Sequence Alignment

- **Global:**

- **Local:**
  - Smith-Waterman (1981)
  - Useful when commonality is small and global alignment is meaningless. Often unaligned portions “mask” short stretches of aligned portions. Example: comparing long stretches of anonymous DNA; aligning proteins that share only some motifs or domains.

- **Dynamic Programming (DP) based.**
Why gaps?

Example: Finding the gene site for a given (eukaryotic) cDNA requires “gaps”.

What is cDNA? cDNA = Copy DNA

![Diagram showing the process of transcription, mRNA, translation, and reverse transcription leading to cDNA and protein]
How to score mismatches?
BLAST & FASTA

FASTA

[Lipman, Pearson ’85, ‘88]

Basic Local Alignment Search Tool

[Altschul, Gish, Miller, Myers, Lipman ’90]
BLAST Overview

- Program(s) to search all sequence databases
- Tremendous Speed/Less Sensitive
- Statistical Significance reported
- WWWBLAST, QBLAST (send now, retrieve results later), Standalone BLAST, BLASTcI3 (Client version, TCP/IP connection to NCBI server), BLAST URLAPI (to access QBLAST, no local client)
**CHAPTER ELEVEN**  Assessing Pairwise Sequence Similarity: BLAST and FASTA

---

**Figure 11.7** The initiation of a BLAST search. The search begins with query words of a given length (here, three amino acids) being compared against a scoring matrix to determine additional three-letter words "in the neighborhood" of the original query word. Any occurrences of these neighborhood words in sequences within the target database are then investigated. See text for details.
BLAST Strategy & Improvements

- Lipman et al.: speeded up finding “runs” of “hot spots”.
- Eugene Myers ’94: “Sublinear algorithm for approximate keyword matching”.
- Karlin, Altschul, Dembo ’90, ’91: “Statistical Significance of Matches”
Why Gaps?

Example: Aligning HIV sequences.
BLAST Variants

- **Nucleotide BLAST**
  - **Standard blastn**
  - **MEGABLAST** (Compare large sets, Near-exact searches)
  - **Short Sequences** (higher E-value threshold, smaller word size, no low-complexity filtering)

- **Protein BLAST**
  - **Standard blastp**
  - **PSI-BLAST** (Position Specific Iterated BLAST)
  - **PHI-BLAST** (Pattern Hit Initiated BLAST; reg expr. Or Motif search)
  - **Short Sequences** (higher E-value threshold, smaller word size, no low-complexity filtering, PAM-30)

- **Translating BLAST**
  - **Blastx**: Search nucleotide sequence in protein database (6 reading frames)
  - **Tblastn**: Search protein sequence in nucleotide dB
  - **Tblastx**: Search nucleotide seq (6 frames) in nucleotide DB (6 frames)
BLAST Cont’d

- **RPS BLAST**
  - Compare protein sequence against Conserved Domain DB; Helps in predicting rough structure and function

- **Pairwise BLAST**
  - blastp (2 Proteins), blastn (2 nucleotides), tblastn (protein-nucleotide w/ 6 frames), blastx (nucleotide-protein), tblastx (nucleotide w/6 frames- nucleotide w/ 6 frames)

- **Specialized BLAST**
  - Human & Other finished/unfinished genomes
  - *P. falciparum*: Search ESTs, STSs, GSSs, HTGs
  - VecScreen: screen for contamination while sequencing
  - IgBLAST: Immunoglobin sequence database
BLAST Credits

- Stephen Altschul
- Jonathan Epstein
- David Lipman
- Tom Madden
- Scott McGinnis
- Jim Ostell
- Alex Schaffer
- Sergei Shavirin
- Heidi Sofia
- Jinghui Zhang
Databases used by BLAST

 Protein
  - nr (everything), swissprot, pdb, alu, individual genomes

 Nucleotide
  - nr, dbest, dbsts, htgs (unfinished genomic sequences), gss, pdb, vector, mito, alu, epd

 Misc
BLAST Parameters and Output

- Type of sequence, nucleotide/protein
- Word size, $w$
- Gap penalties, $p_1$ and $p_2$
- Neighborhood Threshold Score, $T$
- Database to search, $D$
- Scoring Matrix, $M$

- Score Threshold, $S$
- E-value Cutoff, $E$
- Number of hits to display, $H$
- Score $s$ and E-value $e$
  - E-value $e$ is the expected number of sequences that would have an alignment score greater than the current score $s$. 
Scoring Matrix to Use

