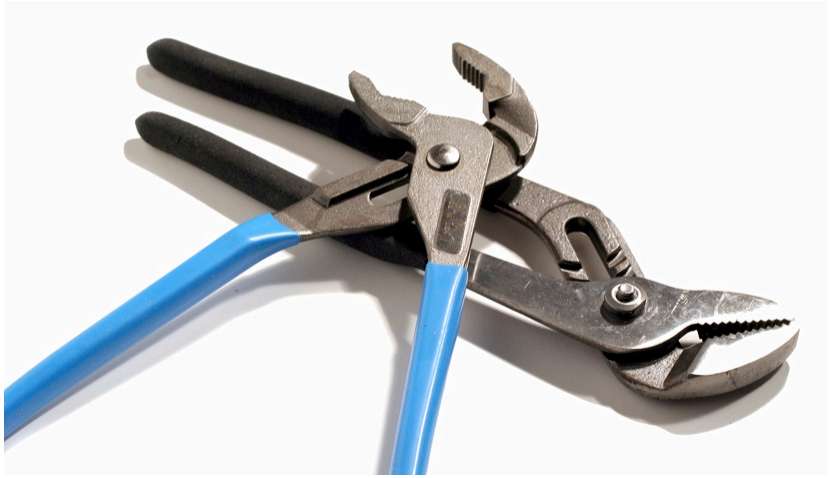
BLAST

Finding Function By Sequence Similarity



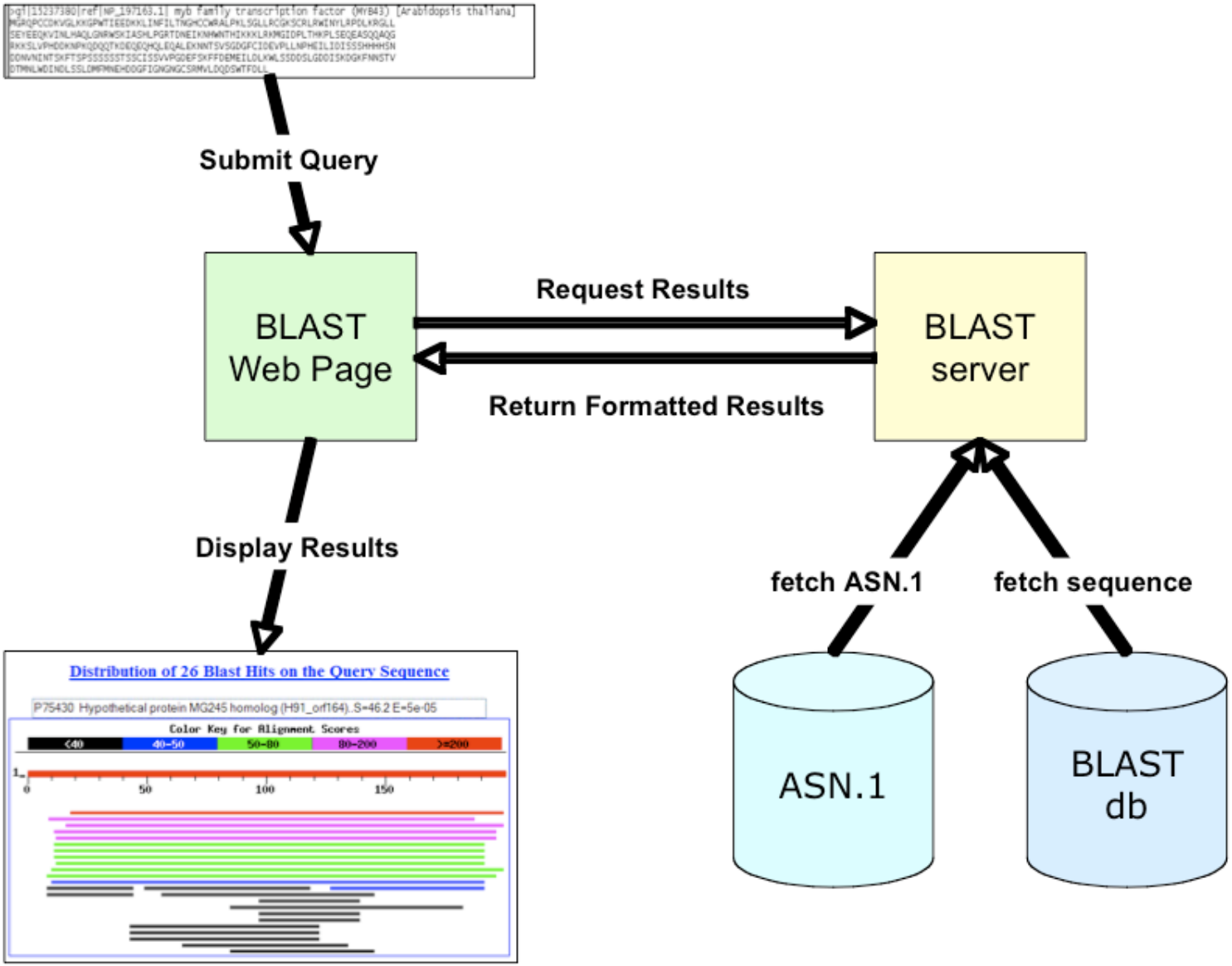
Concepts of Sequence Similarity Searching

- The premise:
- One sequence by itself is not informative; it must be analyzed by comparative methods against existing sequence databases to develop hypothesis concerning relatives and function.

The BLAST algorithm

- The BLAST programs (Basic Local Alignment Search Tools) are a set of sequence comparison algorithms introduced in 1990 that are used to search sequence databases for optimal local alignments to a query.

  - Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.
  - Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." NAR 25:3389-3402.


Submit Query
BLAST Web Page
Request Results
Return Formatted Results
BLAST server
Display Results
fetch ASN.1
fetch sequence
ASN.1
BLAST db
Distribution of 26 Blast Hits on the Query Sequence
Color Key for Alignment Scores

What BLAST tells you ...

- BLAST reports surprising alignments
  - Different than chance
- Assumptions
  - Random sequences
  - Constant composition
- Conclusions
  - Surprising similarities imply evolutionary homology

Evolutionary Homology: descent from a common ancestor
Does not always imply similar function

Basic **L**ocal **A**lignment **S**earch **T**ool

- Widely used similarity search tool
- Heuristic approach based on Smith Waterman algorithm
- Finds best local alignments
- Provides statistical significance
- www, standalone, and network clients

BLAST programs

Program	Description
blastp	Compares an amino acid query sequence against a protein sequence database.
blastn	Compares a nucleotide query sequence against a nucleotide sequence database.
blastx	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
tblastn	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
tblastx	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

more BLAST programs

Program		Notes
Megablast	Contiguous	Nearly identical sequences
	Discontiguous	Cross-species comparison
Position Specific	PSI-BLAST	Automatically generates a position specific score matrix (PSSM)
	RPS-BLAST	Searches a database of PSI-BLAST PSSMs

nucleotide only

protein only

BLAST Algorithm

- Scoring of matches done using scoring matrices
- Sequences are split into words (default n=3)
  - Speed, computational efficiency
- BLAST algorithm extends the initial "seed" hit into an HSP
  - HSP = high scoring segment pair = Local optimal alignment

Sequence Similarity Searching – The statistics are important

- Discriminating between real and artifactual matches is done using an estimate of probability that the match might occur by chance.

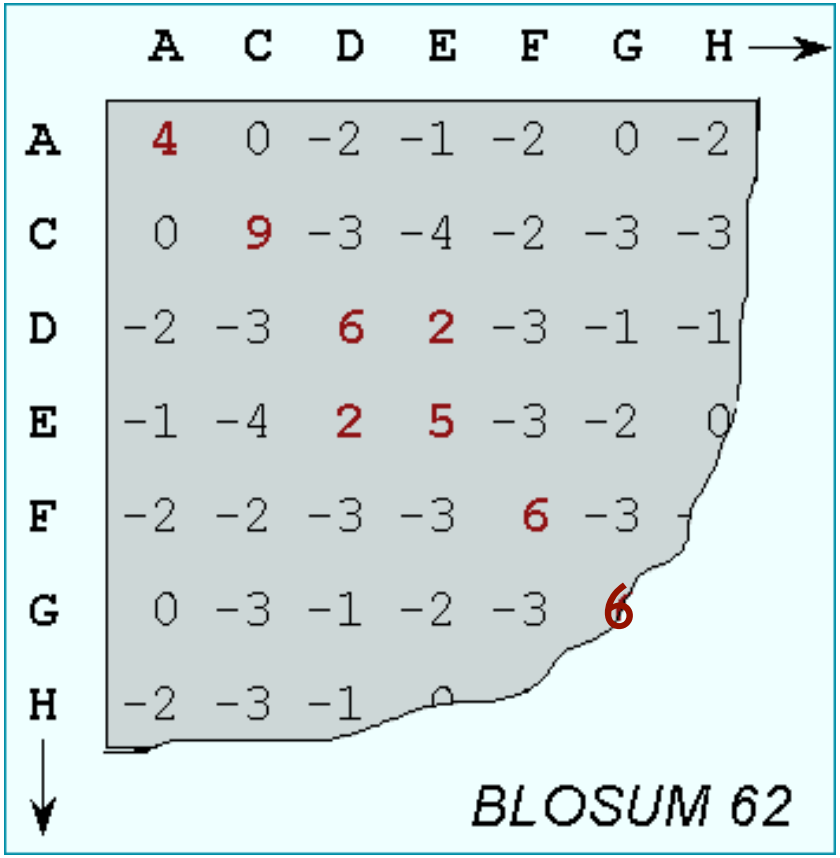
- We'll talk more about the meaning of the scores (S) and e-values (E) that are associated with BLAST hits

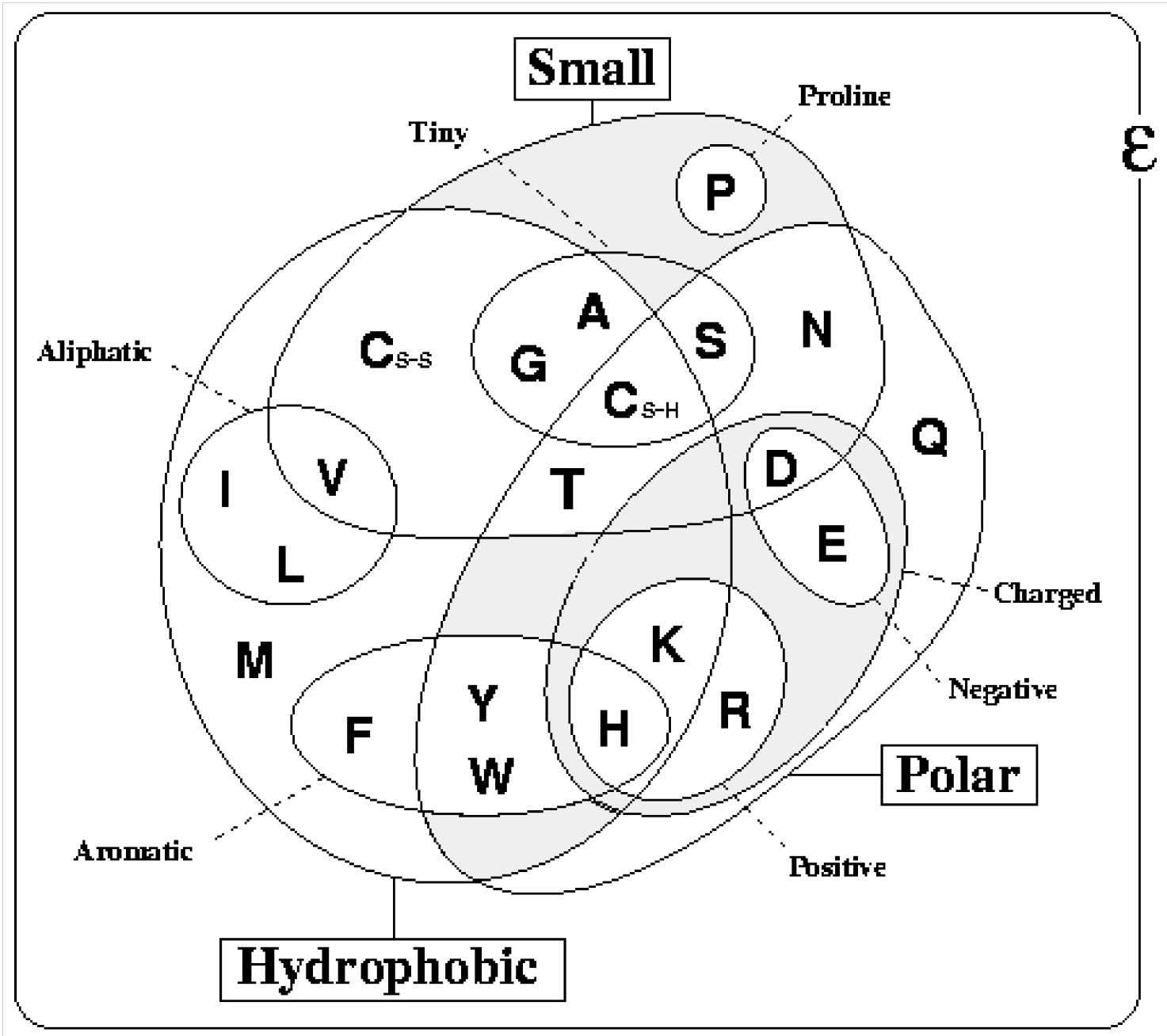
Where does the score (S) come from?

