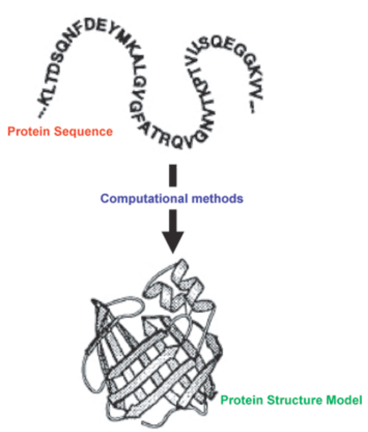
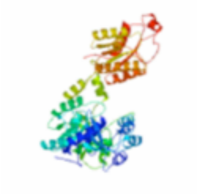
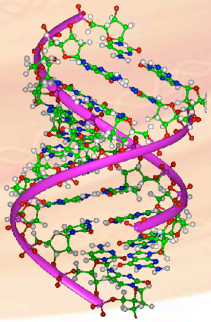




Bioinformatics Computational Methods 1 - BIOL 6308



Million years
3.4 5.1
H aactggtacg tggtcg tacgaactgg H
C aaccggttcg cggtcg ttcgaaccgg C
G aacctgttcg ctgtcg ttcgaacctg G
O atctggtacg tggtcg tacgatctgg O
Million years
7.0 2.2

November 19th 2013

http://155.33.203.128/cleslin/home/teaching6308F2013.php

Last Time

- Entropy and Information Content
- Exact vs. Heuristic
- Dynamic Programming
  - Needleman and Wunsch
  - Smith-Waterman
- Heuristic Methods for Sequence Database Searching
  - FASTA

Heuristics for Large-Scale Database Searching

- Exact solution to sequence alignment are computationally demanding
  - *O(mn)* too slow for large databases with high query traffic
- In practice algorithms are used that run much faster
  - Heuristic methods do fast approximation to dynamic programming
  - At the expense of possibly missing some significant hits due to the heuristics employed
- Such algorithms are usually seed-and-extend approaches in which first small exact matches are found which are then extended to obtain long inexact ones
  - FASTA
    - We know
  - BLAST
    - We'll learn about now

BLAST

BLAST

- Of all the sequence alignment algorithms, BLAST (basic local alignment search tool) is the most common
  - Cited by 43931 2012(November) – 48619 2013 (November)
- Typically used to compare one query nucleotide/protein sequence against a **database of sequences**
  - Uncover similarities and sequence matches
  - The success and popularity stems from combination of
    - **Speed**
    - **Sensitivity**
    - **Statistical assessment of the results**
- **Web version:**
  - **http://blast.ncbi.nlm.nih.gov/Blast.cgi**
- **Standalone version**
  - **ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/**

Different Usages of BLAST

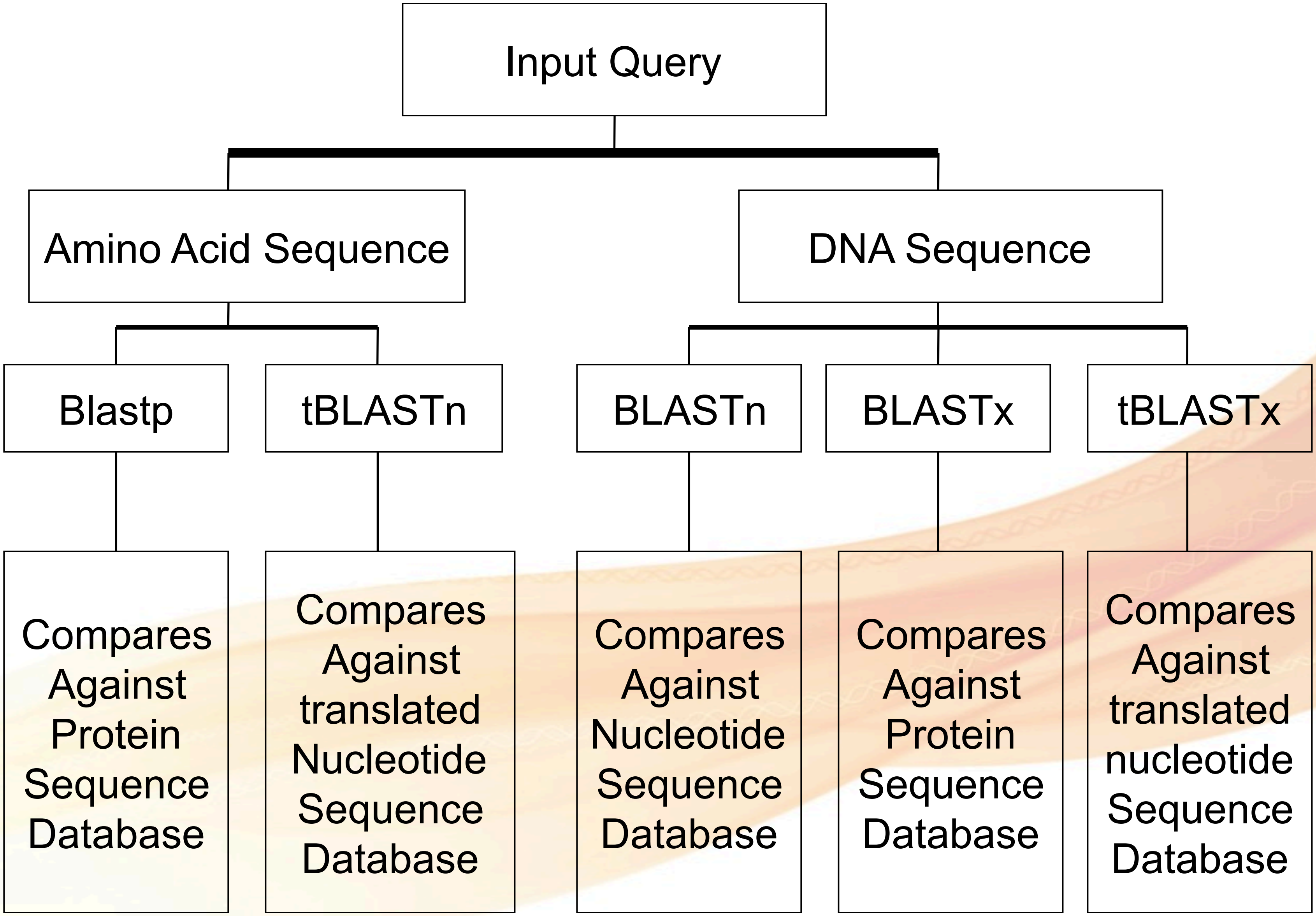
- If you are:
  - Looking for species
    - Sequencing DNA from unknown species
      - BLAST may help identify the correct species or homologous species
  - Looking for domains
    - BLAST a protein sequence (or a translated nucleotide sequence
      - BLAST will look for known domains in the query sequence
  - Mapping DNA to a known chromosome
    - Sequencing a gene from a known species but have no idea of the chromosome location, BLAST can help you
    - BLAST will show you the position of the query sequence in relation to the hit sequences
  - Annotations
    - Used to map annotations from one organism to another
    - Look for common genes in two related species

BLAST Method

- Heuristic method to find the high-scoring locally optimal alignments between a query sequence and a database
- Algorithm and family of programs rely on the statistics of gapped and un-gapped sequence alignments
- The statistics allow the probability of obtaining an alignment with a particular score to be estimated
- BLAST is unlikely to be as sensitive as a full dynamic programming algorithm
  - However, the underlying statistics provide a direct estimate of the significance of any match found
  - There's no comparison in terms of speed when compared to dynamic programming

Types of BLAST

- BLASTn
  - Nucleotide vs Nucleotide
- BLASTp
  - Protein vs Protein
- BLASTx
  - NT translated in all reading frames vs Protein
- tBLASTn
  - Protein vs NT database dynamically translated in all reading frames
- tBLASTx
  - NT translated 6 frames vs NT translated 6 frames
- PSI-BLAST
  - Compares a protein sequence to a protein database
  - Performs the comparison in an iterative fashion in order to detect homologs that are evolutionarily distant
  - Uses a dynamically calculated scoring matrix from the actual BLAST search
- BLAST2seq
  - Compares two protein or two nucleotide sequences


Input Query
Amino Acid Sequence
DNA Sequence
Blastp
tBLASTn
BLASTn
BLASTx
tBLASTx
Compares Against Protein Sequence Database
Compares Against translated Nucleotide Sequence Database
Compares Against Nucleotide Sequence Database
Compares Against Protein Sequence Database
Compares Against translated nucleotide Sequence Database

An Overview of BLAST

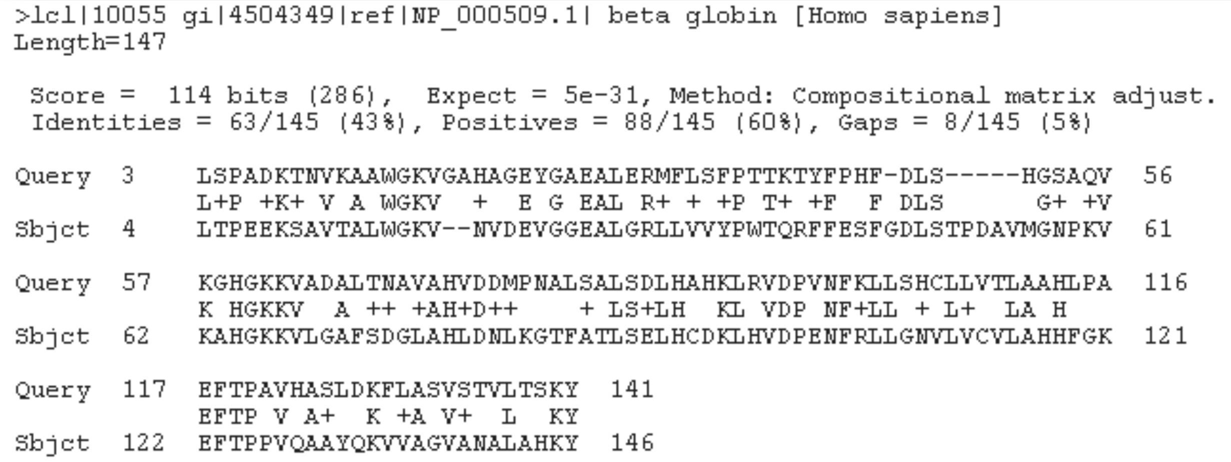
BLAST To the Rescue

```
>gi|4504347|ref|NP_000549.1| alpha 1 globin [Homo sapiens]
    MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKK
    VADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHA
    SLDKFLASVSTVLTSKYR

>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVK
    AHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTP
    PVQAAYQKVVAGVAN ALAHKYH
```

Dynamic programming or BLAST can be used

Antoine van Kampen

But BLAST Finds The Result Much Faster



```
>lcl|10055 gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
Length=147

 Score =  114 bits (286),  Expect = 5e-31, Method: Compositional matrix adjust.
 Identities = 63/145 (43%), Positives = 88/145 (60%), Gaps = 8/145 (5%)

Query  3    LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSAQV  56
            L+P +K+ V A WGKV  +  E G EAL R+ + +P T+ +F  F DLS      G+ +V
Sbjct  4    LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61

Query  57   KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA  116
            K HGKKV  A ++ +AH+D++    + LS+LH  KL VDP NF+LL + L+  LA H
Sbjct  62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121

Query  117  EFTPAVHASLDKFLASVSTVLTSKY  141
            EFTP V A+  K +A V+  L  KY
Sbjct  122  EFTPPVQAAYQKVVAGVANALAHKY  146
```

Adopted from Antoine van Kampen

Understanding the Steps in BLAST

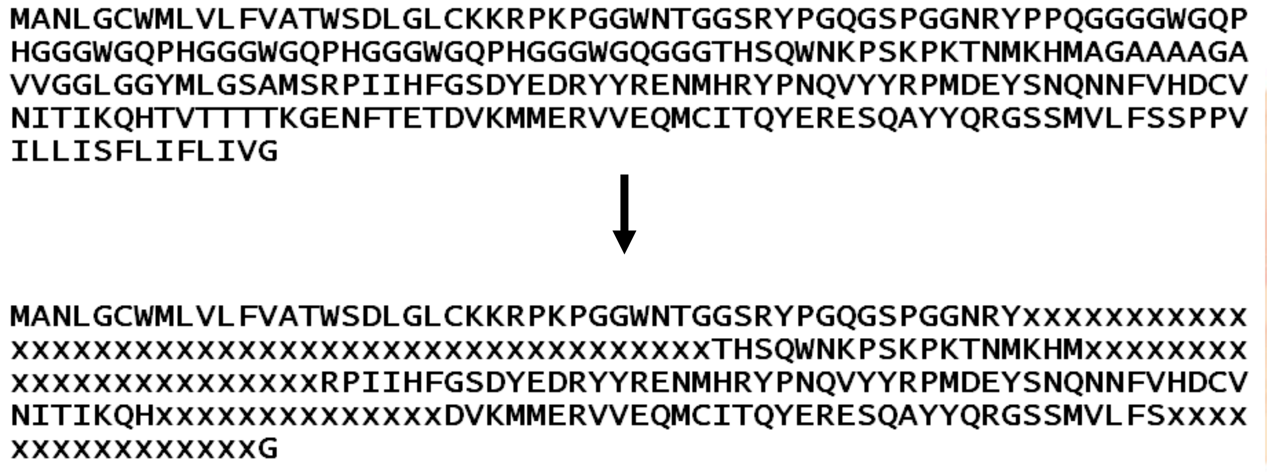
Complexity Filtering – 1st Step

- Masks segments of query sequence with low compositional complexity
  - For:
    - proteins : SEG [Wootton & Federhen]
    - DNA : DUST [Tatusov and Lipman]
- Real biological sequences have many regions where one or a few characters are over-represented (so-called low complexity regions):

```
ATGGPTIVLLVAAAAAAAAAAGPTPGLILW
     |||| |||||
EVVIKPSMCDHAAAATAAAAALCMKFC
```

- Regions bias the alignment because they tend to align with each other
  - Even absolutely unrelated sequences will have regions of false “similarity”
  - Filtering can eliminate statistically significant, but biologically uninteresting hits from the BLAST output
    - Hits against common acidic, basic, or proline-rich regions
- Leaving more biologically interesting regions of the query sequence available for specific matching against the database sequences

Masking Low Complexity Regions in Proteins (PSEG program)



```
MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQP
HGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGA
VVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
NITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPV
ILLISFLIFLIVG
```

↓

```
MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxTHSQWNKPSKPKTNMKHMxxxxxxxx
xxxxxxxxxxxxxxxRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
NITIKQHxxxxxxxxxxxxxxDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSxxxx
xxxxxxxxxxxxG
```

Filter (Mask for lookup table only) – 1st Step

- This option masks only for purposes of constructing the lookup table used by BLAST
  - -F “m D” (dna) or -F “m S” (proteins) (Legacy flags)
  - 2 phases of BLAST
    - Finding the hits based upon a lookup table
    - Extend those hits
- This option will “Mask for lookup table only”
  - No hits during lookup table generation have low-complexity sequence
- BLAST extensions are preformed w/o masking
  - Can be extended through low-complexity sequence
- Mask Lower Case Letters
  - Give regions or your sequence you liked masked in lower case letters
    - Allows you to customize what is filtered
      - U T

Nucleotide Words – 2nd Step

GTACTGGACATGGACCCTACAGGAACGTATACGTAAG Query

11-mer

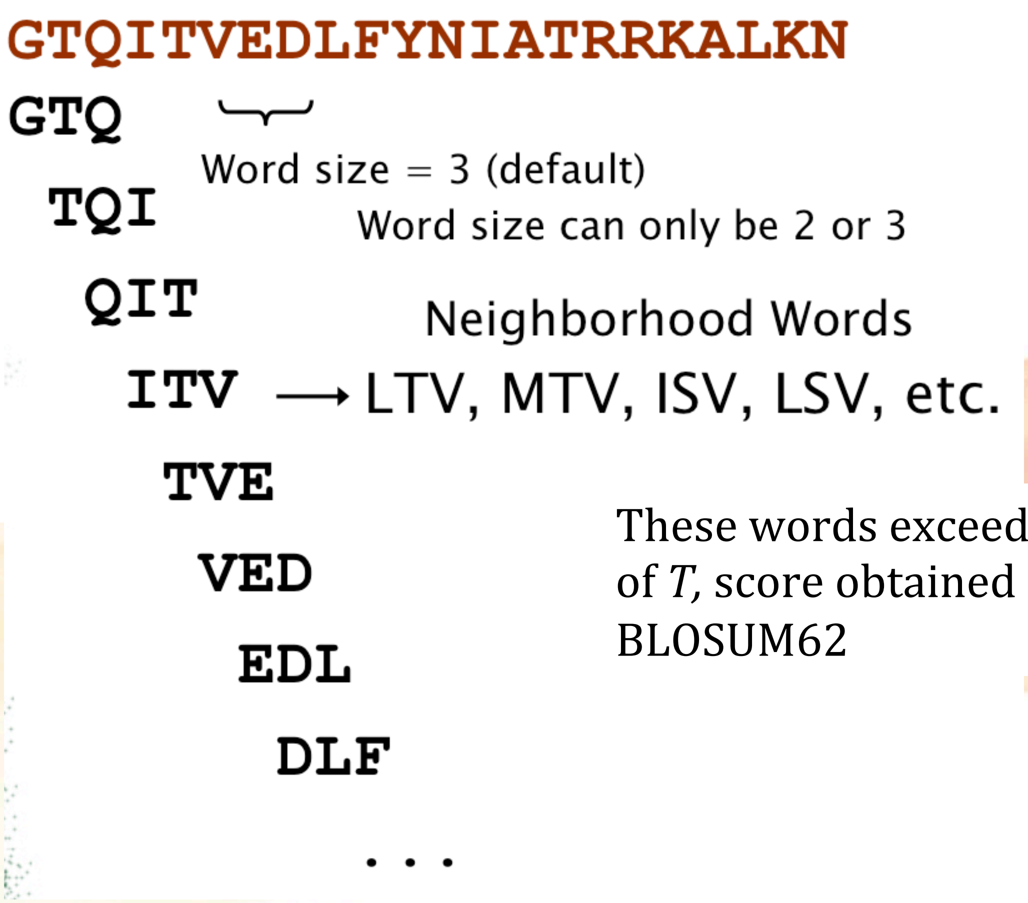
Make a lookup table of words

```
GTACTGGACAT
GTACTGGACATGGACCCTACAGGAACGT
 TACTGGACATG

  ACTGGACATGG

   CTGGACATGGA

    TGGACATGGAC
    TGGACATGGACCCTACAGGAACGTATAC
     GGACATGGACC

      GACATGGACCC

       ACATGGACCCT

          . . .
        CATGGACCCTACAGGAACGTATACGTAA

            .    .    .
```

WORD SIZE	Def.	Min.
blastn	11	7
megablast	28	12

Protein Words- 2nd Step

Make a lookup table of words



GTQITVEDLFYNIATRRKALKN

GTQ

Word size = 3 (default)

TQI

Word size can only be 2 or 3

QIT

Neighborhood Words

ITV → LTV, MTV, ISV, LSV, etc.

