Modeling and Searching for Non-Coding RNA

W.L. Ruzzo
http://www.cs.washington.edu/homes/ruzzo
http://www.cs.washington.edu/homes/ruzzo/courses/gs541/10sp

GENOME 541 Syllabus

"… *protein* and *DNA* sequence analysis … to determine the "periodic table of biology," i.e., the list of *proteins* …, which can be regarded as the first stage in…"

No mention of RNA…

The Message

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

Functionally important, functionally diverse

Structurally complex

New tools required

alignment, discovery, search, scoring, etc.

Rough Outline

Today
- Noncoding RNA Examples
- RNA structure prediction

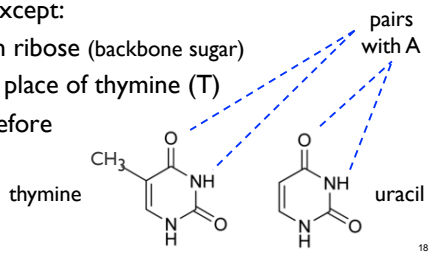
Lecture 2
- RNA "motif" models
- Search

Lecture 3
- Motif discovery
- Applications

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

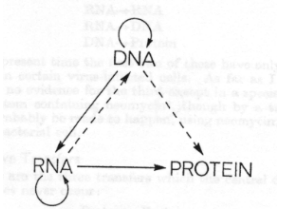
Central Dogma of Molecular Biology

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.



"Classical" RNAs

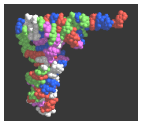
rRNA - ribosomal RNA (~4 kinds, 120-5k nt)

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



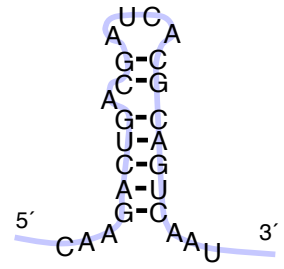
RNA Secondary Structure:
RNA makes helices too

Base pairs

A-U

C-G

U-G



Usually *single* stranded

Bacteria

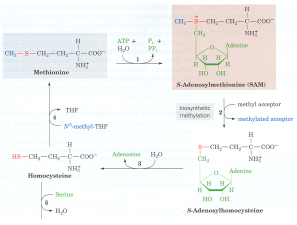
Triumph of proteins

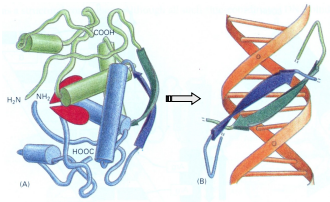
~ 80% of genome is coding DNA

Functionally diverse

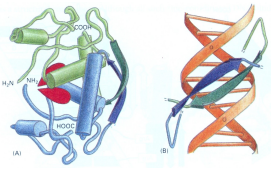
- receptors
- motors
- catalysts
- regulators (Monod & Jakob, Nobel prize 1965)
- …

Proteins catalyze & regulate biochemistry





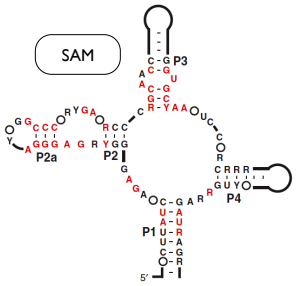
Alberts, et al, 3e.



Not the only way!

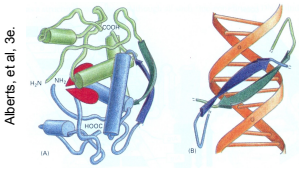
Protein way

Riboswitch alternative



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

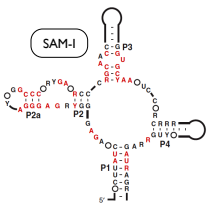
Alberts, et al, 3e.



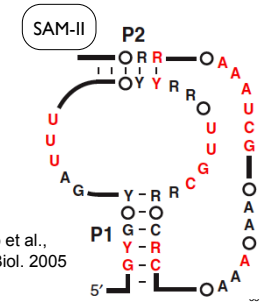
Not the only way!

