CSE 527

Autumn 2006

Lectures 15-16

RNA

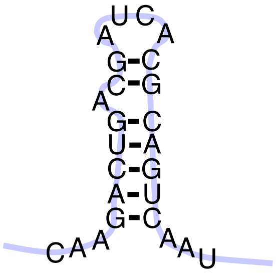
Secondary Structure Prediction

RNA Secondary Structure: RNA makes helices too

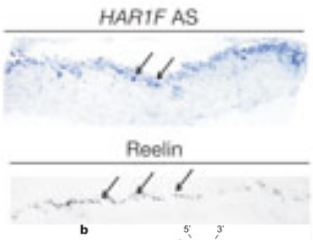
Base pairs

A-U

C-G



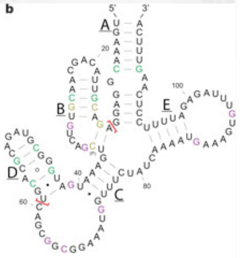
Fastest Human Gene?



a

```
Position              20        30        40        50
Human       AGACGTTACAGCAACGTGTCAGCTGAAATGATGGGCGTAGACGCACGT
Chimpanzee  AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Gorilla     AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Orang-utan  AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Macaque     AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT
Mouse       AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT
Dog         AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Cow         AGAAATTACAGCAATTCATCAGCTGAAATTATAGGTGTAGACACATGT
Platypus    ATAAATTACAGCAATTTATCAAATGAAATTATAGGTGTAGACACATGT
Opossum     AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Chicken     AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Fold        (((((((.....))))).....))[[[[[.(((.(((....))).)))
Pair symbol lmnopqr     rqpon     ml      rstuvwx    xwvutsr
```

b



Origin of Life?

Life needs

- information carrier: DNA
- molecular machines, like enzymes: Protein
- making proteins needs DNA + RNA + proteins
- making (duplicating) DNA needs proteins

Horrible circularities! How could it have arisen in an abiotic environment?

Origin of Life?

RNA can carry information too
(RNA double helix)

RNA can form complex structures

RNA enzymes exist (ribozymes)

The “RNA world” hypothesis:
1st life was RNA-based

Outline

Biological roles for RNA

What is “secondary structure?

How is it represented?

Why is it important?

Examples

Approaches

RNA Structure

Primary Structure: Sequence

Secondary Structure: Pairing

Tertiary Structure: 3D shape

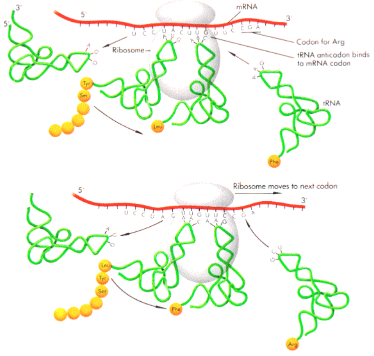
RNA Pairing

Watson-Crick Pairing

C - G	~ 3 kcal/mole
A - U	~ 2 kcal/mole
“Wobble Pair” G - U	~1 kcal/mole

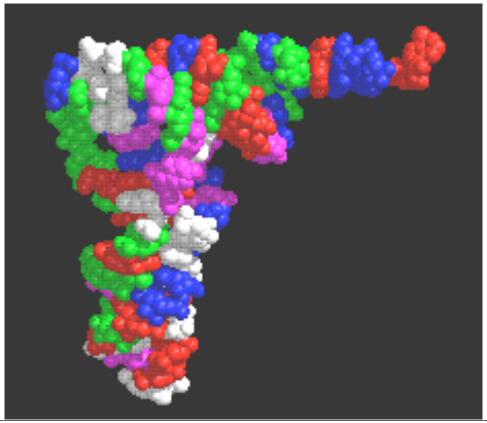
Non-canonical Pairs (esp. if modified)

Ribosomes


3'
5'
5'
mRNA
3'
U C C U A C U U G U U C C G A
A U G
A A C
A G G
Ribosome
Codon for Arg
tRNA anticodon binds to mRNA codon
G A A
tRNA
Tyr
Ser
Leu
Phe
5'
Ribosome moves to next codon
3'
U C C U A C U U G U U C C G A
A A C A A G
A C G
G C U
Leu
Tyr
Ser
Phe
Arg

