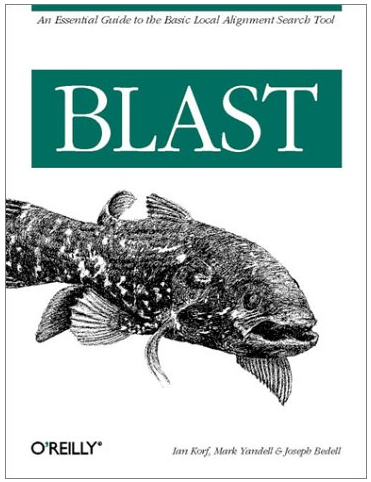
V2 Pairwise sequence alignment

- methods of sequence alignment: dynamic programming vs BLAST
- substitution matrices (PAM / BLOSUM)
- significance of alignments
- BLAST, algorithmus – parameters – output http://www.ncbi.nih.gov



Similarity of amino acids

Margaret Dayhoff: similarity of amino acids = observed frequency of exchanges of two amino acids in one position of homologous sequences

Margaret Dayhoff
http://www.nlm.nih.gov/changingthefaceofmedicine/gallery/photo_76_7.html

Express similarity as $\log_2$ odds ratio, also called *lod score*.

Lod score : compute logarithm with respect to the basis 2 ($\log_2$) of the ratio of the observed frequency of exchanges q_{jj} by the frequency of exchanges expected by chance (which is the product of their frequences p_i and p_j).

$$s_{ij} = \log \frac{q_{ij}}{p_i p_j}$$

Lod score = 0 → observed and expected frequencies are equal

> 0 → exchange is observed more often than expected (AA pair similar)

< 0 → unlikely exchange

Similarity of amino acids

Example: let the relative frequency of Methionine and Leucine be 0.01 and 0.1.

Randomly, we expect 1/1000 pairs of exchanges Met – Leu.

If the observed frequency of exhanges is 1/500, then the ratio is 2/1.

By taking the logarithm wrt. basis 2, this yields a *lod score* of +1 or 1 bit.

Usually one computes *nats* (natural logarithm),
multiplies the values by a scaling factor
and rounds them to integer values.

→ **substition matrices** PAM and BLOSUM.

PAM250 Matrix

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

Computing the raw score of an alignment

If 2 sequences differ in 2 (or more) positions, one likes to compute the likelihood that mutation A takes place in position 1 AND mutation B in position etc.

One needs to consider log (A × B), where × stands for AND.

Use general relationship log (A × B) = log A + log B

→ The score of an alignment is simply the **sum** of the matrix entries for the pairs of amino acids (or nucleotides) of an alignment:

Sequence 1: `TCCPSIVARSN`

Sequence 2: `SCCPSISARNT`

`1 12 12 6 2 5 -1 2 6 1 0` → Alignment score = 46

Dayhoff Matrix (1)

– compiled by Margaret.O. Dayhoff from the statistical frequency of amino acid substitutions in pairwise sequence alignments

– her data set only included pairs of closely related protein sequences (> 85% identity). These can be aligned unambigously.

– she converted substitution frequencies into the 20 x 20 matrix for the likelihood of mutations to occur.

– This matrix is named PAM 1.

An evolutionary distance of 1 PAM (point accepted mutation) means that there is 1 point mutation per 100 residues.
In other words, the two sequences are 99% identical.

Dayhoff Matrix (2)

Using PAM1 one can generate matrices for larger evolutionary distances by multiplying the matrix several times with itself.

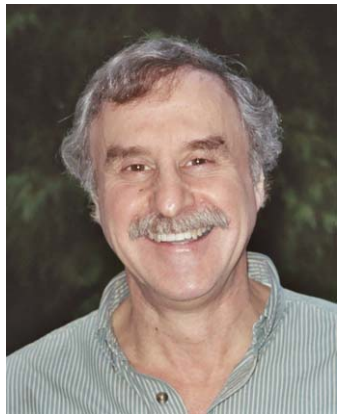
PAM250:

- 2,5 mutations per each residue
- Corresponds to 20% identical positions in two sequences, or 80% substitutions.
- PAM250 is the default matrix in many alignment packages.

BLOSUM Matrix

Limitation of the Dayhoff matrix:
Substitution rates were derived from alignments of closely related sequences where not so much evolution “has taken place”.

S. Henikoff and J.G. Henikoff used later multiple sequence Alignments of more remotely related sequences
➔ **Blosum-Matrix** from aligned “blocks” (no gaps)



Steven Henikoff

This was feasible once enough sequences and algorithms for multiple sequence alignment became available.

Advantages:
- larger data set (there are more remotely related sequences than closely related sequences)
- multiple alignments are more robust than pairwise alignments

BLOSUM Matrix (2)

The BLOSUM matrices (BLOcks SUbstitution Matrix) are based on the BLOCKS database.

The BLOCKS database contains blocks (gap-free amino acid signatures) that are characteristic for one protein family.

Different BLOSUM matrices are obtained by varying the lower threshold of required identity.

e.g. the BLOSUM80 matrix is derived from blocks with > 80% identity

Which matrix should be used?

Sequence closely related (low PAM, high Blosum)

Sequences remotely related (high PAM, low Blosum)

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250
Less divergent ←		→ *More divergent*

Good default values: PAM250, BLOSUM62

Scoring of gaps

Also insertions and deletions need to be scored …

Distinguish the mechanism of **gap opening**:

```
aaagaaa
aaa-aaa
```

from the **extension** of an existing gap:

```
aaaggggaaa
aaa----aaa
```

Idea: gaps are often found in loops of protein structures.

Rather extend an existing gap than opening many small gaps.

Different alignment packages suggest reasonable default values.

Needleman-Wunsch algorithm

- General algorithm for globally aligning two sequences
- Gives alignment with guaranteed best score

- NW uses a matrix representation

All possible pairings of residues (nucleotides or amino acids) are represented in the 2-dimensional lattice.

The possible alignments correspond to different paths through the lattice.

- algorithm has 3 steps:
    1 Initialisation
    2 fill out
    3 trace back

Needleman-Wunsch algorithm: initialize

task: align the two words “COELACANTH” and “PELICAN” of length

m = 10 and n = 7. Construct $(m + 1) \times (n + 1)$ matrix.

Assign to elements of first row and column the values $-m \times gap$ und $-n \times gap$.

Pointers in these fields point at the origin.

		C	O	E	L	A	C	A	N	T	H
	0	-1 ←	-2 ←	-3 ←	-4 ←	-5 ←	-6 ←	-7 ←	-8 ←	-9 ←	-10 ←
P	↑ -1										
E	↑ -2										
L	↑ -3										
I	↑ -4										
C	↑ -5										
A	↑ -6										
N	↑ -7										

Needleman-Wunsch Algorithm: fill out

Fill out all matrix fields with entries and pointers due to simple operations that consider the values of the diagonal, vertical, and horizontal neighbor cells.

match score: value in diagonal cell top left + score for current position (+1 oder -1)

horizontal gap score: value of left cell + gap score (-1)

vertical gap score: value of top cell + gap score (-1).

Assign maximum of the 3 options to the current cell. Pointer points to the maximum.

		C	O	E	L	A	C	A	N	T	H
	0	-1 ←	-2 ←	-3 ←	-4 ←	-5 ←	-6 ←	-7 ←	-8 ←	-9 ←	-10 ←
P	↑ -1	↖ -1	↖ -2								

max(-1, -2, -2) = -1

max(-2, -2, -3) = -2

When two options yield the maximum, orient pointer along diagonal.

Needleman-Wunsch Algorithm: trace back

Trace-back yields alignment.

Start at right bottom corner and follow pointers to the top left corner.

```
COELACANTH
-PELICAN--
```

		C	O	E	L	A	C	A	N	T	H
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
P	-1	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
E	-2	-2	-2	-1	-2	-3	-4	-5	-6	-7	-8
L	-3	-3	-3	-2	0	-1	-2	-3	-4	-5	-6
I	-4	-4	-4	-3	-1	-1	-2	-3	-4	-5	-6
C	-5	-3	-4	-4	-2	-2	0	-1	-2	-3	-4
A	-6	-4	-4	-5	-3	-1	-1	1	0	-1	-2
N	-7	-5	-5	-5	-4	-2	-2	0	2	1	0

BLAST –
Basic Local Alignment Search Tool

• finds best score local alignment of an input sequence against all sequences of a database.

• BLAST algorithm is very fast, much faster than dynamic programming.

• BLAST can be used to search very large databases because it uses a pre-indexed database.

• BLAST is sufficiently sensitive for most purposes

• BLAST Is robust – Default parameters typically work

BLAST Algorithm, step 1

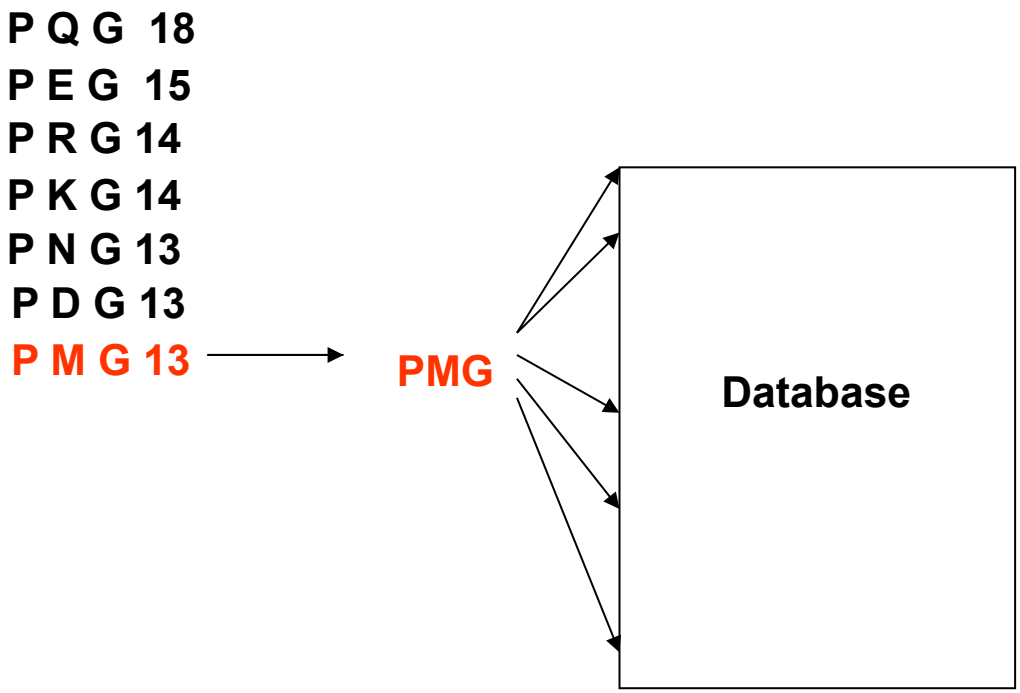
- Given a word of length *w* (*w* = 3 for proteins) and a given substitution matrix
- Generate a list of all words (*w*-mers) that yield a score > *T* when aligned with the input *w*-mer

test sequence **L N K C K T P Q G Q R L V N Q**

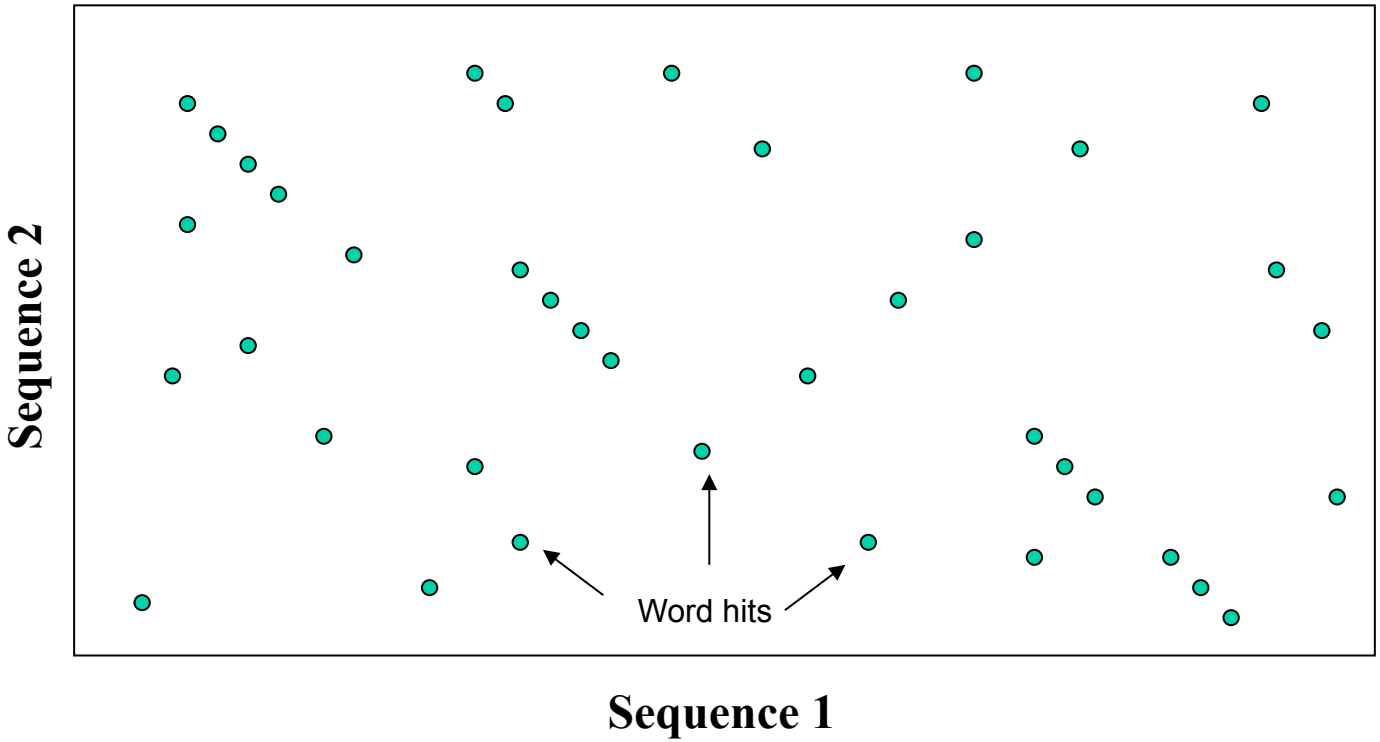
P Q G 18	**Word**	
P E G 15	**related words**	
P R G 14		
P K G 14		
P N G 13		
P D G 13		
P M G 13		
P Q A 12	below threshold (T=13)	
P Q N 12		
etc.		

BLAST Algorithmus, step 2

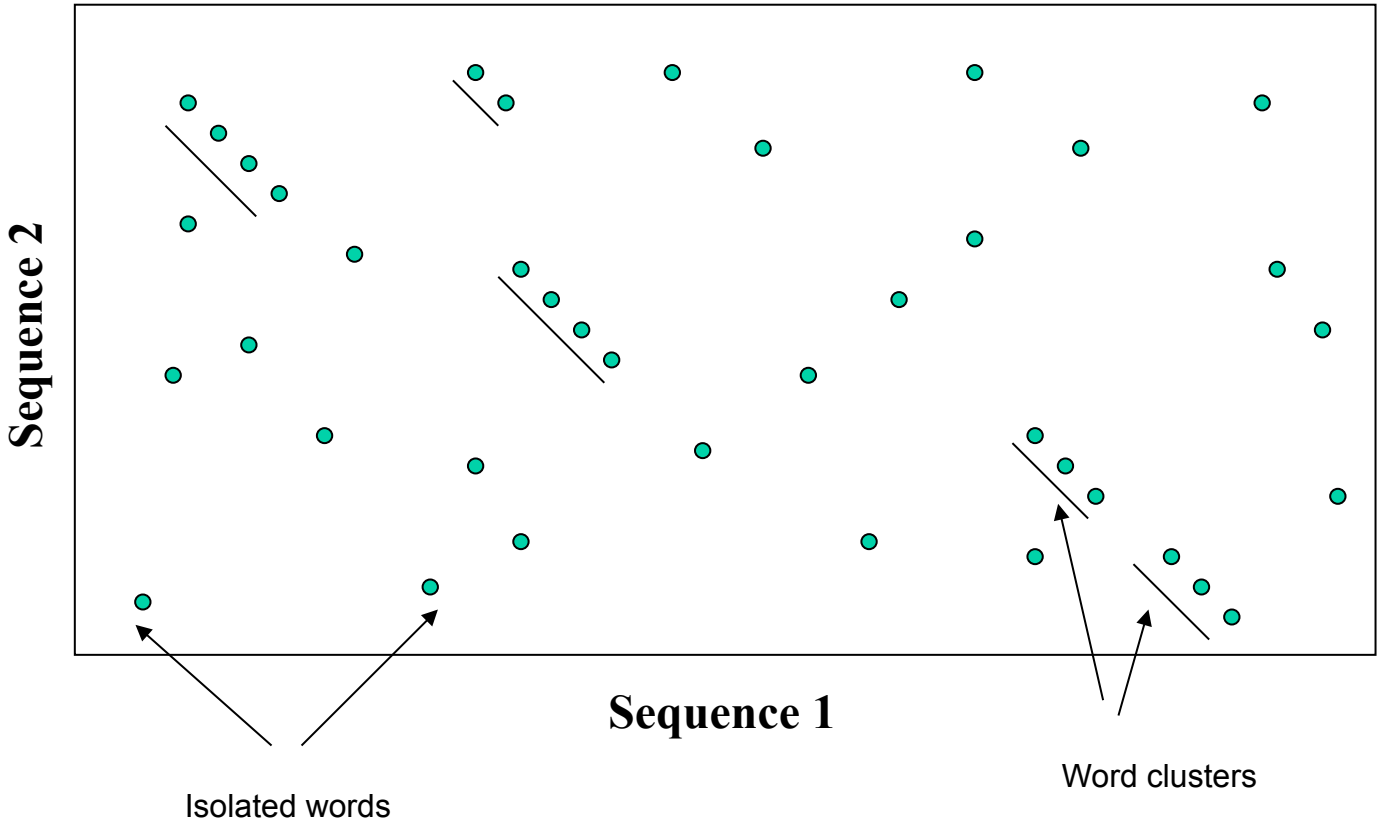
each related word points at positions in the database (hit list).



Seeding



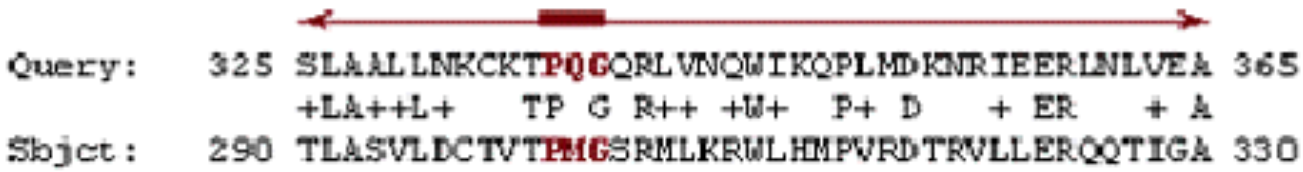
Seeding



BLAST algorithmus: extension

- After finding seeds, BLAST tries to extend the seed in both directions by adding further positions as long as the added score is favorable (exeeds some threshold)

- When maximal extension is reached, trim the alignment to the piece with the best score



High-scoring Segment Pair (HSP)

Related 3-letter words

BLOSUM62		PAM200	
Word	score	Word	score
RGD	17	RGD	18
KGD	14	RGE	17
QGD	13	RGN	16
RGE	13	KGD	15
EGD	12	RGQ	15
HGD	12	KGE	14
NGD	12	HGD	13
RGN	12	KGN	13
AGD	11	RAD	13
MGD	11	RGA	13
RAD	11	RGG	13
RGQ	11	RGH	13
RGS	11	RGK	13
RND	11	RGS	13
RSD	11	RGT	13
SGD	11	RSD	13
TGD	11	WGD	13

Comment:
The choice of the Substitution matrix and the choice of the cut-off will affect the seeding step.

PSI-BLAST

“Position-Specific Iterated BLAST”

- idea: amino acid substitution frequency should depend on the local “environment” of the residue in the protein structure
- PSI-BLAST package starts from a BLAST search with gaps
- PSI-BLAST collects all significant alignments to set up different substitution matrices for each position.
- Use the new substitution matrices in subsequent database searches
- PSI-BLAST can be used iteratively until no new alignments are found

BLAST output (1)

Please wait ...

```
BLASTP 2.2.2 [Dec-14-2001]


Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
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= blast.seq  [Unknown form], 261 bases, 32E76D0B checksum.
         (261 letters)

Database: swissprot
           101,602 sequences; 37,315,215 total letters

Searching..................................................done
```

BLAST output (2)

Small E-value shows that the hit is likely not a random hit

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

swissprot:CTRB_HUMAN Chymotrypsinogen B precursor (EC 3.4.21.1).    433  e-121
swissprot:CTR2_CANFA Chymotrypsinogen 2 precursor (EC 3.4.21.1).    386  e-107
swissprot:CTRB_RAT Chymotrypsinogen B precursor (EC 3.4.21.1).      383  e-106
swissprot:CTRB_BOVIN Chymotrypsinogen B (EC 3.4.21.1).              348  4e-96
swissprot:CTRA_BOVIN Chymotrypsinogen A (EC 3.4.21.1).              330  1e-90
swissprot:CTRA_GADMO Chymotrypsin A precursor (EC 3.4.21.1).        286  2e-77
```

BLAST output (3)

swissprot:CO2_HUMAN	Complement C2 precursor (EC 3.4.21.43) (C3/C...	55	1e-07
swissprot:CO2_MOUSE	Complement C2 precursor (EC 3.4.21.43) (C3/C...	53	3e-07
swissprot:ACH2_LONAC	Achelase II protease (EC 3.4.21.-).	52	1e-06
swissprot:GD_DROME	Serine protease gd precursor (EC 3.4.21.-) (G...	46	4e-05
swissprot:ACRO_CAPHI	Acrosin (EC 3.4.21.10) (Fragment).	39	0.009
swissprot:CTRP_PENMO	Chymotrypsin (EC 3.4.21.1) (Fragment).	36	0.047
swissprot:VSPA_CERCE	Cerastotin (EC 3.4.21.-) (Fragments).	35	0.098
swissprot:EL2B_HORSE	Neutrophil elastase 2B (EC 3.4.21.-) (Prote...	35	0.13
swissprot:CERC_SCHMA	Cercarial protease precursor (EC 3.4.21.-) ...	34	0.26
swissprot:EL2A_HORSE	Neutrophil elastase 2A (EC 3.4.21.-) (Prote...	33	0.42
swissprot:HPT_RABIT	Haptoglobin beta chain (Fragment).	31	1.4
swissprot:NMT1_ASPPA	NMT1 protein homolog.	30	4.8

Low scores with high E-values show that these are no good hits.

Karlin-Altschul statistics: E-value

Karlin und Altschul derived a formula to score the significance of an alignment:

$$E = kmne^{-\lambda S}$$

When searching against database with *n* sequences
using an input sequence with *m* words,

one expects *E* alignments occurring by chance for a given score.

λS : scaled substitution matrix
k: constant

Some rules of thumb

E-value (expectation value)

$E \leq 0{,}0001$	accurate hit
$0{,}0001 \leq E \leq 0{,}02$	sequences likely homologous
$0{,}02 \leq E \leq 1$	homology is possible, but not certain
$E \geq 1$	match is likely due to chance

Various BLAST packages

Program	Database	Query	Typical uses
BLASTN	Nucleotide	Nucleotide	Mapping oligonucleotides, cDNAs and PCR products to a genome, screening repetitive elements; cross-species sequence exploration; annotating genomic DNA sequencing reads
BLASTP	Protein	Protein	Identifying common regions between proteins; collecting related proteins for phylogenetic analyses
BLASTX	Protein	Nucleotide translated into protein	Finding protein-coding genes in genomic DNA; determining if a cDNA corresponds to a known protein
TBLASTN	Nucleotide translated	Protein	Identifying transcripts, potentially from multiple organisms, similar to a given protein; mapping a protein to genomic DNA into protein
TBLAST	Nucleotide translated into protein	Nucleotide translated into protein	Cross-species gene prediction at the genome or transcript level; searching for genes missed by traditional methods or not yet in protein databases

BLAST output (4)

```
>swissprot:CTRB_HUMAN Chymotrypsinogen B precursor (EC 3.4.21.1).
          Length = 263

 Score =  433 bits (1222), Expect = e-121
 Identities = 220/263 (83%), Positives = 252/263 (95%), Gaps = 2/263 (0%)

Query: 1   MAFIWLLSCYALLGTTFGCGVNAIHPVLTGLSKIVNGEEAVPGTWPWQVTLQDRSGFHFC 60
           MAF+WLLSC+ALLGTTFGCGV AIHPVL+GLS+IVNGE+AVPG+WPWQV+LQD++GFHFC
Sbjct: 1   MAFLWLLSCWALLGTTFGCGVPAIHPVLSGLSRIVNGEDAVPGSWPWQVSLQDKTGFHFC 60

Query: 61  GGSLISEDWVVTAAHCGVRTSEILIAGEFDQGSDEDNIQVLRIAKVFKQPKYSILTVNND 120
           GGSLISEDWVVTAAHCGVRTS++++AGEFDQGSDE+NIQVL+IAKVFK PK+SILTVNND
Sbjct: 61  GGSLISEDWVVTAAHCGVRTSDVVVAGEFDQGSDEENIQVLKIAKVFKNPKFSILTVNND 120

Query: 121 ITLLKLASPARYSQTISAVCLPSVDDD--AGSLCATTGWGRTKYNANKSPDKLERAALPL 178
           ITLLKLA+PAR+SQT+SAVCLPS DDD  AG+LCATTGWG+TKYNANK+PDKL++AALPL
Sbjct: 121 ITLLKLATPARFSQTVSAVCLPSADDDFPAGTLCATTGWGKTKYNANKTPDKLQQAALPL 180

Query: 179 LTNAECKRSWGRRLTDVMICGAASGVSSCMGDSGGPLVCQKDGAYTLVAIVSWASDTCSA 238
           L+NAECK+SWGRR+TDVMIC  ASGVSSCMGDSGGPLVCQKDGA+TLV IVSW SDTCS
Sbjct: 181 LSNAECKKSWGRRITDVMICAGASGVSSCMGDSGGPLVCQKDGAWTLVGIVSWGSDTCST 240

Query: 239 SSGGVYAKVTKIIPWVQKILSSN 261
           SS GVYA+VTK+IPWVQKIL++N
Sbjct: 241 SSPGVYARVTKLIPWVQKILAAN 263
```

BLAST output (5)

```
>swissprot:VSP5_TRIMU Mucrofibrase 5 precursor (EC 3.4.21.-).
          Length = 257

 Score =  103 bits (280), Expect = 3e-22
 Identities = 74/232 (31%), Positives = 110/232 (46%), Gaps = 10/232 (4%)

Query: 34   IVNGEEAVPGTWPWQVTLQDRSGFHFCGGSLISEDWVVTAAHCGVRTSEILIAGEFDQGS 93
            I+ G+E      P+ V +      + CGG+LI+E+WV+TAAHC     EI +     +
Sbjct: 25   IIGGDECNINEHPFLVLVYYDD--YQCGGTLINEEWVLTAAHCNGENMEIYLGMHSKKVP 82

Query: 94   DEDNIQVLRIAKVFKQPKYSILTVNNDITLLKLASPARYSQTISAVCLPSVDDDAGSLCA 153
            ++D  + +   K F     +    N DI L++L  P R S  I+ + LPS     GS+C
Sbjct: 83   NKDRRRRVPKEKFFCDSSKNYTKWNKDIMLIRLNRPVRKSAHIAPLSLPSSPPSVGSVCR 142

Query: 154  TTGWGRTKYNANKSPDKLERAALPLLTNAECKRSW-GRRLTDVMICGA--ASGVSSCMGD 210
              GWG          PD    A + LL    C+ ++ G   T   +C      G  SC GD
Sbjct: 143  IMGWGTISPTKVTLPDVPRCANINLLDYEVCRAAYAGLPATSRTLCAGILEGGKDSCGGD 202

Query: 211  SGGPLVCQKDGAYTLVAIVSWASDTCS-ASSGGVYAKVTKIIPWVQKILSSN 261
            SGGPL+C  +G +    IVSW  D C+     G+Y  V   + W++ I++ N
Sbjct: 203  SGGPLIC--NGQFQ--GIVSWGGDPCAQPHEPGLYTNVFDHLDWIKGIIAGN 250
```

BLAST output (6)

```
>swissprot:HPT_RABIT Haptoglobin beta chain (Fragment).
          Length = 40

 Score = 31.3 bits (74), Expect = 1.4
 Identities = 15/41 (36%), Positives = 22/41 (53%), Gaps = 1/41 (2%)

Query: 34 IVNGEEAVPGTWPWQVTLQDRSGFHFCGGSLISEDWVVTAA 74
          I+ G     G++PWQ  +  R      G +LISE W++T A
Sbjct: 1  IIGGSLDAKGSFPWQAKMVSRHNL-VTGATLISEQWLLTTA 40
```

```
>swissprot:NMT1_ASPPA NMT1 protein homolog.
          Length = 342

 Score = 29.6 bits (69), Expect = 4.8
 Identities = 11/34 (32%), Positives = 22/34 (64%)

Query: 72  TAAHCGVRTSEILIAGEFDQGSDEDNIQVLRIAK 105
           TA  CG+  ++ +I G+ D G   +N+Q++ +A+
Sbjct: 137 TAVRCGMNVTKAIIRGDIDAGIGLENVQMVELAE 170
```

Although high portion of identical and positive positions, both matches have unfavorable E-values because fragments are very short.

Summary

Pairwise sequence alignment is routine nowadays, but not trivial.

Dynamic programming (e.g. Smith-Waterman or Needleman-Wunsch) guarantees to find the optimal alignment.
(Note that the scoring function is only a model of biological evolution).

Much faster alignments are produced by BLAST and related programs (BLAT).

BLAST gives robust and useful results for protein sequences.

Multiple sequence alignments can detect more remote relationships and provide a better functional understanding of sequences