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
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
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.
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
80% of genome is coding DNA
Functionally diverse
  receptors
  motors
  catalysts
  regulators (Monod & Jakob, Nobel prize 1965)
...
Proteins Catalyze Biochemistry: Met Pathways

\[
\text{Methionine} \xrightarrow{\text{1}} S\text{-Adenosylmethionine (SAM)}
\]

\[
\text{Homocysteine} \xrightarrow{\text{3}} S\text{-Adenosylhomocysteine}
\]

\[
\text{H}_{2}O \xrightarrow{\text{5}} \text{Serine}
\]

\[
\text{ATP} + P_{i} + H_{2}O \rightarrow \text{1}
\]

\[
\text{biosynthetic methylation}
\]

\[
\text{methylation acceptor} \xrightarrow{\text{2}} \text{methylated acceptor}
\]
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
Not the only way!

Protein way  Riboswitch alternatives

Corbino et al., Genome Biol. 2005
Fuchs et al., NSMB 2006
Weinberg et al., RNA 2008
Not the only way!

Protein way

Riboswitch alternatives

Corbino et al., Genome Biol. 2005
Fuchs et al., NSMB 2006
Weinberg et al., RNA 2008
Meyer, et al., BMC Genomics 2009
And in other bacteria, a riboswitch senses SAH.
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
Chloroflexus aurantiacus
Geobacter metallireducens
Geobacter sulphurreducens

Chloroflexi
 δ-Proteobacteria

γ-Proteobacteria
β-Proteobacteria
α-Proteobacteria
ε-Proteobacteria

Spirochaetes
Chlamydiae
Actinobacteria
(high GC)
Cyanobacteria

Firmicutes
(low GC)

Used by CMfinder
Found by scan
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

E.coli
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 5-10 years
Vertebrate ncRNAs

mRNA, tRNA, rRNA, … of course

PLUS:

snRNA, spliceosome, snoRNA, telomerase, 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
Hundreds now known in human
    may regulate 1/3-1/2 of all genes
    development, stem cells, cancer, infectious disease,
    …
“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 Example: Xist

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

E Torarinsson, M Sawera, JH Havgaard, M Fredholm, J Gorodkin, "Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure."


1800 candidates from 36970 (of 100,000) pairs

CMfinder

Torarinsson, Yao, Wiklund, Bramsen, Hansen, Kjems, Tommerup, Ruzzo and Gorodkin. Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions.


6500 candidates in ENCODE alone (better FDR, but still high)
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 ~ 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/U\}

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

\[
i < j - 4, \text{ and } \quad \begin{cases}
\text{no sharp turns} \\
\text{if } i \cdot j \text{ & } i' \cdot j' \text{ are two different pairs with } i \leq i', \text{ then}
\end{cases}
\]

\[
\begin{align*}
&j < i', \text{ or} \\
&i < i' < j' < j
\end{align*}
\]

\[
\begin{cases}
\text{2nd pair follows 1st, or is} \\
\text{nested within it;} \\
\text{no “pseudoknots.”}
\end{cases}
\]
RNA Secondary Structure: Examples

- base pair
- sharp turn
- crossing
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

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

“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:
A Computation Order

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

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

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

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

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
Computing one cell: $\text{OPT}[2,18] = ?$

$$\text{OPT}(i,j) = \begin{cases} 0 & \text{if } i \geq j - 4 \\ \max\left\{ \text{OPT}[i, j-1], 1 + \max_i \left( \text{OPT}[i, t-1] + \text{OPT}[t+1, j-1] \right) \right\} & \text{otherwise} \end{cases}$$

Case 1:

$2 \geq 18 - 4$? no.

Case 2:

$B_{18}$ unpaired? Always a possibility; then $\text{OPT}[2,18] \geq 3$

GGAAAACCCCAAAAGGGGU
((....))(....)....
Computing one cell: $\text{OPT}[2,18] = ?$

Case 3, $2 \leq t < 18 - 4$:
- $t = 2$: no pair

$\text{OPT}(i, j) = \begin{cases} 
0 & \text{if } i \geq j - 4 \\
\max \left\{ \text{OPT}[i, j - 1], 1 + \max_t (\text{OPT}[i, t - 1] + \text{OPT}[t + 1, j - 1]) \right\} & \text{otherwise}
\end{cases}$
Computing one cell: $\text{OPT}[2,18] = ?$

$$\text{OPT}(i,j) = \begin{cases} 
0 & \text{if } i \geq j - 4 \\
\max \left\{ \text{OPT}[i, j - 1], 1 + \max_i (\text{OPT}[i, t - 1] + \text{OPT}[t + 1, j - 1]) \right\} & \text{otherwise}
\end{cases}$$

Case 3, $2 \leq t < 18-4$: $t = 3$: no pair
Computing one cell: $\text{OPT}[2,18] = ?$

Case 3, $2 \leq t < 18 - 4$:

$t = 4$: yes pair

$\text{OPT}[2,18] \geq 1 + 0 + 3$

$\text{OPT}(i,j) = \begin{cases} 
0 & \text{if } i \geq j - 4 \\
\max \left\{ \text{OPT}[i, j - 1], 1 + \max_t (\text{OPT}[i, t - 1] + \text{OPT}[t + 1, j - 1]) \right\} & \text{otherwise}
\end{cases}$
Computing one cell: \( \text{OPT}[2, 18] = ? \)

\[
\text{OPT}(i, j) = \begin{cases} 
0 & \text{if } i \geq j - 4 \\
\max \left\{ \text{OPT}[i, j - 1], 1 + \max_t (\text{OPT}[i, t - 1] + \text{OPT}[t + 1, j - 1]) \right\} & \text{otherwise}
\end{cases}
\]

Case 3, \(2 \leq t < 18-4\):
\( t = 5 \): yes pair
\( \text{OPT}[2, 18] \geq 1 + 0 + 3 \)

\( \text{GGAAACCCAAAGGGGU} \)
\( \cdots (\cdots ((\cdots )))) \)
Computing one cell: $OPT[2,18] = ?$

$OPT(i,j) = \begin{cases} 
0 & \text{if } i \geq j - 4 \\
\max \left\{ \begin{array}{c}
OPT[i, j-1] \\
1 + \max_t (OPT[i, t-1] + OPT[t+1, j-1])
\end{array} \right. & \text{otherwise}
\end{cases}$

Case 3, $2 \leq t < 18-4$:
$t = 6$: yes pair
$OPT[2,18] \geq 1 + 0 + 3$

GGAAACCACAAGGGU

\[ \ldots (.((((\ldots)))) \]
Computing one cell: $\text{OPT}[2,18] = ?$

$$\text{OPT}(i,j) = \begin{cases} 
0 & \text{if } i \geq j - 4 \\
\max \left\{ \text{OPT}[i,j-1], \max_{t} (\text{OPT}[i,t-1] + \text{OPT}[t+1,j-1]) \right\} & \text{otherwise}
\end{cases}$$
Computing one cell: $\text{OPT}[2,18] = ?$

$\text{OPT}(i,j) = \begin{cases} 
0, & \text{if } i \geq j - 4 \\
\max \left\{ \begin{array}{l}
\text{OPT}[i,j-1] \\
1 + \max_t (\text{OPT}[i,t-1] + \text{OPT}[t+1,j-1])
\end{array} \right\}, & \text{otherwise}
\end{cases}$

Case 3, $2 \leq t < 18-4$:
$t = 8$: no pair
Computing one cell: \( \text{OPT}[2,18] = ? \)

\[
\begin{align*}
\text{OPT}(i,j) &= \begin{cases} 
0 & \text{if } i \geq j - 4 \\
\max \left\{ \begin{array}{l}
\text{OPT}[i,j - 1] \\
1 + \max_t (\text{OPT}[i,t - 1] + \text{OPT}[t + 1, j - 1])
\end{array} \right\} & \text{otherwise}
\end{cases}
\end{align*}
\]

Case 3, \( 2 \leq t < 18-4: \)
\( t = 11: \) yes pair
\( \text{OPT}[2,18] \geq 1 + 2 + 0 \)

GGAAACCCAAAGGGGU
((......))(......)

(not shown: \( t=9,10,12,13 \))
Overall, Max = 4
several ways, e.g.:
GGAAAACCCAAAGGGGU
..(...(((....))))

tree shows trace back:
square = case 3
octagon = case 1

\[
\text{OPT}(i, j) = \begin{cases} 
0 & \text{if } i \geq j - 4 \\
\max \left\{ \text{OPT}[i, j-1], 1 + \max_t (\text{OPT}[i, t-1] + \text{OPT}[t+1, j-1]) \right\} & \text{otherwise}
\end{cases}
\]
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

\[ E(i,j) = \text{energy of pairs in optimal pairing of } r_i \ldots r_j \]
\[ E(i,j) = \infty \text{ for all } i, j \text{ with } i \geq j-4; \text{ otherwise} \]
\[ E(i,j) = \min \text{ of:} \]
\[ \begin{cases} E(i,j-1) & \text{energy of } k-j \text{ pair} \\ \min \{ E(i,k-1) + e(r_k, r_j) + E(k+1,j-1) | i \leq k < j-4 \} & \end{cases} \]

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, 1

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

\[ V(i,j) = \text{as above, but forcing pair } i \cdot j \]

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

\[ W(i,j) = \min( W(i,j-1), \min \{ W(i,k-1) + V(k,j) \mid i \leq k < j-4 \} ) \]
V(i,j) = min(\(e_h(i,j)\), \(e_s(i,j) + V(i+1,j-1)\), \(V_{BI}(i,j)\), \(V_{M}(i,j)\))

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

\(V_{BI}(i,j) = \min \{ e_{bi}(i,j,i',j') + V(i', j') \mid i < i' < j' < j \& i'-i+j-j' > 2 \} \)

Time: \(O(n^4)\)

\(O(n^3)\) possible if \(e_{bi}(.)\) 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)
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”