CSE 527
Computational Biology

RNA: Function, Secondary Structure Prediction, Search, Discovery

The Message

Cells make lots of ~~RNA~~ *noncoding* RNA

Functionally important, functionally diverse

Structurally complex

New tools required

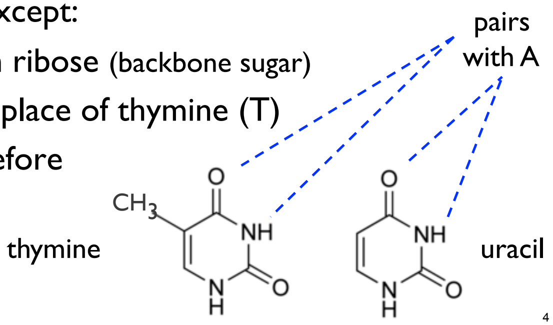
alignment, discovery, search, scoring, etc.

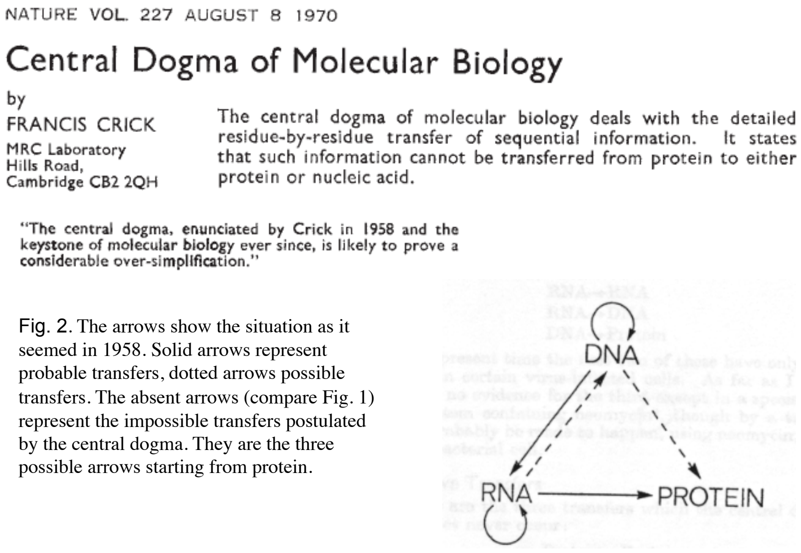
RNA

DNA: DeoxyriboNucleic Acid

RNA: RiboNucleic Acid

- Like DNA, except:
- Lacks OH on ribose (backbone sugar)
- Uracil (U) in place of thymine (T)
- A, G, C as before





NATURE VOL. 227 AUGUST 8 1970

Central Dogma of Molecular Biology

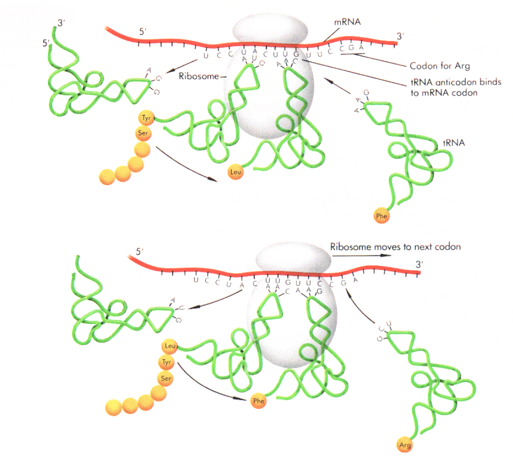
by
FRANCIS CRICK
MRC Laboratory
Hills Road,
Cambridge CB2 2QH

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

"The central dogma, enunciated by Crick in 1958 and the keystone of molecular biology ever since, is likely to prove a considerable over-simplification."

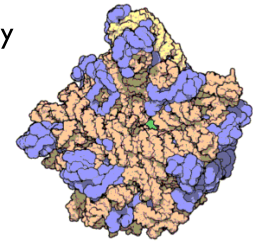
Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

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
A A G
tRNA
Tyr
Ser
Leu
Phe
Ribosome moves to next codon
5′
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

Ribosomes

1974 Nobel prize to Romanian biologist George Palade (1912-2008) for discovery in mid 50's



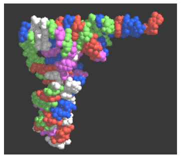
50-80 proteins

3-4 RNAs (half the mass)

Catalytic core is RNA

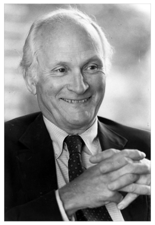
Of course, mRNAs and tRNAs (messenger & transfer RNAs) are critical too

Transfer RNA



The "adapter" coupling mRNA to protein synthesis.

Discovered in the mid-1950s by Mahlon Hoagland (1921-2009, left), Mary Stephenson, and Paul Zamecnik (1912-2009; Lasker award winner, right).