- **PAM 40**: Short alignments with high similarity (70-90%)
- **PAM 160**: Members of a protein family (50-60%)
- **PAM 250**: Longer alignments (divergent sequences) (~30%)

- **BLOSUM90**: Short alignments with high similarity (70-90%)
- **BLOSUM80**: Members of a protein family (50-60%)
- **BLOSUM62**: Finding all potential hits (30-40%)
- **BLOSUM30**: Longer alignments (divergent sequences) (<30%)
Main Ideas in BLAST

- Break sequence into words and look for words in database
- Find hotspots where many words find hits and look more closely
- Instead of looking for *approximate* hits of words …
- … find *exact* hits of nearby words
BLAST algorithm: Phase 1

Phase 1: For each word in query, get words (w=3) within threshold T

Example: for a query sequence

```
...FSGTWYA...
```

Consider a word **GTW** in the query

Get list of words (w=3) close to **GTW**:

```
ATW, GSW, ...
```
Use BLOSUM to score word hits

|   | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| A | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| R | -1| 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| N | -2| 0 | 6 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| D | -2| -2| 1 | 6 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| C | 0 | -3| -3| -3| 9 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Q | -1| 1 | 0 | 0 | -3| 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| E | -1| 0 | 0 | -4| 2 | 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| G | 0 | -2| 0 | -1| -3| -2| -2| 6 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| H | -2| 0 | 1 | -1| -3| 0 | 0 | -2| 8 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| I | -1| -3| -3| -3| -1| -3| -4| -3| -3| 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| L | -1| -2| -3| -4| -1| -2| -3| -4| -3| 2 | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| K | -1| 2 | 0 | -1| 1 | 1 | 2 | -1| -3| -2| 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| M | -1| -2| -2| -3| -1| 0 | -2| -3| -2| 1 | 2 | -1| 5 |   |   |   |   |   |   |   |   |   |   |   |   |
| F | -2| -3| -3| -3| -2| -3| -3| -3| -1| 0 | 0 | -3| 6 |   |   |   |   |   |   |   |   |   |   |   |   |
| P | -1| -2| -2| -1| -3| -1| -1| -2| -2| -3| -3| -3| -3| -3| -3| -3| -1| -2| -3| -4| 7 |   |   |
| S | 1 | -1| 1 | 0 | -1| 0 | 0 | 0 | 1 | -2| -2| 0 | -1| -2| -1| -4|   |   |   |   |   |   |   |
| T | 0 | -1| 0 | -1| -1| -1| -1| -2| -2| -1| -1| -1| -2| -1| 1 | 5 |   |   |   |   |   |   |   |
| W | -3| -3| -4| -4| -2| -2| -2| -3| -2| -3| -1| 1 | -4| -3| -2| H |   |   |   |   |   |   |   |
| Y | -2| -2| -2| -3| -2| -1| -2| -3| 2 | -1| -1| -2| -1| 3 | -3| -2| -3| -2| 2 | 7 |   |   |
| V | 0 | -3| -3| -3| -1| -2| -2| -3| -3| 3 | 1 | -2| 1 | -1| -2| -2| 0 | -3| -1| 4 |   |   |

Page 73
Phase 1: Find list of similar words

- Find list of words of length $w$ (here $w = 3$) and distance at least $T$ (here $T = 11$)
  - GTW 22
  - GSW 18
  - ATW 16
  - NTW 16
  - GTY 13
  - GNW 10
  - GAW 9
BLAST: Phases 2 & 3

- Phase 2: Scan database for hits and find HotSpots

- Phase 3:
  - Extend good hit in either direction.
  - Keep track of the score (use a scoring matrix)
  - Stop when the score drops below some cutoff.

KENFDKARFSGTWYAMAKKDPEG  50
MKGLDIQKVAGTWYS  LAMAASD  44

RBP (query)
lactoglobulin (hit)
BLAST: Threshold vs # Hits & Extensions

Fig. 4.12  page 118
Word Size

- **Blastn**: $w = 7, 11, \text{ or } 15$.
  - $w=15$ gives fewer matches and is faster than $w=11$ or $w=7$.

- **Megablast**: $w = 28 \text{ to } 64$.
  - Megablast is VERY fast for finding closely related DNA sequences!
Scores: Follow Extreme Value Distribution

\[ E = Kmn e^{\lambda S} \]

- \( m,n = \text{seq length} \)
- \( S = \text{Raw Score} \)
- \( K \approx \text{Search space} \)

\[ S' = \frac{(\lambda S - \ln K)}{\ln 2} \]

- \( S' = \text{Bit Score} \)

\[ p = 1 - e^{-E} \]

\( p = \text{p-value} \)
E-value versus P-value

<table>
<thead>
<tr>
<th>E-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.9999546</td>
</tr>
<tr>
<td>5</td>
<td>0.99326205</td>
</tr>
<tr>
<td>2</td>
<td>0.86466472</td>
</tr>
<tr>
<td>1</td>
<td>0.63212056</td>
</tr>
<tr>
<td>0.1</td>
<td>0.09516258</td>
</tr>
<tr>
<td>0.05</td>
<td>0.04877058</td>
</tr>
<tr>
<td>0.001</td>
<td>0.00099950</td>
</tr>
<tr>
<td>0.0001</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

**E-values are easier to interpret;**

If query is short aa sequence, then use very large E-value;
Sometimes even meaningful hits have large E-values.
BLAST: Steps

- Choose your sequence
- Choose your tool
- Choose your database
- Select parameters, if needed
- Interpret your results
NCBI Handbook, Eds. McEntyre, Ostell
Graphical Overview of BLAST Results

Distribution of 41 Blast Hits on the Query Sequence

Mouse-over to show defining and scores. Click to show alignments.

Color Key for Alignment Scores:
- <40
- 40-50
- 50-80
- 80-200
- >=200

Alignment scores distribution across the query sequence.
### List of hits with one line descriptions

<table>
<thead>
<tr>
<th>Sequences producing significant alignments:</th>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) <img src="image1.png" alt="image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) <img src="image2.png" alt="image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g116365</td>
<td>sp</td>
<td>P26374</td>
</tr>
<tr>
<td>g121431607</td>
<td>sp</td>
<td>P24386</td>
</tr>
<tr>
<td>g1585775</td>
<td>sp</td>
<td>P37727</td>
</tr>
<tr>
<td>g113628686</td>
<td>sp</td>
<td>Q61598</td>
</tr>
<tr>
<td>g1729566</td>
<td>sp</td>
<td>P39958</td>
</tr>
<tr>
<td>g113628613</td>
<td>sp</td>
<td>Q97556</td>
</tr>
<tr>
<td>g113632299</td>
<td>sp</td>
<td>P50397</td>
</tr>
<tr>
<td>g11707888</td>
<td>sp</td>
<td>P50398</td>
</tr>
<tr>
<td>g1121108</td>
<td>sp</td>
<td>P21856</td>
</tr>
<tr>
<td>g121903424</td>
<td>sp</td>
<td>P50396</td>
</tr>
<tr>
<td>g113628612</td>
<td>sp</td>
<td>Q97555</td>
</tr>
<tr>
<td>g11707886</td>
<td>sp</td>
<td>P31150</td>
</tr>
<tr>
<td>g113638228</td>
<td>sp</td>
<td>P50395</td>
</tr>
<tr>
<td>g11707891</td>
<td>sp</td>
<td>P50399</td>
</tr>
<tr>
<td>g11723467</td>
<td>sp</td>
<td>Q10305</td>
</tr>
<tr>
<td>g1585776</td>
<td>sp</td>
<td>P32864</td>
</tr>
<tr>
<td>g110720243</td>
<td>sp</td>
<td>Q09381</td>
</tr>
<tr>
<td>g12498411</td>
<td>sp</td>
<td>Q43938</td>
</tr>
<tr>
<td>g11135401</td>
<td>sp</td>
<td>Q9XBQ9</td>
</tr>
<tr>
<td>g11135075</td>
<td>sp</td>
<td>Q05519</td>
</tr>
<tr>
<td>g11135195</td>
<td>sp</td>
<td>P57112</td>
</tr>
<tr>
<td>g122575022</td>
<td>sp</td>
<td>Q8TJ87</td>
</tr>
<tr>
<td>g139153156</td>
<td>sp</td>
<td>P49488</td>
</tr>
<tr>
<td>g123788</td>
<td>sp</td>
<td>P30599</td>
</tr>
<tr>
<td>g12498412</td>
<td>sp</td>
<td>P75499</td>
</tr>
<tr>
<td>g15487891</td>
<td>sp</td>
<td>P36225</td>
</tr>
<tr>
<td>g1586602</td>
<td>sp</td>
<td>P37747</td>
</tr>
</tbody>
</table>
Pairwise alignment result of human beta globin and myoglobin

Query = HBB; Subject = MB

Myoglobin RefSeq

Information about this alignment: score, expect value, identities, positives, gaps...