TVE

VED

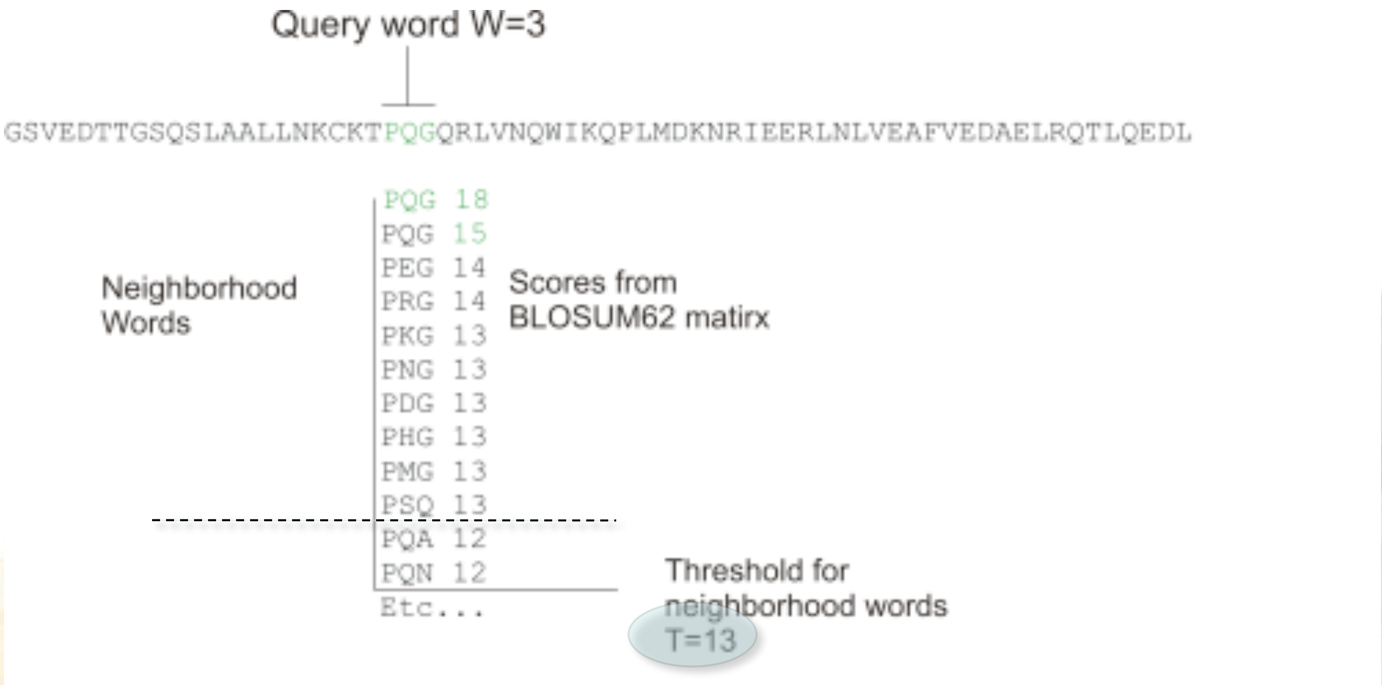
EDL

DLF

. . .

These words exceed the threshold of *T*, score obtained using BLOSUM62

Neighbor Words in Proteins – 3rd Step



Query word W=3

GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL

Neighborhood Words

PQG	18
PQG	15
PEG	14
PRG	14
PKG	13
PNG	13
PDG	13
PHG	13
PMG	13
PSQ	13
PQA	12
PQN	12

Etc...

Scores from BLOSUM62 matirx

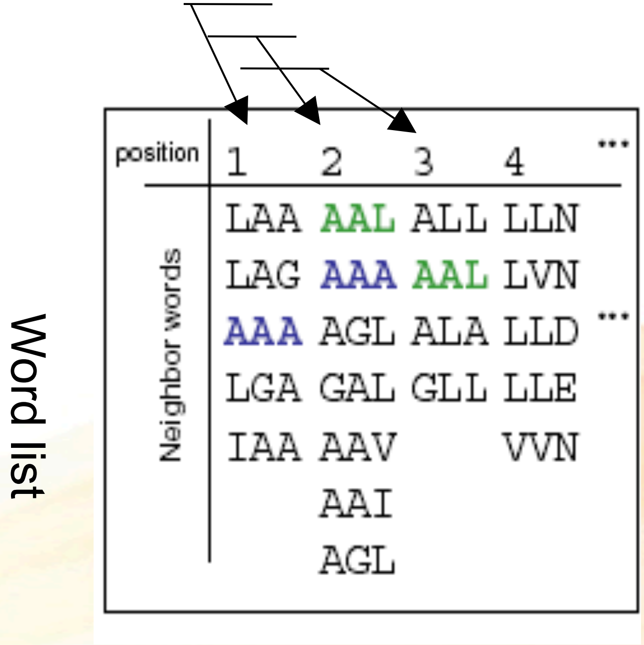
Threshold for neighborhood words T=13

We know how BLOSUM62 and other matrices are generated

Hash Table – 4th step

- Now the algorithm organizes words into Hash Table

Query: LAALLN…………



Word list

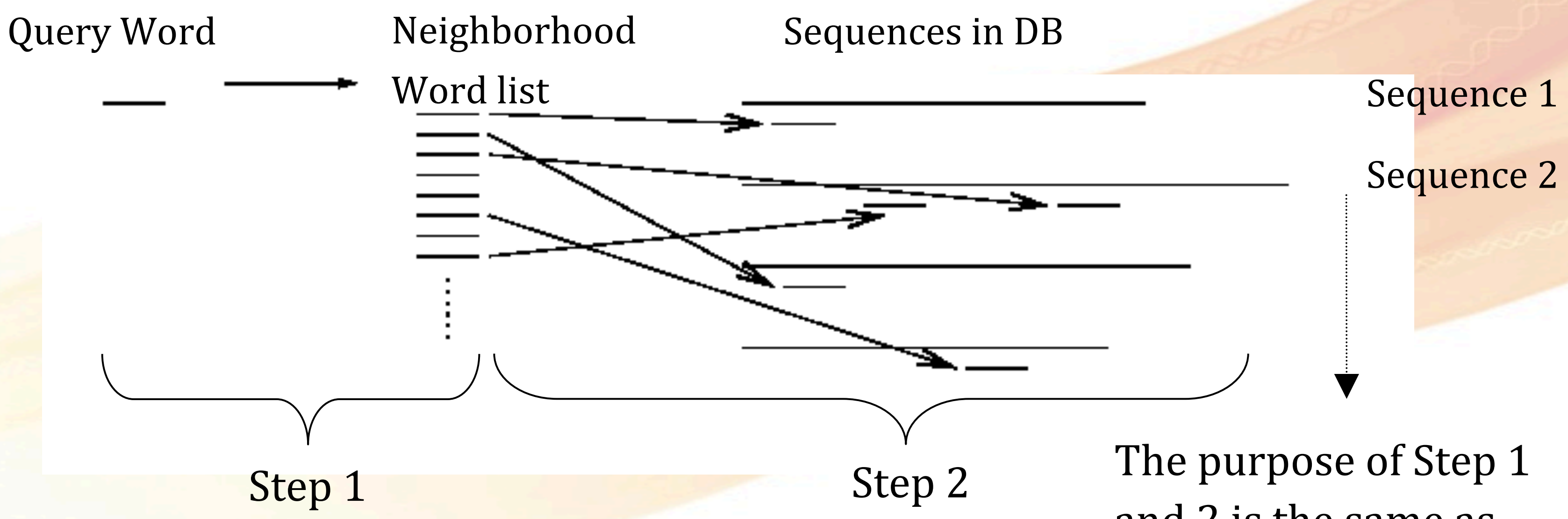
position	1	2	3	4	…
Neighbor words	LAA	AAL	ALL	LLN	
	LAG	AAA	AAL	LVN	
	AAA	AGL	ALA	LLD	…
	LGA	GAL	GLL	LLE	
	IAA	AAV		VVN	
		AAI			
		AGL			

Hash Table

word	position
AAA	1,2,15,16…
AAL	2,3,10,11…
LAA	1,5,7, …
GLL	3,8,34,…
VVN	4,21,25,…
:	:

Scanning the Database – 5th Step

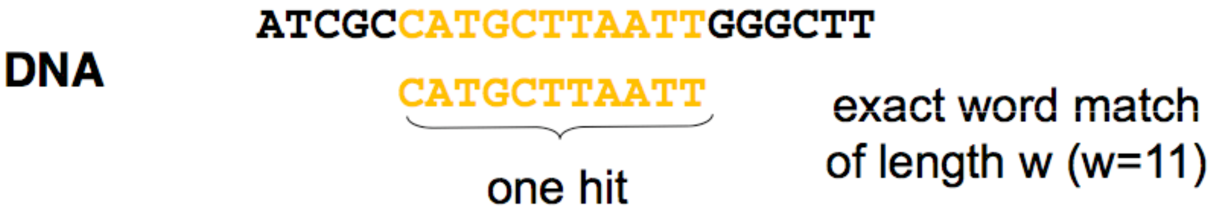
- Scanning DB
- For each words list, identify all exact matches with DB sequences



The purpose of Step 1 and 2 is the same as FASTA algorithm

This step generates seed for the 6th step

Minimum Requirements for a Seed - 6th Step



DNA

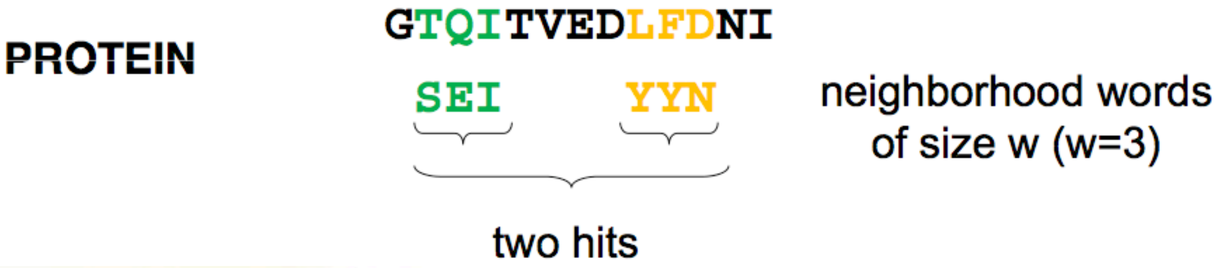
ATCGCCATGCTTAATTGGGCTT

CATGCTTAATT

one hit

exact word match of length w (w=11)

- •Nucleotide BLAST requires one exact match
- •Protein BLAST requires two neighboring matches within 40 aa



PROTEIN

GTQITVEDLFDNI

SEI YYN

two hits

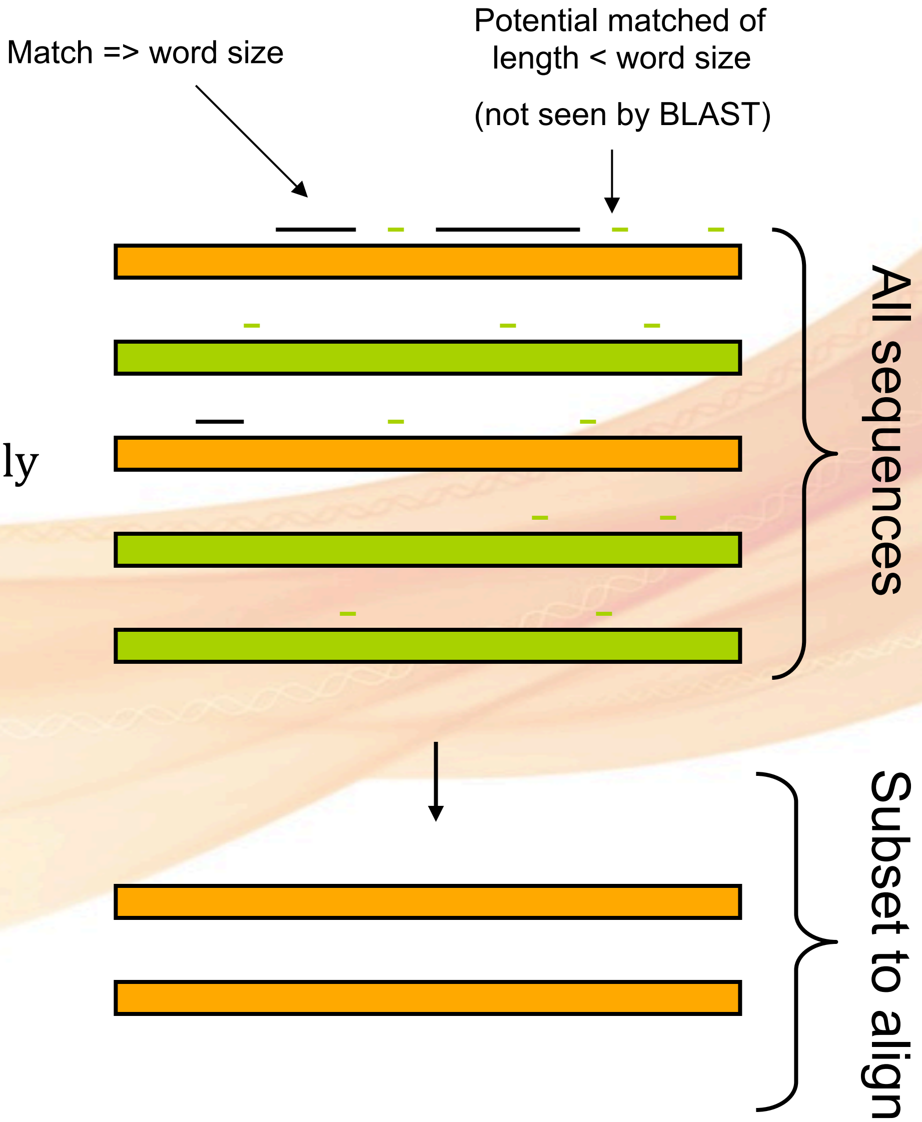
neighborhood words of size w (w=3)

BLASTn

- Alignment matrix:
  - Perfect match: **1**
  - Mismatch: **-3**

- Notice: All mismatched are equally penalized:
  - E.g. A:G == A:C == A:A

- Heuristics:
  - Perfect match “word” of the size: 7, 11 (default) or 15.

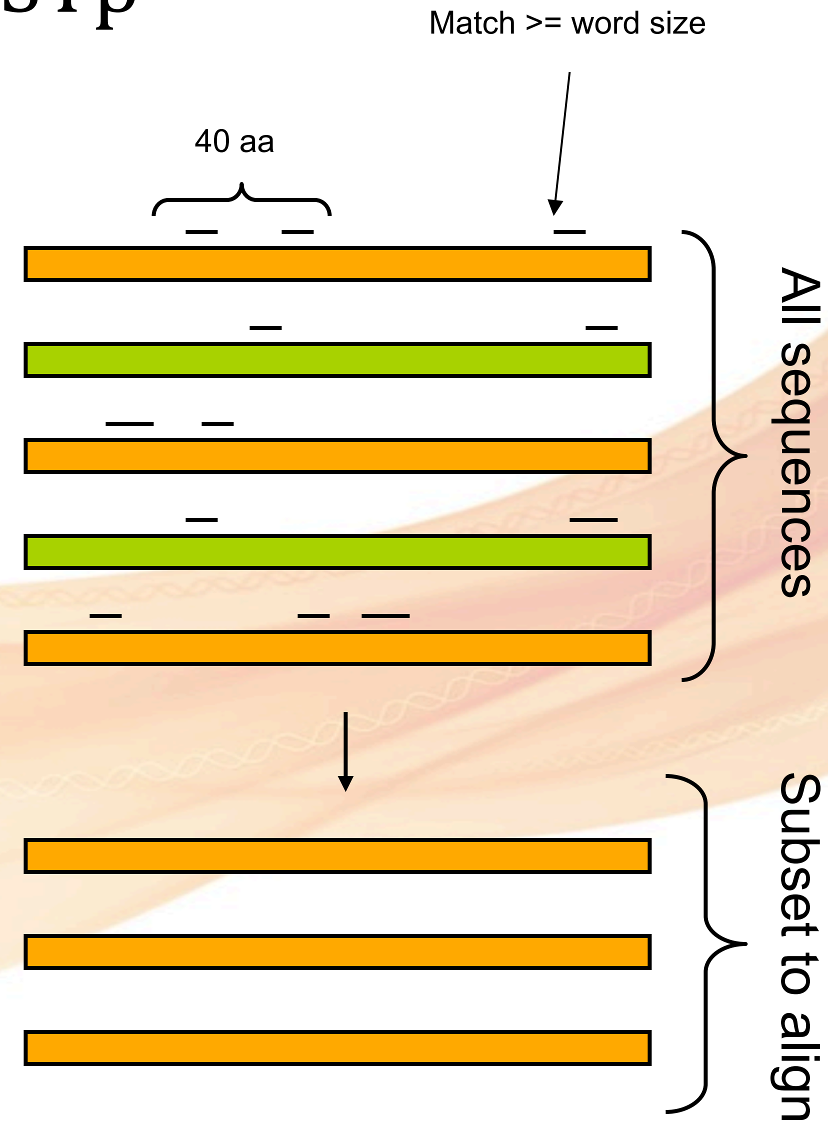


BLASTp

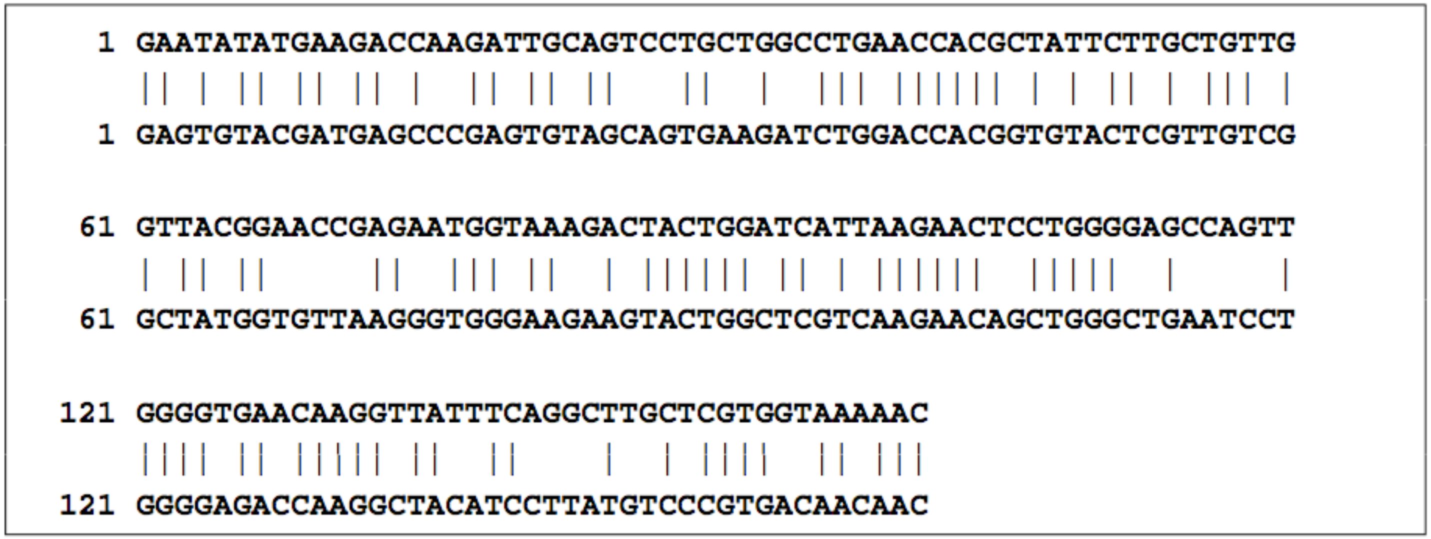
- Alignment matrix:
  - PAM and BLOSUM-series (default: BLOSUM 62)

- Notice: These alignment matrices incorporate knowledge about protein evolution

- Heuristics:
  - 2 x "Near match" within a windows (two hit method)
  - Default word length: 3 aa
  - Default window length: 40 aa
  - -A flag command line



An Alignment of DNA Sequences that BLAST Would Not Find Using W = 11



```
  1 GAATATATGAAGACCAAGATTGCAGTCCTGCTGGCCTGAACCACGCTATTCTTGCTGTTG
    || | || || || |  || || ||   ||  |  ||| |||||| | | || | ||| |
  1 GAGTGTACGATGAGCCCGAGTGTAGCAGTGAAGATCTGGACCACGGTGTACTCGTTGTCG

 61 GTTACGGAACCGAGAATGGTAAAGACTACTGGATCATTAAGAACTCCTGGGGAGCCAGTT
    | || ||     ||  ||| ||  | |||||| || | ||||||  |||||  |     |
 61 GCTATGGTGTTAAGGGTGGGAAGAAGTACTGGCTCGTCAAGAACAGCTGGGCTGAATCCT

121 GGGGTGAACAAGGTTATTTCAGGCTTGCTCGTGGTAAAAAC
    |||| || ||||| ||  ||    |  | ||||  || |||
121 GGGGAGACCAAGGCTACATCCTTATGTCCCGTGACAACAAC
```

Why ?

