RNA Secondary Structure

CSE 417
W.L. Ruzzo

The Double Helix

The “Central Dogma” of Molecular Biology

DNA → RNA → Protein

Non-coding RNA

- Messenger RNA - codes for proteins
- Non-coding RNA - all the rest
  - Before, say, mid 1990’s, 1-2 dozen known (critically important, but narrow roles: e.g. ribosomal and transfer RNA, splicing, SRP)
  - Since mid 90’s dramatic discoveries
    - Regulation, transport, stability/degradation
    - E.g. “microRNA”: hundreds in humans
    - E.g. “riboswitches”: thousands in bacteria
DNA structure: dull

...ACCGCTAGATG...

...TGGCGATCTAC...

RNA Structure: Rich

• RNA’s fold, and function
• Nature uses what works

Why is structure Important?

• For protein-coding, similarity in sequence is a powerful tool for finding related sequences
  – e.g. “hemoglobin” is easily recognized in all vertebrates
• For non-coding RNA, many different sequences have the same structure, and structure is most important for function.
  – So, using structure plus sequence, can find related sequences at much greater evolutionary distances

Q: What’s so hard?

A: Structure often more important than sequence
6S mimics an open promoter

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

“Central Dogma” = “Central Chicken & Egg”?

Was there once an “RNA World”?

**DNA** → **RNA** → **Protein**
6.5 RNA Secondary Structure

RNA. String $B = b_1 b_2 ... b_n$ over alphabet $\{A, C, G, U\}$.

Secondary structure. RNA is single-stranded so it tends to loop back and form base pairs with itself. This structure is essential for understanding behavior of molecule.

Examples.

Free energy. Usual hypothesis is that an RNA molecule will form the secondary structure with the optimum total free energy, approximately by number of base pairs.

Goal. Given an RNA molecule $B = b_1 b_2 ... b_n$, find a secondary structure $S$ that maximizes the number of base pairs.
RNA Secondary Structure: Subproblems

First attempt. $OPT(j)$ = maximum number of base pairs in a secondary structure of the substring $b_1 b_2 ... b_j$.

**Difficulty.** Results in two sub-problems.

- Finding secondary structure in: $b_1 b_2 ... b_{t-1}$.
- Finding secondary structure in: $b_t b_{t+1} ... b_{n-1}$.

Dynamic Programming Over Intervals

Notation. $OPT(i, j)$ = maximum number of base pairs in a secondary structure of the substring $b_i b_{i+1} ... b_j$.

- Case 1. If $i ≥ j - 4$.
  - $OPT(i, j) = 0$ by no-sharp turns condition.
- Case 2. Base $b_j$ is not involved in a pair.
  - $OPT(i, j) = OPT(i, j-1)$
- Case 3. Base $b_j$ pairs with $b_t$ for some $i ≤ t < j - 4$.
  - non-crossing constraint decouples resulting sub-problems
  - $OPT(i, j) = 1 + \max_t \{ OPT(i, t-1) + OPT(t+1, j-1) \}$

**Remark.** Same core idea in CKY algorithm to parse context-free grammars.

Bottom Up Dynamic Programming Over Intervals

Q. What order to solve the sub-problems?
A. Do shortest intervals first.

```c
// RNA(b_1, ..., b_n) {
for k = 5, 6, ..., n-1
  for i = 1, 2, ..., n-k
    j = i + k
    Compute M[i, j]
  
return M[1, n] using recurrence
}
```

Running time. $O(n^3)$.

E.g.: $OPT(1, 6) = 1$: `CUCCGG`  
E.g.: $OPT(6, 16) = 2$: `GUUGCAAUGUC`