7.91 / 7.36 / BE.490

Lecture #6

Mar. 11, 2004

Predicting RNA Secondary Structure

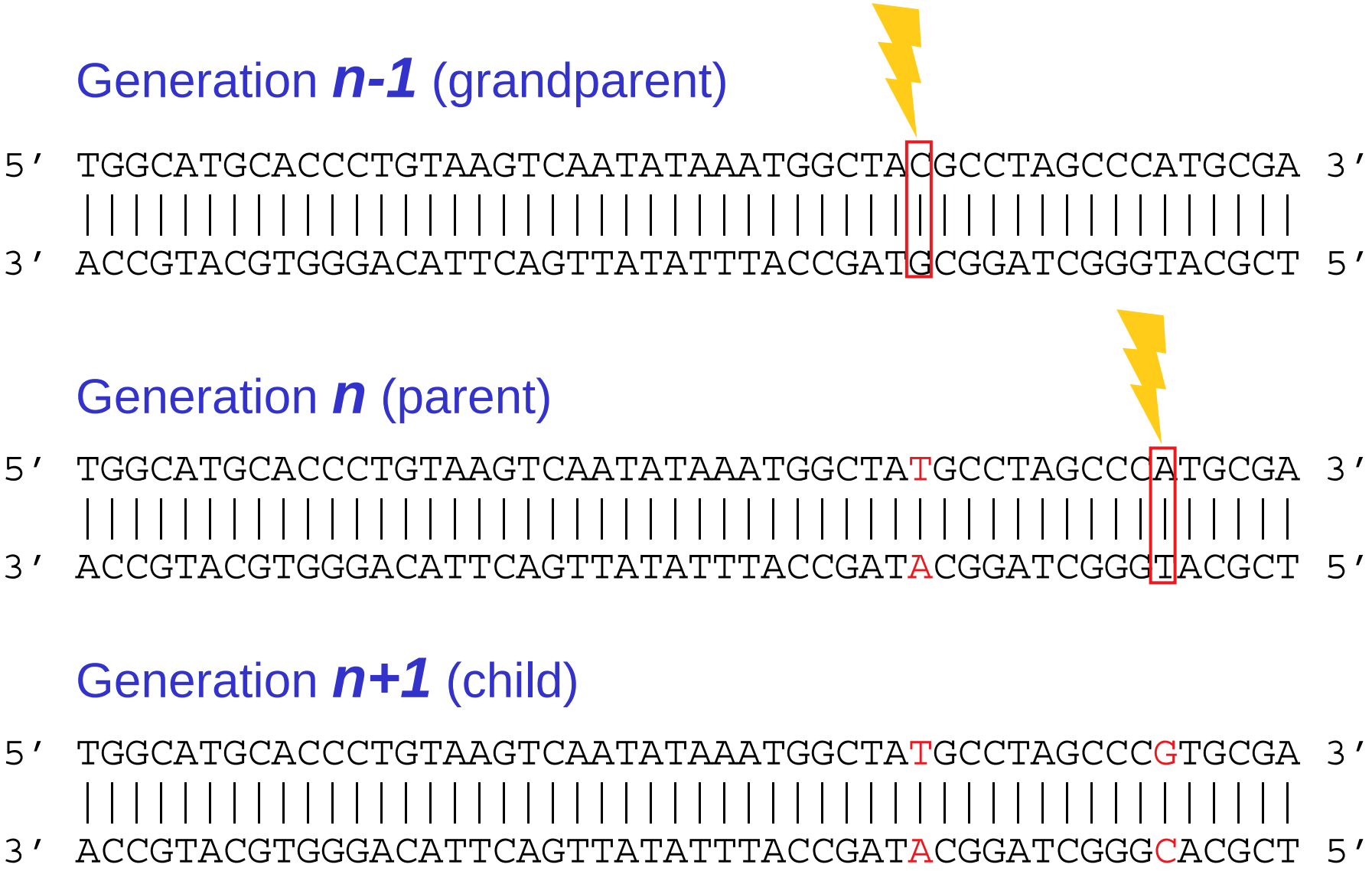
Chris Burge

Review of Markov Models & DNA Evolution

- CpG Island HMM
- The Viterbi Algorithm
- Real World HMMs
- Markov Models for DNA Evolution

Ch. 4 of Mount

DNA Sequence Evolution



What is a *Markov* Model (aka *Markov* Chain)?

Classical Definition

A discrete stochastic process $X_1, X_2, X_3, \ldots$ which has the Markov property:

$$P(X_{n+1} = j \mid X_1=x_1, X_2=x_2, \ldots X_n=x_n) = P(X_{n+1} = j \mid X_n=x_n)$$

(for all x_i, all j, all n)

In words:

A random process which has the property that the future (next state) is conditionally independent of the past given the present (current state)

Markov - a Russian mathematician, ca. 1922

DNA Sequence Evolution is a Markov Process

No selection case

S_n = base at generation n

$$P_{ij} = P(S_{n+1} = j \mid S_n = i)$$

$$P = \begin{pmatrix} P_{AA} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & P_{CC} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & P_{GG} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & P_{TT} \end{pmatrix}$$