After Initial Words Found, Do Extension - 7th Step

- Extend hits in both directions
  - Each time alignment is extended, an alignment increases/decreases
  - Extension is continued until drop-off score reaches a threshold
    - Ensures alignment is not extended to regions where only very poor alignment occurs
    - Permits the match to cross a region of marginal homology or mismatching (e.g. a small intron in tBLASTn) if it flanks a region of high similarity
- If the Score of the alignment receives a score above threshold its reported

How Does Extension Work?

- Once search space is seeded, alignments generated by extending in both directions from the seed
- Example:

```
The quick brown fox jumps over the lazy dog.
The quiet brown cat purrs when she sees him.
```

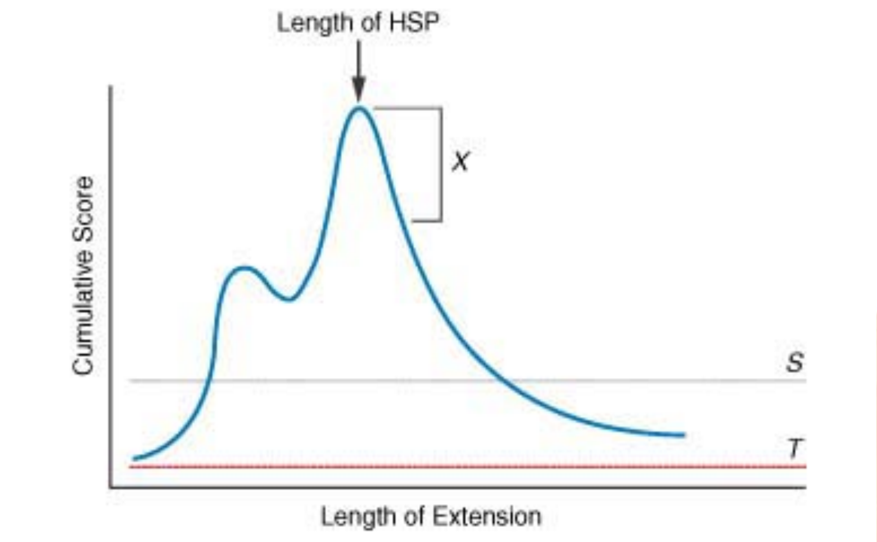
- Can align first six characters
- How far should we continue?

Extension (Continued)

- *X* parameter – How much is score allowed to drop off after last maximum?
- Example (assume identical scores +1 and mismatch scores -1)
- See *X* on the next slide

```
The quick brown fox jump
The quiet brown cat purr
123 45654 56789 876 5654 <- score
000 00012 10000 123 4345 <-drop off score (X)
```

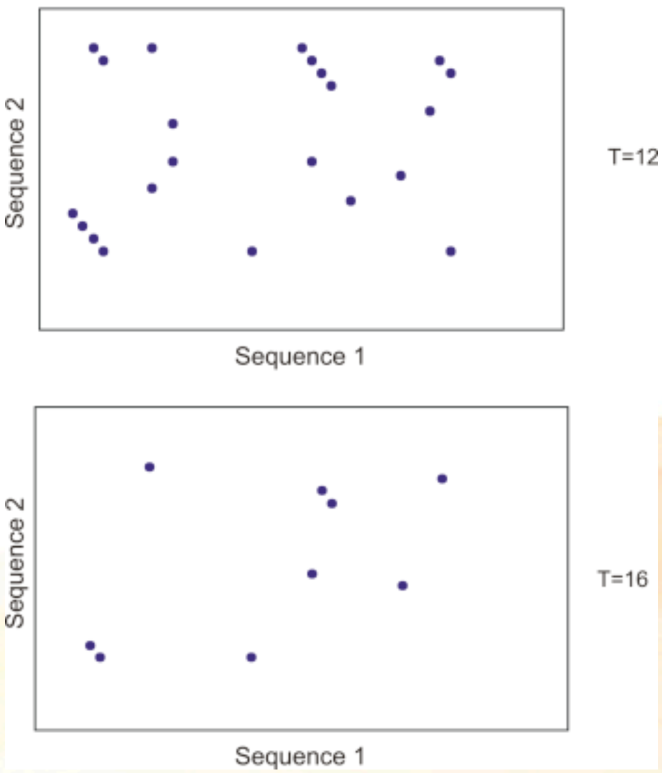
Extending the High Scoring Segment Pair (HSP)



Minimum Score (S)

Neighborhood Score Threshold (*T*)

Neighborhood Threshold - Proteins

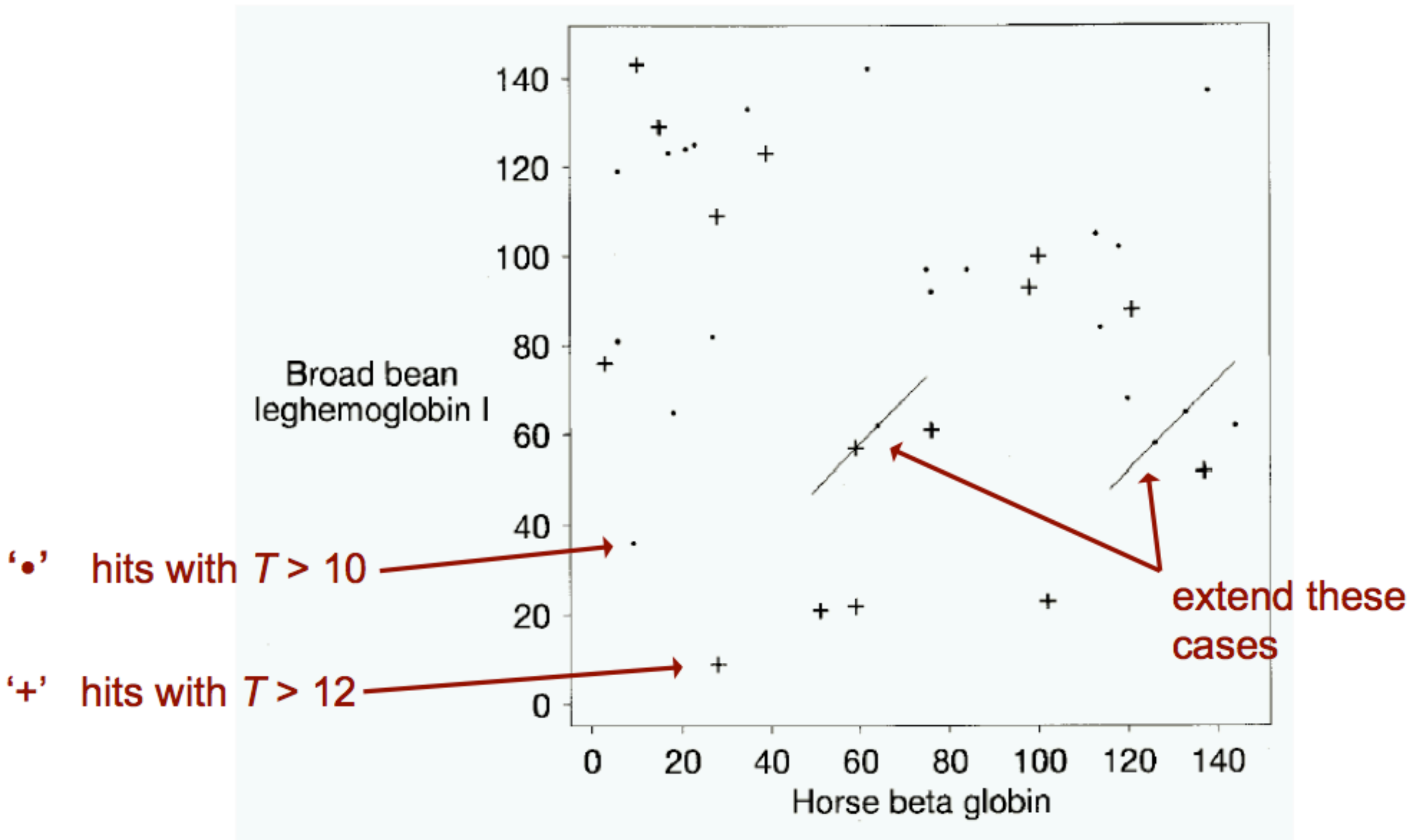


- The program declares a hit if the word taken from the query sequence has a score >= T **when a scoring matrix** is used
- Allows word size (W) to be kept high (for speed) without sacrificing sensitivity
- If T is increased
  - # of background hits is **reduced and the program will run faster**
  - Possible limit search space by increasing *T* or Word Size (*W*)
    - Increase the speed of BLAST
    - Loss of sensitivity
      - Legacy Flags
        - » `-f` from command line for *T*
        - » `-W` for Word Size

The Two Hit Method

- Extension step typically accounts for 90% of BLAST's execution time
- Key idea:
  - Do extension only when there are two hits on the same diagonal within distance A of each other
  - To maintain sensitivity, lower T parameter
    - More single hits found
    - But only small fraction have associated 2nd hit
- Main parameter controlling the sensitivity vs. running-time trade-off is T
  - Small T: greater sensitivity, more hits to expand
  - Large T: lower sensitivity, fewer hits to expand

The Two Hit Method



'•' hits with *T* > 10

'+' hits with *T* > 12

extend these cases

Figure from: Altschul et al. *Nucleic Acids Research* 25, 1997

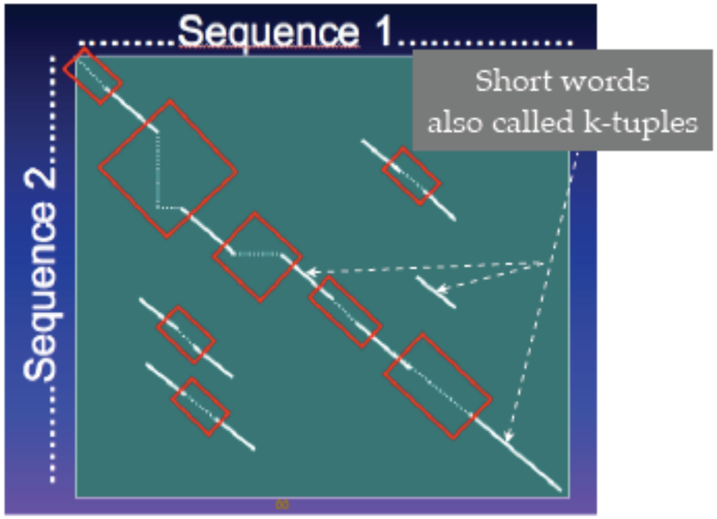
Word Size for DNA

- Changing word size has a great impact on the seeded sequence space
  -W 11 is default for DNA, 3 for protein
  - Change the word size to find sequence matches which would otherwise not be found using default parameters
- For instance, word size can be decreased when searching for primers or short nucleotides
  - BLASTn decrease word size of 11 to 7
  - Increase the E-value to 1000
  - Turn of complexity filtering

Keys to Speed with BLAST

- **Use word matching and prior indexing**
- BLAST local alignment is slow (like dynamic programming)
  - Only a small part of total search space is analyzed
  - B/c positions of all possible dataset word matches are indexed and stored prior to the BLAST search
    - Relevant parts of **search space are reached quickly**
- Tradeoff is in accuracy and certainty
- Occasionally matches will be missed
  - When they are **distant enough** and dispersed enough that **no local word pairs match well enough**

Only a fraction of the whole DP matrix is computed



The E-value

- *False positive expectation value*
- **Describes number of “hits” one can "expect" to see just by chance when searching a database of a particular size**
- Decreases exponentially as the Similarity Score (S) increases
  - Inverse relationship
  - Higher the Similarity Score, the lower the E-value
- Describes random background noise that exists for matches between two sequences
- Used as a convenient way to create a significance threshold for reporting results
- When E-value increased from 10 prior , a larger list with more low-similarity scoring hits can be reported

BLAST Match Statistics

- What should I use?
- It suffices to look at the E-value
  - Likelihood that matches of this score or better would be found by chance in the search
  - This is an expectation value, **not a P-value**
- # of matches of this score or better that would be expected if the database were composed of random sequences
- **Scores (bit scores) are independent of dataset size**
  - They simply measure the path weight of the specific alignment found
- **E-values are DEPENDENT on database size**
  - This is intuitive: in a random dataset, the more data - more likely find a match of a given score or higher

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

E value (Karlin-Altschul statistics)

- $\boldsymbol{E = K\, m\, n\, e^{-\lambda S}}$
  - ***K*** is a scaling factor (constant)
  - ***m*** is the length of the query sequence
  - ***n*** is the length of the database sequence
  - ***λ*** is the decay constant
  - ***S*** is the similarity score
- If ***S*** increases, ***E*** decreases exponentially
- If the decay constant increases, E decreases exponentially
- If ***m n*** increases the "search space" increases
  - Then there is a greater chance for a random "hit" and ***E*** increases
  - A larger database will increase ***E***
    - However, larger query sequence often decreases E. Why???

The parameters *K* and *lambda* can be thought of simply as natural scales for the search space size and the scoring system respectively

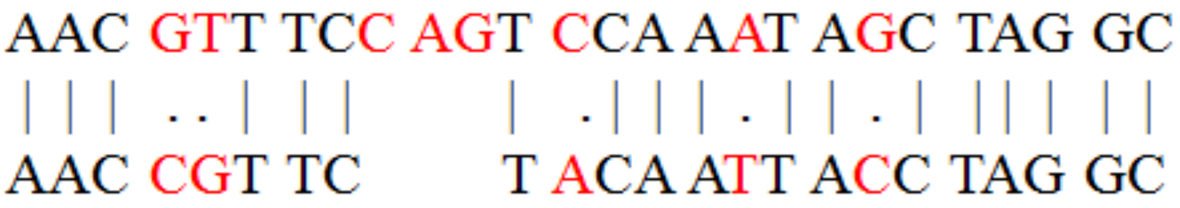
Score

- Score is a numerical value that describes the overall quality of an alignment
- Score indicates how good the alignment sequence are
- Higher numbers correspond to higher similarity of alignment

…Score

- For nucleotides, each identical match is given the same score, and all mismatches are given a penalty (negative) score
- Match : 1
- Mismatch : -2
- Gap opening : -2
- Gap extension ： -3

Example 1



```
AAC GTT TCC AGT CCA AAT AGC TAG GC
||| ..| ||    | .|| |.| |.| ||| ||
AAC CGT TC    T ACA ATT ACC TAG GC
```

