

# **RNA structure determination**

**Experimental techniques**

**&**

**Computational prediction**

# Why study RNA structure?

Biological function highly depends upon RNA folding:  
The structure of an RNA molecule determines both the function of the molecule and the mechanism behind that function.

From Felden 2007:

"Proper functioning of RNAs requires the formation of intricate three-dimensional (3D) structures, as well as the ability to efficiently interconvert between multiple functional states."

RNA structure might reveal RNA's role in the origin and evolution of life on earth

RNA structure might function as drug target.  
Example: stem-loop II motif in RNA element of SARS virus genome (M.P. Robertson 2005)

Functions of RNA include:

- coding
- information transfer
- catalytic activities

# **Structure determination**

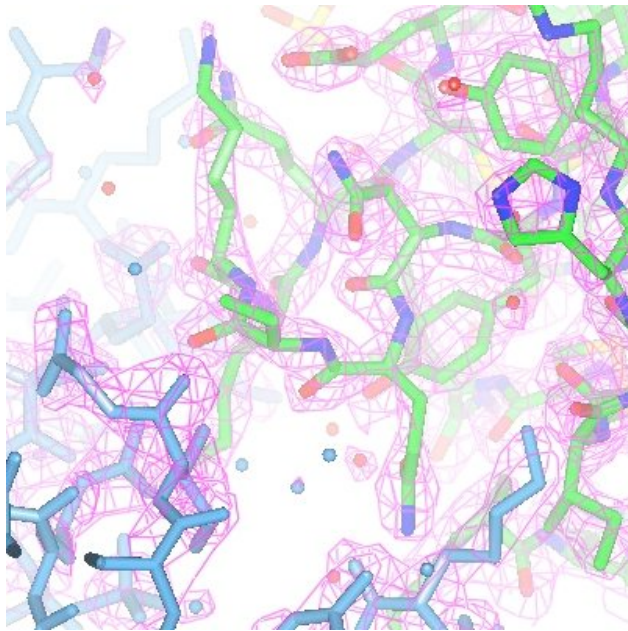
## **Experimental techniques**

**B. Felden. Curr. Opinion in Microbiology (2007). 10:286-291**

# High resolution methods

X-ray crystallography

Nuclear Magnetic Resonance (NMR) spectroscopy

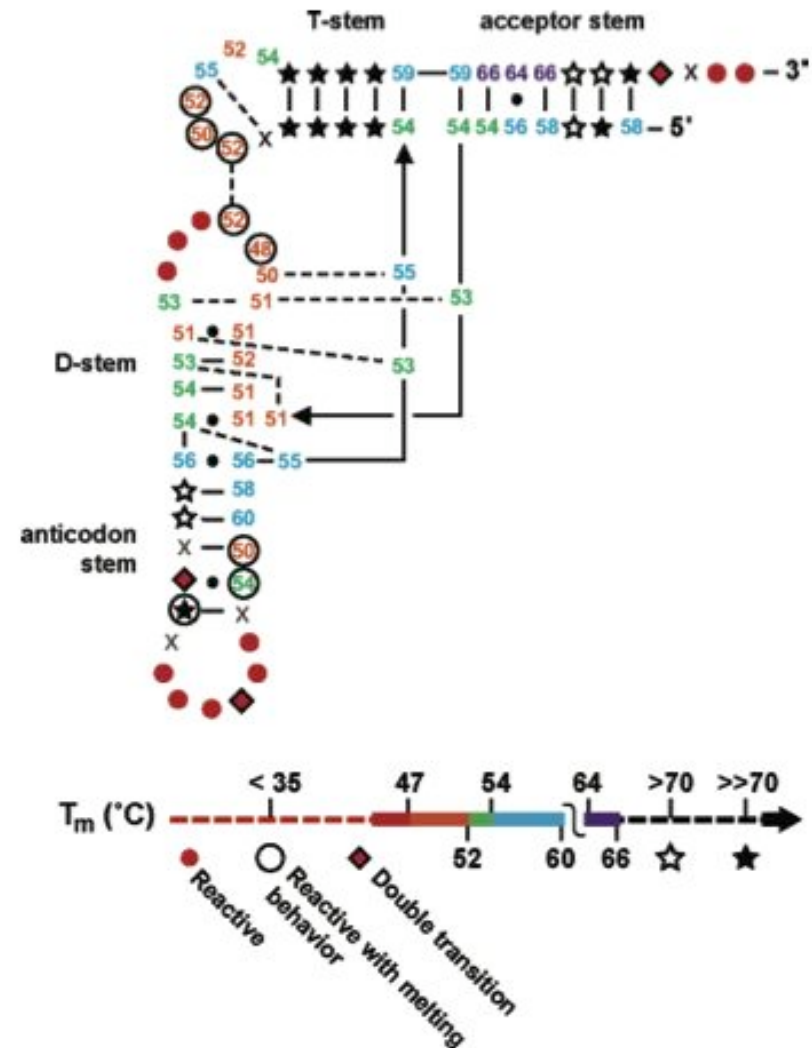


The CryoEM facility has a 300KV, helium-cooled, FEI Polara FEG and a FEI 200KV Sphera microscope.

**Cryo-electron microscopy (Cryo-EM)**

# Low(er) resolution methods

- Chemical/enzymatic probing
- Thermal denaturation (melting studies)
- Mass spectrometry
- RNA engineering



Selective 2'-Hydroxyl Acylation analyzed by Primer Extension (SHAPE chemistry)

# Structure determination

## Computational structure prediction

### Review articles:

- M. Zuker. *Curr. Opin. in Structural Biology* (2000). 10:303-310.
- F. Major & R. Griffey. *Curr. Opin. in Structural Biology* (2001). 11:282-286.
- P.P. Gardner & R. Giegerich. *BMC Bioinformatics* (2004). 5:140.
- D.H. Mathews. *J. Mol. Biol.* (2006). 359:526-532.
- D.H. Mathews & D.H. Turner. *Curr. Opin. in Struct. Biol.* (2006). 16:270-278.
- Y. Ding. *RNA* (2006). 12:323-331.

# Base pair maximization

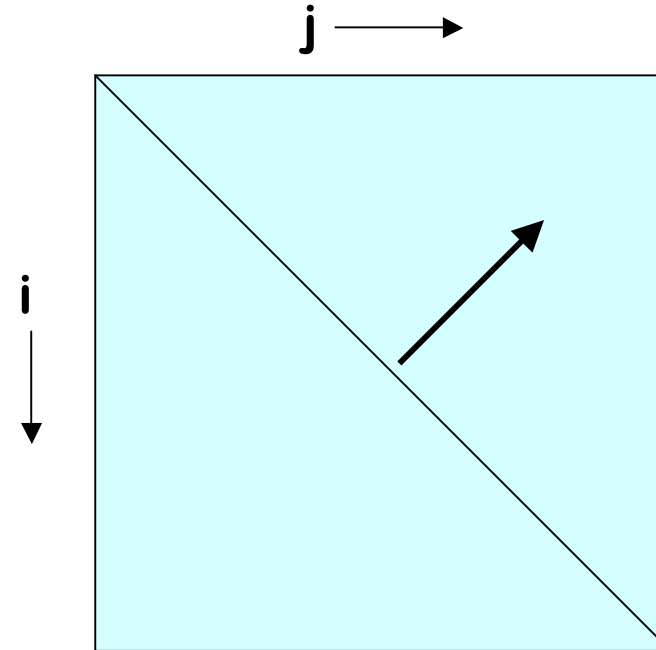
Nussinov & Jakobson 1980

Dynamic programming

-- fill stage

-- traceback

Recursion rules



$$F(i, j) = \max \begin{cases} s(x_i, x_j) + F(i + 1, j - 1) \\ F(i, k) + F(k + 1, j) \end{cases} \quad i \leq k < j$$

Initialize

$F(i, i) = 0$  for  $i = 1$  to  $L$

$F(i, i-1) = 0$  for  $i = 2$  to  $L$

Base pair score

$s(x_i, x_j) = 1$  if  $i$  and  $j$  base pair, 0 otherwise





# Solution

		j →																
		1													15			
i ↓	1	A	0	0	0	0	0	1	2	2	3	3	3	4	4	4	5	
	C	0	0	0	0	0	1	2	2	2	2	3	4	4	4	4	5	
	C		0	0	0	0	1	1	1	1	2	3	3	3	3	3	4	
	A			0	0	0	0	0	0	1	2	2	2	2	2	2	3	
	A				0	0	0	0	0	1	1	1	1	2	2	2	3	
	G					0	0	0	0	0	0	0	0	1	2	3	3	
	G						0	0	0	0	0	0	0	1	2	3	3	
	G							0	0	0	0	0	0	1	2	3	3	
	U								0	0	0	0	0	1	2	3	3	
	U									0	0	0	0	1	2	3	3	
	G										0	0	0	1	2	3	3	
	G											0	0	1	2	3	3	
	A												0	1	2	3	3	
	A														0	1	2	3
	15	C															0	0

Initialization

Traceback

Number of base pairs in optimal solution: 5

Actual base pairs:  
 [(2,7),(3,6),(8,15),  
 (9,14),(10,13)]

# Thermodynamics

- $\Delta G = \Delta H - T\Delta S$   
 $\Delta H$  is enthalpy,  $\Delta S$  is entropy, and  $T$  is the temperature in Kelvin.
- Molecular interactions, such as hydrogen bonds, van der Waals and electrostatic interactions contribute to the  $\Delta H$  term.  $\Delta S$  describes the change of order of the system.
- Thus, both molecular interactions as well as the order of the system determine the direction of a chemical process.
- For any nucleic acid solution, it is extremely difficult to calculate the free energy from first principle
- Biophysical methods can be used to measure free energy changes

# Thermodynamics

- **Gibbs Free Energy,  $G$**
- **Describes the energetics of biomolecules in aqueous solution. The change in free energy,  $\Delta G$ , for a chemical process, such as nucleic acid folding, can be used to determine the direction of the process:**
- **$\Delta G=0$ : equilibrium**
- **$\Delta G>0$ : unfavorable process**
- **$\Delta G<0$ : favorable process**
- **Thus the natural tendency for biomolecules in solution is to minimize free energy of the entire system (biomolecules + solvent).**