$\vec{q}^{\,n} = (q_A, q_C, q_G, q_T)$ = vector of prob's of bases at gen. n

Handy relations: $\vec{q}^{\,n+1} = \vec{q}^{\,n} P$ $\quad \vec{q}^{\,n+k} = \vec{q}^{\,n} P^k$

Limit Theorem for Markov Chains

S_n = base at generation n $\quad P_{ij} = P(S_{n+1} = j \,|\, S_n = i)$

If $P_{ij} > 0$ for all i,j (and $\sum_j P_{ij} = 1$ for all i)

then there is a unique vector $\vec{r}$ such that

$\vec{r} = \vec{r}P$ and $\lim_{n\to\infty} \vec{q}P^n = \vec{r}$ (for any prob. vector $\vec{q}$)

$\vec{r}$ is called the “stationary” or “limiting” distribution of P

See Ch. 4, Taylor & Karlin, An Introduction to Stochastic Modeling, 1984 for details

Stationary Distribution Examples

2-letter alphabet: R = purine, Y = pyrimidine

Stationary distributions for:

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \qquad 0 < p < 1$$

$$P' = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \qquad 0 < p < 1,\ 0 < q < 1$$

How are mutation rates measured?

How does entropy change when a Markov transition matrix is applied?

If limiting distribution is uniform, then entropy increases

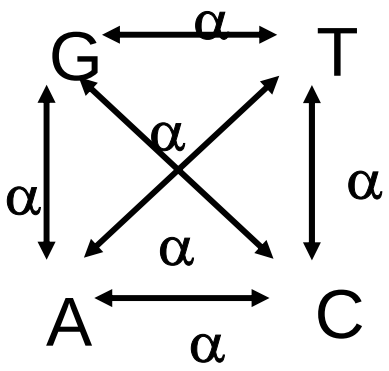
(analogous to 2nd Law of Thermodynamics)

However, this is not true in general (why not?)

How rapidly is the stationary distribution approached?

Jukes-Cantor Model

Courtesy of M. Yaffe



Assume each nucleotide equally likely
to change into any other nt,
with rate of change=α.
Overall rate of substitution = 3α
…so if G at t=0, at t=1, $P_{G(1)}=1-3\alpha$

and $P_{G(2)}=(1-3\alpha)P_{G(1)}+\alpha\,[1-P_{G(1)}]$

Expanding this gives $P_{G(t)}=1/4+\boxed{(3/4)e^{-4\alpha t}}$

Can show that this gives $K = -3/4\ \ln[1-(4/3)(p)]$

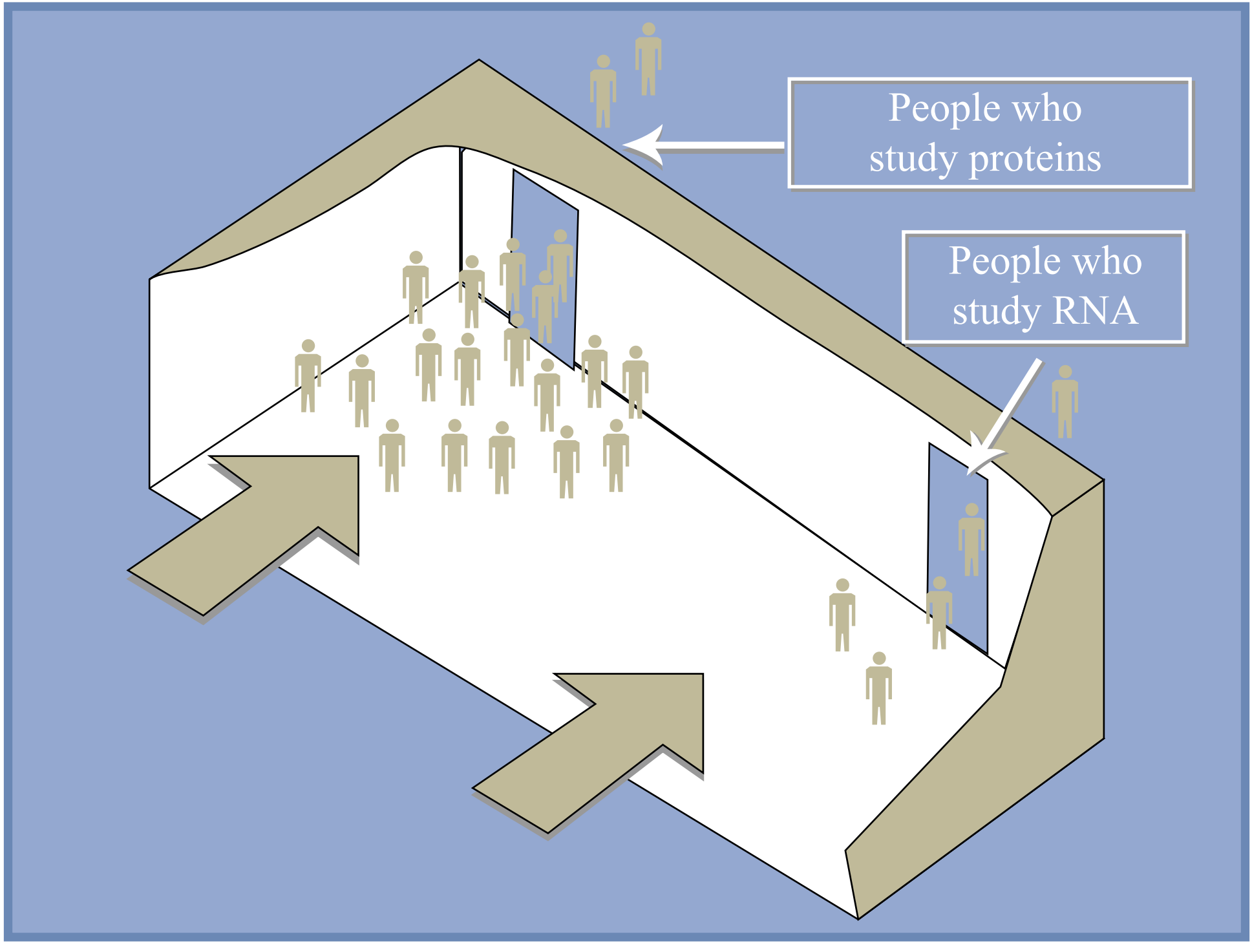
K = true number of substitutions that have occurred,
P = fraction of nt that differ by a simple count.