Matches (+1): 18

Mismatches (-2): 5

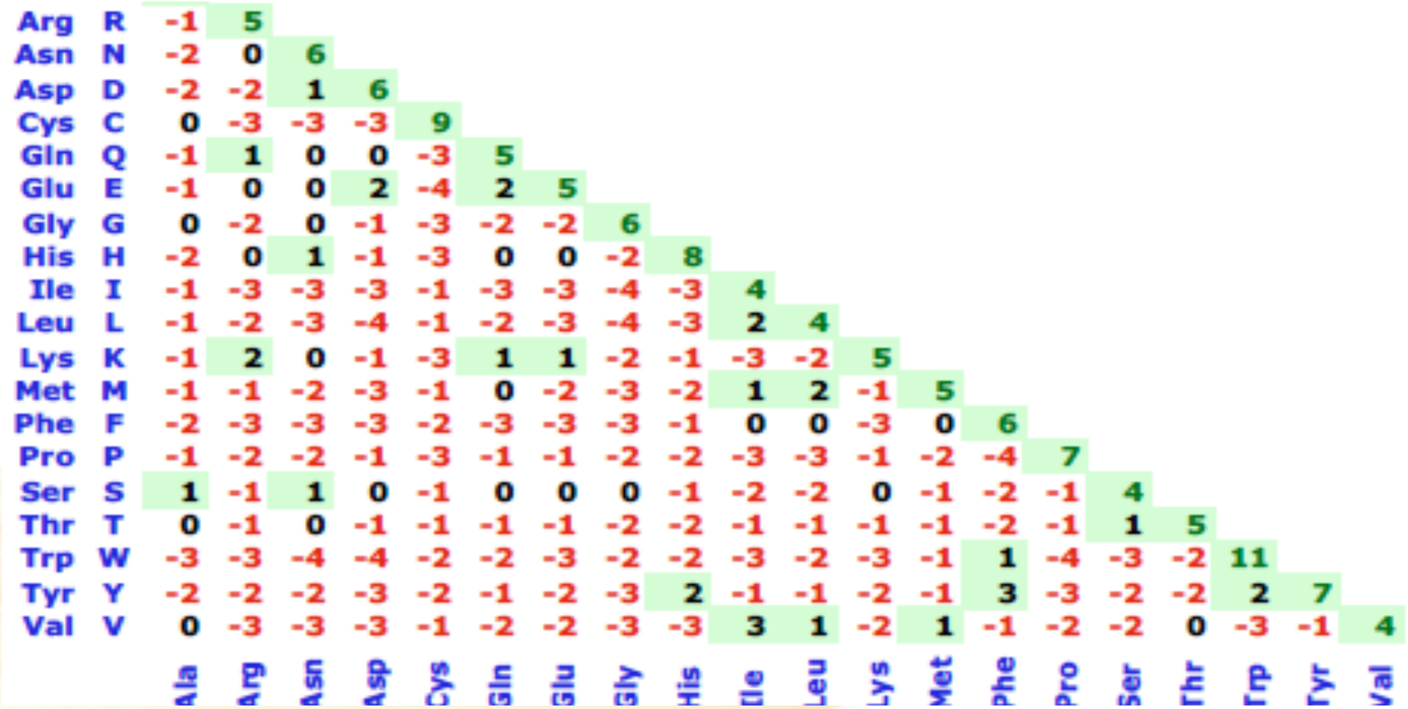
Gaps (opening -2, extension -1): 1, 2

Score = 18 * (+1) + 5 * (-2) +1 * (– 2) + 2 * (-1) = 4

…Score

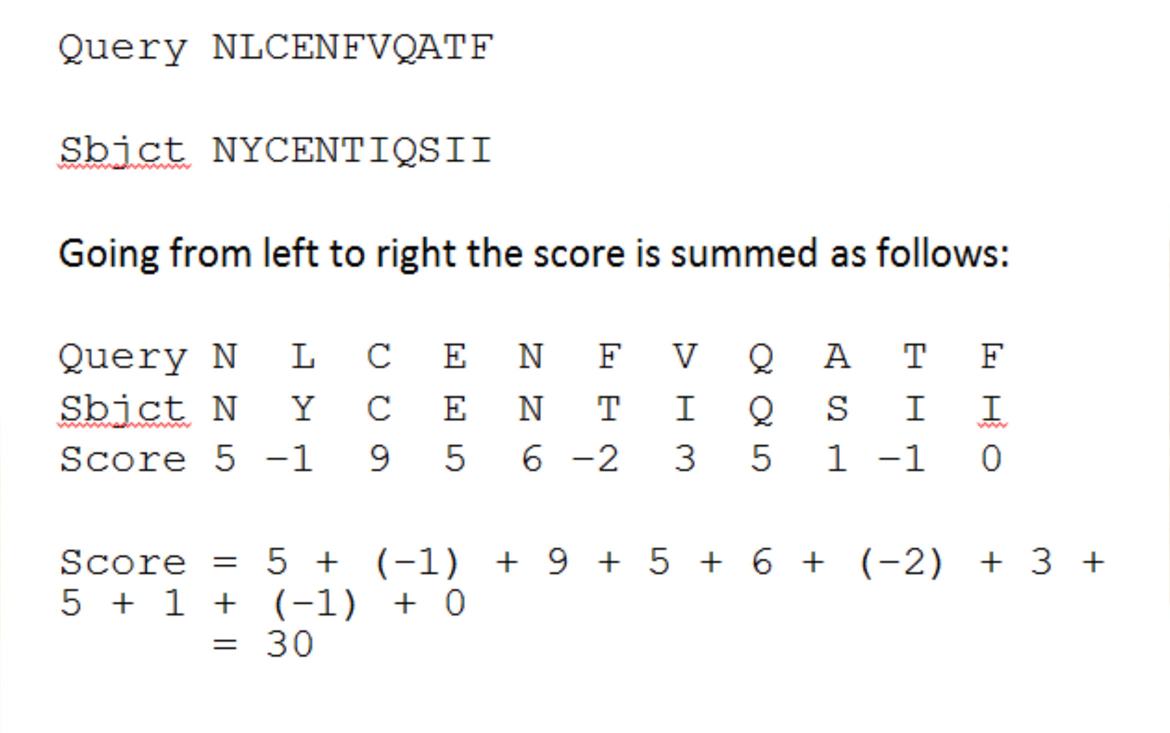
- For amino acid, blosum62 scoring matrices are used to obtain the S value
- BLOSUM = **BLO**cks **SU**bstitution **M**atrix
- Gap opening scoring is -4 and extension is -1

Blosum62 Scoring Matrix



		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Example 2

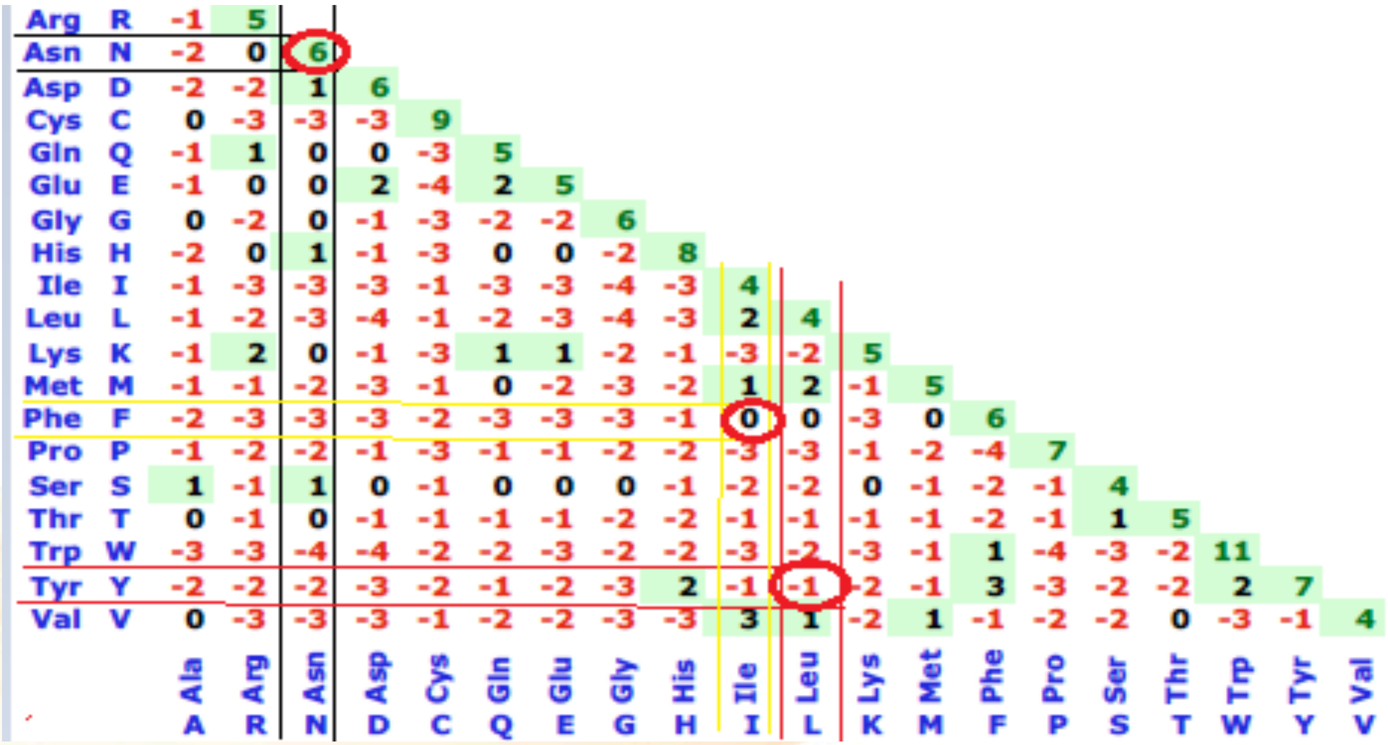


```
Query NLCENFVQATF

Sbjct NYCENTIQSII
```

Going from left to right the score is summed as follows:

```
Query N  L  C  E  N  F  V  Q  A  T  F
Sbjct N  Y  C  E  N  T  I  Q  S  I  I
Score 5 -1  9  5  6 -2  3  5  1 -1  0

Score = 5 + (-1) + 9 + 5 + 6 + (-2) + 3 +
5 + 1 + (-1) + 0
      = 30
```

How to Get Score From Blosum62 ?



		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Bit-Score

- Log-scaled version of a score
- Normalized score expressed in *bits*
- *Formula :*

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

- Where λ and *K* are parameters that characterize the expected distribution of S for the scoring system used (scoring matrix dependent)

Example 3

As referred from Example 2,
score = 30

##########From Bottom of the Blast report
Gapped

Lambda	K
0.318	0.14

##########

S' = (λ S - ln K) / ln 2
= (0.318 * 30 - ln (0.14) / ln 2
= 16.5998

E-value

- An expectation value, *E* for the alignment is calculated as:
- $\boldsymbol{E = m\, n\, 2^{-s'}}$

where *m* is the length of the database
n is the length of the query sequence
S′ is the normalized bit score from above

Example 4

As referred from Example 3, bit-score = 16.5998
assume m = 11 , n = 11

$E = mn2^{-s'}$
$= 11 * 11 * 2^{-16.5998}$
$= 1.22 \times 10^{-03}$

…E-value

- There is another way to calculate E-value without having bit-score.

- $\boldsymbol{E = K\, m\, n\, e^{-\lambda S}}$
- Where S is the score

  λ and K are constant parameter,

  m is the length of the database

  n is the length of the query sequence

Example 5

- As referred from Example 2, Score = 30
- λ = 0.318 and K= 0.14.
- Assume, m = 11 and n = 11
- $\boldsymbol{E = K\, m\, n\, e^{-\lambda S}}$

  $= 0.14 * 11 * 11 * e^{-(0.318 * 30)}$

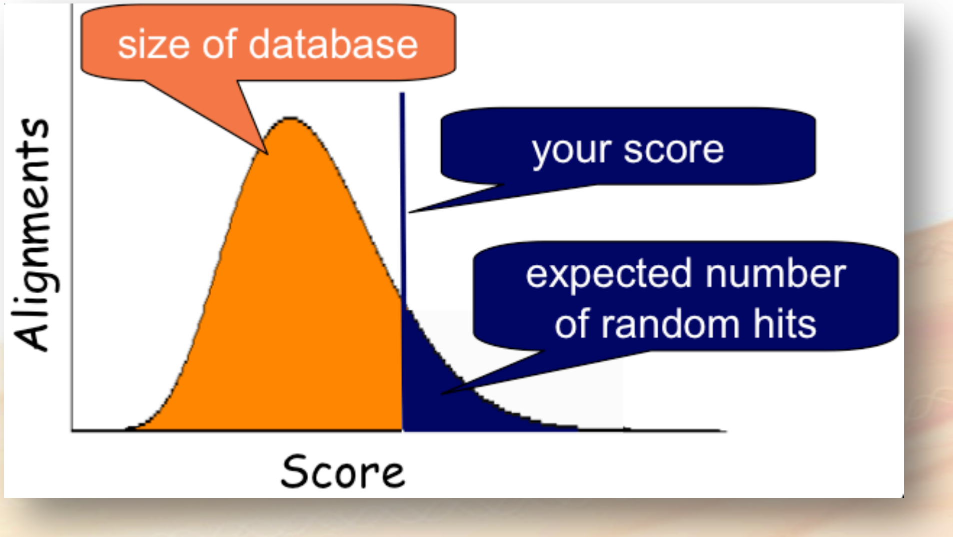
  $\mathbf{= 1.22 \times 10^{-03}}$

BLAST Alignment Statistics

- In a million entry database would leave
  - E = 0.001 1000 entries **due** to **chance**
  - E = 1e−6 would only leave **one** entry due to **chance**

BLAST Alignment Statistics

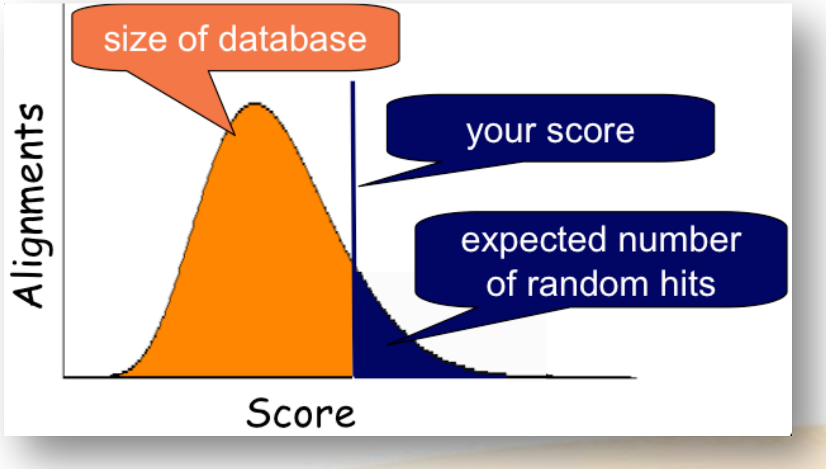
- Scores of local alignments between two random sequences follow the Extreme Value Distribution



Expect Value

E = number of database hits you expect to find by chance

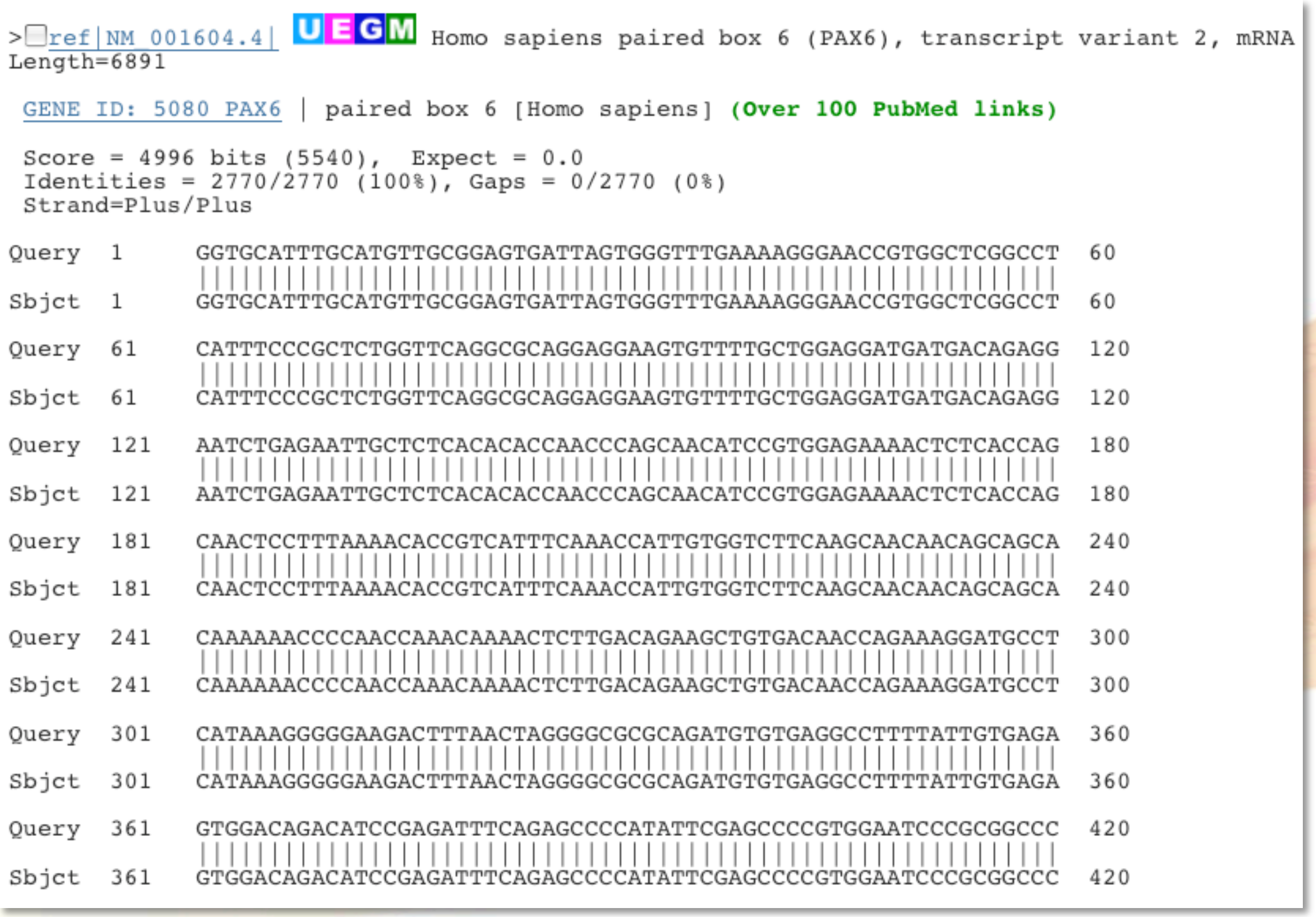
Database Searching: E-values in BLAST



- BLAST uses pre-computed extreme value distributions to calculate E-values from alignment scores
- Why - certain combinations of substitution matrices and gap penalties
- This also means that the fit is based on a different data set than the one you are working on

