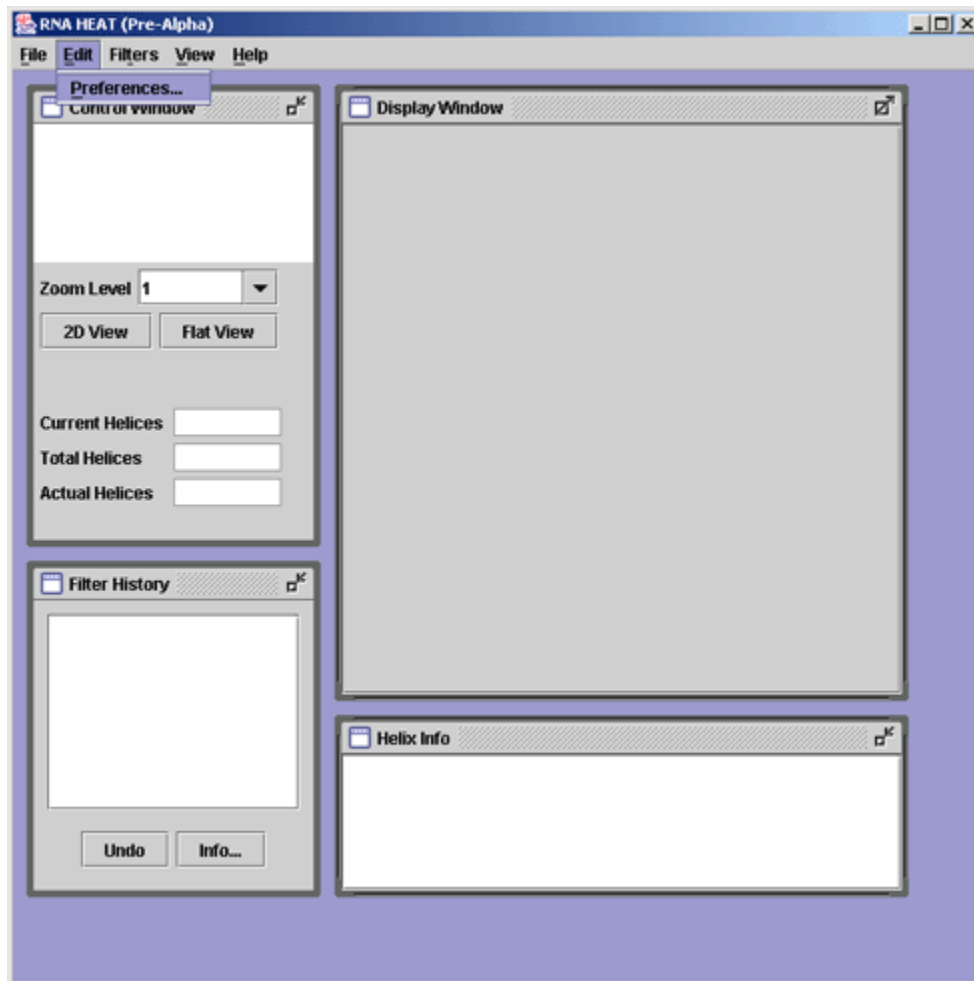
Ability to render the flat (histogram) view of the RNA helices.	10%	80%	10%
Resolve OutofMemoryError issue when displaying 23S structure.	10%	90%	0%
Add interactive capability with the display window.	100%	0%	0%
Add dialog boxes for each filter to accept user arguments.	90%	0%	10%
Implement of the Help -> About box	100%	0%	0%
Implement the help viewer system	100%	0%	0%
Write the help/user documentation	98%	2%	0%
Generate JavaDocs	30%	30%	40%
Generate PostScript Outputs	0%	0%	100%

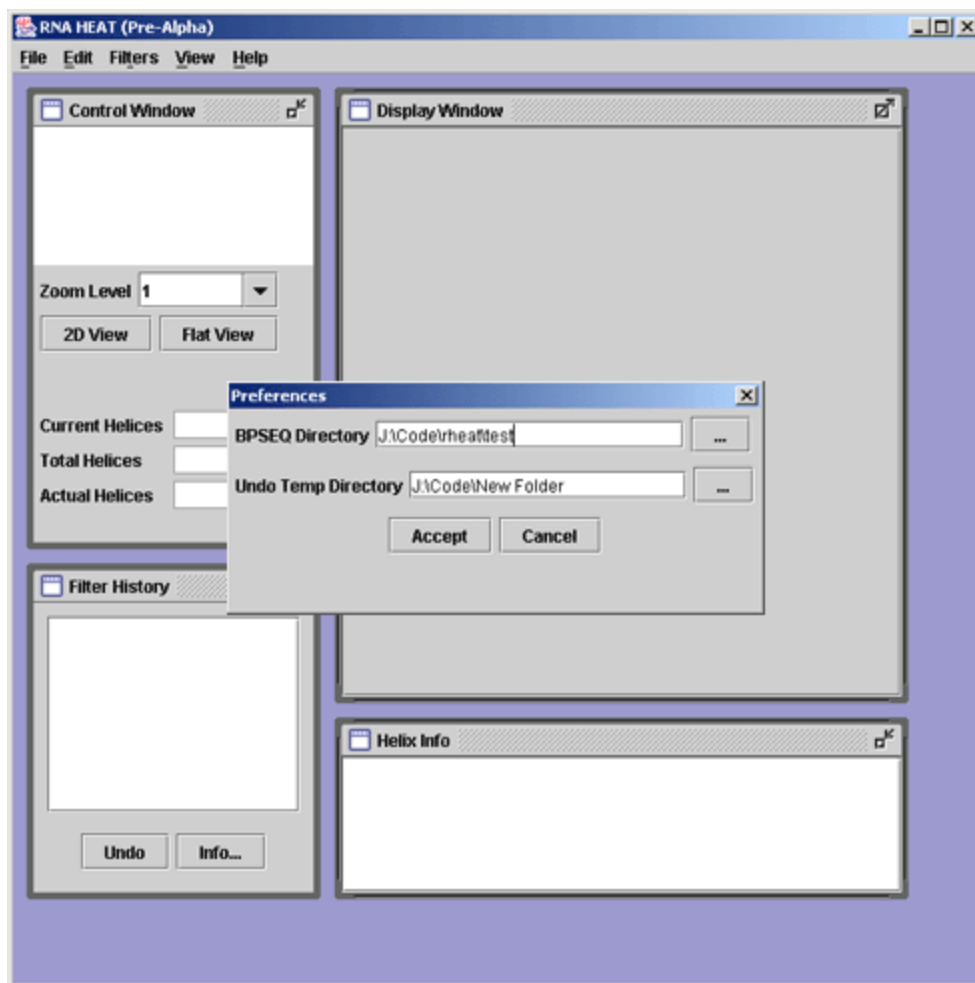
Appendix B: User Documentation

RNA HEAT: Getting Started

Setting Folder Preferences:

Folder preferences should be set before opening sequence files. To set these preferences, click **Edit | Preferences**. For **BPSEQ Directory**, specify the location of a folder where RNA sequence files are to be found. For **Undo Temp Directory**, specify the location of an undo folder. Once these folders are selected, opening a sequence file will automatically bring up the BPSEQ Directory, and undo data will be stored in the Undo Temp Directory. These preferences can be changed for different folders.



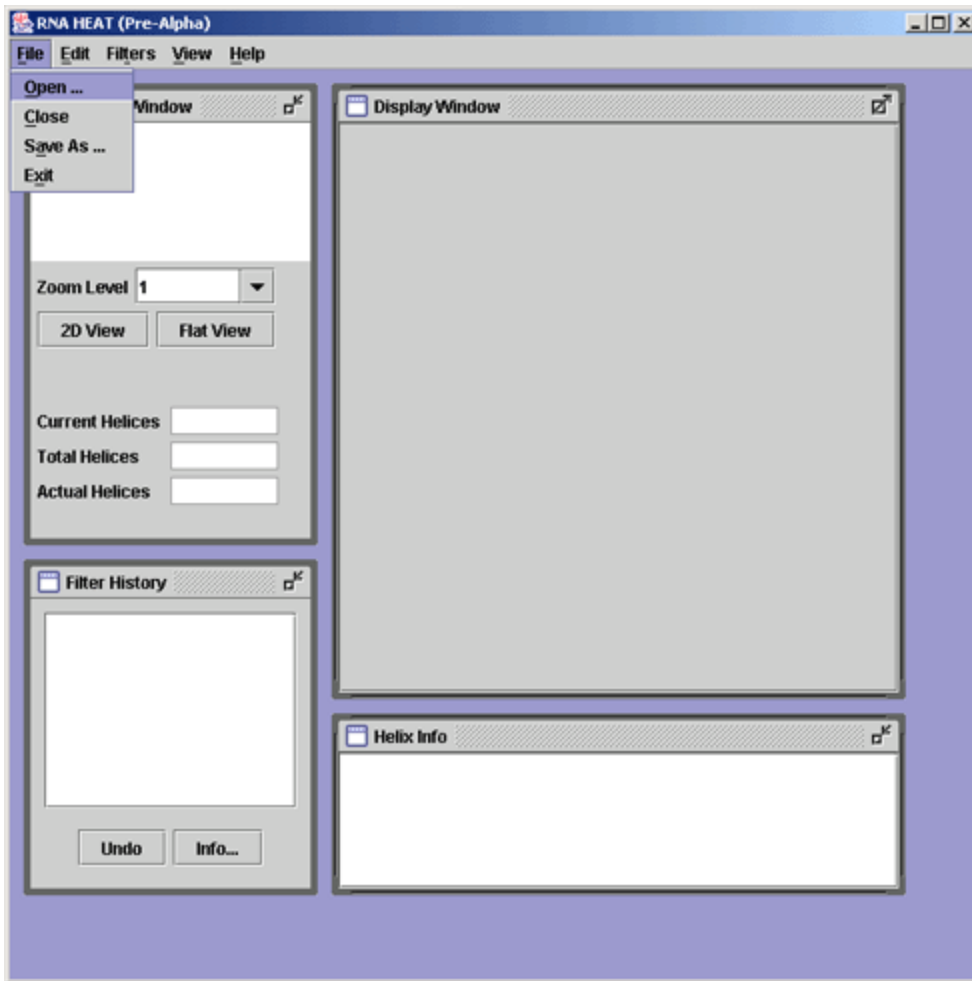


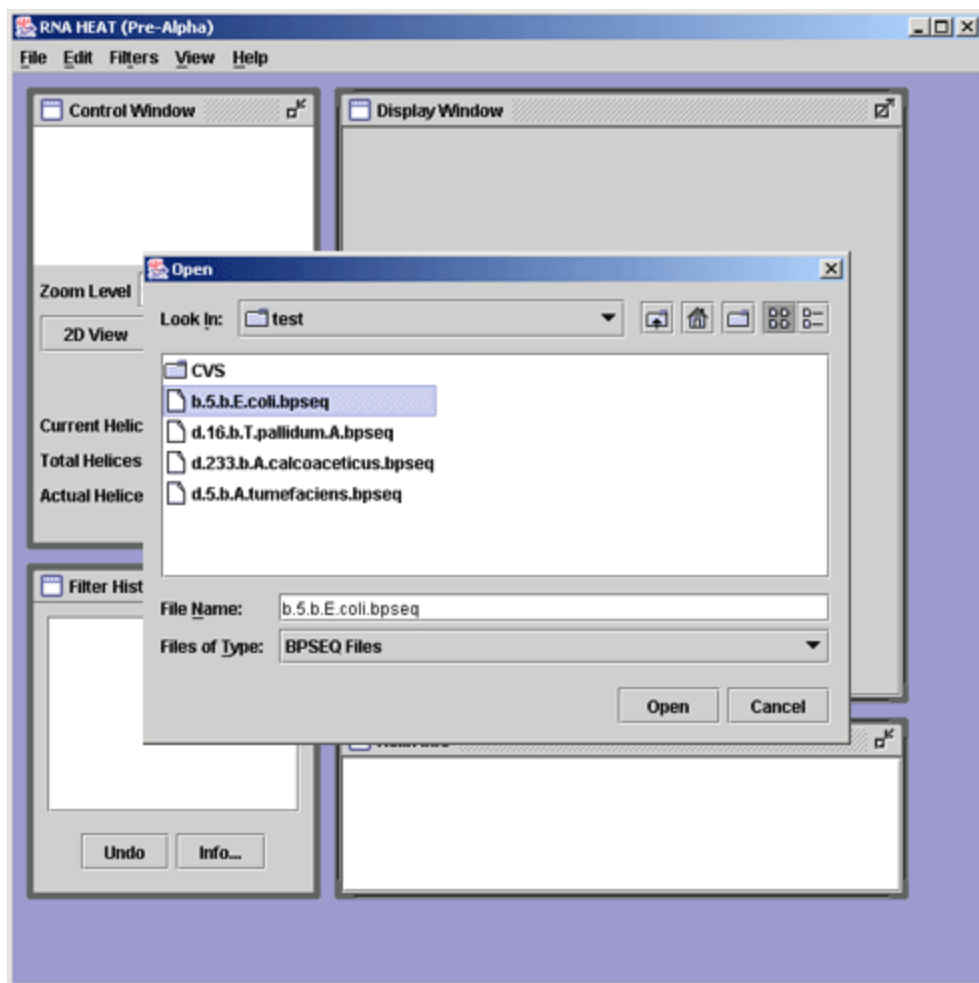
If you have not yet set any preferences, or if your preferences folder no longer exists at the location specified, you may encounter the following message. RNA HEAT will then choose a folder for you, which you can change with the steps given above.



Opening a Sequence File:

To open a sequence, select **File | Open...** from the menu bar. Open a file of type ".bpseq". Information regarding the the sequence is displayed in the Control and Display windows.

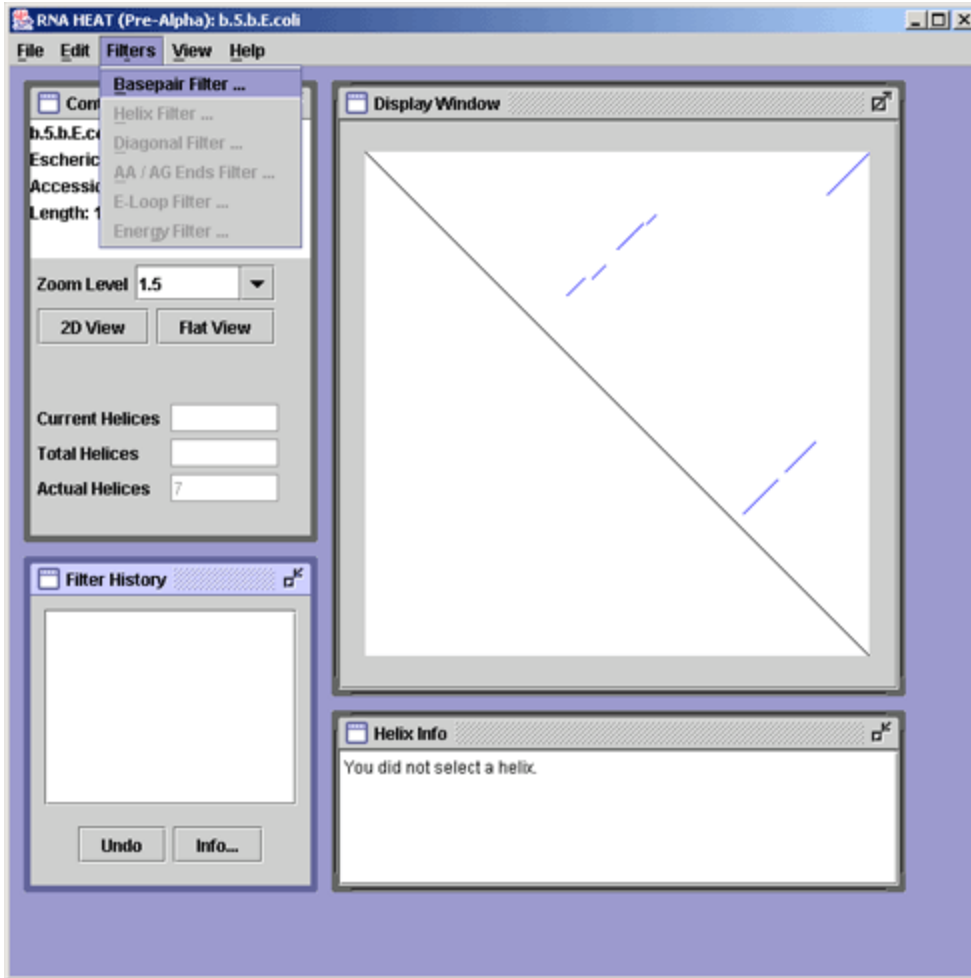


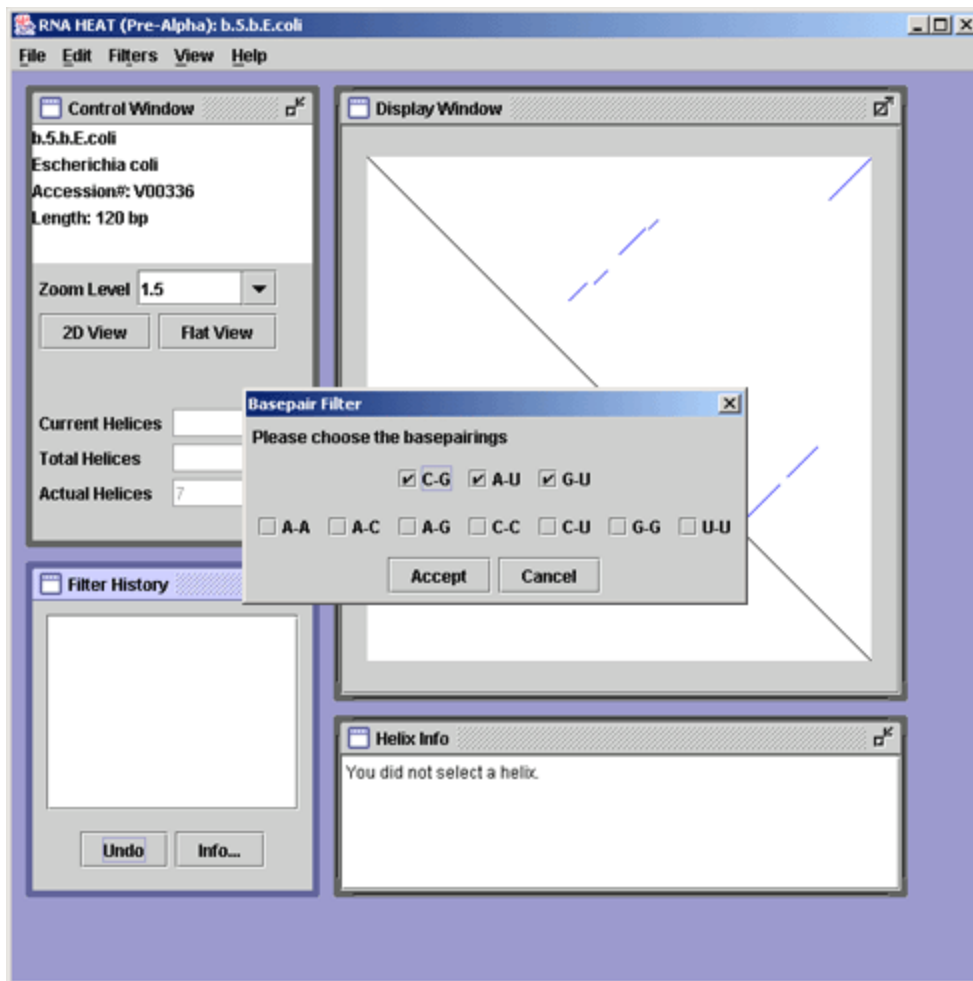


Applying Filters:

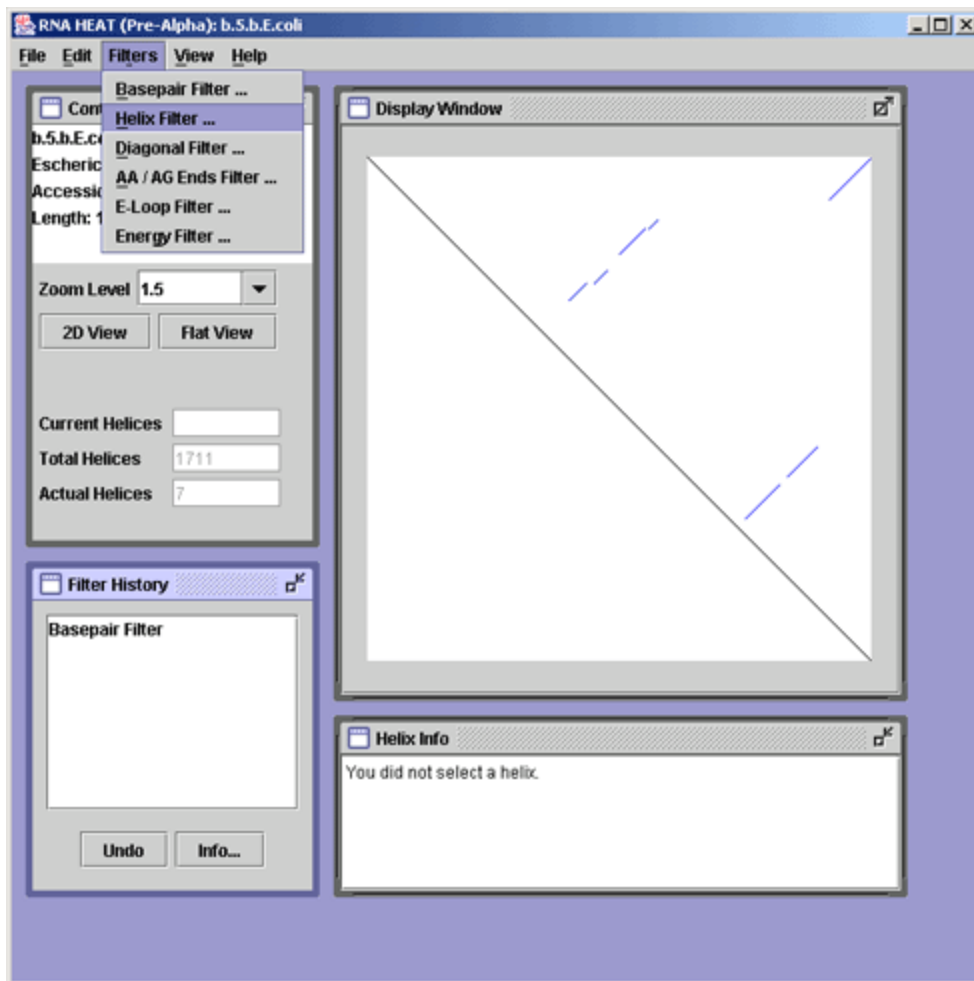
To apply a filter, select **Filters** from the menu bar, and select an enabled filter. Specify any parameters required by the filter. Results of the filter application should be displayed in the Display Window. The **Basepair Filter** must be applied first to a sequence that has been opened before any other filters.

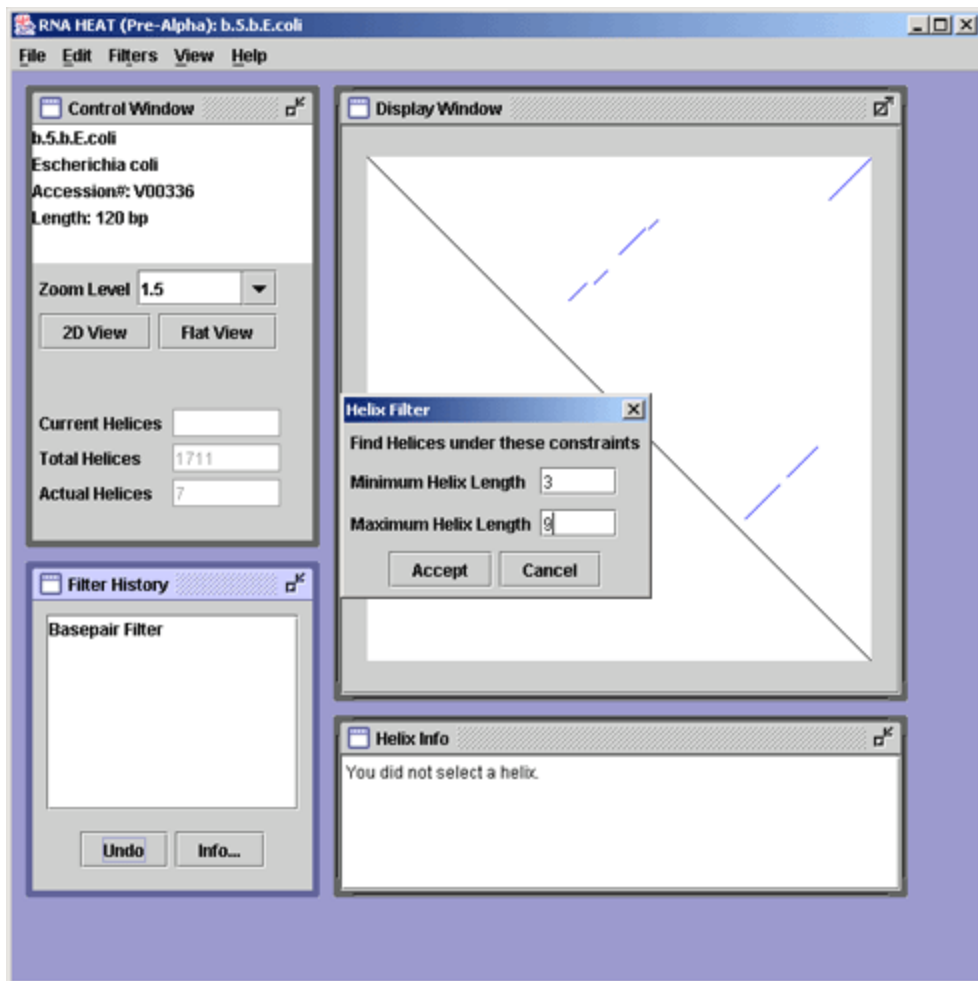
Applying the **Basepair Filter**:

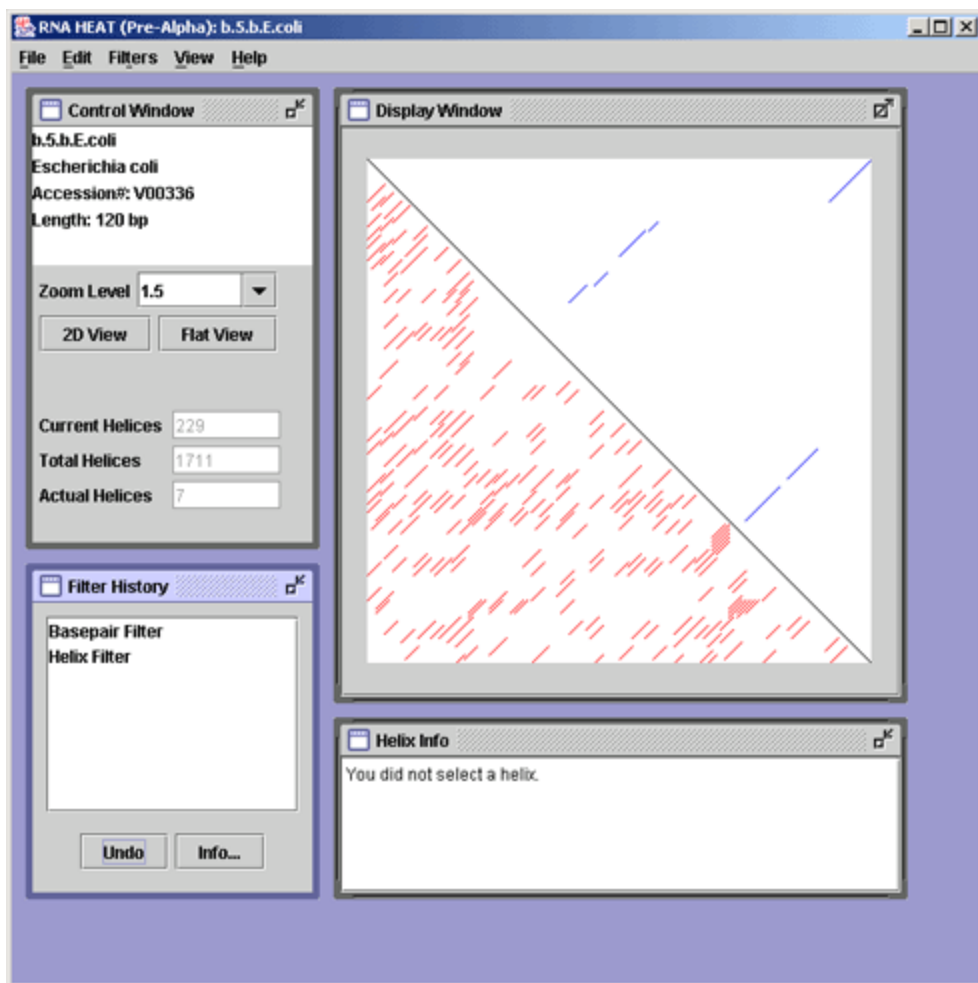




Applying the **Helix Filter** after the **Basepair Filter**:







To view the results in more/less detail, adjust the **Zoom Level**.

For more information on how to use the RNA HEAT desktop and menu bar, see the appropriate section of the Help Contents.

RNA HEAT: The Control Window



The Control Window is Minimizable.

Sequence Information Display:

The Control Window's Sequence Information Display shows information about the sequence that are given in its .bpseq file, such as the file name, organism name, accession number, and length.

Zoom Level :

Changing the zoom level will cause the current figure in the **Display Window** to be reduced or enlarged. The zoom can be adjusted from ten times smaller (zoom level at 0.1) to ten times larger (zoom level at 100).

2D View / Flat View:

2D View gives a two dimensional view of helices in the **Display Window**. This implies that the sequence runs down both the x and y directions, and each line in the plane represents a helix. The Flat View is a one-dimensional representation, where the sequence runs down a single axis. This view has not yet been implemented in the current version of RNA HEAT.

Current Helices:

This window displays the current number of helices present in the lower portion of the 2D View in the Display Window, as a result of applying any filter(s).

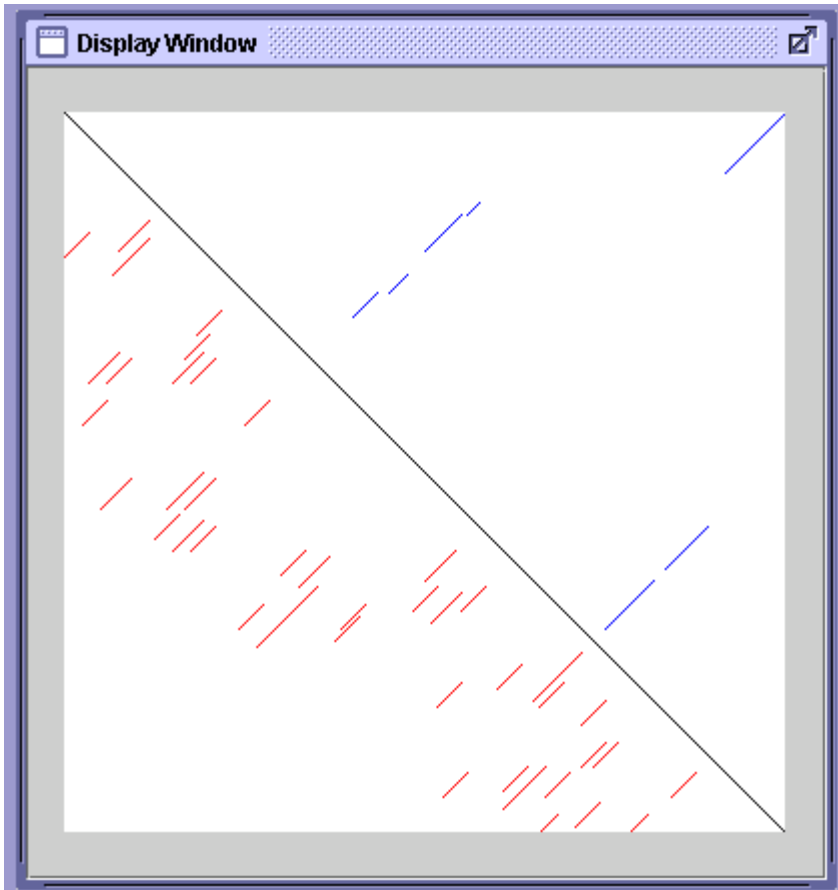
Total Helices :

This window displays the total number of helices possible: the same number of helices that result after initially applying the **Basepair Filter**.

Actual Helices :

This window displays the actual number of helices found for a particular sequence, and is the number of helices present in the upper portion of the 2D View in the **Display Window**.

RNA HEAT: The Display Window

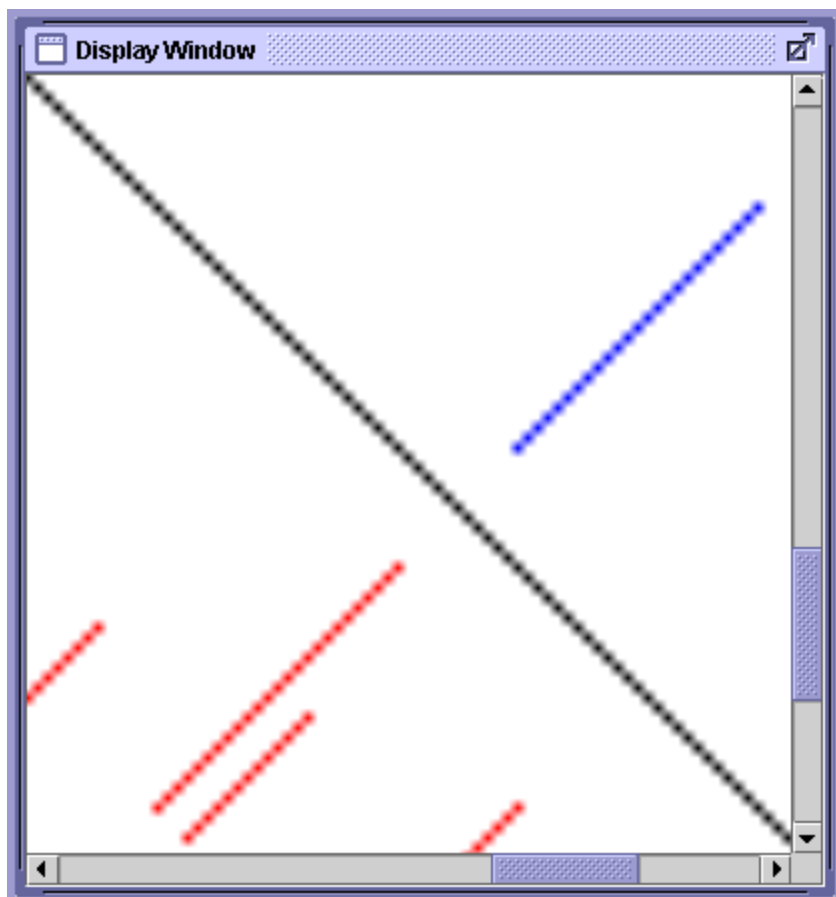


The Display Window is Maximizable.

2D View Display:

In the 2D View Display, the current sequence is assumed to run down both the x and y-axes 5' to 3'. The red lines represent the possible helices that the sequence could form after applying a filter. The blue lines represent the actual helices that have been found for the current sequence. Therefore, if a possible helix matches an actual helix, the red and blue lines will be reflections of each other across the black diagonal line.

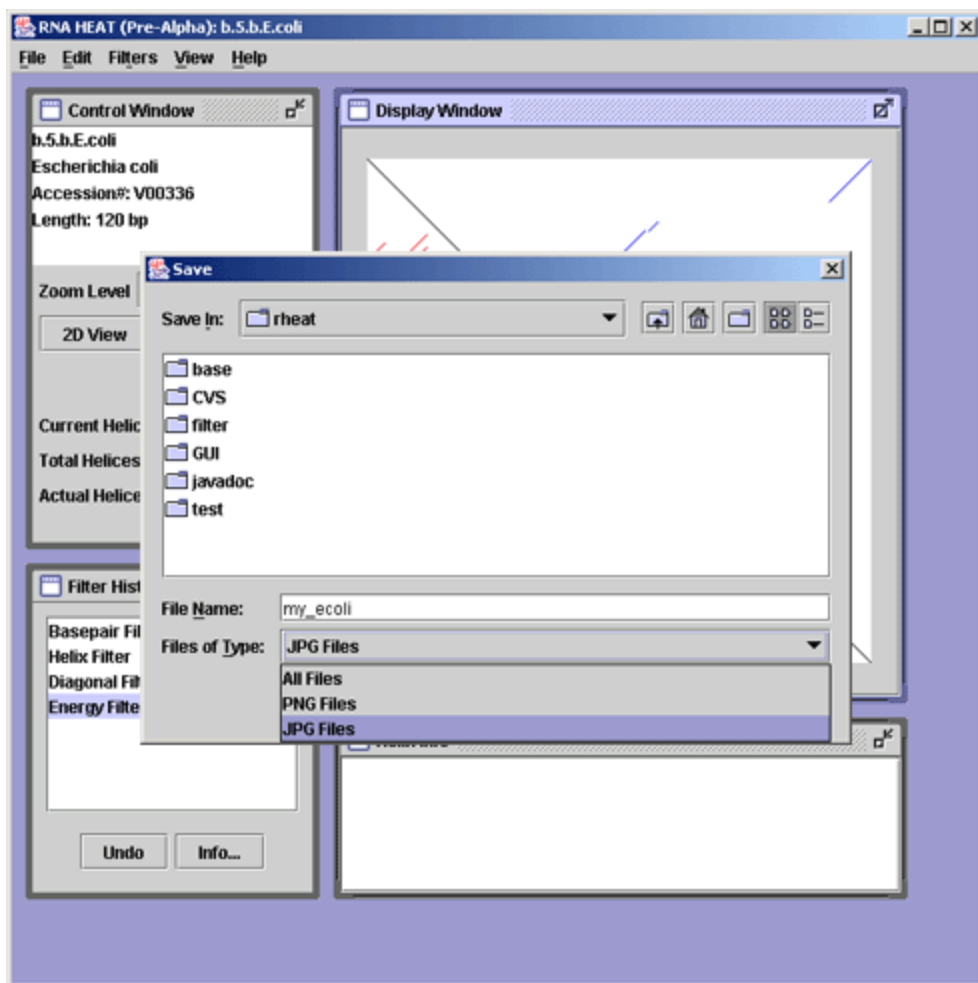
Zoomed-in View:



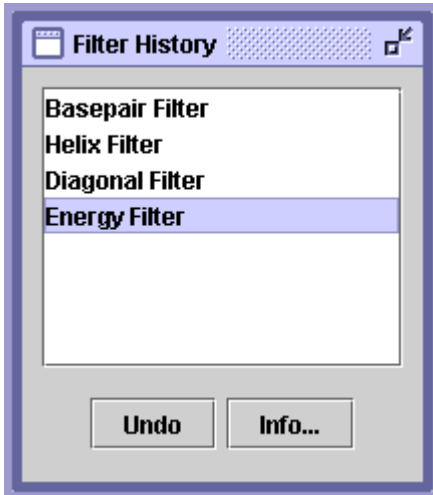
After changing the **Zoom Level** in the **Control Window** to enlarge a graph in the display window, the graph will resemble the above figure. Each "dot" does not necessarily represent one base pair. To see more information about a possible helix (i.e. a red helix), click on any part of the helix. Detailed information about this helix will appear in the **Helix Info Window**.