Captures general behaviour…

Predicting RNA Secondary Structure

- Review of RNA structure
  - Motivation: ribosome gallery, miRNAs/siRNAs
- Predicting 2° structure by energy minimization
- Predicting 2° structure by covariation
- Finding non-coding RNA genes

Read Mount, Ch. 5

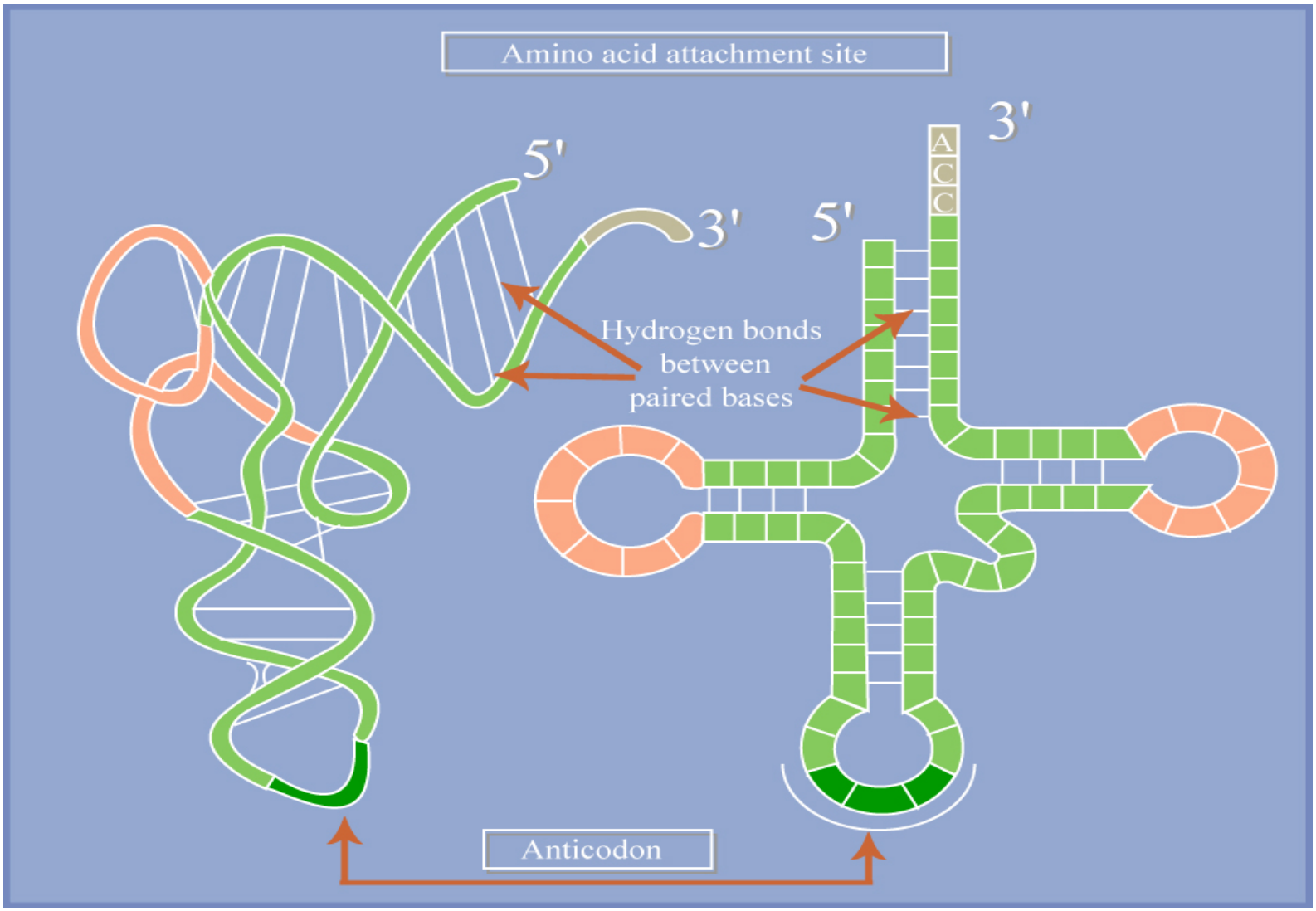

People who
study proteins
People who
study RNA

Types of Functional RNAs

- tRNAs
- rRNAs
- mRNAs
- snRNAs
- snoRNAs
- RNaseP
- SRP RNA
- tmRNA
- miRNAs
- siRNAs

