

7.91 / 20.490 / 6.874 / HST.506

7.36 / 20.390 / 6.802

C. Burge Lecture #10

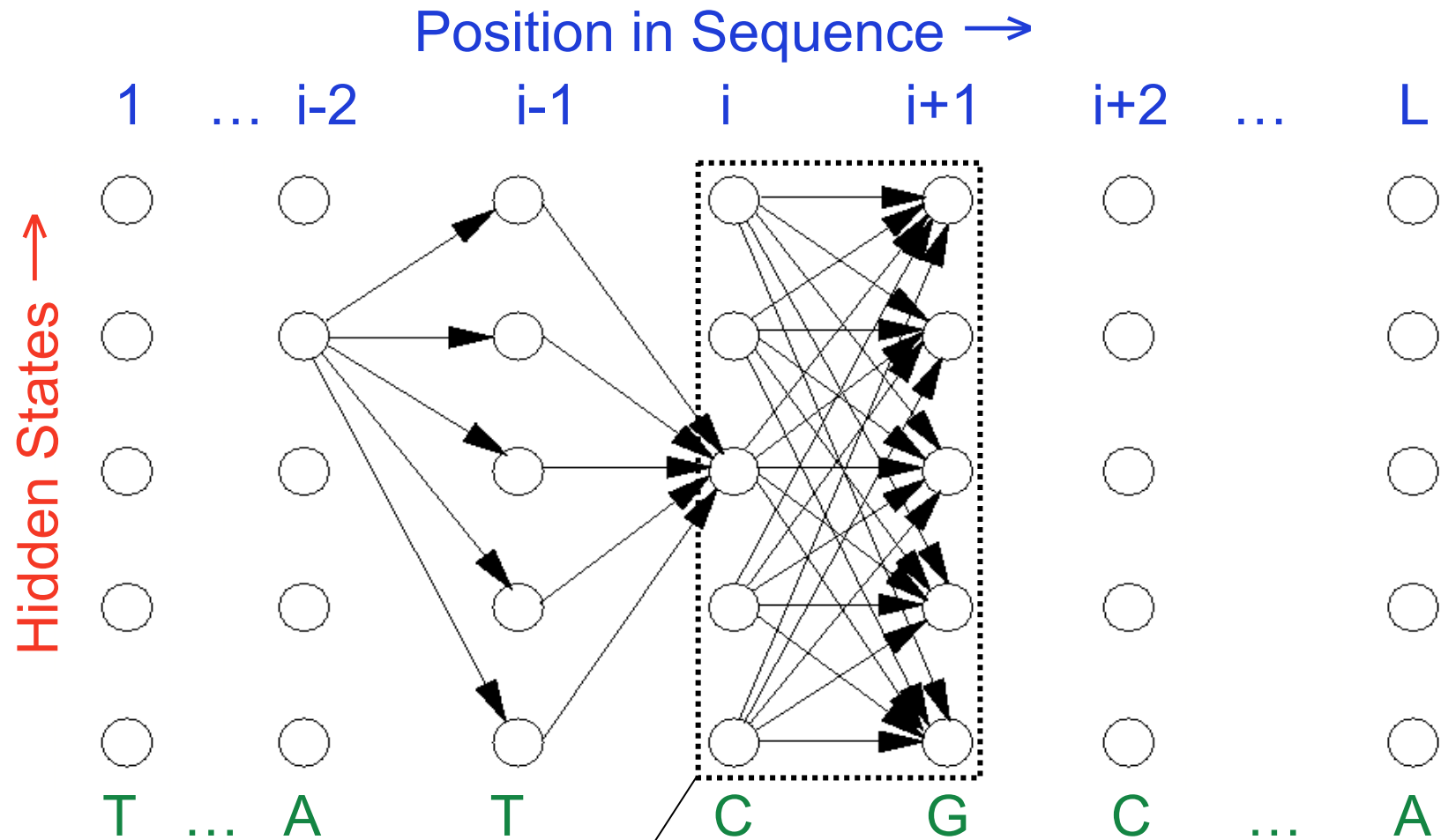
March 13, 2014

RNA Secondary Structure - Biological Functions & Prediction

Hidden Markov Models of Genomic & Protein Features

- Hidden Markov Model terminology
- Viterbi algorithm
- Examples
 - CpG Island HMM
 - TMHMM (transmembrane helices)

“Trellis” Diagram for Viterbi Algorithm



“Initiation probabilities” π_j

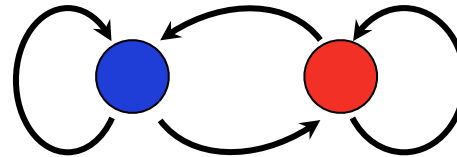
Rabiner notation

CpG Island HMM

$P_g = 0.99, P_i = 0.01$

$P_{gg} = 0.99999, P_{ig} = 0.001$

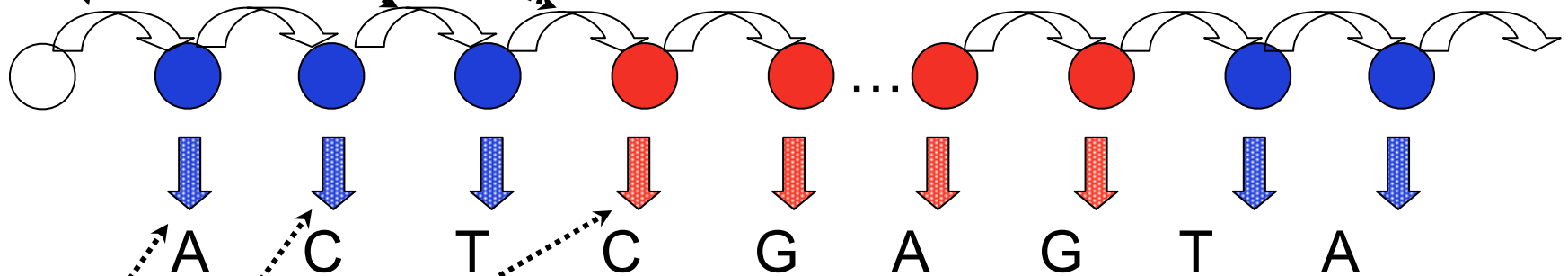
Genome



$P_{ii} = 0.999$

“Transition probabilities” a_{ij}

$P_{gi} = 0.00001$ **Island**



“Emission Probabilities” $b_j(k)$

	<u>C</u>	<u>G</u>	<u>A</u>	<u>T</u>
CpG Island:	0.3	0.3	0.2	0.2
Genome:	0.2	0.2	0.3	0.3

More Viterbi Examples

What is the optimal parse of the sequence for the CpG island HMM defined previously?

- $(ACGT)_{10000}$
- $A_{1000}C_{80}T_{1000}C_{20}A_{1000}G_{60}T_{1000}$

Powers of 1.5:

N =	20	40	60	80
$(1.5)^N =$	3×10^3	1×10^7	3×10^{10}	1×10^{14}

Real World HMMs

“Profile HMM” with insertions/deletions

A. Sequence alignment

N	•	F	L	S
N	•	F	L	S
N	K	Y	L	T
Q	•	W	-	T

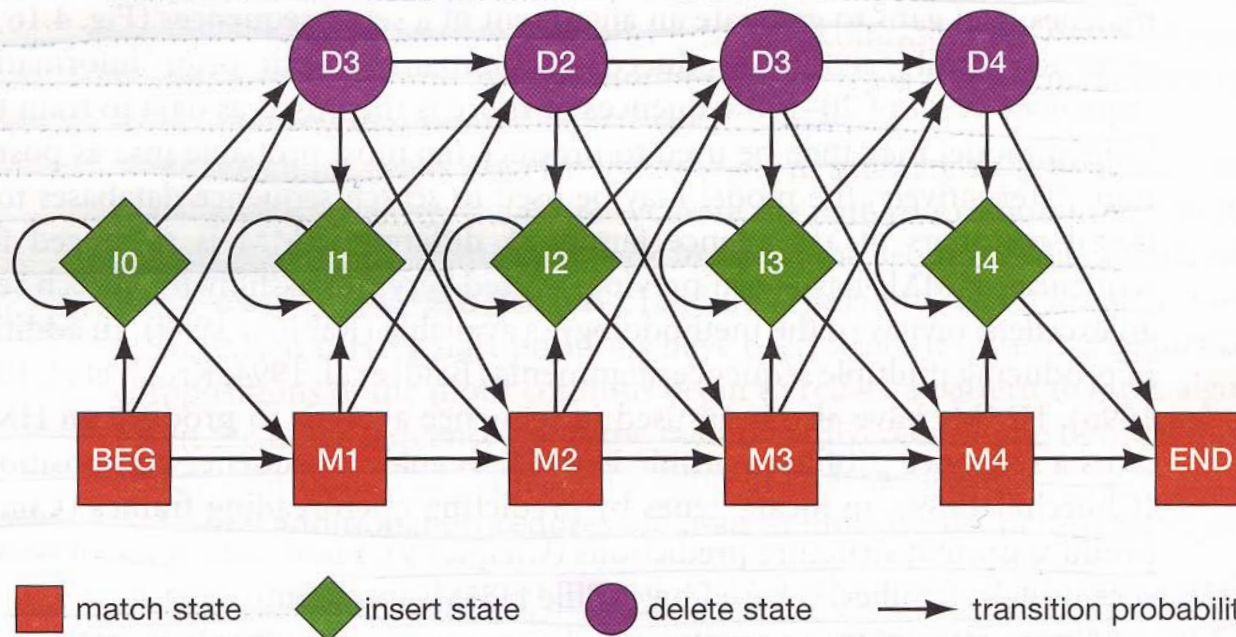
RED POSITION REPRESENTS ALIGNMENT IN COLUMN

GREEN POSITION REPRESENTS INSERT IN COLUMN

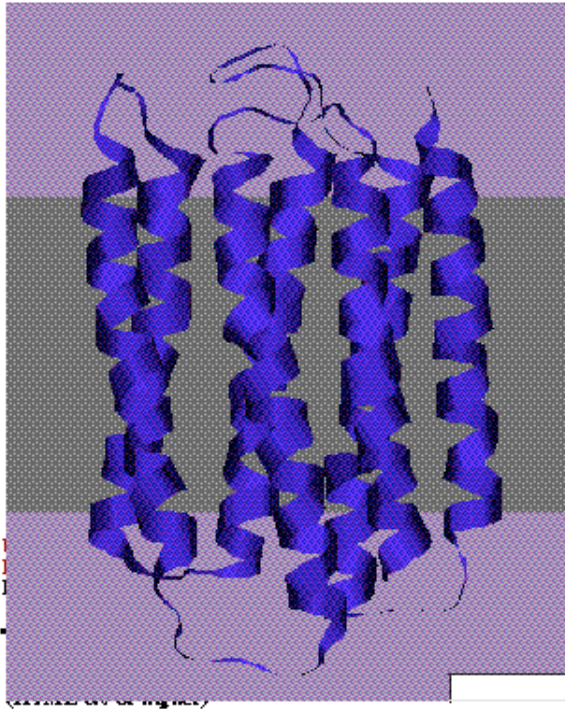
PURPLE POSITION REPRESENTS DELETE IN COLUMN

Of course, can have insertion/
deletion states for HMM models of
DNA/RNA as well

B. Hidden Markov model for sequence alignment



© Cold Spring Harbor Laboratory Press. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.



TMHMM (v. 2.0)

Prediction of transmembrane helices in proteins

[Help/Information](#) (updated Sept 13, 2001)

One of the [World Wide Web Prediction Servers](#)
from the [Center for Biological Sequence Analysis](#)

limit each submission to at most 4000 proteins.
each large submission.

OR by pasting sequence(s) in [FASTA](#) format:

x

Output format: ☒ Extensive, with graphics
☐ Extensive, no graphics
☐ One line per protein

Other options: ☐ Use old model (version 1)

Submit

Clear

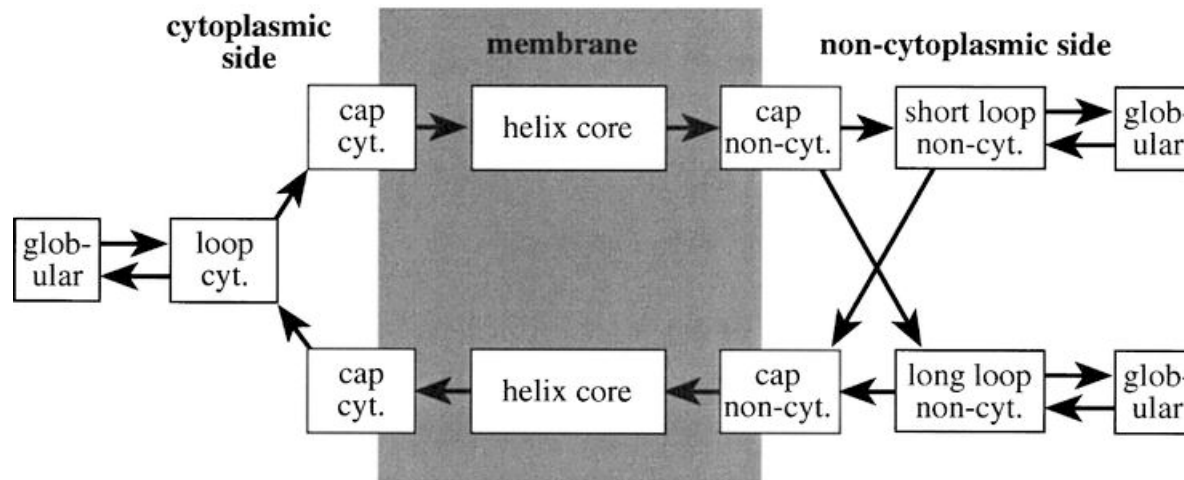
Correctly predicts ~97%
of transmembrane helices
according to authors

A. Krogh et al. *J. Mol. Biol.* 2001

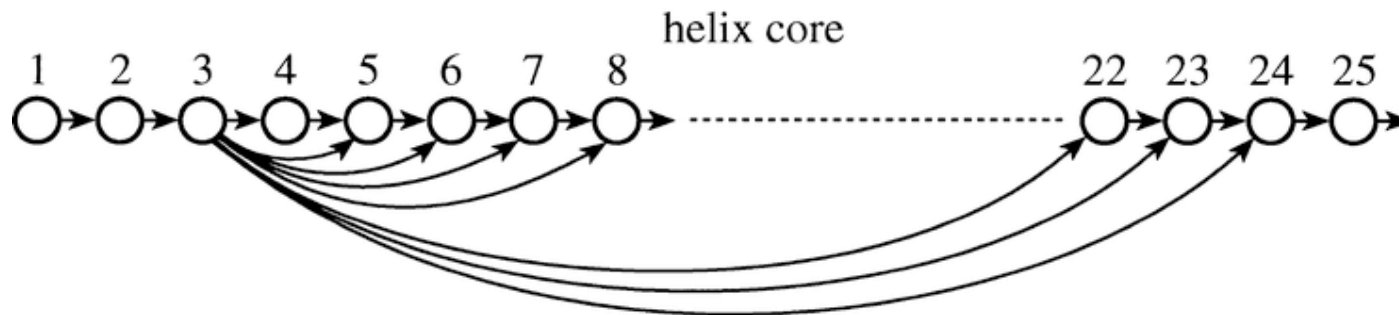
© [Center for Biological Sequence Analysis](#). All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

Architecture of TMHMM

(a)



(c)

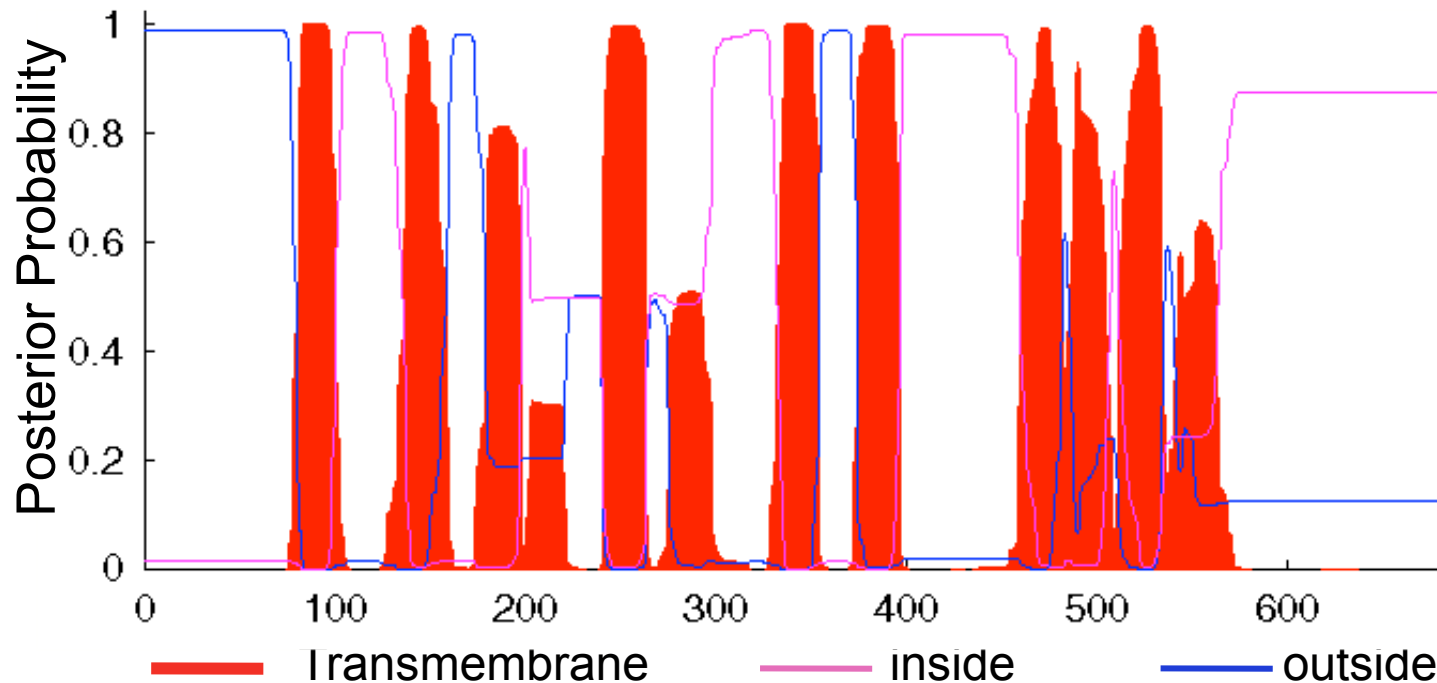
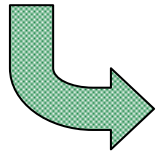


Courtesy of Biomedical Informatics Publishing Group. Used with permission.