Watson, Gilman, Witkowski, & Zoller, 1992

tRNA 3d Structure



tRNA - Alt. Representations

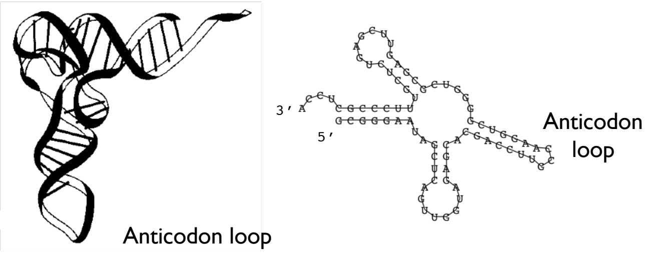
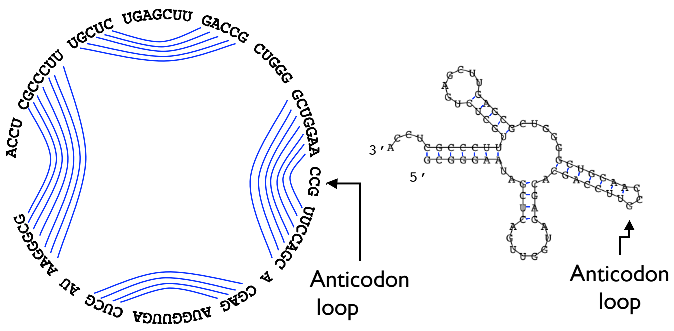


Figure 1: a) The spatial structure of the phenylalanine tRNA form yeast

b) The secondary structure extracts the most important information about the structure, namely the pattern of base pairings.

tRNA - Alt. Representations



“Classical” RNAs

tRNA - transfer RNA (~61 kinds, ~ 75 nt)

rRNA - ribosomal RNA (~4 kinds, 120-5k nt)

snRNA - small nuclear RNA (splicing: U1, etc, 60-300nt)

RNaseP - tRNA processing (~300 nt)

RNase MRP - rRNA processing; mito. rep. (~225 nt)

SRP - signal recognition particle; membrane targeting (~100-300 nt)

SECIS - selenocysteine insertion element (~65nt)

6S - ? (~175 nt)

Semi-classical RNAs

(discovery in mid 90’s)

tmRNA - resetting stalled ribosomes

Telomerase - (200-400nt)

snoRNA - small nucleolar RNA (many varieties; 80-200nt)

Recent discoveries

microRNAs (Nobel prize 2006, Fire & Mello)

riboswitches

many ribozymes

regulatory elements

…

Hundreds of families

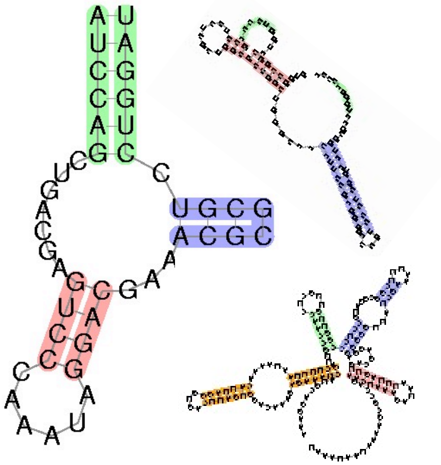
Rfam release 1, 1/2003: 25 families, 55k instances

Rfam release 7, 3/2005: 503 families, 300k instances

Why?

RNA’s fold, and function

Nature uses what works



Noncoding RNAs

Dramatic discoveries in last 5 years

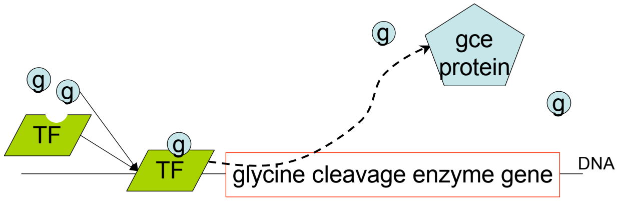
- *100s* of new families
- Many roles: Regulation, transport, stability, catalysis, …

1% of DNA codes for protein, but 30% of it is copied into RNA, i.e. ncRNA >> mRNA

Example: Glycine Regulation

How is glycine level regulated?

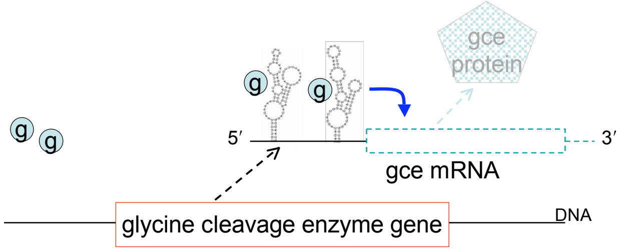
Plausible answer:



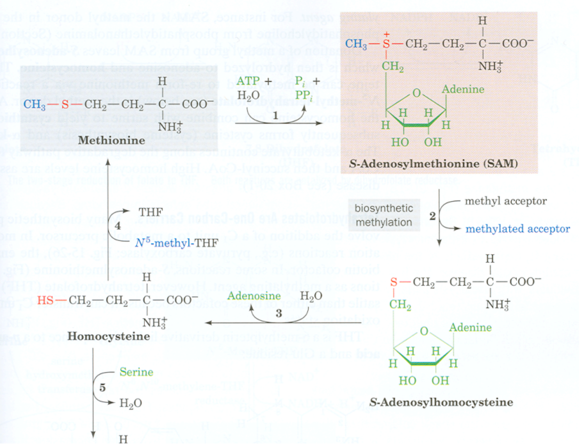
transcription factors (proteins) bind to DNA to turn nearby genes on or off

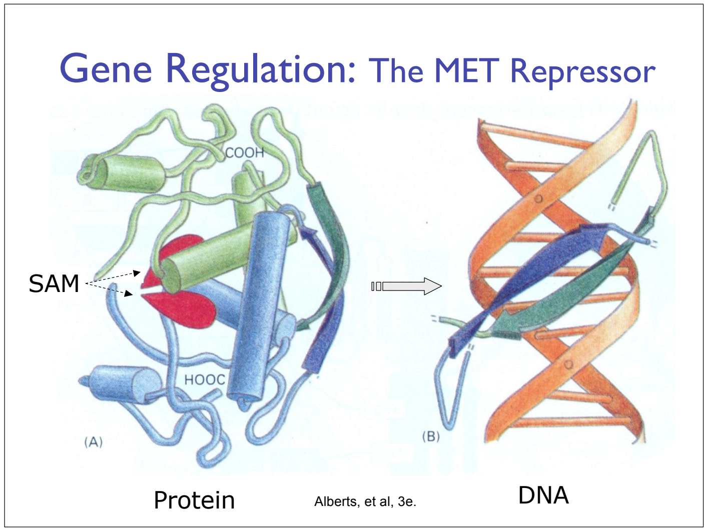
The Glycine Riboswitch

Actual answer (in many bacteria):



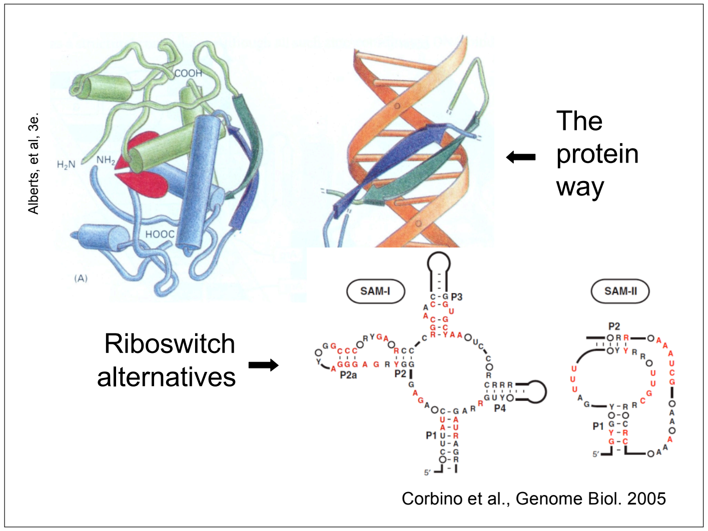
Mandal et al. Science 2004





Gene Regulation: The MET Repressor

Alberts, et al, 3e.

