CSE280B Basics & Non-coding RNA

Vineet Bafna
Pavel A. Pevzner
Topics in Bioinformatics

- This class has few required homeworks.
- All reading is optional, but recommended.
  - Some background in algorithms is helpful.
- 1 Final Exam, and 1 research project
- Goal:
  - Pose research problems with minimum preparation
- Lectures will be given by instructors and by students
The scope/syllabus
Life begins with Cell

- A cell is the smallest structural unit of an organism that is capable of independent functioning.
- All cells have some common features.
All Life depends on 3 critical molecules

- **Protein**
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body’s major components (e.g. hair, skin, etc.)

- **DNA**
  - Holds all information about the cell.

- **RNA**
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
The molecules of life and Bioinformatics

- DNA, RNA, and Proteins can all be represented as strings!
- DNA/RNA are string over a 4 letter alphabet (A,C,G,T/U).
- Protein Sequences are strings over a 20 letter alphabet.
- This allows us to store and query them as text.
The Central Dogma

DNA -> RNA -> Protein

- DNA can replicate.
- Information coded in the sequence of base pairs in DNA is passed to molecules of RNA.
- Information in RNA is passed to proteins. It never passes from proteins to nucleic acids.
Transcription

- Transcription is the process of 'transcribing' or copying a gene from DNA to RNA.
Translation

- The transcript goes outside the nucleus and is translated into a protein.
- Therefore, the consequence of a change in the environment of a cell is a change in transcription, or a change in translation.
**The Central Dogma**

- However, not all genes are translated!
Example: tRNA
Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
Topic I

- Could it be that, there is a large trove of non-coding RNA, that is not discovered yet?
  - The answer seems to be YES

- How can we look for ncRNA?
  - Will be covered in the first few lectures
mRNA

- Q: Cells in different organs behave differently. How?

- While non-coding RNA is important, a large fraction of the cellular RNA is messenger RNA (mRNA), encoding for a protein.
- The transcription process is regulated in many genes.
Non-topics

- The amount of mRNA produced is regulated by proteins (transcription factors) that bind upstream of the gene.
  - How do we find these regions?
- Quantitative measurement of mRNA can be done using gene chips.
What regulates the transcription factors?

- A signalling cascade of proteins regulates transcription factors (proteins).
- How can we identify the active proteins?
- How can we identify signalling pathways?
Topic II

- Mass Spectrometry is a key technology for measuring active proteins, their interactions, and their post-translational modifications.
- Computation plays a key role in interpreting mass spectrometry data.
Back to the genome

- Recall that
Novel ncRNAs are abundant: Ex: miRNAs

- miRNAs were the second major story in 2001 (after the genome).
- Subsequently, many other non-coding genes have been found.
ncRNA gene finding

- Possible that many undiscovered ncRNA exist, and that RNA are as important as protein coding genes.
- Computational methods for discovering ncRNA are not mature.
- What are the clues to non-coding genes?
  - Look for signals selecting start of transcription and translation. Non-coding genes are transcribed by Pol III
  - Non-coding genes have structure. Look for genomic sequences that fold into an RNA structure
- Structure: Given a sequence, what is the structure into which it can fold with minimum energy?
RNA structure: Basics

- Key: RNA is single-stranded. Think of a string over 4 letters, A,C,G, and U.
- The complementary bases form pairs.
- Base-pairing defines a secondary structure. The base-pairing is usually non-crossing.
RNA structure: pseudoknots

- Sometimes, unpaired bases in loops form ‘crossing pairs’. These are pseudoknots.

**Figure 1.** A simple pseudoknot. In a pseudoknot, nucleotides inside a hairpin loop pair with nucleotides outside the stem-loop.
RNA structure prediction

- Any set of non-crossing base-pairs defines a secondary structure.
- Abstract Question:
  - Given an RNA string find a structure that maximizes the number of non-crossing base-pairs
  - Incorporate the true energetics of folding
  - Incorporate Pseudo-knots
ncRNA discovery

- Q: Given genomic DNA, discover all regions likely to be ncRNA
- ncRNA (unlike other DNA) should have secondary structure
  - Approach: Find all substrings that fold into a low energy structure.
Unfortunately...

Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs

Elena Rivas and Sean R. Eddy*

Department of Genetics, Washington University, St. Louis, MO 63110, USA

Received on August 4, 1999; revised on December 15, 1999; accepted on December 21, 1999

- Random DNA (with high GC content) often folds into low-energy structures.
- What other signals determine non-coding genes?
Discovering ncRNA

1. Consider each ncRNA family separately. Compute features that are distinct from other sequences.
ncRNA: miRNA

- ncRNA ~22 nt in length
- Pairs to sites within the 3’ UTR, specifying translational repression.
- Similar to siRNA (involved in RNAi)
- Unlike siRNA, miRNA do not need perfect base complementarity
- Until recently, no computational techniques to predict miRNA
  - Most predictions based on cloning small RNAs from size fractionated samples
• Given a pair of conserved sequences, are they conserved because they encode ncRNA?

• Q: How would you compute such conserved pairs in the first place?
Comparative Approach to discovering ncRNA

• Given a query ncRNA (sequence & structure), compute all homologs that are similar in sequence and structure.

• How can you do it efficiently?
A combinatorial problem

- **Input:**
  - A string over A,C,G,U
  - A pairs with U, C pairs with G

- **Output:**
  - A subset of possible base-pairs of maximum size such that
    - No two base-pairs intersect

- **How can we compute this set efficiently?**
Nussinov’s algorithm

2. Let $W(i, j)$ be the score of the best structure of the subsequence from $i$ to $j$.

$$
\begin{align*}
\text{for } i = n \text{ down to 1} & \{ \\
& \text{ for } j = i+1 \text{ to } n \{ \\
& W(i, j) = \max \left\{ B(r_i, r_j) + W(i+1, j-1), \\
& \quad \quad W(i, j-1), \\
& \quad \quad W(i+1, j) \\
& \quad \quad W(i, k) + W(k+1, j) \quad i \leq k < j \right\} \\
& \} \\
& \} \\
\end{align*}
$$
for i = n downto 1  
    for j = i+1 to n  
    
    \[ W(i, j) = \max \begin{cases} 
        B(r_i, r_j) + W(i + 1, j - 1), \\
        W(i, j - 1), \\
        W(i + 1, j), \\
        W(i, k) + W(k + 1, j) 
    \end{cases} \]  

if (1) { 
    S(i, j) = / 
} else if (2) 
    S(i, j) = | 
else if(3) 
    S(i, j) = - 
else 
    S(i, j) = k
Procedure print_RNA(i,j) {
    if S(i,j) = / {
        print "(i,j)";
        print_RNA(i+1,j-1);
    } else if (S(i,j) = -) {
        print_RNA(i+1,j);
    } else if (S(i,j) = |) {
        print_RNA(i,j-1);
    } else {
        k=S(i,j)
        print_RNA(i,k);
        print_RNA(k+1,j);
    }
}
RNA structure: example

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
RNA Structure: Details
Base-pairing & Loops

- Base-pairs arise from complementary nucleotides
- Single-stranded
- Stack is when 2 base-pairs are contiguous
- Loops arise when there are unpaired bases.
- They are characterized by the number of base-pairs that close it.
  - Hairpin: closed by 1 base-pair
  - Bulge/Interior Loops (2 base-pairs)
  - Multiple Internal loops (k base-pairs)
Scoring Loops, multi-loops

- Zuker-Turner Energy Rules
  - http://www.bioinfo.rpi.edu/~zukerm/rna/energy/node2.html
- Stacking Energies
- Energy for Bulges and Interior Loops
- Energy for Multi-loops
Other tricks for obtaining structure

- Alignment and Covariance

![Diagram showing RNA structures and alignment scores]

**Figure 1.** Sequence Alignment Scoring versus Structural Alignment Scoring
RNA: unsolved problems

- The structure problem is still unsolved.
  - De novo prediction does not work as well.
  - Co-variance models require prior alignment.

- Many undiscovered non-coding genes
  - miRNA, and others have only just been discovered.
  - Very hard to detect signal for these genes
  - Random sequence folds into low energy structures.