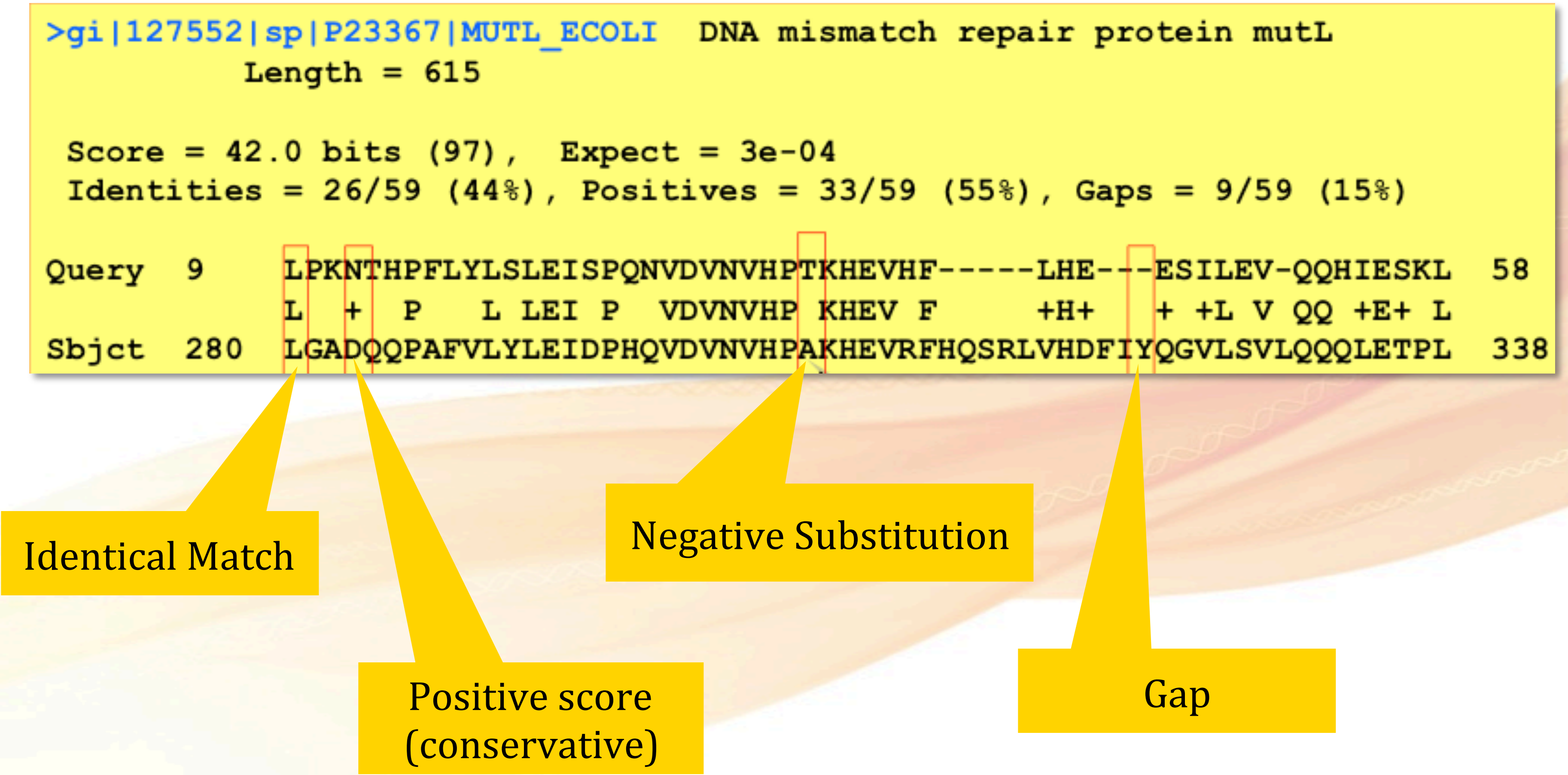
- A word of caution: **BLAST tends to overestimate the significance of its matches**
- E-values from BLAST are fine for identifying sure hits
  - Be careful using E-values to judge if a marginal hit can be trusted (e.g., you may want to use E-values of 10^{-4} to 10^{-5})

BLAST Output: DNA Alignments



```
>□ref|NM_001604.4| UEGM Homo sapiens paired box 6 (PAX6), transcript variant 2, mRNA
Length=6891

 GENE ID: 5080 PAX6 | paired box 6 [Homo sapiens] (Over 100 PubMed links)

 Score = 4996 bits (5540),  Expect = 0.0
 Identities = 2770/2770 (100%), Gaps = 0/2770 (0%)
 Strand=Plus/Plus

Query  1      GGTGCATTTGCATGTTGCGGAGTGATTAGTGGGTTTGAAAAGGAACCGTGGCTCGGCCT  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1      GGTGCATTTGCATGTTGCGGAGTGATTAGTGGGTTTGAAAAGGAACCGTGGCTCGGCCT  60

Query  61     CATTTCCCGCTCTGGTTCAGGCGCAGGAGGAAGTGTTTTGCTGGAGGATGATGACAGAGG  120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61     CATTTCCCGCTCTGGTTCAGGCGCAGGAGGAAGTGTTTTGCTGGAGGATGATGACAGAGG  120

Query  121    AATCTGAGAATTGCTCTCACACACCAACCCAGCAACATCCGTGGAGAAAACTCTCACCAG  180
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121    AATCTGAGAATTGCTCTCACACACCAACCCAGCAACATCCGTGGAGAAAACTCTCACCAG  180

Query  181    CAACTCCTTTAAAACACCGTCATTTCAAACCATTGTGGTCTTCAAGCAACAACAGCAGCA  240
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181    CAACTCCTTTAAAACACCGTCATTTCAAACCATTGTGGTCTTCAAGCAACAACAGCAGCA  240

Query  241    CAAAAAACCCCAACCAAACAAAACTCTTGACAGAAGCTGTGACAACCAGAAAGGATGCCT  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241    CAAAAAACCCCAACCAAACAAAACTCTTGACAGAAGCTGTGACAACCAGAAAGGATGCCT  300

Query  301    CATAAAGGGGGAAGACTTTAACTAGGGGCGCGCAGATGTGTGAGGCCTTTTATTGTGAGA  360
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301    CATAAAGGGGGAAGACTTTAACTAGGGGCGCGCAGATGTGTGAGGCCTTTTATTGTGAGA  360

Query  361    GTGGACAGACATCCGAGATTTCAGAGCCCCATATTCGAGCCCCGTGGAATCCCGCGGCCC  420
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361    GTGGACAGACATCCGAGATTTCAGAGCCCCATATTCGAGCCCCGTGGAATCCCGCGGCCC  420
```

BLAST Output: Protein Alignments



```
>gi|127552|sp|P23367|MUTL_ECOLI  DNA mismatch repair protein mutL
          Length = 615

 Score = 42.0 bits (97),  Expect = 3e-04
 Identities = 26/59 (44%), Positives = 33/59 (55%), Gaps = 9/59 (15%)

Query  9    LPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHF-----LHE---ESILEV-QQHIESKL  58
            L  +  P   L LEI P  VDVNVHP KHEV F     +H+   + +L V QQ +E+ L
Sbjct  280  LGADQQPAFVLYLEIDPHQVDVNVHPAKHEVRFHQSRLVHDFIYQGVLSVLQQQLETPL  338
```

Identical Match

Positive score (conservative)

Negative Substitution

Gap

Biologically Meaningful BLAST Hits

- Low E-value does not always imply “biologically meaningful”
- No strict rule how to choose E-value
- Trade-off between *sensitivity* and *specificity*
- E-value < 10^{-6} – **almost surely homologous**
  - **Will miss remote homologues**
- Between 10^{-2} and 10^{-6} – **probably homologous**
- Between 10^{-2} and 10 – might or might not be interesting…
  - Use your own judgments
  - Pairwise sequence comparison cannot detect remote homologues reliably when sequence identity drops below 20-35%
  - Need more sophisticated approaches

Advanced Parameters (Legacy Blast)

- `-G` Cost to open gap [Integer] default = 5 for nucleotides 11 proteins
- `-E` Cost to extend gap [Integer] default = 2 nucleotides 1 proteins
- `-q` Penalty for nucleotide mismatch [Integer] default = -3
- `-r` reward for nucleotide match [Integer] default = 1
- `-e` expect value [Real] default = 10
- `-W` wordsize [Integer] default = 11 nucleotides 3 proteins
- `-y` Dropoff (X) for BLAST extensions in bits (default if zero) default = 20 for BLASTn, 7 for other programs
- `-Z` final X dropoff value for gapped alignment (in bits)50 for BLASTn, 25 for other programs

- Easy to determine new flags in BLAST+ with Perl script provided by NCBI

Biological Relevance

- Were you looking for a **short region of nearly identical sequence** or a **longer region of general similarity**?
  - If it's a protein
    - Are the mismatches conservative ones?
    - Is it a shared domain?
    - Is there a shared motif?
    - Are key residues conserved i.e. (active site or ligand binding sites)
  - If it's DNA
    - Are the matching regions important structural components of the gene?
    - Are these matching regions just introns or flanking regions?
    - Why am I using DNA?

Borderline Similarity (1)

- What about matches with E-values in the 0.5 - range?
  - Depends on what you’re looking for?
    - Length dependent
      - Long genes with short matches probably not significant
      - But what about primers/probes
        - » Searching much shorter sequence matches which will have much higher E-values?
        - » Have to set the E-value to 1000 and Wordsize 7
        - » You can still potentially miss hits?

Borderline Similarity (2)

- Retest the sequences you're concerned with
  - Similarity is transitive:
    - If A~B, B~C, then A~C
    - How could we do this?
    - We could use Perl for this project, right!
  - But be careful hits might not be transitive with multi-domain proteins

The Twilight Zone

- Identity (proteins) drops into 20-35%
  - Estimates of statistical significance fail to distinguish between related and unrelated sequences
  - Relatedness cannot be distinguished from random match
    - referred to as the twilight zone of pairwise sequence alignment
    - pinkish area of the overlap you will see in the upcoming lab

- Proteins with similar structure/function that have pairwise sequence identity below 20-35% can score lower than structurally and functionally dissimilar proteins

Be Careful of Sequence Titles

A Sequence to BLAST

```
>YP_003565295
MKAKLIQYVVYDAECRLFKSVNQHFDRKHLNRFLRLLTHAGGATFTIVIACLLLFLYPSSV
AYACAFSLAVSHIPVAIAKKLYPRKRPYIQLKHTKVLENPLKDHSFPSGHTTAIFSLVTP
LMIVYPAFAAVLLPLAVMVGISRIYLGLHYPTDVMVGLILGIFSGAVALNIFLT
```

- NCBI--> BLAST --> Protein BLAST
- Write in accession
- Note the database: default is nr
- Algorithm: BLASTp is the standard one used for protein sequences
- Algorithm parameters: just under the BLAST button on the bottom of the page
  - Maximum number of sequences to display (default = 100)
  - Expect threshold: maximum e-value to display (default = 10)
  - Scoring matrix and gap penalties
  - Low complexity filters
- When in doubt, accept the default

BLAST Online Results (1)

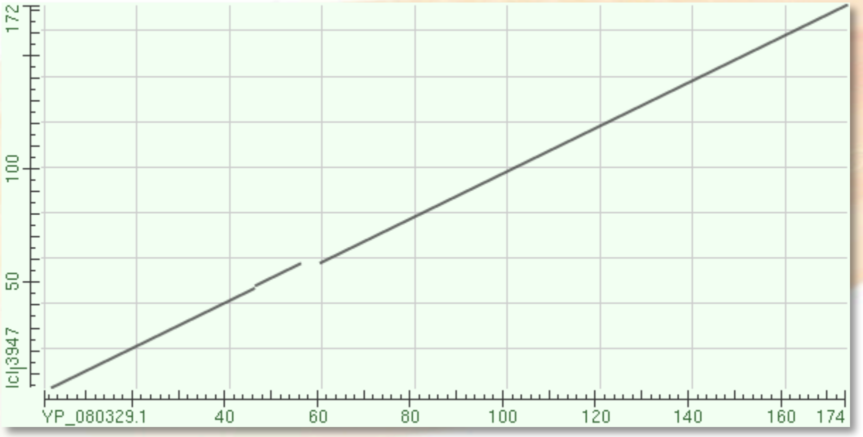
- Conserved Domains Database (CDD) automatically searched
  - ex. PAP2-like superfamily, which can be clicked on for more info
- Graphic Summary: Representation of top 100 hits
  - Length bars illustrates extent of homologous region
  - Color shows alignment scores (bit scores)
  - Mouse-over bars to get info about each hit
- Descriptions: A list of the sequences hitting your query
  - Has a link:
    - to that gene
    - gene name and (often) species, bit score, e-value

BLAST Online Results (2)

- Alignments: Detailed alignments for each hit
  - Get percentage of identities and positives
  - Number of gap bases, sequence length
  - “Expect” = e-value
  - Aligned query and subject sequences, with identities and positives shown between them, and position numbers in the sequences
- One of the top hits is annotated incorrectly:
  - Everything else including the CDD domain information says this is a phsophatase
  - But this hit says it is ribosomal protein S2 from *Bacillus licheniformis* (what's new!)

BLAST Two Sequences

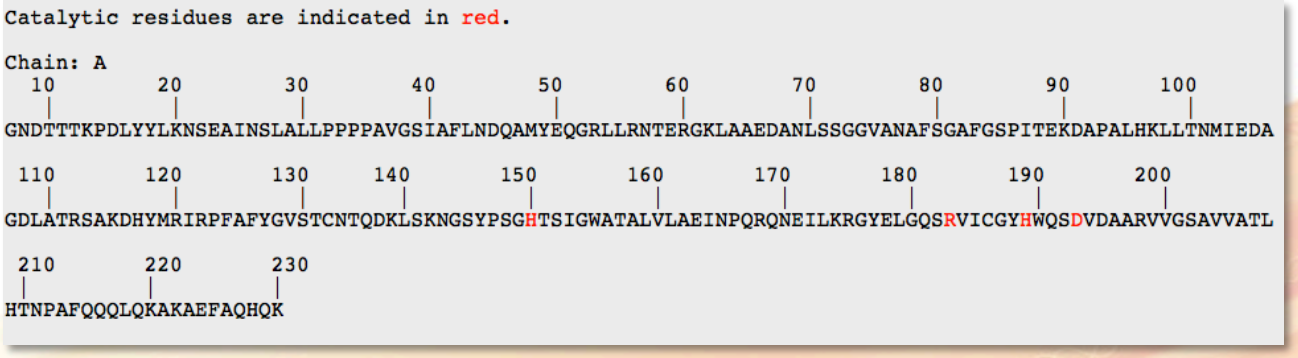
- Get the *B. licheniformis* ribosomal S2 protein sequence (one of the top hits)
- Or, you can just paste in the reference number: YP_080329.1
- Now do a BLAST2Seq
  - Results: same as with regular BLAST
    - Also has Dot matrix view: click it
  - Full length alignment, with 2 small gaps



- Blast 2 Sequences is a great tool!

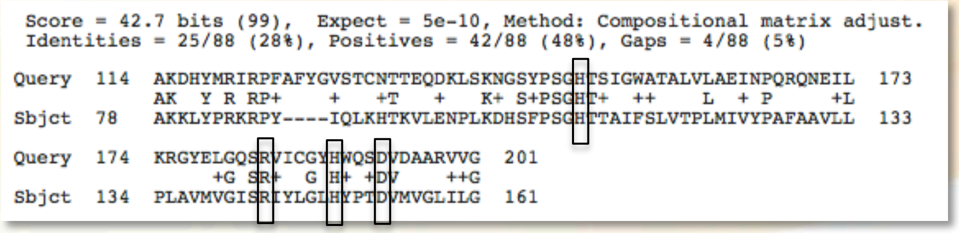
Getting Catalytic Residues

- 1d2tA - Acid phosphatase
  - http://www.ncbi.nlm.nih.gov/protein/1D2T_A
  - http://www.ebi.ac.uk/thornton-srv/databases/CSA/



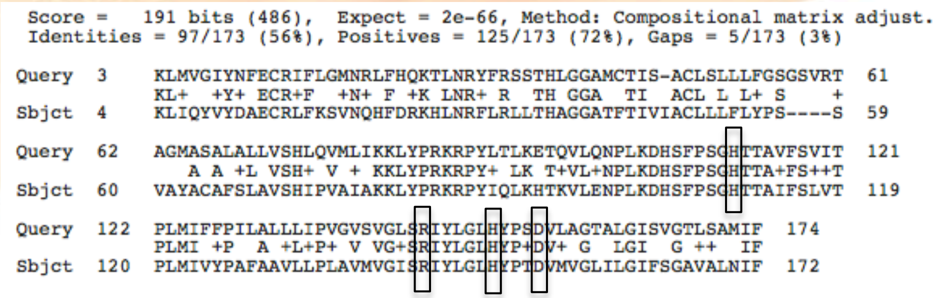
```
Catalytic residues are indicated in red.

Chain: A
  10        20        30        40        50        60        70        80        90       100
   |         |         |         |         |         |         |         |         |         |
GNDTTTKPDLYYLKNSEAINSLALLPPPPAVGSIAFLNDQAMYEQGRLLRNTERGKLAAEDANLSSGGVANAFSGAFGSPITEKDAPALHKLLTNMIEDA

 110       120       130       140       150       160       170       180       190       200
   |         |         |         |         |         |         |         |         |         |
GDLATRSAKDHYMRIRPFAFYGVSTCNTQDKLSKNGSYPSGHTSIGWATALVLAEINPQRQNEILKRGYELGQSRVICGYHWQSDVDAARVVGSAVVATL

 210       220       230
 |         |         |
HTNPAFQQQLQKAKAEFAQHQK
```

YP_003565295 vs 1d2t_A



```
 Score = 42.7 bits (99),  Expect = 5e-10, Method: Compositional matrix adjust.
 Identities = 25/88 (28%), Positives = 42/88 (48%), Gaps = 4/88 (5%)

Query  114  AKDHYMRIRPFAFYGVSTCNTTEQDKLSKNGSYPSGHTSIGWATALVLAEINPQRQNEIL  173
            AK  Y R RP+    +   +T   +   K+ S+PSGHT+  ++    L  + P     +L
Sbjct  78   AKKLYPRKRPY----IQLKHTKVLENPLKDHSFPSGHTTAIFSLVTPLMIVYPAFAAVLL  133

Query  174  KRGYELGQSRVICGYHWQSDVDAARVVG  201
                 +G SR+  G H+ +DV    ++G
Sbjct  134  PLAVMVGISRIYLGLHYPTDVMVGLILG  161
```

YP_003565295 vs YP_080329.1



```
 Score =  191 bits (486),  Expect = 2e-66, Method: Compositional matrix adjust.
 Identities = 97/173 (56%), Positives = 125/173 (72%), Gaps = 5/173 (3%)

Query  3    KLMVGIYNFECRIFLGMNRLFHQKTLNRYFRSSTHLGGAMCTIS-ACLSLLLFGSGSVRT  61
            KL+  +Y+ ECR+F  +N+ F +K LNR+ R  TH GGA  TI  ACL L L+ S    +
Sbjct  4    KLIQYVYDAECRLFKSVNQHFDRKHLNRFLRLLTHAGGATFTIVIACLLLFLYPS----S  59

Query  62   AGMASALALLVSHLQVMLIKKLYPRKRPYLTLKETQVLQNPLKDHSFPSGHTTAVFSVIT  121
               A A +L VSH+ V + KKLYPRKRPY+ LK T+VL+NPLKDHSFPSGHTTA+FS++T
Sbjct  60   VAYACAFSLAVSHIPVAIAKKLYPRKRPYIQLKHTKVLENPLKDHSFPSGHTTAIFSLVT  119

Query  122  PLMIFFPILALLLIPVGVSVGLSRIYLGLHYPSDVLAGTALGISVGTLSAMIF  174
            PLMI +P  A +L+P+ V VG+SRIYLGLHYP+DV+ G  LGI  G ++  IF
Sbjct  120  PLMIVYPAFAAVLLPLAVMVGISRIYLGLHYPTDVMVGLILGIFSGAVALNIF  172
```

More Gene Names

- Do the names of the top BLAST hits agree with each other?
- They should:
  - But there are always annotation errors, and our knowledge of gene function increases over time
  - With some sloppiness due to different naming conventions practiced by different scientists
- Here we have a classic case of misnaming

More Gene Names

- Why is the one of the top hits a ribosomal protein S2?
  - Ribosomal proteins are **highly conserved in evolution**
  - Query: Ribosomal protein S2 AND Bacillus licheniformis[ORGN]
  - No homology exists between:
    - [YP_080329.1] and the actual ribosomal protein S2 found in Bacillius licheniformis [YP_091458]
- The other names are similar - although not identical
  - What is “PAP2”?
    - A quick Google search shows that it stands for “phosphatidic acid phosphatase”
      - fits the other names well
  - There is probably some uncertainty about its exact function, given the variety of names and the “family protein” designation in several of them

Remember it’s up to you the bioinformatician to scrutinize these alignments!

