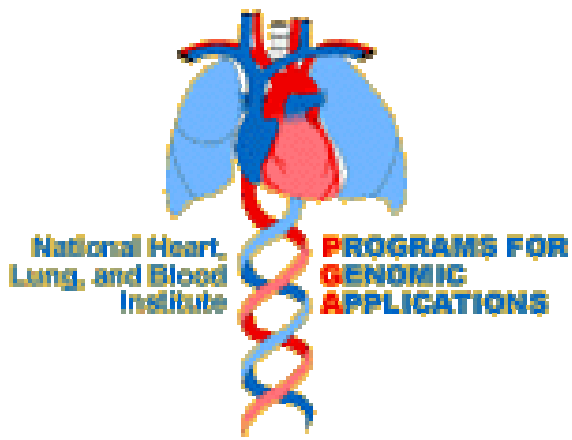
BLAST:

Basic Local Alignment Search Tool

Jonathan M. Urbach
Bioinformatics Group
Department of Molecular Biology



Topics to be covered:

- BLAST as a Sequence Alignment Tool
- Uses of BLAST
- Types of BLAST
- How BLAST works
    - Scanning for 'hits'
    - Scoring with Substitution Matrices
- Common Databases for Use with BLAST available at NCBI
- Interpretation of Blast Results
- Blast options: on the net or on **your** computer
- Learning More About BLAST,
- A BLAST demo

gi|13325078|gb|AAG33875.2| (AF232004) HrpL [Pseudomonas syringae pv. tomato]
Length = 184

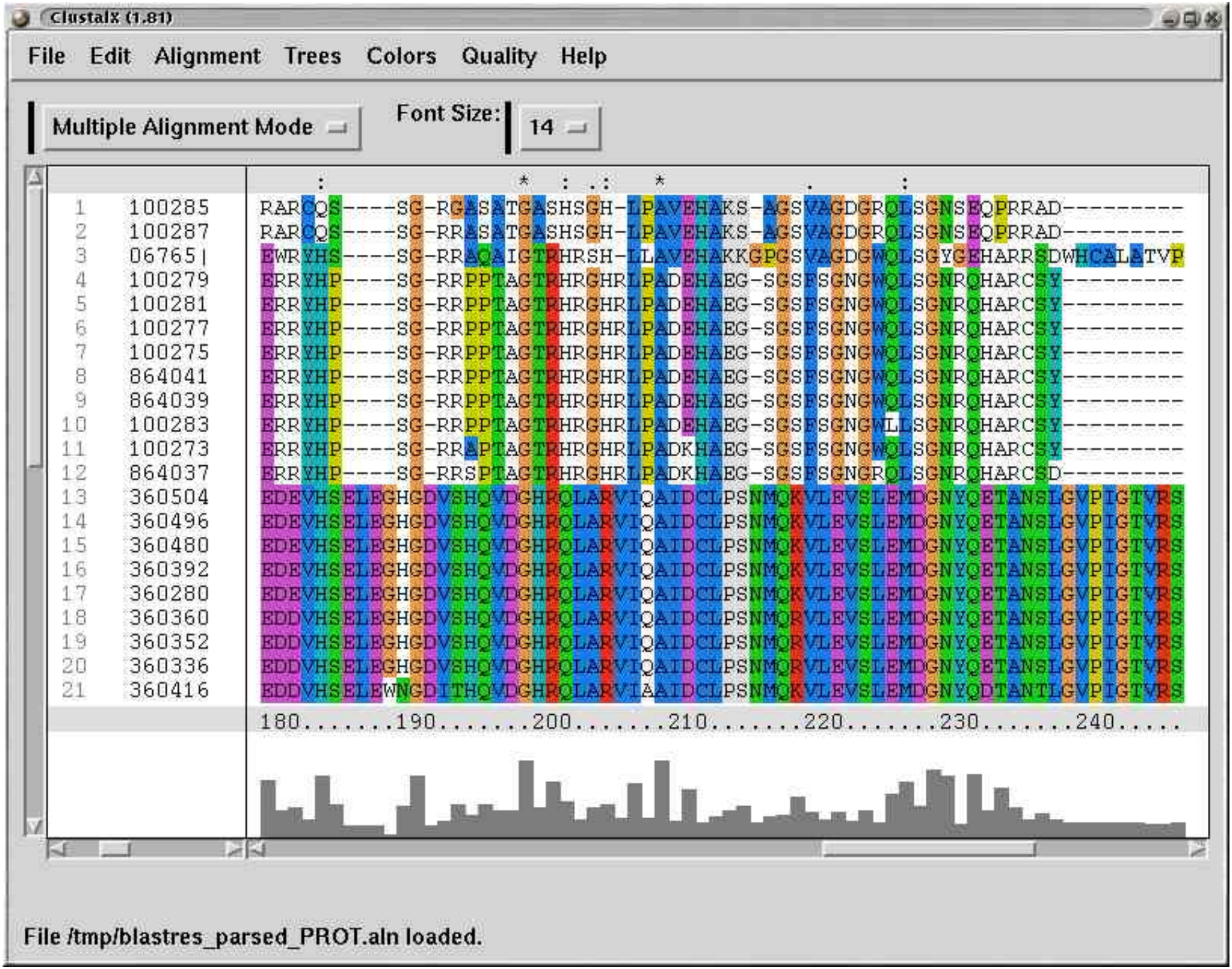
Score = 347 bits (889), Expect = 4e-95
Identities = 182/184 (98%), Positives = 183/184 (98%)

```
Query: 1   MFQKIVILDSTQPRQPSSSAGIRQMTADQIQMLRAFIQKRVMNPDDVDDILQCVFLEALR 60
           MFQKIVILDSTQPRQPSSSAGIRQMTADQIQMLRAFIQKRVMNPDDVDDILQCVFLEALR
Sbjct: 1   MFQKIVILDSTQPRQPSSSAGIRQMTADQIQMLRAFIQKRVMNPDDVDDILQCVFLEALR 60

Query: 61  NEHKFQHASKPQTWLCGIALNLIRNHFRKMYRQPYQESWEDEVHSELEGHGDVSHQVDGH
           NEHKFQHASKPQTWLCGIALNLIRNHFRKMYRQPYQESWEDEVHSELEGHGDVSHQV+GH
Sbjct: 61  NEHKFQHASKPQTWLCGIALNLIRNHFRKMYRQPYQESWEDEVHSELEGHGDVSHQVEGH 12

Query: 121 RQLARVIQAIDCLPSNMQKVLEVSLEMDGNYQETANSLGVPIGTVRSRLSRARVQLKQQI 180
           RQLARVIQAIDCLPSNMQKVLEVSLEMDGNYQETANSLGVPIGTVRSRLS ARVQLKQQI
Sbjct: 121 RQLARVIQAIDCLPSNMQKVLEVSLEMDGNYQETANSLGVPIGTVRSRLSGARVQLKQQI 180

Query: 181 DPFA 184
           DPFA
Sbjct: 181 DPFA 184
```


ClustalX (1.81)
File Edit Alignment Trees Colors Quality Help
Multiple Alignment Mode
Font Size:
14
1 100285 RARCQS----SG-RGASATGASHSGH-LPAVEHAKS-AGSVAGDGRQLSGNSEQPRRAD---------
2 100287 RARCQS----SG-RRASATGASHSGH-LPAVEHAKS-AGSVAGDGRQLSGNSEQPRRAD---------
3 06765| EWRVHS----SG-RRAQAIGTRHRSH-LLAVEHAKKGPGSVAGDGWQLSGYGEHARRSDWHCALATVP
4 100279 ERRVHP----SG-RRPPTAGTRHRGHRLPADEHAEG-SGSFSGNGWQLSGNRQHARCSY---------
5 100281 ERRVHP----SG-RRPPTAGTRHRGHRLPADEHAEG-SGSFSGNGWQLSGNRQHARCSY---------
6 100277 ERRVHP----SG-RRPPTAGTRHRGHRLPADEHAEG-SGSFSGNGWQLSGNRQHARCSY---------
7 100275 ERRVHP----SG-RRPPTAGTRHRGHRLPADEHAEG-SGSFSGNGWQLSGNRQHARCSY---------
8 864041 ERRVHP----SG-RRPPTAGTRHRGHRLPADEHAEG-SGSFSGNGWQLSGNRQHARCSY---------
9 864039 ERRVHP----SG-RRPPTAGTRHRGHRLPADEHAEG-SGSFSGNGWQLSGNRQHARCSY---------
10 100283 ERRVHP----SG-RRPPTAGTRHRGHRLPADEHAEG-SGSFSGNGWLLSGNRQHARCSY---------
11 100273 ERRVHP----SG-RRAPTAGTRHRGHRLPADKHAEG-SGSFSGNGWQLSGNRQHARCSY---------
12 864037 ERRVHP----SG-RRSPTAGTRHRGHRLPADKHAEG-SGSFSGNGRQLSGNRQHARCSD---------
13 360504 EDEVHSELEGHGDVSHQVDGHRQLAPVIQAIDCLPSNMQKVLEVSLEMDGNYQETANSLGVPIGTVRS
14 360496 EDEVHSELEGHGDVSHQVDGHRQLAPVIQAIDCLPSNMQKVLEVSLEMDGNYQETANSLGVPIGTVRS
15 360480 EDEVHSELEGHGDVSHQVDGHRQLAPVIQAIDCLPSNMQKVLEVSLEMDGNYQETANSLGVPIGTVRS
16 360392 EDEVHSELEGHGDVSHQVDGHRQLAPVIQAIDCLPSNMQKVLEVSLEMDGNYQETANSLGVPIGTVRS
17 360280 EDEVHSELEGHGDVSHQVDGHRQLAPVIQAIDCLPSNMQKVLEVSLEMDGNYQETANSLGVPIGTVRS
18 360360 EDDVHSELEGHGDVSHQVDGHRQLAPVIQAIDCLPSNMQRVLEVSLEMDGNYQETANSLGVPIGTVRS
19 360352 EDDVHSELEGHGDVSHQVDGHRQLAPVIQAIDCLPSNMQRVLEVSLEMDGNYQETANSLGVPIGTVRS
20 360336 EDDVHSELEGHGDVSHQVDGHRQLAPVIQAIDCLPSNMQRVLEVSLEMDGNYQETANSLGVPIGTVRS
21 360416 EDDVHSELEWNGDITHQVDGHRQLAPVIAAIDCLPSNMQKVLEVSLEMDGNYQDTANTLGVPIGTVRS
180.....190.....200.....210.....220.....230.....240.....
File /tmp/blastres_parsed_PROT.aln loaded.

Sequence Alignment Tools

Database Searching:

BLAST:
NCBI, Web Interface: http://www.ncbi.nlm.nih.gov/BLAST/
WuBLAST http://blast.wustl.edu
FASTA: http://www.ebi.ac.uk/fasta3/
Smith-Waterman
Par-Align: http://dna.uio.no/search/

Multiple Sequence Alignment:
CLUSTALW: http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html
DiAlign, Web Interface: http://genomatix.gsf.de/cgi-bin/dialign/dialign.pl
MSA:http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/msa.html
Web Interface: http://bioweb.pasteur.fr/seqanal/interfaces/msa-simple.html

Uses of BLAST:

Query a database for sequences similar to an input sequence.

Uses of BLAST:

Query a database for sequences similar to an input sequence.

- **Identify previously characterized sequences.**

Uses of BLAST:

Query a database for sequences similar to an input sequence.

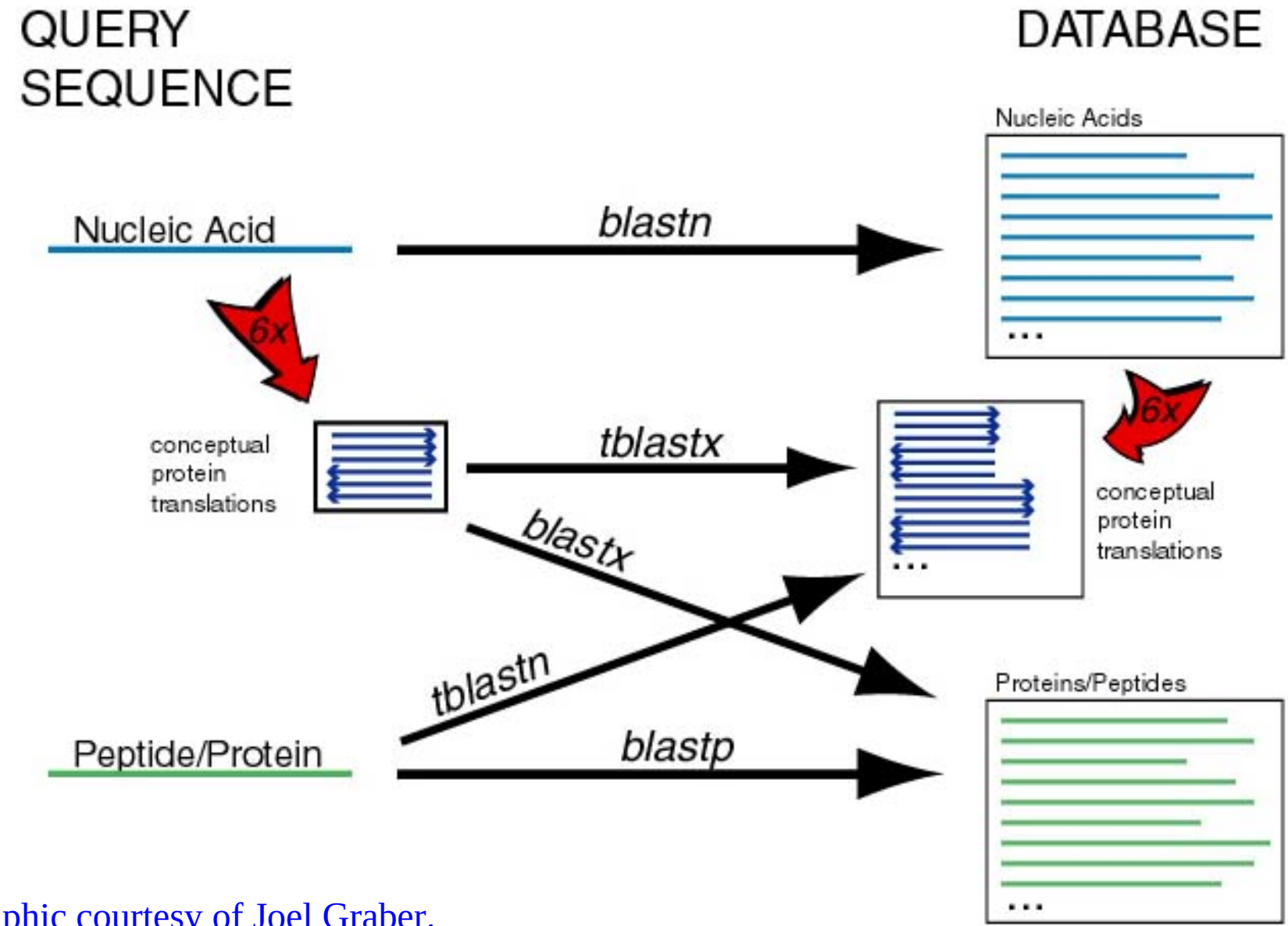
- **Identify previously characterized sequences.**
- **Find phylogenetically related sequences.**

Uses of BLAST:

Query a database for sequences similar to an input sequence.

- **Identify previously characterized sequences.**
- **Find phylogenetically related sequences.**
- **Identify possible functions based on similarities to known sequences.**

Types of BLAST:



Graphic courtesy of Joel Graber.

How BLAST Works

(1) BLAST scans database for 'words' of a predetermined length (a 'hit') with some minimum threshold parameter, *T*.

(2) BLAST then extends the hit until the score falls below the maximum score yet attained minus some value *X*.

Altschul, S. F. et al., *Nucleic Acids Research*, **25**, 3389-3402 (1997)

Query:

MVSHSAAQAYSMLTNSEFVSMWSAESCRTPLCSVNNSYFPGALGDEKTT

Query:

MVSHSAAQAYSMLTNSEFVSMWSAESCRTPLCSVNNSYFPGALGDEKTT

Use 2 or 3-letter words...

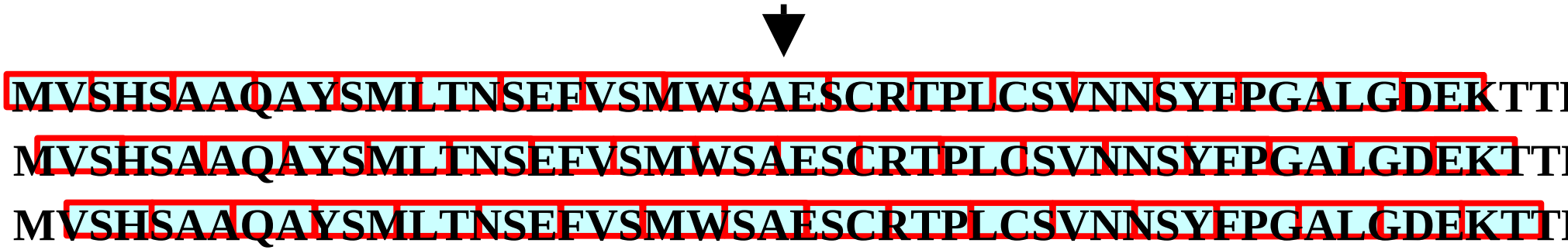
MVSHSAAQAYSMLTNSEFVSMWSAESCRTPLCSVNNSYFPGALGDEKTT

MVSHSAAQAYSMLTNSEFVSMWSAESCRTPLCSVNNSYFPGALGDEKTT

MVSHSAAQAYSMLTNSEFVSMWSAESCRTPLCSVNNSYFPGALGDEKTT

Query:

MVSHSAAQAYSMLTNSEFVSMWSAESCRTPLCSVNNSYFPGALGDEKTT

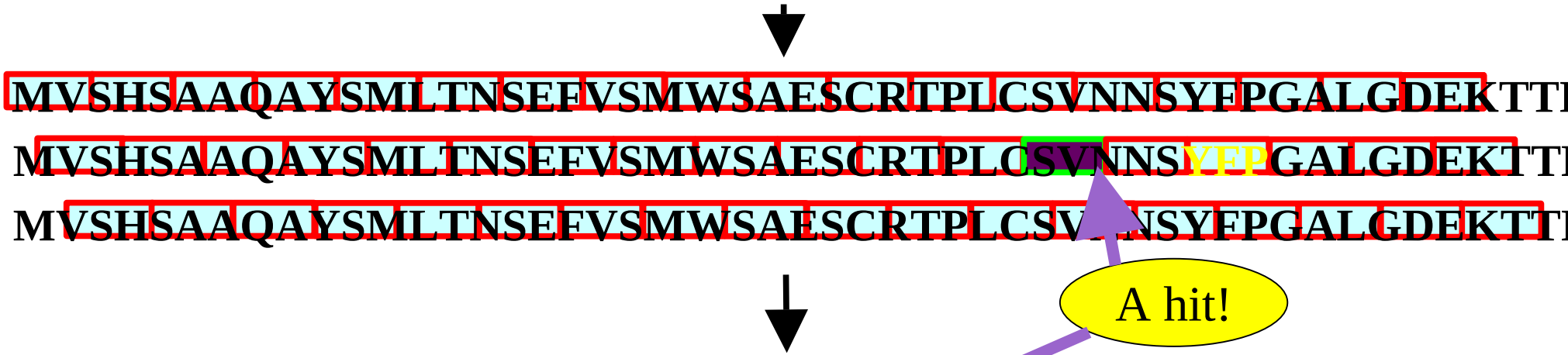


Scan against subject sequence:

>gi|507311|gb|AAA25685.1| aminoglycoside 6'-N-acetyltransferase
MTEHDLAMLYEWLNRSHIVEWWGGEEARPTLADVQEQYLPSVLAQESVTPYIAMLNGEPIG
SGDGWWEEETDPGVRGIDQSLANASQLGKGLGTKLVRALVELLFNDPEVTKIQTDPSPSNLR
GFERQGTVTTPDGPAVYMVQTRQAFERTRSDA

Query:

MVSHSAAQAYSMLTNSEFVSMWSAESCRTPLCSVNNSYFPGALGDEKTT

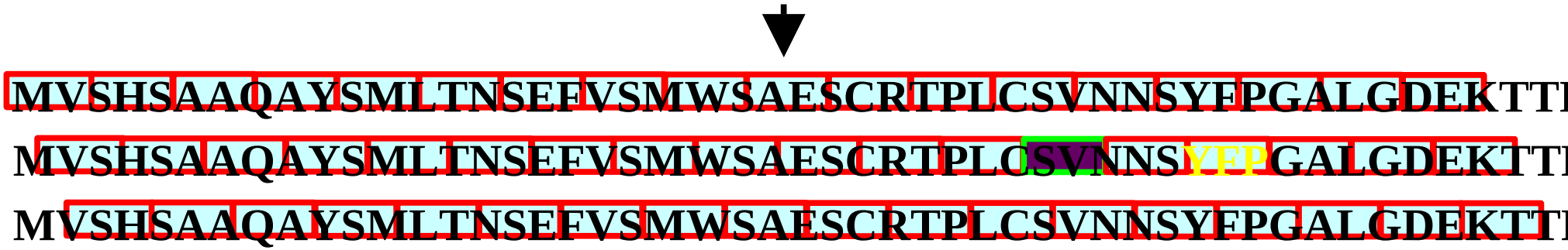


>gi|507311|gb|AAA25685.1| aminoglycoside 6'-N-acetyltransferase

MTEHDLAMLYEWLNRSHIVEWWGGEEARPTLADVQEQYLPSVLAQESVTPYIAMLNGEPIG
SGDGWWEEETDPGVRGIDQSLANASQLGKGLGTKLVRALVELLFNDPEVTKIQTDPSPSNLR
GFERQGTVTTPDGPAVYMVQTRQAFERTRSDA

Query:

MVSHSAAQAYSMLTNSEFVSMWSAESCRTPLCSVNNSYFPGALGDEKTT



>gi|507311|gb|AAA25685.1| aminoglycoside 6'-N-acetyltransferase

MTEHDLAMLYEWLNRSHIVEWWGGEEARPTLADVQEQYLPSVLAQESVTPYIAMLNGEPIG
SGDGWWEEETDPGVRGIDQSLANASQLGKGLGTKLVRALVELLFNDPEVTKIQTDPSPSNLR
GFERQGTVTTPDGPAVYMVQTRQAFERTRSDA

Extension:

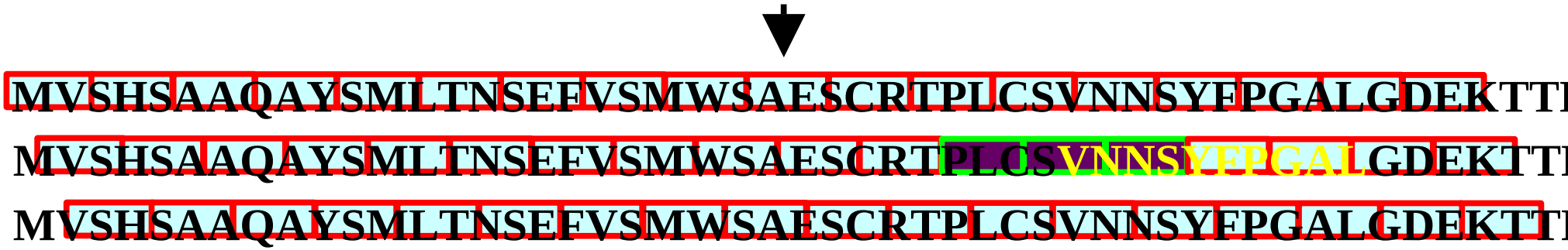
Query: YFP

Y P

Sbjct: YLP

Query:

MVSHSAAQAYSMLTNSEFVSMWSAESCRTPLCSVNNSYFPGALGDEKTT



>gi|507311|gb|AAA25685.1| aminoglycoside 6'-N-acetyltransferase

MTEHDLAMLYEWLNRSHIVEWWGGEEARPTLADVQEQYLPSVLAQESVTPYIAMLNGEPIG
SGDGWWEEETDPGVRGIDQSLANASQLGKGLGTKLVRALVELLFNDPEVTKIQTDPSPSNLR
GFERQGTVTTPDGPAVYMVQTRQAFERTRSDA

Extension:

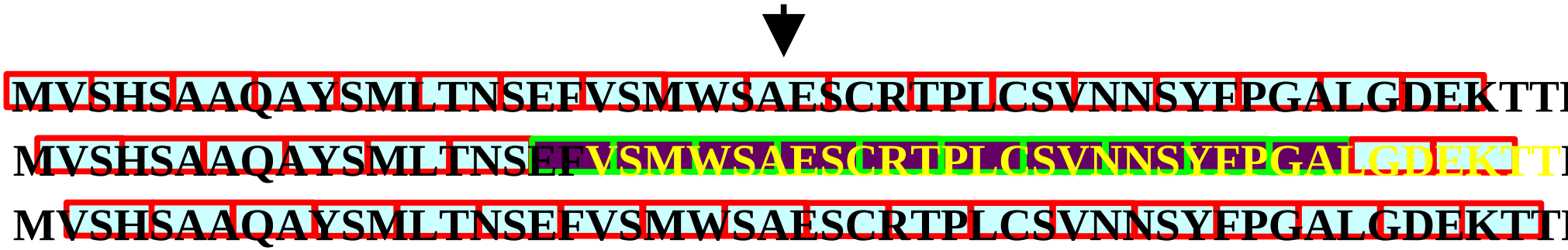
Query: **VNNSYFPGAL**

V Y P L

Sbjct: **VQEQYLPSVL**

Query:

MVSHSAAQAYSMLTNSEFVSMWSAESCRTPLCSVNNSYFPGALGDEKTT



>gi|507311|gb|AAA25685.1| aminoglycoside 6'-N-acetyltransferase

MTEHDLAMLYEWLNRSHIVEWWGGEEARPTLADVQEQYLPSVLAQESVTPYIAMLNGEPIG
SGDGWWEEETDPGVRGIDQSLANASQLGKGLGTKLVRALVELLFNDPEVTKIQTDPSPSNLR
GFERQGTVTTPDGPAVYMVQTRQAFERTRSDA

Extension:

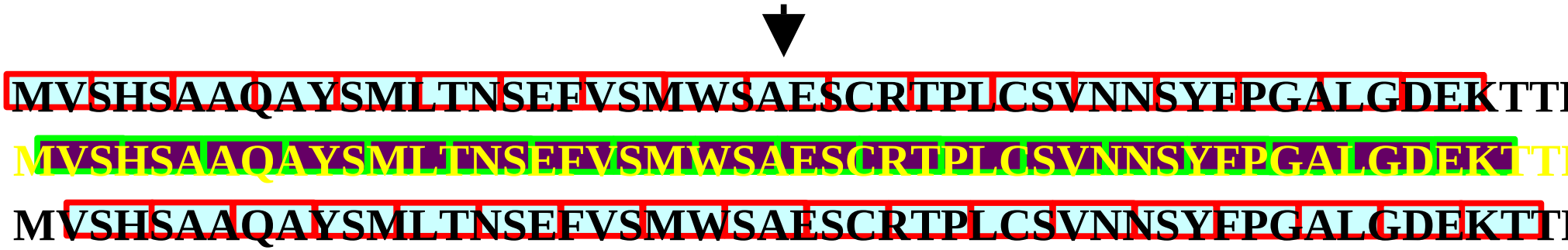
Query: VSMWSAESCRTPLCSVNNSYFPGALGDEKTT

V W E R L V Y P L E T

Sbjct: VEWWGGEEARPTLADVQEQYLPSVLAQESVT

Query:

MVSHSAAQAYSMLTNSEFVSMWSAESCRTPLCSVNNSYFPGALGDEKTT



>gi|507311|gb|AAA25685.1| aminoglycoside 6'-N-acetyltransferase

MTEHDLAMLYEWLNRSHIVEWWGGEEARPTLADVQEQYLPSVLAQESVTPYIAMLNGEPIG
SGDGWWEEETDPGVRGIDQSLANASQLGKGLGTKLVRALVELLFNDPEVTKIQTDPSPSNLR
GFERQGTVTTPDGPAVYMVQTRQAFERTRSDA

Extension:

Query: MVSHSAAQAYSMLTNSEFVSMWSAESCRTPLCSVNNSYFPGALGDEKTTKVI
M H A Y L S V W E R L V Y P L E T I L
Sbjct: MTEHDLAMLYEWLNRSHIVEWWGGEEARPTLADVQEQYLPSVLAQESVTPY

HSP: A High-Scoring Segment Pair

Towards BLAST Scoring

- **Expected negative score for alignment of two random residues.**
- **Maximal score for a perfect match.**
- **Combinations of residues that can commonly substitute for one another in proteins may have positive score.**

```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =   0.6979, Expected =  -0.5209
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1 -4
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
```

BLAST Scoring

- **Nominal HSP scores (*S*) are sums of scores from substitution matrices.**

- **Nominal scores are normalized to give 'bit scores' (*S'*):**

(I) $$S' = \frac{\Leftrightarrow S \mathrm{B} \ln K}{\ln 2}$$

K and λ are statistical parameters *that relate the calculated score to the probability finding a hit with at least that score.*

1. **Allows comparison of alignments scored by different methods**

Relating Scores to Probability: E-Values and P-Values

The expected number of HSPs with scores of at least S is *given by the following equations:*

(II) $$E = Kmn\,e^{-\lambda S}$$

(III) $$E = mn\,2^{-S'}$$

K and λ are statistical *'normalization' parameters.*
m and *n* are the lengths of the query *sequence and database.*
S is a nominal score.
S' is a bit score.

P-Values are the likelihood of finding a match with a score of at least S:

(IV) $$P = 1 - e^{-E}$$

Substitution Matrices

Scores in the substitution matrix are expressed in 'log-odds' format:

$$(V) \qquad s_{ij} = \ln\left(\frac{q_{ij}}{p_i p_j}\right)/\lambda$$

q_{ij} = target frequency
pi, pj = frequency those residues appear by chance
λ = normalization parameter

1. **The more frequently the substitution occurs, the higher the score.**
1. **The less frequently the residue occurs in the sequence as a whole, the higher the score.**

Substitution Matrices

- Derived from empirically observed substitution frequencies
- Higher scores for substitution with similar residues.
- Random substitutions give negative scores

Types of Substitution Matrices

Each tailored to a specific degree of evolutionary divergence.

PAM Matrices:

'Percent Accepted Mutation'

start with closely related sequences, and extrapolate substitution probabilities for more distantly related sequences.

1 PAM unit=1 mutation event per 100 bases.

e.g.: PAM 100 tailored for 100 mutation events per 100 bases.

Barker, W.C. & Dayhoff, M.O. *Atlas of Protein Sequence and Structure,* pp 101-110, National Biomedical Research Foundation (1972).

Types of Substitution Matrices

BLOSUM Matrices

'BLOck SUbstitution Matrix'

Values inferred from sequences sharing a maximum of the given value.

e.g.: BLOSUM62 derived from sequences no more than 62% identical.