"Classical" RNAs

rRNA - ribosomal RNA (~4 kinds, 120-5k nt)
tRNA - transfer RNA (~61 kinds, ~ 75 nt)
RNaseP - tRNA processing (~300 nt)
snRNA - small nuclear RNA (splicing: U1, etc, 60-300nt)

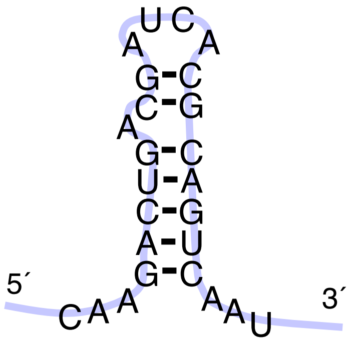
a handful of others

RNA Secondary Structure:
RNA makes helices too

Base pairs

A-U
C-G

U-G



Usually *single* stranded

Bacteria

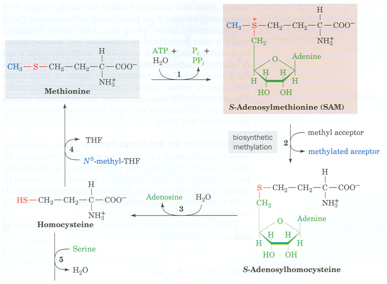
Triumph of proteins

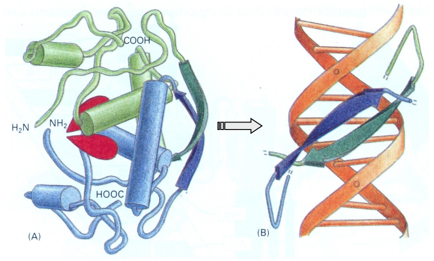
80% of genome is coding DNA

Functionally diverse

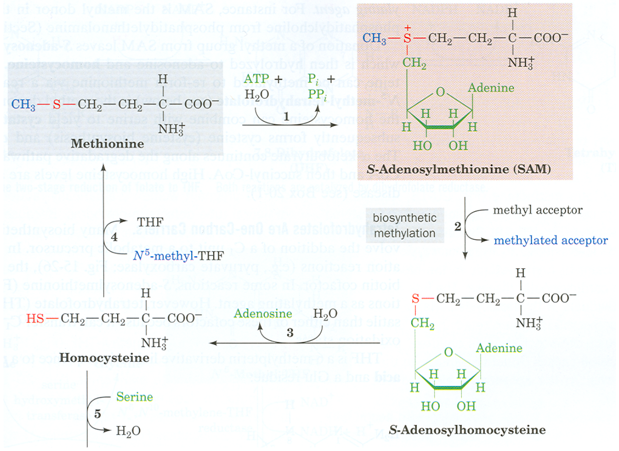
- receptors
- motors
- catalysts
- regulators (Monod & Jakob, Nobel prize 1965)
- …

Proteins catalyze & regulate biochemistry



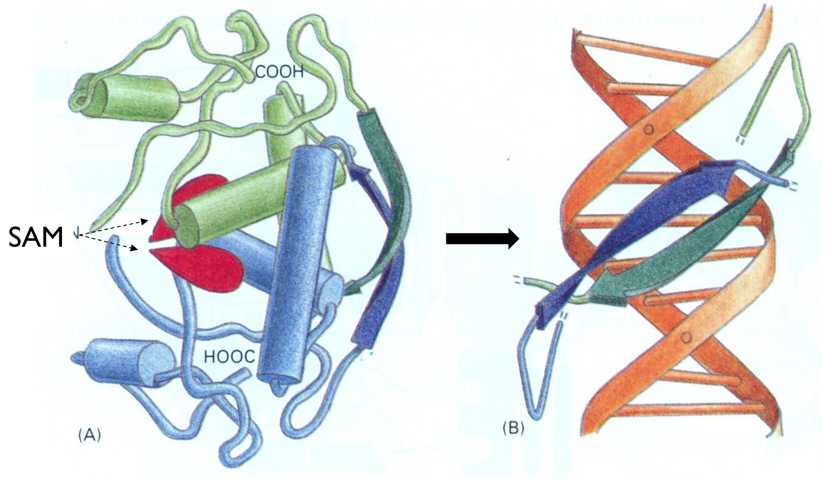


Met Pathways



• • •

Gene Regulation: The MET Repressor

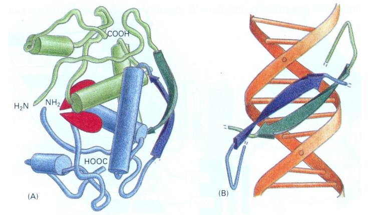


Protein

Alberts, et al, 3e.