Protein way

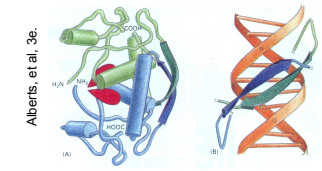
Riboswitch alternatives



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

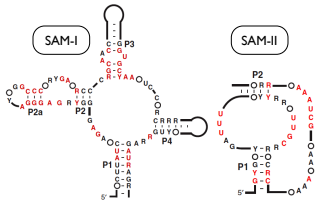


Corbino et al., Genome Biol. 2005



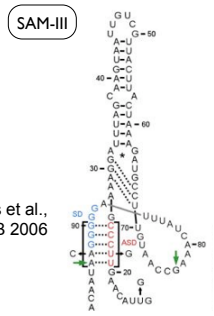
Not the only way!

Protein way ← Riboswitch alternatives

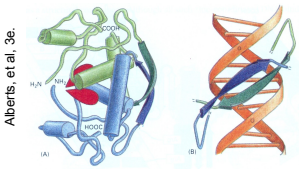


Grundy, Epshtein, Winkler et al., 1998, 2003

Corbino et al., Genome Biol. 2005

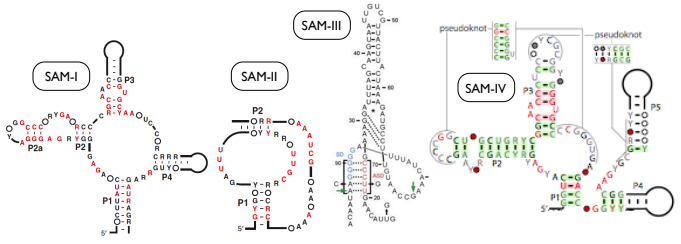


Fuchs et al., NSMB 2006



Not the only way!

Protein way ← Riboswitch alternatives

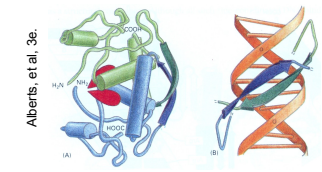


Grundy, Epshtein, Winkler et al., 1998, 2003

Corbino et al., Genome Biol. 2005

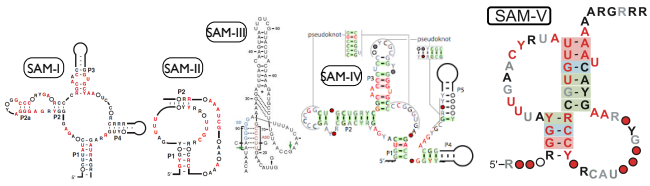
Fuchs et al., NSMB 2006

Weinberg et al., RNA 2008



Not the only way!

Protein way ← Riboswitch alternatives



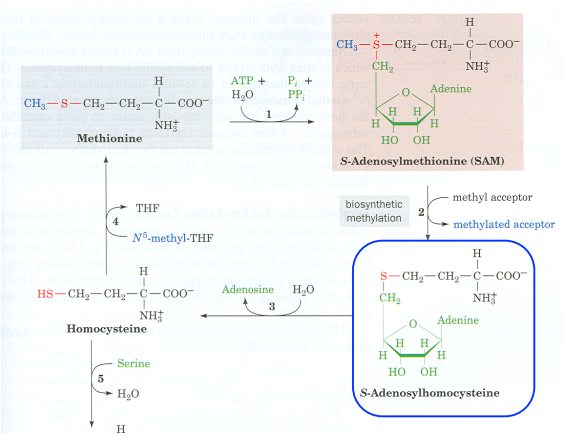
Grundy, Epshtein, Winkler et al., 1998, 2003