Henikoff, S. & Henikoff, J.G., *Proc. Natl. Acad. Sci., USA,* ***89, 10915-10919 (1992).***

Comparing Substitution Matrices

Similar Evolutionary Distances

PAM 120<----> BLOSUM80

PAM160<----> BLOSUM62

PAM250 <----> BLOSUM45

BLOSUM more tolerant to hydrophobic than PAM but less tolerant to hydrophilic substitutions.

```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =   0.6979, Expected =  -0.5209
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1 -4
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2  0  0 -2 -1 -1 -1 -1 -1 -4
```

BLAST Databases at NCBI:

Database	Description	DNA	Protein
nr	All non-redundant GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or HTGS sequences).	4	4
month	All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.	4	4
dbest	Non-redundant database of GenBank+EMBL+DDBJ EST Divisions.	4	
dbsts	Non-redundant database of GenBank+EMBL+DDBJ STS Divisions.	4	
mouse ests	The non-redundant Database of GenBank+EMBL+DDBJ EST Divisions limited to the organism mouse.	4	
human ests	The Non-redundant Database of GenBank+EMBL+DDBJ EST Divisions limited to the organism human.	4	
other ests	The non-redundant database of GenBank+EMBL+DDBJ EST Divisions all organisms except mouse and human.	4	
yeast	Yeast (Saccharomyces cerevisiae) genomic nucleotide sequences. Not a collection of all Yeast nucelotides sequences, but the sequence fragments from the Yeast complete genome.	4	4
E. coli	E. coli (Escherichia coli) genomic nucleotide sequences.	4	4
pdb	Sequences derived from the 3-dimensional structure of proteins.	4	4
kabat [kabatnuc]	Kabat's database of sequences of immunological interest. For more information http://immuno.bme.nwu.edu/	4	4
patents	Nucleotide sequences derived from the Patent division of GenBank.	4	4
vector	Vector subset of GenBank(R), NCBI, (ftp://ncbi.nlm.nih.gov/pub/blast/db/ directory).	4	
swissprot	The last major release of the SWISS-PROT protein sequence database (no updates). These are uploaded to our system when they are received from EMBL.		4
alu	Translations of select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences. It is available at ftp://ncbi.nlm.nih.gov/pub/jmc/alu . See "Alu alert" by Claverie and Makalowski, Nature vol. 371, page 752 (1994) .		4

Adapted From:http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/query_tutorial.html

Interpreting Blast Results

>gi|6580755|gb|AAF18265.1|U22895_1 (U22895) alternative sigma factor AlgU [Azotobacter vinelandii]
Length = 193

Score = 334 bits (857), Expect = 2e-91
Identities = 180/192 (93%), Positives = 189/192 (97%)

Query: 1
MLTQEQDQQLVERVQRGDKRAFDLLVLKYQHKILGLIVRFVHDAQEAQDVAQEAFIKAYR 60
ML QEQDQQLVERVQRGD+RAFDLLVLKYQHKILGLIVRFVHDA EAQDVAQEAFIKAYR
Sbjct: 1 MLNQEQDQQLVERVQRGDRRAFDLLVLKYQHKILGLIVRFVHDAHEAQDVAQEAFIKAYR 60

Query: 61 ALGNFRGDSAFYTWLYRIAINTAKNHLVARGRRPPDSDVTAEDAEFFEGDHALKDIESPE 120
ALGNFRGDSAFYTWLYRIAINTAKNHLVARGRRPPDSDV+A DAEF+EGDHALKDIESPE
Sbjct: 61 ALGNFRGDSAFYTWLYRIAINTAKNHLVARGRRPPDSDVSAGDAEFYEGDHALKDIESPE 120

Query: 121 RAMLRDEIEATVHQTIQQLPEDLRTALTLREFEGLSYEDIATVMQCPVGTVRSRIFRARE 180
R++LRDEIEATVH+TIQQLPEDLRTALTLREF+GLSYEDIA+VMQCPVGTVRSRIFRARE
Sbjct: 121 RSLLRDEIEATVHRTIQQLPEDLRTALTLREFDGLSYEDIASVMQCPVGTVRSRIFRARE 180

Query: 181 AIDKALQPLLRE 192
AIDKALQPLL+E

Interpreting Blast Results

Hit name

>gi|6580755|gb|AAF18265.1|U22895_1 (U22895) alternative sigma factor AlgU [Azotobacter vinelandii]
Length = 193

Score = 334 bits (857), Expect = 2e-91
Identities = 180/192 (93%), Positives = 189/192 (97%)

Query: 1
MLTQEQDQQLVERVQRGDKRAFDLLVLKYQHKILGLIVRFVHDAQEAQDVAQEAFIKAYR 60
ML QEQDQQLVERVQRGD+RAFDLLVLKYQHKILGLIVRFVHDA EAQDVAQEAFIKAYR
Sbjct: 1 MLNQEQDQQLVERVQRGDRRAFDLLVLKYQHKILGLIVRFVHDAHEAQDVAQEAFIKAYR 60

