Bioinformatics Chapter 8.

Sequence Alignment and Database Searching

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

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

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

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

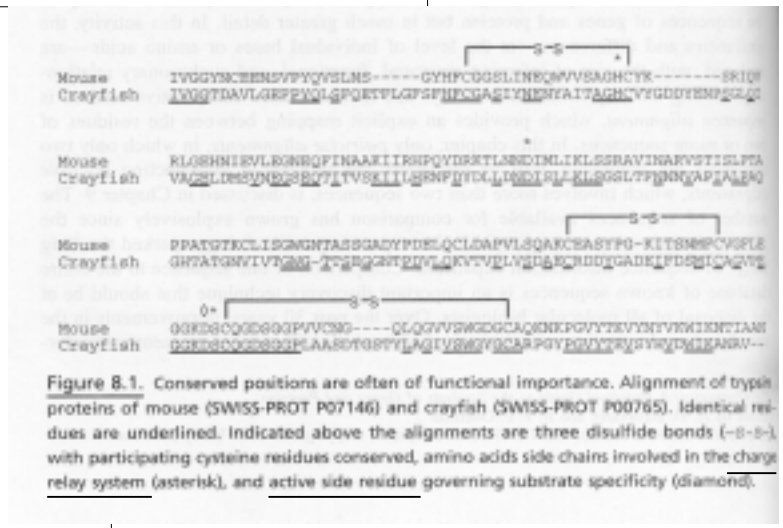
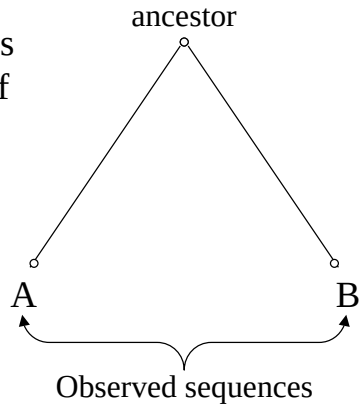


Figure 8.1. Conserved positions are often of functional importance. Alignment of trypsin proteins of mouse (SWISS-PROT P07146) and crayfish (SWISS-PROT P00765). Identical residues are underlined. Indicated above the alignments are three disulfide bonds (–S–S–) with participating cysteine residues conserved, 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*

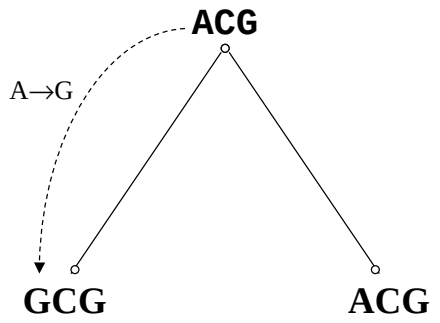


Alignment Jargon

```
GCG
 ||
ACG
```

- 1 mismatch
- 2 matches

Substitution

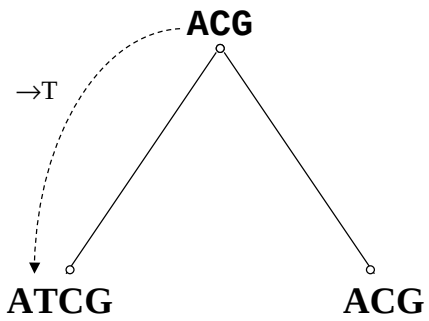


Alignment Jargon

```
ATCG
|  ||
A-CG
```

- 0 mismatches
- 3 matches
- 1 gap

Insertion

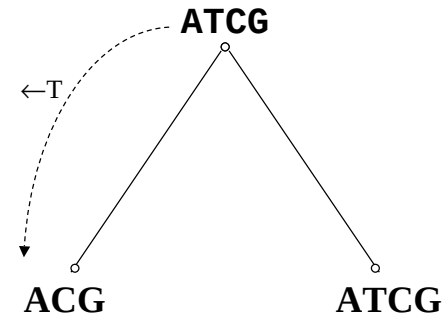


Alignment Jargon

```
ATCG
|  ||
A-CG
```

- 0 mismatches
- 3 matches
- 1 gap

Deletion



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

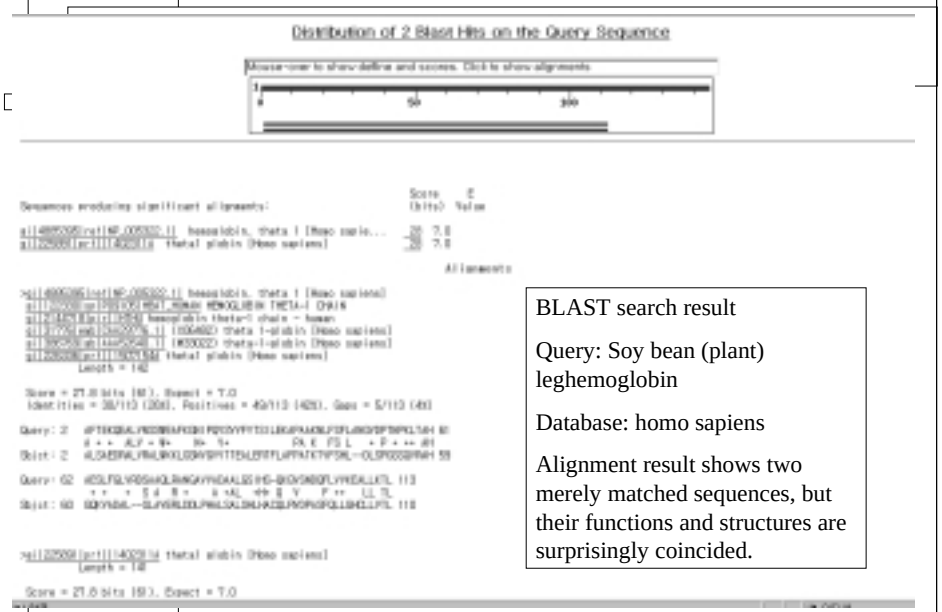
Local Alignment vs. Global Alignment

- Local Alignment
  - Attempts to align regions of sequences with the highest density of matches. In doing so, one of 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

		A	C	C	A	C	A	C	A
	0	0	0	0	0	0	0	0	0
A	0	1	0	0	1	0	1	0	1
C	0	0	2	1	0	2	0	2	0
A	0	1	0	1	2	0	3	1	3
C	0	0	2	1	0	3	1	4	2
C	0	0	0	3	2	1	2	2	3
A	0	1	0	1	4	2	2	1	3
T	0	0	0	0	2	3	1	1	1
A	0	1	0	0	1	1	4	2	2

		A	C	C	A	C	A	C	A
	0	m	m	m	m	m	m	m	m
A	m	1	-1	-3	-5	-7	-9	-11	-13
C	m	-1	2	0	-2	-4	-6	-8	-10
A	m	-3	0	1	1	-1	-3	-5	-7
C	m	-5	-2	1	0	2	0	-2	-4
C	m	-7	-4	-1	0	1	1	1	-1
A	m	-9	-6	-3	0	-1	2	0	2
T	m	-11	-8	-5	-2	-1	0	1	0
A	m	-13	-10	-7	-4	-1	0	-1	2



Optimal Global Sequence Alignment

```
Human-ZCr       MATGQKLMRAVRVFEFGGPEVLKLRSDIAVPIPKDHQVLIKVHACGVNPVETYIRSGTYS
Ecoli-QOR       ------MATRIEFHKHGGPEVLQA-VEFTPADPAENEIQVENKAIGINFIDTYIRSGLYP
                      .  .     ******.   . .   * .... .  .* *.* ..****** *

Human-ZCr       RKPLLPYTPGSDVAGVIEAVGDNASAFKKGDRVFTSSTISGGYAEYALAADHTVYKLPEK
Ecoli-QOR       -PPSLPSGLGTEAAGIVSKVGSGVKHIKAGDRVVYAQSALGAYSSVHNIIADKAAILPAA
                  * **   *.. **..  **  .    * ****  .  .  * *.            **

Human-ZCr       LDFKQGAAIGIPYFTAYRALIHSACVKAGESVLVHGASGGVGLAACQIARAYGLKILGTA
Ecoli-QOR       ISFEQAAASFLKGLTVYYLLRKTYEIKPDEQFLFHAAAGGVGLIACQWAKALGAKLIGTV
                . * * **  .  * *  * ..  .*  *  * * *.***** *** *.* * *..**

Human-ZCr       GTEEGQKIVLQNGAHEVFNHREVNYIDKIKKYVGEKGIDIIIEMLANVNLSKDLSLLSHG
Ecoli-QOR       GTAQKAQSALKAGAWQVINYREEDLVERLKEITGGKKVRVVYDSVGRDTWERSLDCLQRR
                ** .  .  *. ** .* * **   ....*   * * . .. . .   .  * *  .

Human-ZCr       GRVIVVG-SRGTIEINPRDTMAKES----SIIGVTLFSSTKEEFQQYAAALQAGMEIGWL
Ecoli-QOR       GLMVSFGNSSGAVTGVNLGILNQKGSLYVTRPSLQGYITTREELTEASNELFSLIASGVI
                * ..  * * *..        . .      .   .  . .*.**  . .  * .  * .

Human-ZCr       KPVIGSQ--YPLEKVAEAHENIIHGSGATGKMILLL
Ecoli-QOR       KVDVAEQQKYPLKDAQRAHE-ILESRATQGSSLLIP
                *  .  *  ***      *** *.    . *  .*.
```

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

Modular Nature of Proteins

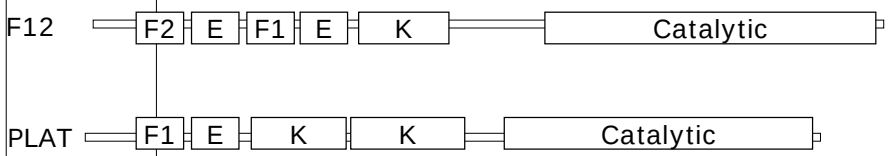


Figure 8.3. Modular structure of two proteins involved in blood clotting. Schematic representation of the modular structure of human tissue plasminogen activator and coagulation factor XII. A module labeled C is shared by several proteins involved in blood clotting. F1 and F2 are frequently repeated 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
   |||
STRQNVINPWAAQ
```

S = 3a

S=alignment score
a=match score

Measuring Alignment Quality

… allow some mismatches

```
QTRPQNVLNPP
   ||| ||
STRQNVINPWAAQ
```

S = 5a - 1b

S=alignment score
a=match score
b=mismatch penalty

Measuring Alignment Quality

…and finally, introduce some gaps

```
QTRPQNVLNPP
  || ||| ||
STR-QNVINPWAAQ
```

S = 7a - 1b -1c

S=alignment score
a=match score
b=mismatch penalty
c=gap penalty

Dot-matrix Representations

- Figure 8.4

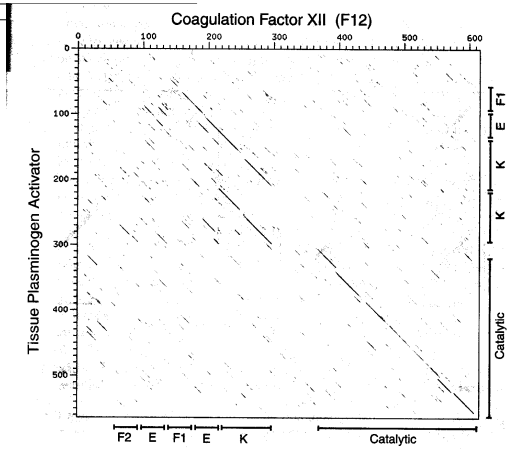


Figure 8.4. Dot matrix sequence comparison. Dot matrix comparison of the amino acid sequences of human coagulation factor XII (F12; SWISS-PROT P00748) and tissue plasminogen activator (PLAT; SWISS-PROT P00750). The figure was generated using the dotter program (Sonnhammer and Durban, 1996).

Dot-matrix Representations

- Figure 8.5

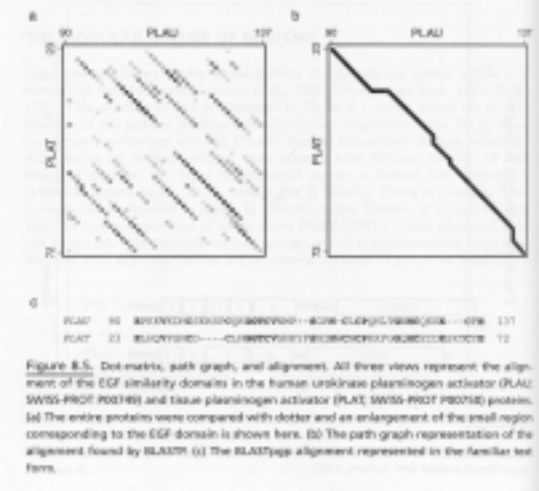


Figure 8.5. Dot-matrix, path graph, and alignment. All three views represent the alignment of the EGF similarity domains in the human urokinase plasminogen activator (PLAU; SWISS-PROT P00749) and tissue plasminogen activator (PLAT; SWISS-PROT P00750) proteins. (a) The entire proteins were compared with 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 BLASTpgp alignment represented in the familiar text form.

Local Alignments

- Many problems in computer science can be reduce 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

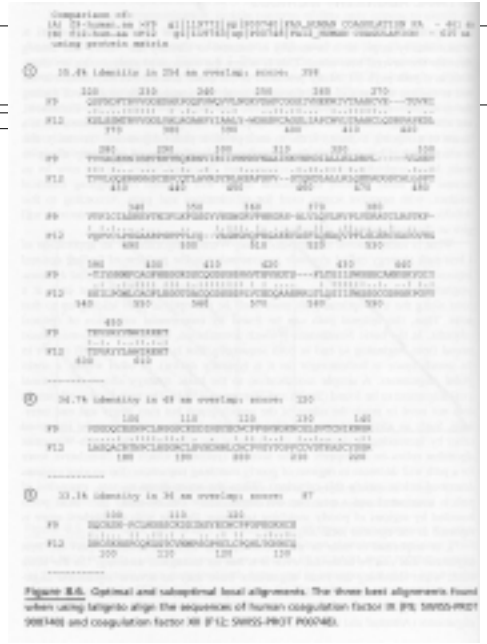
		T	C	A	T	G
	0	0	0	0	0	0
C	0	0	1	0	0	0
A	0	0	0	2	1	0
T	0	1	0	0	3	2
T	0	1	1	0	2	3
G	0	0	1	1	1	3

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

TCAT*G
*CATTG

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
- Ref : Smith, T. F. and Watermann, M. S. (1981)
  Identification of common molecular subsequence.
  Journal of Molecular Biology, **147**: 196-197.



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	-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	-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

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																			
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	2	-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	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	1									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-2	-3	-3	-3	-1	-3	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-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	-1	-2	1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

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,
λ 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 $E = N / 2^{S'}$, where $N = nm$

BLAST & FASTA

- Using heuristic algorithms
- Word based match
- Faster than Smith &Waterman
- **BLAST**: www.ncbi.nlm.nih.gov/BLAST
  Ref: Altschul, S. F., *et al.* (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403-410.
- **FASTA**: www.ebi.ac.uk/fasta3
  Ref: Wilbur, W. J. and Lipman, D. J. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences USA* **80**, 726-30.

- Figure 8.9

a

```
The best scores are:                                      initn init1  opt z-sc   E(59248)
gi|1706794|sp|P49789|FHIT_HUMAN FRAGILE HISTIDINE       996 996   996 1350.4 0
gi|1703339|sp|P49776|APH1_SCHPO BIS(5'-NUCLEOSYL)       431 395   395 536.2  2.8e-23
gi|1724025|sp|P49775|YD15_YEAST HYPOTHETICAL 24.8       290 171   316 428.1  2.9e-17
gi|1724021|sp|Q11066|YHIT_MYCTU HYPOTHETICAL 20.0       178 178   184 250.7  2.2e-07
gi|417124|sp|Q04344|HIT_YEAST HIT1 PROTEIN (ORF U       159 104   157 216.2  1.8e-05
gi|418447|sp|P32084|YHIT_SYNP7 HYPOTHETICAL 12.4        139 139   140 195.0  0.00028
gi|1351828|sp|P47378|YHIT_MYCGR HYPOTHETICAL 15.6       132 132   133 183.9  0.0012
gi|1169826|sp|P43424|GAL7_RAT GALACTOSE-1-PHOSPHA        97  97   128 169.7  0.0072
gi|418446|sp|P32083|YHIT_MYCHR HYPOTHETICAL 13.1        102 102   119 166.8  0.01
gi|1708543|sp|P49773|IPK1_HUMAN PROTEIN KINASE C         87  87   118 164.5  0.014
gi|1724020|sp|P49774|YHIT_MYCLE HYPOTHETICAL 17.0       131  82   117 161.6  0.02
gi|1724019|sp|P53795|YHIT_CAEEL HYPOTHETICAL HIT-        98  98   116 161.5  0.02
gi|1170581|sp|P16436|IPK1_BOVIN PROTEIN KINASE C         86  86   115 160.4  0.023
gi|1730188|sp|Q03249|GAL7_MOUSE GALACTOSE-1-PHOSP        87  87   120 159.3  0.027
gi|1177047|sp|P42856|ZB14_MAIZE 14 KD ZINC-BINDIN       132  79   112 156.3  0.04
gi|120908|sp|P07902|GAL7_HUMAN GALACTOSE-1-PHOSPH        78  78   117 154.8  0.048
gi|1177046|sp|P42855|ZB14_BRAJU 14 KD ZINC-BINDIN       115  76   110 154.5  0.05
gi|140775|sp|P26724|YHIT_AZORR HYPOTHETICAL 13.2        115  65   109 152.6  0.064
gi|1169825|sp|P31764|GAL7_HABIN GALACTOSE-1-PHOSP        62  62   104 137.9  0.42
gi|113999|sp|P16550|APA1_YEAST 5',5'''-P-1,P-4-TE       108  66   103 137.1  0.47
```

b

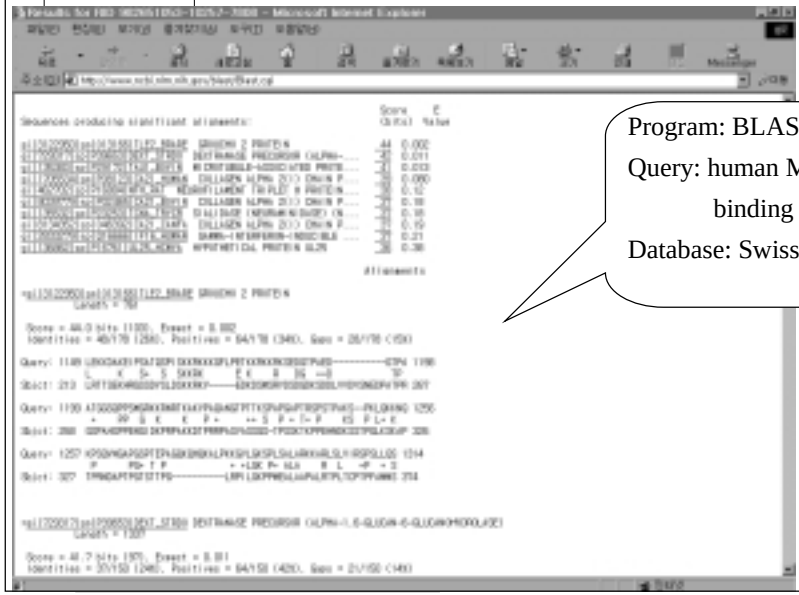
```
>>gi|1169826|sp|P43424|GAL7_RAT GALACTOSE-1-PHOSPHATE UR (379 aa)
initn:   97 init1:   97  opt:  128 z-score: 169.7 E(): 0.0072
Smith-Waterman score: 128;    30.8% identity in 107 aa overlap

                                        10        20        30
HIT                                 MSFRFG-QHLIKPSVVFLKTELSFALVNRKPV
                                    ...: X:...  . : .  ..::    :
GalT   VWASNFLPDIAQREERSQQTYHNQHGKPLLLKYGHQELLRKERLVLTSSYTWIVLVPFWAV
     190       200       210       220       230       240

               40        50        60        70        80
HIT    VPGHVLVCPLRPVERFHDLRPDEVADLFQTTQRVGTVVEKHFHGTSLTFSM--QDGP---
       : ..:. : : :.:. .: : :  :: .: ... :  .. X. :: .:  . .:
GalT   WPFQYLLLPRRHVQRLPELTPAERDDLASTMKKLLTKYDNLFE-TSFPYSMGWHGAPMGL
     250       260       270       280       290       300

         90        100       110       120       130       140
HIT    EAGQTVKH--VHVHVLPRKAGDFHRNDSIYEELQKHDKEDFPASWRSEEEMAAEAAALRV
       ..: : : :..:  :
GalT   KTGAYCDHWQLHAHYYPLLRSATVRKFMVGYEMLAQAQRDLTPEQAAERLRVLPEVHYC
       310       320       330       340       350       360
```

Figure 8.9. 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.

FASTA

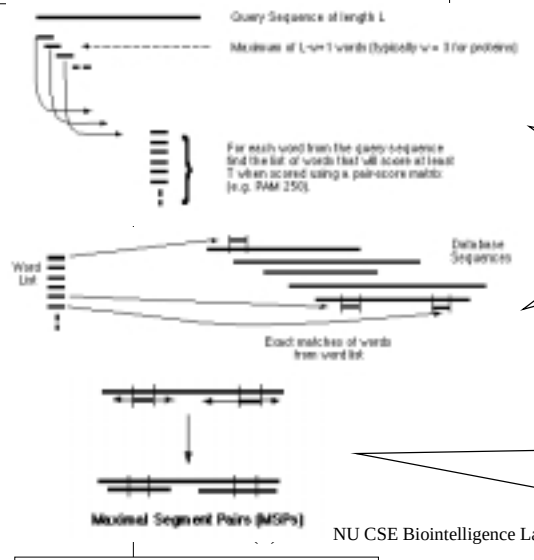
- Score using a substitution matrix
  - ♦Applies a substitution matrix to the 10 best regions
  - ♦The matrix encapsulates the biological significance of word matches
  - ♦Single best subalignment - init1
  - ♦db strings with init1 < threshold are filtered

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

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

Protein Sequence Databases for use with BLAST

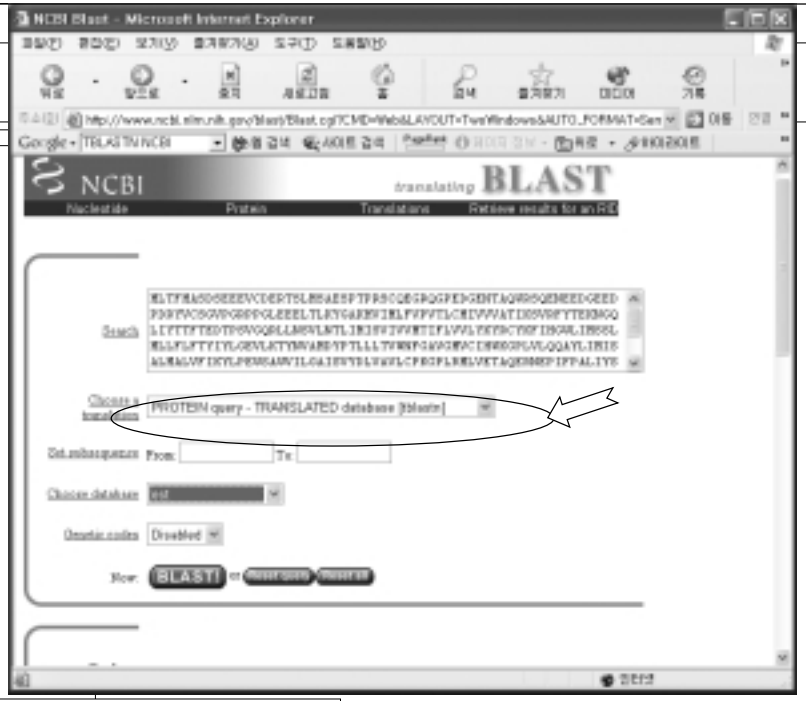
Databas	Description
nr	Non-redundant merge of SWISS-PROT, PIR, PRF, and proteins derived form GenBank coding sequences and PDB atomic coordinates
month	Subset of *nr* witch is new or modified within the last 30 days
swissprot	The SWISS-PROT database
pdb	Amino acid sequences parsed from atomic coordinates of three-dimensional structures
ecoli	Complete set of proteins encoded by the *E. coli* genome
yeast	Complete set of proteins encoded by the *S. cerevisiae* genome
drosoph	Complete set of proteins encoded by the *E. melanogaster* genome

Nucleotide Sequence Databases for use with BLAST

Databas	Description
nr	Non-redundant Genbank, excluding the EST, STS, and GSS divisions
month	Subset of *nr*, witch is new or modified within the last 30 days
est	Genbank EST division (expressed sequence tags)
sts	Genbank STS division (sequence tagged sites)
htgs	Genbank HTG division (high-throughput genomic sequences)
gss	Genbank GSS division (genome survey sequences)
ecoli	Complete genome sequence of *E. coli*
yeast	Complete genome sequence of *S. cerevisiae*
drosoph	Complete genome sequence of *E. melanogaster*
mito	Complete genome sequence of vertebrate mitochondria
alu	Collection of primate Alu repeat sequences
vector	Collection of popular cloning vectors

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.



- 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 sequence 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.

a

Sequences producing High-scoring Segment Pairs:

	Reading Frame	High Score	Smallest Sum Probability P(N)	N
gb|AA056325|AA056325 zf53a03.s1 Soares retina N2b4HR N...	+3	724	3.4e-102	2
gb|T03796|T03796 IB913 Infant brain, Bento Soares ...	+3	567	2.6e-78	2
gb|AA260597|AA260597 mx76g09.r1 Soares mouse NML Mus m...	+2	239	4.9e-53	4
gb|H86456|H86456 yt01b06.s1 Homo sapiens cDNA clon...	+2	323	4.3e-52	4
gb|N24576|N24576 yx72a04.s1 Homo sapiens cDNA clon...	+1	365	5.5e-47	2
gb|AA265273|AA265273 mx91c12.r1 Soares mouse NML Mus m...	+2	239	6.4e-41	2
gb|AA237206|AA237206 mx18e01.r1 Soares mouse NML Mus m...	+3	159	1.5e-40	3
gb|R14600|R14600 yf34b10.r1 Homo sapiens cDNA clon...	+1	278	1.5e-40	2
gb|AA200706|AA200706 mu03f12.r1 Soares mouse 3NbMS Mus...	+1	343	1.9e-40	1
gb|AA045064|AA045064 zk77f12.s1 Soares pregnant uterus...	-3	269	2.3e-37	2
gb|AA087434|AA087434 mm28a04.r1 Stratagene mouse skin ...	+3	322	3.6e-37	1
gb|R05907|R05907 ye93h02.r1 Homo sapiens cDNA clon...	+3	252	7.7e-37	2
gb|AA268820|AA268820 vb01c10.r1 Soares mouse NML Mus m...	+2	234	7.7e-35	2
gb|AA162310|AA162310 mn44a07.r1 Beddington mouse embry...	+1	134	8.3e-34	3
gb|N27820|N27820 yx54h10.r1 Homo sapiens cDNA clon...	+3	154	7.8e-29	2
gb|AA234907|AA234907 zs38f03.r1 Soares NbHMPu S1 Homo ...	+2	155	1.8e-28	2
gb|AA231081|AA231081 mw11d11.r1 Soares mouse 3NMEI2 5 ...	+3	134	8.8e-23	2
gb|H91652|H91652 ys80c04.s1 Homo sapiens cDNA clon...	-3	215	3.7e-22	1
gb|H50532|H50532 yo30h08.s1 Homo sapiens cDNA clon...	-2	211	1.2e-21	1
gb|AA150236|AA150236 zl03c01.r1 Soares pregnant uterus...	+1	159	5.0e-21	2
gb|AA144382|AA144382 mr15d12.r1 Soares mouse 3NbMS Mus...	+3	159	7.6e-21	2
➔ gb|AA390557|AA390557 LD09473.5prime LD Drosophila Embr...	+3	130		
gb|AA210480|AA210480 mo86b03.r1 Beddington mouse embry...	+2	128		
gb|H19012|H19012 ym44b02.r1 Homo sapiens cDNA clon...	+2	134		
gb|AA283084|AA283084 zt14g09.s1 Soares NbHTGBC Homo sa...	-3	175		
gb|H25759|H25759 yl49d01.s1 Homo sapiens cDNA clon...	-2	185		
gb|H33787|H33787 EST110123 Rattus sp. cDNA 5' end ...	+1	137		
gb|AA201988|AA201988 LD05058.5prime LD Drosophila Embr...	+3	175		
gb|AA263526|AA263526 LD06652.5prime LD Drosophila Embr...	+1	167		
gb|H46340|H46340 yj52c04.s1 Homo sapiens cDNA clon...	-1	151		
gb|AA246675|AA246675 LD05588.5prime LD Drosophila Embr...	+2	117		
gb|AA282899|AA282899 zt14g09.r1 Soares NbHTGBC Homo sa...	+3	118		
gb|AA247705|AA247705 csh0941.seq.F Human fetal heart, ...	+3	56		

b

```
gb|AA390557|AA390557 LD09473.5prime LD Drosophila Embryo Drosophila
          melanogaster cDNA clone LD09473 5'
          Length = 659

 Score = 130 (60.4 bits), Expect = 1.6e-20, Sum P(2) = 1.6e-20
 Identities = 25/60 (41%), Positives = 40/60 (66%), Frame = +3

Query:   105 TIKSVRFYTEKNGQLIYTFTFTEDTPSVGQRLLNSVLNTLIMISVIVVMVIFLVVLYKYRC 164
             +I S+ FY   +  L+YT F E +P    +  +++ ++LI++SV+VVMT  L+VLYK RC
Sbjct:   480 SINSISFYNSTDVYLLYTPFHEQSPEPSVKPWSALQSSLILMSVVVVMTFLLIVLYKKRC 659

 Score = 117 (54.3 bits), Expect = 1.6e-20, Sum P(2) = 1.6e-20
 Identities = 23/30 (76%), Positives = 27/30 (90%), Frame = +1

Query:    75 LKEELTLKYGAKHVIMLFVPVTLCMIVVVA 104
             +KEE  LKYGA+HVI LFVPV+LCM+VVVA
Sbjct:   391 MSEEQGLKYGAQHVIKLFVPVSLCMLVVVA 480
```

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 8.11

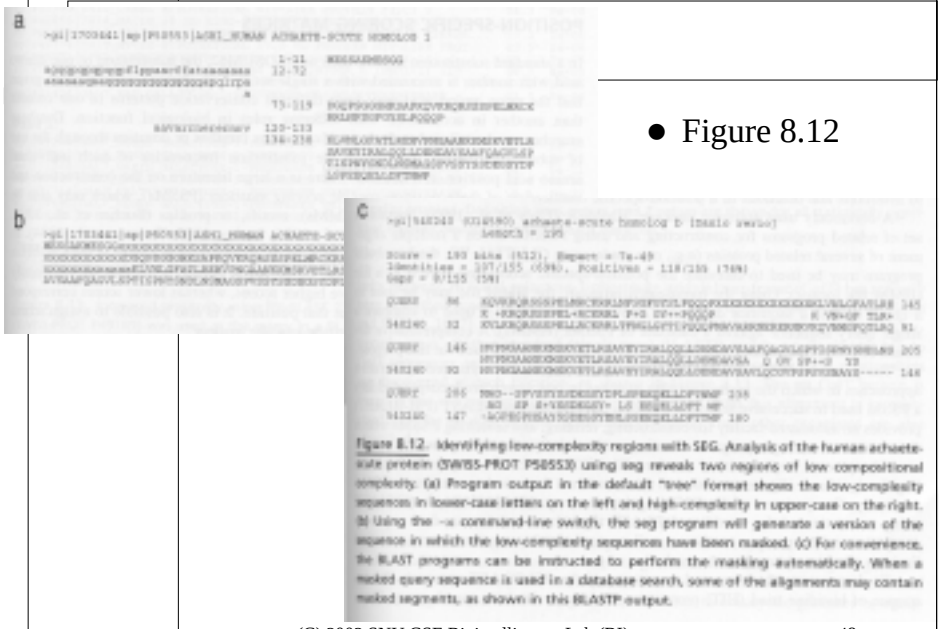
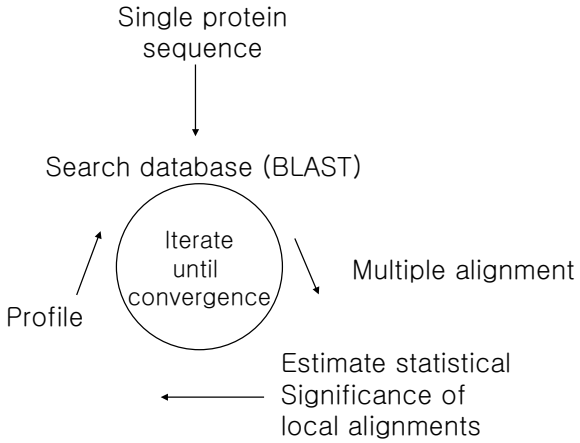


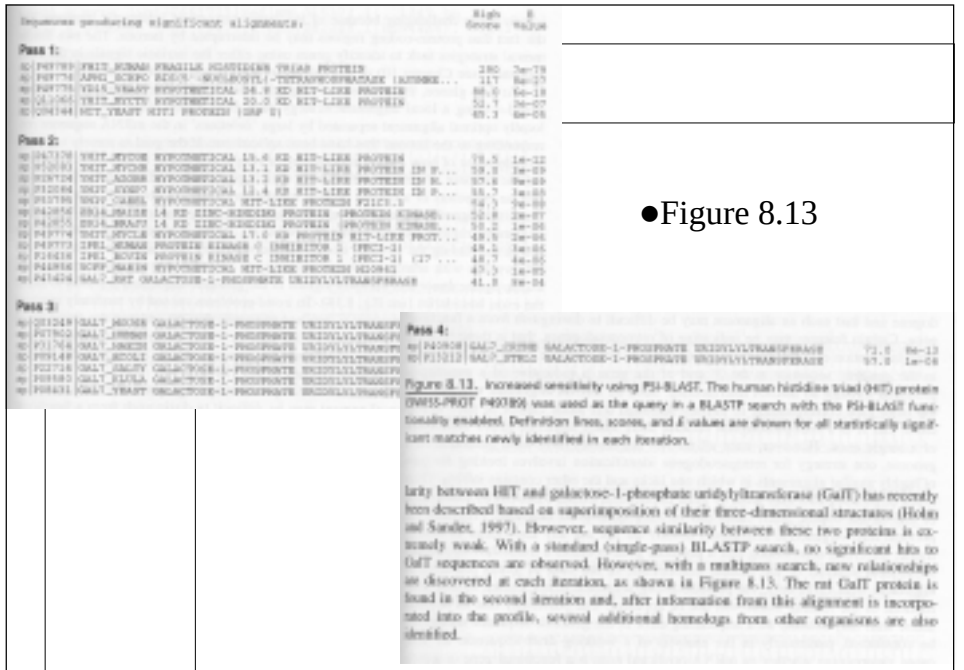
Figure 8.12. Identifying low-complexity regions with SEG. Analysis of the human achaete-scute protein (SWISS-PROT P50553) using seg reveals two regions of low compositional complexity. (a) Program output in the default "free" format shown the low-complexity sequences in lower-case letters on the left 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.

- Figure 8.12

Position Specific Scoring Matrices

- Position-Specific Iterated BLAST (PSI-BLAST)





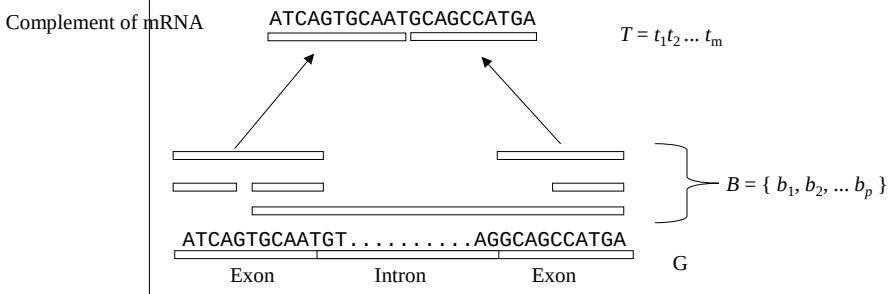
●Figure 8.13

Figure 8.13. Increased sensitivity using PSI-BLAST. The human histidine triad (HIT) protein (DWISS-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.

larity between HIT and galactose-1-phosphate uridylyltransferase (GalT) has recently been described based on superimposition of their three-dimensional structures (Holm and Sander, 1997). However, sequence similarity between these two proteins is extremely weak. With a standard (single-pass) BLASTP search, no significant hits to GalT sequences are observed. However, with a multipass search, new relationships are discovered at each iteration, as shown in Figure 8.13. The rat GalT protein is found in the second iteration and, after information from this alignment is incorporated into the profile, several additional homologs from other organisms are also identified.

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 betwe en *C* and *T* is minimal among all sets of blocks from B.



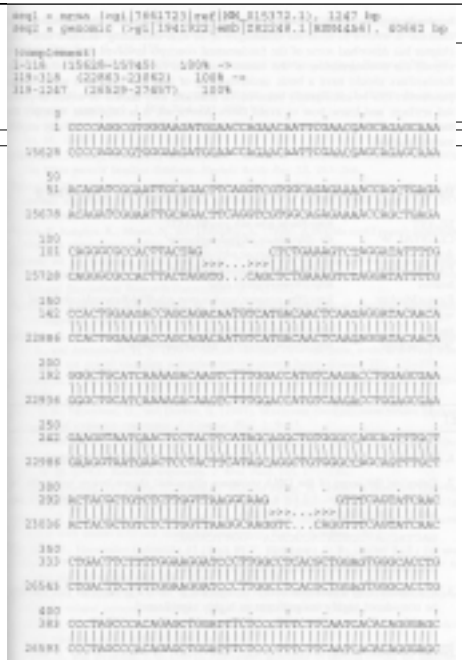


Figure 8.14. Spliced alignment. The sim4 program was used to align a novel human mRNA (RefSeq NM_015372) to the genomic sequence of a cosmid from chromosome 22 (EMBL Z82248). Three exons were identified on the complementary (the third one has been truncated for brevity). The ">>>" symbols indicate splice 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