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

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

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

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):

Gene Regulation: The Met Repressor

Mandal et al. Science 2004
Two SAM Riboswitches

6S mimics an open promoter

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 \( \sim 3 \text{ kcal/mole} \)
- A - U \( \sim 2 \text{ kcal/mole} \)
- “Wobble Pair” G - U \( \sim 1 \text{ kcal/mole} \)

*Non-canonical Pairs* (esp. if modified)

---

**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 \( i\text{ }j \) s.t.

\[
\begin{align*}
&\text{i }< \text{j-4, and i' }< \text{j'} < \text{j} \quad \text{no sharp turns} \\
&\text{if } i\text{ }j \text{ & i' }j' \text{ are two different pairs with } i \leq i', \text{ then} \\
&\quad \begin{cases} 
\text{j }< \text{i'}, \text{ or } \text{i' }< \text{j'} < \text{j} & \text{2nd pair follows 1st, or} \\
\text{i' }< \text{j'} < \text{j} & \text{is nested within it; no "pseudoknots."}
\end{cases}
\end{align*}
\]
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 crystallography, NMR)
Nussinov: Max Pairing

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

Time: \( O(n^3) \)

“Optimal pairing of \( r_i \ldots r_j \)”

Two possibilities

J Unpaired:
Find best pairing of \( r_i \ldots r_{j-1} \)

J Paired:
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?

Pair-based Energy Minimization

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

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
- Hairpin loop
- Stack
- Bulge
- Interior loop
- Multiloop
**Base Pairs and Stacking**

- uracil
- thymine
- cytosine
- guanine
- adenine

**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-

**Loop Examples**

**Zuker: Loop-based Energy, I**

- \( W(i,j) = \text{energy of optimal pairing of } r_i \ldots r_j \)
- \( V(i,j) = \text{as above, but forcing pair } i^*j \)
- \( W(i,j) = V(i,j) = \infty \text{ for all } i, j \text{ with } i \geq j-4 \)
- \( W(i,j) = \min(W(i,j-1), \min \{ W(i,k-1) + V(k,j) \mid i \leq k < j-4 \}) \)
Zuker: Loop-based Energy, II

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

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

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

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

\( O(n^3) \) possible if \( \text{ebi}(.) \) 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.

Example of suboptimal folding

Black dots: pairs in opt fold

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

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

Latest estimates suggest ~50-75% of base pairs predicted correctly in sequences of up to ~300nt. Definitely useful, but obviously imperfect.

Task 2: Motif

Description

How to model an RNA “Motif”? Conceptually, start with a profile HMM:
from a multiple alignment, estimate nucleotide/insert/delete
preferences for each position
given a new seq, estimate likelihood that it could be generated by
the model, & align it to the model

mostly G del ins all G

How to model an RNA “Motif”? Add “column pairs” and pair emission probabilities
for base-paired regions
RNA Motif Models

“Covariance Models” (Eddy & Durbin 1994)
aka profile stochastic context-free grammars
aka hidden Markov models on steroids
Model position-specific nucleotide preferences and base-pair preferences

Pro: accurate
Con: model building hard, search slow

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
RNA “motifs” (seq + 2-ary struct) well-captured by “covariance models”