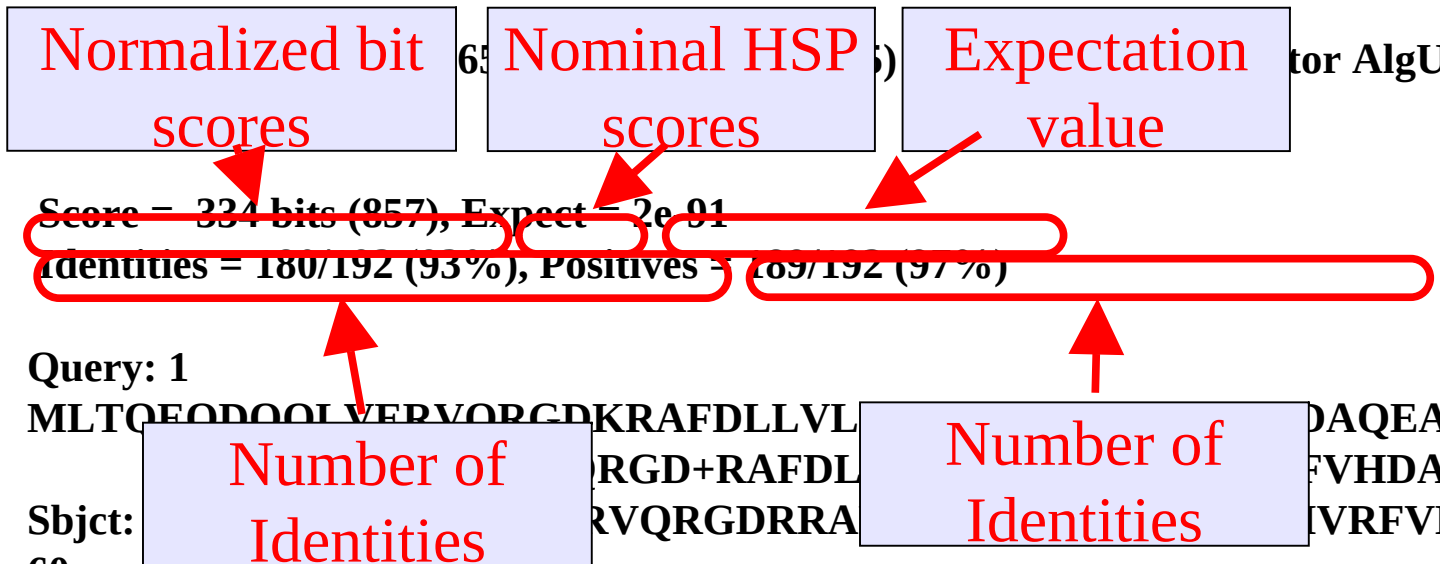
Query: 61 ALGNFRGDSAFYTWLYRIAINTAKNHLVARGRRPPDSDVTAEDAEFFEGDHALKDIESPE 120
ALGNFRGDSAFYTWLYRIAINTAKNHLVARGRRPPDSDV+A DAEF+EGDHALKDIESPE
Sbjct: 61 ALGNFRGDSAFYTWLYRIAINTAKNHLVARGRRPPDSDVSAGDAEFYEGDHALKDIESPE 120

Query: 121 RAMLRDEIEATVHQTIQQLPEDLRTALTLREFDGLSYEDIASVMQCPVGTVRSRIFRARE 180
R++LRDEIEATVH+TIQQLPEDLRTALTLREF+GLSYEDIA+VMQCPVGTVRSRIFRARE
Sbjct: 121 RSLLRDEIEATVHRTIQQLPEDLRTALTLREFDGLSYEDIASVMQCPVGTVRSRIFRARE 180

Alignment with query sequence

Query: 181 AIDKALQPLLRE 192
AIDKALQPLL+E

Interpreting Blast Results



65) tor AlgU [Azotobacter vinelandii]

Score = 334 bits (857), Expect = 2e-91
Identities = 180/192 (93%), Positives = 189/192 (97%)

Query: 1 MLTQFQDQQLVERVQRGDKRAFDLLVL DAQEAQDVAQEAFIKAYR 60
RGD+RAFDL FVHDA EAQDVAQEAFIKAYR
Sbjct: RVQRGDRRA IVRFVHDAHEAQDVAQEAFIKAYR 60

Query: 61 ALGNFRGDSAFYTWLYRIAINTAKNHLVARGRRPPDSDVTAEDAEFFEGDHALKDIESPE 120
ALGNFRGDSAFYTWLYRIAINTAKNHLVARGRRPPDSDV+A DAEF+EGDHALKDIESPE
Sbjct: 61 ALGNFRGDSAFYTWLYRIAINTAKNHLVARGRRPPDSDVSAGDAEFYEGDHALKDIESPE 120

Query: 121 RAMLRDEIEATVHQTIQQLPEDLRTALTLREFEGLSYEDIATVMQCPVGTVRSRIFRARE 180
R++LRDEIEATVH+TIQQLPEDLRTALTLREF+GLSYEDIA+VMQCPVGTVRSRIFRARE
Sbjct: 121 RSLLRDEIEATVHRTIQQLPEDLRTALTLREFDGLSYEDIASVMQCPVGTVRSRIFRARE 180

Query: 181 AIDKALQPLLRE 192
AIDKALQPLL+E

BLAST: On the Net, and On Your Computer

Advantages/Disadvantages of Net Based Blast:
(1) Use databases hosted remotely at NCBI.
(2) Little/No setup required.
(3) But, Cannot use a customized database.

Advantages/Disadvantages of Local Microcomputer-Based Blast:
(1) Can Use a Customized Database.
(2) Better suited to scripting / automation or when a large number of queries will be performed (UNIX).
(3) But, Requires some setup and computer expertise.

BLAST: On the Net, and On Your Computer

On the Net:
http://www.ncbi.nlm.nih.gov/BLAST/

On Your Computer:
UNIX/MacOS/Windows
ftp://ncbi.nlm.nih.gov/blast/executables/

NCBI Tools for UNIX
ftp://ncbi.nlm.nih.gov/toolbox/

WUBLAST
http://blast.wustl.edu

Learning More about BLAST

How Blast Works:
Altschul, S.F. et al., *Nucleic Acids Research*, **25**, 3389-3402 (1997).

Scoring Schemes:
Karlin, S., and Altschul, S.F., *Proc. Natl. Acad. Sci.*, 87, 2264-2268 (1990).

Henikoff, S., and Henikoff, J.G., *Proc. Natl. Acad. Sci.*, **89**, 10915-10919 (1992).

http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

Online Tutorial
http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html