Lecture 17: Heuristic methods for sequence alignment: BLAST and FASTA

Fall 2019
November 14, 2019
Motivation

- Smith-Waterman algorithm too slow for searching large sequence databases
- Most sequences are not homologous to the query so there is no need for the alignment
- Use heuristic methods:
  - FASTA
  - BLAST
FASTA

- Idea: in order for two sequences to be similar, need a run of identical letters
- Only sequences that have high scoring segments need to be aligned

Finding all matching k-mers

• Store the positions of all k-mers in the query sequence
• Scan for matches with those k-mers
• k=2 for protein sequences.
• Need an efficient method for retrieving k-mer positions: hash table!
Assume that x and y are very similar

**Assumption:** \#gaps(x, y) < k

We can align x and y more efficiently:

Time, Space: O(n × k)
Bounded Dynamic Programming

Initialization:

\[ S(i,0), S(0,j) \text{ undefined for } i, j > k \]

Iteration:

For \( i = 1, \ldots, n \)

For \( j = \max(1, i - k), \ldots, \min(n, i+k) \)

\[ S(i, j) = \max \left\{ \begin{array}{l}
S(i - 1, j - 1) + s(x_i, y_j) \\
S(i, j - 1) - d, \text{ if } j > i - k \\
S(i - 1, j) - d, \text{ if } j < i + k
\end{array} \right\} \]

Can extend to the affine gap case
**BLAST: Basic Local Alignment and Search Tool**

- Similar idea to FASTA:
  - Does not compute alignments that do not look promising
  - Fast alignment: does not consider complete edit graph
- Difference from FASTA: seed matches do not need to match exactly
- Great improvement in speed, with a modest decrease in sensitivity with respect to SW.

---

Outline of BLAST

- Eliminate low complexity regions
- Create an “index” of 3-mers from the query sequence
- Search for hits in the database
- Extend hits into HSSPs: High Scoring Segment Pairs (ungapped)
- Extend HSSPs into a local alignment
- Compute “E-values”
BLAST algorithm (cont’d)

List of 3-mers that have a similarity level exceeding some threshold when compared to 3-mers in the query.

Query: KRHRKVLRDNIQGITKPAIRRLARRGGGVKRISGLIYEEETRGVLKIFLENVIRD

k-mer

Neighborhood k-mers:
- GVK 18
- GAK 16
- GIK 16
- GGK 14
- GLK 13
- GNK 12
- GRK 11
- GEK 11
- GDK 11

Neighborhood score threshold (T = 13)

Query: 22 VLRDNIQGITKPAIRRLARRGGGVKRISGLIYEEETRGVLK 60
+++DN +G + IR L G+K I+ L+ E+ RG++K
Sbjct: 226 IIKDNGRGFSGKQIRNLNYGIGLKVIAVL-EKHRGIIK 263

High-scoring segment pair (HSSP)
Indexing the query

• Compile a list of high-scoring 3-mers (have a similarity level exceeding some threshold when compared to 3-mers in the query)
• Typical size of list: 50 times the length of the sequence
Original BLAST

- **Extending a hit**
  - *Ungapped* extensions until score (sum of subst. matrix elements) falls below some threshold

- **Output**
  - All local ungapped alignments with score > threshold
The seed **GGTC** initiates an alignment

Extension with no gaps

Output:

GTAAGGTCC
GTTAGGTCC
Gapped BLAST

- Search for seed. THEN:
- Extend with gaps in a band around seed (anchor) until score < threshold
- Result:
  GTAAGGTCCAGT
  GTTAGGTC–AGT
BLAST: Locally Maximal Segment Pairs

- A segment pair is **locally maximal** if its score can’t be improved by extending or shortening.
- BLAST finds all locally maximal segment pairs with scores above some threshold.
- Output: Statistically significant locally maximal segment pairs (more likely to be of biological interest).
We want to assign probabilities to BLAST scores.

\[ p = P(\text{two random bases are equal}) \]

- The event of a mismatch followed by \( t \) matches has probability \((1 - p)p^t\)
- There are \( mn \) places to begin the event.
- The expected # of such events: \( mn (1 - p)p^t \)
- Rare event: use a Poisson distribution with mean

\[
f_\lambda(k) = \frac{\lambda^k e^{-\lambda}}{k!}
\]

\[
P(\text{alignment of length } t \text{ or longer}) = 1 - P(\text{no such event}) = 1 - \exp\left(-mn (1 - p)p^t\right)
\]

This is known as the \textit{extreme value distribution}.
BLAST statistics

- # hits with score greater than $\theta$ has mean $E(\theta) = Kmne^{-\lambda\theta}$; $K$ is a constant, $m,n$ are the lengths of the two compared sequences.
**Sample BLAST output**

- **Blast of human beta globin protein against zebra fish**

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Score (bits)</th>
<th>Value E</th>
</tr>
</thead>
<tbody>
<tr>
<td>171</td>
<td>3e-44</td>
</tr>
<tr>
<td>170</td>
<td>7e-44</td>
</tr>
<tr>
<td>170</td>
<td>7e-44</td>
</tr>
<tr>
<td>168</td>
<td>3e-43</td>
</tr>
</tbody>
</table>

ALIGNMENTS

<table>
<thead>
<tr>
<th>Score = 171 bits (434), Expect = 3e-44</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identities = 76/148 (51%), Positives = 106/148 (71%), Gaps = 1/148 (0%)</td>
</tr>
</tbody>
</table>

Query: 1
```
MVHTPEEKSAVTALWGKVNVDEVGGEALGRRLLVYEPWTQFFESFDLSTPDAMGNPK
```

Sbjct: 1
```
MVEWTDARISGLWGKLNIDEIGPQLSRCLIVYPWTQRYFATFQNLSSPAAIMGNPK
```

Query: 61
```
VKAHGKKVLGAFAFDGLAHLDNKGFATLSELHCDDKLHVDPPFRILLSNVLVCLAHF
```

Sbjct: 61
```
VAAHGRTVMMGGLERAIKNMDVKNAYAALSVMHSEKLHVDPPFRILLSDCTVCAAMKFG
```

Query: 121 KE-FTPPVQAAYKVVAVANGALAHKYH
```
+ F VQ A+QK +A V +AL +YH
```

Sbjct: 121 QAGFNADVQEAQKFLAVVVSALCQYH
### Sample BLAST output

#### Blast of human beta globin DNA against human DNA

<table>
<thead>
<tr>
<th>gi</th>
<th>Accession</th>
<th>Description</th>
<th>Score</th>
<th>E Value</th>
<th>(bits) Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>19849266</td>
<td>gb</td>
<td>AF487523.1</td>
<td>Homo sapiens gamma A hemoglobin (HBG1)</td>
<td>289</td>
<td>1e-75</td>
</tr>
<tr>
<td>183868</td>
<td>gb</td>
<td>M11427.1</td>
<td>HUMHBG3E Human gamma-globin mRNA, 3' end</td>
<td>289</td>
<td>1e-75</td>
</tr>
<tr>
<td>44887617</td>
<td>gb</td>
<td>AY534688.1</td>
<td>Homo sapiens A-gamma globin (HBG1) ge...</td>
<td>280</td>
<td>1e-72</td>
</tr>
<tr>
<td>31726</td>
<td>emb</td>
<td>V00512.1</td>
<td>HSGGL1 Human messenger RNA for gamma-globin</td>
<td>260</td>
<td>1e-66</td>
</tr>
<tr>
<td>38683401</td>
<td>ref</td>
<td>NR_001589.1</td>
<td>Homo sapiens hemoglobin, beta pseud...</td>
<td>151</td>
<td>7e-34</td>
</tr>
<tr>
<td>18462073</td>
<td>gb</td>
<td>AF339400.1</td>
<td>Homo sapiens haplotype PB26 beta-glob...</td>
<td>149</td>
<td>3e-33</td>
</tr>
</tbody>
</table>

**ALIGNMENTS**

>gi|28380636|ref|NG_000007.3| Homo sapiens beta globin region (HBB0) on chromosome 11

Length = 81706

Score = 149 bits (75), Expect = 3e-33

Identities = 183/219 (83%)

Strand = Plus / Plus

Query: 267 ttgggagatgcacaaagcacttggtatgctcaagggcacctttgcccagctgatgcaa 326

Sbjct: 54409 ttcggaagaagctgttactgtgacctcagggacccctttgtacactgcgtaagctgacc 54468

Query: 327 ctgcactgtgaccaagctgcactgctgatctgactggactgagc 365

Sbjct: 54469 ctgcactgtaacaagctgcacgtgagaccccttgagactgcactctcgaacttc 54507
Flavors of BLAST

- blastn: Nucleotide-nucleotide
- blastp: Protein-protein
- blastx: Translated query vs. protein database (all six frames)
- tblastn: Protein query vs. translated database (finding homologous protein coding regions in unannotated nucleotide sequences – ESTs and draft genomes)
- tblastx: Translated query vs. translated database (6 frames each - expensive)
Versions of BLAST

- PSI-BLAST
  - Find members of a protein family or build a custom position-specific score matrix
- Megablast:
  - Search longer sequences with fewer differences
Timeline

- 1970: Needleman-Wunsch global alignment algorithm
- 1981: Smith-Waterman local alignment algorithm
- 1985: FASTA
- 1990: BLAST (basic local alignment search tool)
- 2000s: BLAST has become too slow in - faster algorithms are developed
  - Pattern Hunter
  - BLAT
- 2009: BLAST+: a complete re-write of BLAST
- 2010: Next generation sequencing: need super fast algorithms!
PatternHunter

- BLAST: matches short consecutive sequences (consecutive seed)
- A seed of length 11:
  
  11111111111

  Each 1 represents a “match”

- PatternHunter: matches short non-consecutive sequences (spaced seed)
- Increases sensitivity by locating homologies that would otherwise be missed
- Example (a spaced seed of length 18 w/ 11 “matches”):
  
  111010010100110111

  Each 0 represents a “don’t care”, so there can be a match or a mismatch
BLAT: BLAST-Like Alignment Tool

- BLAT builds an index of the database and scans through the query sequence, whereas BLAST builds an index of the query and then scans through the database.

- What is the advantage of this approach?

- Potential disadvantage?

---

Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002.
BLAT: BLAST-Like Alignment Tool

- Builds an index of all 11-mers that are not heavily involved in repeat regions (what are the tradeoffs in the choice of k-mer size?)
BLAT: BLAST-Like Alignment Tool

- Designed to find matches for sequences that are closely related (originally built to help annotate vertebrate genomes)
- Models splice junction dimers and performs “spliced alignment” (BLAST will generate matches to individual exons).
- Much faster than BLAST
Some might wonder why in the year 2002 the world needs another sequence alignment tool.
A BLAST demo...