Source: Chaturvedi, Navaneet, Sudhanshu Shanker, et al. "Hidden Markov Model for the Prediction of Transmembrane Proteins using MATLAB." *Bioinformatics* 7, no. 8 (2011): 418.

TMHMM Output for Mouse Chloride Channel CLC6

Optimal
Parse



RNA Secondary Structure

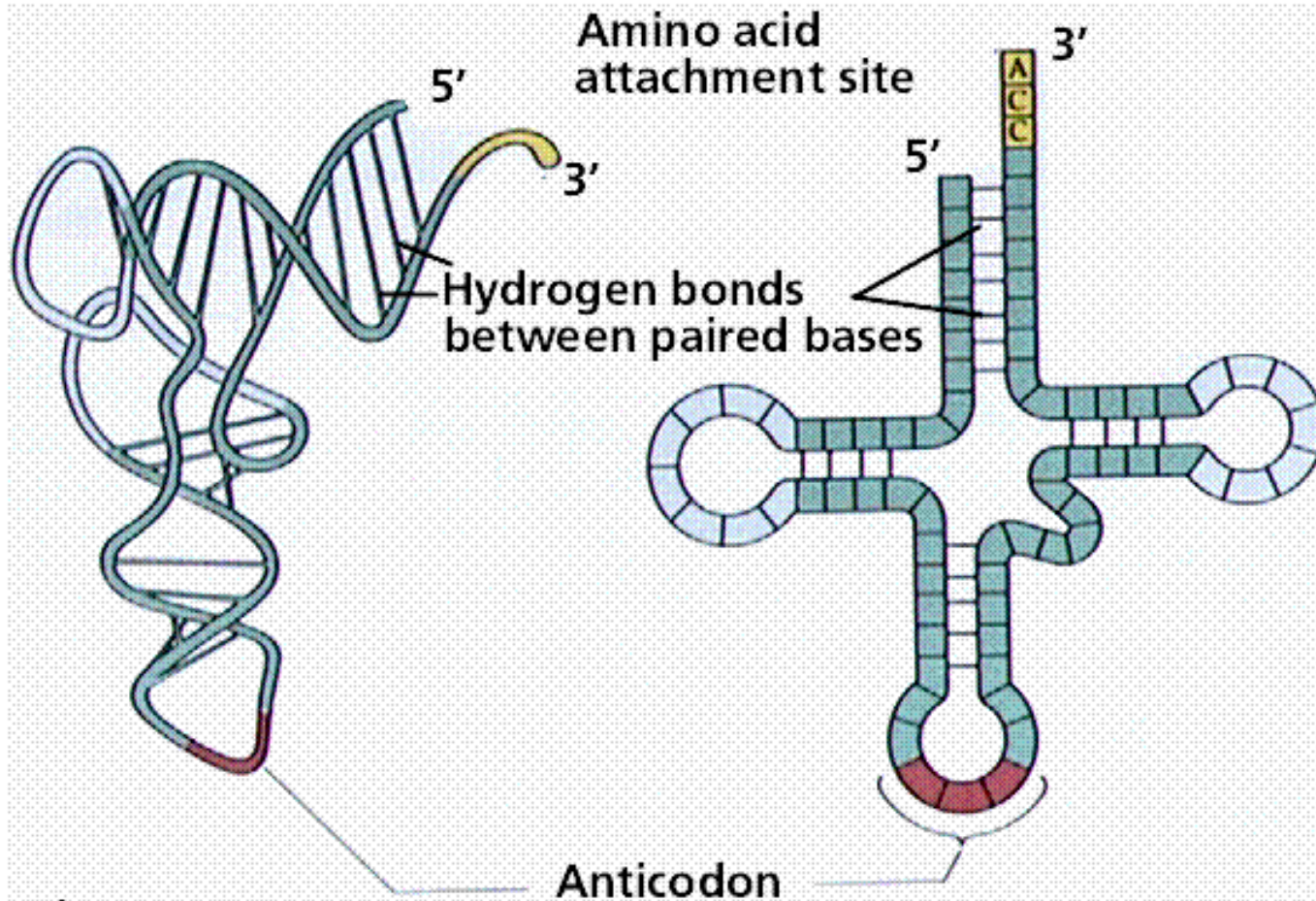
- Biological examples of RNA structure
- Predicting 2^o structure by covariation
- Predicting 2^o structure by energy minimization

Readings

NBT Primer on RNA folding, Z&B Ch. 11.9

RNA Secondary and Tertiary Structure

Example: tRNA



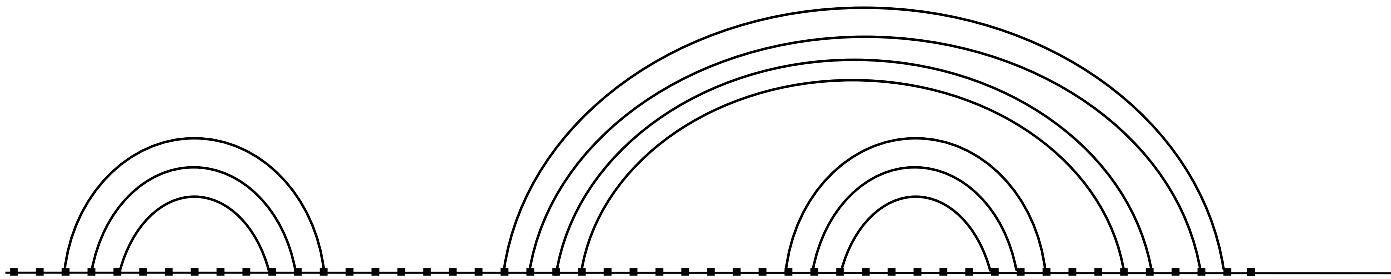
© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

RNA Secondary Structure Notation

Parentheses notation

..(((.....))).....((((.....))).))...