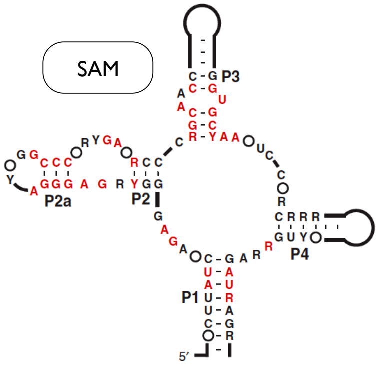
DNA

Alberts, et al, 3e.



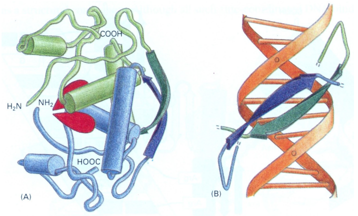
← The protein way

Riboswitch alternative



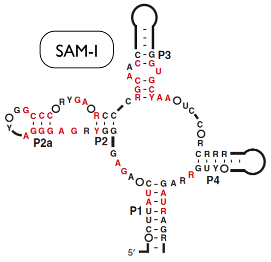
Grundy & Henkin, Mol. Microbiol 1998
Epshtein, et al., PNAS 2003
Winkler et al., Nat. Struct. Biol. 2003

Alberts, et al, 3e.

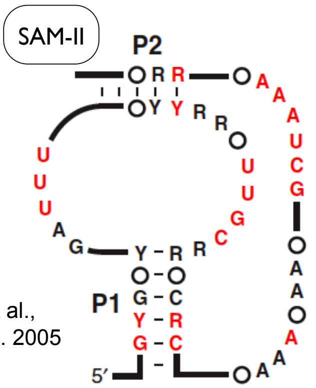


← The protein way

Riboswitch alternatives

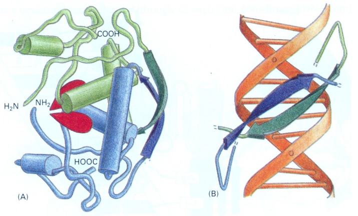


Grundy, Epshtein, Winkler et al., 1998, 2003



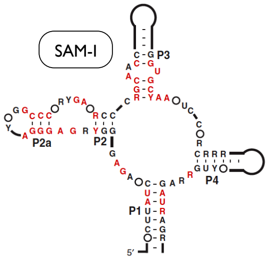
Corbino et al., Genome Biol. 2005

Alberts, et al, 3e.

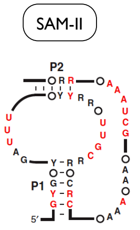


← The protein way

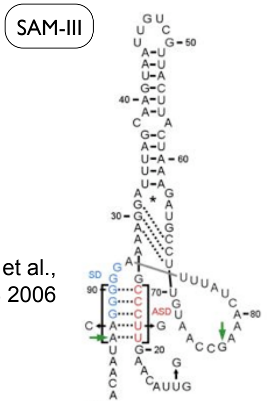
Riboswitch alternatives



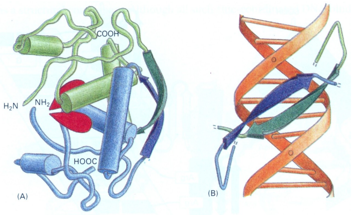
Grundy, Epshtein, Winkler et al., 1998, 2003



Corbino et al., Genome Biol. 2005



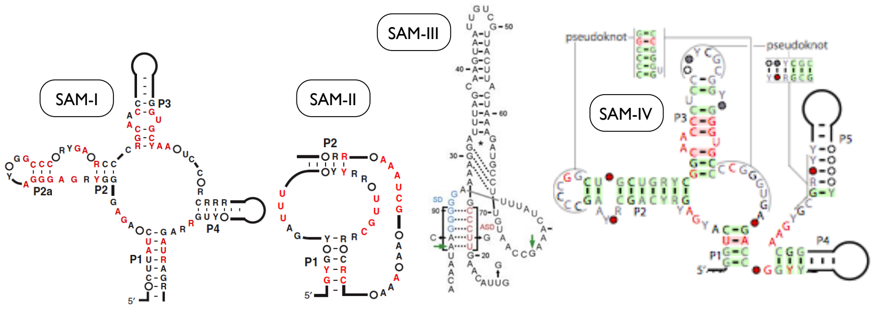
Fuchs et al., NSMB 2006



Alberts, et al, 3e.

