Bioinformatics Chapter 8.

Sequence Alignment and Database Searching

오 석 준
서울대 바이오정보기술연구센터
E-mail: sjbaugh@cbit.snu.ac.kr

장 병 탁
서울대 컴퓨터공학부 바이오지능연구실 &
바이오정보기술연구센터
E-mail: btzhang@cse.snu.ac.kr

Sirk-June Augh
Center for Bioinformation Technology (CBIT)
Seoul National University

Byoung-Tak Zhang
Center for Bioinformation Technology (CBIT) &
Biointelligence Laboratory
School of Computer Science and Engineering
Seoul National University

This material is available at http://bi.snu.ac.kr/ &
http://cbit.snu.ac.kr/
Outline

- The Evolutionary Basis of Sequence Alignment
- The Modular Nature of Proteins
- Optimal Alignment Methods
- Substitution Scores and Gap Penalties
- Statistical Significance of Alignments
- Database Similarity Searching
- FASTA
- BLAST
- Database Searching Artifacts
- Position-Specific Scoring Matrices
- Spliced Alignments
Similarity vs. Homology

Similarity:
- An observable quantity that might be expressed, as say, percent identity or some other suitable measure.

Homology:
- Refers to a conclusion drawn from these data that two genes share a common evolutionary history.
Conserved Positions are often of Functional Importance

**Figure 8.1.** Conserved positions are often of functional importance. Alignment of trypsin proteins of mouse (SWISS-PROT P07146) and crayfish (SWISS-PROT P00075). Identical residues are underlined. Indicated above the alignments are three disulfide bonds (–S–S–), with participating cysteine residues conserved, amino acids side chains involved in the charge relay system (asterisk), and active side residue governing substrate specificity (diamond).
Alignment Jargon

Evolutionarily related sequences differ from one other because of several processes:

- **Substitutions**
- **Insertions**
- **Deletions**
Alignment Jargon

GCG
||
ACG

- 1 mismatch
- 2 matches

Substitution

A → G

GCG

ACG

ACG
Alignment Jargon

\textbf{ATCG} \quad \textbf{ACG}
\quad ||
\quad \textbf{A–CG}

- 0 mismatches
- 3 matches
- 1 gap

Insertion

\rightarrow T

(C) 2002 SNU CSE Biointelligence Lab (BI)
Alignment Jargon

ATCG

\[ \begin{array}{c}
| & | & | \\
A & C & G \\
\end{array} \]

- 0 mismatches
- 3 matches
- 1 gap

Deletion

\[ \begin{array}{c}
A & C & G \\
\end{array} \]

\[ \begin{array}{c}
A & T & C & G \\
\end{array} \]

\[ \begin{array}{c}
A & C & G \\
\end{array} \]

\[ \begin{array}{c}
A & T & C & G \\
\end{array} \]
Alignment vs. Prediction: When are Alignment Reliable?

- New sequences are aligned against all sequences in DB
  - Hints toward structural/functional relationships + functional insights
- Fundamental question
  - When is the sequence similarity high enough that one may infer a structural/functional similarity from the pairwise alignment of two sequences?
  - Given the detected overlap in a sequence segment, can a similarity threshold be defined that sifts out cases where the inference will be reliable?
- Answer
  - It depends on the structural/functional aspect one wants to investigate
  - Different for each task
- Prediction
  - When alignment alone is not enough to lead a reliable inference
Local Alignment vs. Global Alignment

- **Local Alignment**
  - Attempts to align regions of sequences with the highest density of matches. In doing so, one or more islands of subalignments are created in the aligned sequences.

- **Global Alignment**
  - Attempts to match as many characters as possible, from end to end, in a set of two or more sequences.
## Local Alignment vs. Global Alignment

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
</tr>
<tr>
<td>A</td>
<td>m</td>
<td>1</td>
<td>-1</td>
<td>-3</td>
<td>-5</td>
<td>-7</td>
<td>-9</td>
<td>-11</td>
</tr>
<tr>
<td>C</td>
<td>m</td>
<td>-1</td>
<td>2</td>
<td>0</td>
<td>-2</td>
<td>-4</td>
<td>-6</td>
<td>-8</td>
</tr>
<tr>
<td>A</td>
<td>m</td>
<td>-3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-3</td>
<td>-5</td>
</tr>
<tr>
<td>C</td>
<td>m</td>
<td>-5</td>
<td>-2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>C</td>
<td>m</td>
<td>-7</td>
<td>-4</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>m</td>
<td>-9</td>
<td>-6</td>
<td>-3</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>m</td>
<td>-11</td>
<td>-8</td>
<td>-5</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>m</td>
<td>-13</td>
<td>-10</td>
<td>-7</td>
<td>-4</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
</tr>
</tbody>
</table>
BLAST search result
Query: Soy bean (plant) leghemoglobin
Database: homo sapiens
Alignment result shows two merely matched sequences, but their functions and structures are surprisingly coincided.
### Optimal Global Sequence Alignment

<table>
<thead>
<tr>
<th>Human-ZCr</th>
<th>Ecoli-QOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATGQKLMRAVRVFEGGPEVLKLRSIDAVPIPKDHQVLIKVHACGVNPVETYIRSHTS</td>
<td>-------MATUREFHKGPEVLQA-VEFTPADPAEENIQVENKAINFDITYIRSGLY</td>
</tr>
<tr>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>RKLPLLGYTPGDVG1EAVGDNASTAFKKGDRVFTSSTSTSGYAEYALADDHTYKLPK</td>
<td>-PPSLPSGLGTEAAGIVSKVGSGVKHIAGDRVVAQSLAGYSSVNH1ADKAILPA</td>
</tr>
<tr>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>LDFKQGAAILIPYFTPAYRALIHSACVKAGESVAVHGASGGVGLACQIARAYGLKILGA</td>
<td>ISFEQAAASFLKGLTVYLLRKYIIPDEQFLFAAAAGVVGLACQWAKALGANLGTIV</td>
</tr>
<tr>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>GTEEGKQIVLQNGAHEVFNHEHNYIDKIIVVKEGIDIIEMLANVNLSDKLSLLSIG</td>
<td>GTAQKAQSLAKGAQVINSRREDVSLKEITGGKVVRVYDSSVRGDRTWEIILCLQRH</td>
</tr>
<tr>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>GRTVIVVG-SRGTIEINPRDTMAKES---SIIGVTLESTSKEEEOQYAAALQAEGNELW</td>
<td>GLMVSGVSNSSQAVTOVNGLINQKSLVTPRSLQYITTBEELTEASNELFSLIASVI</td>
</tr>
<tr>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>KPVIGSO--YPEKVAEHENIIGSGATGKMILL</td>
<td>KVDVAEQQKYLDAQRAHE-ILERATQGSLLLIP</td>
</tr>
<tr>
<td>**</td>
<td>**</td>
</tr>
</tbody>
</table>

**Figure 8.2.** Optimal global sequence alignment. Alignment of the amino acid sequences of human zeta-crystallin (SWISS-PROT Q08257) and *E. coli* quinone oxidoreductase (SWISS-PROT P28304). It is an optimal global alignment produced by the CLUSTAL W program (Higgins et al., 1996). Identical residues are marked by asterisks below the alignment, and dots indicate conserved residues.
Prediction of Functional Features

- Two protein sequences share sequence similarity
  - These proteins share common function?
- New sequence identity threshold is required for prediction of function
  - Threshold used for structural problems can not be used.
- Solution
  - Split each sequence into a number of subsequences
  - The fraction of aligned site per alignment
Figure 8.3. Modular structure of two proteins involved in blood clotting. Schematic representation of the modular structure of human tissue plasminogen activator and coagulation factor XII. A module labeled C is shared by several proteins involved in blood clotting. F1 and F2 are frequently repeated units that were first seen in fibronectin. E is a module resembling epidermal growth factor. A module known as a “kringle domain” is denoted K.
Measuring Alignment Quality

Good alignments should have …

- “many” exact matches
- “few” mismatches
  - “many” of the mismatches should be similar residues
- “few” gaps
Measuring Alignment Quality

Begin with...
Longest Exact Match

QTRPQNVLNPP

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

STRQNVINPWAAQ

S = 3a

S = alignment score
a = match score
Measuring Alignment Quality

... allow some mismatches

\[
\text{QTRPQNVLNPP} \quad \text{||| |||} \quad \text{STRQNVINPWAAQ}
\]

\[
S = 5a - 1b
\]

S = alignment score
a = match score
b = mismatch penalty
Measuring Alignment Quality

...and finally, introduce some gaps

\[ S = 7a - 1b - 1c \]

QTRPQNVLNPP

\[
\begin{array}{cccccc}
& & & & & \\
\mid & \mid & \mid & \mid & \\
\end{array}
\]

STR-QNVINPWAAQ

S=alignment score
\[ a= \text{match score} \]
\[ b= \text{mismatch penalty} \]
\[ c= \text{gap penalty} \]
Dot-matrix Representations

Figure 8.4

Figure 8.4. Dot matrix sequence comparison. Dot matrix comparison of the amino acid sequences of human coagulation factor XII (F12; SWISS-PROT P00748) and tissue plasminogen activator (PLAT; SWISS-PROT P00750). The figure was generated using the dotter program (Sonnhammer and Durban, 1996).
Dot-matrix Representations

Figure 8.5

Figure 8.5. Dot-matrix, path graph, and alignment. All three views represent the alignment of the EGF similarity domains in the human urokinase plasminogen activator (PLAU; SWISS-PROT P00749) and tissue plasminogen activator (PLAT; SWISS-PROT P00750) proteins. (a) The entire proteins were compared with dotter and an enlargement of the small region corresponding to the EGF domain is shown here. (b) The path graph representation of the alignment found by BLASTP. (c) The BLASTp alignment represented in the familiar text form.
Local Alignments

- Many problems in computer science can be reduced to the task of finding the **optimal path** through a graph.
- Some positive incremental scores will be used for aligning **identical residues**, with negative scores used for substitutions and gaps.
- Finding optimal local alignment
  - Dynamic programming.
  - Needleman-Wunsch algorithm (Needleman and Wunsch, 1970)
  - Smith-Waterman algorithm (Smith and Waterman, 1981)
  - Viterbi decoding algorithm
Dynamic Programming

- General optimization technique
- Application environment
  - Problem can be recursively subdivided into two similar subproblems of smaller size
  - Solution can be obtained by piecing together the solutions to the two subproblems
- Example: finding shortest path
  - Problem: [A → C → B] → [A → C] and [C → B]
  - Solution: [A → C] + [C → B] → [A → B]
Smith-Waterman Algorithm (1)

- Finding local alignments
- Using dynamic programming

\[
s(i, j) = \max \begin{cases} 
S(i-1, j-1) + s(i,j) \\ 
S(i-1, j) - GP \\ 
S(i, j-1) - GP \\ 
VG(i-2, j) - GEP \\ 
HG(i, j-2) - GEP \\ 
0 
\end{cases}
\]

TCAT*G
*CATTGG
Smith-Waterman Algorithm (2)

- Best results and slow performance
- Can grasp results missed by BLAST or FASTA.
- Available on web: spiral.genes.nig.ac.jp/homolgy/ssearch-e.shtml
Comparison of:
(A) F9-human.aa >F9  gl|119772|sp|P00740|PAI2_HUMAN COAGULATION FA - 461 aa
(B) f12-hum.aa =F12  gl|119763|sp|P00748|PAI2_HUMAN COAGULATION - 615 aa

using protein matrix

1

35.4% identity in 254 aa overlap; score: 358

220 230 240 250 260 270

F9 QSFDTPTRVQGGERDKPQQSFQPVVLNGKVKDGFCGSIVKEKTVAAKCF---TVKIVILAA

F12 KLSKNTTVGGYTVLSMGAPYTVADLY-WSHKFCASLLAPCWLTVAAHLQDPRFEDL

370 380 390 400 410 420

280 290 300 310 320 330

F9 TVVWGRHNMTCRTTVQKRNIVRIPHRNVAAIDKYNHIALALLDFQV---TVALNVY

F12 TVVLQGERANHNSCEPYCQUTLAKVYSHLHAASFY---STQDLALNLISQZADQSCALLSF

430 440 450 460 470 480

340 350 360 370 380

F9 VTPICIAKETYIAFEPLKGGSTVQGWRVPHGTS-ALVDQYLYRVLVQDRLCTDCSTK---

F12 VQCVCSQGAAARPSPHTTCQ---VAGWGHQPDEAETASPQIAQVQPLILBCEECSAPTVHG

490 500 510 520 530

390 400 410 420 430 440

F9 -TIVYNDFCAKFMGGKQSCQDLQSGNPTVEVTS---PLTVIGWNGKECMKNGQY

F12 KELPSMVCGMLQNLDQGQSEHSHGQGHPCDOQAKRPQLQVILLNWCGSGHMACEFQY

540 550 560 570 580 590

450

F9 TKVSVYHVNWIKET

F12 TVMAYTAMNHP

600 610

2

34.7% identity in 49 aa overlap; score: 120

100 110 120 130 140

F9 VDGQQKQNERPCMLGSGCSEDINSYCCCPQFCRGCELDPPNCE

F12 LASQACRTNPCLMGRCLVESWHRLECPVQVTGFFCDVIVQASCTYDR

190 200 210 220

3

33.3% identity in 36 aa overlap; score: 87

100 110 120

F9 DQCHSH-PCLANGSKCRDINSTEBKCPFQRKREK

F12 DQCHSHSPQKGGQTVSNMSPQCLPQHLTQNNCEQ

100 110 120 130

--------

Figure 8.6. Optimal and suboptimal local alignments. The three best alignments found when aligning the sequences of human coagulation factor IX (F9; SWISS-PROT P00740) and coagulation factor XII (F12; SWISS-PROT P00748).
Scoring Matrix

- A **unitary matrix** is used for base pairs
  - Each position can be given a score of +1 if it matches and a score of zero if it does not.

- **Substitution matrices** are used for amino acid alignments.
  - Certain amino acids can substitute easily for one another in related proteins because of their similar physicochemical properties.
Substitution Scores

- **Point accepted mutation (PAM) model of evolution**
  - A unit of *evolutionary divergence* in which 1% of the amino acids have been changed
  - Log-odds approach
    - The substitution scores in the matrix are proportional to the natural log of the ratio of target frequencies to background frequencies

- **BLOSUM substitution matrices**
  - Use of a different strategy for estimating the target frequencies
  - The underlying data are derived from the BLOCKS database, which contains local alignments ("blocks")
Gap Penalties

- Affine gap penalties
  - For a gap with length $n$:
    - Penalty = $G + Ln$
    - $G$: gap-opening penalty (around 10-15)
    - $L$: gap-extension penalty (around 1 or 2)
PAM Scoring Matrices

- Point accepted mutation (PAM) model of evolution
  - A unit of *evolutionary divergence* in which 1% of the amino acids have been changed
  - Log-odds approach
    - The substitution scores in the matrix are proportional to the natural log of the ratio of target frequencies to background frequencies

- Parameters were estimated from a small sample of closely related proteins.
The PAM250 Scoring Matrix

|   | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| A | 2 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| R | -2| 6 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| N | 0 | 0 | 2 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| D | 0 | -1| 2 | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| C | -2| -4| -5| 12|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Q | 0 | 1 | 1 | 2 | -5| 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| E | 0 | -1| 1 | 3 | -5| 2 | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |
| G | 1 | -3| 0 | 1 | -3| -1| 0 | 5 |   |   |   |   |   |   |   |   |   |   |   |   |
| H | -1| 2 | 2 | 1 | -3| 3 | 1 | -2| 6 |   |   |   |   |   |   |   |   |   |   |   |
| I | -1| -2| -2| -2| -2| -2| -3| -2| 5 |   |   |   |   |   |   |   |   |   |   |   |
| L | -2| -3| -3| -4| -6| -2| -3| -4| -2| 2 | 6 |   |   |   |   |   |   |   |   |   |
| K | -1| 3 | 1 | 0 | -5| 1 | 0 | -2| 0 | -2| -3| 5 |   |   |   |   |   |   |   |   |
| M | -1| 0 | -2| -3| -5| -1| -2| -3| -2| 2 | 4 | 0 | 6 |   |   |   |   |   |   |   |
| F | -3| -4| -3| -6| -4| -5| -5| -5| -2| 1 | 2 | -5| 0 | 9 |   |   |   |   |   |   |   |
| P | 1 | 0 | 0 | -1| -3| 0 | -1| 0 | 0 | -2| -3| -1| -2| -5| 6 |   |   |   |   |   |
| S | 1 | 0 | 1 | 0 | 0 | -1| 0 | 1 | -1| -1| -3| 0 | -2| -3| 1 | 2 |   |   |   |   |
| T | 1 | -1| 0 | 0 | -2| -1| 0 | 0 | -1| 0 | -2| 0 | -1| -3| 0 | 1 | 3 |   |   |   |
| W | -6| 2 | -4| -7| -8| -5| -7| -7| -3| -5| -2| -3| -4| 0 | -6| -2| -5| 17|   |   |
| Y | -3| -4| -2| -4| 0 | -4| -4| -5| 0 | -1| -1| -4| -2| 7 | -5| -3| -3| 0 | 10|   |
| V | 0 | -2| -2| -2| -2| -2| -2| -1| -2| 4 | 2 | -2| -2| 1 | -1| -1| 0 | -6| -2| 4 |   |

(C) 2002 SNU CSE Biointelligence Lab (BI)
BLOSUM Scoring Matrices

- Construct a database of “blocks”: ungapped, aligned conserved regions of proteins
- Cluster sequences within a block that are more similar than a chosen threshold (e.g., 62% for BLOSUM62)
- Represent each cluster of sequences using their “average” sequence
- After the averaging procedure, each modified block consists of average sequence(s) and sequences that did not cluster with the others
- Examine all pairs of sequences in the modified blocks, tabulate $p_{ij}$ the probability of observing aa $i$ in one sequence and aa $j$ in a second sequence.
The BLOSUM62 Scoring Matrix

|   | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| A | 4 | 1 | 5 | 6 | 7 | 8 | 9 | 10| 11| 12| 13| 14| 15| 16| 17| 18| 19| 20| 21| 22| 23|
| R | -1| 1 | 5 | 6 | 7 | 8 | 9 | 10| 11| 12| 13| 14| 15| 16| 17| 18| 19| 20| 21| 22| 23|
| N | -2| 0 | 1 | 6 | 7 | 8 | 9 | 10| 11| 12| 13| 14| 15| 16| 17| 18| 19| 20| 21| 22| 23|
| D | -2| -2| 1 | 6 | 7 | 8 | 9 | 10| 11| 12| 13| 14| 15| 16| 17| 18| 19| 20| 21| 22| 23|
| C | -3| -3| -3| 9 | 10| 11| 12| 13| 14| 15| 16| 17| 18| 19| 20| 21| 22| 23| 24| 25| 26|
| Q | -1| 0 | 0 | -3| 5 | 6 | 7 | 8 | 9 | 10| 11| 12| 13| 14| 15| 16| 17| 18| 19| 20| 21| 22|
| E | -1| 0 | 2 | -4| 2 | 5 | 8 | 11| 14| 17| 20| 23| 26| 29| 32| 35| 38| 41| 44| 47| 50| 53|
| I | -1| -3| -3| -3| -1| -3| -6| -10| -14| -18| -22| -26| -30| -34| -38| -42| -46| -50| -54| -58| -62| -66|
| L | -1| -2| -3| -4| -1| -2| -3| -4| -5| -6| -7| -8| -9| -10| -11| -12| -13| -14| -15| -16| -17| -18| -19|
| K | -1| 2 | 0 | -1| -3| 1 | 1 | 4 | 7 | 10| 13| 16| 19| 22| 25| 28| 31| 34| 37| 40| 43| 46|
| M | -1| -1| -2| -3| -1| 0 | -2| -3| -2| 2 | 5 | 8 | 11| 14| 17| 20| 23| 26| 29| 32| 35| 38|
| F | -2| -3| -3| -2| -3| -1| -3| -3| -1| 1 | 2 | 4 | 6 | 8 | 10| 12| 14| 16| 18| 20| 22| 24|
| P | -1| -2| -2| -1| -3| -1| -2| -2| -2| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3|
| S | 1 | -1| 1 | 0 | -1| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0|
| T | 0 | -1| 0 | -1| -1| -1| -1| -1| -1| -1| -1| -1| -1| -1| -1| -1| -1| -1| -1| -1| -1| -1|
| W | -3| -3| -4| -4| -2| -2| -3| -2| -3| -2| -3| -2| -3| -2| -3| -2| -3| -2| -3| -2| -3| -2|
| Y | -2| -2| -2| -2| -3| -2| -2| -3| -2| -3| -2| -3| -2| -3| -2| -3| -2| -3| -2| -3| -2| -3|
| V | 0 | -3| -3| -3| -1| -2| -2| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3|
Statistical Significance: E-values

- Let $P_i$ be the frequency of aa $i$. For unrelated sequences, an alignment of $i$ with $j$ has probability $P_{ij}$.
- Given $P_{ij}$ and $s_{ij}$, we can calculate normalized scores ("bit scores") from the raw score, $S$:

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

$K$ is a function of the database size,
$\lambda$ is a function of the scoring matrix

- When 2 random sequences of length $m$ and $n$ are aligned, the expected number of HSPs with normalized scores greater than $S'$ is approximately

$$E = \frac{N}{2^{S'}}$$

where $N = nm$
BLAST & FASTA

- Using heuristic algorithms
- Word based match
- Faster than Smith & Waterman
  

- **FASTA**: [www.ebi.ac.uk/fasta3](http://www.ebi.ac.uk/fasta3)
  
Figure 8.9

(a) The best scores are:

<table>
<thead>
<tr>
<th>Query</th>
<th>Identity</th>
<th>Opt</th>
<th>Z-score</th>
<th>E-value</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi 1706794</td>
<td>996 996</td>
<td>130.5</td>
<td>4.0</td>
<td>59248</td>
<td>996 996</td>
</tr>
<tr>
<td>gi 1703339</td>
<td>935 935</td>
<td>28.2</td>
<td>2.3</td>
<td>5206</td>
<td>431 395</td>
</tr>
<tr>
<td>gi 1723425</td>
<td>290 171</td>
<td>420.1</td>
<td>2.9</td>
<td>158</td>
<td>316 250.7</td>
</tr>
<tr>
<td>gi 1723430</td>
<td>104 157</td>
<td>21.6</td>
<td>1.8</td>
<td>0.0002</td>
<td>315 216.2</td>
</tr>
<tr>
<td>gi 417124</td>
<td>159 140</td>
<td>195.0</td>
<td>0.0002</td>
<td>0.0002</td>
<td>157 190.5</td>
</tr>
<tr>
<td>gi 1169826</td>
<td>128 169.7</td>
<td>77</td>
<td>0.0072</td>
<td>0.0072</td>
<td>183 169.7</td>
</tr>
<tr>
<td>gi 414846</td>
<td>119 166.8</td>
<td>102</td>
<td>0.01</td>
<td>0.0012</td>
<td>102 166.8</td>
</tr>
<tr>
<td>gi 170554</td>
<td>118 164.5</td>
<td>87</td>
<td>0.014</td>
<td>0.0012</td>
<td>87 164.5</td>
</tr>
<tr>
<td>gi 1742020</td>
<td>112 168.7</td>
<td>131</td>
<td>0.02</td>
<td>0.02</td>
<td>131 167.1</td>
</tr>
<tr>
<td>gi 1742039</td>
<td>98</td>
<td>116.1</td>
<td>98</td>
<td>0.02</td>
<td>98 116.1</td>
</tr>
<tr>
<td>gi 1170581</td>
<td>115 160.4</td>
<td>86</td>
<td>0.023</td>
<td>0.023</td>
<td>86 160.4</td>
</tr>
<tr>
<td>gi 1730188</td>
<td>87 159.3</td>
<td>87</td>
<td>0.027</td>
<td>0.027</td>
<td>87 159.3</td>
</tr>
<tr>
<td>gi 1177047</td>
<td>112 156.3</td>
<td>79</td>
<td>0.04</td>
<td>0.04</td>
<td>79 156.3</td>
</tr>
<tr>
<td>gi 1177046</td>
<td>117 154.8</td>
<td>78</td>
<td>0.048</td>
<td>0.048</td>
<td>78 154.8</td>
</tr>
<tr>
<td>gi 1177045</td>
<td>115 154.5</td>
<td>76</td>
<td>0.05</td>
<td>0.05</td>
<td>76 154.5</td>
</tr>
<tr>
<td>gi 140775</td>
<td>109 152.6</td>
<td>65</td>
<td>0.064</td>
<td>0.064</td>
<td>65 152.6</td>
</tr>
<tr>
<td>gi 1169825</td>
<td>104 137.9</td>
<td>62</td>
<td>0.42</td>
<td>0.42</td>
<td>62 137.9</td>
</tr>
<tr>
<td>gi 113999</td>
<td>103 137.1</td>
<td>66</td>
<td>0.47</td>
<td>0.47</td>
<td>66 137.1</td>
</tr>
</tbody>
</table>

(b) Output of a FASTA search. (a) Hit list from a FASTA search with human histidine triad (HIT) protein (SWISS-PROT P49789) as the query against the swissprot database. The search was performed using ktup = 1. (b) Optimal local alignment of the query to one of the database entries (marked by arrow in hit list) containing the sequence of rat galactose-1-phosphate uridylyltransferase (GalT). Although the sequence similarity is weak, these proteins have been shown to share structural similarity.

(C) 2002 SNU CSE Biointelligence Lab (BI)
FASTA

- Score using a substitution matrix
  - Applies a substitution matrix to the 10 best regions
  - The matrix encapsulates the biological significance of word matches
  - Single best subalignment - init1
  - db strings with init1 < threshold are filtered
BLAST Result

Program: BLASTP
Query: human MYB binding protein
Database: Swissprot
The BLAST Search

For the query, find the list of high scoring words of length W.

Compares the word list to the database and identifies exact matches.

For each word match, extends the alignment in both directions to find alignments that score greater than a threshold of value S.
## BLAST Programs

<table>
<thead>
<tr>
<th>Program</th>
<th>Query</th>
<th>Database</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASTP</td>
<td>Protein</td>
<td>Protein</td>
<td>Uses substitution matrix for finding distant relationship relationships; SEG filtering available</td>
</tr>
<tr>
<td>BLASTN</td>
<td>Nucleotide</td>
<td>Nucleotide</td>
<td>Tuned for very high-scoring matches, not distant relationships</td>
</tr>
<tr>
<td>BLASTX</td>
<td>Nucleotide (translated)</td>
<td>Protein</td>
<td>Useful for analysis of new DNA sequences and ESTs</td>
</tr>
<tr>
<td>TBLASTN</td>
<td>Protein</td>
<td>Nucleotide (translated)</td>
<td>Useful for finding unannotated coding regions in database sequences</td>
</tr>
<tr>
<td>TBLASTX</td>
<td>Nucleotide (translated)</td>
<td>Nucleotide (translated)</td>
<td>May be useful for EST analysis, but computationally intensive</td>
</tr>
</tbody>
</table>

(C) 2002 SNU CSE Biointelligence Lab (BI)
### Protein Sequence Databases for use with BLAST

<table>
<thead>
<tr>
<th>Databases</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>nr</em></td>
<td>Non-redundant merge of SWISS-PROT, PIR, PRF, and proteins derived form GenBank coding sequences and PDB atomic coordinates</td>
</tr>
<tr>
<td><em>month</em></td>
<td>Subset of <em>nr</em> witch is new or modified within the last 30 days</td>
</tr>
<tr>
<td><em>swissprot</em></td>
<td>The SWISS-PROT database</td>
</tr>
<tr>
<td><em>pdb</em></td>
<td>Amino acid sequences parsed from atomic coordinates of three-dimensional structures</td>
</tr>
<tr>
<td><em>ecoli</em></td>
<td>Complete set of proteins encoded by the <em>E. coli</em> genome</td>
</tr>
<tr>
<td><em>yeast</em></td>
<td>Complete set of proteins encoded by the <em>S. cerevisiae</em> genome</td>
</tr>
<tr>
<td><em>drosoph</em></td>
<td>Complete set of proteins encoded by the <em>E. melanogaster</em> genome</td>
</tr>
</tbody>
</table>
# Nucleotide Sequence Databases for use with BLAST

<table>
<thead>
<tr>
<th>Databases</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>nr</code></td>
<td>Non-redundant Genbank, excluding the EST, STS, and GSS divisions</td>
</tr>
<tr>
<td><code>month</code></td>
<td>Subset of <code>nr</code>, which is new or modified within the last 30 days</td>
</tr>
<tr>
<td><code>est</code></td>
<td>Genbank EST division (expressed sequence tags)</td>
</tr>
<tr>
<td><code>sts</code></td>
<td>Genbank STS division (sequence tagged sites)</td>
</tr>
<tr>
<td><code>htgs</code></td>
<td>Genbank HTG division (high-throughput genomic sequences)</td>
</tr>
<tr>
<td><code>gss</code></td>
<td>Genbank GSS division (genome survey sequences)</td>
</tr>
<tr>
<td><code>ecoli</code></td>
<td>Complete genome sequence of <em>E. coli</em></td>
</tr>
<tr>
<td><code>yeast</code></td>
<td>Complete genome sequence of <em>S. cerevisiae</em></td>
</tr>
<tr>
<td><code>drosoph</code></td>
<td>Complete genome sequence of <em>E. melanogaster</em></td>
</tr>
<tr>
<td><code>mito</code></td>
<td>Complete genome sequence of vertebrate mitochondria</td>
</tr>
<tr>
<td><code>alu</code></td>
<td>Collection of primate Alu repeat sequences</td>
</tr>
<tr>
<td><code>vector</code></td>
<td>Collection of popular cloning vectors</td>
</tr>
</tbody>
</table>
The protein product of the Alzheimer’s disease susceptibility gene (GenBank L43964) was used as the query in a TBLASTN search against the est database. The goal was to identify cDNA clones from other organisms that may represent homologs of the human gene.
Portion of the hit list showing the 25 best hits. Each sequence is identified by GenBank accession number and a portion of the definition line. The reading frame and score of the best HSP are shown, together with the sum probability of a chance occurrence. The value in the last column gives the number of HSPs that were used in the sum probability calculation. At least 10 sequences from mouse and one from *Drosophila* may be seen on the hit list.

Match to the conceptual translation of the *Drosophila* EST sequence (GenBank AA390557). Two HSPs were found, each in a different reading frame. Identical residues are echoed to the central line, and plus (+) symbols indicate pairs of nonidentical amino acids with positive substitution scores.
Figure 8.11

Output of a TBLASTN search. The protein product of the Alzheimer's disease susceptibility gene (GenBank L43964) was used as the query in a TBLASTN search against the EST database. The goal was to identify cDNA clones from other organisms that may represent homologs of the human gene. (a) Portion of the hit list showing the 25 best hits. Each sequence is identified by GenBank accession number and a portion of the definition line. The reading frame and score of the best HSP are shown, together with the sum probability of a chance occurrence. The value in the last column gives the number of HSPs that were used in the sum probability calculation. At least 10 sequences from mouse and one from Drosophila may be seen on the hit list. (b) Match to the conceptual translation of the Drosophila EST sequence (GenBank AA390557). Two HSPs were found, each in a different reading frame. Identical residues are echoed to the central line, and plus (+) symbols indicate pairs of nonidentical amino acids with positive substitution scores.
Figure 8.12. Identifying low-complexity regions with SEG. Analysis of the human achaete-scute protein (SWISS-PROT P50553) using SEG reveals two regions of low compositional complexity. (a) Program output in the default "tree" format shows the low-complexity sequences in lower-case letters on the left and high-complexity in upper-case on the right. (b) Using the -x command-line switch, the SEG program will generate a version of the sequence in which the low-complexity sequences have been masked. (c) For convenience, the BLAST programs can be instructed to perform the masking automatically. When a masked query sequence is used in a database search, some of the alignments may contain masked segments, as shown in this BLASTP output.
Position Specific Scoring Matrices

- Position-Specific Iterated BLAST (PSI-BLAST)

  Single protein sequence

  Search database (BLAST)

  Profile

  Iterate until convergence

  Multiple alignment

  Estimate statistical significance of local alignments
Sequences producing significant alignments:

**Pass 1:**
- sp P49779 HHT_HUMAN FRAGILE HISTIDINE TRIAD PROTEIN 290 7e-70
- sp P49776 AP81_SCHD B (S'-NUCLEOSYLA)-TETRAPHOSPHATASE (ASYMMETRICAL) 117 8e-27
- sp P49775 YDA5_YES T HYPOTHETICAL 24.8 KD HIT-LIKE PROTEIN 88.0 6e-18
- sp Q14666 HIT_HYCTU HYPOTHETICAL 20.0 KD HIT-LIKE PROTEIN 52.9 3e-07
- sp Q04344 HIT_HYCTU HYPOTHETICAL HIT1 PROTEIN (OFF U) 45.3 4e-05

**Pass 2:**
- sp P47378 YHT_MYCE HYPOTHETICAL 15.6 KD HIT-LIKE PROTEIN 70.5 1e-12
- sp P32083 YHT_MYCHR HYPOTHETICAL 13.1 KD HIT-LIKE PROTEIN IN P... 59.0 3e-09
- sp P47374 YHT_AEGH HYPOTHETICAL 13.2 KD HIT-LIKE PROTEIN IN R... 57.6 9e-09
- sp P47372 YHT_SYNC HYPOTHETICAL 12.4 KD HIT-LIKE PROTEIN IN P... 55.7 3e-08
- sp P47371 YHT_CELE HYPOTHETICAL HIT-LIKE PROTEIN P21C 3. 54.3 9e-08
- sp P47373 YHT_MYCE HYPOTHETICAL 17.0 KD PROTEIN HIT-LIKE PROTEIN... 49.5 2e-06
- sp P47373 IPK1_HUMAN PROTEIN KINASE C INHIBITOR 1 (PKCI-1) 49.1 3e-06
- sp P16350 IPK1_BOVIN PROTEIN KINASE C INHIBITOR 1 (PKCI-1) (17 ... 48.7 4e-05
- sp P47374 YCPF_NAEIN HYPOTHETICAL HIT-LIKE PROTEIN NAO961 47.3 1e-05
- sp P47374 GAL7_RAT GALACTOSE-1-PHOSPHATE URIDYLYLTRANSFERASE 41.0 9e-04

**Pass 3:**
- sp Q03249 GAL7_MOUSE GALACTOSE-1-PHOSPHATE URIDYLYLTRANSFERASE 41.0 9e-04
- sp P79002 GAL7_HUMAN GALACTOSE-1-PHOSPHATE URIDYLYLTRANSFERASE 41.0 9e-04
- sp P79002 GAL7_HUMAN GALACTOSE-1-PHOSPHATE URIDYLYLTRANSFERASE 41.0 9e-04
- sp P47374 GAL7_HAEIN GALACTOSE-1-PHOSPHATE URIDYLYLTRANSFERASE 41.0 9e-04
- sp P47374 GAL7_RCOLL GALACTOSE-1-PHOSPHATE URIDYLYLTRANSFERASE 41.0 9e-04
- sp P47374 GAL7_SALTY GALACTOSE-1-PHOSPHATE URIDYLYLTRANSFERASE 41.0 9e-04
- sp P47374 GAL7_KLULA GALACTOSE-1-PHOSPHATE URIDYLYLTRANSFERASE 41.0 9e-04
- sp P47374 GAL7_KLULA GALACTOSE-1-PHOSPHATE URIDYLYLTRANSFERASE 41.0 9e-04

**Figure 8.13.** Increased sensitivity using PSI-BLAST. The human histidine triad (HIT) protein (SWISS-PROT P49789) was used as the query in a BLASTP search with the PSI-BLAST functionality enabled. Definition lines, scores, and E values are shown for all statistically significant matches newly identified in each iteration.

Table showing alignments and scores for various proteins.

- **Pass 1:**
  - SNU CSE Biointelligence Lab (BI) 2002
Spliced Alignments Problem

- Given strings $G$ (genomic sequence) and $T$ (target sequence), and a set $B$ of substrings of $G$, find a set of non-overlapping strings from $B$ whose concatenation $C$ fits $T$ the best; i.e., the edit distance between $C$ and $T$ is minimal among all sets of blocks from $B$. 

Complement of mRNA

$G = \text{ATCAGTGCAATGCAGCCATGA}$

$T = \text{t}_1\text{t}_2 \ldots \text{t}_m$

$B = \{ \text{b}_1, \text{b}_2, \ldots, \text{b}_p \}$

(C) 2002 SNU CSE Biointelligence Lab (BI)
Figure 8.14. Spliced alignment. The sim4 program was used to align a novel human mRNA (RefSeq NM_015372) to the genomic sequence of a cosmid from chromosome 22 (EMBL Z82248). Three exons were identified on the complementary (the third one has been truncated for brevity). The “>>>” symbols indicate splice sites found at the exon/intron boundaries.
BIOLOGY IN THE FUTURE

Systems biology
Neuroimmunology

Bioinformatics
Data mining
Combinatorial chemistry
Synthetic biology

Biochip & microfluidics (LOC)
Biocomputation