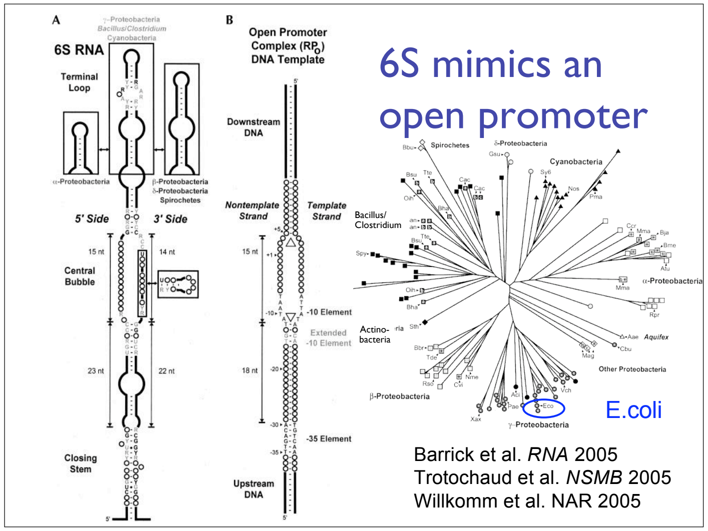


Alberts, et al, 3e.

The protein way

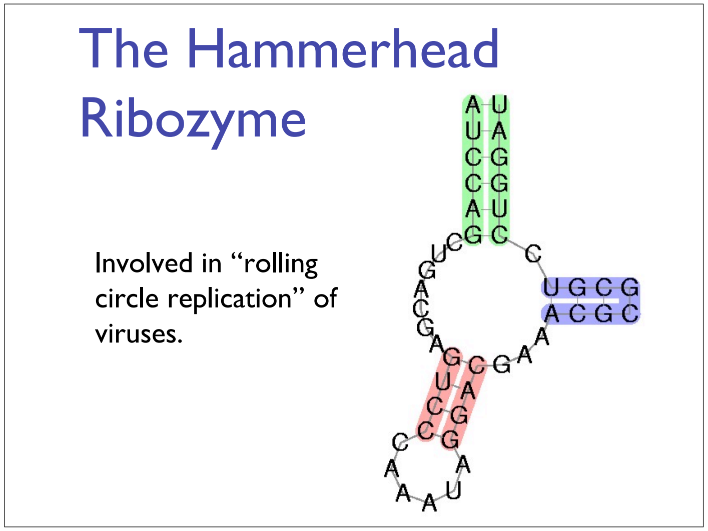
Riboswitch alternatives

Corbino et al., Genome Biol. 2005



6S mimics an open promoter

Barrick et al. *RNA* 2005
Trotochaud et al. *NSMB* 2005
Willkomm et al. NAR 2005



The Hammerhead Ribozyme

Involved in "rolling circle replication" of viruses.

Wanted

Good structure prediction tools

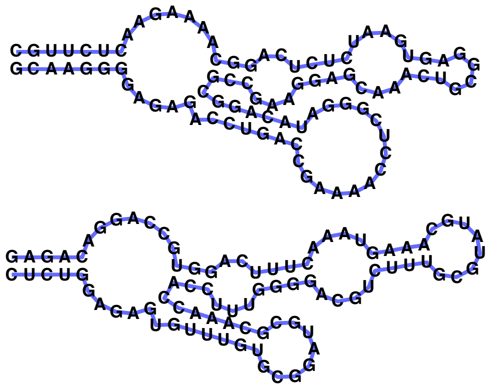
Good motif descriptions/models

Good, fast search tools
 (“RNA BLAST”, etc.)