- The quality of each pair-wise alignment is represented as a score and the scores are ranked.
- Scoring matrices are used to calculate the score of the alignment base by base (DNA) or amino acid by amino acid (protein).
- The alignment score will be the sum of the scores for each position.

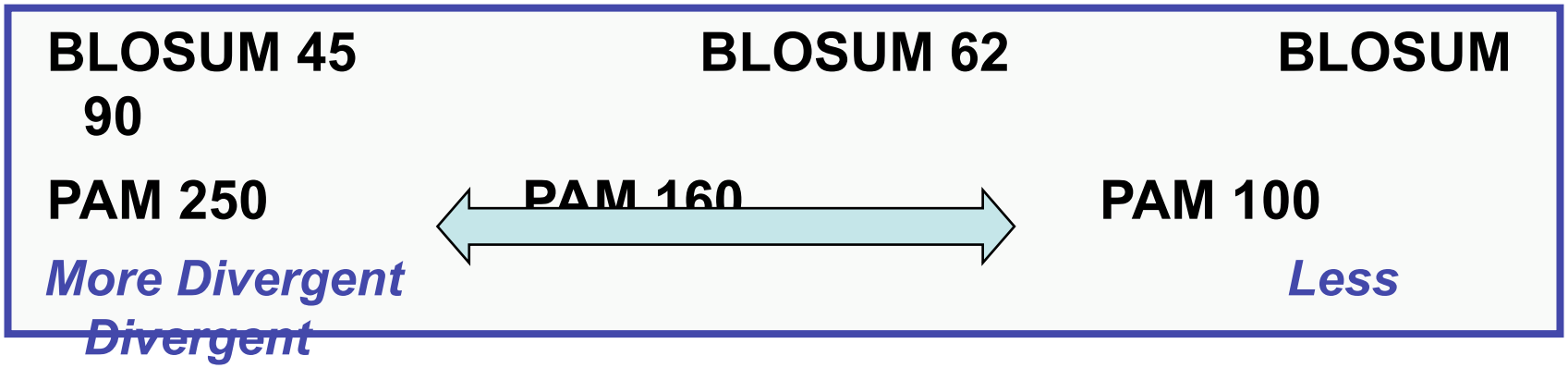
What's a scoring matrix?