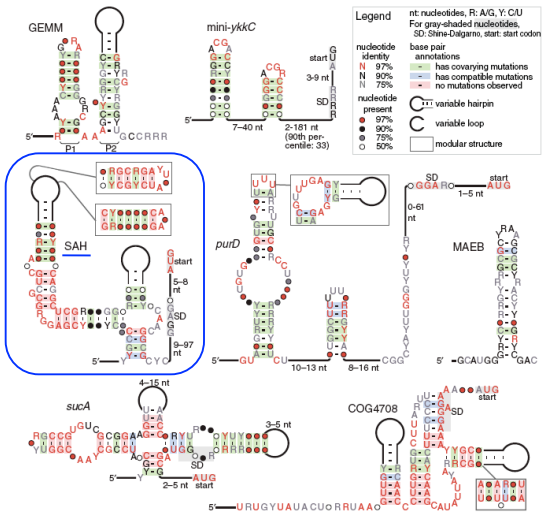
Corbino et al., Genome Biol. 2005

Fuchs et al., NSMB 2006

Weinberg et al., RNA 2008

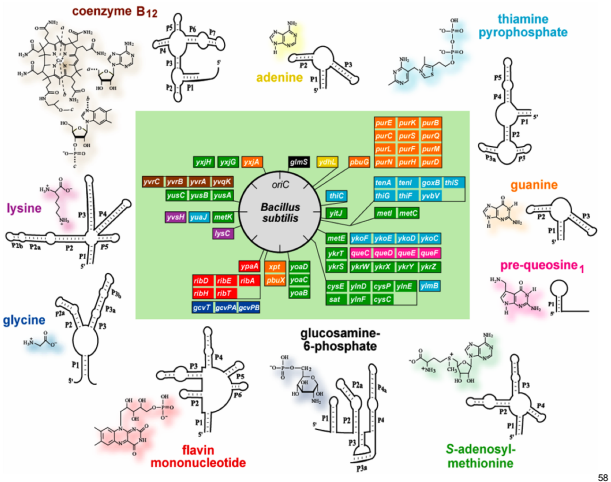
Meyer, etal., BMC Genomics 2009





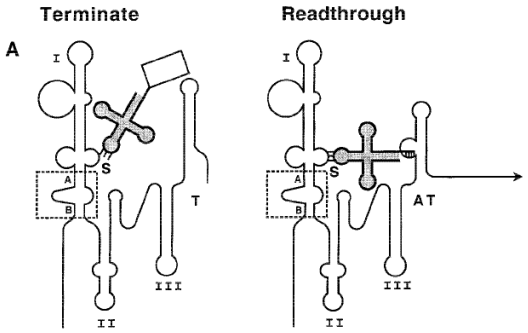
Riboswitches

- ~ 20 ligands known; multiple nonhomologous solutions for some
- dozens to hundreds of instances of each
- TPP known in archaea & eukaryotes
- one known in bacteriophage
- on/off; transcription/translation; splicing; combinatorial control
- In some bacteria, more riboregulators identified than protein TFs
- all found since ~2003



58

ncRNA Example: T-boxes



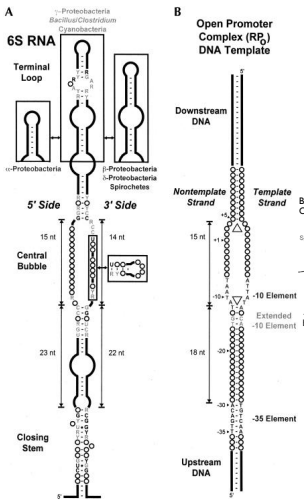
ncRNA Example: 6S

medium size (175nt)

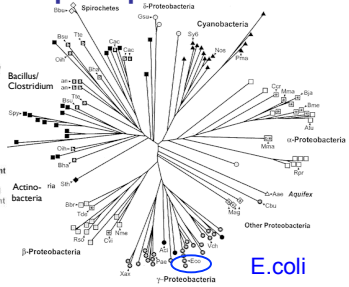
structured

highly expressed in E. coli in certain growth conditions

sequenced in 1971; function unknown for 30 years



6S mimics an open promoter



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

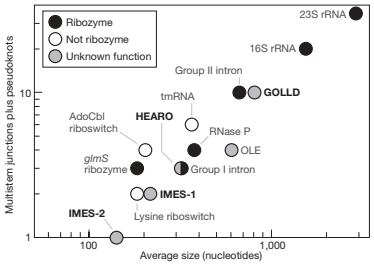
64

nature Vol 462|3 December 2009|doi:10.1038/nature08586

LETTERS

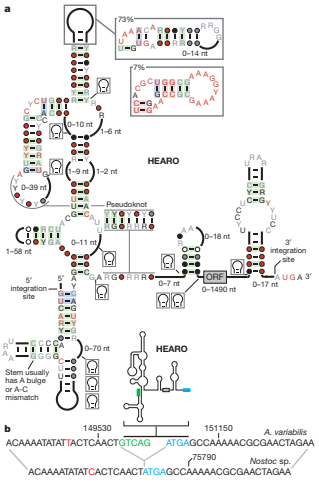
Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis

Zasha Weinberg[1,2], Jonathan Perreault[2], Michelle M. Meyer[2] & Ronald R. Breaker[1,2,3]

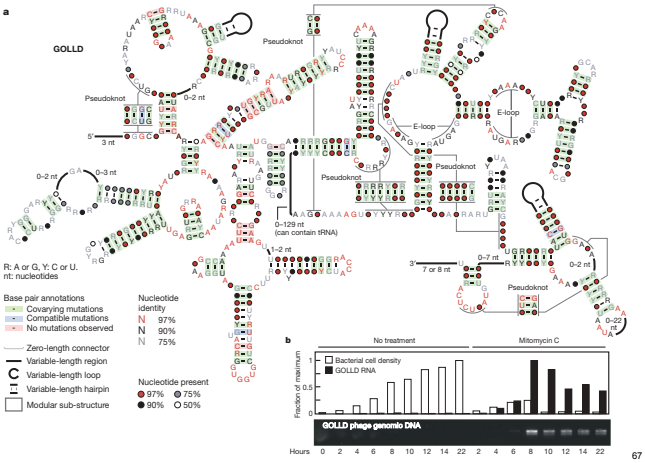


65

RNAs of unusual size and complexity



66

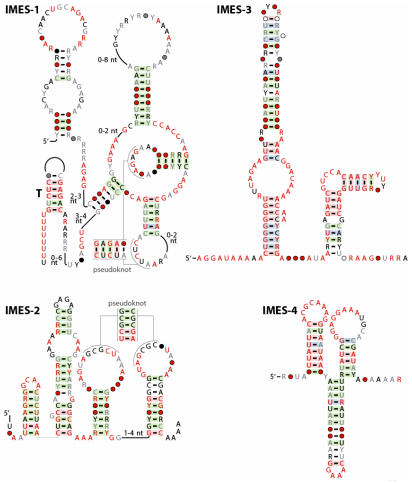


67

RNAs of unusual abundance

More abundant than 5S rRNA

From unknown marine organisms



Summary: RNA in Bacteria

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

Regulation of MANY genes involves RNA

In some species, we know identities of more ribo-regulators than protein regulators

Dozens of classes & thousands of new examples in just last 5 years

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, …

77

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,…

79

siRNA

2006 Nobel Prize
Fire & Mello

"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

Human Predictions

Evofold
S Pedersen, G Bejerano, A Siepel, K Rosenbloom, K Lindblad-Toh, ES Lander, J Kent, W Miller, D Haussler, "Identification and classification of conserved RNA secondary structures in the human genome." PLoS Comput. Biol., 2, #4 (2006) e33.
48,479 candidates (~70% FDR?)

RNAz
S Washietl, IL Hofacker, M Lukasser, A Hüttenhofer, F Stadler, "Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome." Nat. Biotech., 23, #11 (2005) 1383-90.
30,000 structured RNA elements
1,000 conserved in vertebrates.
~1/3 in introns of protein genes, ~1/6 in UTRs
~1/2 located near any known gene

FOLDALIGN
E Torarinsson, M Sawera, JH Havgaard, J Fredholm, J Gorodkin, "Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure."
Genome Research, (2006) 885-9.
1800 predicted from 36970 (of 4,000,000) pairs

CMfinder
Torarinsson, Yao, Wiklund, Bramsen, Hansen, Kjems, Tommerup, Ruzzo and Gorodkin. Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions.
Genome Research, Feb 2008, 18(2):242-251 PMID: 18096747
6500 candidates in ENCODE alone (better FDR, but still high)

Thousands of Predictions

Bottom line?

