

RNA Folding

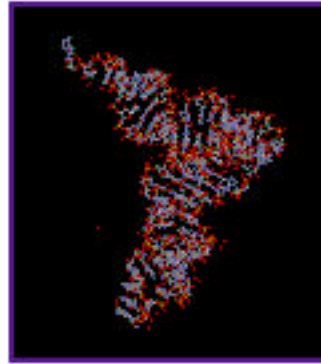
May 22, 2002

Sequence Analysis & Consulting Service

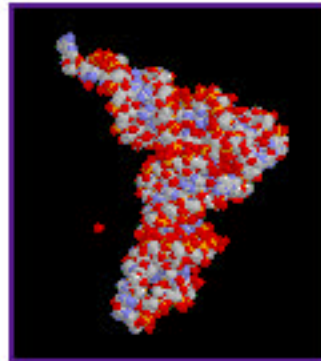
Seminar Series

Susan Johns

(johns@cgl.ucsf.edu)



view 1



view 2



view 3

First, copy over to your account the necessary data. Enter the following commands. [The first one does the actual copying. The second one moves you into this area.]

```
rnas_setup<rtm>  
cd rna_struc<rtm>
```

As a result of these commands, you are now in a sub-directory called `rna_struc` where the necessary files are located.

RNA

A linear, usually single-stranded polymer of ribonucleotides, each containing the sugar ribose in association with a phosphate group and one of the four nitrogenous bases: adenine, guanine, cytosine, or uracil. RNA is found in all living cells; in prokaryotic and eukaryotic cells, it encodes the information needed to synthesize proteins (i.e., it copies "instructions" that it receives from DNA); in certain viruses, it serves as the genome.

RNA is not simply long strands of nucleotides, rather, intra-strand base pairing produce complex structures.

There are four major classes of RNA which can be found in most organisms.

- * mRNA - messenger RNA - encodes for the formation of one or more proteins
- * tRNA - transfer RNA - small ~ 80 bases sequences that translate mRNA into amino acid sequences
- * rRNA - ribosomal RNA - form ribosomes along with ribosomal proteins
- * viral RNA

RNA structure examples

The following RNA representations are from the 1TRA x-ray structure. The data shown is of a transfer RNA (1TRA) deposited in the PDB database in 1986. The sequence is composed of 76 bases. Images were created with RasMol.

RNA Secondary structure

RNA secondary structure is similar to the alignment of other protein or nucleotide sequences, except that the sequence folds back on itself and "complementary bases" pair rather than identical or similar bases. An RNA secondary structure is a simplification of a complex 3 dimensional folding of the sequence.

The stability of a particular secondary structure is a function of a number of factors:

- * number of GC versus AU and GU base pairs
 - G and C pair with a triple hydrogen bond
 - A and U pair with a double hydrogen bond
 - G and U pair with a single hydrogen bond
- * length of stem region
- * size hairpin loop regions
- * number of unpaired bases

There are a number of different types of folding patterns that can appear in a RNA structure.

stem	hairpin loop	pseudo knot
bulge loop	internal loop	branch loop

RNA Secondary structure

The stability of a RNA secondary structure is quantified as the amount of free energy either released or used by forming base pairs. Positive free energy values require work to form the configuration, negative ones release energy.

Free energies are additive so the total free energy of a secondary structure can be determined by adding up all the component free energy values. The more negative the resulting total value, the more likely the formation of the secondary structure.

RNA secondary structure prediction algorithms have been written that seek to determine the base pair configuration with the minimum possible free energy. Empirical energy parameters are used to do this. These parameters summarize the free energy changes associated with all possible base pairing combinations.

These algorithms don't work well for determining pseudo knot structures.

RNA Web Resources

There are a number of web sites that allow prediction of RNA secondary structure. Since there are different types of RNA, and such a determination can be CPU intensive, it is wise to check out the web site to see if it will handle the intended sequence.

possible sites

* Zuker's MFOLD

500 bases immediate, 3000 bases batch - raw sequence

SSU rRNA Secondary Structure Prediction and Alignment

for rRNA requires 1000 to 2500 bases - fasta format

* Vienna RNA package

300 bases - raw sequence

*RNA folding

RNA folding prediction started out with the finding of repeat regions in RNA sequences. There are a number of programs that allow do this.

EMBOSS programs

einverted
etandem

equicktandem
palindrome

GCG program

stemloop

MFOLD

MFOLD is one of the most heavily used programs for predicting RNA secondary structure. Its author, Dr. Michael. Zuker, currently of Rensselaer Polytechnic Institute, has been working in this area for some time.

The a version of the program is included in GCG, available at a number of web sites and there is a standalone version for multi-user platforms as well. The standalone version will be used here. This is due to the fact that this version of the program is more powerful than the one in GCG.

MFOLD goes through and determines a large number of sub-optimal folded structures and their associated postscript files. This means that running any sequence over 200 bases may generate a large number of files that take up a lot of disk space.

MFOLD

MFOLD uses nearest neighbor thermodynamics rules and specific constraints to either force or prohibit base pairs. It is attempting to take into account stacking interactions between adjacent helices. Pseudo knots are not considered.

The program can be run with a number of parameters that can be adjusted to improve the prediction.

In normal operation MFOLD produces a number of files. The most significant of these are the energy dot plot and a number of folded structures.

The program is supposed to be able to generate gif images of the results, but that aspect of the software isn't working on socrates at the moment. Only postscript images are created. For each run, 12 overhead files are created as well as 3 for the energy plot and 6 for each possible structure.

MFOLD

MFOLD is very simple to use on the command line for a normal run. The program accepts genbank, embl, fasta and intelligentsics formatted files as input.

mfold SEQ=sequence_name

The program can be made to produce html result files. The first file contain a text version of the structure and links to the MFOLD documentation. File 2 contains information on loop free-energy decomposition values for the run. The third file only contains the text version of the found structure.

mfold SEQ=sequence_name RUN_TYPE=html

MFOLD

MFOLD run with a longer input sequence produces multiple structures.

MFOLD

MFOLD can eat up lots of disk space. Therefore, if you want to use longer sequences, work in the /usr/tmp location. Create a directory there, move over your data and run the program. Print off the postscript files to look at and then remove the data. [Files only stay in this location for 2 weeks.]

A sequence of 1687 bases generated 189 output files, of which 29 were postscript structure files.

URL

To better understand the output files, look at the example results given in the on-line documentation.

URLs used in the seminar

1TRA.pdb file

<http://www.sacs.ucsf.edu/Training/rnastruc/1TRA-pdb.html>

Dr. Walter's materials:

http://www.finchcms.edu/biochem/Walters/rna_folding.html

Drs. Nelson and Istrail's materials:

<http://www.santafe.edu/~pth/rna.html>

Dr. Michael Zuker's materials:

<http://bioinfo.math.rpi.edu/~zukerm/Bio-5495/RNAfold-html/rnafold.html>

Zuker's mfold site:

<http://bioinfo.math.rpi.edu/~mfold/rna/form1.cgi>

rRNA secondary structure prediction

<http://www.cse.ucsc.edu/research/compbio/ssurrna.html>

Vienna RNA package:

<http://www.tbi.univie.ac.at/~ivo/RNA/>

EMBOSS repeat software documentation:

<http://www.sacs.ucsf.edu/Documentation/emboss/einverted.html>

<http://www.sacs.ucsf.edu/Documentation/emboss/equicktandem.html>

<http://www.sacs.ucsf.edu/Documentation/emboss/etandem.html>

<http://www.sacs.ucsf.edu/Documentation/emboss/palindrome.html>

GCG repeat software documentation:

<http://www.sacs.ucsf.edu/Documentation/gcghelp/stemloop.html>

MFOLD documentation:

<http://www.sacs.ucsf.edu/Documentation/mfold/>

<http://www.sacs.ucsf.edu/Documentation/mfold/node10.html>

<http://www.sacs.ucsf.edu/Documentation/mfold/node11.html>

<http://www.sacs.ucsf.edu/Documentation/mfold/node7.html>

<http://www.sacs.ucsf.edu/Documentation/mfold/node14.html>

<http://www.sacs.ucsf.edu/Documentation/mfold/node15.html>

<http://www.sacs.ucsf.edu/Documentation/mfold/node16.html>

complete package
energy dot plot
foldings
program operation
example 1
example 2
example 3