- Substitution matrices are used for amino acid alignments.
  - each possible residue substitution is given a score
- A simpler unitary matrix is used for DNA pairs (+1 for match, -2 mismatch)




Small
Proline
Tiny
ε
P
A
S
N
Aliphatic
Cs-s
G
Cs-H
Q
D
I
V
T
L
E
Charged
K
M
Negative
Y
R
F
H
W
Polar
Aromatic
Positive
Hydrophobic

BLOSUM vs PAM



- BLOSUM 62 is the default matrix in BLAST 2.0. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix.

What do the Score and the e-value really mean?

- The quality of the alignment is represented by the Score (S).
- The score of an alignment is calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (PAM, BLOSUM) whereas gap scores are assigned empirically .
- The significance of each alignment is computed as an E value (E).
- Expectation value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

Notes on E-values

- Low E-values suggest that sequences are homologous
  - Can't show non-homology
- Statistical significance depends on both the size of the alignments and the size of the sequence database
  - Important consideration for comparing results across different searches
  - E-value increases as database gets bigger
  - E-value decreases as alignments get longer

Homology: Some Guidelines

- Similarity can be indicative of homology
- Generally, if two sequences are significantly similar over entire length they are likely homologous
- Low complexity regions can be highly similar without being homologous
- Homologous sequences not always highly similar

Suggested BLAST Cutoffs

Take Home Message:
Always look at your alignments

- Source: Chapter 11 – Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins
- For nucleotide based searches, one should look for hits with E-values of 10-6 or less and sequence identity of 70% or more
- For protein based searches, one should look for hits with E-values of 10-3 or less and sequence identity of 25% or more

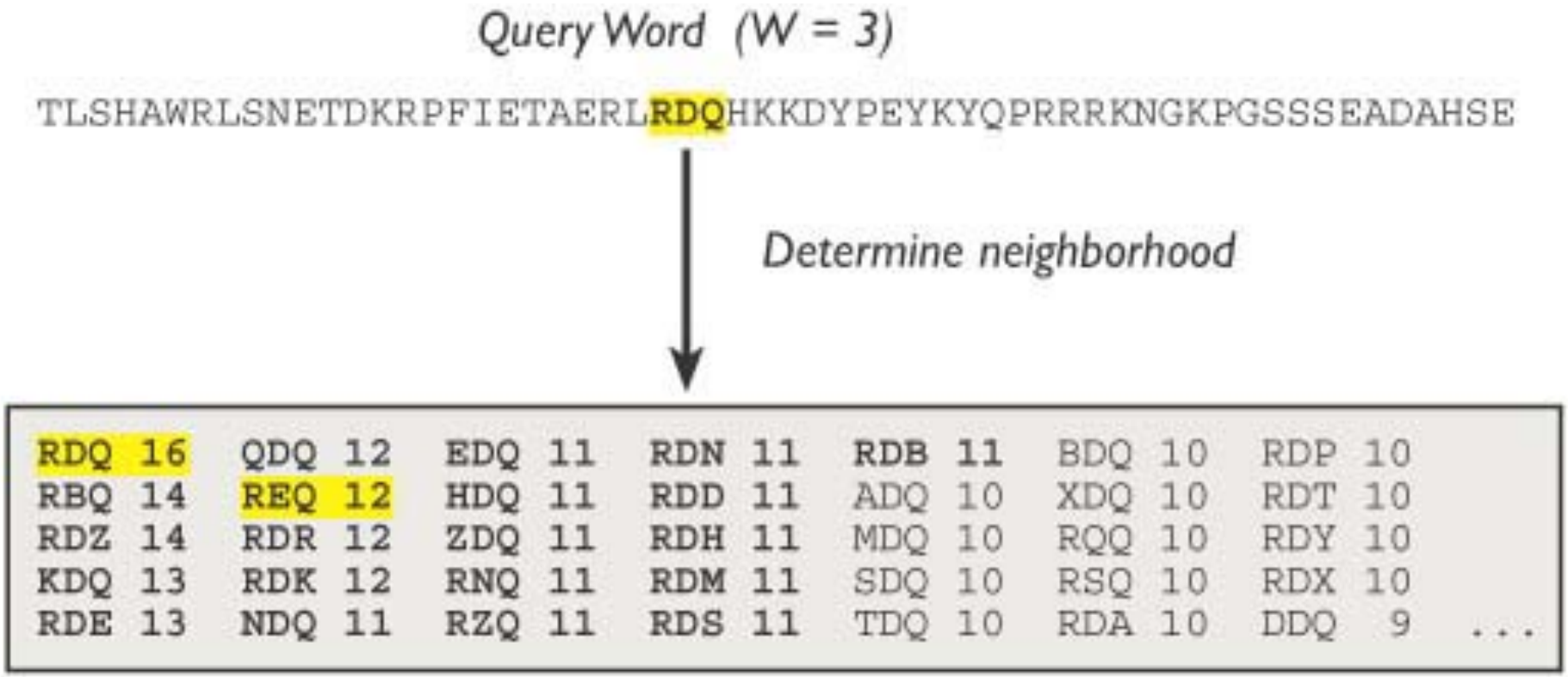
BLAST Algorithm

- Scoring of matches done using scoring matrices
- Sequences are split into words (default n=3)
  - Speed, computational efficiency
- BLAST algorithm extends the initial "seed" hit into an HSP
  - HSP = high scoring segment pair = Local optimal alignment

How Does BLAST Really Work?

- The BLAST programs improved the overall speed of searches while retaining good sensitivity (important as databases continue to grow) by breaking the query and database sequences into fragments ("words"), and initially seeking matches between fragments.
- Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of "S".