← The protein way

Riboswitch alternatives ↓

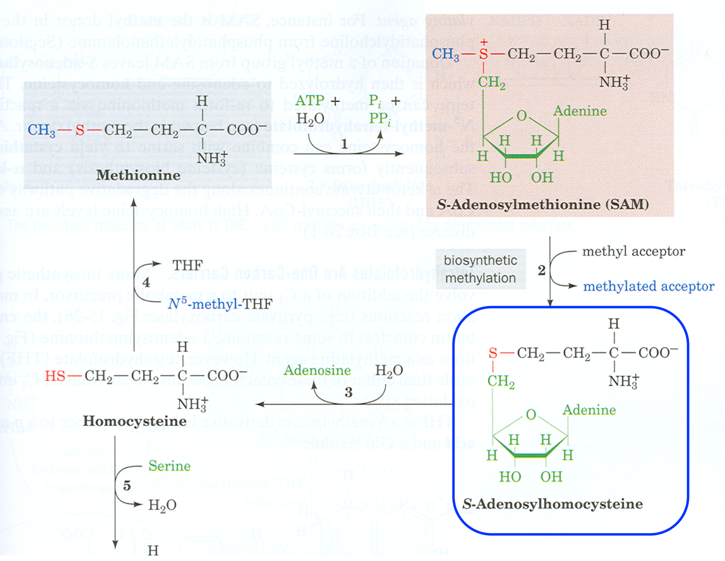


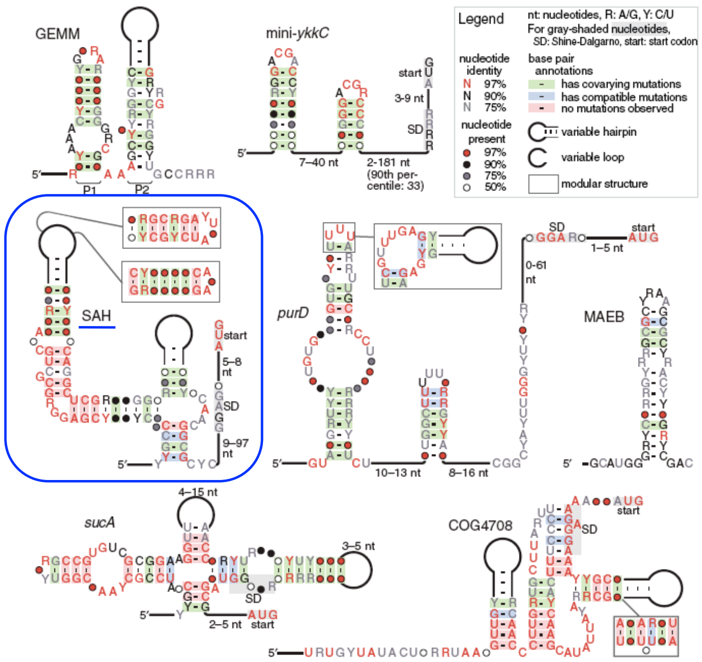
Grundy, Epshtein, Winkler et al., 1998, 2003

Corbino et al., Genome Biol. 2005

Fuchs et al., NSMB 2006

Weinberg et al., RNA 2008

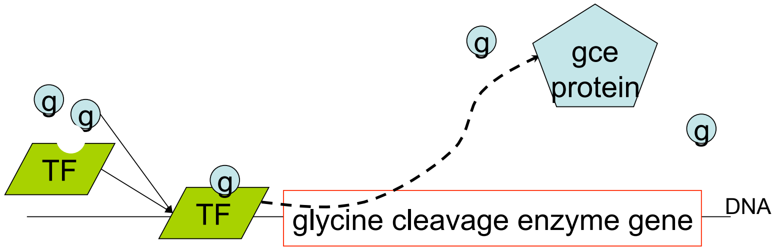




Example: Glycine Regulation

How is glycine level regulated?

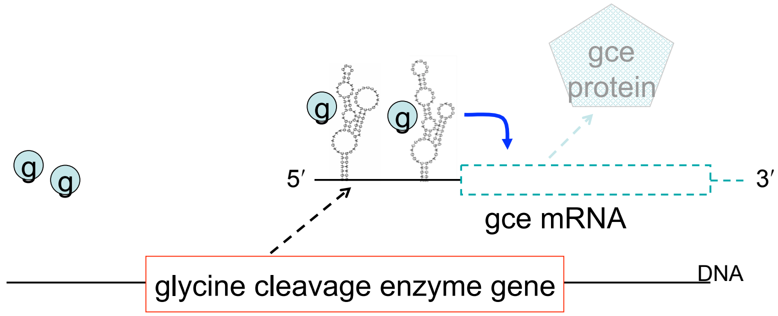
Plausible answer:



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

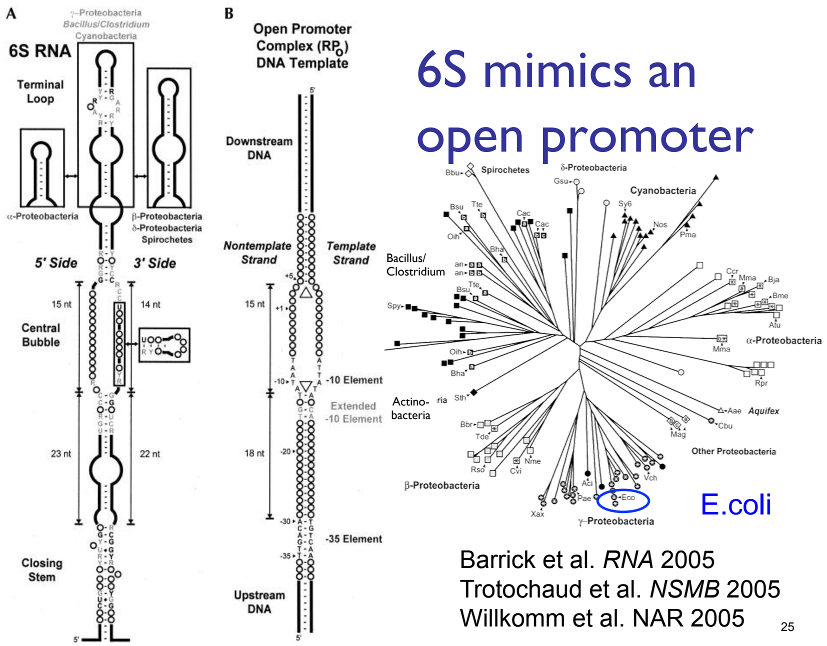
The Glycine Riboswitch

Actual answer (in many bacteria):

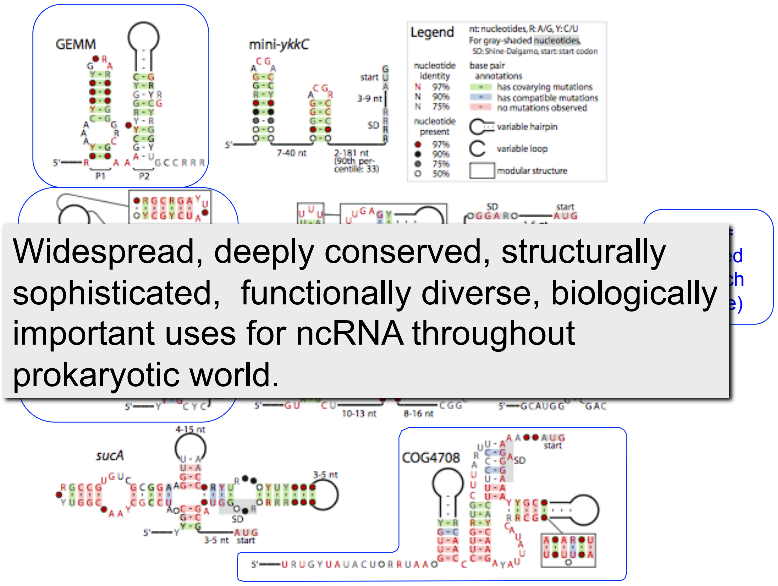


Mandal et al. Science 2004

6S mimics an open promoter



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



Widespread, deeply conserved, structurally sophisticated, functionally diverse, biologically important uses for ncRNA throughout prokaryotic world.

Weinberg, et al. Nucl. Acids Res., July 2007 35: 4809-4819.

Vertebrates

Bigger, more complex genomes

<2% coding

But >5% conserved in sequence?

And 50-90% transcribed?

And *structural* conservation, if any, invisible
(without proper alignments, etc.)

What's going on?

Vertebrate ncRNAs

mRNA, tRNA, rRNA, … of course

PLUS:

snRNA, spliceosome, snoRNA, teleomerase, microRNA, RNAi, SECIS, IRE, piwi-RNA, XIST (X-inactivation), ribozymes, …

MicroRNA

1st discovered 1992 in C. elegans

2nd discovered 2000, also C. elegans

- *and* human, fly, everything between

21-23 nucleotides

- literally fell off ends of gels

Hundreds now known in human

- may regulate 1/3-1/2 of all genes
- development, stem cells, cancer, infectious diseases,…

siRNA

“Short Interfering RNA”

Also discovered in *C. elegans*

Possibly an antiviral defense, shares machinery with miRNA pathways

Allows artificial repression of most genes in most higher organisms

Huge tool for biology & biotech

ncRNA Characteristics

Often low levels

Can come from anywhere

- Sense, antisense, introns, intergenic

Often poorly conserved

- CDS : neutral ~ 10 : 1 vs ncRNA : neutral ~ 1.2 : 1

May suggest “transcriptional noise”

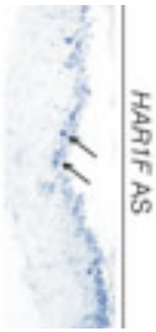
Noise?

HOWEVER:

Sometimes capped, spliced, polyA+

Some known ncRNAs are intronic (e.g. some miRNAs, all snoRNAs)

Sometimes very precisely localized to specific compartments, cell types, developmental stages, (esp. dev & neuronal …)



Conservation?

Neutral rate underestimated?