Standalone BLAST

BLAST result

```
BLASTP 2.2.18 [Mar-02-2008]


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= metL gi|16131778|ref|NP_418375.1| aspartokinase II and
homoserine dehydrogenase II; bifunctional: aspartokinase II
(N-terminal); homoserine dehydrogenase II (C-terminal) [Escherichia
coli K12]
         (810 letters)

Database: /Users/jvanheld/rsa-
tools/data/genomes/Escherichia_coli_K12/genome/NC_000913.faa
           4242 sequences; 1,351,322 total letters

Searching.........done

                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gi|16131778|ref|NP_418375.1| aspartokinase II and homoserine deh...  1596   0.0
gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I (N-te...   344   2e-95
gi|16131850|ref|NP_418448.1| aspartokinase III, lysine sensitive...   122   7e-29
gi|16128228|ref|NP_414777.1| gamma-glutamate kinase [Escherichia...    31   0.28

>gi|16131778|ref|NP_418375.1| aspartokinase II and homoserine
           dehydrogenase II; bifunctional: aspartokinase II
           (N-terminal); homoserine dehydrogenase II (C-terminal)
           [Escherichia coli K12]
          Length = 810

 Score = 1596 bits (4132), Expect = 0.0
 Identities = 810/810 (100%), Positives = 810/810 (100%)
```

- Query: *E.coli* protein MetL, a bifunctional enzyme combining aspartokinase and homoserine dehydrogenase activities.
- Database: all proteins from *Escherichia coli K12.*
- BLAST result file starts with a summary of
  - parameters used for the search
  - The matching sequences and the score of each match.

Jacques van Helden

BLAST Result – First Match

```
>gi|16131778|ref|NP_418375.1| aspartokinase II and homoserine
           dehydrogenase II; bifunctional: aspartokinase II
           (N-terminal); homoserine dehydrogenase II (C-terminal)
           [Escherichia coli K12]
          Length = 810

 Score = 1596 bits (4132), Expect = 0.0
 Identities = 810/810 (100%), Positives = 810/810 (100%)

Query: 1   MSVIAQAGAKGRQLHKFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINW 60
            MSVIAQAGAKGRQLHKFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINW
Sbjct: 1   MSVIAQAGAKGRQLHKFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINW 60

Query: 61  LKLSQTDRLSAHQVQQTLRRYQCDLISGLLPAEEADSLISAFVSDLERLAALLDSGINDA 120
            LKLSQTDRLSAHQVQQTLRRYQCDLISGLLPAEEADSLISAFVSDLERLAALLDSGINDA
Sbjct: 61  LKLSQTDRLSAHQVQQTLRRYQCDLISGLLPAEEADSLISAFVSDLERLAALLDSGINDA 120

Query: 121 VYAEVVGHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAERAAQPQVDEGLSYPLLQQLL 180
            VYAEVVGHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAERAAQPQVDEGLSYPLLQQLL
Sbjct: 121 VYAEVVGHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAERAAQPQVDEGLSYPLLQQLL 180

Query: 181 VQHPGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADP 240
            VQHPGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADP
Sbjct: 181 VQHPGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADP 240

Query: 241 RKVKDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQGSTRIER 300
            RKVKDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQGSTRIER
Sbjct: 241 RKVKDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQGSTRIER 300

Query: 301 VLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQLL 360
            VLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQLL
Sbjct: 301 VLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQLL 360
```

- First match is the query sequence itself (metL)
- Not surprising since we scanned the set of all E.coli proteins with a protein from E.coli
- The E-value (0) means that, with this level of similarity; one would expect 0 false positive by chance

Jacques van Helden

BLAST Result – Second Match

```
>gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I
           (N-terminal); homoserine dehydrogenase I (C-terminal)
           [Escherichia coli K12]
          Length = 820

 Score =  344 bits (882), Expect = 2e-95
 Identities = 247/821 (30%), Positives = 410/821 (49%), Gaps = 44/821 (5%)

Query: 16  KFGGSSLADVKCYLRVAGIMAEYSQPDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74
            KFGG+S+A+ + +LRVA I+   ++   +  V+SA    TN L+  ++ + + + +   +
Sbjct: 5   KFGGTSVANAERFLRVADILESNARQGQVATVLSAPAKITNHLVAMIEKTISGQDALPNI 64

Query: 75  QQTLRRYQCDLISGLLPAEEADSL--ISAFVSDLERLAALLDSGIN------DAVYAEVV 126
                R +  +L++GL  A+    L  +  FV         +  GI+      D++ A ++
Sbjct: 65  SDAERIF-AELLTGLAAAQPGFPLAQLKTFVDQEFAQIKHVLHGISLLGQCPDSINAALI 123

Query: 127 GHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183
              GE  S  +M+ VL  +G     +D  E L A      +   + E        ++   H
Sbjct: 124 CRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183

Query: 184 PGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243
                +++ GF + N  GE V+LGRNGSDYSA  + A        IW+DV GVY+ DPR+V
Sbjct: 184 ---MVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQV 240

Query: 244 KDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ-----GSTRI 298
             DA LL  +   EA EL+   A VLH RT+ P++  +I   ++ +  P       G++R
Sbjct: 241 PDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300

Query: 299 ERVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQ 358
            E  L      + +++ +++ +     P  +        +   + RA++  + +   +
Sbjct: 301 EDELP----VKGISNLNNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLITQSSSEY 356

Query: 359 LLQFCYTSEVADSALKILDEA-------GLPGELRLRQGLALVAMVGAGVTRNPLHCHRF 411
             + FC        A + + E        GL   L + + LA++++VG G+        +F
Sbjct: 357 SISFCVPQSDCVRAERAMQEEFYLELKEGLLEPLAVTERLAIISVVGDGMRTLRGISAKF 416
```

- Second match is another bifunctional protein, product of the gene thrA
- Contains the same two domains as metA (aspartokinase and homoserine dehydrogenase).
- Alignment covers almost the complete sequences (820 aa), with 30% identities and 49% similarity.
- E-value is very low (2e-95), indicating that thrA and metL are likely to be true homologs.

Jacques van Helden

BLAST Result – Third Match

```
>gi|16131850|ref|NP_418448.1| aspartokinase III, lysine sensitive;
           aspartokinase III, lysine-sensitive [Escherichia coli
           K12]
          Length = 449

 Score =  122 bits (307), Expect = 7e-29
 Identities = 121/452 (26%), Positives = 194/452 (42%), Gaps = 25/452 (5%)

Query: 16  KFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINWLK-LSQTDRLSAHQV 74
           KFGG+S+AD     R A I+    +   ++V+SA+   TN L+   + L   +R    +
Sbjct: 8   KFGGTSVADFDAMNRSADIVLSDANVR-LVVLSASAGITNLLVALAEGLEPGERF---EK 63

Query: 75  QQTLRRYQCDLISGLLPAEEADSLISAFVSDLERLAALLDSGINDAVYAEVVGHGEVWSA 134
              +R  Q  ++  L         I   + ++  LA       + A+  E+V HGE+ S
Sbjct: 64  LDAIRNIQFAILERLRYPNVIREEIERLLENITVLAEAAALATSPALTDELVSHGELMST 123

Query: 135 RLMSAVLNQQGLPAAWLDAREFLRA-ERAAQPQVDEGLSYPLLQQLLVQHPGKRLVVT-G 192
            L   +L ++ + A W D R+ +R  +R   + + D      L    L+    + LV+T G
Sbjct: 124 LLFVEILRERDVQAQWFDVRKVMRTNDRFGRAEPDIAALAELAALQLLPRLNEGLVITQG 183

Query: 193 FISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKVKDACLLPLL 252
           FI   N G T  LGR GSDY+A  +      SRV IW+DV G+Y+ DPR V  A  +  +
Sbjct: 184 FIGSENKGRTTTLGRGGSDYTAALLAEALHASRVDIWTDVPGIYTTDPRVVSAAKRIDEI 243

Query: 253 RLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQGSTRI---------ERVLA 303
              EA+E+A   A VLH  TL P   S+I + +  S  P  G T +          R LA
Sbjct: 244 AFAEAAEMATFGAKVLHPATLLPAVRSDIPVFVGSSKDPRAGGTLVCNKTENPPLFRALA 303

Query: 304 SGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQLLQFC 363
                 ++T H    L      A     LA   I   L       +A+        L
Sbjct: 304 LRRNQTLLTLHSLNMLHSRGFLAEVFGILARHNISVDLITTSEVSVAL-------TLDTT 356

Query: 364 YTSEVADSAL--KILDEAGLPGELRLRQGLALVAMVGAGVTRNPLHCHRFWQQLKGQPVE 421
            ++   D+ L   +L E      + + +GLALVA++G  +++        +  L+   +
Sbjct: 357 GSTSTGDTLLTQSLLMELSALCRVEVEEGLALVALIGNDLSKACGVGKEVFGVLEPFNIR 416
```

- Third match is the product of the gene lysC: aspartokinase III
- Protein contains the aspartokinase domain, but not the homoserine dehydrogenase.
- Alignment only extends over the first half of the query protein (453aa)
- On this segment, there is a good level of identity (26%) and similarity (42%)
- E-value is very low (7e–29), indicating that the two domains are likely to be true homologs.

BLAST Result – Fourth Match

```
>gi|16128228|ref|NP_414777.1| gamma-glutamate kinase [Escherichia
           coli K12]
          Length = 367

 Score = 31.2 bits (69), Expect = 0.28
 Identities = 17/56 (30%), Positives = 29/56 (51%)

Query: 194 ISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKVKDACLL 249
           I+ N+A  T  +    +D  +     LAG  ++ + +D  G+Y+ADPR    A L+
Sbjct: 133 INENDAVATAEIKVGDNDNLSALAAILAGADKLLLLTDQKGLYTADPRSNPQAELI 188
```

- Fourth match is a gamma-glutamate kinase, product of proB
- Same level of identity (30%) and similarity (51%) as the second match (thrA)
- Match only extends over 56aa, whereas the alignment between thrA and metL extends over 821aa
- E-value is quite high (0.28) suggesting that the similarity could be an artefact.
- **This clearly illustrates the fact that the important parameter to evaluate the significance of an alignment is the E-value, not the percentage of similarity !**

BLAST Result – Summary

```
 Database: /Users/jvanheld/rsa-
  tools/data/genomes/Escherichia_coli_K12/genome/NC_000913.faa
    Posted date:  Sep 8, 2004 12:13 PM
  Number of letters in database: 1,351,322
  Number of sequences in database:  4242

Lambda     K      H
   0.320    0.136    0.397

Gapped
Lambda     K      H
   0.267    0.0410    0.140

Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 2,199,628
Number of Sequences: 4242
Number of extensions: 96525
Number of successful extensions: 290
Number of sequences better than  1.0: 4
Number of HSP's better than  1.0 without gapping: 4
Number of HSP's successfully gapped in prelim test: 0
Number of HSP's that attempted gapping in prelim test: 279
Number of HSP's gapped (non-prelim): 5
length of query: 810
length of database: 1,351,322
effective HSP length: 92
effective length of query: 718
effective length of database: 961,058
effective search space: 690039644
effective search space used: 690039644
T: 11
A: 40
X1: 16 ( 7.4 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (21.8 bits)
S2: 65 (29.6 bits)
```

- The last part of the BLAST result gives some statistics about the search:
  - Number of hits
  - Number of sequences in the DB
  - …

ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid57779/NC_000913.faa

BLAST Running Time

- Running Time

The length of Query : 153

DB size : 5997 sequences

Algorithm	Running Time
D.P	16.989 [s]
FASTA	0.618 [s]
BLAST	0.118 [s]

PC : Pentium 4

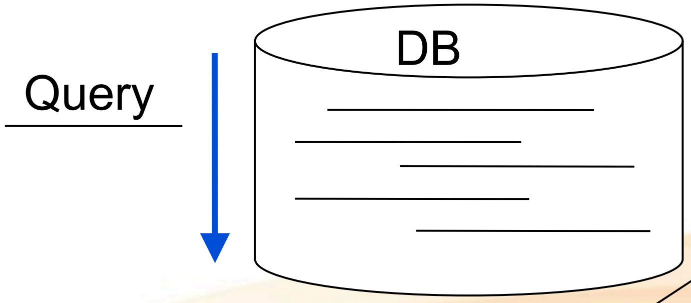
By Dr. Takeshi Kawabata

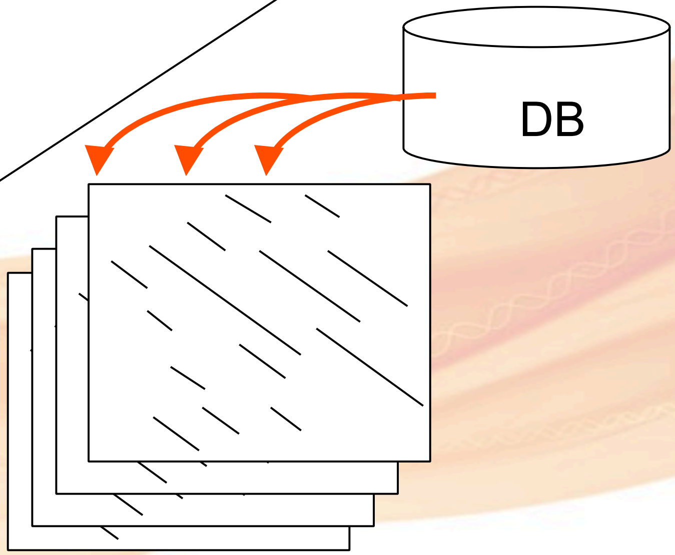
Nara Sentan Gijyutu University

FASTA vs BLAST

BLAST

Compare the query and sequences in DB with the same threshold.





FASTA

Compare the query and a sequence one by one
And compare each result

Conclusion

Algorithm	Sensitivity	Running Time
D.P	1	3
FASTA	2 better DNA	2
BLAST	2 better proteins	1

Comparison of Algorithm

- Dynamic Programming
  - Most sensitive result
    - D.P uses **all information** of two sequence
  - Running time is slow
    - D.P compute **the useless area** for computing the optimal sequence

Comparison of Algorithm

- FASTA
  - Less sensitive than D.P and BLAST (protein search)
  - FASTA uses **partial information** to **speed up** the computation
- Running time is faster D.P
  - The same reason as the above

Comparison of Algorithms

- BLAST
  - More sensitive than FASTA (protein search)
- Faster than FASTA
  - Because BLAST evaluates the entire DB with the same threshold based on statistics
  - BLAST **eliminate noises** and **reduces** the running time

Using BLAST to Get Quick Answers to Bioinformatics Problems

Task	BLAST method	Trad. Method
Predict protein function (1)	Perform BLASTp on PIR or Swiss-Prot database	Perform wet-lab experiment
Predict protein function (2)	Perform tBLASTn on NR database	Perform wet-lab experiment
Predict protein structure	Perform BLASTp against PDB	Structure prediction software, x-ray crystall., NMR

Using BLAST to Get Quick Answers to Bioinformatics Problems

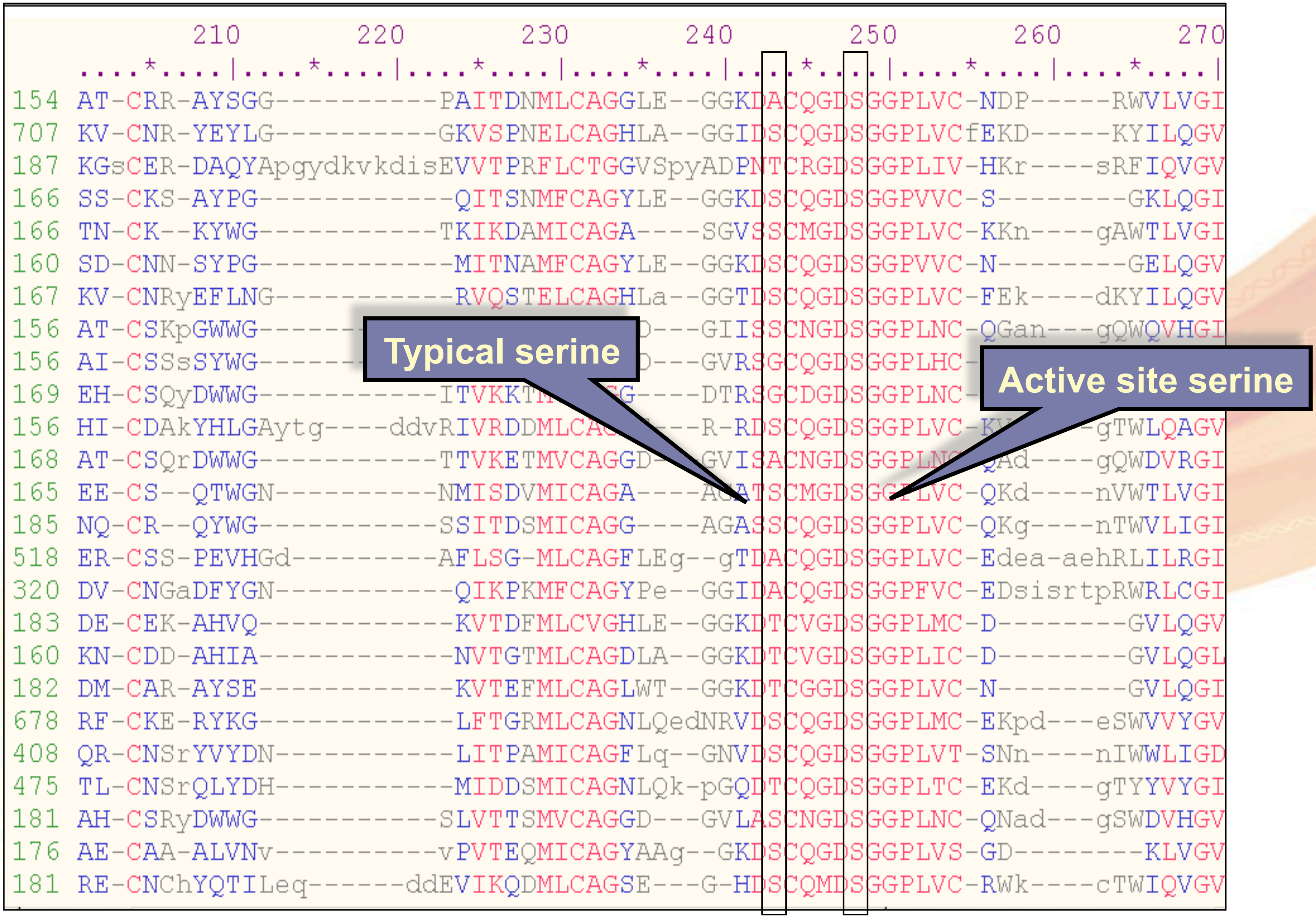
Task	BLAST method	Trad. Method
Locate genes in a genome	Divide genome into 2-5 kb sequences. Perform BLASTx against NR protein datbase	Run gene prediction software. Perform microarray/RNAseq analysis
Find distantly related proteins	Perform psi-BLAST	No traditional method
Identify DNA sequence	Perform BLASTn	Screen genomic DNA library

PSI-BLAST

PSI-position specific iterative BLAST – Used for
Finding new family members that do not hit the original query

What do you think we could use as a model?