Good, fast motif discovery tools
 (“RNA MEME”, etc.)

Importance of structure makes last 3 hard

Why is RNA hard to deal with?



A: *Structure* often more important than *sequence*

Task 1: Structure Prediction

RNA Pairing

Watson-Crick Pairing	
C - G	~ 3 kcal/mole
A - U	~ 2 kcal/mole
“Wobble Pair” G - U	~ 1 kcal/mole
Non-canonical Pairs (esp. if modified)	

Definitions

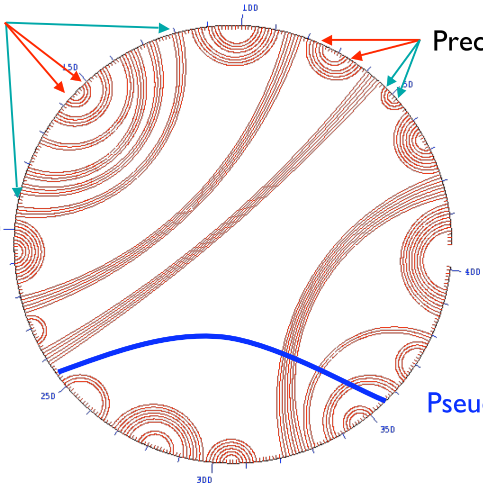
Sequence $^{5'}\, r_1\ r_2\ r_3 \ldots r_n\, ^{3'}$ in {A, C, G, T}

A Secondary Structure is a set of pairs i•j s.t.

i < j-4, and } no sharp turns

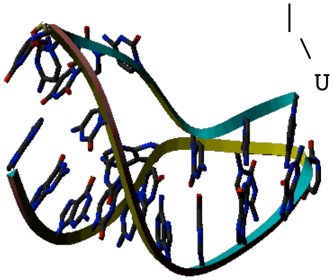
if i•j & i'•j' are two different pairs with i ≤ i', then

j < i', or
i < i' < j' < j
} 2nd pair follows 1st, or is nested within it; no "pseudoknots."



A Pseudoknot

```
                      A-C
                     /   \
3' - A-G-G-C-U        U
     U-C-C-G-A-G-G-G
     |           C-C-C - 5'
      \         /
       U-C-U-C
```



Approaches to Structure Prediction

Maximum Pairing
- + works on single sequences
- + simple
- - too inaccurate

Minimum Energy
- + works on single sequences
- - ignores pseudoknots
- - only finds "optimal" fold

Partition Function
- + finds all folds
- - ignores pseudoknots

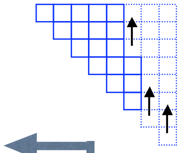
Nussinov: Max Pairing

B(i,j) = # pairs in optimal pairing of $r_i \dots r_j$

B(i,j) = 0 for all i, j with i ≥ j-4; otherwise

B(i,j) = max of:

$$\begin{cases} B(i,j-1) \\ \max \{ B(i,k-1)+1+B(k+1,j-1) \mid i \le k < j-4 \text{ and } r_k\text{-}r_j \text{ may pair}\} \end{cases}$$



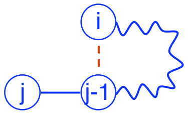
Time: $O(n^3)$

"Optimal pairing of $r_i \dots r_j$"

Two possibilities

J Unpaired:

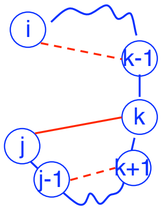
Find best pairing of $r_i \dots r_{j-1}$



J Paired:

Find best $r_i \dots r_{k-1}$ +

best $r_{k+1} \dots r_{j-1}$ plus 1



Why is it slow?

Why do pseudoknots matter?