Saving the Display Window Graph:

The Display Window graph can be saved as either a .JPG or .PNG file using **File | Save As ...** in the menu bar:



RNA HEAT: The Filter History Window



Filter History Display:

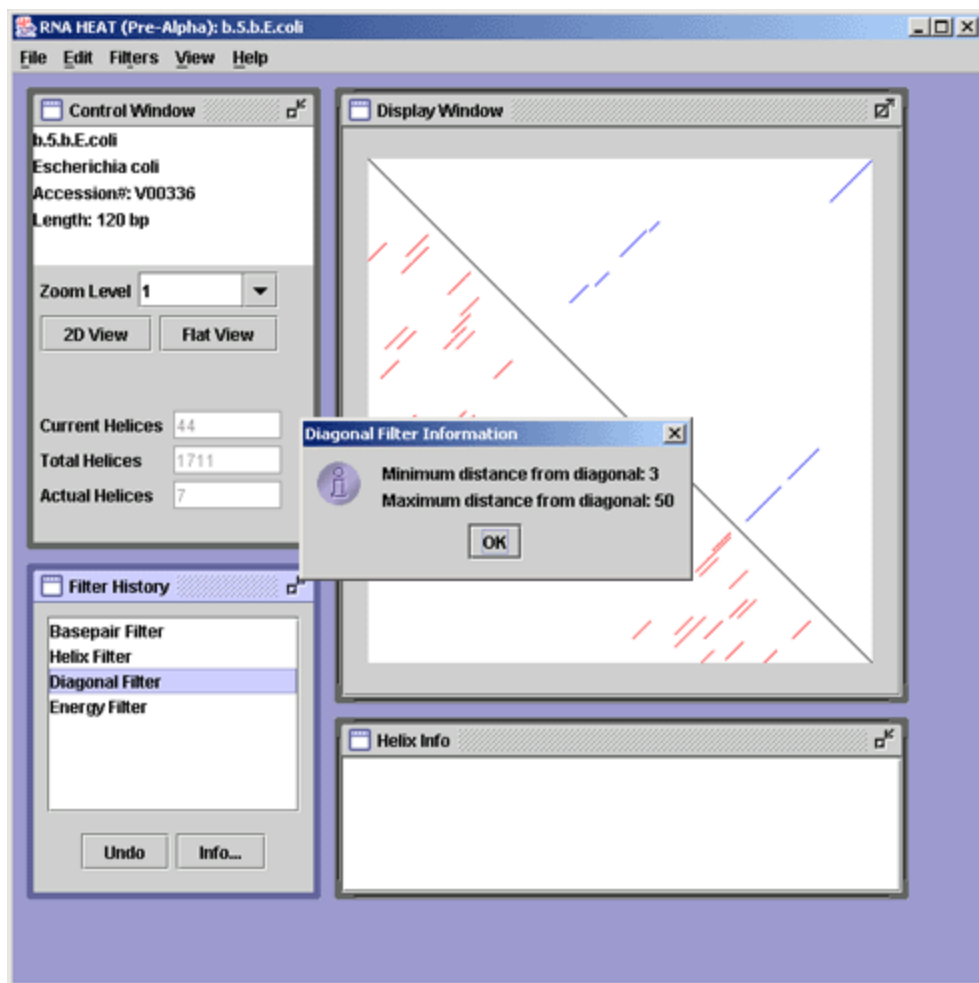
The Filter History Display keeps track of all the filters that have been applied since a sequence was opened.

Undo:

To undo any filter(s), click on the filter and then the **Undo** button. This will return the the helices to their state before this filter was applied. The changes will be reflected in the **Display Window**.

Info...:

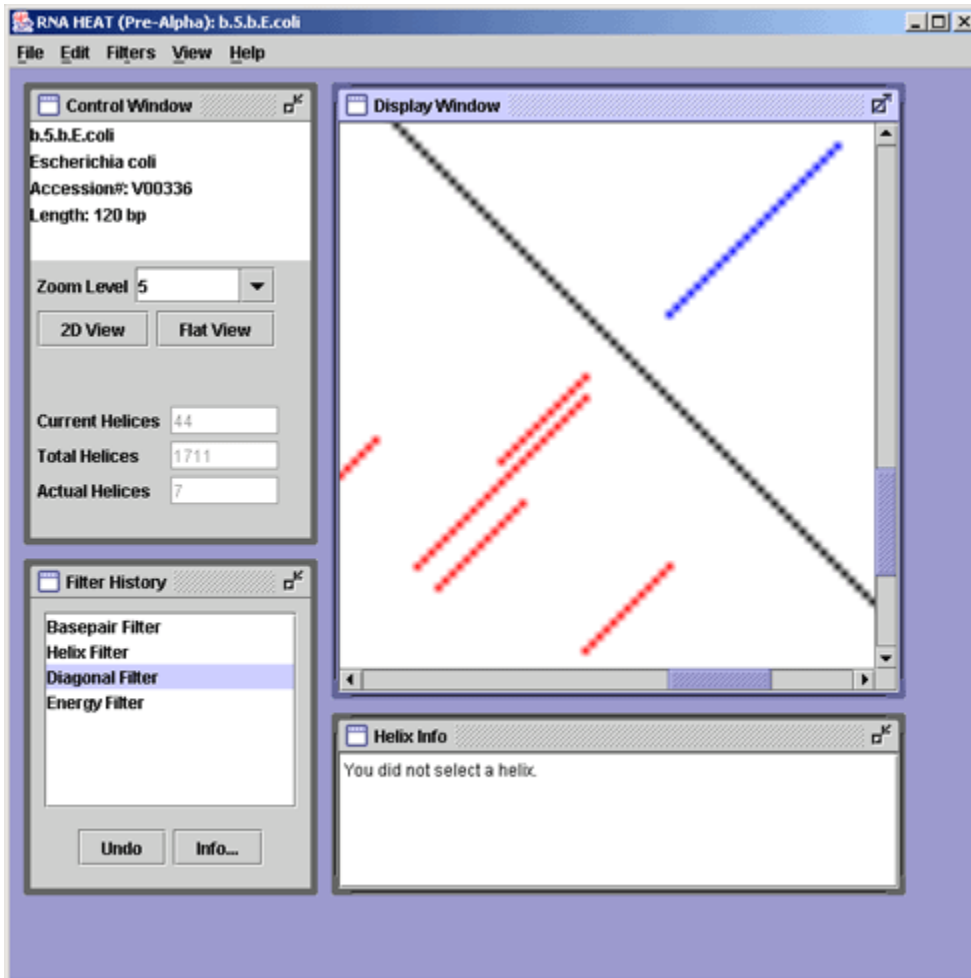
To see specific information about the parameters associated with a filter, click on the filter, and then the **Info...** button. A box will appear with parameters that were entered for the filter:

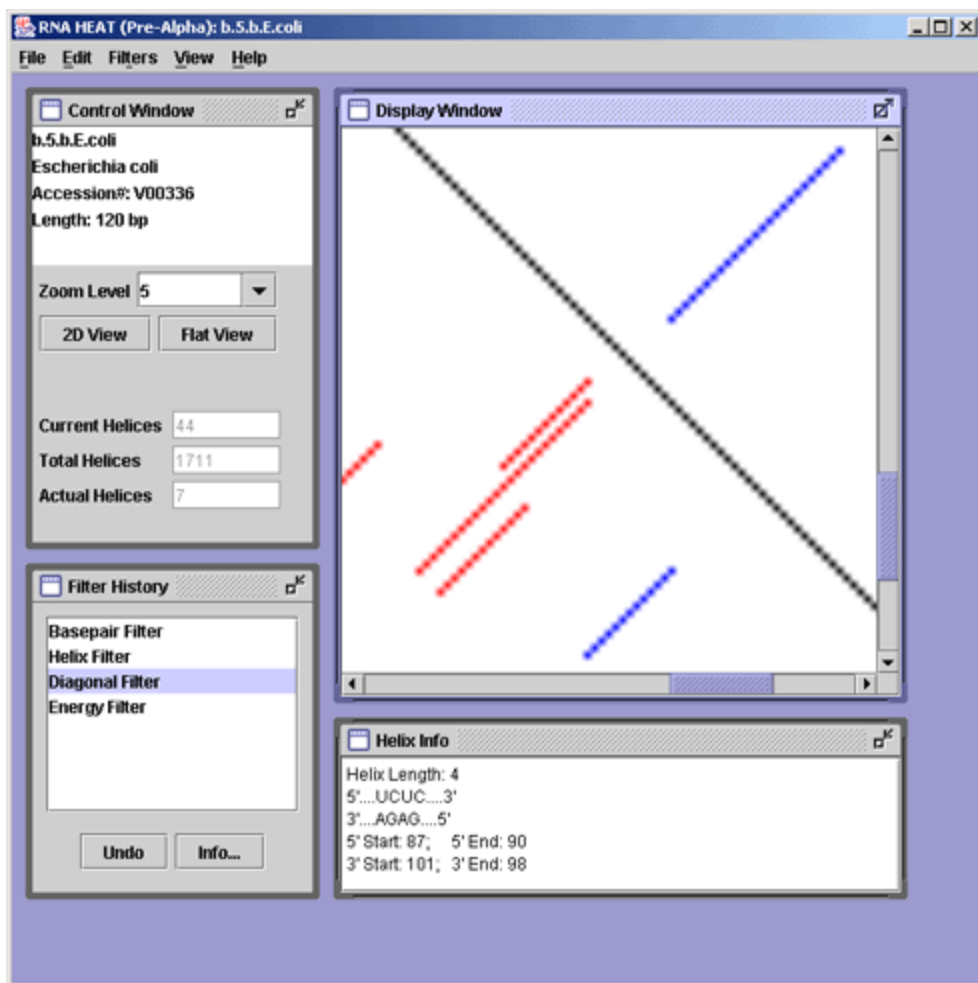


RNA HEAT: The Helix Info Window

Displaying Information of a Particular Helix:

Clicking on a red helix below the diagonal line in the **Display Window** will cause its helix information to be displayed in the Helix Info Window. The color of selected helix will become blue. Helices above the diagonal line that were already blue (the actual helices) do not display information. Helix information includes the helix length, the base-pairs of the helix, the 5' nucleotide number at the beginning of the helix, the 3' nucleotide number at the beginning of the helix, the 5' nucleotide number at the end of the helix, and the 3' nucleotide number at the end of the helix:

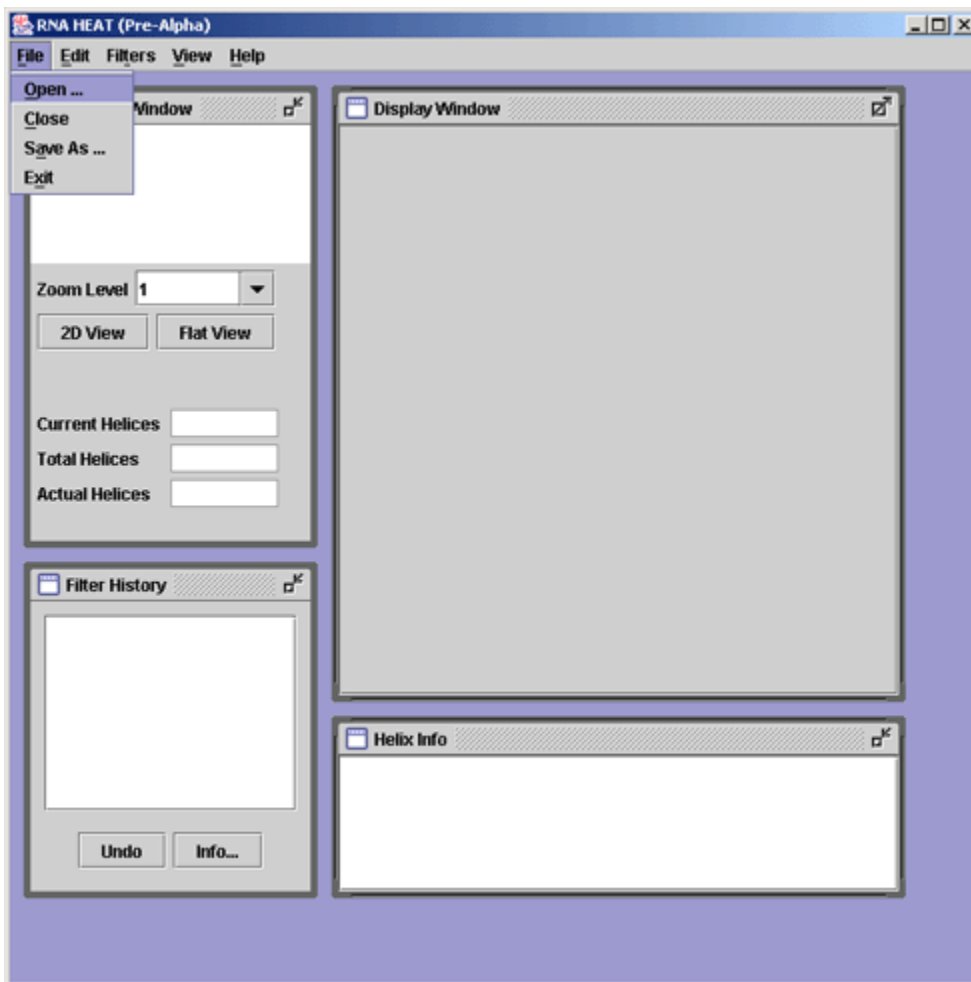


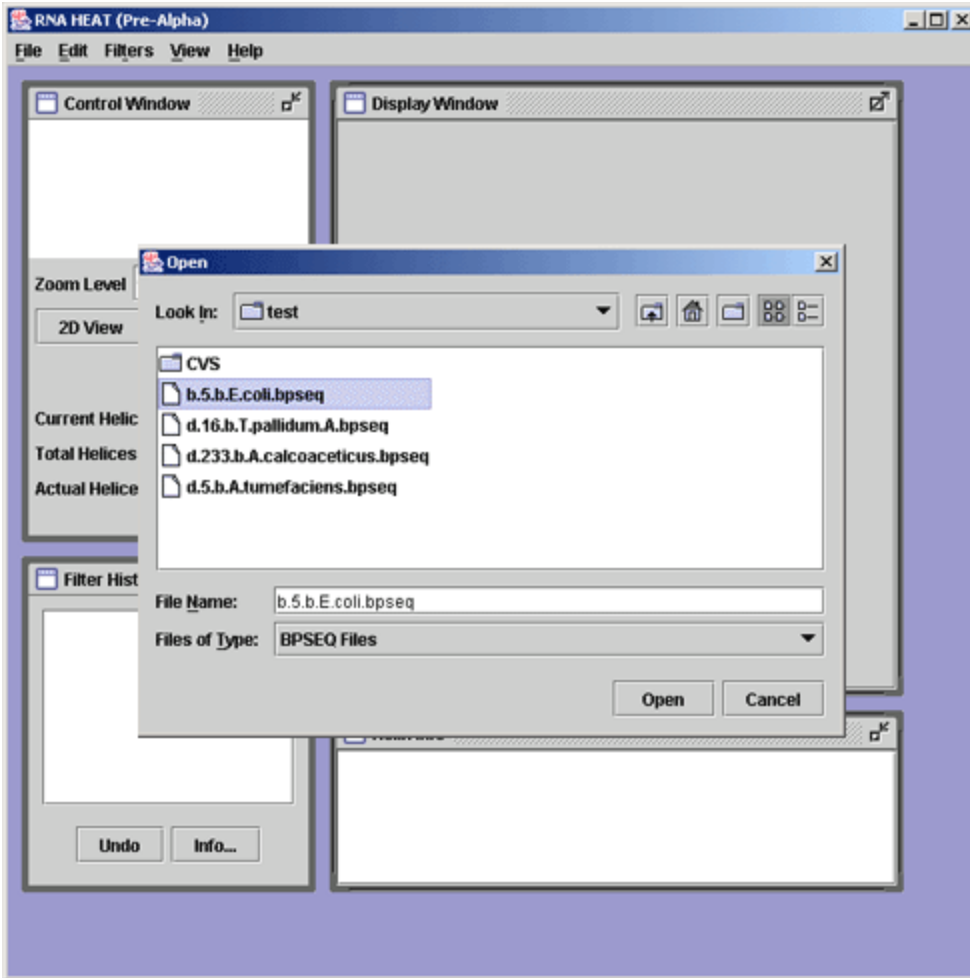


RNA HEAT: The Menu Bar: File

Open ...:

Opens a sequence file for helix acquisition/elimination. Open a file of type ".bpseq". Information regarding the sequence is displayed in the Control and Display windows. The Open dialog box will automatically look first for sequence files in the folder that was specified using **Edit | Preferences**.

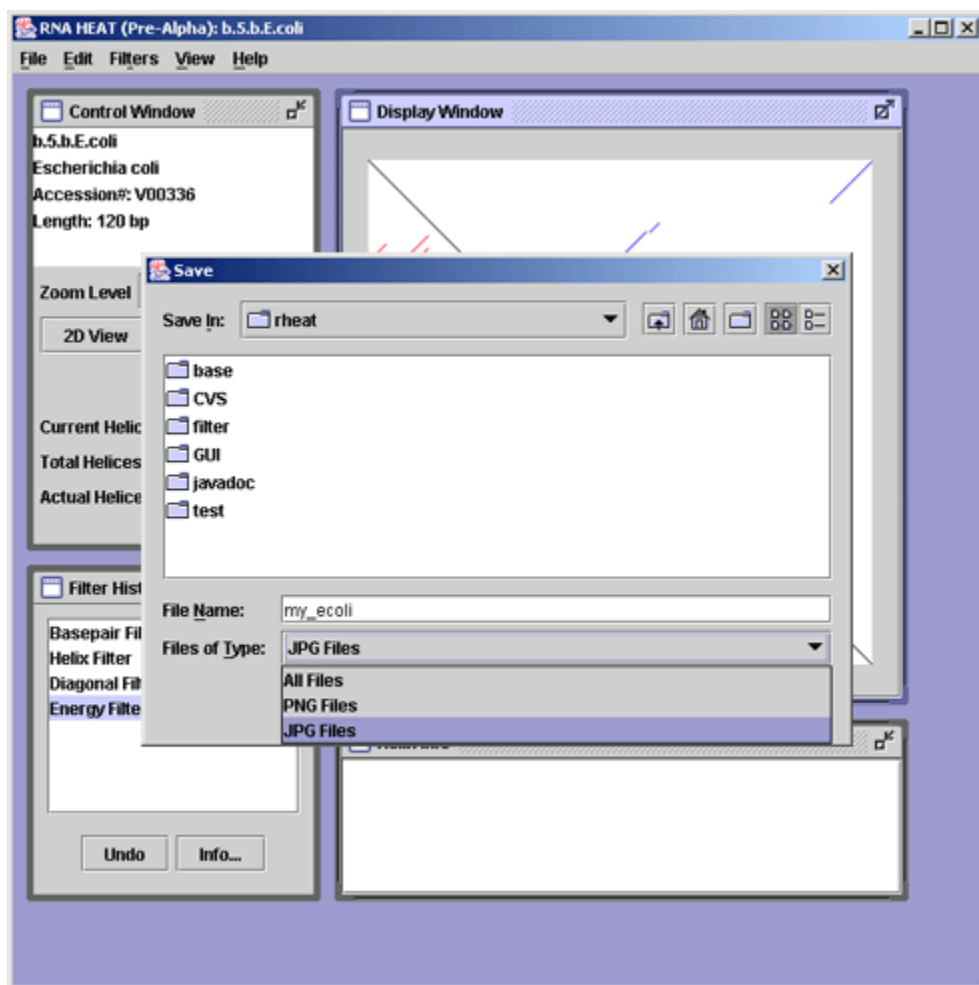


**Close:**

Closes the current sequence file. All windows are cleared.

Save As ...:

Saves the current graph in the Display Window as either a .JPG or .PNG file.

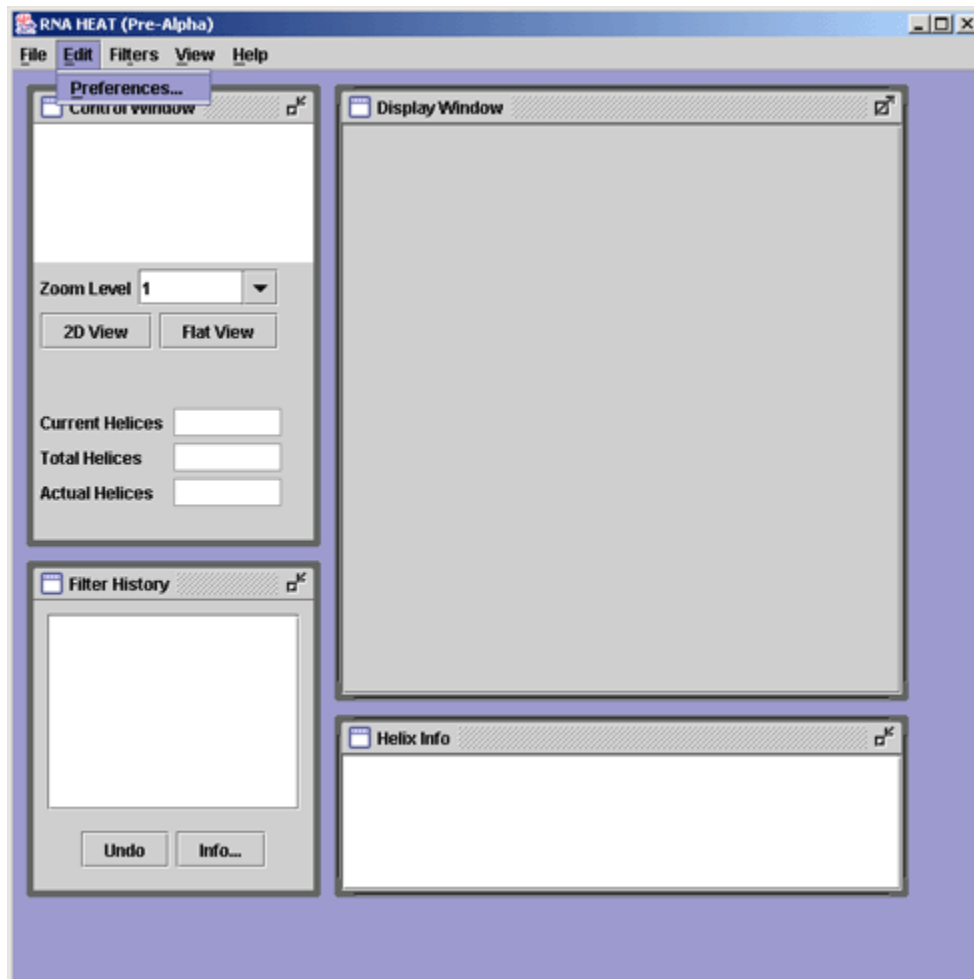
**Exit:**

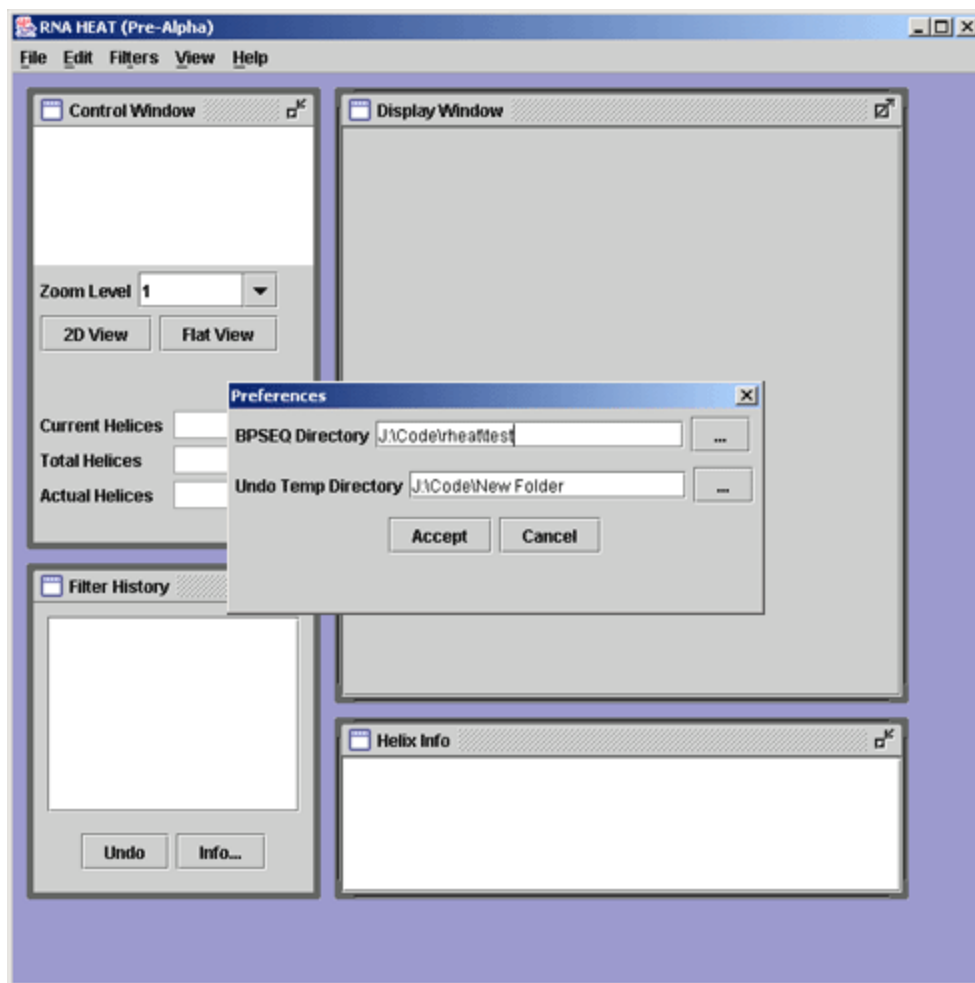
Exit RNA HEAT.

RNA HEAT: The Menu Bar: Edit

Preferences:

Sets folder preferences. For **BPSEQ Directory**, specify the location of a folder where RNA sequence files are to be found. For **Undo Temp Directory**, specify the location of an undo folder. Once these folders are selected, opening a sequence file will automatically bring up the BPSEQ Directory, and undo data will be stored in the Undo Temp Directory. These preferences can be changed for different folders.





