Heuristic Methods for Sequence Database Searching

BMI/CS 576
www.biostat.wisc.edu/bmi576/
Colin Dewey
colin.dewey@wisc.edu
Fall 2015
### BLAST Results

**Sequences producing significant alignments:**

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gb</td>
<td>AAN84548.1</td>
<td>beta globin chain variant [Homo sapiens]</td>
<td>90.6</td>
</tr>
<tr>
<td>gb</td>
<td>AAK29639.1</td>
<td>AF349114_1 beta globin chain variant [Homo sapiens]</td>
<td>90.6</td>
</tr>
<tr>
<td>gb</td>
<td>AAF00489.1</td>
<td>AF181989_1 hemoglobin beta subunit variant [Homo sapiens]</td>
<td>90.6</td>
</tr>
<tr>
<td>gb</td>
<td>AAA35952.1</td>
<td>beta-globin</td>
<td>90.6</td>
</tr>
<tr>
<td>gb</td>
<td>AAX37051.1</td>
<td>hemoglobin beta [synthetic construct]</td>
<td>90.6</td>
</tr>
<tr>
<td>gb</td>
<td>AAR96398.1</td>
<td>hemoglobin beta [Homo sapiens]</td>
<td>90.1</td>
</tr>
<tr>
<td>gb</td>
<td>AAL68978.1</td>
<td>AP083883_1 mutant beta-globin [Homo sapiens]</td>
<td>90.1</td>
</tr>
<tr>
<td>gb</td>
<td>AAX29557.1</td>
<td>hemoglobin beta [synthetic construct]</td>
<td>90.1</td>
</tr>
<tr>
<td>ref</td>
<td>NP_000509.1</td>
<td>beta globin [Homo sapiens] &gt;ref</td>
<td>XP_508242.1</td>
</tr>
<tr>
<td>sp</td>
<td>P02024</td>
<td>HBB_GOGO</td>
<td>Hemoglobin subunit beta (Hemoglobin beta cha</td>
</tr>
<tr>
<td>gb</td>
<td>AAD19696.1</td>
<td>beta-globin chain [Homo sapiens]</td>
<td>90.1</td>
</tr>
<tr>
<td>em</td>
<td>CAA26204.1</td>
<td>beta-globin [Pan troglodytes]</td>
<td>89.7</td>
</tr>
<tr>
<td>gb</td>
<td>AAN16468.1</td>
<td>hemoglobin beta chain variant Hb.Sinai-Bel Air [H</td>
<td>89.7</td>
</tr>
<tr>
<td>gb</td>
<td>ABG47031.1</td>
<td>hemoglobin [Homo sapiens]</td>
<td>89.7</td>
</tr>
<tr>
<td>gb</td>
<td>ABA19233.1</td>
<td>hemoglobin beta [Homo sapiens]</td>
<td>89.7</td>
</tr>
<tr>
<td>em</td>
<td>CAA43421.1</td>
<td>beta-globin [Gorilla gorilla]</td>
<td>89.3</td>
</tr>
<tr>
<td>gb</td>
<td>AAY46275.1</td>
<td>beta globin chain [Homo sapiens]</td>
<td>89.3</td>
</tr>
<tr>
<td>gb</td>
<td>AAK20080.1</td>
<td>mutant beta globin [Homo sapiens]</td>
<td>89.3</td>
</tr>
<tr>
<td>gb</td>
<td>AAN11321.1</td>
<td>hemoglobin beta chain variant Hb-I_Toulouse [Homo sapiens]</td>
<td>89.3</td>
</tr>
<tr>
<td>gb</td>
<td>AAG46184.1</td>
<td>mutant beta-globin [Homo sapiens] &gt;gb</td>
<td>AAG46184...</td>
</tr>
<tr>
<td>gb</td>
<td>ABX52138.1</td>
<td>hemoglobin, beta (predicted) [Papio anubis]</td>
<td>88.4</td>
</tr>
<tr>
<td>gb</td>
<td>AAD30656.1</td>
<td>mutant beta-globin [Homo sapiens]</td>
<td>88.0</td>
</tr>
<tr>
<td>pdb</td>
<td>1HBA</td>
<td>Chain B, High-Resolution X-Ray Study Of Deoxyhemog...</td>
<td>86.7</td>
</tr>
</tbody>
</table>
Heuristic Alignment Motivation

- $O(mn)$ too slow for large databases with high query traffic
- **Heuristic algorithm**: an algorithm that isn’t guaranteed to find the optimal solution, but that is efficient and finds good solutions in practice
- heuristic methods do fast approximation to dynamic programming
  - FASTA [Pearson & Lipman, 1988]
Heuristics Alignment Motivation

• consider the task of searching SWISS-PROT against a query sequence:
  – say our query sequence is 362 amino-acids long
  – SWISS-PROT release 38 contained 29,085,265 amino acids
  – finding local alignments via dynamic programming would entail $O(10^{10})$ matrix operations
• many servers handle thousands of such queries a day (NCBI > 100,000)
BLAST Overview

