A practical, probabilistic method for de novo assembly evaluation

MohammadReza Ghodsi, Christopher M. Hill, Irina Astrolovskaya, Henry Lin, Dan Sommer, Mihai Pop

Center for Bioinformatics and Computational Biology
University of Maryland, College Park

Introduction
A theoretical foundation for assembly evaluation lays in Gene Myers work [1] where he suggested that correct assembly must be consistent with statistical characteristics of the data generation process. Probabilistically, the observed set of DNA sequences (reads) corresponds to outcomes from a set of independent trials. In each trial, a position is uniformly sampled from the genome, therefore, the probability of read \( i \), \( p_i \) of length / being generated from an assembly is:

\[
p_i = \frac{n_i}{2(L - l)}
\]

where \( n_i \) is the number of times read \( i \) matches length of assembly (and rev-comp)

Medvedev [2] shows that the true genome maximizes the product of length of assembly (and rev-comp)

We extend the error-free model to take into account errors in the reads. Furthermore, our framework allows for any type of data that can modeled as a probability to be used (mate-pairs).

Practical considerations: errors
All current technologies for sequencing DNA fragments have a small, but significant probability of error.

The probability of each read, having been generated from any position \( j \) of the reference, is a value between zero and one, depending on the number of differences between the sequence of the read and the sequence of the reference at position \( j \).

\[
p_i = \sum_{j=1}^{2(L-1)} p_{i,j} \frac{p_{i,j}}{2(L - l)}
\]

Calculated using a dynamic programming algorithm equivalent to the forward algorithm for Hidden Markov Models

Practical considerations: mate-pairs
If assembly \( A \) consists of contigs (or scaffolds), then a pair of mated reads is supported by \( A \) if both reads match to the same contig (or scaffold) with correct orientation and distance.

\[
Pr[\text{read } i \text{ and its mate } i^* \text{ align at position } j_1 \text{ and } j_2 \mid \text{assembly } A] = \Pr[\text{read } i \text{ aligns here}] \cdot \Pr[\text{insert of this size}] \cdot \Pr[\text{read } i^* \text{ aligns here}]
\]

Thus, the probability that mated reads \( i \) and \( i^* \) were generated from \( A \) is equal to the sum of the probabilities of each mate pair alignment / length of the assembly.

Practical considerations: sampling
If the set of reads is too large, we can use a random subset (a sample) of the reads, \( R' \subseteq R \), to estimate the probability of the reads, given the assembly.

\[
\text{Geometric Mean } (R') = \left( \prod_{i \in R'} p_i \right)^{1/|R'|}
\]

Our metric: \( \log \text{GM}(R') \)

Probability approximation using alignment tools
When it is impractical to calculate exact probabilities for large sets of reads, we can approximate these probabilities using fast and memory-efficient alignment search programs (Bowtie2 [3]).

instead of summing over all possible positions in the assembly, only sum the probabilities of good alignments.

Bowtie2 yields similar probabilities and is often an order of magnitude faster in practice.

Results: assembly competitions
We evaluated our method by using data from popular assembly bake-offs [4] [5].

Overall, our rankings correlate with the assembly competitions findings. When we evaluate assemblies using more than just raw sequence data, such as including mate pair information, we get a more similar evaluation.

Human Chr13 assembly rankings

<table>
<thead>
<tr>
<th>Time</th>
<th># reads sampled per trial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling reads</td>
<td>20m 41s</td>
</tr>
<tr>
<td>Bowtie2</td>
<td>16m 26s</td>
</tr>
<tr>
<td>Building index</td>
<td>2m 48s</td>
</tr>
<tr>
<td>Aligning reads</td>
<td>1m 27s</td>
</tr>
<tr>
<td>total</td>
<td>20m 41s</td>
</tr>
</tbody>
</table>

Future Directions
Our metric can be used to tune assembly parameters resulting in better assemblies.

Our framework can be extended to take into account any observational data that can be turned into a probability (quality values, optical maps).

Availability
Source code: [http://sourceforge.net/p/assembly-eval/]
Supports fasta/fastq and mate pair data with varying insert sizes/orientations

References

Chris Hill cmhill@umiacs.umd.edu, Mihai Pop mpop@umiacs.umd.edu