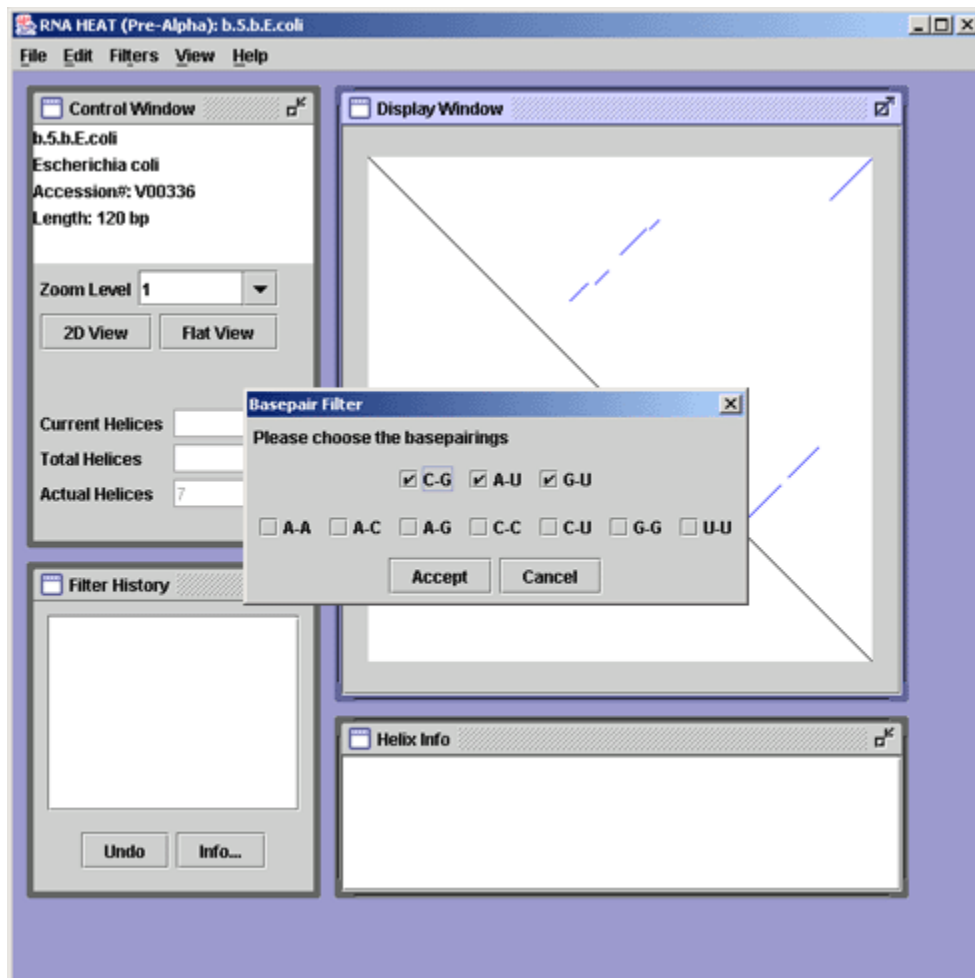
If you have not yet set any preferences, or if your preferences folder no longer exists at the location specified, you may encounter the following message. RNA HEAT will then choose a folder for you, which you can change with the steps given above.



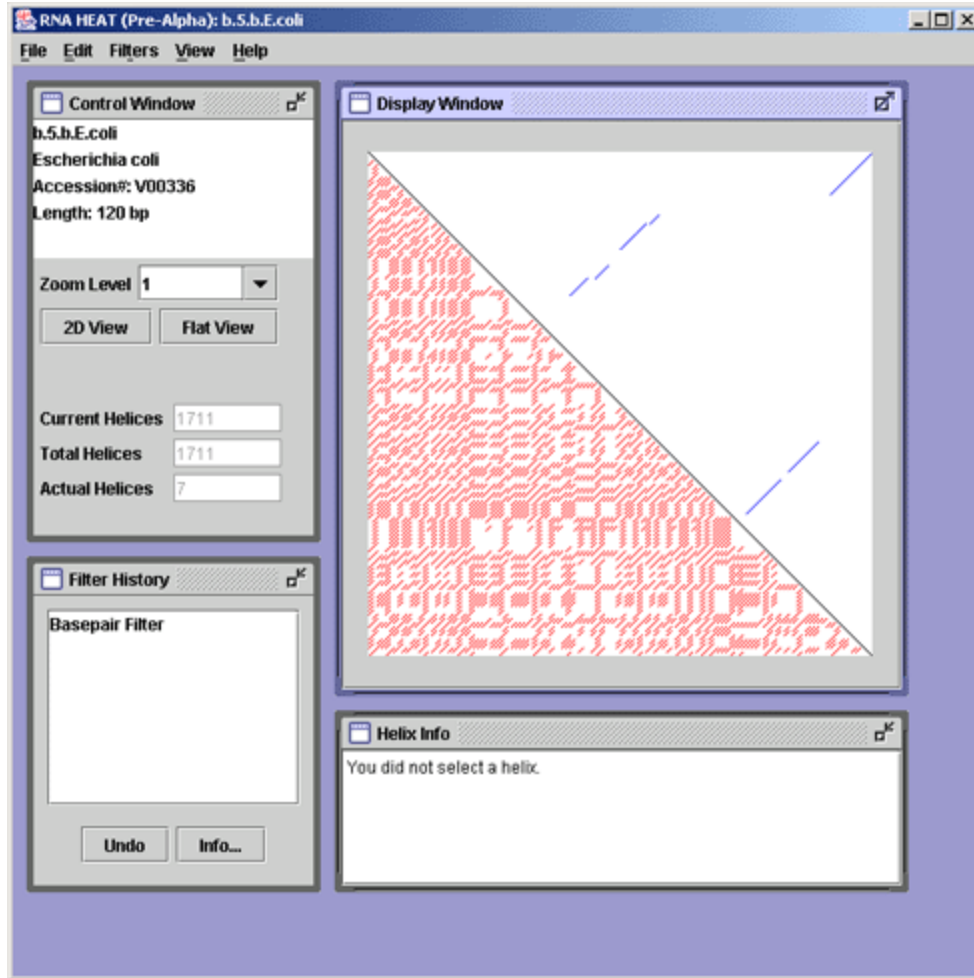
RNA HEAT: The Menu Bar: Filters

Basepair Filter ...:

This filter must be applied before any other filters. Applying the Basepair Filter acquires all the possible helices that the other filters can work with. The type of base-pairs that can be counted in a helix can be set as a parameter for the filter. The default option for the parameter are the most common pairings (C-G, A-U, G-U). Because this filter is the basic filter for all other filters, the helices that are acquired will not automatically be shown in the display window. If you would like to see the resulting helices after applying the basepair filter, double click on the display window.



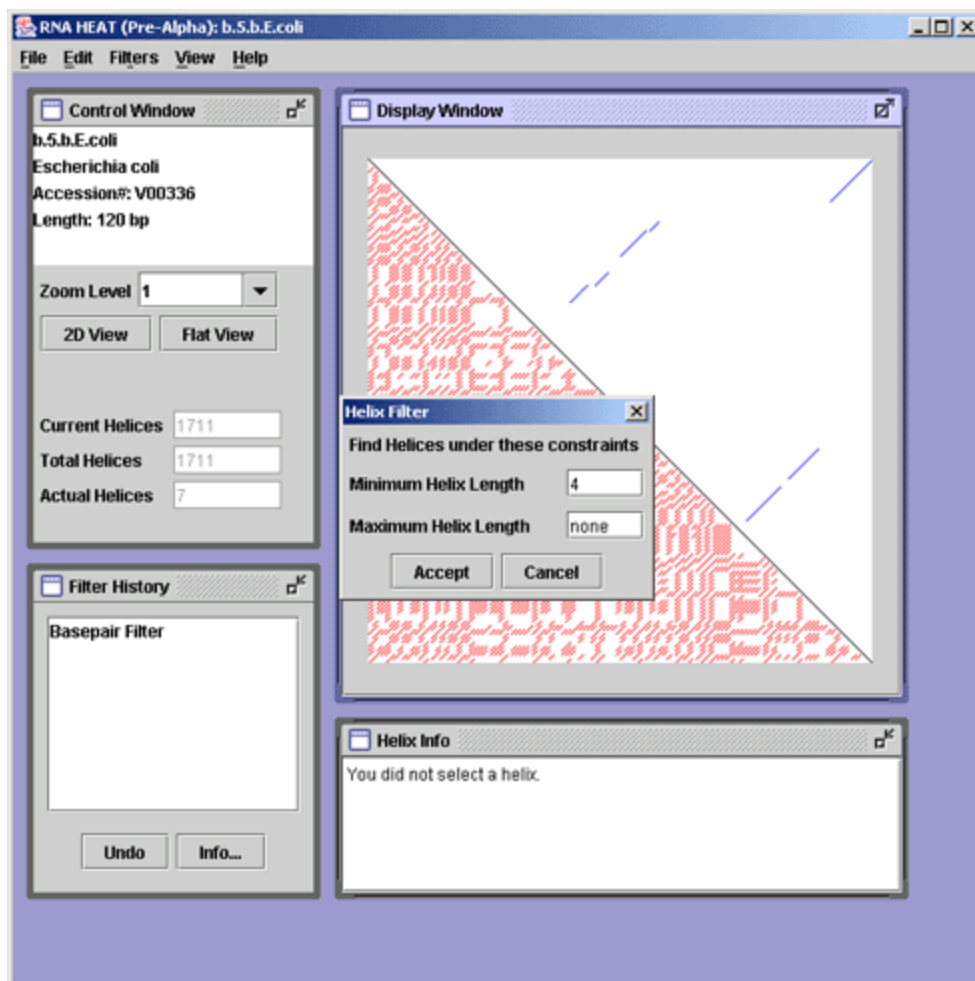
After double-clicking in display window:

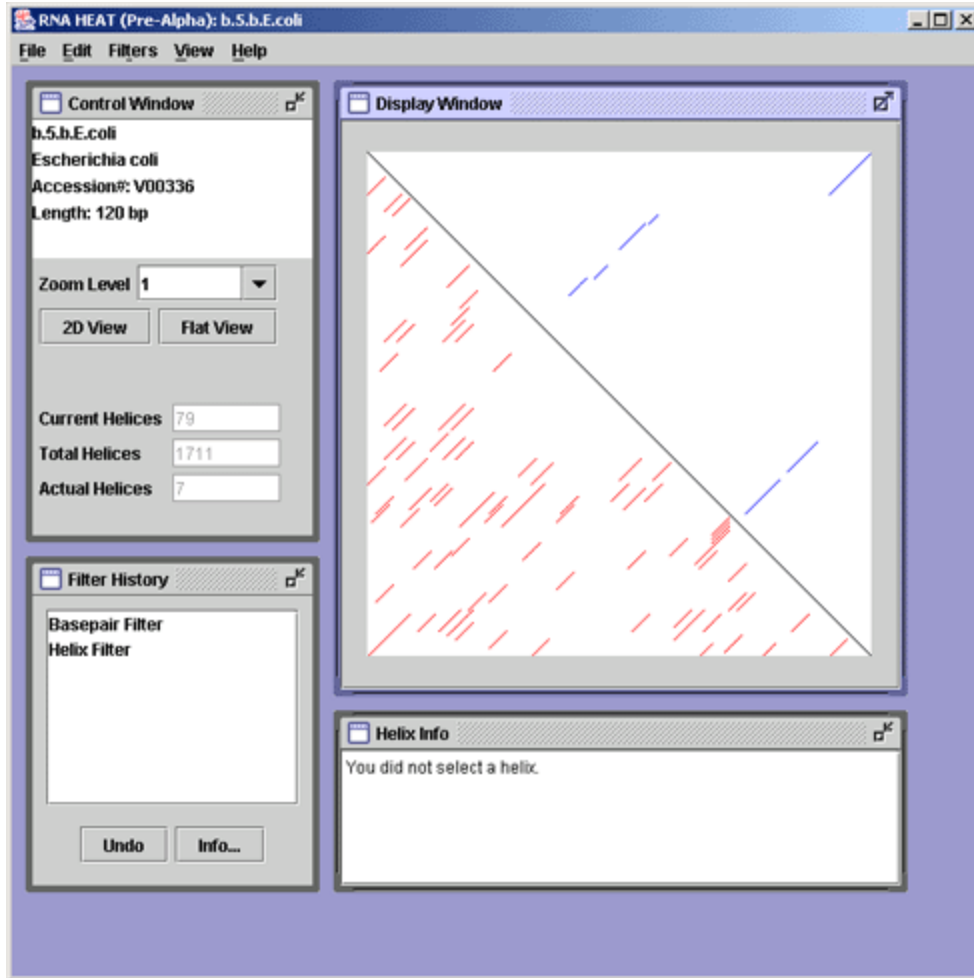


Helix Filter ...:

This helix keeps the helices that are in a specified length range, and eliminates those that are not within the range. The minimum and maximum length values can be set as parameters. If the maximum length can be arbitrarily long, leave the parameter as "none".