• **Basic Local Alignment Search Tool**
• BLAST heuristically finds high scoring local alignments
• typically used to search a query sequence against a database of sequences
• key tradeoff made: sensitivity vs. speed

\[ \text{sensitivity} = \frac{\text{# significant matches detected}}{\text{# significant matches in DB}} \]
Overview of BLAST Algorithm

• given: query sequence $q$, word length $w$, word score threshold $T$, segment score threshold $S$
  – compile a list of “words” (of length $w$) that score at least $T$ when compared to words from $q$
  – scan database for matches to words in list
  – extend all matches to seek high-scoring alignments
• return: alignments scoring at least $S$
Determining Query Words

Given:

- query sequence: QLNFSAGW
- word length $w = 2$ (default for protein usually $w = 3$)
- word score threshold $T = 9$

Step 1: determine all words of length $w$ in query sequence ($w$-mers)

QL  LN  NF  FS  SA  AG  GW
Determining Query Words

Step 2: determine all words that score at least $T$ when compared to a word in the query sequence

words from sequence query words w/ $T=9$
QL $QL=9$
LN $LN=10$
NF $NF=12, NY=9$
...
SA none
...
Scanning the database

• Search database for all occurrences of query words
• Approach:
  – index database sequences into table of words (pre-compute this)
  – index query words into table (at query time)
Extending Hits

- BLAST extends hits into local alignments
- The original version of BLAST extended each hit separately

![Graph showing the relationship between Broad bean leghemoglobin I and Horse beta globin](Image)

11
Extending Hits in Original Blast

- extend hits in both directions (without allowing gaps)
- terminate extension in one direction when score falls certain distance below best score for shorter extensions

\[ \text{score}(c) \geq \text{score}(b) - \varepsilon \]

- return segment pairs scoring at least \( S \)
How to choose w and T?

- Tradeoff between running time and sensitivity
- Sensitivity

\[
\text{sensitivity} = \frac{\# \text{ significant matches found}}{\# \text{ of significant matches in DB}}
\]

- T
  - small T: greater sensitivity, more hits to expand
  - large T: lower sensitivity, fewer hits to expand

- w
  - Larger w: fewer query word seeds, lower time for extending, but more possible words (\(20^w\) for AAs)
The Two-Hit Method

• extension step typically accounts for 90% of BLAST’s execution time
• key idea: do extension only when there are two hits on the same diagonal within distance A of each other
• to maintain sensitivity, lower $T$ parameter
  – more single hits found
  – but only small fraction have associated 2nd hit
The Two-Hit Method

Figure from: Altschul et al. Nucleic Acids Research 25, 1997
Gapped BLAST

• trigger gapped alignment if two-hit extension has a sufficiently high score
• find length-11 segment with highest score; use central pair in this segment as seed
• run DP process both forward & backward from seed
• prune cells when local alignment score falls a certain distance below best score yet
Gapped BLAST

Figure from: Altschul et al. Nucleic Acids Research 25, 1997
# BLAST Programs

<table>
<thead>
<tr>
<th>Program</th>
<th>Query</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASTP</td>
<td>Protein</td>
<td>Protein</td>
</tr>
<tr>
<td>BLASTN</td>
<td>DNA</td>
<td>DNA</td>
</tr>
<tr>
<td>BLASTX</td>
<td>Translated DNA</td>
<td>Protein</td>
</tr>
<tr>
<td>TBLASTN</td>
<td>Protein</td>
<td>Translated DNA</td>
</tr>
<tr>
<td>TBLASTX</td>
<td>Translated DNA</td>
<td>Translated DNA</td>
</tr>
</tbody>
</table>
PSI (*Position Specific Iterated*) BLAST

- basic idea
  - use results from BLAST query to construct a *profile matrix*
  - search database with profile instead of query sequence
  - iterate
### A Profile Matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.4</td>
</tr>
<tr>
<td><strong>R</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td><strong>D</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td><strong>N</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.2</td>
<td></td>
</tr>
<tr>
<td><strong>C</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-3.1</td>
<td></td>
</tr>
</tbody>
</table>

**Amino acids**

**Sequence positions**
PSI BLAST: Searching with a Profile

- aligning profile matrix to a simple sequence
  - like aligning two sequences
  - except score for aligning a character with a matrix position is given by the matrix itself – not a substitution matrix

```
C   N   A   R
A
R
D
N
C
```
PSI BLAST: Constructing the Profile Matrix

Figure from: Altschul et al. Nucleic Acids Research 25, 1997
BLAST Notes

• It’s heuristic: may miss some good matches
• It’s fast: empirically, 10 to 50 times faster than Smith-Waterman
• large impact:
  – NCBI’s BLAST server handles more than 100,000 queries a day
  – most used bioinformatics program in the world