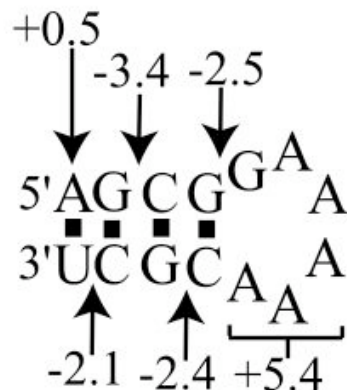
## RNA folding

**Equilibrium between strands in folded/unfolded state. Lowest free energy structure is the most represented conformation at equilibrium**

# Free energy minimization (MFE)

Zuker & Stiegler 1981

Nearest-neighbor rules: free energies assigned to base pair stacks and to loops (unpaired regions). In helices, energy contributions depend on a base pair and its adjacent pair.



$$\begin{aligned}\Delta G_{37}^{\circ} \text{ total} &= 0.5 - 2.1 - 3.4 - 2.4 - 2.5 + 5.4 \\ &= -4.5 \text{ kcal/mol}\end{aligned}$$

Energy parameters from:  
Mathews et al. 1999

Dynamic programming solution ( $O(N^3)$ )  
Two matrices:  $W$  &  $V$

$W(i,j)$  = the minimum free energy of all possible admissible structures formed from the subsequence  $S(i,j)$

$V(i,j)$  = the minimum free energy of all possible admissible structures formed from  $S(i,j)$  in which  $S_i$  and  $S_j$  base pair with each other.

Implementations:

mFold (now unafold)

RNAfold (in Vienna RNA package)

# Limitations

The accuracy of MFE methods is limited:

- free energy nearest-neighbor model is incomplete (e.g. motifs)
- some effects on stability are non-nearest-neighbor (bulge loops and single non-canonical pairs)
- not all RNA sequences are at equilibrium (i.e. kinetics might also be important!)
- Topological limitations (e.g. no pseudoknots!)
- RNA sequence might have multiple conformations (e.g. riboswitches, tRNA)

Need for suboptimal structure prediction:

RNAsubopt

- For sequence length  $N \sim 1.8^N$  structures
- Sequence 100 nucleotides  $\sim 3 \times 10^{25}$  structures. Suppose calculate 1000 structs/sec it would take  $10^{15}$  years!!!
- Zuker & Stiegler 1981, Mathews et al. 1999: heuristic set of suboptimal structures
- Wuchty 1999: all suboptimal structures within an energy increment above the MFE

# Partition function

McCaskill 1990

A partition function is a quantity that encodes the statistical properties of a system in thermodynamic equilibrium. It connects the mechanics to thermodynamics

$$Q = \sum e^{-\Delta G/RT}$$

Sum over all possible secondary structures

R = gas constant

T = absolute temperature

Given the partition function, the probability of a given base pair is:

$$P = \frac{\sum e^{-\Delta G/RT}}{Q}$$

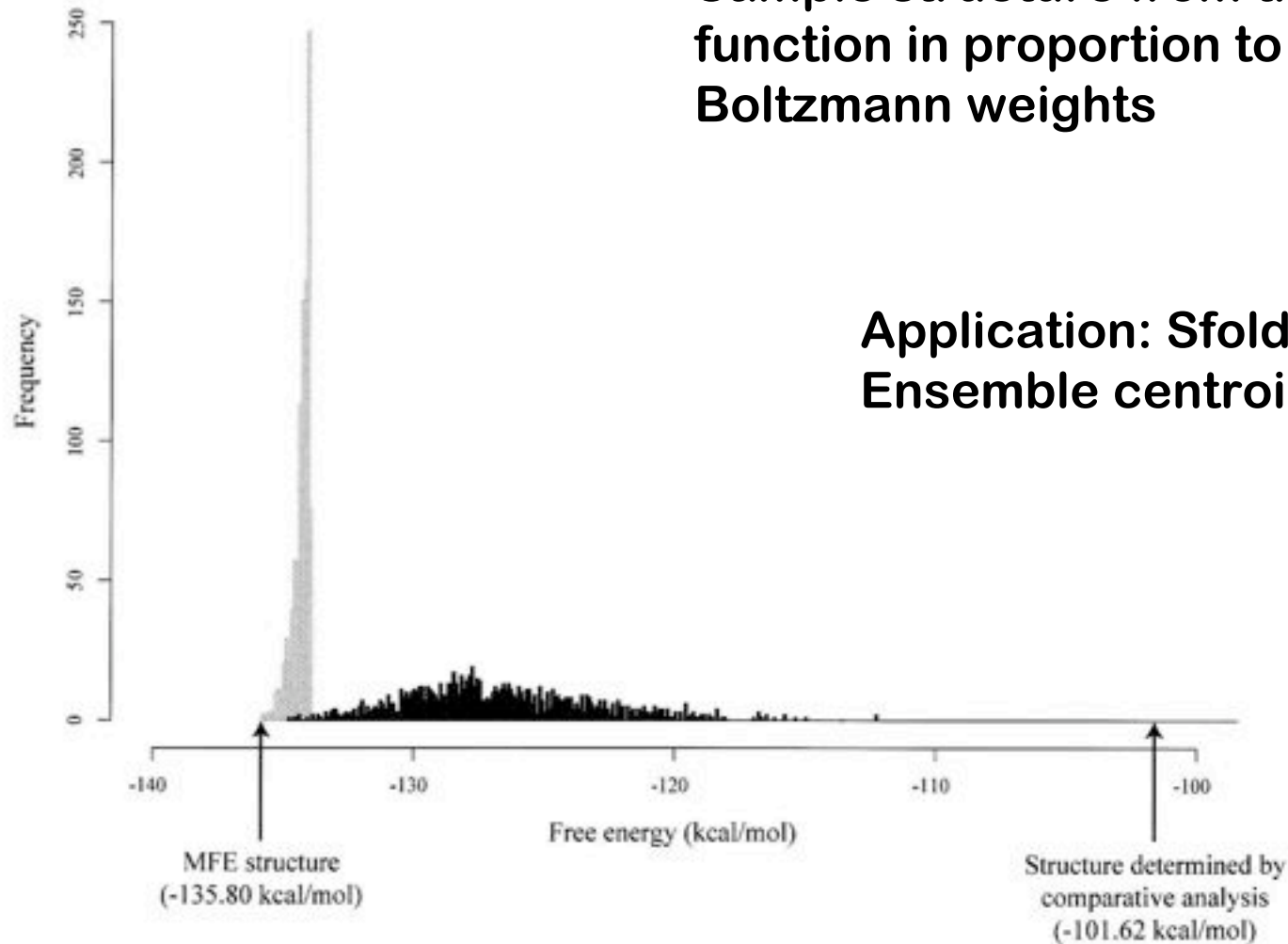
Sum over all secondary structures that contain the given base pair

# Statistical sampling

Ding & Lawrence 2003

Sample structure from the partition function in proportion to their Boltzmann weights

Application: Sfold  
Ensemble centroid prediction



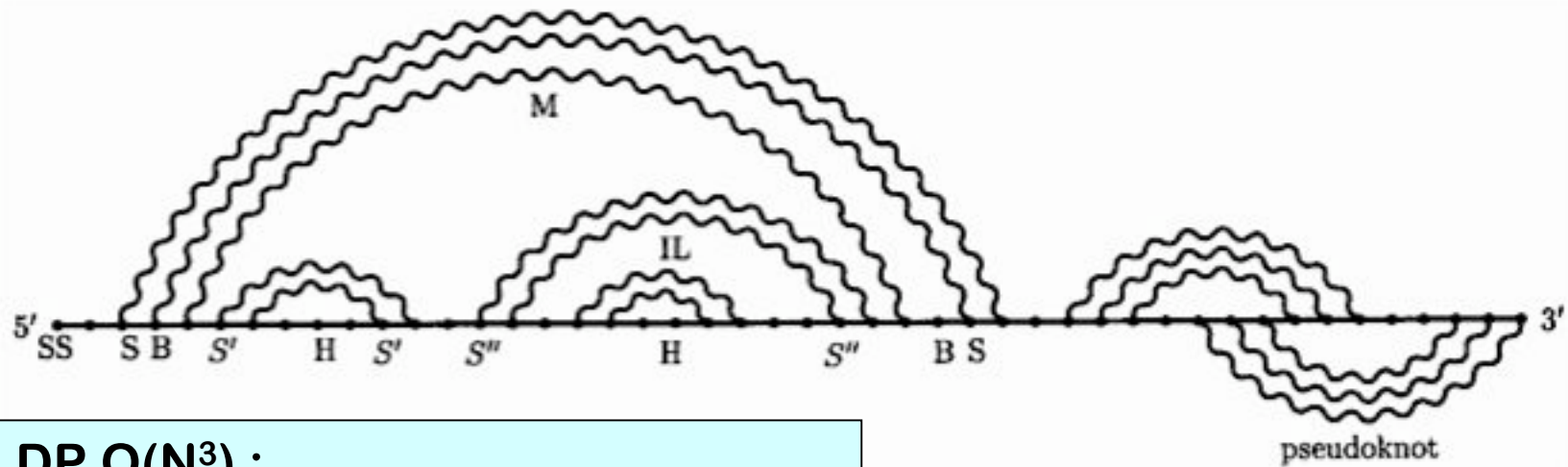
# Pseudoknot prediction

PKNOTS: E. Rivas & S.R. Eddy 1999

Two matrices  $v_x$  and  $w_x$ , like normal MFE

Complex recursion rules to include pseudoknots

Complexity:  $O(N^6)$



Normal DP  $O(N^3)$  :

2x sequence length = 8x computer time

Pseudoknots  $O(N^6)$  :

2x sequence length = 64x computer time!

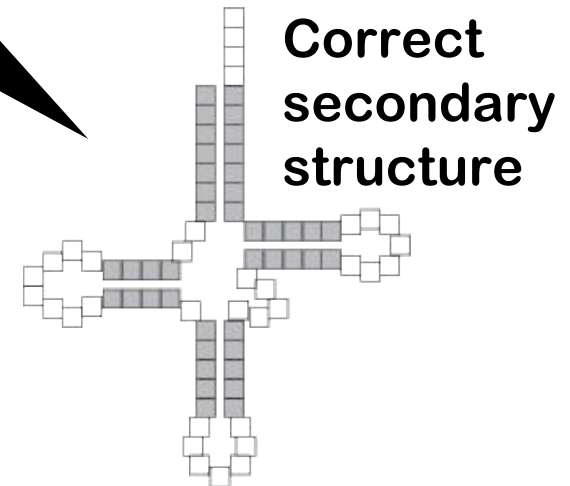
Other approximations  
to predict pseudoknots



# Comparative approaches

```
(((((.....))))).(((.....))).....((((.....)))))).....  
GCCCGGAUAGCUCAGUC-GGU--AGAGCAGGGGAUUGAAAAUCCCGUGUCCUUGGUUCGAUCCGAGUCCGGGCACCA  
GCCGAAAUAGCUCAGUU-GGG--AGAGCGUUAGACUGAAGAUCUAAAGGUCCUGGUUCAUCCCGGGUUUCGGCACCA  
GCCGCCGUAGCUCAGC--GGG--AGAGCGCCCGGCUGAAGACCGGGUGGUCCGGGGUUCGAAUCCCGCGGGCGGCACCA  
GGCCAGGUAGCUCAGU-CGGU-AUGAGCGUCCGCCUGAAAAGCGGAAGGUCGGCGGUUCGAUCCCGCCUUGGCCACCA  
GGUUCAGUAGCUCAGU-UGGU--AGAGCAAUGGAUUGAAGCUCCAUGUGUCGGCAGUUCGACUCUGUCCUGAACCACCA
```

**Input = multiple related sequences (homologues)**  
**Two steps:**  
-- sequence alignment  
-- prediction of common structure for all sequences



# Covariation

Covariation, detectable at the sequence level, is caused by compensatory mutations (mostly isosteric base pairs)

**Mutual information**

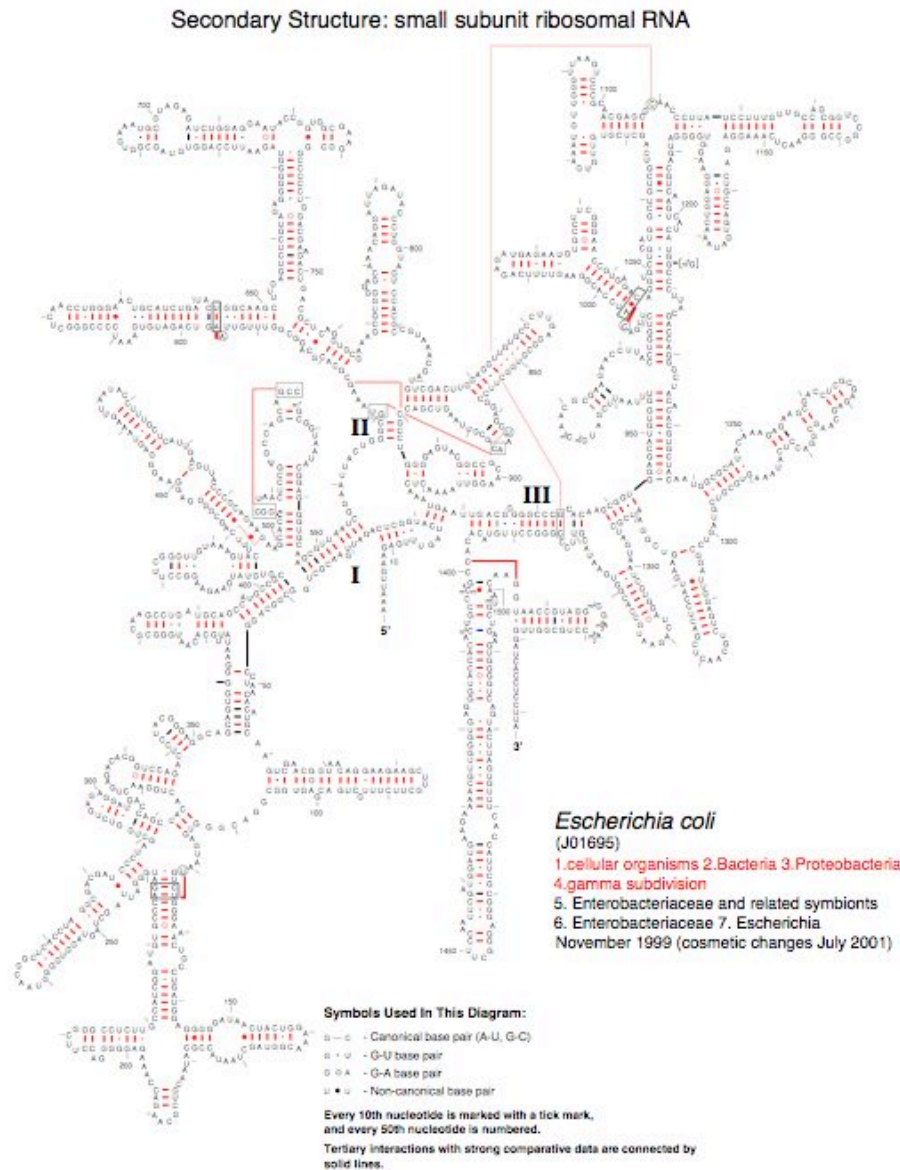
$$H(i, j) = F_{i,j}(N_1, N_2) \log_2 \frac{F_{i,j}(N_1, N_2)}{f_i(N_1) f_j(N_2)}$$

$$N_1, N_2 \in U, C, A, G$$

$F_{i,j}(N_1, N_2)$       Frequency of the (N1,N2) nucleotide pair in columns i and j

$f_i(N_1), f_j(N_2)$       Frequencies of nucleotides N1 and N2 in columns i and j