The following example keeps helices that are length 4 base-pairs and over, and eliminates those helices with length less than 4:

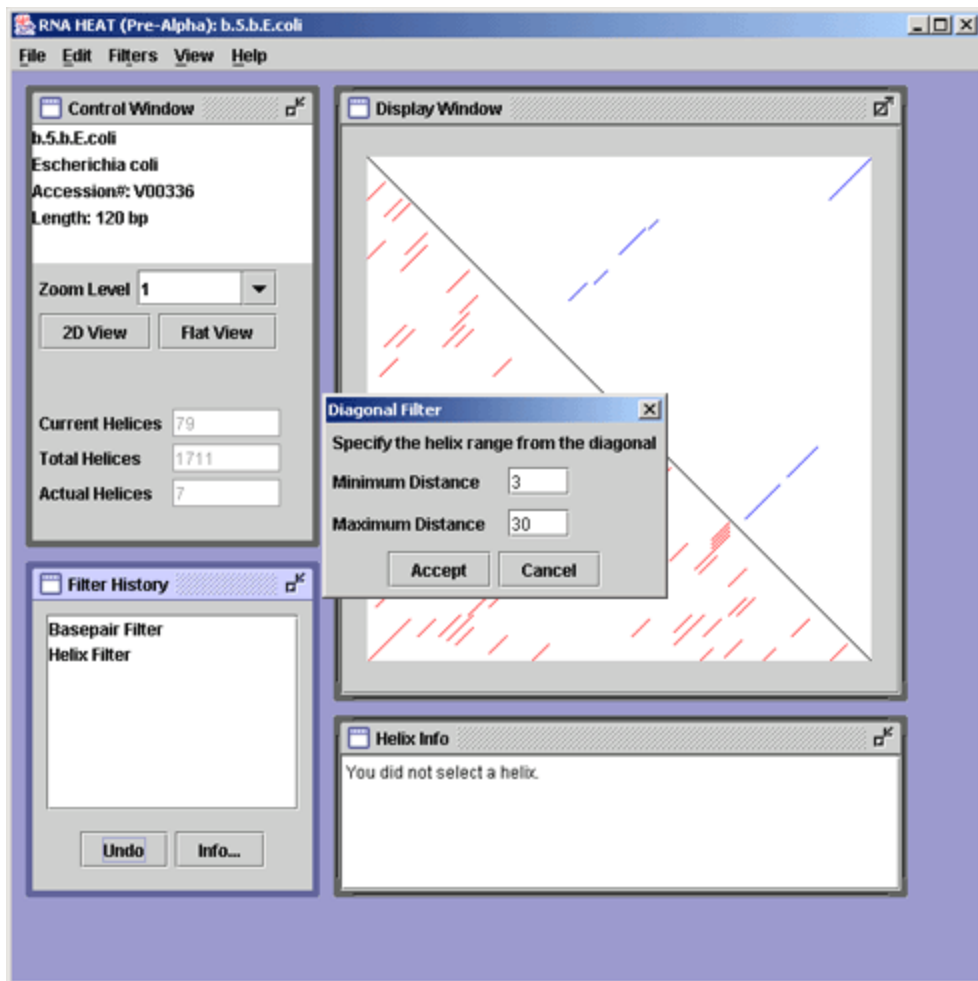


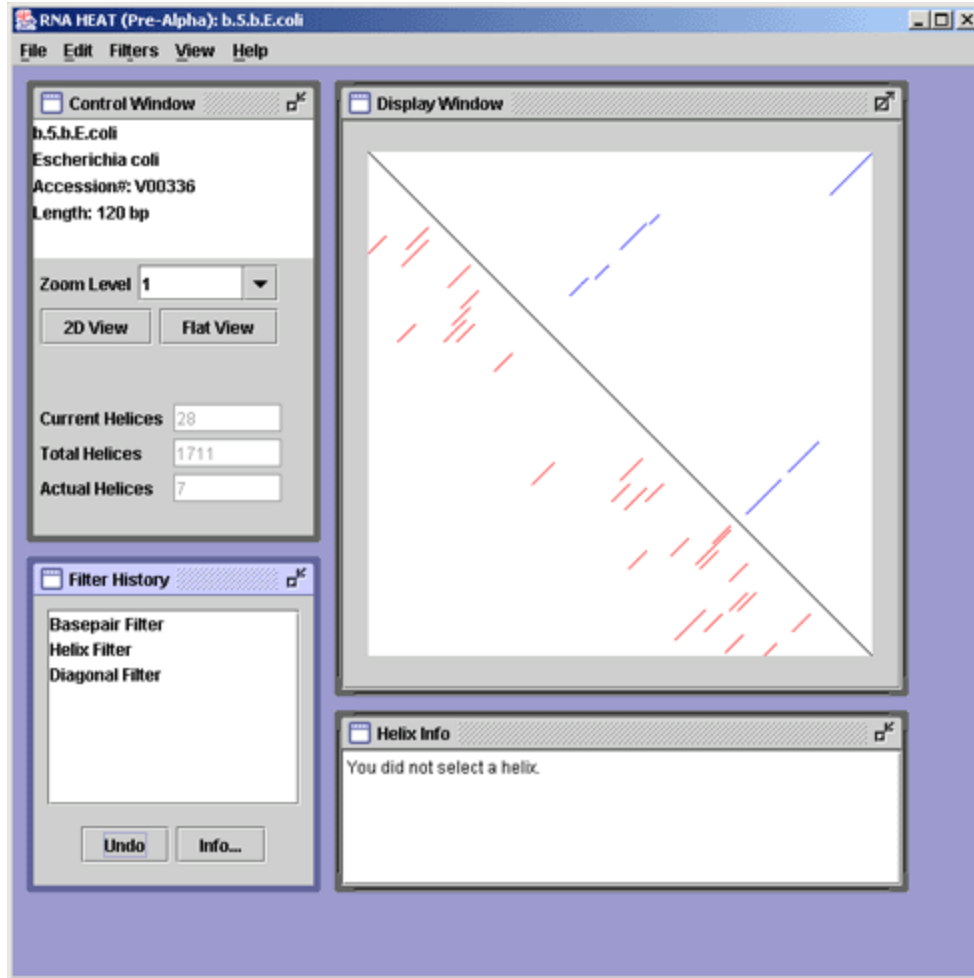


Diagonal Filter ...:

The diagonal filter keeps helices that are a certain distance from the diagonal line. Helices that are close to the diagonal line are short-range helices. Short-range helices are formed by base-pairing portions of the sequence that are close to each other. The minimum and maximum distance from the diagonal can be set as parameters. The minimum distance must be 1 or more. If the maximum distance is an arbitrarily large distance, leave the parameter set to "none".

The following example keeps helices that have more than 3 nucleotides between the 5' and 3' nucleotides at the beginning of the helix, and eliminates those helices with a distance more of than 30 nucleotides between the starting nucleotides:

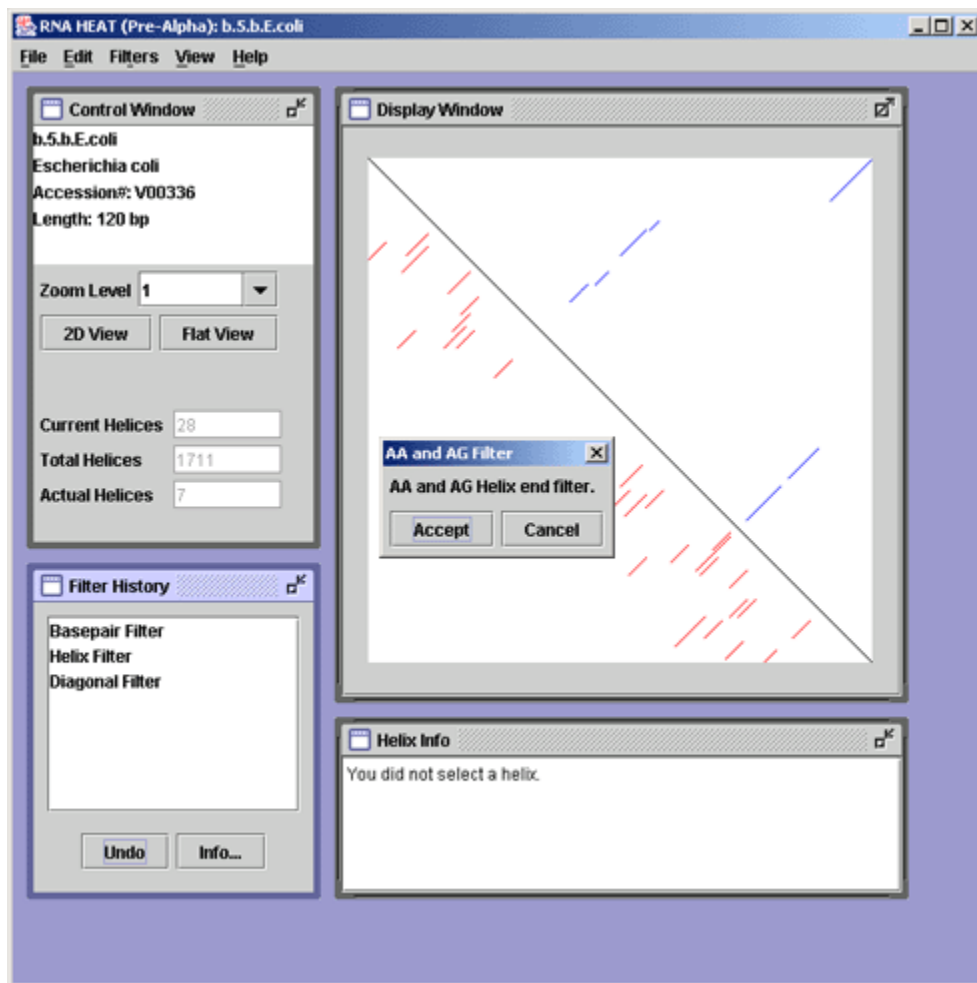


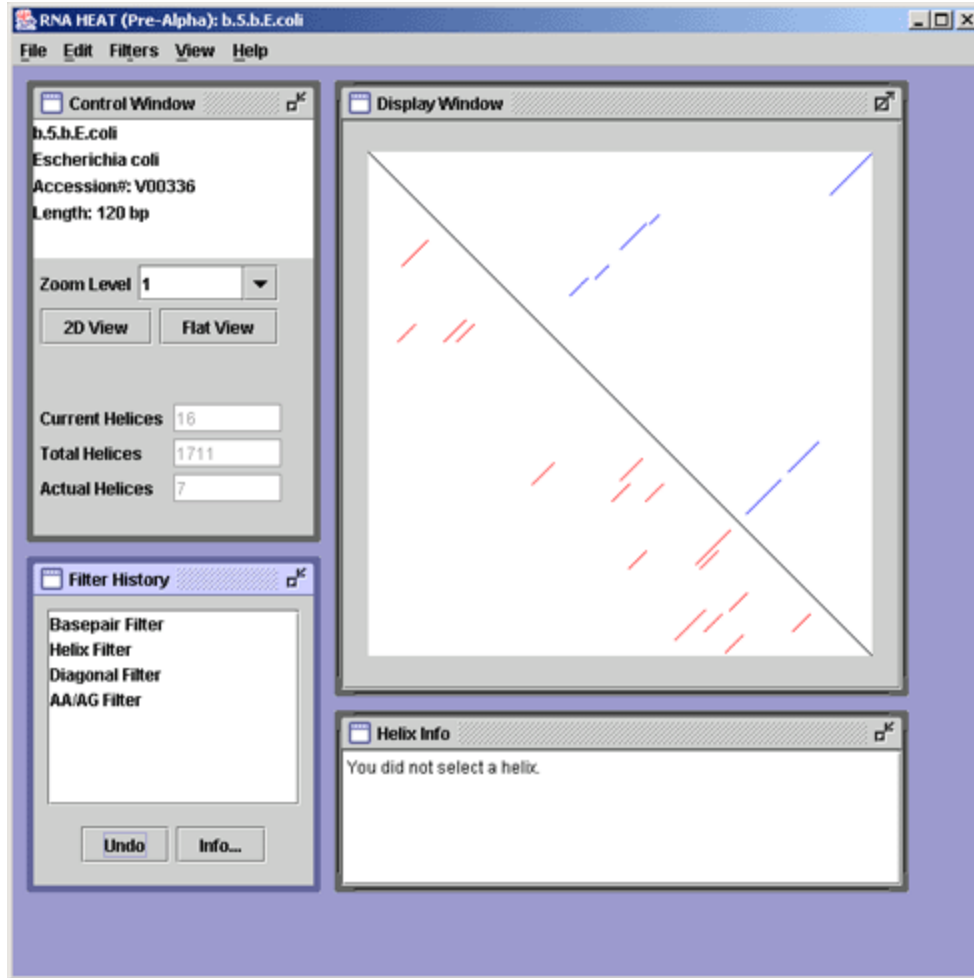


AA / AG Ends Filter ...:

This filter keeps helices that have A's, or an A and a G, just beyond the end of the helix, or just before the beginning of the helix. This particular pattern is a motif found at the ends of known helices. For more information, see the article

[AA.AG@Helix.Ends: A:A and A:G Base-pairs at the Ends of 16 S and 23 S rRNA Helices.](#)





E-Loop Filter ...:

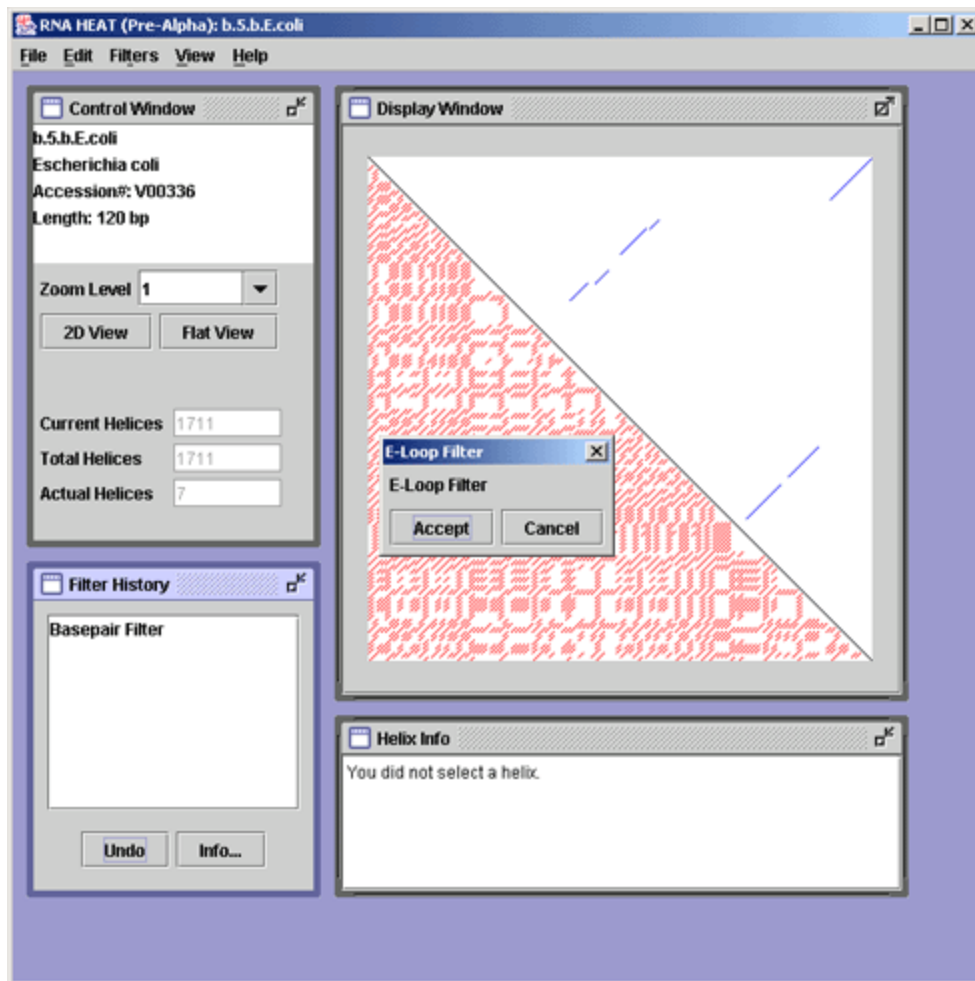
This filter keeps helices that have this pattern at the end of the helix:

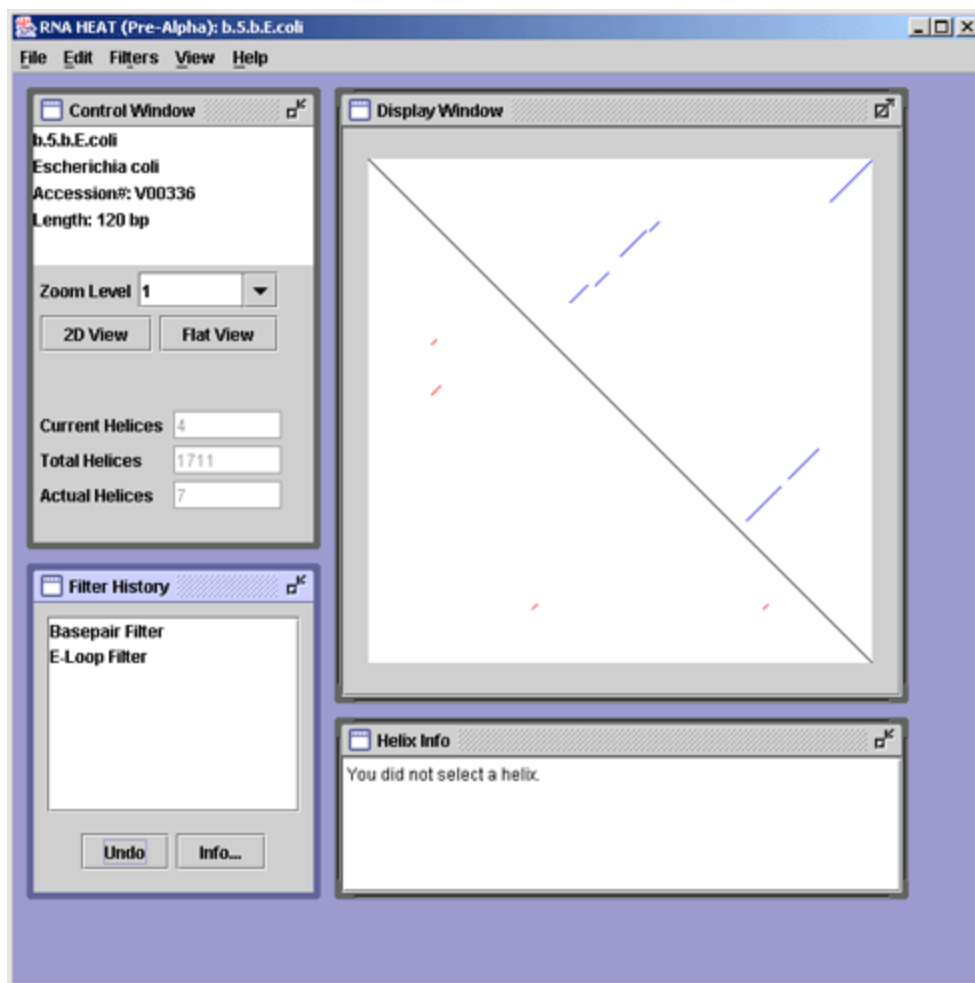
```

5'-NNGAA-3'
| |
3'-NNAUGN-5'

```

and eliminates those that do not. "NN" signifies the last 2 base-pairs at the end of the helix. This particular pattern is commonly found at the ends of known helices.





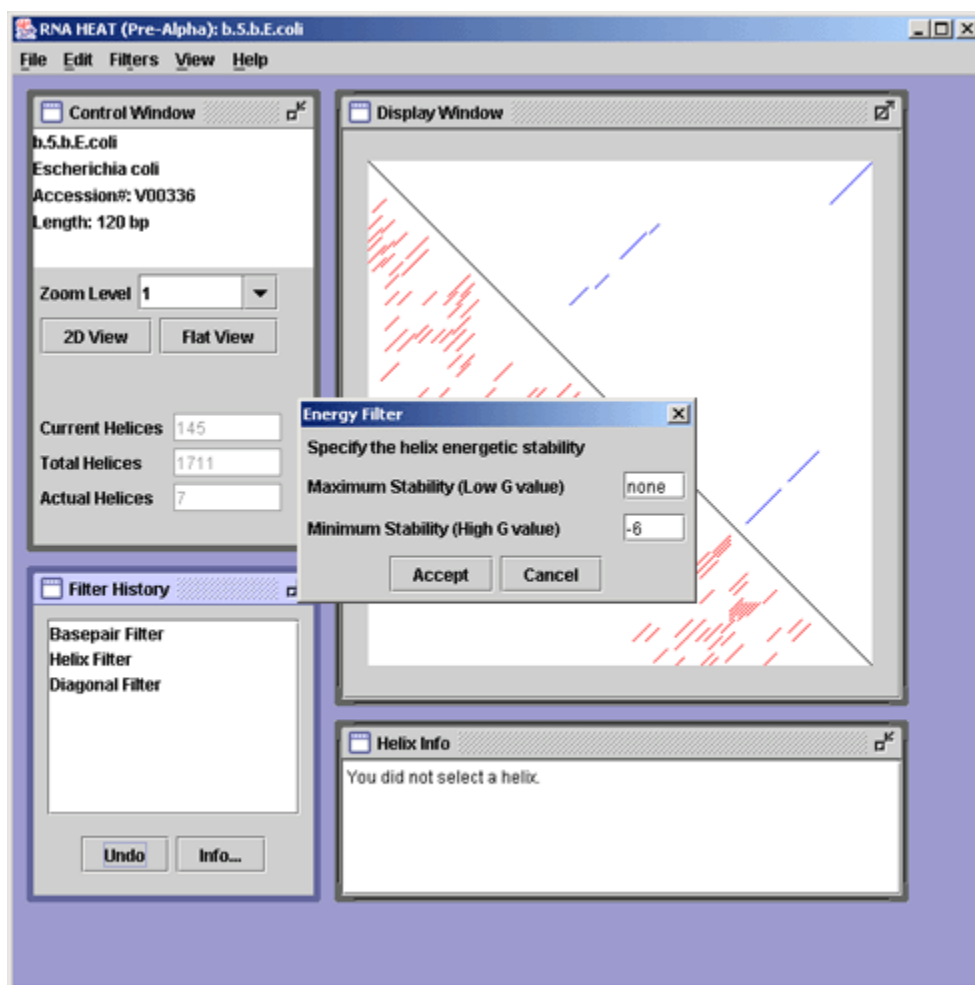
Energy Filter ...:

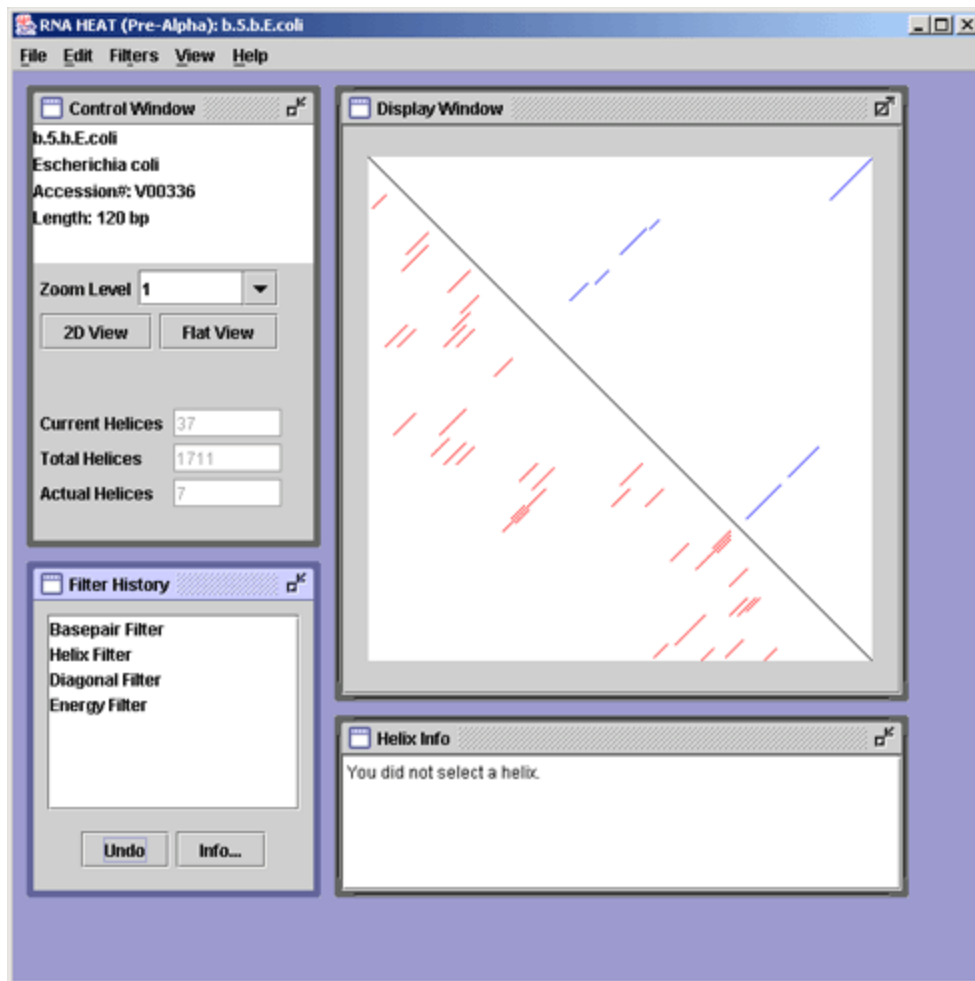
The energy filter eliminates helices that are not within a range of stability (delta G) values, and keeps those that are within the range. If helices are allowed to be arbitrarily stable (delta G arbitrarily negative), leave Maximum Stability at "none". Energies are calculated for every two consecutive base-pairs in a helix and summed to find the energy value of the whole helix. The energy value for consecutive base-pairs are looked up in a predetermined energy table. For RNA HEAT, the following table of energies is used. Non-canonical base-pairs are rare and do not have experimental delta G values, so they have been set to large number (+10), and appear in the table as an "X". The energies are given in KCal.

Basepair	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
AA	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
AC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
AG	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
AU	X	X	X	- 1.10	X	X	- 2.10	X	X	- 2.20	X	- 1.40	- 0.90	X	- 0.60	X
CA	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
CC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
CG	X	X	X	-	X	X	-	X	X	-	X	-	-	X	-	X

				2.10			2.40			3.30		2.10	2.10		1.40	
CU	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GA	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GC	X	X	X	2.20	X	X	3.30	X	X	3.40	X	2.50	2.40	X	1.50	X
GG	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GU	X	X	X	1.40	X	X	2.10	X	X	2.50	X	1.30	1.30	X	0.50	X
UA	X	X	X	0.90	X	X	2.10	X	X	2.40	X	1.30	1.30	X	1.00	X
UC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
UG	X	X	X	0.60	X	X	1.40	X	X	1.50	X	0.50	1.00	X	0.30	X
UU	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

The following example keeps helices with a delta G of -6 to -infinity, and eliminates helices with delta G of greater than -6.





Complex Distance Filter ...:

RNA HEAT: The Menu Bar: View

Display Window :

Sets the Display Window as active desktop window.

Control Window :

Sets the Control Window as active desktop window. Restores to default size if minimized beforehand.

History Window:

Sets the Filter History Window as active desktop window. Restores to default size if minimized beforehand.

Info Window:

Sets the Helix Info Window as active desktop window. Restores to default size if minimized beforehand.