Pair-based Energy Minimization

E(i,j) = energy of pairs in optimal pairing of $r_i \dots r_j$

E(i,j) = ∞ for all i, j with i ≥ j-4; otherwise

E(i,j) = min of:

$$\begin{cases} E(i,j-1) \\ \min \{ E(i,k-1) + e(r_k, r_j) + E(k+1,j-1) \mid i \le k < j-4 \} \end{cases}$$

energy of j-k pair

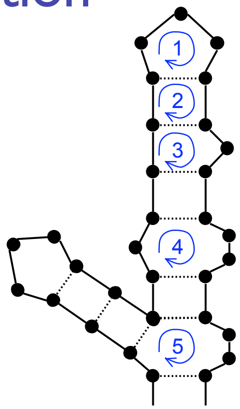
Time: $O(n^3)$

Loop-based Energy Minimization

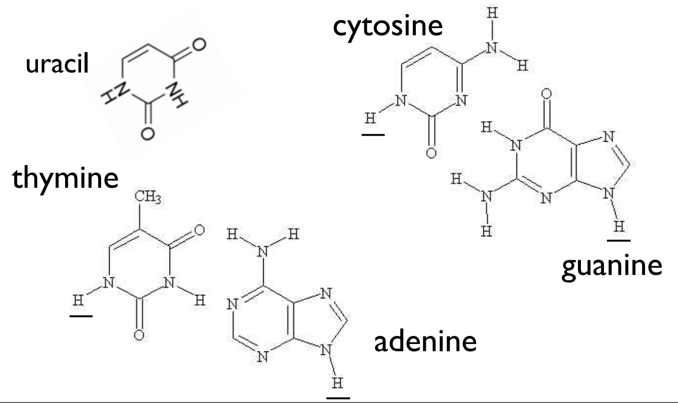
Detailed experiments show it's more accurate to model based on loops, rather than just pairs

Loop types

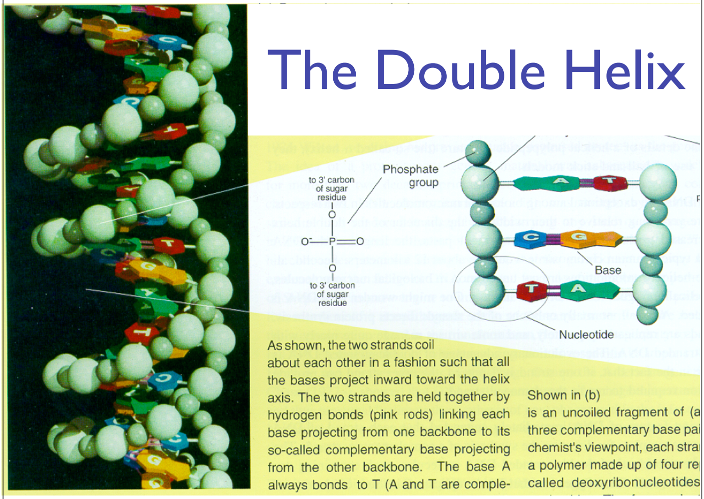
- Hairpin loop
- Stack
- Bulge
- Interior loop
- Multiloop



Base Pairs and Stacking



The Double Helix

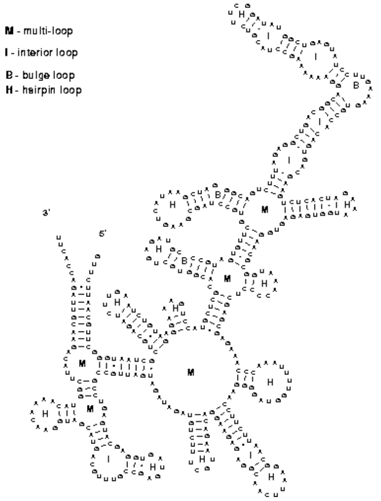


As shown, the two strands coil about each other in a fashion such that all the bases project inward toward the helix axis. The two strands are held together by hydrogen bonds (pink rods) linking each base projecting from one backbone to its so-called complementary base projecting from the other backbone. The base A always bonds to T (A and T are comple-

Shown in (b) is an uncoiled fragment of (a three complementary base pai chemist's viewpoint, each stra a polymer made up of four re called deoxyribonucleotides

Loop Examples

Bacillus subtilis RNase P RNA



Zuker: Loop-based Energy, I

W(i,j) = energy of optimal pairing of r_i ... r_j

V(i,j) = as above, but forcing pair i•j

W(i,j) = V(i,j) = ∞ for all i, j with i ≥ j-4

$$W(i,j) = \min\Big(W(i,j-1),\ \min\{\, W(i,k-1)+V(k,j) \mid i \le k < j-4 \,\}\Big)$$

Zuker: Loop-based Energy, II

hairpin stack bulge/ interior multi- loop

$V(i,j) = \min(eh(i,j), es(i,j)+V(i+1,j-1), VBI(i,j), VM(i,j))$

$VM(i,j) = \min \{ W(i,k)+W(k+1,j) \mid i < k < j \}$

$VBI(i,j) = \min \{ ebi(i,j,i',j') + V(i', j') \mid i < i' < j' < j \;\&\; i'-i+j-j' > 2 \}$

bulge/ interior

Time: $O(n^4)$

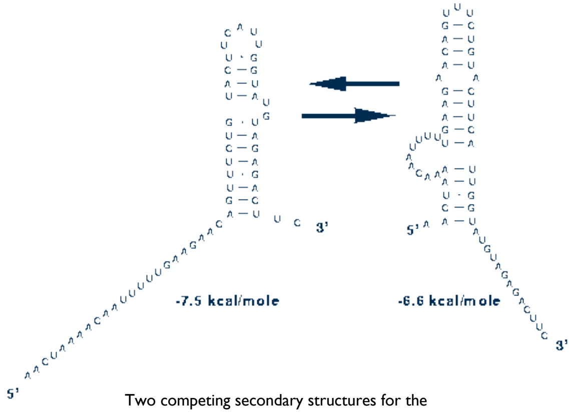
$O(n^3)$ possible if ebi(.) is "nice"

Suboptimal Energy

There are always alternate folds with near-optimal energies. Thermodynamics: populations of identical molecules will exist in different folds; individual molecules even flicker among different folds

Mod to Zuker's algorithm finds subopt folds

McCaskill: more elaborate dyn. prog. algorithm calculates the "partition function," which defines the probability distribution over all these states. (Key addition: recurrence must count each possibility exactly once.)

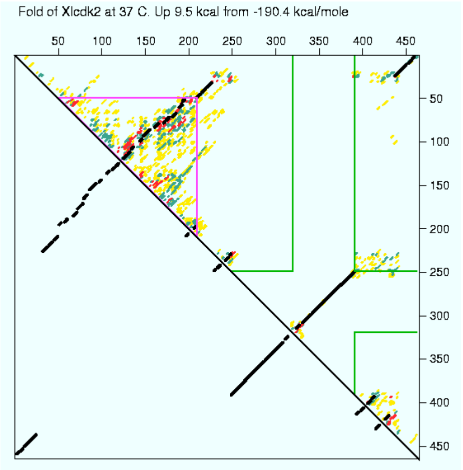


Two competing secondary structures for the Leptomonas collosoma spliced leader mRNA.

Example of suboptimal folding

Black dots: pairs in opt fold

Colored dots: pairs in folds 2-5% worse than optimal fold



Accuracy

Latest estimates suggest ~50-75% of base pairs predicted correctly in sequences of up to ~300nt

Definitely useful, but obviously imperfect

Approaches to Structure Prediction

Maximum Pairing
+ works on single sequences
+ simple
- too inaccurate

Minimum Energy
+ works on single sequences
- ignores pseudoknots
- only finds "optimal" fold

Partition Function
+ finds all folds
- ignores pseudoknots

Approaches, II

Comparative sequence analysis
+ handles all pairings (incl. pseudoknots)
- requires several (many?) aligned, appropriately diverged sequences

Stochastic Context-free Grammars
Roughly combines min energy & comparative, but no pseudoknots

Physical experiments (x-ray crystalography, NMR)

Summary

RNA has important roles beyond mRNA

Many unexpected recent discoveries

Structure is critical to function

True of proteins, too, but they're easier to find, due, e.g., to codon structure, which RNAs lack

RNA secondary structure can be predicted (to useful accuracy) by dynamic programming

Next time: RNA "motifs" (seq + 2-ary struct) well-captured by "covariance models"