Bioinformatics

Pairwise sequence alignment

Bioinformatics - lectures

- Introduction
- Information networks
- Protein information resources
- Genome information resources
- DNA sequence analysis
- Pairwise sequence alignment
- Multiple sequence alignment
- Secondary database searching
- Analysis packages
- Protein structure modelling

Pairwise sequence alignment

- database searching
- alphabets and complexity
- algorithms and programs
- sequences and sub-sequences
- identity and similarity
- dotplot
- local and global similarity
- pairwise database searching

Database searching

- Database search can take a form of text queries or sequence similarity searches.
- Text queries are problematic due to missing annotations in many sequences.
- query sequence = probe
  searched sequence = subject
- The purpose of searches is to identify evolutionary relationships (homology) from sequence similarity. Important for search of analogous family members in different species.

Alphabets and complexity

- A sequence consists of letters from an alphabet.
- The complexity of the alphabet is defined by the number of letters it contains:
  - DNA = 4
  - EST = 5
  - proteins = 20
- Special letters can be used for ambiguous bases (N) or residues (X). Sequence searching programs must be able to deal with them.

Algorithms and programs

- Algorithm is a set of steps that define a certain computational process.
- Program is a the implementation of the algorithm.
- Same algorithm may be implemented in many programs.

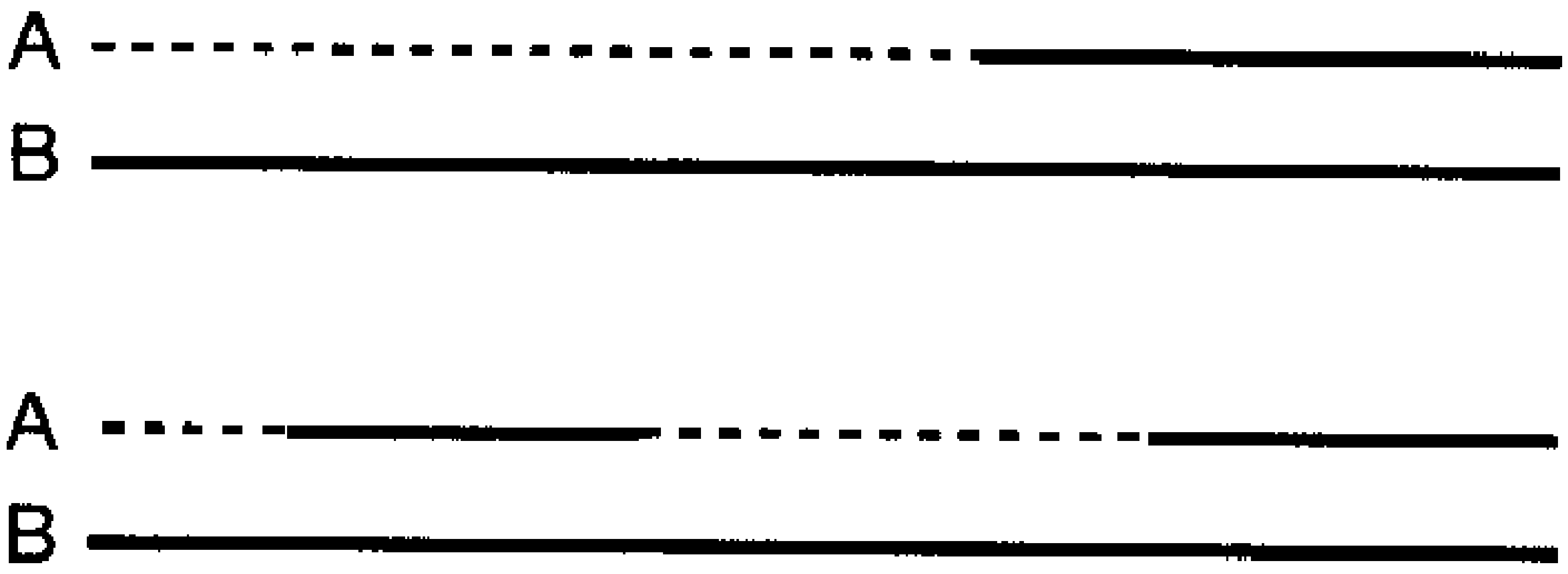
Sequences and sub-sequences

- Alignment of two short sequences:

```
Unaligned                                 score = 6
Sequence 1 (query)         AGGVLIIQVG
                           ||||||
Sequence 2 (subject)       AGGVLIQVG

Aligned                                   score = 9
Sequence 1 (query)         AGGVLIIQVG
                           |||||| |||
Sequence 2 (subject)       AGGVLI-QVG
```

- Score increases by the insertion of a gap. The gap increases the number of aligned identical residues.

Alignment of a sub-sequence with full sequence



Identity and similarity

- Introduction of gaps solely to maximise identities is not biologically meaningful.
- Scoring penalties are introduced to minimise opening and extension of gaps.
- Unitary matrix (counting identities) is replaced by similarity matrix (counting similarities) = high-scoring matches are replaced by biologically meaningful low-scoring matches.
- Diagnostic power of similarity matrices is higher.

(a)

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

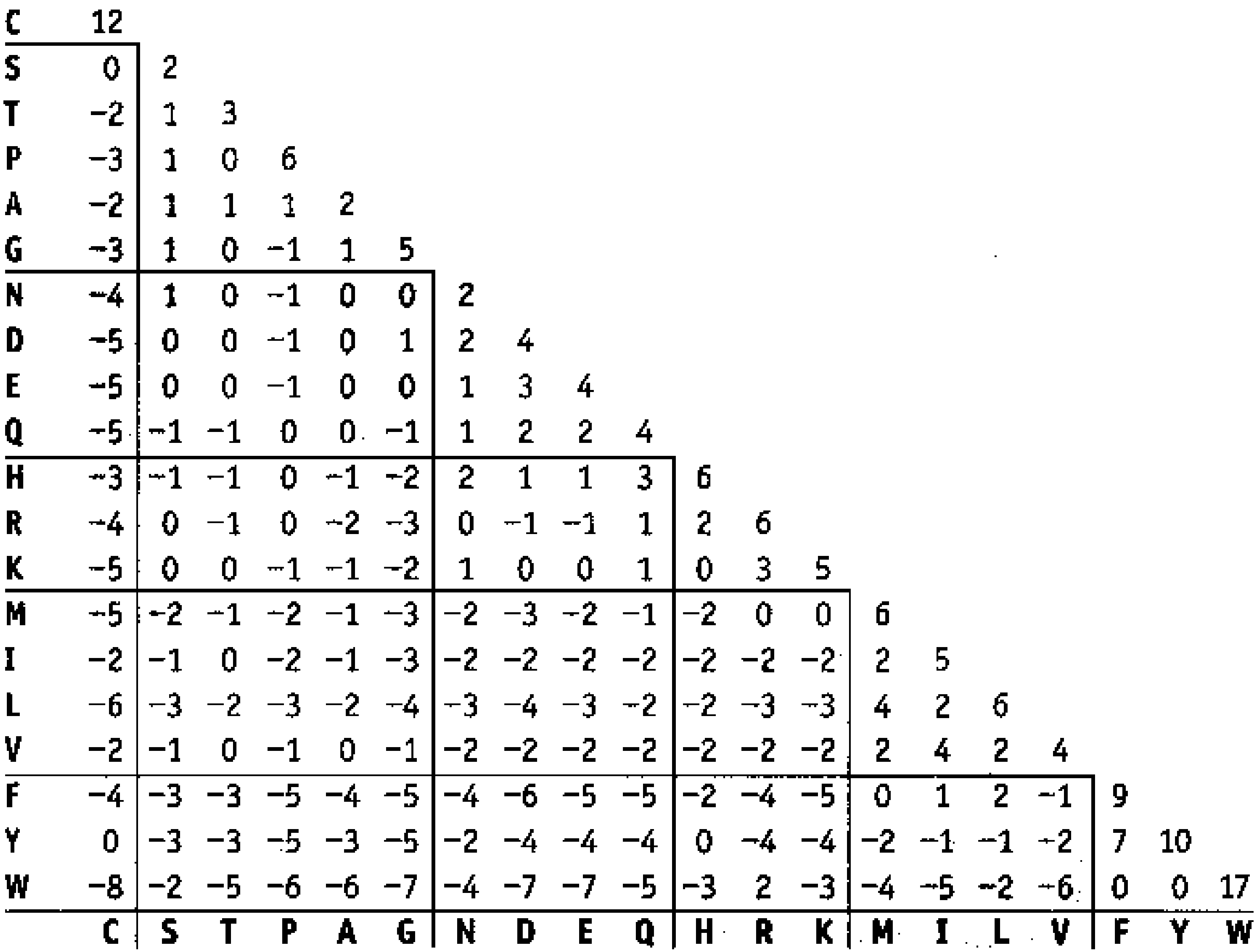
(b)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	B	Z	X
C	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Identity and similarity

- Dayhoff Mutation Data Matrix
  - score is based on the concept of Point Accepted Mutation (PAM)
  - evolutionary distance 1 PAM = probability of a residue mutating during a distance in which 1 point mutation is accepted per 100 residues
  - 250 PAM matrix - similarity score equivalent to 20% matches remaining between two sequences = suitable for identification of similarities in twilight zone
  - limitation: derived from alignment of sequences >85% identical

Mutation Data Matrix for 250 PAMs



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

Identity and similarity

- BLOSUM matrices
  - BLOcks SUbstitution Matrix
  - derived from blocks of aligned sequences in BLOCKS database - represents distant relationships implicitly
  - bias from identical sequences is removed by clustering
  - BLOSUM62 = matrix derived from sequences clustered at 62% or greater identity

Identity and similarity

- Statistical measures of alignment significance
  - performing sequence alignment computationally = creating match according to mathematical model
  - adjustable parameters: gap penalties, impact of sequence length, effect of alphabet complexity
  - level of confidence to constructed alignment is quantified by statistical parameters:

    probability (p) - probability that the constructed alignment arose by chance [should approach 0]

    expected frequency (E) - number of hits one can expect to see by chance [should be <0.001]

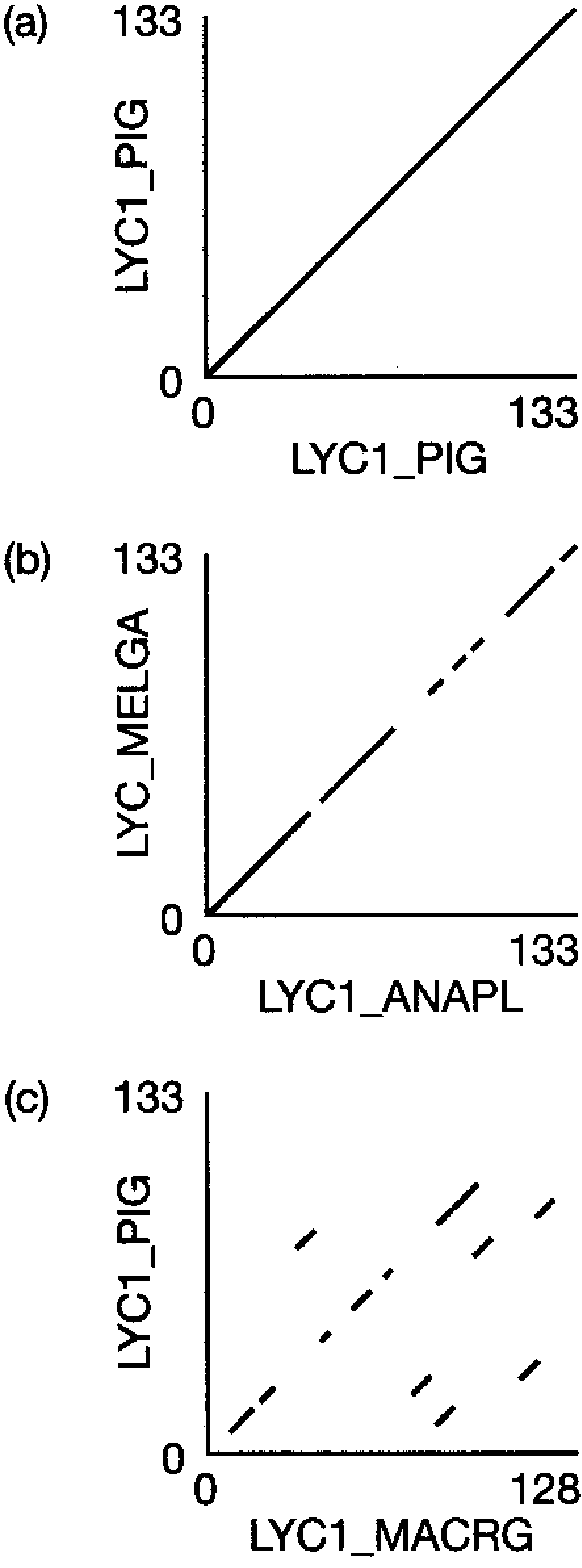
Example hit list from a database search

```
                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

sp|P51698|LINB_PSEPA (LINB)1,3,4,6-tetrachloro-1,4-cyclohexadie...  616  e-176
sp|Q50642|YP79_MYCTU (RV2579..)Hypothetical 33.7 kDa protein Rv...  450  e-126
sp|P27652|LUCI_RENRE Renilla-luciferin 2-monooxygenase (EC 1.13...  218  2e-56
sp|Q50600|YJ33_MYCTU (RV1833C..)Hypothetical 32.2 kDa protein R...  102  8e-22
sp|Q50670|YM96_MYCTU (RV2296..)Putative haloalkane dehalogenase...   93  7e-19
sp|P22643|HALO_XANAU (DHLA)Haloalkane dehalogenase (EC 3.8.1.5)...   87  5e-17
sp|P34913|HYES_HUMAN (EPHX2)Soluble epoxide hydrolase (SEH) (EC...   49  2e-05
sp|O07214|YR15_MYCTU (RV2715..)Hypothetical 36.9 kDa protein Rv...   47  5e-05
sp|Q50599|YI34_MYCTU (RV1834..)Hypothetical 31.7 kDa protein Rv...   45  2e-04
sp|O31158|PRXC_PSEFL (CPO..)Non-heme chloroperoxidase (EC 1.11....   45  2e-04
sp|P22862|ESTE_PSEFL Arylesterase (EC 3.1.1.2) (Aryl-ester hydr...   44  6e-04
sp|P23106|XYLF_PSEPU (XYLF)2-hydroxymuconic semialdehyde hydrol...   40  0.008
sp|P29715|BPA2_STRAU (BPOA2)Non-haem bromoperoxidase BPO-A2 (EC...   39  0.011
sp|P49323|PRXC_STRLI (CPO..)Non-heme chloroperoxidase (EC 1.11....   37  0.054
sp|P54549|YQJL_BACSU (YQJL)Hypothetical 28.2 kDa protein in GLN...   36  0.093
sp|P48972|MYBB_MOUSE (MYBL2..)Myb-related protein B (B-Myb).[Mu...   36  0.12
sp|Q55921|PRXC_SYNY3 (SLR0314)Putative non-heme chloroperoxidas...   36  0.16
sp|Q9JZR6|PIP_NEIMB (PIP..)Proline iminopeptidase (EC 3.4.11.5)...   36  0.16
sp|O13912|YDW6_SCHPO (SPAC23C11.06C)Hypothetical 60.1 kDa prote...   34  0.47
sp|Q59695|ACOC_PSEPU (ACOC)Dihydrolipoamide acetyltransferase c...   34  0.62
sp|P46544|PIP_LACDE (PEPIP)Proline iminopeptidase (EC 3.4.11.5)...   33  1.1
sp|P46542|PIP_LACDL (PIP..)Proline iminopeptidase (EC 3.4.11.5)...   32  1.4
sp|P10244|MYBB_HUMAN (MYBL2..)Myb-related protein B (B-Myb).[Ho...   30  9.2
sp|Q15811|ITSN_HUMAN (ITSN..)Intersectin (SH3 domain-containing...   30  9.2
```

Dotplot

- The most basic visual method for comparison of two sequences.
- Separates noise (random dots) from the signal (adjacent dots).
- Identical sequences are represented by single central diagonal line, similar sequences by a broken diagonal and dissimilar sequences by random dots.
- Advanced dotplots utilise similarity matrices for calculation of cell scores.

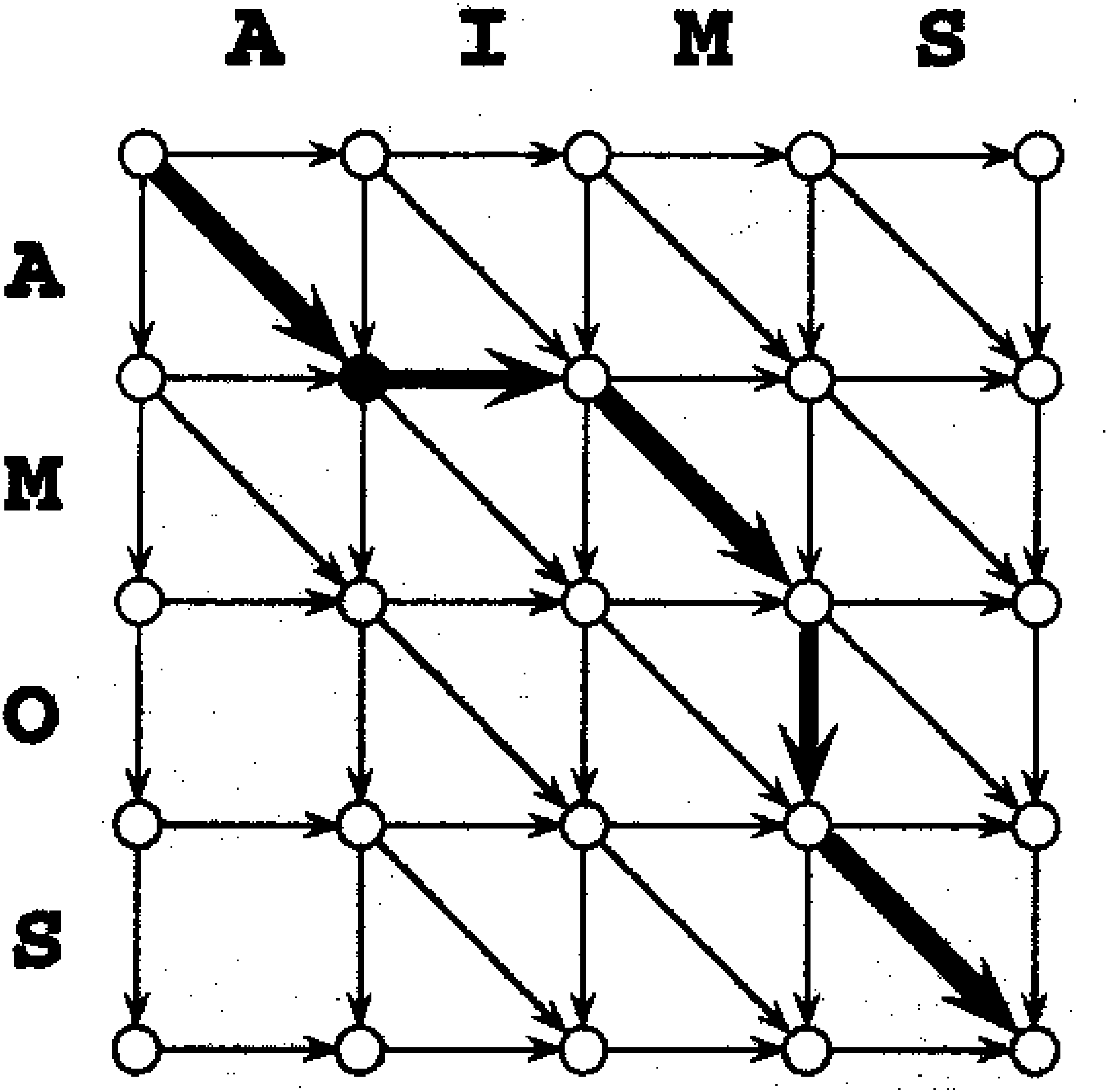
	M	T	F	R	D	L	L	S	V	S	F	E	G	P	R	P	D	S	S	A	G	G	S	S	A	G	G
M	X																										
T		X																									
F			X								X																
R				X											X												
D					X												X										
L						X	X																				
L						X	X																				
S								X		X								X	X				X	X			
V									X																		
S								X		X								X	X				X	X			
F			X								X																
E												X															
G													X								X	X				X	X
P														X		X											
R				X											X												
P														X		X											
D					X												X										
S								X		X								X	X				X	X			
S								X		X								X	X				X	X			
A																				X					X		
G													X								X	X				X	X
G													X								X	X				X	X


(a)
133
LYC1_PIG
0
0
133
LYC1_PIG
(b)
133
LYC_MELGA
0
0
133
LYC1_ANAPL
(c)
133
LYC1_PIG
0
0
128
LYC1_MACRG

Local and global similarity

- Alignments are mathematical models whose behaviour can be modified through the use of adjustable parameters. The models constructed by dynamic programming algorithms - finding solution of a problem by solving smaller, but similar sub-problems.

- Global alignment - considers similarity across the entire sequence.
- Local alignment - considers similarity in parts of sequences only.



Alignment

AIM-S
A-MOS

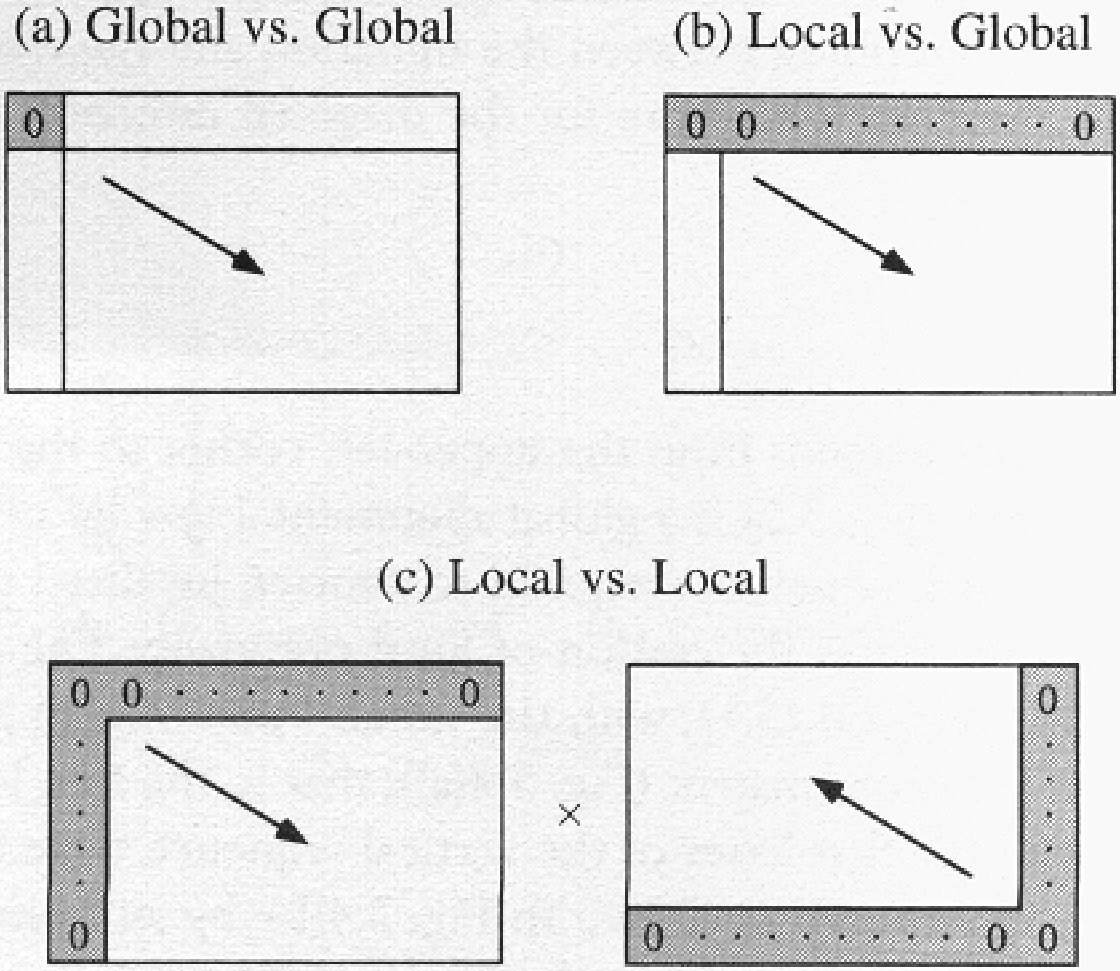
Local and global similarity

- Global alignment
  - Needleman and Wunsch algorithm
  - suitable for sequences similar across most of their length (usually closely related)

  - 1. construction of 2D similarity matrix (“dotplot”)
  - 2. successive summation of the cells in the matrix starting from N-terminal end → progressing through the sequence
  - 3. construction of maximum-match path through the entire sequence

Local and global similarity

- Local alignment
  - Smith-Waterman algorithm
  - suitable for distantly related sequences displaying local regions of similarity (functionally-relevant or structurally-relevant)
  - each point of the matrix defines the end point of a potential alignment = edge cells of the matrix are initialised to 0
  - possibility for ending the alignment are calculated for every cell
  - algorithm is much faster compared to global similarity algorithms



(a) Global vs. Global

(b) Local vs. Global

(c) Local vs. Local

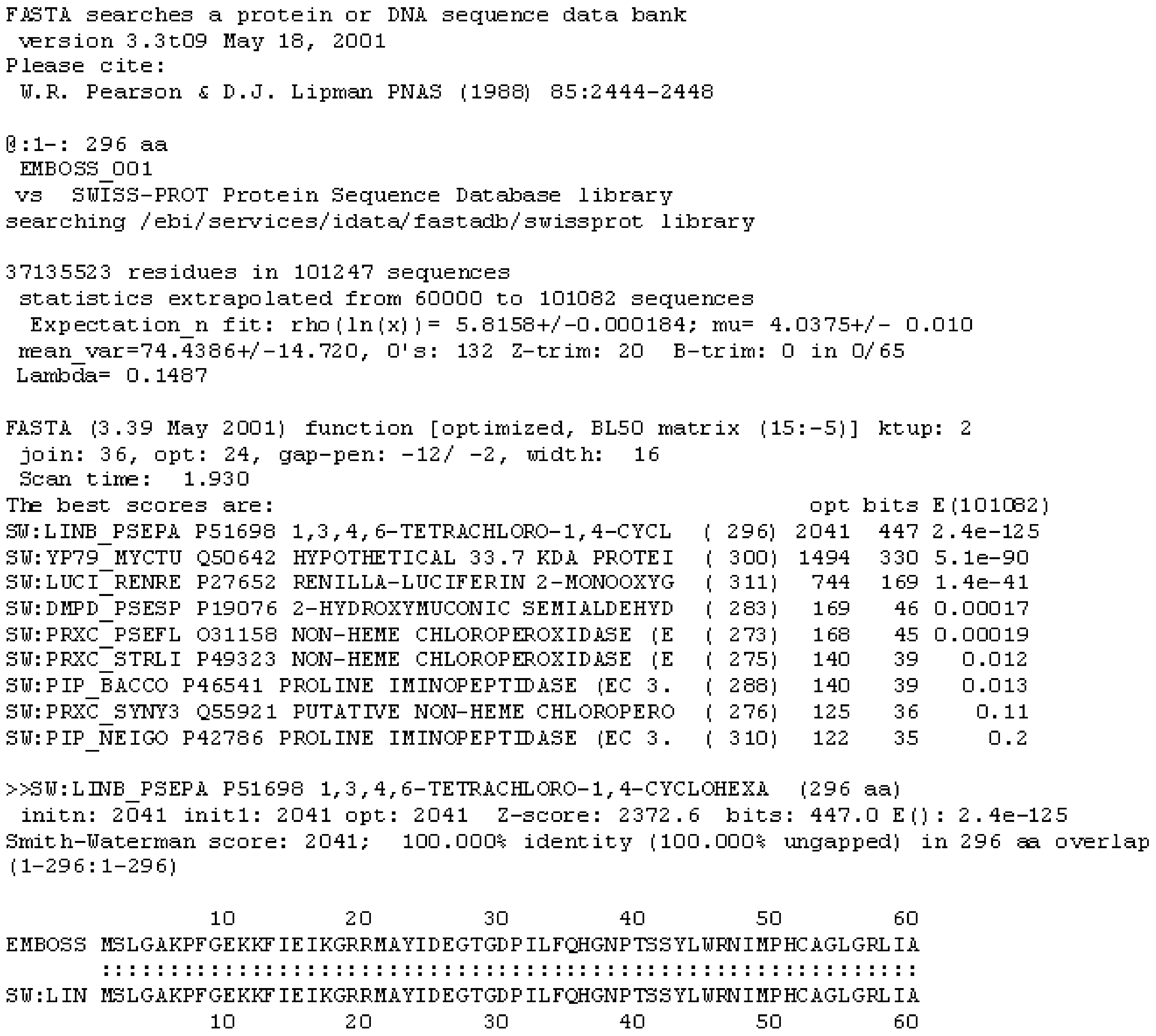
Pairwise database searching

- Extension of the pairwise sequence alignments.
- Large database searches can not be performed using the original Needleman and Wunsch or Smith-Waterman algorithms due to time limitations.
- Very fast local-similarity search methods employing heuristics = FastA and BLAST. These methods concentrates on finding short identical matches.

Pairwise database searching

- FastA
  - algorithm by Lipman and Pearson (1985)
  - identifies short words (k-tuples) common to both sequences
  - k-tuples for proteins: 1-2 residues
  - k-tuples for DNA: up to 6 bases
  - k-tuples lying close to each other on the same diagonal joined by heuristics → gapped alignments computed by dynamic programming

Output from FastA search

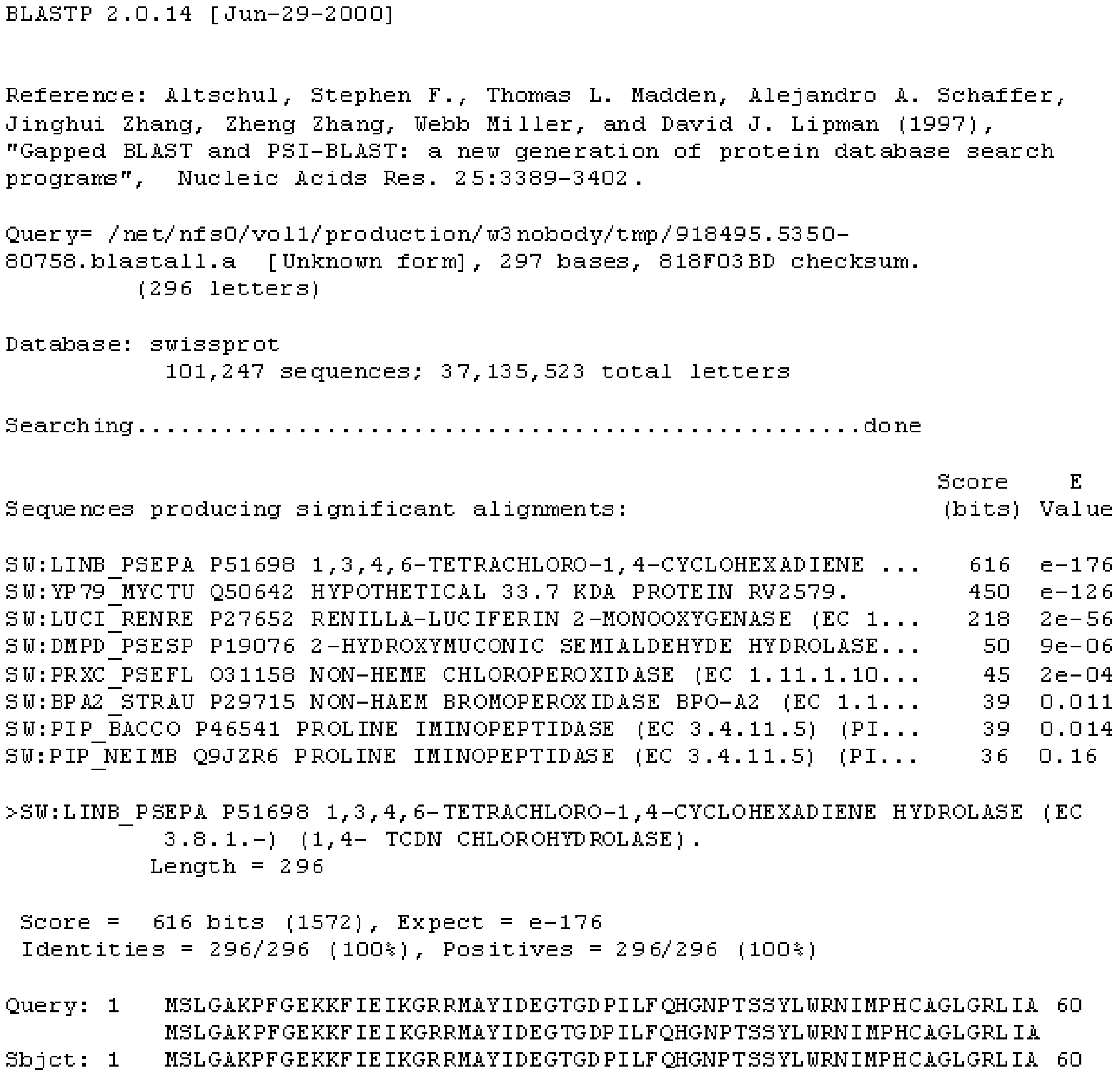


```
FASTA searches a protein or DNA sequence data bank
 version 3.3t09 May 18, 2001
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

@:1-: 296 aa
 EMBOSS_001
 vs  SWISS-PROT Protein Sequence Database library
searching /ebi/services/idata/fastadb/swissprot library

37135523 residues in 101247 sequences
 statistics extrapolated from 60000 to 101082 sequences
  Expectation_n fit: rho(ln(x))= 5.8158+/-0.000184; mu= 4.0375+/- 0.010
 mean_var=74.4386+/-14.720, 0's: 132 Z-trim: 20  B-trim: 0 in 0/65
 Lambda= 0.1487

FASTA (3.39 May 2001) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 36, opt: 24, gap-pen: -12/ -2, width:  16
 Scan time:  1.930
The best scores are:                                      opt bits E(101082)
SW:LINB_PSEPA P51698 1,3,4,6-TETRACHLORO-1,4-CYCL  ( 296) 2041  447 2.4e-125
SW:YP79_MYCTU Q50642 HYPOTHETICAL 33.7 KDA PROTEI  ( 300) 1494  330 5.1e-90
SW:LUCI_RENRE P27652 RENILLA-LUCIFERIN 2-MONOOXYG  ( 311)  744  169 1.4e-41
SW:DMPD_PSESP P19076 2-HYDROXYMUCONIC SEMIALDEHYD  ( 283)  169   46 0.00017
SW:PRXC_PSEFL O31158 NON-HEME CHLOROPEROXIDASE (E  ( 273)  168   45 0.00019
SW:PRXC_STRLI P49323 NON-HEME CHLOROPEROXIDASE (E  ( 275)  140   39   0.012
SW:PIP_BACCO P46541 PROLINE IMINOPEPTIDASE (EC 3.  ( 288)  140   39   0.013
SW:PRXC_SYNY3 Q55921 PUTATIVE NON-HEME CHLOROPERO  ( 276)  125   36    0.11
SW:PIP_NEIGO P42786 PROLINE IMINOPEPTIDASE (EC 3.  ( 310)  122   35     0.2

>>SW:LINB_PSEPA P51698 1,3,4,6-TETRACHLORO-1,4-CYCLOHEXA  (296 aa)
 initn: 2041 init1: 2041 opt: 2041  Z-score: 2372.6  bits: 447.0 E(): 2.4e-125
Smith-Waterman score: 2041;  100.000% identity (100.000% ungapped) in 296 aa overlap
(1-296:1-296)

               10        20        30        40        50        60
EMBOSS MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SW:LIN MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIA
               10        20        30        40        50        60
```

Pairwise database searching

- BLAST
  - Basic Local Alignment Search Tool
  - algorithm by Altschul *et al.* (1990)
  - identifies short ungapped sub-sequences (segment pairs) of the same length
  - sub-sequences are extended using dynamic programming to obtain local alignments - high scoring pairs (HSPs)
  - improved algorithm by Altschul *et al.* (1997) - produces gapped alignments
  - algorithm very fast - most commonly used for databases searching

Output from BLAST search



```
BLASTP 2.0.14 [Jun-29-2000]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= /net/nfs0/vol1/production/w3nobody/tmp/918495.5350-
80758.blastall.a  [Unknown form], 297 bases, 818F03BD checksum.
         (296 letters)

Database: swissprot
           101,247 sequences; 37,135,523 total letters

Searching..................................................done


                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

SW:LINB_PSEPA P51698 1,3,4,6-TETRACHLORO-1,4-CYCLOHEXADIENE ...   616  e-176
SW:YP79_MYCTU Q50642 HYPOTHETICAL 33.7 KDA PROTEIN RV2579.        450  e-126
SW:LUCI_RENRE P27652 RENILLA-LUCIFERIN 2-MONOOXYGENASE (EC 1...   218  2e-56
SW:DMPD_PSESP P19076 2-HYDROXYMUCONIC SEMIALDEHYDE HYDROLASE...    50  9e-06
SW:PRXC_PSEFL O31158 NON-HEME CHLOROPEROXIDASE (EC 1.11.1.10...    45  2e-04
SW:BPA2_STRAU P29715 NON-HAEM BROMOPEROXIDASE BPO-A2 (EC 1.1...    39  0.011
SW:PIP_BACCO P46541 PROLINE IMINOPEPTIDASE (EC 3.4.11.5) (PI...    39  0.014
SW:PIP_NEIMB Q9JZR6 PROLINE IMINOPEPTIDASE (EC 3.4.11.5) (PI...    36  0.16

>SW:LINB_PSEPA P51698 1,3,4,6-TETRACHLORO-1,4-CYCLOHEXADIENE HYDROLASE (EC
           3.8.1.-) (1,4- TCDN CHLOROHYDROLASE).
          Length = 296

 Score =  616 bits (1572), Expect = e-176
 Identities = 296/296 (100%), Positives = 296/296 (100%)

Query: 1   MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIA 60
           MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIA
Sbjct: 1   MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIA 60
```