CENG 734
Advanced Topics in Bioinformatics

Week 3
RNA-Seq

Fall 2010-2011
Quiz #1

- Construct a de Bruijn graph for the following 4 reads using 3-mers.
  - AGTTCA
  - ATCAGT
  - GTACAG
  - CAGTTT

- Simplify linear paths to single nodes

- Identify tips and bulges

- Do not create reverse complement (i.e., twin) nodes.
From last week

- A consistency-based consensus algorithm for de novo and reference-guided sequence assembly of short reads
- Assembly based on multiple-sequence alignment (utilizes consistency and uses progressive alignment)
- Alignment guided by the alignment graph of short-reads
Alignment graph

- Nodes: sequence segments, i.e. substrings of short sequence reads!
- The segments partition the sequence read
- Edges only between segments belonging to different reads, i.e., an $n$-partite graph if we have $n$ reads. The number of vertices may be larger than $n$.
- The score of aligning the segments are used as edge weights.
- A subset of the edges may represent an alignment: the maximum trace (NP-hard)
Multi-read alignment algorithm

- **Step 1:** compute alignments between reads
  - Use dynamic programming
  - Speed-up using predicted read positions

- **Step 2:** construct the alignment graph
  - By segmenting the alignments of reads into ungapped alignments

- **Step 3:** consistency extension
  - Use triplet extension similar to T-Coffee

- **Step 4:** graph-based progressive alignment
  - Progressive alignment guided by UPGMA tree
RNA-Seq

- Sequencing transcripts
- May align the reads
  - To a genome
    - Confirm introns, exons etc.
    - Gene identification
  - To an mRNA library
  - De novo assembly
- Many applications
- Advantages over microarray
RNA-Seq

- Problem definition
Isoforms
Single reads
Paired end reads
## Statistical Modeling

### Terms

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I$</td>
<td>Total number of different types of transcripts in the sample.</td>
</tr>
<tr>
<td>$J$</td>
<td>Total number of different types of reads.</td>
</tr>
<tr>
<td>$\theta_i$</td>
<td>The abundance of transcript type $i$, $i = 1, \ldots, I$.</td>
</tr>
<tr>
<td>$\theta$</td>
<td>The isoform abundance vector $[\theta_1, \theta_2, \ldots, \theta_I]$.</td>
</tr>
<tr>
<td>$s_j$</td>
<td>Read type $j$, $j = 1, \ldots, J$.</td>
</tr>
<tr>
<td>$n_{i,j}$</td>
<td>The number of reads $s_j$ that are generated from transcripts $i$.</td>
</tr>
<tr>
<td>$n_j$</td>
<td>The number of read $s_j$ that are generated from all the transcripts, i.e. $n_j = \sum_{i=1}^{I} n_{i,j}$.</td>
</tr>
<tr>
<td>$a_{i,j}$</td>
<td>Up to proportionality, the sampling rate of $n_{i,j}$, i.e., the rate that read $s_j$ is generated from each individual transcript $i$.</td>
</tr>
<tr>
<td>$a_j$</td>
<td>The sampling rate vector $[a_{1,j}, a_{2,j}, \ldots, a_{I,j}]$ for read $s_j$.</td>
</tr>
<tr>
<td>$\theta \cdot a_j$</td>
<td>The sampling rate of $n_j$, i.e., the rate that read $s_j$ is generated from all the transcripts.</td>
</tr>
<tr>
<td>$A$</td>
<td>The $J \times I$ matrix of the sampling rates ${a_{i,j}}_{i=1,j=1}^{I,J}$.</td>
</tr>
<tr>
<td>$c_i$</td>
<td>The copy number of the $i^{th}$ transcript in the sample.</td>
</tr>
<tr>
<td>$l_i$</td>
<td>The length of the $i^{th}$ transcript in the sample.</td>
</tr>
<tr>
<td>$n$</td>
<td>The total number of reads.</td>
</tr>
</tbody>
</table>
Problem

- Given reads \((n_{ij})\), sequence/length information on mRNAs \((l_i)\), and some statistics about paired-end reads estimate the abundance of each mRNA \((\theta_i)\)
Summary

- The methods are based on assumptions on the sampling rate, $a_{ij}$
- Uniform sampling
  - 0 or $n$
- Insert length model
  - Filter out some inserts based on empirical length distribution
Uniform model

- Discuss from the paper
Insert length model

- Discuss from the paper
Estimation

\[
\begin{align*}
\text{maximize} & \quad n^T \log(A\theta) - \text{sum}(A\theta) \\
\text{s.t.} & \quad \theta \geq 0
\end{align*}
\]
Computational efficiency

- Reduce individual reads into read categories
- Prove this is sufficient statistics for estimation
Example

- From paper
Discussion

● Possible improvements?
Next week

- Genome annotation

Reading:

- Annotation confidence score for genome annotation: a genome comparison approach
- ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles