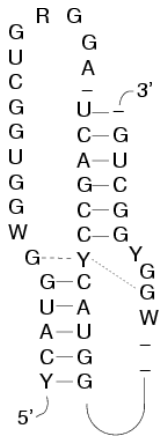
MIfold: tutorial

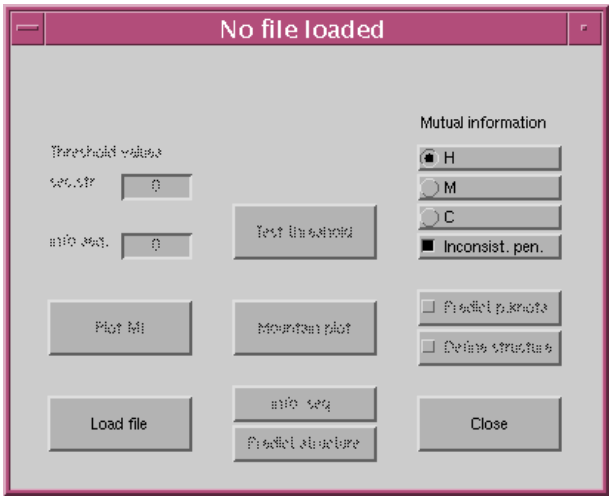
Eva Freyhult[1], Vincent Moulton, Paul Gardner

1. Download the alignment of the prion protein mRNA pseudoknot (Barrette *et al.*, 2001). The pseudoknot has the structure;

```
(((((.....................[[[[[[...)))))........]]]]]]
```



2. Open matlab and in the command window type `MIfold`. This command opens the MIfold GUI.



3. Most options are not available yet, because no alignment has been loaded. The only options that can be set before loading is selecting what kind of mutual information that should be used in the MIfold analysis. This can be changed also after loading. The default option is

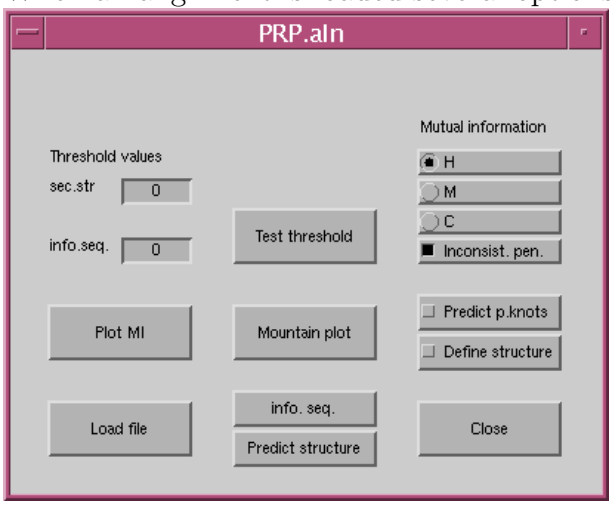
[1]Eva.Freyhult@lcb.uu.se

the classical mutual information (H) (**?**) with inconsistent sequences penalty (q).

Don't change these options, click on the `Load` button, browse to the provided file PRP.aln and load it. The prion protein mRNA alignment should be loaded in a few seconds, larger files might take a couple of minutes to load, depending on the number of sequences and the length of the alignment.

When the alignment is loaded some information is printed out in the matlab command window; the file name, number of sequences, alignment length, average pairwise sequence identity and the prior nucleotide frequencies (by default these are computed from the alignment). (For small alignments this is not appropriate and you should set the prior probabilities manually. For information of how to do this, see the description of ADVANCED USAGE in the README file.)