# Ribosomal RNA



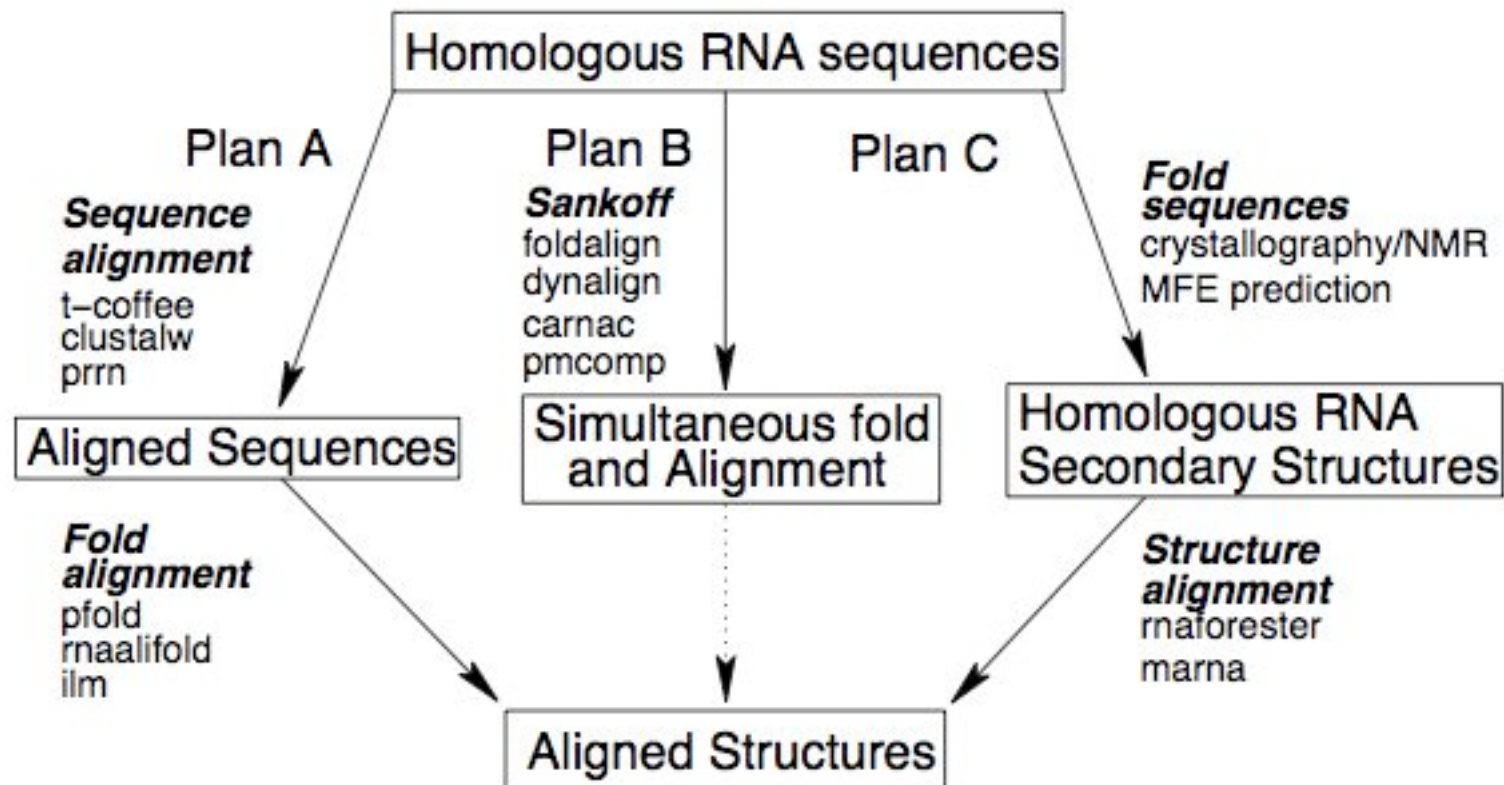
Large-scale alignment and covariation analysis led to accurate prediction of the SSU/LSU rRNA structures.

Using the recently solved crystal structures of the ribosome, the models were shown to be over 97% accurate (Gutell et al. 2002)

Comparative RNA web  
Cannone et al. 2002

Several approaches:

- 1) Simultaneously align and fold
- 2) Predict secondary structure for a given alignment
- 3) Fold sequences, align structures



# Align and fold

D. Sankoff 1985

Simultaneous RNA sequence alignment and folding  
Theoretically ideal, but computationally over-expensive  
Complexity =  $O(n^{3m})$  time and  $O(n^{2m})$  in space  
 $n$  = sequence length and  $m$  = number of sequences

Many approximations exist making the Sankoff method practical (e.g. FOLDALIGN and Dynalign)

**FOLDALIGN: Gorodkin et al. 1997**

Limits the maximum distance between paired nucleotides and the maximum length difference for fragments being aligned

**Dynalign: Mathews & Turner 2002**

Limits the maximum difference in index for a nucleotide in the first sequence that will be aligned to the second sequence