Promoters also evolving rapidly

Sequence/function constraint for RNA ≠ CDS

Alignments are suspect away from CDS

Alignments are not optimized for RNA *structure*

Despite all this, there *is* evidence for purifying selection on ncRNA promoters, splice sites, tissue-specific expression patterns, indels, …

Bottom line?

A significant number of "one-off" examples

Extremely wise-spread ncRNA expression

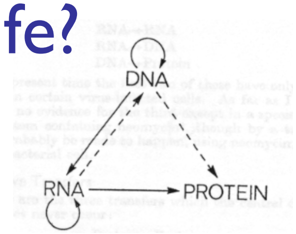
At a minimum, a vast evolutionary substrate

New technology (e.g. RNAseq) exposing more

How do you recognize an interesting one?

Conserved secondary structure

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-directed RNA polymerase

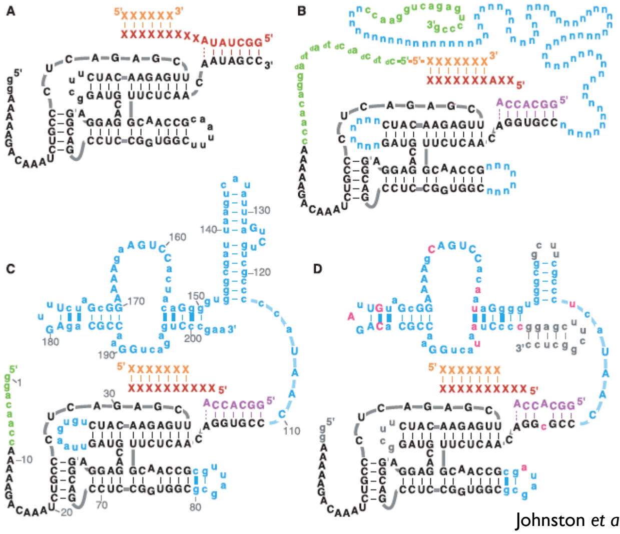
RNA can form complex structures

RNA enzymes exist (ribozymes)

RNA can control, do logic (riboswitches)

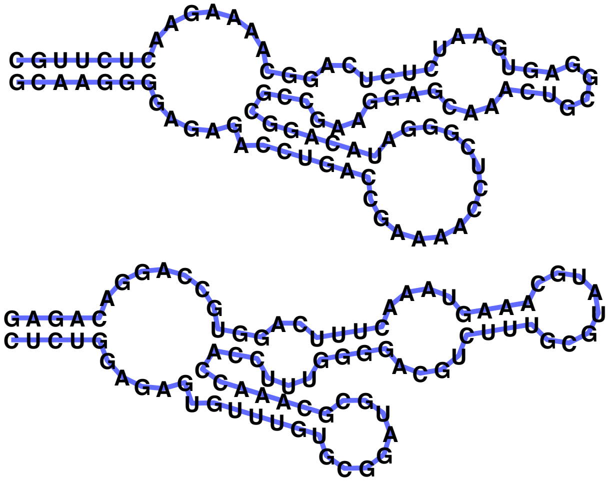
The “RNA world” hypothesis:
1st life was RNA-based

RNA replicase



Johnston *et al., Science,* 2001

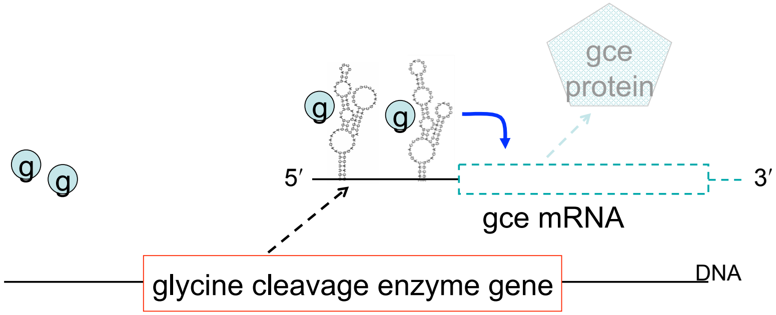
Why is RNA hard to deal with?



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

The Glycine Riboswitch

Actual answer (in many bacteria):



Mandal et al. Science 2004

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

Task 1:
Structure Prediction

RNA Structure

Primary Structure: Sequence

Secondary Structure: Pairing

Tertiary Structure: 3D shape

RNA Pairing

Watson-Crick Pairing

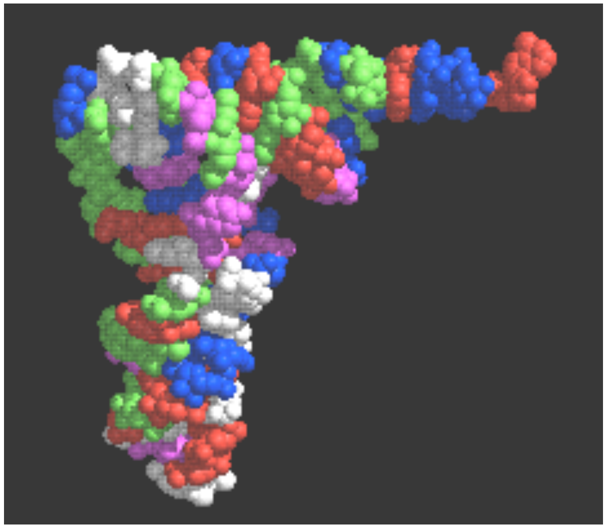
C - G ~ 3 kcal/mole

A - U ~ 2 kcal/mole

“Wobble Pair” G - U ~1 kcal/mole

Non-canonical Pairs (esp. if modified)