The Good News:

Functional RNAs have secondary structure



Composition of the Ribosome

E. coli 70S ribosome - 2.6 x 10^6 daltons

30S subunit - 0.9 x 10^6 daltons

16S rRNA (1542 nts)

21 proteins

50S subunit - 1.7 x 10^6 daltons

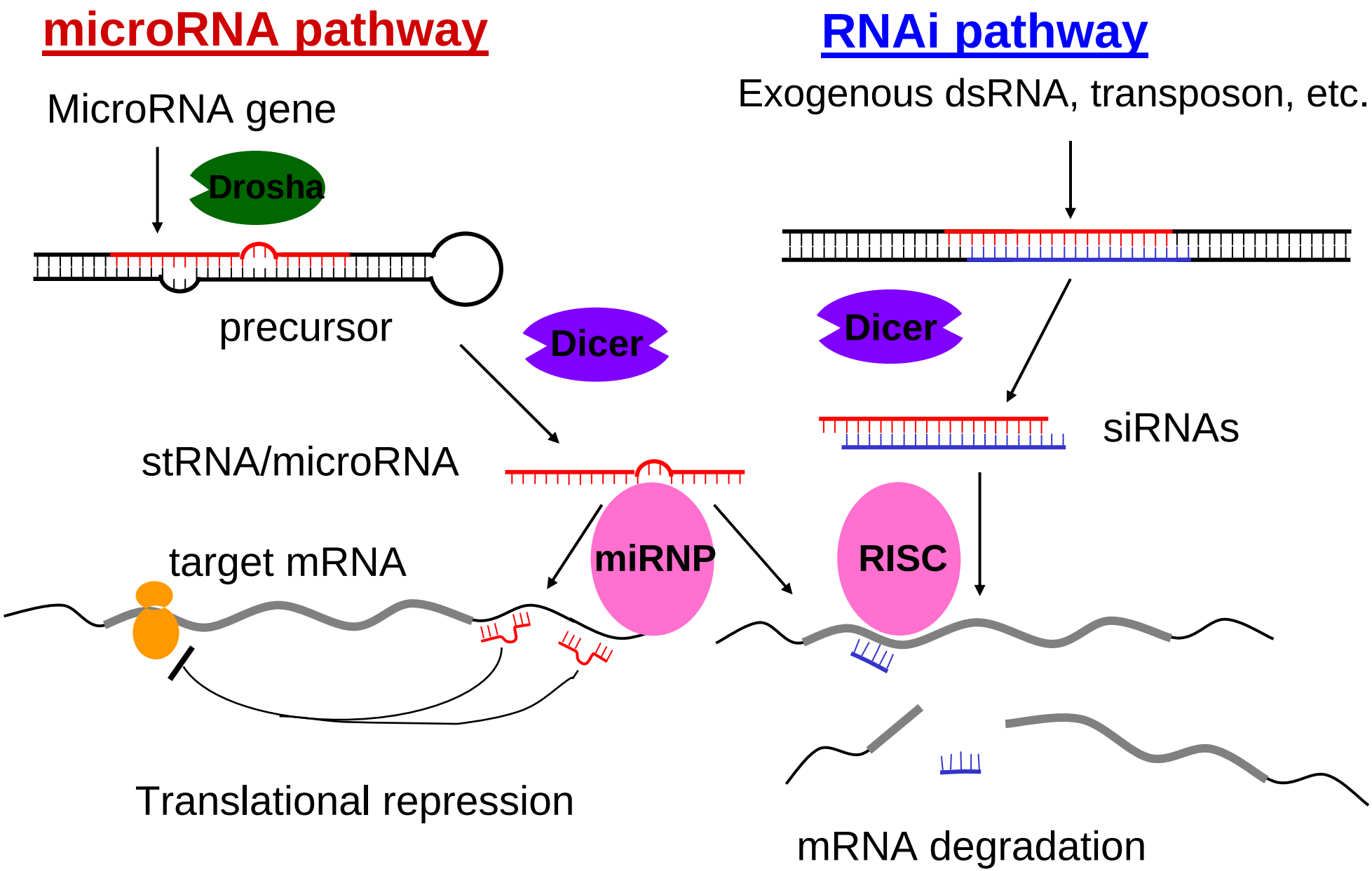
5S rRNA (120 nts)

23S rRNA (2904 nts)

34 proteins

The ribosome is a large macromolecular machine composed of RNA and protein components in a ratio of about 2 to 1. For many years,biochemistry and evolutionary considerations have argued for a central role being played by the rRNAs in the function of the ribosome. Now, in the face of atomic resolution data, the answer is clear - the ribosome is an RNA machine - and that is part of the story that I will tell you about today.

The microRNA and RNAi Pathways



- Soon after the discovery of *let*-7, the Mello, Zamore, and Hannon labs reported that CLICK gene inactivation by RNAi and the control of developmental timing by stRNAs, are interconnected processes that share certain molecular components.