# Predict from alignment

**RNAalifold: Hofacker et al. 2002**

**Extension to Zuker-Stiegler algorithm for computing a consensus structure from RNA alignments**

**Combines thermodynamics and covariation**

**Pfold: Knudsen & Hein 1999, 2003**

**Stochastic context free grammar (SCFG)**

$S \rightarrow aSu \mid uSa \mid cSg \mid gSc$

$S \rightarrow aS \mid cS \mid gS \mid uS$

$S \rightarrow Sa \mid Sc \mid Sg \mid Su$

$S \rightarrow SS$

$S \rightarrow \epsilon$

**R. Dowell & S.R. Eddy 2004**

**Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction**

# Fold and align structures

**RNAforester: Hochsmann et al 2003**

**Tree alignment model**

**Pairwise alignment of two input structures**

**Consensus shapes: J. Reeder & R. Giegerich 2005**

**Abstract shapes: representation of RNA secondary structure that displays the branching pattern of the helices**

**((((..((..(((...))))).((...))..)))**

**Level 5 shape: [[][]]**

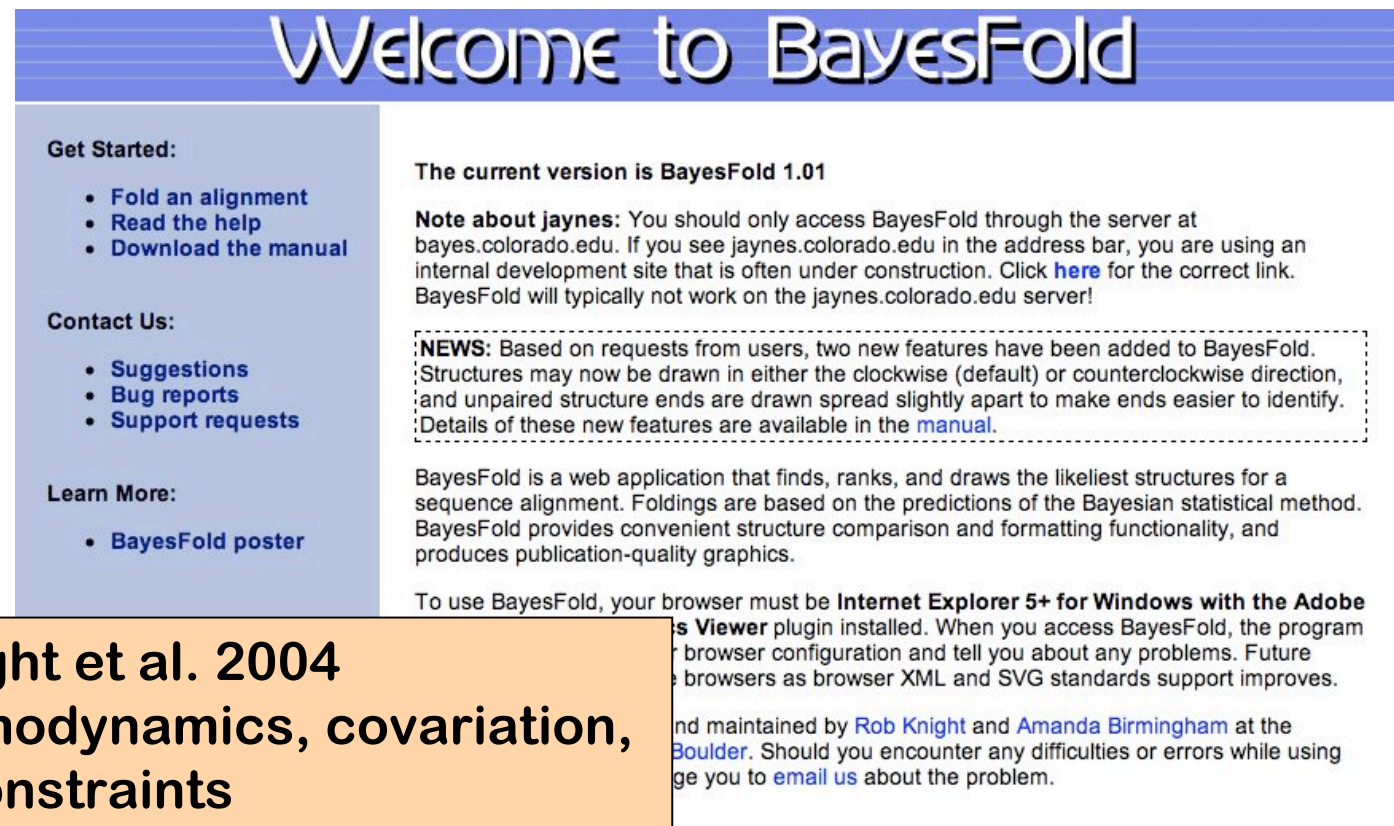
**Level 3 shape: [[]][[]]**

- 1) Generate list of abstract shapes (much shorter than the list of suboptimal structures within a certain energy range)**
- 2) Find the lowest free energy abstract shape common to all sequences**



# Combine multiple sources

**RNAstructure: Mathews et al. 2004**  
Combines thermodynamics, covariation,  
chemical modification constraints



The screenshot shows the BayesFold website homepage. At the top is a blue banner with the text "Welcome to BayesFold" in white. Below the banner is a light blue sidebar on the left containing three sections: "Get Started:" with links for "Fold an alignment", "Read the help", and "Download the manual"; "Contact Us:" with links for "Suggestions", "Bug reports", and "Support requests"; and "Learn More:" with a link for "BayesFold poster". The main content area on the right has a white background. It starts with the text "The current version is BayesFold 1.01". Below this is a "Note about jaynes" section. A dashed box contains a "NEWS" section. Further down is a paragraph describing BayesFold as a web application. At the bottom of the screenshot is a yellow box with text about BayesFold: Knight et al. 2004, combining thermodynamics, covariation, and experimental constraints.