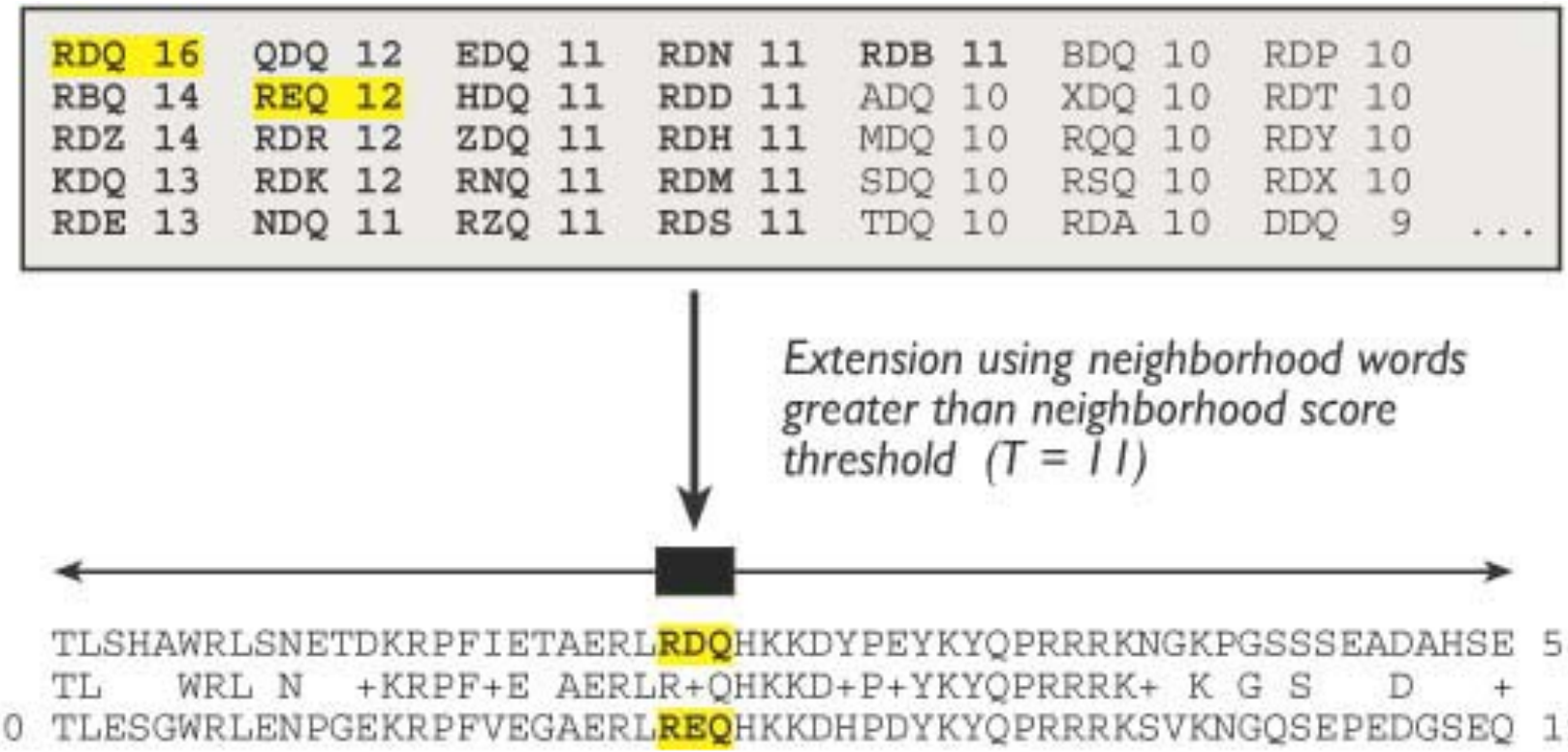
BLAST Algorithm



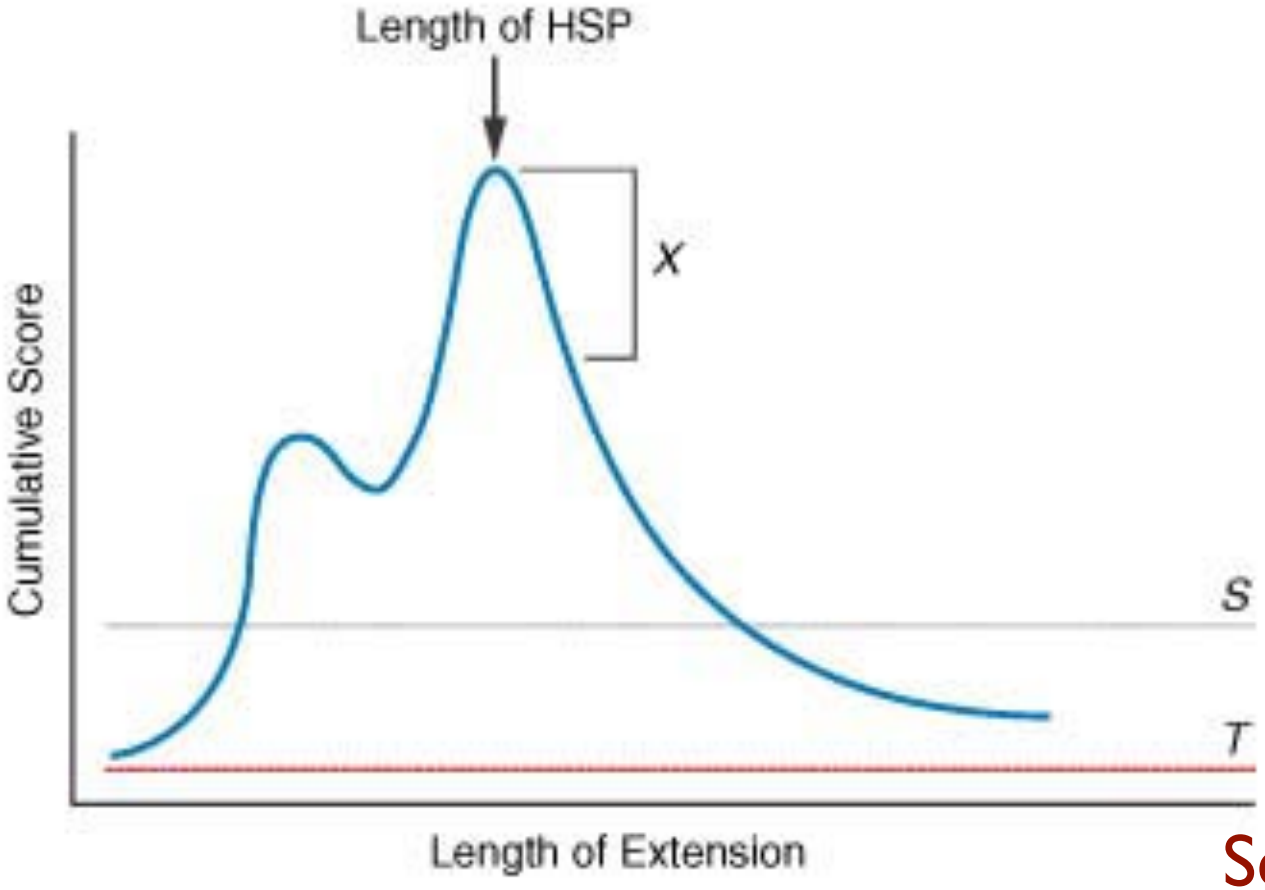
How Does BLAST Really Work?