Position Specific Substitution Rates

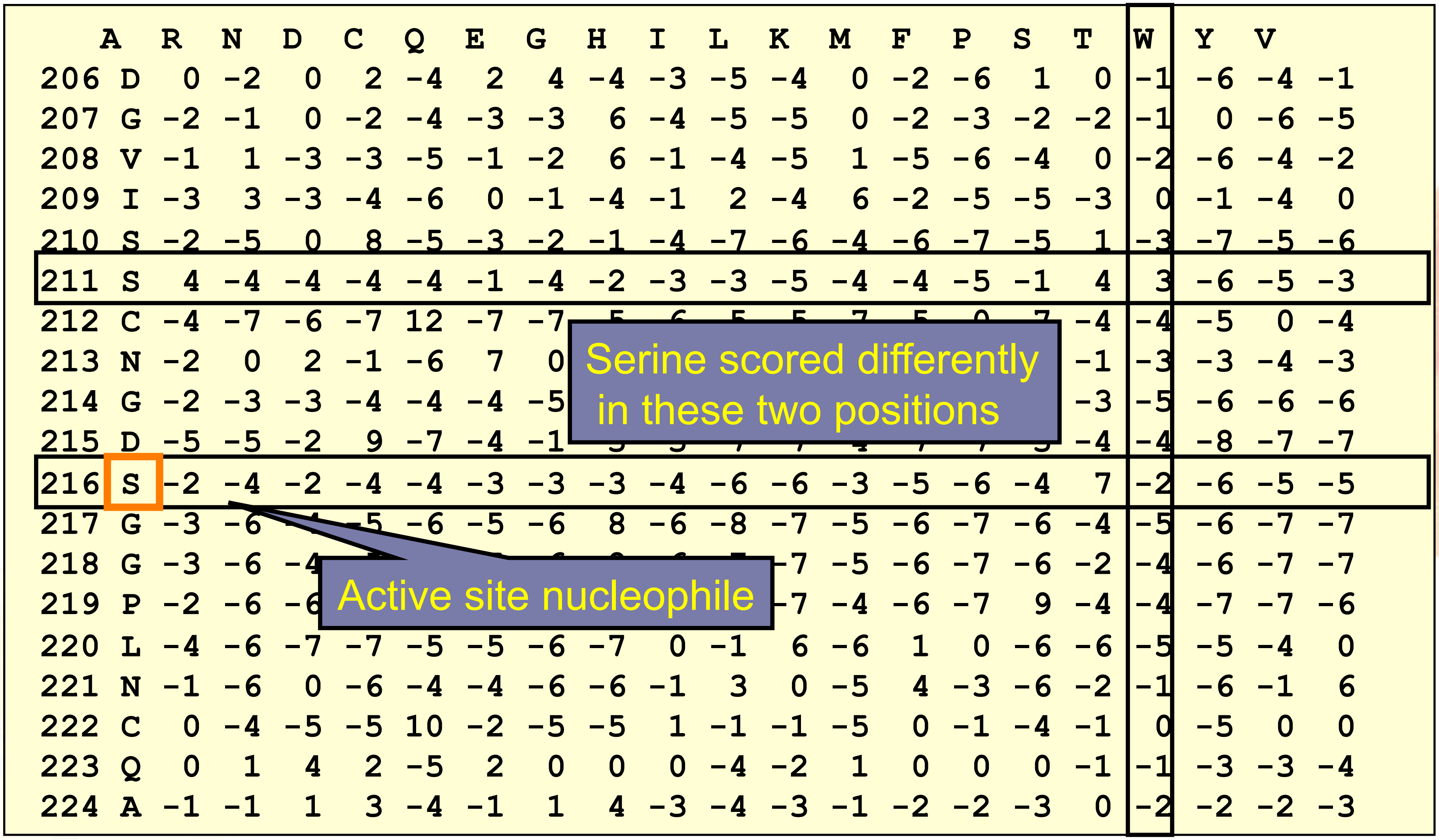


```
           210       220       230       240       250       260       270
    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|
154 AT-CRR-AYSGG----------PAITDNMLCAGGLE--GGKDACQGDSGGPLVC-NDP-----RWVLVGI
707 KV-CNR-YEYLG----------GKVSPNELCAGHLA--GGIDSCQGDSGGPLVCfEKD-----KYILQGV
187 KGsCER-DAQYApgydkvkdisEVVTPRFLCTGGVSpyADPNTCRGDSGGPLIV-HKr----sRFIQVGV
166 SS-CKS-AYPG------------QITSNMFCAGYLE--GGKDSCQGDSGGPVVC-S--------GKLQGI
166 TN-CK--KYWG-----------TKIKDAMICAGA----SGVSSCMGDSGGPLVC-KKn----gAWTLVGI
160 SD-CNN-SYPG------------MITNAMFCAGYLE--GGKDSCQGDSGGPVVC-N--------GELQGV
167 KV-CNRyEFLNG-----------RVQSTELCAGHLa--GGTDSCQGDSGGPLVC-FEk----dKYILQGV
156 AT-CSKpGWWG------                 D---GIISSCNGDSGGPLNC-QGan---gQWQVHGI
156 AI-CSSsSYWG------                 D---GVRSGCQGDSGGPLHC-
169 EH-CSQyDWWG-----------ITVKKT     G----DTRSGCDGDSGGPLNC-
156 HI-CDAkYHLGAytg----ddvRIVRDDMLCA   ---R-RDSCQGDSGGPLVC-K     -gTWLQAGV
168 AT-CSQrDWWG-----------TTVKETMVCAGGD-  GVISACNGDSGGPLN  QAd----gQWDVRGI
165 EE-CS--QTWGN----------NMISDVMICAGA----A ATSCMGDSGG LVC-QKd----nVWTLVGI
185 NQ-CR--QYWG-----------SSITDSMICAGG----AGASSCQGDSGGPLVC-QKg----nTWVLIGI
518 ER-CSS-PEVHGd---------AFLSG-MLCAGFLEg--gTDACQGDSGGPLVC-Edea-aehRLILRGI
320 DV-CNGaDFYGN-----------QIKPKMFCAGYPe--GGIDACQGDSGGPFVC-EDsisrtpRWRLCGI
183 DE-CEK-AHVQ------------KVTDFMLCVGHLE--GGKDTCVGDSGGPLMC-D--------GVLQGV
160 KN-CDD-AHIA------------NVTGTMLCAGDLA--GGKDTCVGDSGGPLIC-D--------GVLQGL
182 DM-CAR-AYSE------------KVTEFMLCAGLWT--GGKDTCGGDSGGPLVC-N--------GVLQGI
678 RF-CKE-RYKG------------LFTGRMLCAGNLQedNRVDSCQGDSGGPLMC-EKpd---eSWVVYGV
408 QR-CNSrYVYDN-----------LITPAMICAGFLq--GNVDSCQGDSGGPLVT-SNn----nIWWLIGD
475 TL-CNSrQLYDH-----------MIDDSMICAGNLQk-pGQDTCQGDSGGPLTC-EKd----gTYYVYGI
181 AH-CSRyDWWG-----------SLVTTSMVCAGGD---GVLASCNGDSGGPLNC-QNad---gSWDVHGV
176 AE-CAA-ALVNv----------vPVTEQMICAGYAAg--GKDSCQGDSGGPLVS-GD--------KLVGV
181 RE-CNChYQTILeq------ddEVIKQDMLCAGSE---G-HDSCQMDSGGPLVC-RWk----cTWIQVGV
```

Typical serine

Active site serine

Position Specific Score Matrix w/ BLAST



		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
206	D	0	-2	0	2	-4	2	4	-4	-3	-5	-4	0	-2	-6	1	0	-1	-6	-4	-1
207	G	-2	-1	0	-2	-4	-3	-3	6	-4	-5	-5	0	-2	-3	-2	-2	-1	0	-6	-5
208	V	-1	1	-3	-3	-5	-1	-2	6	-1	-4	-5	1	-5	-6	-4	0	-2	-6	-4	-2
209	I	-3	3	-3	-4	-6	0	-1	-4	-1	2	-4	6	-2	-5	-5	-3	0	-1	-4	0
210	S	-2	-5	0	8	-5	-3	-2	-1	-4	-7	-6	-4	-6	-7	-5	1	-3	-7	-5	-6
211	S	4	-4	-4	-4	-4	-1	-4	-2	-3	-3	-5	-4	-4	-5	-1	4	3	-6	-5	-3
212	C	-4	-7	-6	-7	12	-7	-7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-4	-4	-5	0	-4
213	N	-2	0	2	-1	-6	7	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-1	-3	-3	-4	-3
214	G	-2	-3	-3	-4	-4	-4	-5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-3	-5	-6	-6	-6
215	D	-5	-5	-2	9	-7	-4	-1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-4	-4	-8	-7	-7
216	S	-2	-4	-2	-4	-4	-3	-3	-3	-4	-6	-6	-3	-5	-6	-4	7	-2	-6	-5	-5
217	G	-3	-6	-4	-5	-6	-5	-6	8	-6	-8	-7	-5	-6	-7	-6	-4	-5	-6	-7	-7
218	G	-3	-6	-4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-7	-5	-6	-7	-6	-2	-4	-6	-7	-7
219	P	-2	-6	-6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-7	-4	-6	-7	9	-4	-4	-7	-7	-6
220	L	-4	-6	-7	-7	-5	-5	-6	-7	0	-1	6	-6	1	0	-6	-6	-5	-5	-4	0
221	N	-1	-6	0	-6	-4	-4	-6	-6	-1	3	0	-5	4	-3	-6	-2	-1	-6	-1	6
222	C	0	-4	-5	-5	10	-2	-5	-5	1	-1	-1	-5	0	-1	-4	-1	0	-5	0	0
223	Q	0	1	4	2	-5	2	0	0	0	-4	-2	1	0	0	0	-1	-1	-3	-3	-4
224	A	-1	-1	1	3	-4	-1	1	4	-3	-4	-3	-1	-2	-2	-3	0	-2	-2	-2	-3

Serine scored differently in these two positions

Active site nucleophile

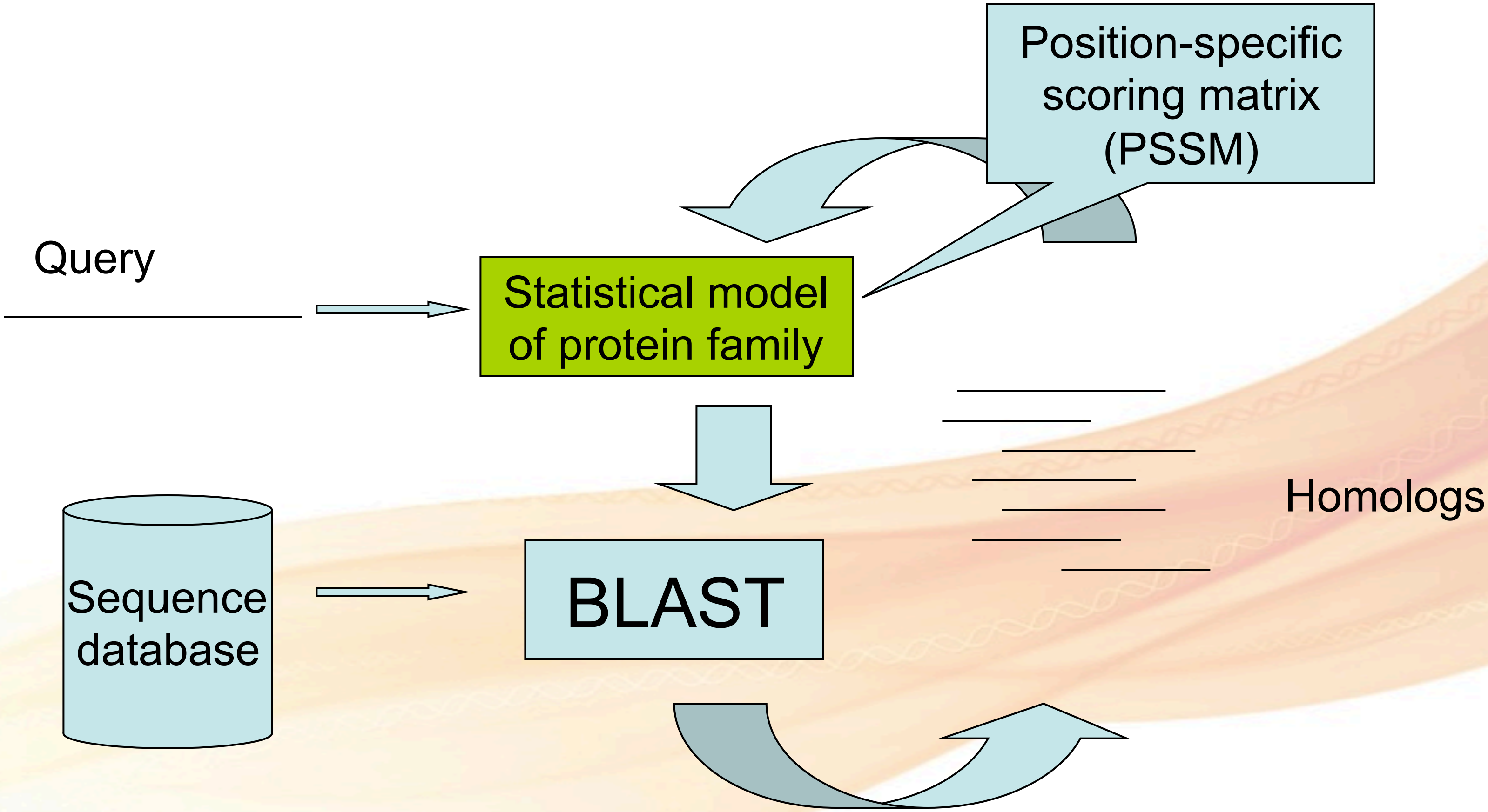
Query Sequence - DGVIS**S**CNGD**S**GGPLNCAQ

Taken from PSI BLAST

PSI-BLAST

- An iterative process:
  - DB searched with an initial query sequence
  - All hits with E-value lower cutoff (default = 0.005) are kept
  - A **PSSM** is constructed automatically from multiple HSPs of initial BLAST search
    - PSSM is then used to search the database again
    - A lower E-value threshold is now used (E=.001)
    - If new sequences with lower e-value than cutoff found - PSSM is updated to include them, and search is run again
  - Eventually, no new sequences are found and the PSI-BLAST search is complete (**converged**)
- Result
  - 1) Obtains **distantly** related sequences
  - 2) Can find the **important residues** that provide function or structure
  - 3) Extensions of the matches, why's this?

Position-specific iterated BLAST



Position-specific scoring matrix (PSSM)