## Welcome to BayesFold

**Get Started:**

- [Fold an alignment](#)
- [Read the help](#)
- [Download the manual](#)

**Contact Us:**

- [Suggestions](#)
- [Bug reports](#)
- [Support requests](#)

**Learn More:**

- [BayesFold poster](#)

The current version is BayesFold 1.01

**Note about jaynes:** You should only access BayesFold through the server at bayes.colorado.edu. If you see jaynes.colorado.edu in the address bar, you are using an internal development site that is often under construction. Click [here](#) for the correct link. BayesFold will typically not work on the jaynes.colorado.edu server!

**NEWS:** Based on requests from users, two new features have been added to BayesFold. Structures may now be drawn in either the clockwise (default) or counterclockwise direction, and unpaired structure ends are drawn spread slightly apart to make ends easier to identify. Details of these new features are available in the [manual](#).

BayesFold is a web application that finds, ranks, and draws the likeliest structures for a sequence alignment. Foldings are based on the predictions of the Bayesian statistical method. BayesFold provides convenient structure comparison and formatting functionality, and produces publication-quality graphics.

To use BayesFold, your browser must be **Internet Explorer 5+ for Windows with the Adobe Flash Viewer** plugin installed. When you access BayesFold, the program will check your browser configuration and tell you about any problems. Future versions will support other browsers as browser XML and SVG standards support improves.

BayesFold is developed and maintained by [Rob Knight](#) and [Amanda Birmingham](#) at the University of Colorado Boulder. Should you encounter any difficulties or errors while using BayesFold, please get you to [email us](#) about the problem.

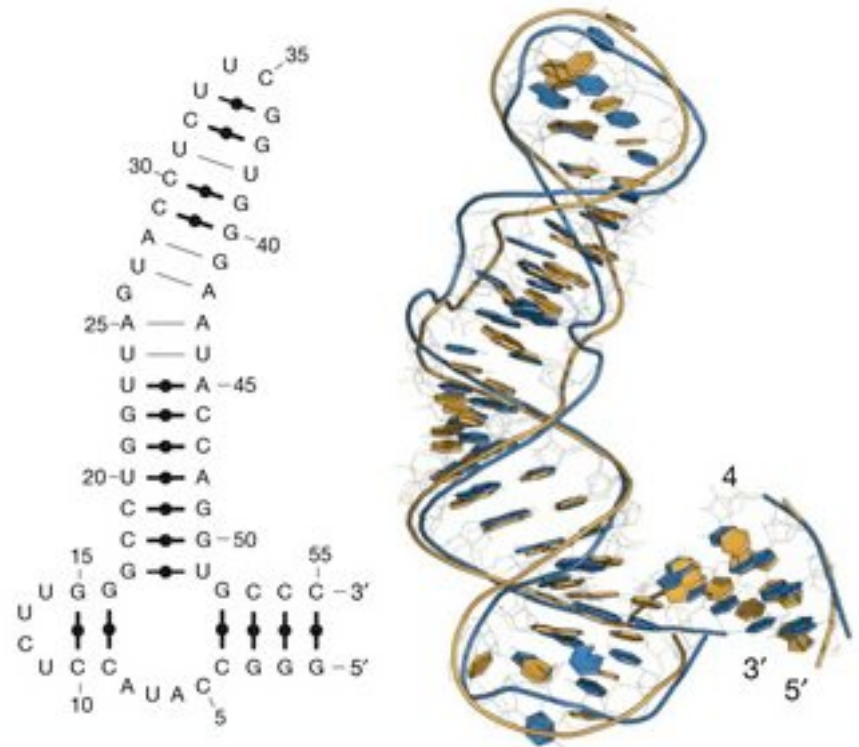
**BayesFold: Knight et al. 2004**  
Combines thermodynamics, covariation,  
experimental constraints



# Latest greatest

M. Parisien and F. Major. Nature (2008). 452:51-55.  
The MC-Fold and MC-Sym pipeline infers RNA  
structure from sequence data

Structure prediction pipeline  
-- secondary structure  
-- tertiary structure  
Based on nucleotide cyclic  
motifs (NCMs)  
Includes all base pairs (not only  
canonical pairs)



# Accuracy

Benchmark study: Garder & Giegerich 2004

Accuracy metrics:

True/False positives, True/False negatives

Sensitivity, positive predictive value, Matthews correlation coefficient, and others

Comparative approaches in general more accurate than single-sequence MFE methods.

-- Several factors limit the accuracy of MFE methods (see earlier slide).

-- Structure & function are more conserved than sequence.

Multiple sequences provide many constraints on the structure.

# Summary

## Experimental approaches

- several experimental techniques are available, ranging in precision
- experiments are difficult, expensive, and time-consuming
- if they work, they are a great source of structural information

## Computational methods are necessary

- single-sequence versus multi-sequence methods
- minimum free energy prediction methods
- suboptimal foldings, partition function, and statistical sampling
- comparative approaches, three strategies
- integration of multiple sources of evidence