tRNA 3d Structure



tRNA - Alt. Representations

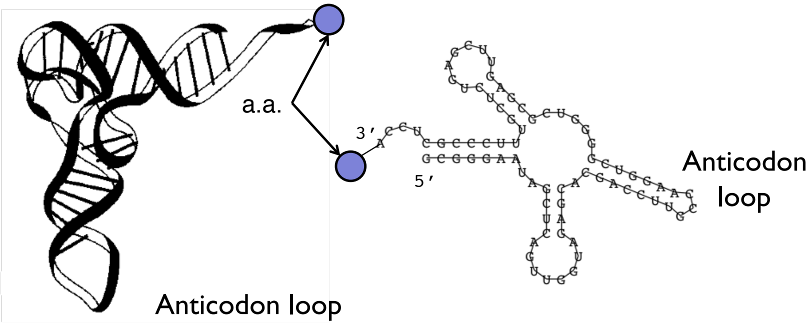
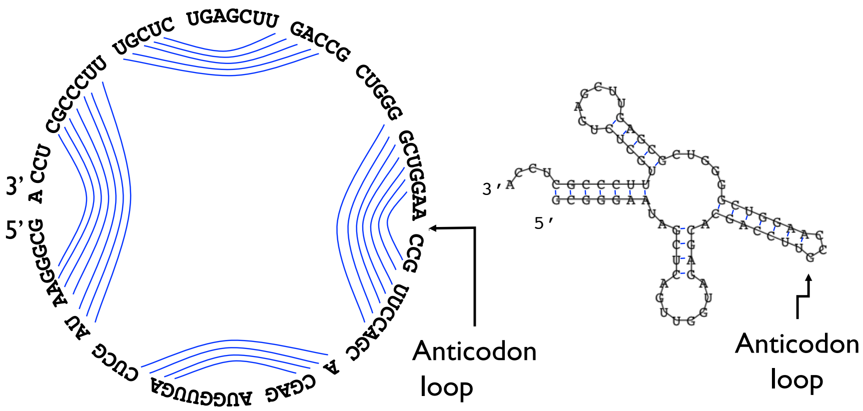


Figure 1: a) The spatial structure of the phenylalanine tRNA form yeast

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

tRNA - Alt. Representations



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

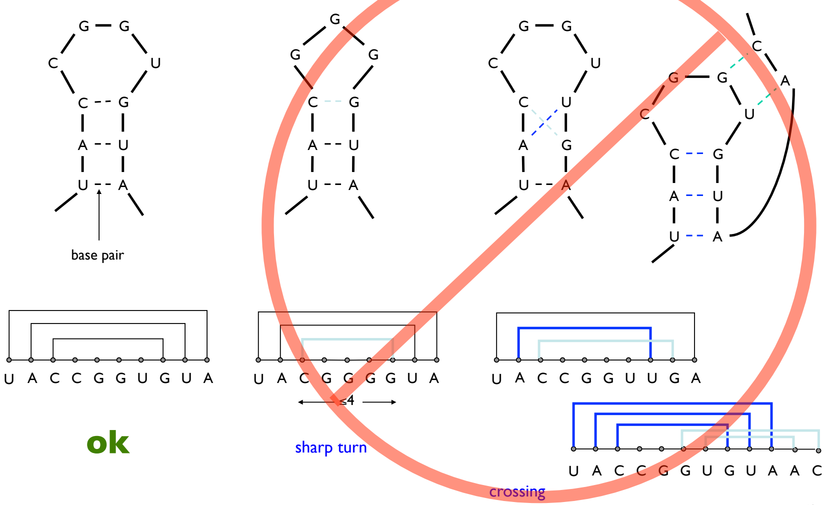
if i•j & i'•j' are two different pairs with i ≤ i', then