Arc ('rainbow') notation

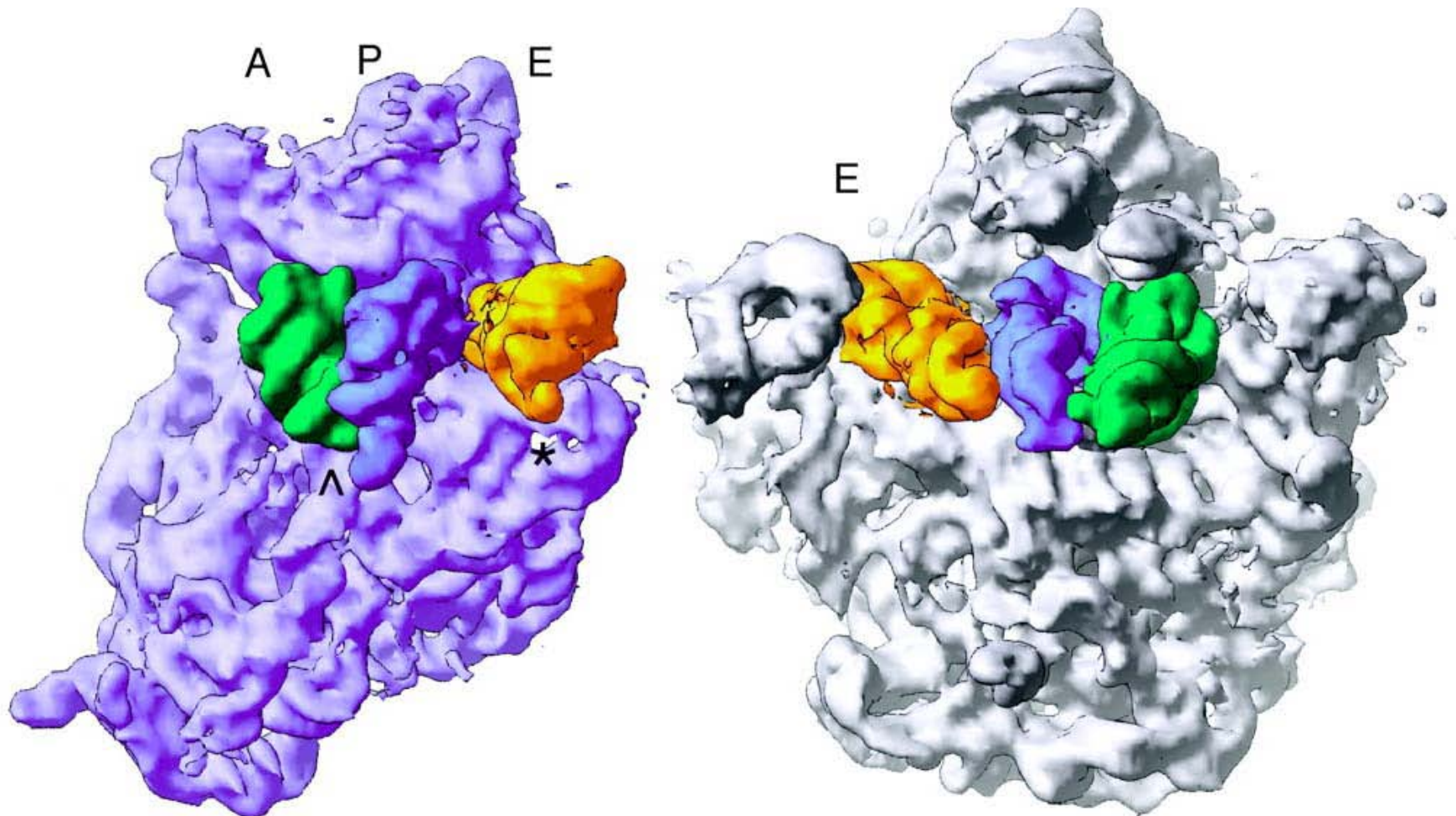


What do these structures look like?

What is the difference between these two structures?

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.
Source: Cate, Jamie H., Marat M. Yusupov, et al. "X-ray Crystal Structures of 70S Ribosome Functional Complexes." *Science* 285, no. 5436 (1999): 2095-104.

Ribosome at 7 Å with tRNAs



© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.
Source: Cate, Jamie H., Marat M. Yusupov, et al. "X-ray Crystal Structures of 70S Ribosome Functional Complexes." *Science* 285, no. 5436 (1999): 2095-104.

Slide courtesy of Rachel Green

Can build useful structures out of RNA

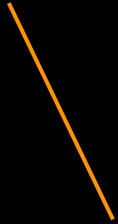
The exit channel for the growing polypeptide

Slide courtesy of Rachel Green

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.
Source: Ban, Nenad, Poul Nissen, et al. "The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution." *Science* 289, no. 5481 (2000): 905-20.

RNA/protein distribution on the 50S ribosome

linguini = protein



fettucini = RNA



© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.
Source: Ban, Nenad, Poul Nissen, et al. "The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution." *Science* 289, no. 5481 (2000): 905-20.

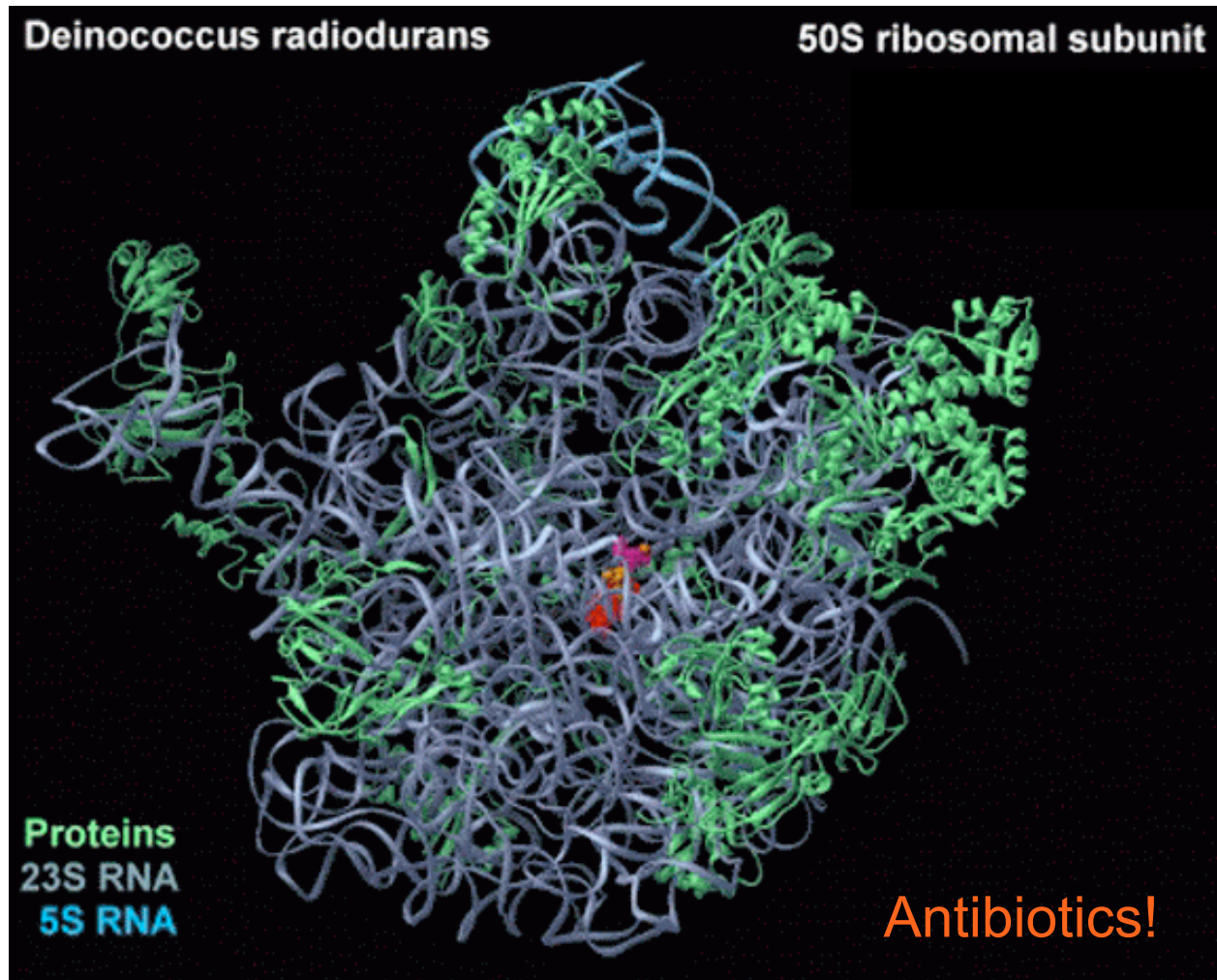
The ribosome is a ribozyme

Nearest proteins
and distances to
active site (Å)

Slide courtesy of Rachel Green

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.
Source: Nissen, Poul, Jeffrey Hansen, et al. "The Structural Basis of Ribosome Activity in Peptide Bond Synthesis." *Science* 289, no. 5481 (2000): 920-30.

What are the practical applications of knowing the ribosome structure?