- The BLAST programs improved the overall speed of searches while retaining good sensitivity (important as databases continue to grow) by breaking the query and database sequences into fragments ("words"), and initially seeking matches between fragments.
- Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of "S".

BLAST Algorithm



Extending the High Scoring Segment Pair (HSP)



Minimum Score (S)

Neighborhood Score Threshold (T)

```
>gb|AAL08419.1|  PTEN [Takifugu rubripes]
Length=412

 Score =  197 bits (501),  Expect = 2e-49, Method: Composition-based stats.
 Identities = 95/100 (95%), Positives = 98/100 (98%), Gaps = 0/100 (0%)

Query  2    IVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI  61
            +VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI
Sbjct  8    MVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI  67

Query  62   YNLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFKQN  101
            YNLCAERHYD AKFNCRVAQYPFEDHNPPQLELIKPF ++
Sbjct  68   YNLCAERHYDAAKFNCRVAQYPFEDHNPPQLELIKPFCED  107


 Score = 83.6 bits (205),  Expect = 4e-15, Method: Composition-based stats.
 Identities = 60/103 (58%), Positives = 68/103 (66%), Gaps = 32/103 (31%)

Query  99   KQNKMLKKDKMFHFWVNTFFIPGPEEV--------------------------------D  126
            KQNKM+KKDKMFHFWVNTFFIPGPEE                                 +
Sbjct  260  KQNKMMKKDKMFHFWVNTFFIPGPEESRDKLENGAVNNADSQQGVPAPGQGQPQSAECRE  319

Query  127  NDKEYLVLTLTkndldkankdkanRYFSPNFKVKLYFTKTVEE  169
            +D++YL+LTL+KND DKANKDKANRYFSPNFKVKL F+KTVEE
Sbjct  320  SDRDYLILTLSKNDRDKANKDKANRYFSPNFKVKLCFSKTVEE  362


>gb|AAH93110.1|  Ptenb protein [Danio rerio]
Length=289

 Score =  197 bits (500),  Expect = 2e-49, Method: Composition-based stats.
 Identities = 95/99 (95%), Positives = 98/99 (98%), Gaps = 0/99 (0%)

Query  3    VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIY  62
            VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHK+HYKIY
Sbjct  9    VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKDHYKIY  68

Query  63   NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFKQN  101
            NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPF ++
Sbjct  69   NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFCED  107
```

BLAST Algorithm

- Scoring of matches done using scoring matrices
- Sequences are split into words (default n=3)
  - Speed, computational efficiency
- BLAST algorithm extends the initial "seed" hit into an HSP
  - HSP = high scoring segment pair = Local optimal alignment

Credits

- Materials for this presentation have been adapted from the following sources:

  NCBI HelpDesk - Field Guide Course Materials

  Bioinformatics: A practical guide to the analysis of genes and proteins

- Questions? Please contact:

Dr. Joanne Fox
Michael Smith Laboratories
joanne@msl.ubc.ca