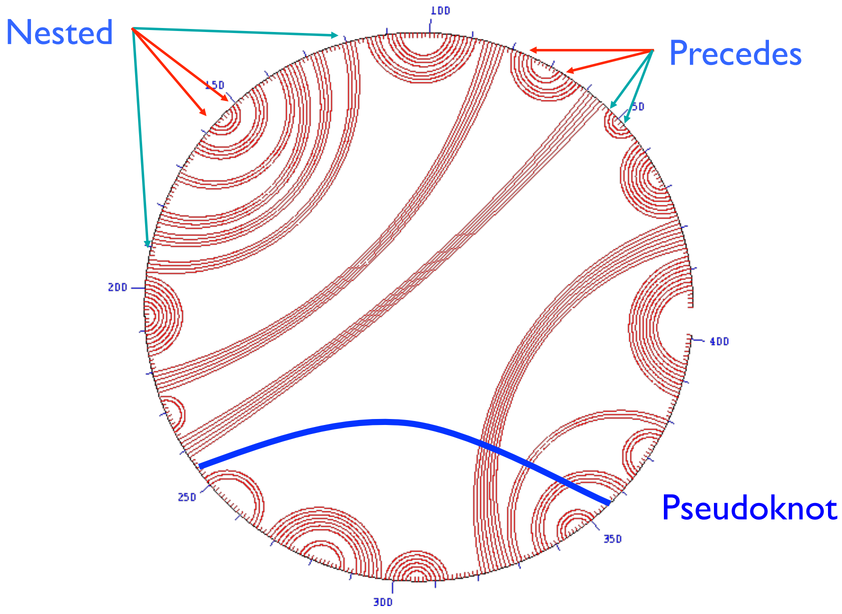
j < i', or

i < i' < j' < j

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

RNA Secondary Structure: Examples





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

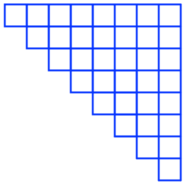
Partition Function
- + finds all folds
- - ignores pseudoknots

Nussinov: Max Pairing

B(i,j) = # pairs in optimal pairing of r_i ... r_j

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

B(i,j) = max of:



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

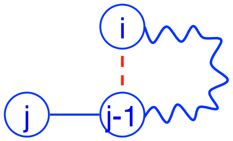
R Nussinov, AB Jacobson, "Fast algorithm for predicting the secondary structure of single-stranded RNA." PNAS 1980.

“Optimal pairing of r_i … r_j”

Two possibilities

j Unpaired:

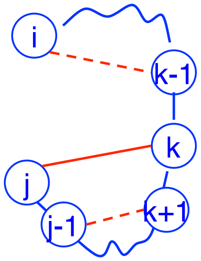
Find best pairing of r_i … r_{j-1}



j Paired (with some k):

Find best r_i … r_{k-1} +

best r_{k+1} … r_{j-1} plus 1

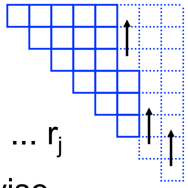


Why is it slow?

Why do pseudoknots matter?

Nussinov:

A Computation Order

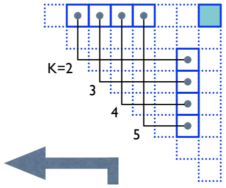


B(i,j) = # pairs in optimal pairing of r_i … r_j

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: $O(n^3)$

Which Pairs?

Usual dynamic programming “trace-back” tells you *which* base pairs are in the optimal solution, not just how many

Pair-based Energy Minimization

E(i,j) = energy of pairs in optimal pairing of r_i … 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 k-j pair

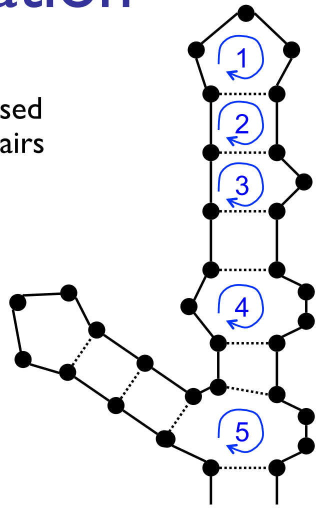
Time: $O(n^3)$

Loop-based Energy Minimization

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

Loop types

1. Hairpin loop
2. Stack
3. Bulge
4. Interior loop
5. Multiloop



Zuker: Loop-based Energy, I

$W(i,j)$ = energy of optimal pairing of $r_i \ldots r_j$

$V(i,j)$ = as above, but forcing pair $i \bullet j$

$W(i,j) = V(i,j) = \infty$ for all i, j with $i \geq j-4$

$$
\begin{aligned}
W(i,j) = \min(&W(i,j-1),\\
&\min\{\, W(i,k-1)+V(k,j) \mid i \leq k < j-4 \,\}\\
&)
\end{aligned}
$$

Zuker: Loop-based Energy, II

hairpin stack bulge/interior multi-loop

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

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

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

bulge/interior

Time: $O(n^4)$

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

Energy Parameters

Q. Where do they come from?

A1. Experiments with carefully selected synthetic RNAs

A2. Learned algorithmically from trusted alignments/structures [Andronescu et al., 2007]

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 (potentially 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 from sequence alone due, e.g., to codon structure, which RNAs lack

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

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