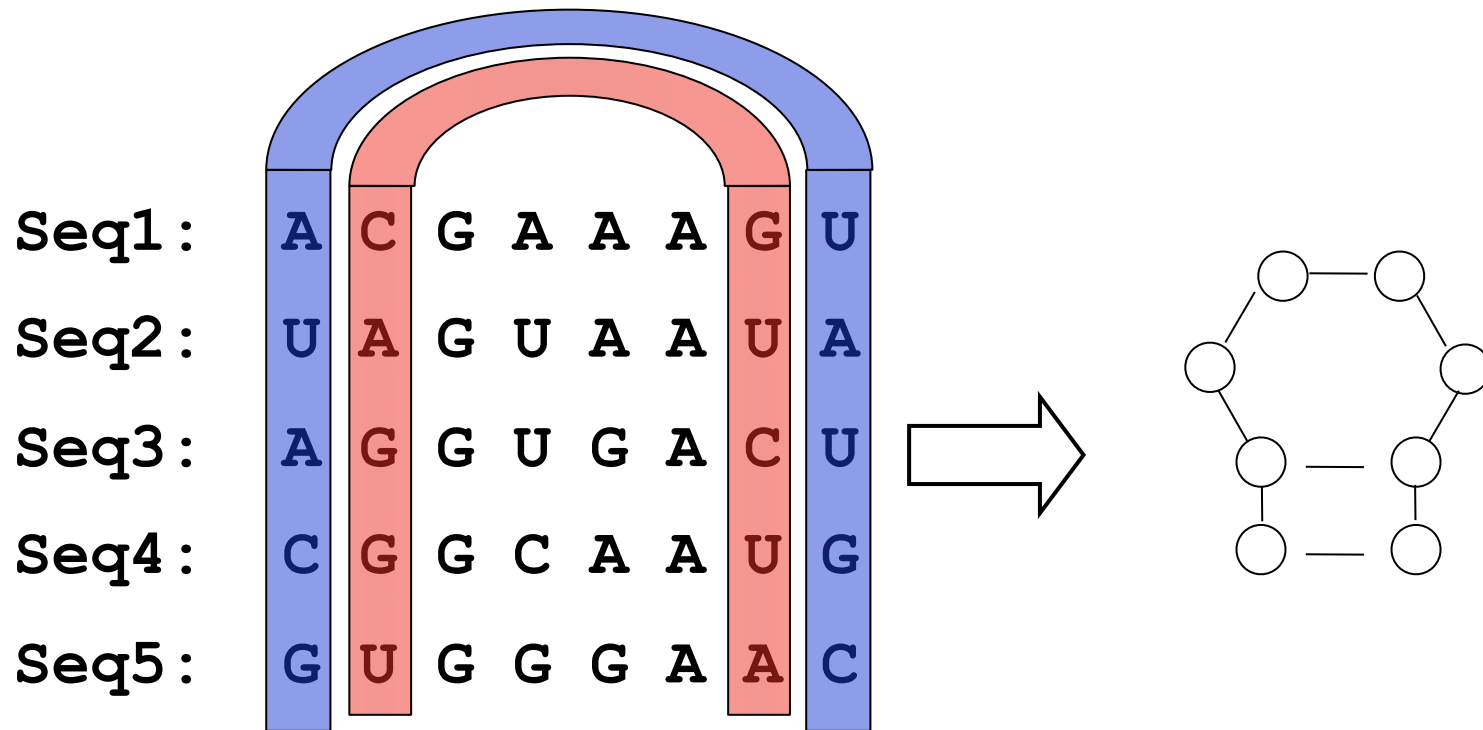


© sources unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

ncRNAs: Challenges for Computational Biology

- Prediction of ncRNA structure
- Identification of ncRNA genes
- Prediction of ncRNA functions

RNA 2^o structure by covariation / compensatory changes



Mutual information statistic for pair of columns in a multiple alignment

$$M_{ij} = \sum_{x,y} f_{x,y}^{(i,j)} \log_2 \frac{f_{x,y}^{(i,j)}}{f_x^{(i)} f_y^{(j)}}$$

$f_{x,y}^{(i,j)}$ = fraction of seqs w/ nt. x in col. i , nt. y in col. j

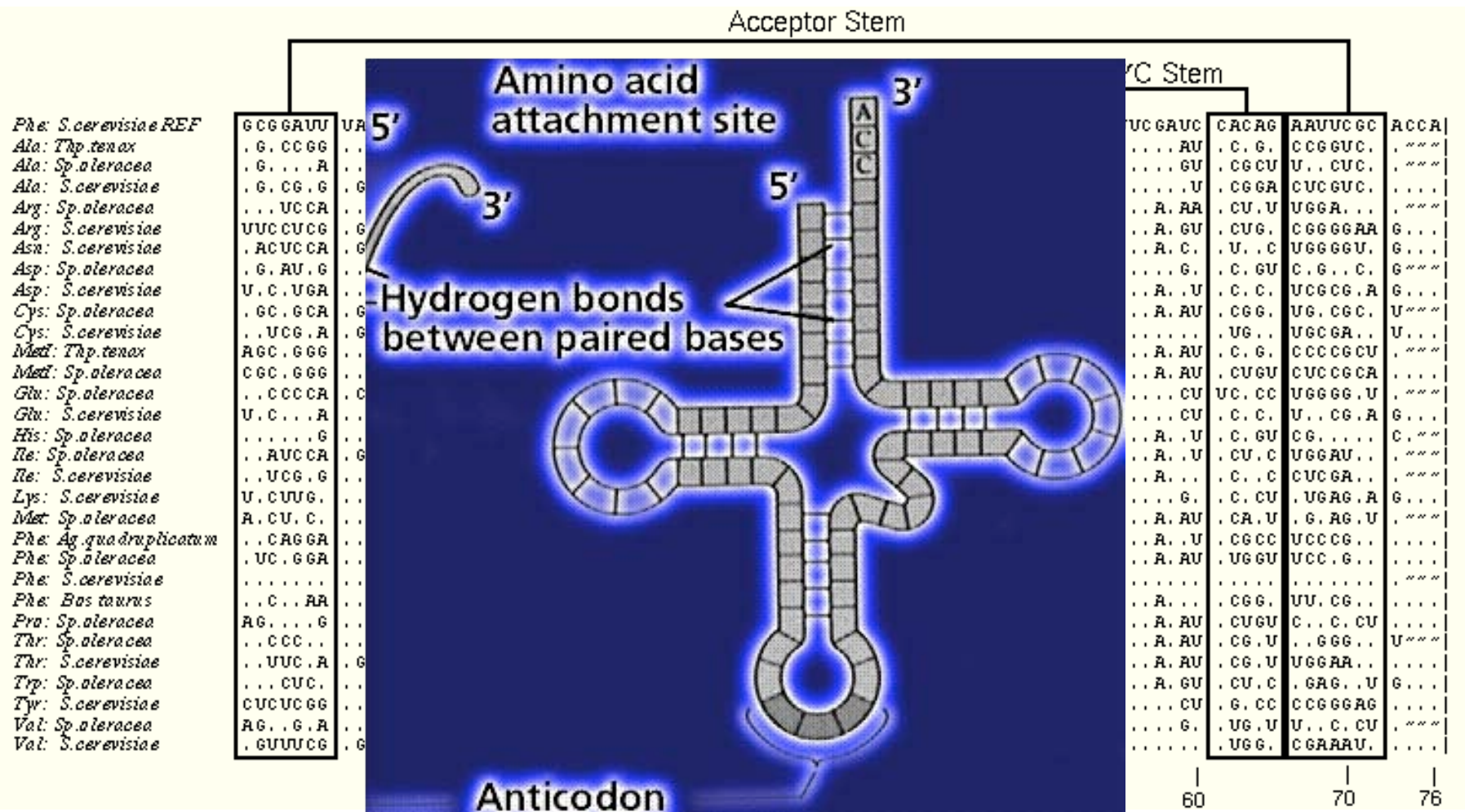
$f_x^{(i)}$ = fraction of seqs w/ nt. x in col. i

sum over $x, y = A, C, G, U$

M_{ij} is maximal (2 bits) if x and y individually appear at random (A,C,G,U equally likely), but perfectly covary (e.g., always complementary)

Could use other measure of dependence (e.g., chi-square statistic)

Inferring 2° structure from covariation



© sources unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

What is needed for accurate inference of RNA secondary structure by covariation?

- Secondary structure more highly conserved than primary sequence
- Sufficient divergence between homologs for many variations to have occurred, but not so much that can't be aligned
- Sufficient number of homologs sequenced

Classes of Non-coding RNAs

- tRNAs
 - rRNAs
 - UTRs
 - snRNAs
 - snoRNAs
 - prok. terminators
 - ...
- RNaseP
 - SRP RNA
 - tmRNA
 - miRNAs
 - lncRNAs
 - riboswitches
 - ...

Energy Minimization Approach

$$\Delta G_{\text{folding}} = G_{\text{unfolded}} - G_{\text{folded}}$$

There are typically many possible folded states

- assumption that minimum energy state(s) will be occupied

$$\Delta G = \Delta H - T\Delta S$$

Enthalpy favors folding

Entropy favors unfolding

What environmental variables affect RNA folding?

How Do Energy Minimization Algorithms Work?