Middle row displays identities; + sign for similar matches

Slide: Courtesy J. Pevsner
Pairwise alignment result of human beta globin and myoglobin: the score is a sum of match, mismatch, gap creation, and gap extension scores

<table>
<thead>
<tr>
<th>Score</th>
<th>18.1 bits</th>
<th>Expect</th>
<th>0.015, Method: Composition-based stats.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identities</td>
<td>11/24 (45%)</td>
<td>Positives</td>
<td>12/24 (50%), Gaps</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query 12</th>
<th>VTALWCKVNNVD--EVCEALCRL</th>
<th>33</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>+WCKV</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>C E I R</td>
<td></td>
</tr>
<tr>
<td>Sbjct 11</td>
<td>VLNWCKVEADIPGHGQEVRLF</td>
<td>34</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>match</th>
<th>4 11 5 6</th>
<th>6 5 4 5 sum of matches: +60</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>mismatch</td>
<td>-1 1 0</td>
<td>-2 -2 -4 0 sum of mismatches: -13</td>
</tr>
<tr>
<td></td>
<td>2 0</td>
<td>3 0</td>
</tr>
<tr>
<td>gap open</td>
<td>11</td>
<td>sum of gap penalties: 12</td>
</tr>
<tr>
<td>gap extend</td>
<td>1</td>
<td>total raw score: 60 13 12 = 35</td>
</tr>
</tbody>
</table>
Pairwise alignment result of human beta globin and myoglobin: the score is a sum of match, mismatch, gap creation, and gap extension scores.

<table>
<thead>
<tr>
<th>Query 12</th>
<th>VTALWCEKNVD--EVCCEALCRL</th>
<th>33</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match</td>
<td>V+EKV D C E L RL</td>
<td></td>
</tr>
<tr>
<td>Score</td>
<td>VINVWCEVADIPCHCQEBVLRLF</td>
<td>34</td>
</tr>
</tbody>
</table>

**Match**
- V matching V earns +4
- T matching L earns -1

**Mismatch**

**Gap open**

**Gap extend**

These scores come from a “scoring matrix”!

Slide: Courtesy J. Pevsner
If $S$ is the (raw) score for a local alignment, the **normalized** score $S'$ (in bits) is given by

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

The parameters $K$ and $\lambda$ depend on the scoring system.
Expect value or E-value

- E-value is **not a probability**, but describes strength of random background noise.
- E-value describes **number of hits** one can “**expect**” to see by chance when searching a database of a particular size.
- It decreases exponentially with the score (S).
- **E-value = 1** means “in a database of current size, one might expect to see one match with a similar score simply by chance. Lower E-value mean more “**significant**” match.
- **WARNING**: Short sequences can be virtually identical and have relatively high E-values.
  - Calculation of E-value takes into account length of query sequence. Since shorter sequences have a high probability of occurring in the database purely by chance, E-values can be high.
BLAST Tutorial

RBP4 and PAEP:
Low bit score, E value 0.49, 24% identity (“twilight zone”).
But they are indeed homologous. Try a BLAST search with PAEP as a query, and find many other lipocalins.
Difficulties with BLAST

- Use human beta globin as a query against human RefSeq proteins, and blastp does not “find” human myoglobin. This is because the two proteins are too distantly related. PSI-BLAST at NCBI as well as hidden Markov models easily solve this problem.

- How can we search using 10,000 base pairs as a query, or even millions of base pairs? Many BLAST-like tools for genomic DNA are available such as PatternHunter, Megablast, BLAT, and BLASTZ.
Rules of Thumb

- Most sequences with significant similarity over their entire lengths are homologous.
- Matches that are > 50% identical in a 20-40 aa region occur frequently by chance.
- Distantly related homologs may lack significant similarity. Homologous sequences may have few absolutely conserved residues.
- A homologous to B & B to C ⇒ A homologous to C.
- Low complexity regions, transmembrane regions and coiled-coil regions frequently display significant similarity without homology.
- Greater evolutionary distance implies that length of a local alignment required to achieve a statistically significant score also increases.
Results of searches using different scoring systems may be compared directly using normalized scores.

If $S$ is the (raw) score for a local alignment, the normalized score $S'$ (in bits) is given by

$$S = \frac{\lambda - \ln(K)}{\ln(2)}$$

The parameters depend on the scoring system.

Statistically significant normalized score,

$$S > \log\left(\frac{N}{E}\right)$$

where $E$-value $= E$, and $N =$ size of search space.