- --

- The most prominent component was the highly-conserved nuclease Dicer, which cleaves double-stranded precursor molecules into stRNAs and siRNAs.

- Essential background on microRNAs

- - Family of small non-coding RNAs found in animals and plants

- Endogenous precursor RNA foldbacks are processed by the enzyme Dicer to mature single-stranded 21 or 22 nt microRNAs

- RNA interference involves longer, perfect duplex RNA (exogenous or endogenous), processed by Dicer to ~21mer siRNAs

- Characterized animal microRNAs direct translational inhibition by basepairing to 3′ UTRs of protein coding mRNAs and are often involved in developmental control. Pairing between microRNA and mRNA is always partial/incomplete (usually multiple bulges/loops). There are typically several microRNA complementary sites per regulated mRNA.

- Plant microRNAs and siRNAs generally have perfect or near-perfect complementarity to mRNAs and can trigger mRNA degradation.

Ways to Predict RNA 2^{o} Structure

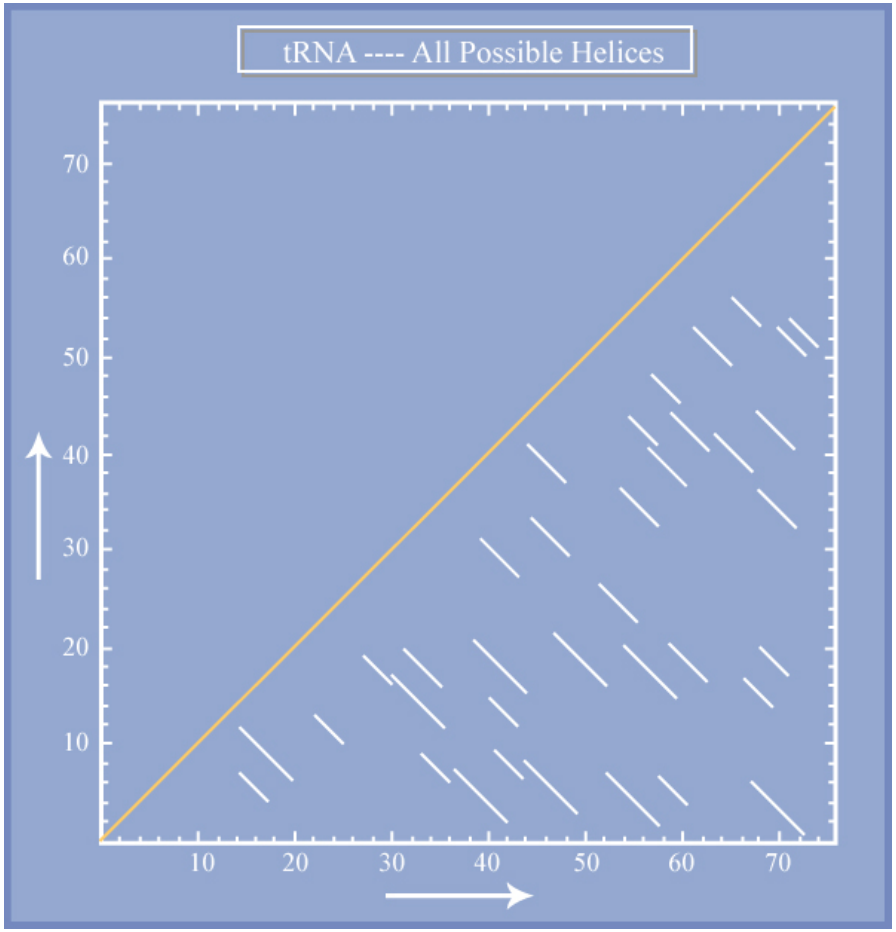
- Dot Plot

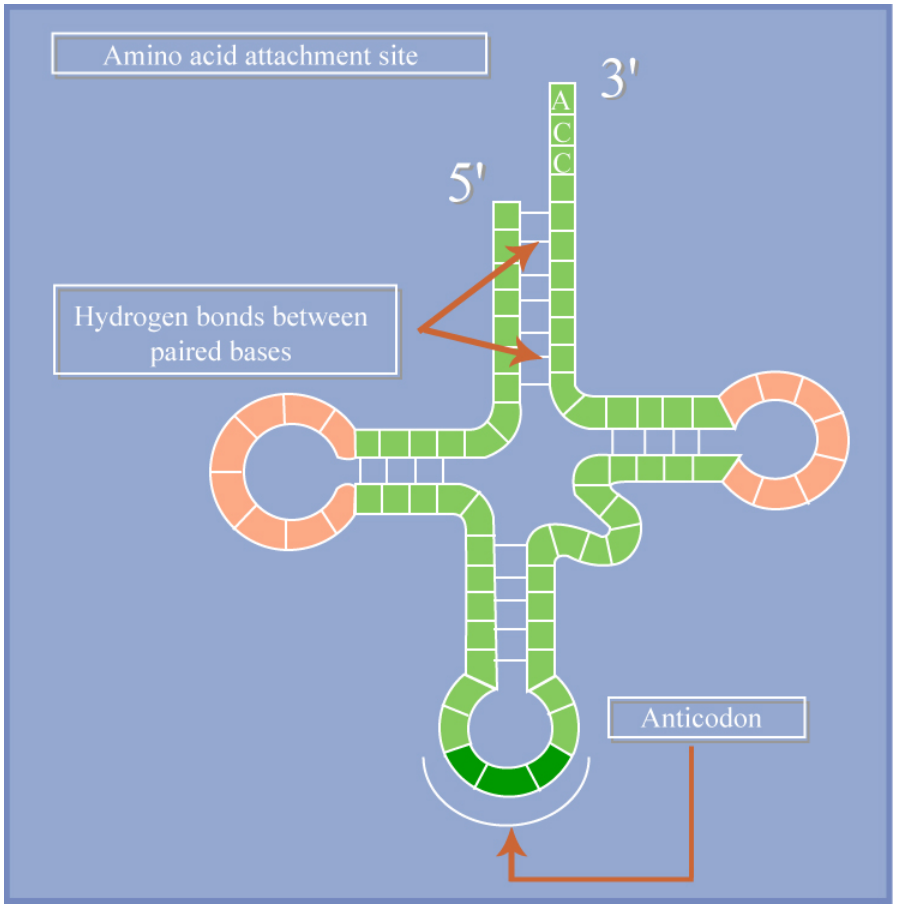
  (+ Dynamic Programming on Helices)

- Energy Minimization

- Covariation

Helices in tRNA





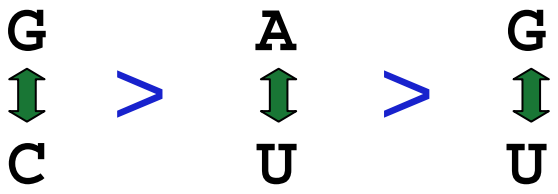
All possible helices

RNA Energetics I

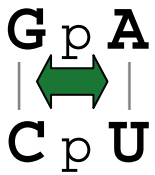
```
   ...CCAUUCAUAG-5'
      ||||||
5'...CGUGAGU.....  3'
```

Free energy of helix formation derives from:

- base pairing:



- base stacking:



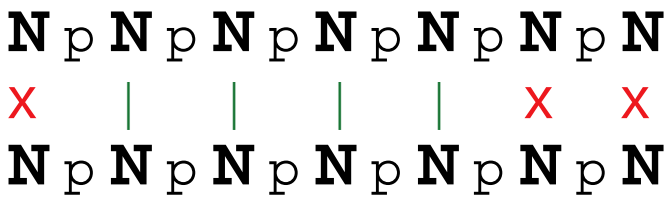
Doug Turner's Energy Rules:

```
                5' --> 3'
                   UX
                   AY
                3' <-- 5'

                      X
Y       A       C       G       U
A       .       .       .   -1.30
C       .       .   -2.40       .
G       .   -2.10       .   -1.00
T   -0.90       .   -1.30       .
```

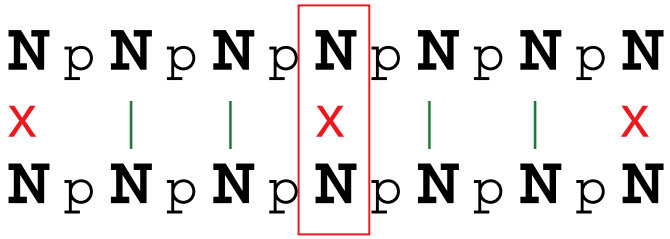
RNA Energetics II

A)



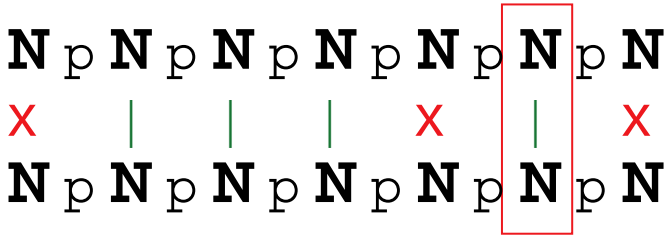
Lots of consecutive base pairs - good

B)



Internal loop - bad

C)



Terminal base pair not stable - bad

Generally A will be more stable than B or C

RNA Energetics III

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

See Mount, Ch. 5

RNA Energetics IV

Folding by Energy Minimization

A clever dynamic programming algorithm is used

- the Zuker algorithm - see Mount, Ch. 5 for details

Gives:

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

Accuracy: ~70-80% of base pairs correct

M. Zuker, a Canadian scientist, now at RPI

Practical Stuff

The Mfold web server:

http://www.bioinfo.rpi.edu/applications/mfold/old/rna/

The Vienna RNAfold package (free for download)