Consider Simple Model: Base Pair Maximization

Scoring System:

+1 for base pair (C:G, A:U)

0 for anything else

Maximizing score equivalent to minimizing folding free energy for a model which assigns same enthalpy to all allowed base pairs (and ignores details such as base stacking, loops, entropy)

Nussinov algorithm: recursive maximization of base pairing

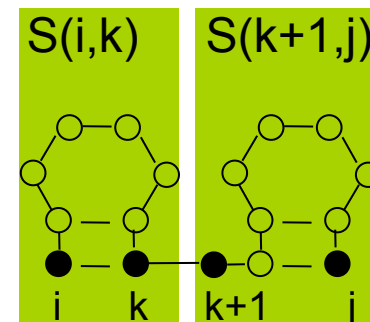
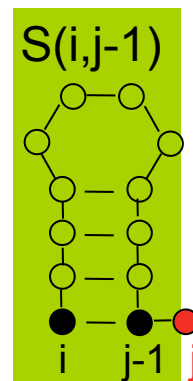
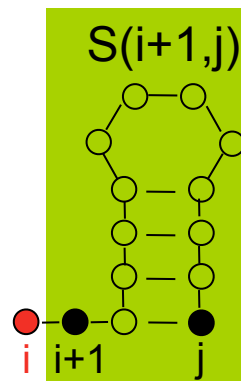
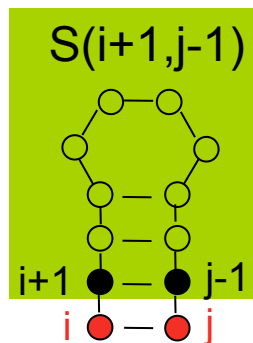
Recursive Maximization of Base Pairing

Given an RNA sequence of length N

Define $S(i,j)$ to be the score of the best structure for the subsequence (i, j)

Notice that $S(i,j)$ can be defined recursively in terms of optimal scores of smaller subsequences of the interval (i,j)

There are four possible ways that the score of the optimal structure on (i,j) can relate to scores of optimal structures of nested subsequences:



1. i, j pair 2. i unpaired 3. j unpaired

4. bifurcation

Courtesy of Macmillan Publishers Limited. Used with permission.

Source: Eddy, Sean R. "How do RNA Folding Algorithms Work?" *Nature Biotechnology* 22, no. 11 (2004): 1457-8.

Eddy, Nature Biotech. 2004

Base Pair Maximization Algorithm

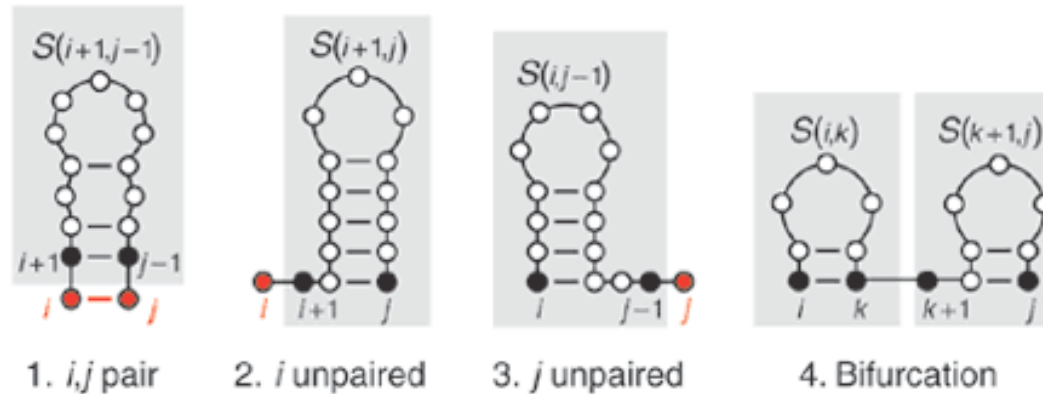
$S(i,j)$ = score of the optimal structure for the subsequence (i, j)

$$S(i,j) = \max \left\{ \begin{array}{ll} S(i+1,j-1) + 1 & \text{(if } i,j \text{ base pair)} \\ S(i+1,j) & \text{(} i \text{ is unpaired)} \\ S(i,j-1) & \text{(} j \text{ is unpaired)} \\ \max_{i < k < j} S(i,k) + S(k+1,j) & \text{(bifurcation)} \end{array} \right.$$

- 1) Initialize an $N \times N$ matrix S with $S(i,i) = S(i,i-1) = 0$
- 2) Fill in $S(i,j)$ matrix recursively from the diagonal up and to the right (keep track of which choice was made at each step)
- 3) Trace back from $S(1,N)$ (upper right corner of matrix) to diagonal to determine optimal structure

Dynamic Programming for Base Pair Maximization

Recursive definition of the best score for a sub-sequence i,j looks at four possibilities:



Dynamic programming algorithm for all sub-sequences i, j , from smallest to largest:

		$j \rightarrow$									
		G	G	G	A	A	A	U	C	C	
$i \downarrow$	G	0									
	G	0	0								
	G		0	0							
	A			0	0						
	A				0	0					
	A					0	0				
	U						0	0			
	C							0	0		
C								0	0		

Initialization;

Courtesy of Macmillan Publishers Limited. Used with permission.

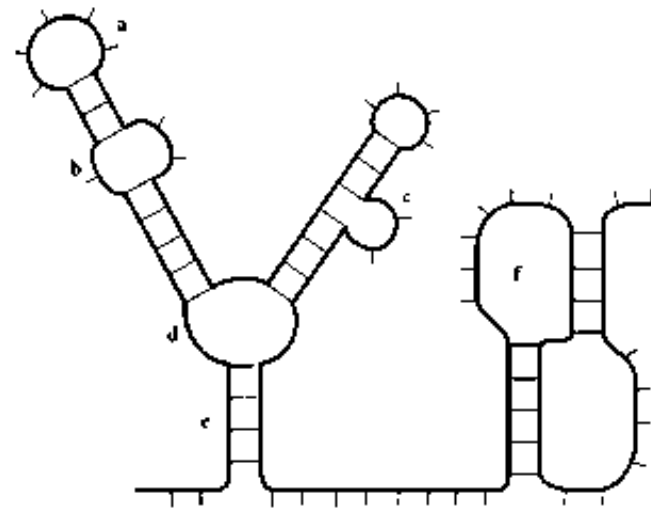
Source: Eddy, Sean R. "How do RNA Folding Algorithms Work?" *Nature Biotechnology* 22, no. 11 (2004): 1457-8.

Eddy, Nature Biotech. 2004

Base Pair Maximization Algorithm Issues

- What is computational complexity of algorithm?
(for sequence of length N)

Answer: Memory - $O(N^2)$ Time - $O(N^3)$

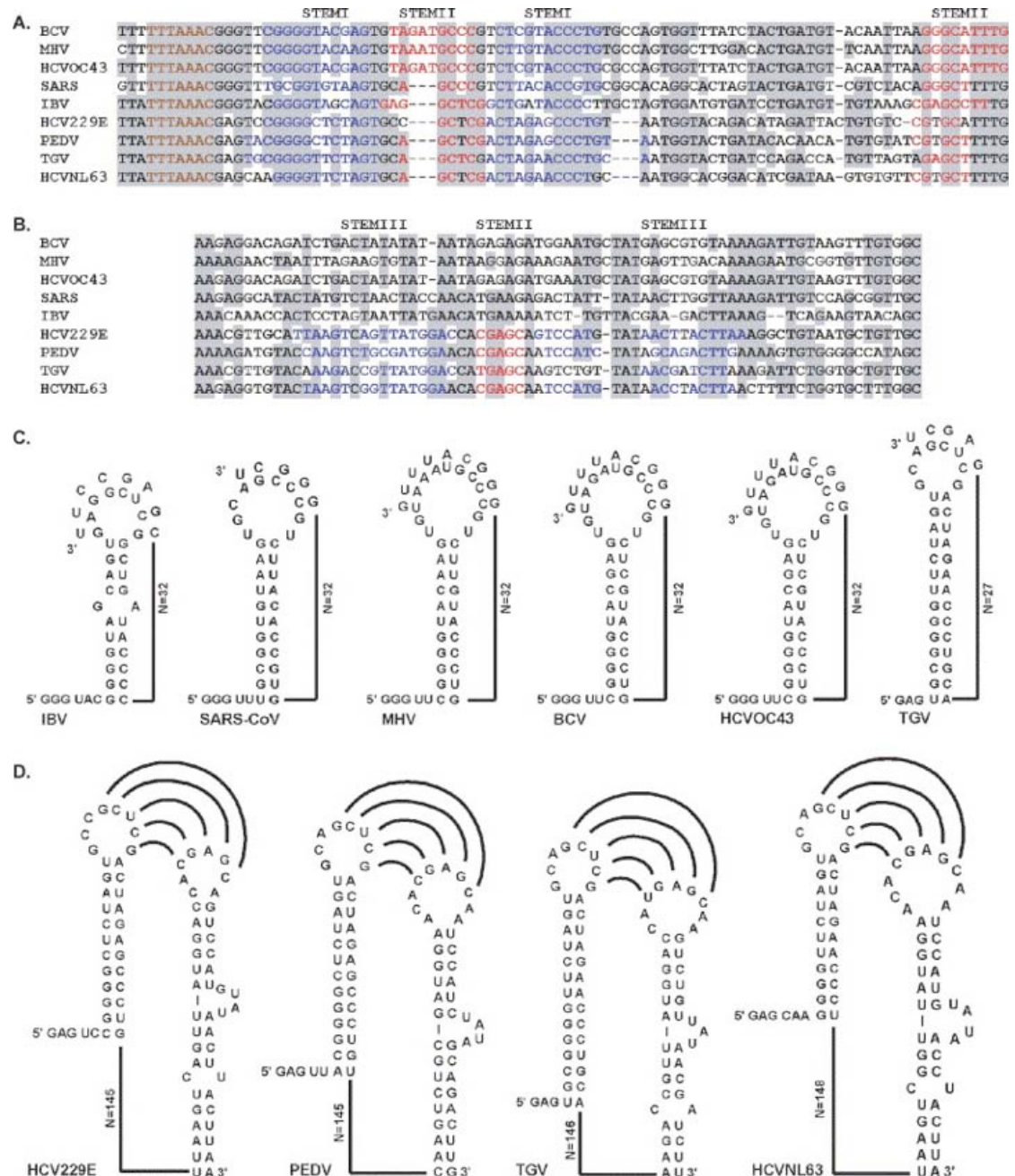


- Can it handle pseudoknots?

© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

Answer: No. Pseudoknots invalidate recursion for $S(i,j)$

Viral Pseudoknots and “Kissing loops”



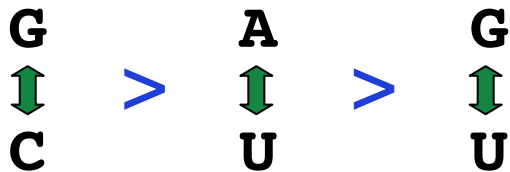
Baranov et al. Virology 2005

Courtesy of Elsevier, Inc., <http://www.sciencedirect.com>. Used with permission.
Source: Baranov, Pavel V., Clark M. Henderson, et al. "Programmed Ribosomal Frameshifting in Decoding the SARS-CoV Genome." *Virology* 332, no. 2 (2005): 498-510.

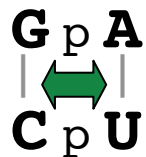
RNA Energetics I

Free energy contributions to helix formation come from:

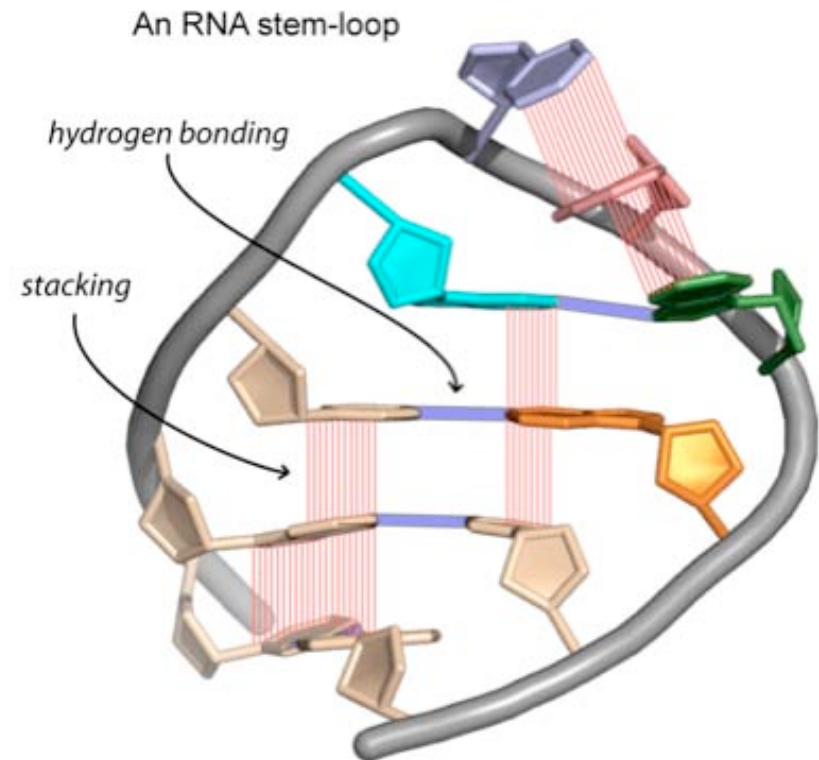
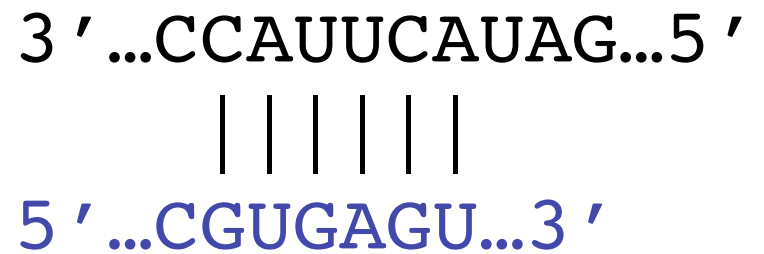
- base pairing:



- base stacking:



Base stacking contributes more to free energy than base pairing

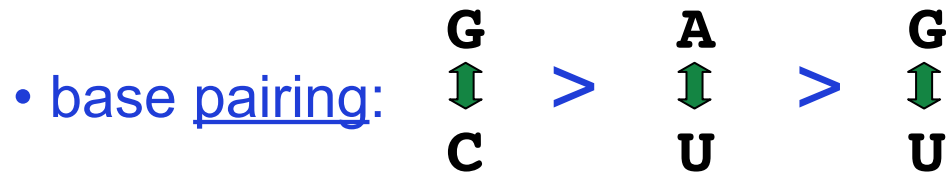


© American Chemical Society. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

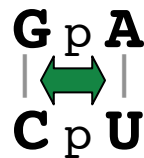
Source: Mohan, Srividya, Chiaolong Hsiao, et al. "RNA Tetraloop Folding Reveals Tension Between Backbone Restraints and Molecular Interactions." *Journal of the American Chemical Society* 132, no. 36 (2010): 12679-89.

RNA Energetics I

Free energy contributions from:



• base stacking:



are combined in
Doug Turner's Energy Rules:

Matrix for each X,Y stacking on
each possibly base pair
or free end

		5' --> 3'		
		UX		
		AY		
		3' <-- 5'		
		<u>X</u>		
<u>Y</u>	A	C	G	U
A	.	.	.	-1.30
C	.	.	-2.40	.
G	.	-2.10	.	-1.00
U	-0.90	.	-1.30	.

RNA Energetics II

Other Contributions to Folding Free Energy

- Hairpin loop destabilizing energies
 - a function of loop length
- Interior and bulge loop destabilizing energies
 - a function of loop length
- Terminal mismatch and base pair energies

RNA Energetics III

Folding by Energy Minimization

A more complex dynamic programming algorithm is used - similar in spirit to the Nussinov base pair maximization algorithm

Gives:

- minimum energy fold
- suboptimal folds (e.g., five lowest ΔG folds)
- probabilities of particular base pairs
- full partition function

Accuracy: ~70% of base pairs correct

Links & References

The Mfold web server:

<http://mfold.rna.albany.edu/?q=mfold/rna-folding-form>

The Vienna RNAfold package (free for download)

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

RNA folding references:

M. Zuker, et al. In *RNA Biochemistry and Biotechnology* (1999)

D.H. Mathews et al. *J. Mol. Biol.* **288**, 911-940 (1999)

Vienna package by Ivo Hofacker

RNA Secondary Structure Prediction by Energy Minimization Summary

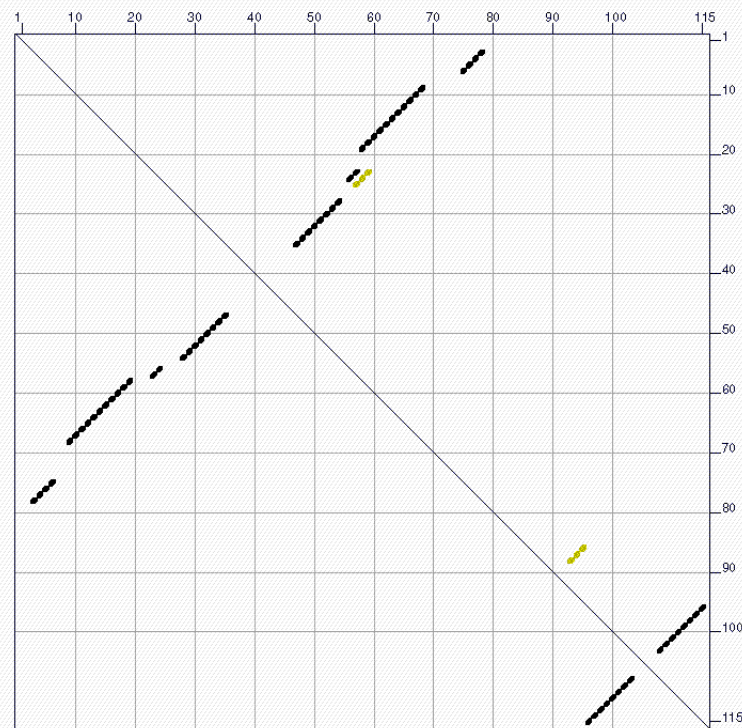
- Assumes folding energy decomposable into independent contributions of small units of structure
- Algorithms are guaranteed to find minimal free energy structure defined by the model
- In practice, algorithms predict ~70% of bp correct
- Errors result from
 - imprecision of the model/parameters
 - differences between *in vitro* and *in vivo* conditions
 - *in vivo* structure may not always have minimum free energy

Sample Mfold Output (Human U5 snRNA)

boxplot_rngby D. Stewart and M. Zuker
© 2002 Washington University

Fold of 02Apr10-15-09-05 at 37 C.