A significant number of "one-off" examples

Extremely wise-spread ncRNA expression

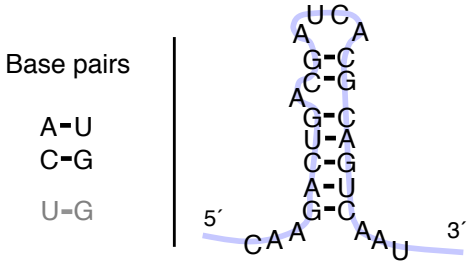
At a minimum, a vast evolutionary substrate

New technology (e.g. RNAseq) exposing more

How do you recognize an interesting one?

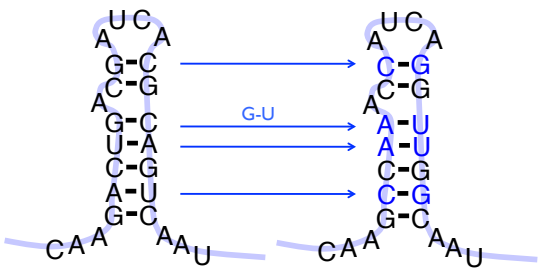
Conserved secondary structure

RNA Secondary Structure:
RNA makes helices too

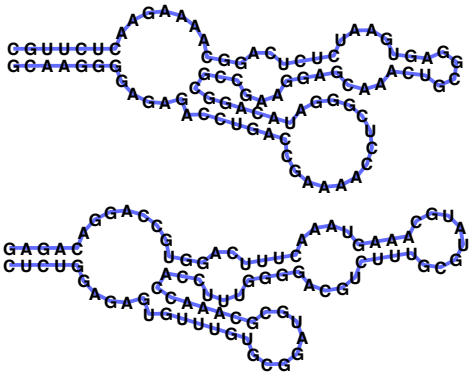


Usually *single* stranded

RNA Secondary Structure: can be fixed while sequence evolves



Why is RNA hard to deal with?



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

Structure Prediction

RNA Structure

Primary Structure: Sequence

Secondary Structure: Pairing

Tertiary Structure: 3D shape

RNA Pairing

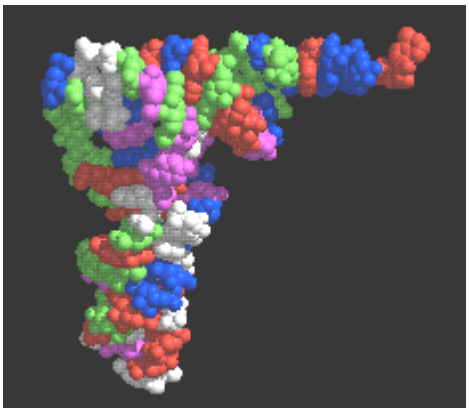
Watson-Crick Pairing

- 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

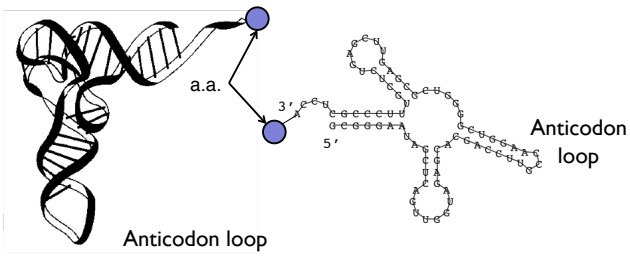
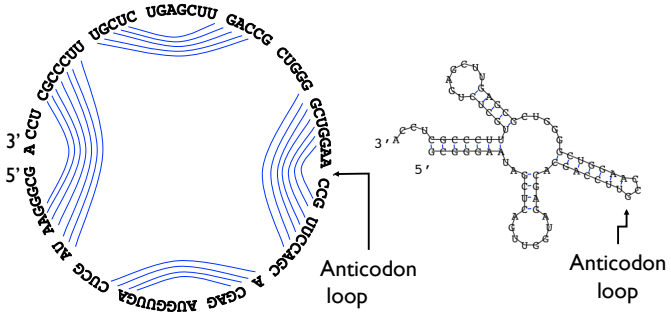


Figure 1: a) The spatial structure of the phenylalanine tRNA form yeast

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 ... r_n $^{3'}$ in {A, C, G, T}