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

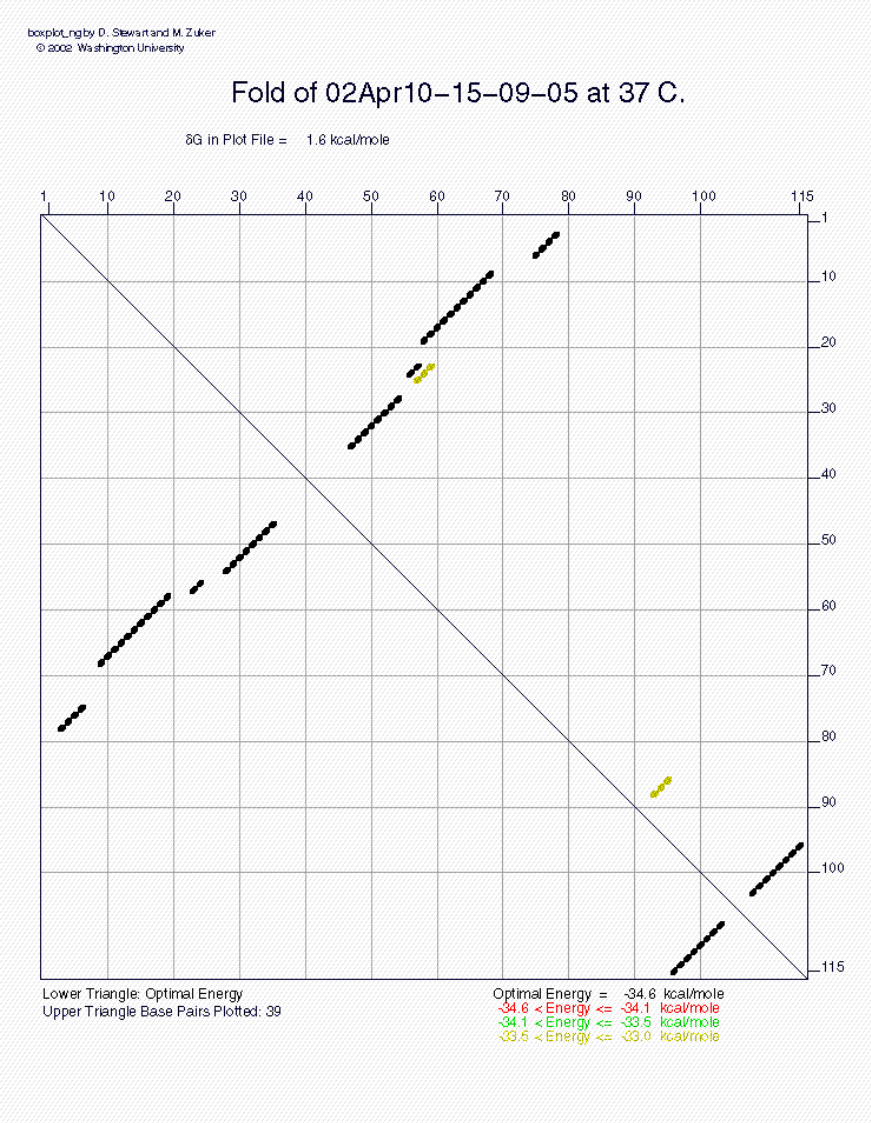
RNA folding references:

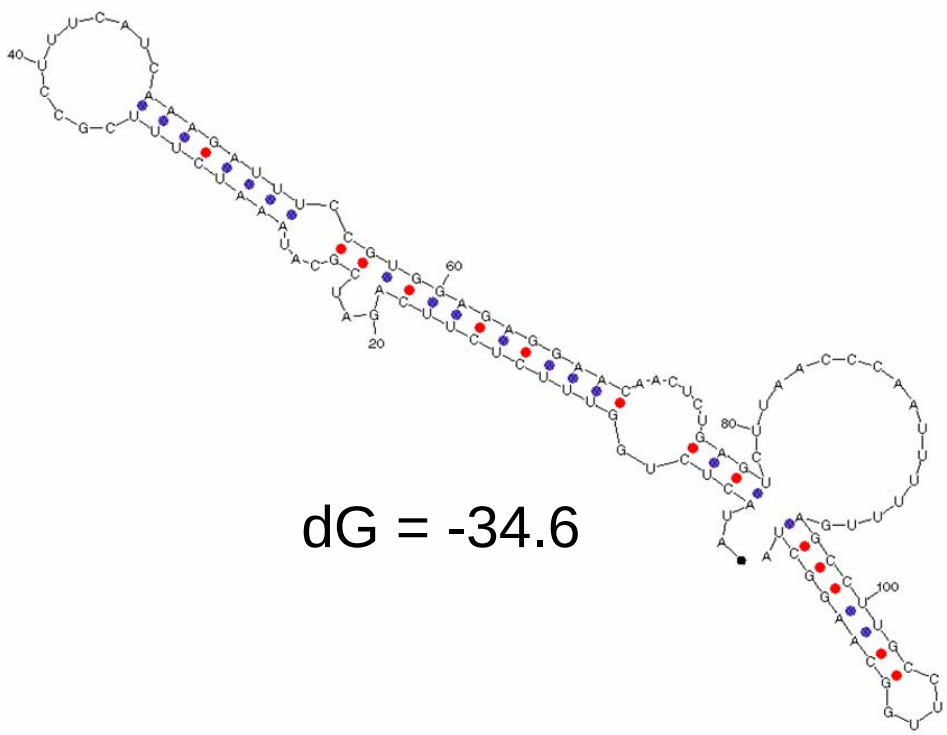
M. Zuker, et al. In *RNA Biochemistry and Biotechnology* (1999)