δG in Plot File = 1.6 kcal/mole

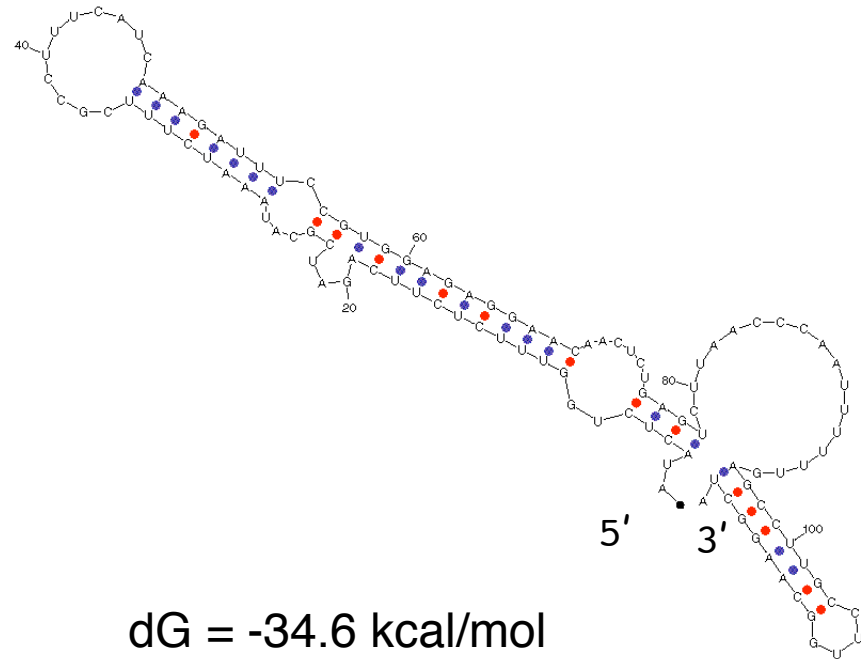


Lower Triangle: Optimal Energy
Upper Triangle Base Pairs Plotted: 39

Optimal Energy = -34.6 kcal/mole
-34.6 < Energy <= -34.1 kcal/mole
-34.1 < Energy <= -33.5 kcal/mole
-33.5 < Energy <= -33.0 kcal/mole

© Washington University. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

Energy dot plot



$dG = -34.6$ kcal/mol

Minimum free energy structure

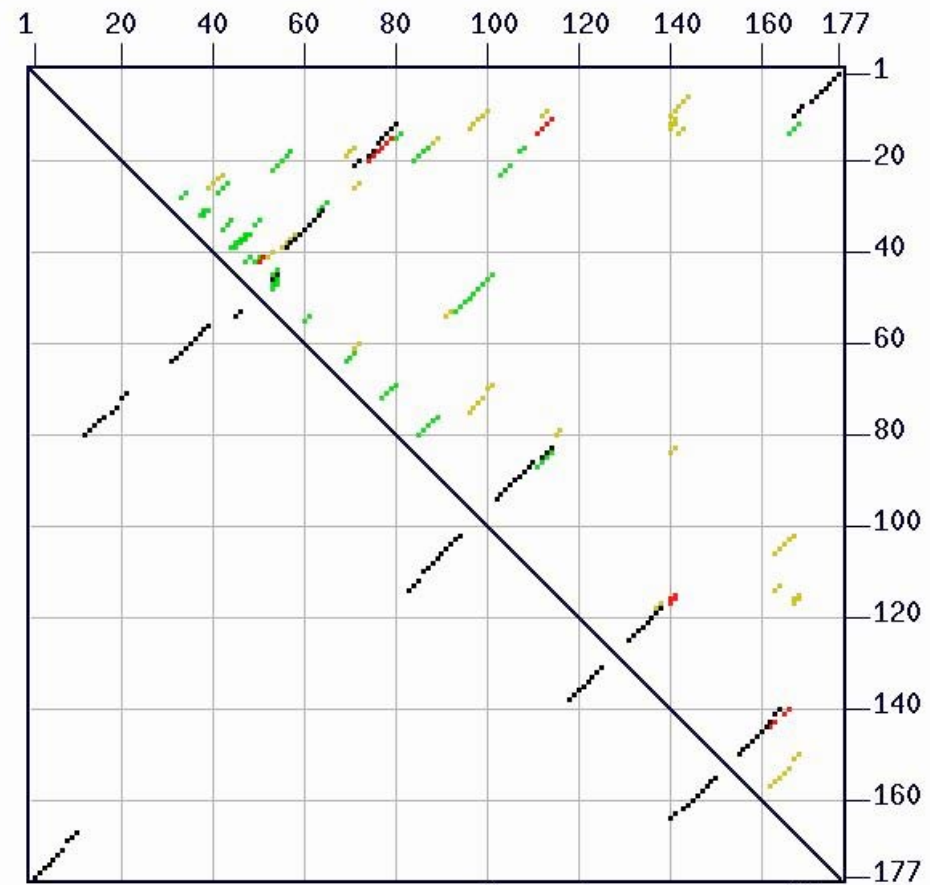
© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

Energy dot plot for a lysine riboswitch

boxplot_ng by D. Stewart and M. Zuker
© 2005 Washington University

Fold of 05Mar30-14-22-34 at 37 C.

deltaG in Plot File = 4.5 kcal/mole



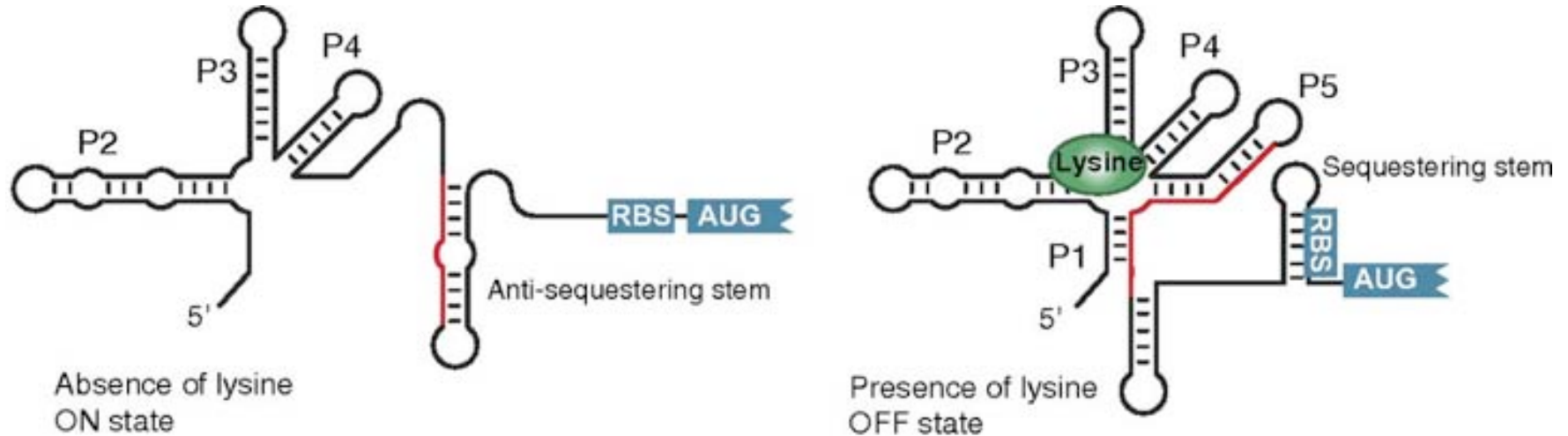
Lower Triangle Shows
Optimal Energy

Upper Triangle
Basepairs Plotted: 230

Optimal energy: -90.6
 $-90.6 < \text{energy} \leq -89.1$
 $-89.1 < \text{energy} \leq -87.6$
 $-87.6 < \text{energy} \leq -86.1$

© Washington University. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

Function of the lysine riboswitch



© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

Lysine interacts with the junctional core of the riboswitch and is specifically recognized through shape-complementarity within the elongated binding pocket and through several direct and K^+ -mediated hydrogen bonds to its charged ends.

Controls expression of enzymes involved in biosynthesis and transport of lysine

Serganov et al. Nature 2008. Caron et al PNAS 2012

MIT OpenCourseWare
<http://ocw.mit.edu>

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology
Spring 2014

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.