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

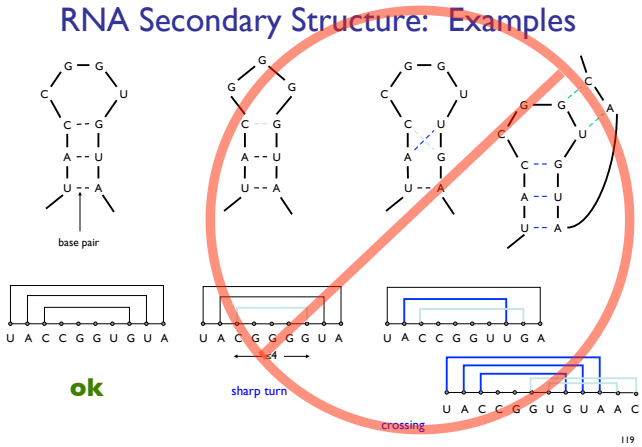
i < j-4, and } no sharp turns

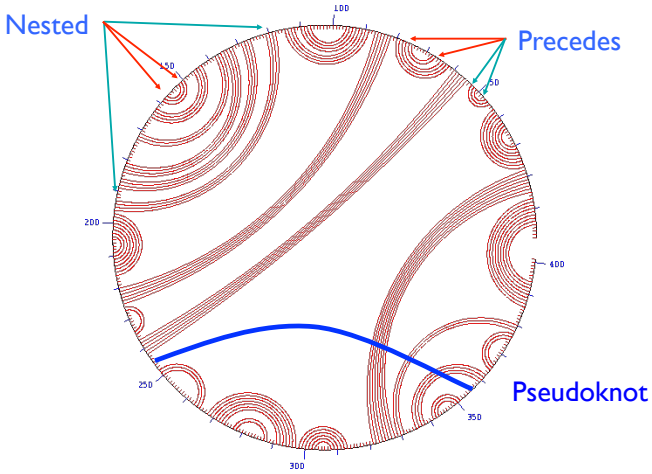
if i•j & i'•j' are two different pairs with i ≤ i', then

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

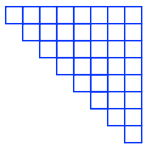
Nussinov: Max Pairing

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:

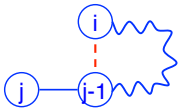
- B(i,j-1)
- max { B(i,k-1)+1+B(k+1,j-1) |
  i ≤ k < j-4 and r_k-r_j may pair}



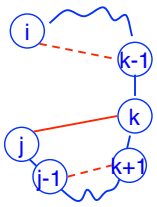
R Nussinov, AB Jacobson, "Fast algorithm for predicting the secondary structure of single-stranded RNA." PNAS 1980.

"Optimal pairing of r_i ... r_j" Two possibilities

j Unpaired:
Find best pairing of r_i ... r_{j-1}



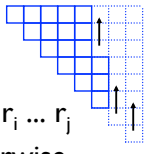
j Paired (with some k):
Find best r_i ... r_{k-1} +
best r_{k+1} ... r_{j-1} plus 1



Why is it slow?

Why do pseudoknots matter?

Nussinov: A Computation Order

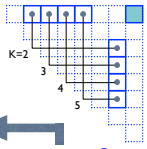


B(i,j) = # pairs in optimal pairing of r_i ... r_j

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

B(i,j) = max of:

- B(i,j-1)
- max { B(i,k-1)+1+B(k+1,j-1) | i ≤ k < j-4 and r_k-r_j may pair}



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

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

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:

- E(i,j-1)
- min { E(i,k-1) + e(r_k, r_j) + E(k+1,j-1) | i ≤ k < j-4 }

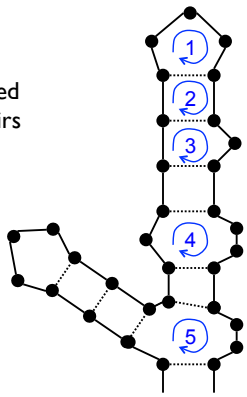
energy of k-j pair

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 ... 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(W(i,j-1),
min { W(i,k-1)+V(k,j) | i ≤ k < j-4 }
)

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"

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]

Single Seq Prediction Accuracy

Mfold, Vienna,... [Nussinov, Zuker, Hofacker, McCaskill]

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"