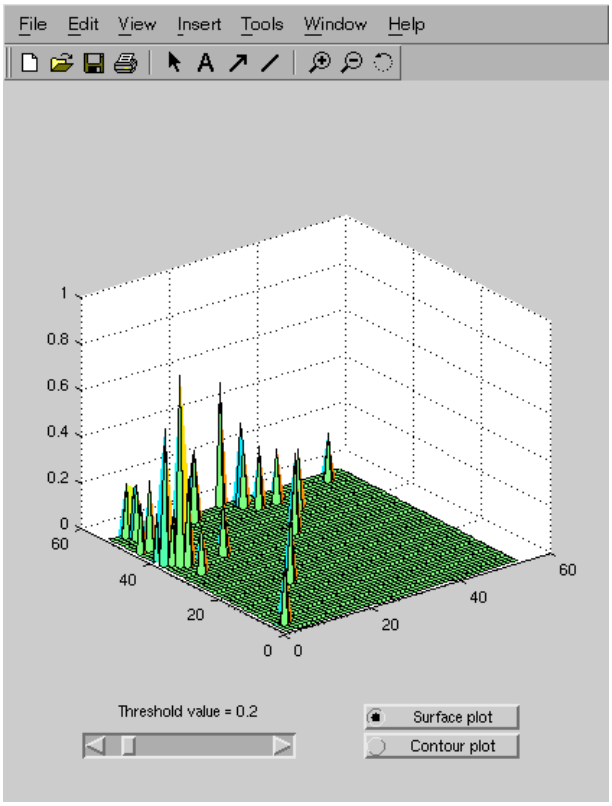
When an alignment is loaded several options should become available;



4. First we will investigate what an appropriate threshold value might be. A mutual information value measures the covariation between two sites, the higher the mutual information value, the stronger the covariation.

   Biological data contains noise and a positive mutual information value is not always due to a conserved base pair. By using a threshold and setting all mutual information values below the threshold value to zero the signal to noise ratio will increase. However, using a too high threshold value will remove true signal as well. By studying the mutual information (MI) landscape we can guess a good starting threshold.

   Click on `plot MI` to get a 3D graph of the mutual information matrix.



In the figure you can zoom and rotate the surface plot by using the options in the figure toolbar. You can also change the threshold value with the slider. Try raising and lowering the threshold value to see what happens to the landscape.

Maybe you prefer a contour plot to a surface plot? Select `Contour plot` and compare the two ways of displaying the mutual information landscape. If you start off with a threshold value of zero you can see that you have several low peaks. If you increase the threshold gradually these will disappear. The threshold value should be set to a value where the small (noise) peaks are eliminated, but yet the signal should not be deleted. For this particular example 0.2 is such a value (if you don't agree try another value and see what happens).

Close the mutual information plot and set the threshold value in the edit box in the MIfold window.

You can use `plot MI` to compare the different mutual information measures and to see how the inconsistent sequences penalty change the mutual information distribution.

5. To display the secondary structure click on the `Mountain plot` button.

To get the predicted structure in dot-bracket notation click on `Predict structure` and the predicted structure will be printed in the matlab command window. Note that the `Predict p.knots` checkbox is not checked and thus no pseudoknots will be predicted.

The mountain plot is incremented by the mutual information associated to the predicted base pairs. This means that a high incrementation corresponds to a base pair with a high covariation. Base pairs with low incrementations should be less trusted, especially if they are lonely pairs.

Below the mountain plot the sequence information content is displayed a base plot. A positive information content (if a nucleotide occurs more often than expected) is shown above the x-axis, a negative information content (a nucleotide is occuring at a lower frequency than expected) is shown below the x-axis.

The most informative sequence (see manuscript for explanation) will be printed in the command window by clicking on the pushbutton `info.seq`.

6. Now check the `Predict p.knots` box and draw the mountain plot again. Observe the difference. A threshold value can be set also for the pseudoknot prediction.

   Pseudoknots are displayed in the mountain plot as lines conneting the base pairing nucleotides. The mountain plot can be switched between incrementing the secondary structure and the pseudoknot. Use the radio buttons to try this.

7. If the sequence is known in advance you can use MIfold to investigate if the structure is supported by compensating mutations. Check the checkbox `Define structure` and enter the structure. For the PRP example above enter the structure:

   `((((((...................[[[[[[..))))))........]]]]]]`.

   If you now ask MIfold to predict the structure only base pairs that agree with the defined structure will be formed. If the mutual information for some base pairs is below the threshold no base pair will be predicted. Use the mountain plot to investigate how well supported the base pairs are.

References

Barrette, I., Poisson, G., Gendron, P. & Major, F. (2001) Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis and pattern searching. *Nucleic. Acids. Res.,* **29**, 753–758.