CSEP 527
Computational Biology

RNA: Function, Secondary Structure Prediction, Search, Discovery
Blue = ORF length threshold; Green = Markov Model threshold

Blue = ORF length-based threshold

MM-based threshold

ORF length-based threshold
CSEP 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.
Rough Outline

Today

Noncoding RNA Examples
RNA structure prediction

Next Time

RNA “motif” models
Search
Motif discovery
RNA

DNA: DeoxyriboNucleic Acid
RNA: RiboNucleic Acid

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

[Chemical structures of thymine and uracil with dashed lines indicating base pairing with A]
RNA Secondary Structure: RNA makes helices too

Base pairs

A-U
C-G
U-G

Usually *single* stranded
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 - ribosomai 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
Ribosomes

Watson, Gilman, Witkowski, & Zoller, 1992
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

Atomic structure of the 50S Subunit from *Haloarcula marismortui*. Proteins are shown in blue and the two RNA strands in orange and yellow. The small patch of green in the center of the subunit is the active site.

- Wikipedia
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).
Bacteria

Triumph of proteins
50-80% of genome is coding DNA
Functionally diverse
  receptors
  motors
  catalysts
  regulators  (Monod & Jakob, Nobel prize 1965)
...

Proteins Catalyze Biochemistry: Met Pathways

Methionine

\[ \text{CH}_3\text{SCH}_2\text{CH}_2\text{C}^-\text{COO}^- \]

\[ \text{S-Adenosylmethionine (SAM)} \]

1. ATP + P_i + H_2O \rightarrow 1

2. methyl acceptor \rightarrow methylated acceptor

3. Adenosine + H_2O \rightarrow Adenine

4. THF \rightarrow N^8\text{-methyl-THF}

5. Serine \rightarrow \text{H}_2\text{O}

\[ \text{Homocysteine} \]

\[ \text{S-Adenosylhomocysteine} \]
Proteins Regulate Biochemistry:
The MET Repressor

Protein  Alberts, et al, 3e.  DNA
Not the only way!

Protein way

Riboswitch alternative

Epshtein, et al., PNAS 2003
Winkler et al., Nat. Struct. Biol. 2003
Not the only way!

Protein way  Riboswitch alternatives

SAM-II

Corbino et al., Genome Biol. 2005

Not the only way!

Protein way

Riboswitch alternatives

SAM-III

Fuchs et al., NSMB 2006


Corbino et al., Genome Biol. 2005

Alberts, et al, 3e.

Corbino et al., Genome Biol. 2005

SAM-I

SAM-II

SAM-III
Not the only way!

Protein way

Riboswitch alternatives

Alberts, et al., 3e.

Corbino et al., Genome Biol. 2005

Weinberg et al., RNA 2008

Fuchs et al., NSMB 2006


SAM-I

SAM-II

SAM-III

SAM-IV
Alberts, et al., 3e.

Corbino et al., Genome Biol. 2005

Fuchs et al., NSMB 2006


Weinberg et al., RNA 2008

Meyer, et al., BMC Genomics 2009

Not the only way!

Protein way

Riboswitch alternatives

SAM-V

SAM-IV

SAM-III

SAM-II

SAM-I

...
And in other bacteria, a riboswitch senses SAH

\[
\text{SAH} \rightarrow \text{methylated acceptor} \rightarrow \text{methyl acceptor}
\]

\[
\text{ATP} + \text{H}_2\text{O} \rightarrow \text{PP}_i + \text{H}_3\text{PO}_4
\]
ncRNA Example: Riboswitches

UTR structure that directly senses/binds small molecules & regulates mRNA
widespread in prokaryotes
some in eukaryotes & archaea, one in a phage
~ 20 ligands known; multiple nonhomologous solutions for some
dozens to hundreds of instances of each
on/off; transcription/translation; splicing; combinatorial control
all found since ~2003; most via bioinformatics
New Antibiotic Targets?