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

Vienna package by Ivo Hofacker

Sample Mfold Output

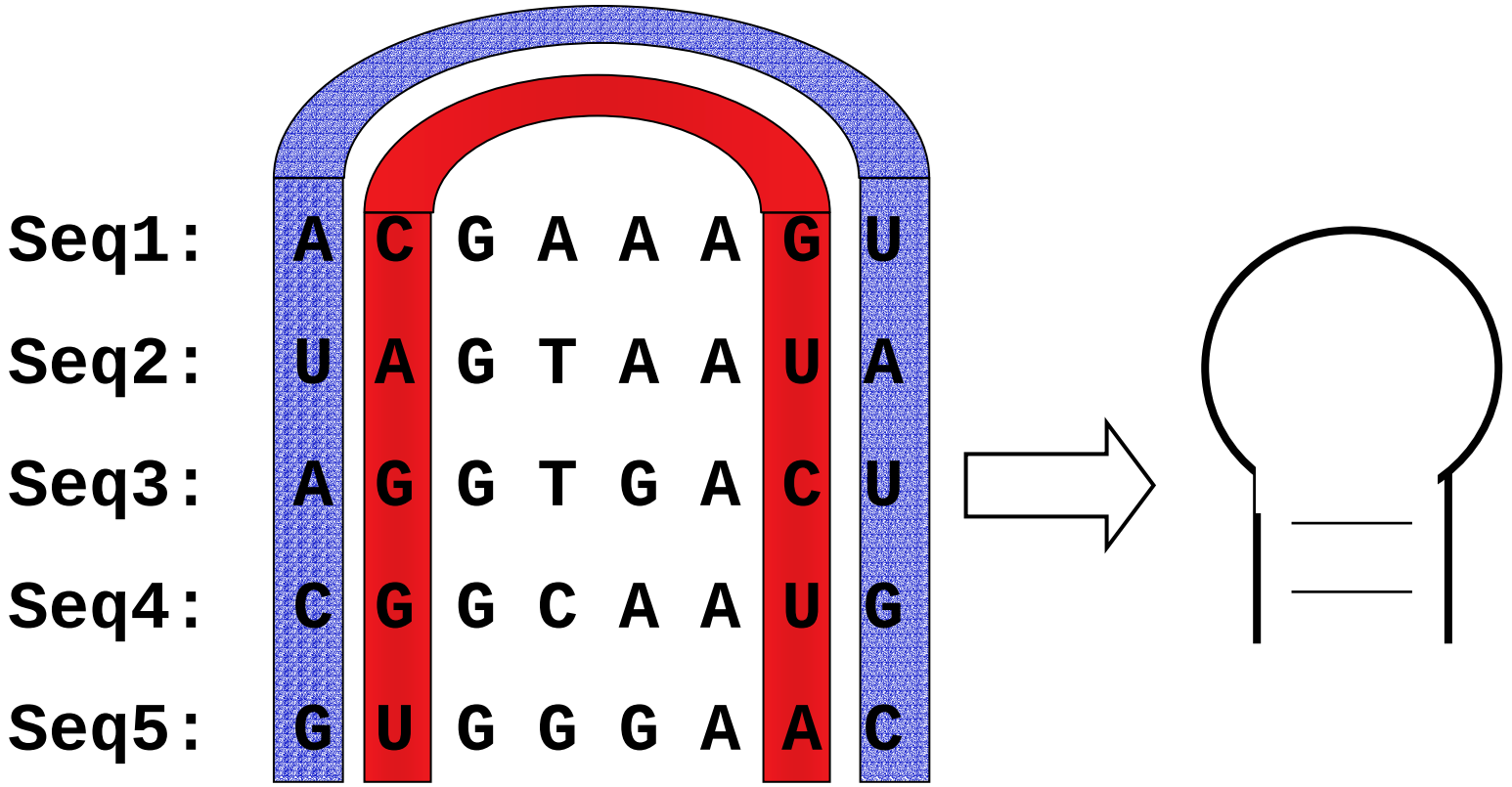




dG = -34.6

http://www.biology.wustl.edu/gcg/mfold.html

Other ways to infer RNA 2^{o} structure



Method of 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 are perfectly correlated (e.g., always complementary)

Inferring 2^{o} structure from covariation

Please see

Brown, T. A. *Genomes.* NY: John Wiley & Sons, 1999.

The ncRNA Gene Finding Problem

Approach 1:

Devise algorithm to find specific family of ncRNAs

- Lowe, T. M. and S. R. Eddy. "A Computational Screen for Methylation Guide snoRNAs in Yeast." *Science* 283 (1999): 1168.

Approach 2:

Devise algorithm to find ncRNAs in general

- Rivas, E., et al. "Computational Identification of Noncoding RNAs in *E. coli* by Comparative Genomics." *Curr. Biol.* 11 (2001): 1369.

Literature Discussion Tues. 3/16

Paper #1:

Kellis, M, N Patterson, M Endrizzi, B Birren, and ES Lander. "Sequencing and Comparison of Yeast Species to Identify Genes and Regulatory Elements." *Nature* 423, no. 6937 (15 May 2003): 241-54.

Part 1 - Finding Genes, etc., pp. 241-247
Part 2 - Regulatory Elements, pp. 247-254

Paper #2:

Rivas, E, RJ Klein, TA Jones, and SR Eddy. "Computational Identification of Noncoding RNAs in E. coli by Comparative Genomics." *Curr Biol.*11, no. 17 (4 September 2001): 1369-73.