Old drugs, new understanding:
- TPP riboswitch ~ pyrithiamine
- lysine riboswitch ~ L-aminoethylcysteine, DL-4-oxalysine
- FMN riboswitch ~ roseoflavin

Potential advantages - no (known) human riboswitches, but often multiple copies in bacteria, so potentially efficacious with few side effects?
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
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 the last ~10-15 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, …
MicroRNA

1st discovered 1992 in C. elegans
2nd discovered 2000, also C. elegans
    and human, fly, everything between – basically all multi-celled plants & animals
21-23 nucleotides
    literally fell off ends of gels
100s – 1000s now known in human
    may regulate 1/3-1/2 of all genes
    development, stem cells, cancer, infectious disease,…
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

2006 Nobel Prize
Fire & Mello
ncRNA Example: Xist

large ($\approx 12$kb)
largely unstructured RNA
required for X-inactivation in mammals
  (Remember calico cats?)

One of many thousands of “Long NonCoding RNAs” (lncRNAs) now recognized, tho most others are of completely unknown significance
Human Predictions

EvoFold
48,479 candidates (~70% FDR?)

RNAz
30,000 structured RNA elements
1,000 conserved across all vertebrates.
~1/3 in introns of known genes, ~1/6 in UTRs
~1/2 located far from any known gene

FOLDALIGN
1800 candidates from 36970 (~1/2000) pairs

CMfinder

Bottom line?

A significant number of “one-off” examples
Extremely wide-spread ncRNA expression
At a minimum, a vast evolutionary substrate
New technology (e.g., RNAseq) exposing
more

How do you recognize an interesting one?

A Clue: Conserved secondary structure
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 \( \sim 3 \text{ kcal/mole} \)
A - U \( \sim 2 \text{ kcal/mole} \)

“Wobble Pair” G - U \( \sim 1 \text{ kcal/mole} \)

Non-canonical Pairs (esp. if modified)
tRNA 3d Structure
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

Anticodon loop

Anticodon loop
Definitions

Sequence $5' r_1 r_2 r_3 ... r_n 3'$ in \{A, C, G, T/U\}

A **Secondary Structure** is a set of pairs $i\cdot j$ s.t.

- $i < j-4$, and
- no sharp turns
- if $i\cdot j$ & $i'\cdot j'$ are two different pairs with $i \leq i'$, then
- $j < i'$, or
- $i < i' < j' < j$
- 2nd pair follows 1st, or is nested within it;
- no “pseudoknots”
- And pairs, not triples, etc.
RNA Secondary Structure: Examples

- base pair
- sharp turn
- crossing
- ok
Approaches to Structure Prediction

Maximun 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
“Optimal pairing of $r_i \ldots r_j$”
Two possibilities

j Unpaired:
Find best pairing of $r_i \ldots r_{j-1}$

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

Why is it slow?
Why do pseudoknots matter?
Nussinov: Max Pairing

\[ B(i,j) = \# \text{ pairs in optimal pairing of } r_i \ldots r_j \]
\[ B(i,j) = 0 \text{ for all } i, j \text{ with } i \geq j-4; \]
Otherwise
\[ B(i,j) = \max \text{ of:} \]
\[ \begin{cases} B(i,j-1) \\ \max \{ B(i,k-1) + 1 + B(k+1,j-1) \mid i \leq k < j-4 \text{ and } r_k - r_j \text{ may pair} \} \end{cases} \]

Nussinov: 
A Computation Order

\[ B(i,j) = \# \text{ pairs in optimal pairing of } r_i \ldots r_j \]

\[ B(i,j) = 0 \text{ for all } i, j \text{ with } i \geq j-4; \text{ otherwise} \]

\[ B(i,j) = \max \text{ of:} \]

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

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
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

\[ E(i,j) = \text{energy of pairs in optimal pairing of } r_i \ldots r_j \]

\[ E(i,j) = \infty \text{ for all } i, j \text{ with } i \geq j-4; \text{ otherwise} \]

\[ E(i,j) = \min \{ E(i,j-1), \min \{ E(i,k-1) + e(r_k, r_j) + E(k+1,j-1) \mid i \leq k < j-4 \} \} \]

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

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

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

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

$W(i,j) = \min(W(i,j-1),$

$\min \{ W(i,k-1) + V(k,j) \mid i \leq k < j-4 \})$
Zuker: Loop-based Energy, II

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

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

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

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]

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 crystallography, NMR)
Two competing secondary structures for the Leptomonas collosoma spliced leader mRNA.
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”