Query

Statistical model of protein family

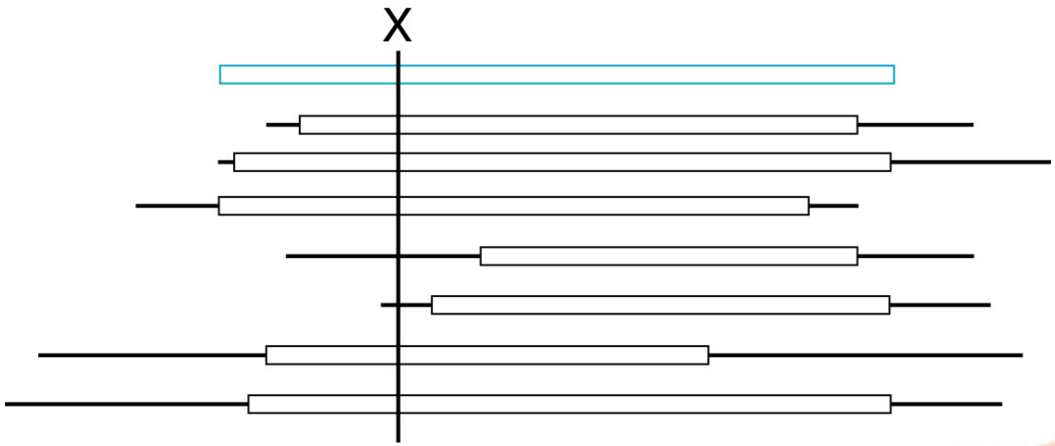
Sequence database

BLAST

Homologs

How PSI-BLAST Generates PSSMs

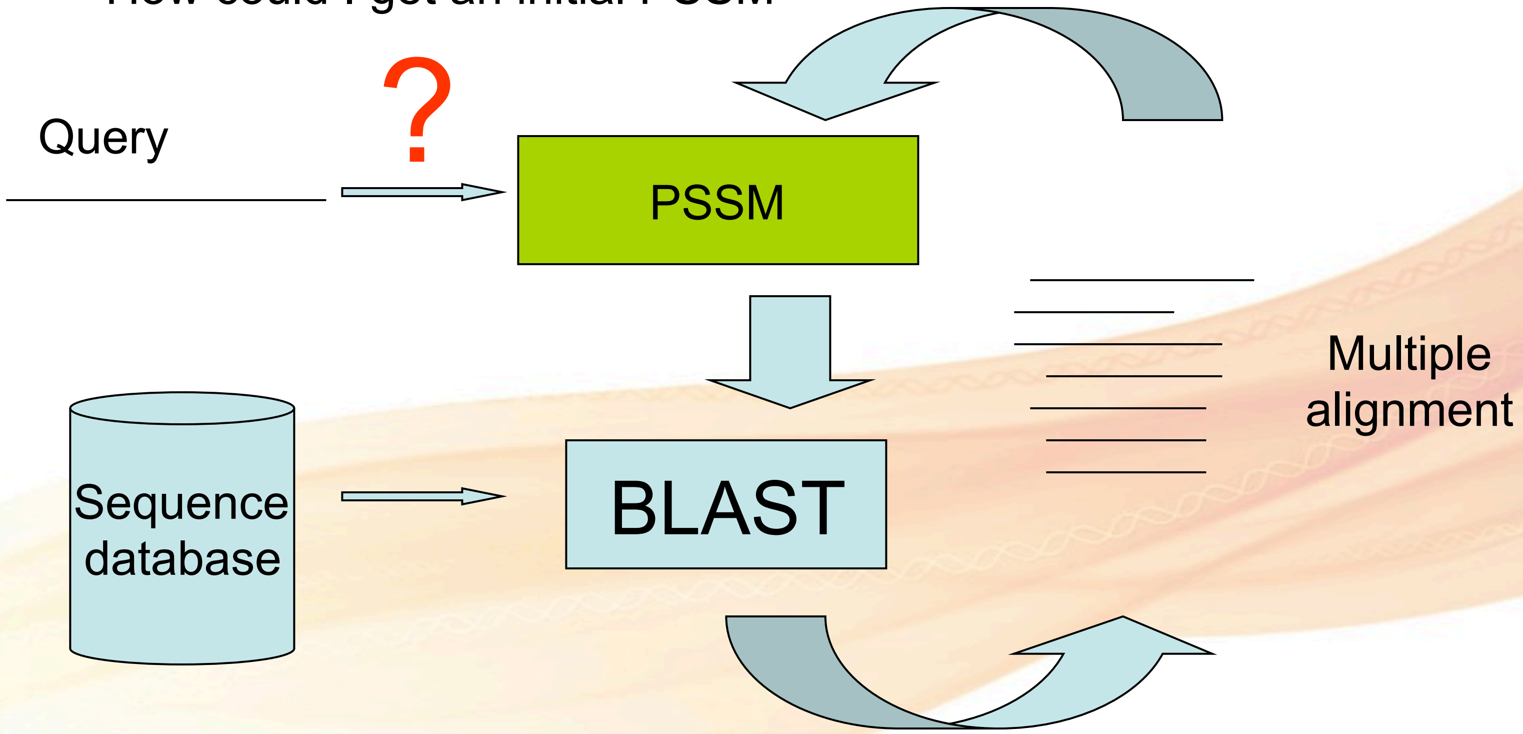
What Blast uses to Construct PSSM



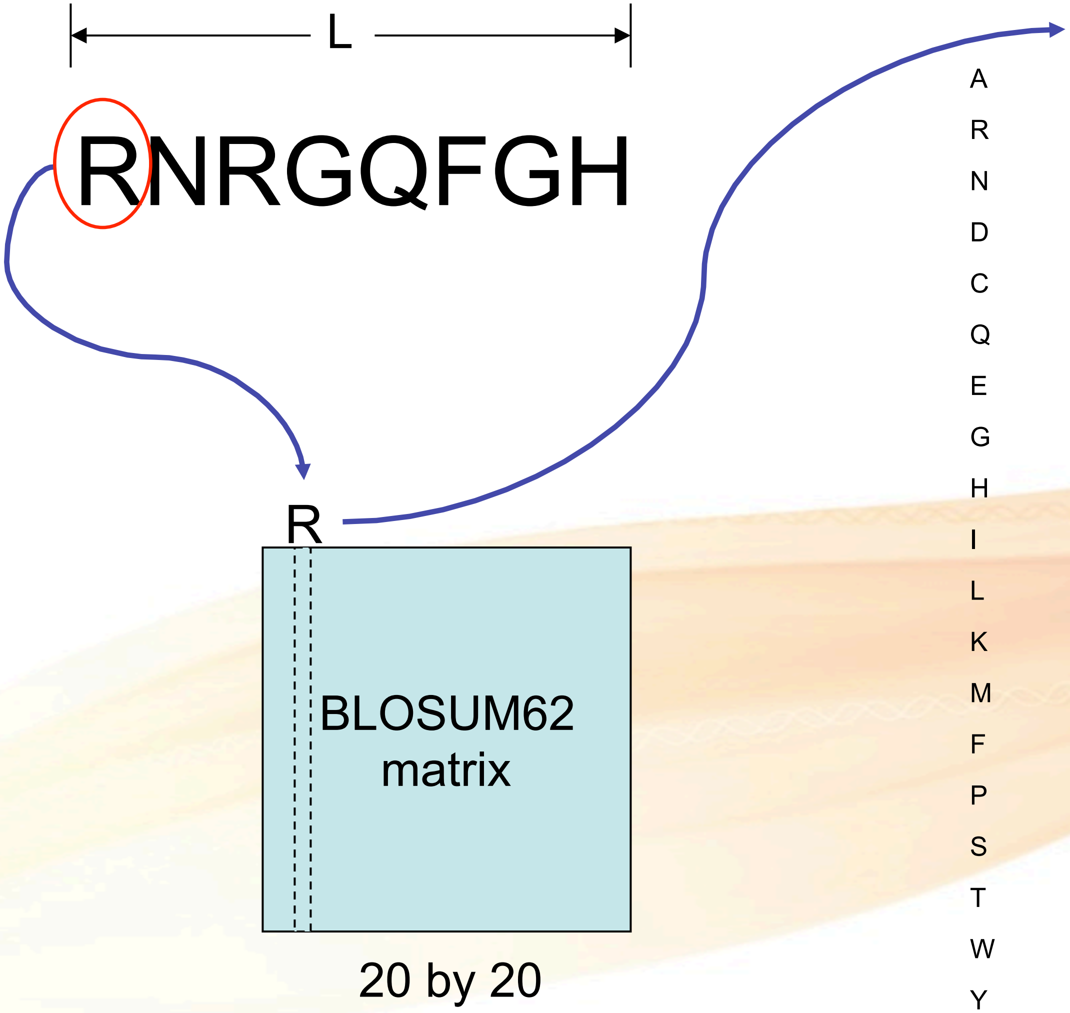
- **PSI-BLAST uses the pairwise sequence alignments from a database search in PSSM construction**
  - The BLAST local alignments to qu**ery sequence** (blue sequence) are shown as rectangles
  - At each **residue position** of the query, a PSSM is constructed using only those sequences whose BLAST alignments involve that position
  - Here, at position X only six residues are considered in order to derive the PSSM, as only six of the sequences (including the query sequence) have local alignments including this position

PSI-BLAST

How could I get an initial PSSM



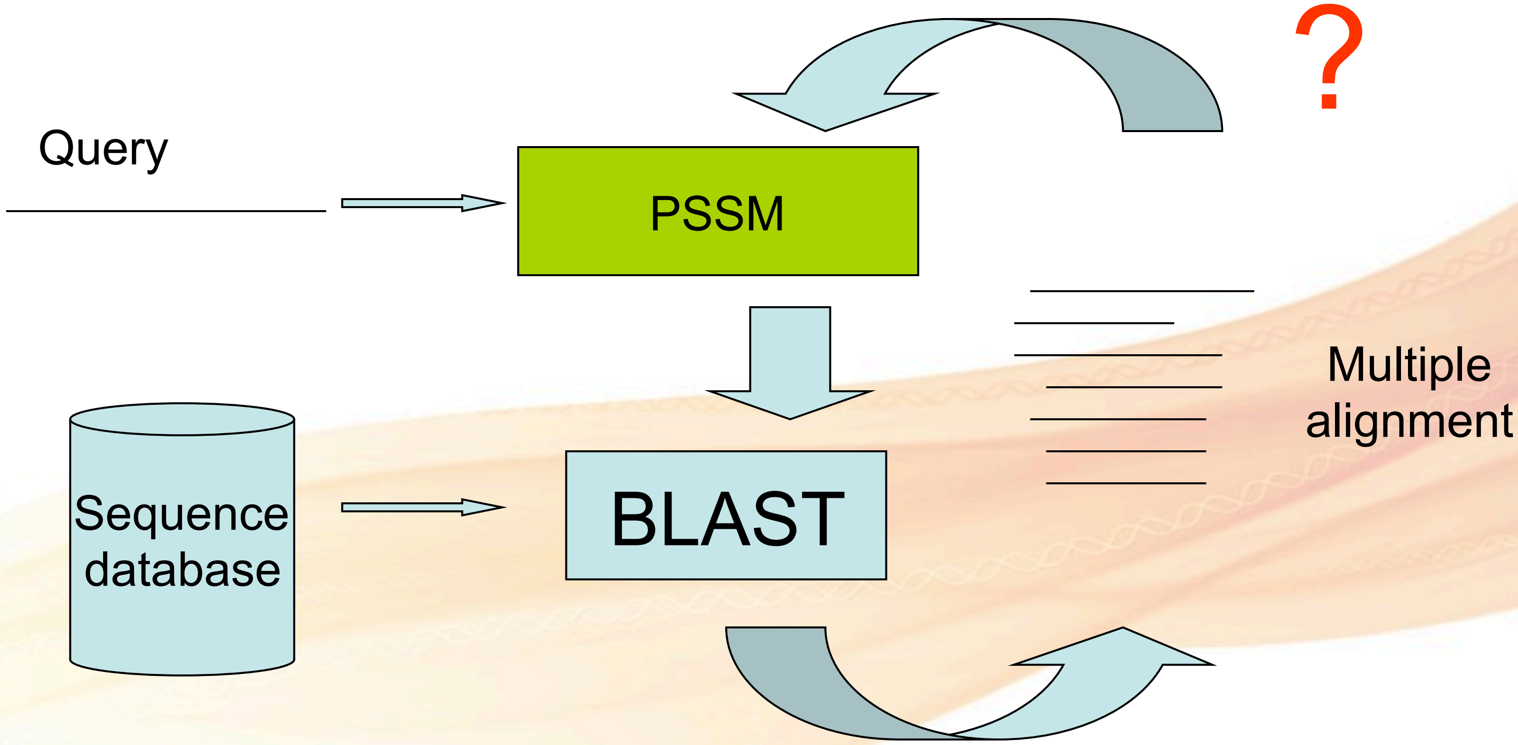
Creating a PSSM from 1 sequence



	R							
A	-1	-2	-1	0	-1	-2	0	-2
R	5	0	5	-2	1	-3	-2	0
N	0	6	0	0	0	-3	0	1
D	-2	1	-2	-1	0	-3	-1	-1
C	-3	-3	-3	-3	-3	-2	-3	-3
Q	1	0	1	-2	5	-3	-2	0
E	0	0	0	-2	2	-3	-2	0
G	-2	0	-2	6	-2	-3	6	-2
H	0	1	0	-2	0	-1	-2	8
I	-3	-3	-3	-4	-3	0	-4	-3
L	-2	-3	-2	-4	-2	0	-4	-3
K	2	0	2	-2	1	-3	-2	-1
M	-1	-2	-1	-3	0	0	-3	-2
F	-3	-3	-3	-3	-3	6	-3	-1
P	-2	-2	-2	-2	-1	-4	-2	-2
S	-1	1	-1	0	0	-2	0	-1
T	-1	0	-1	-2	-1	-2	-2	-2
W	-3	-4	-3	-2	-2	1	-2	-2
Y	-2	-2	-2	-3	-1	3	-3	2
V	-3	-3	-3	-3	-2	-1	-3	-3

20 by L

PSI-BLAST



Creating a PSSM from Multiple Sequences

- Discard columns that contain gaps in the query
- For each column C
  - Compute relative sequence weights
  - Compute PSSM entries, taking into account
    - Observed residues in this column
    - Sequence weights
    - Substitution matrix

Discard Query Gap Columns

```
EEFG----SVDGLVNNA
QKYG----RLDVMINNA
RRLG----TLNVLVNNA
GGIG----PVD-LVNNA
KALG----GFNVIVNNA
ARFG----KID-LIPNA
FEPEGPEKGMWGLVNNA
AQLK----TVDVLINGA
```

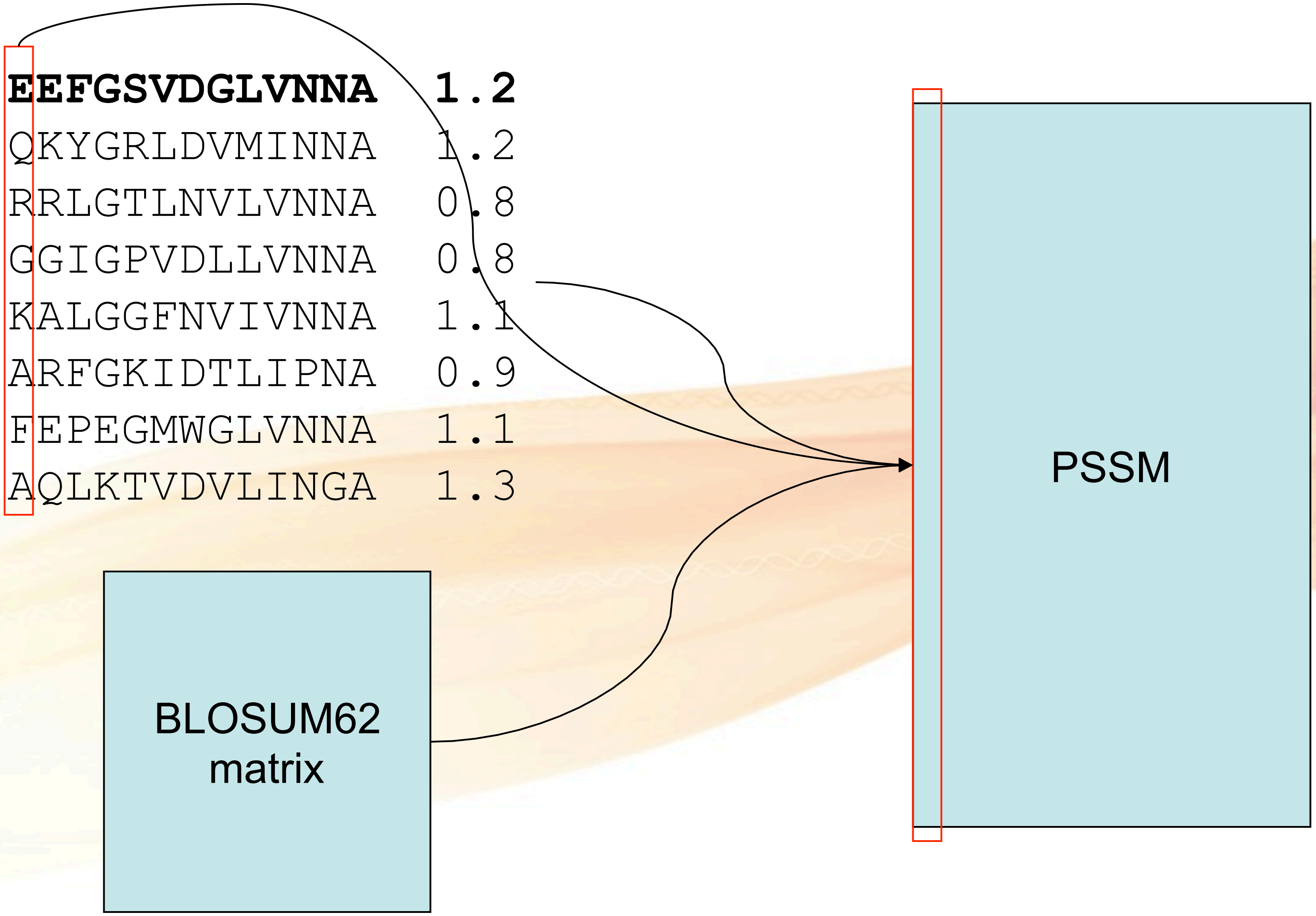
```
EEFGSVDGLVNNA
QKYGRLDVMINNA
RRLGTLNVLVNNA
GGIGPVD-LVNNA
KALGGFNVIVNNA
ARFGKID-LIPNA
FEPEGMWGLVNNA
AQLKTVDVLINGA
```

Compute Sequence eights

- Low weights are assigned to redundant sequences
- High weights are assigned to unique sequences

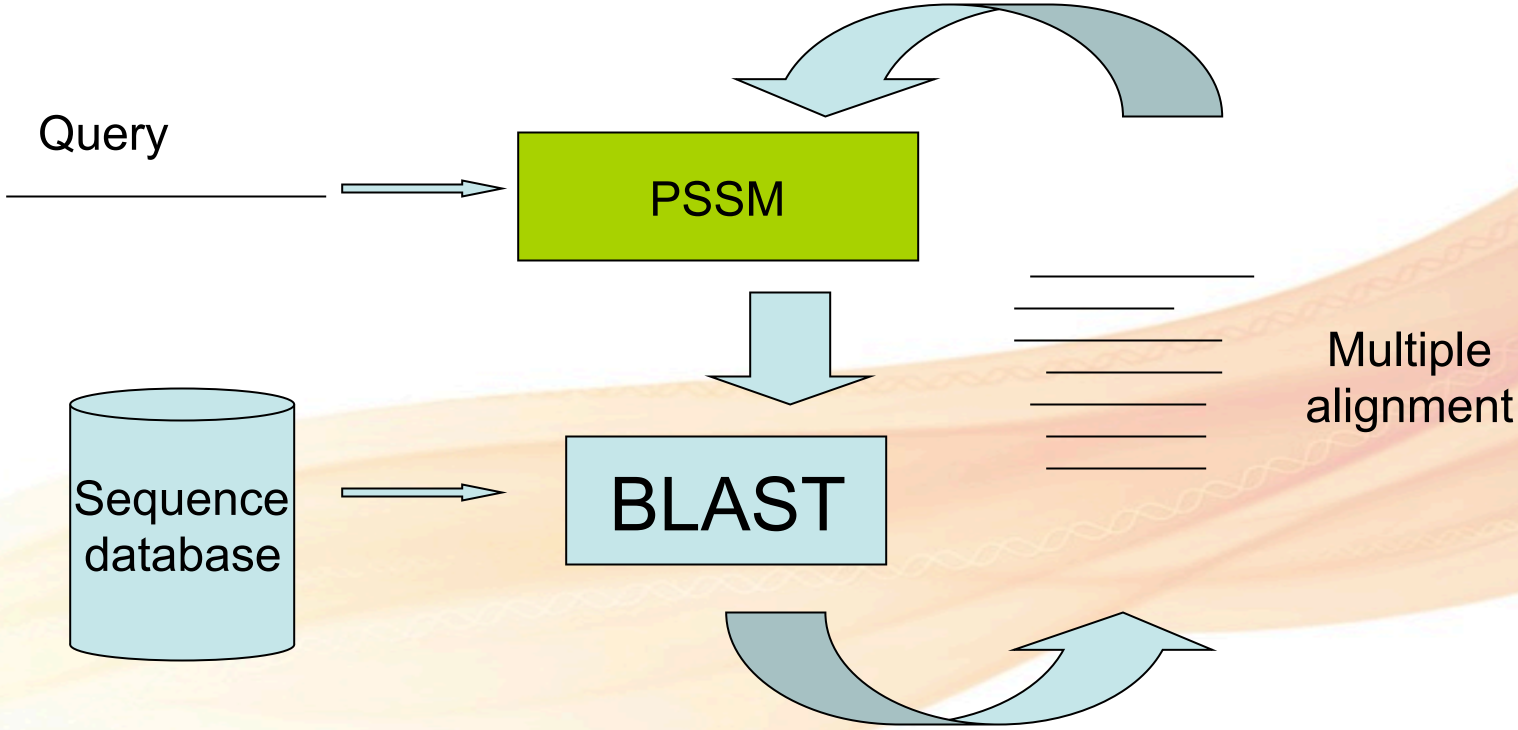
EEFGSVDGLVNNA	**1.2**
QKYGRLDVMINNA	1.2
RRLGTLNVLVNNA	0.8
GGIGPVDLLVNNA	0.8
KALGGFNVIVNNA	1.1
ARFGKIDTLIPNA	0.9
FEPEGMWGLVNNA	1.1
AQLKTVDVLINGA	1.3

Compute PSSM entries



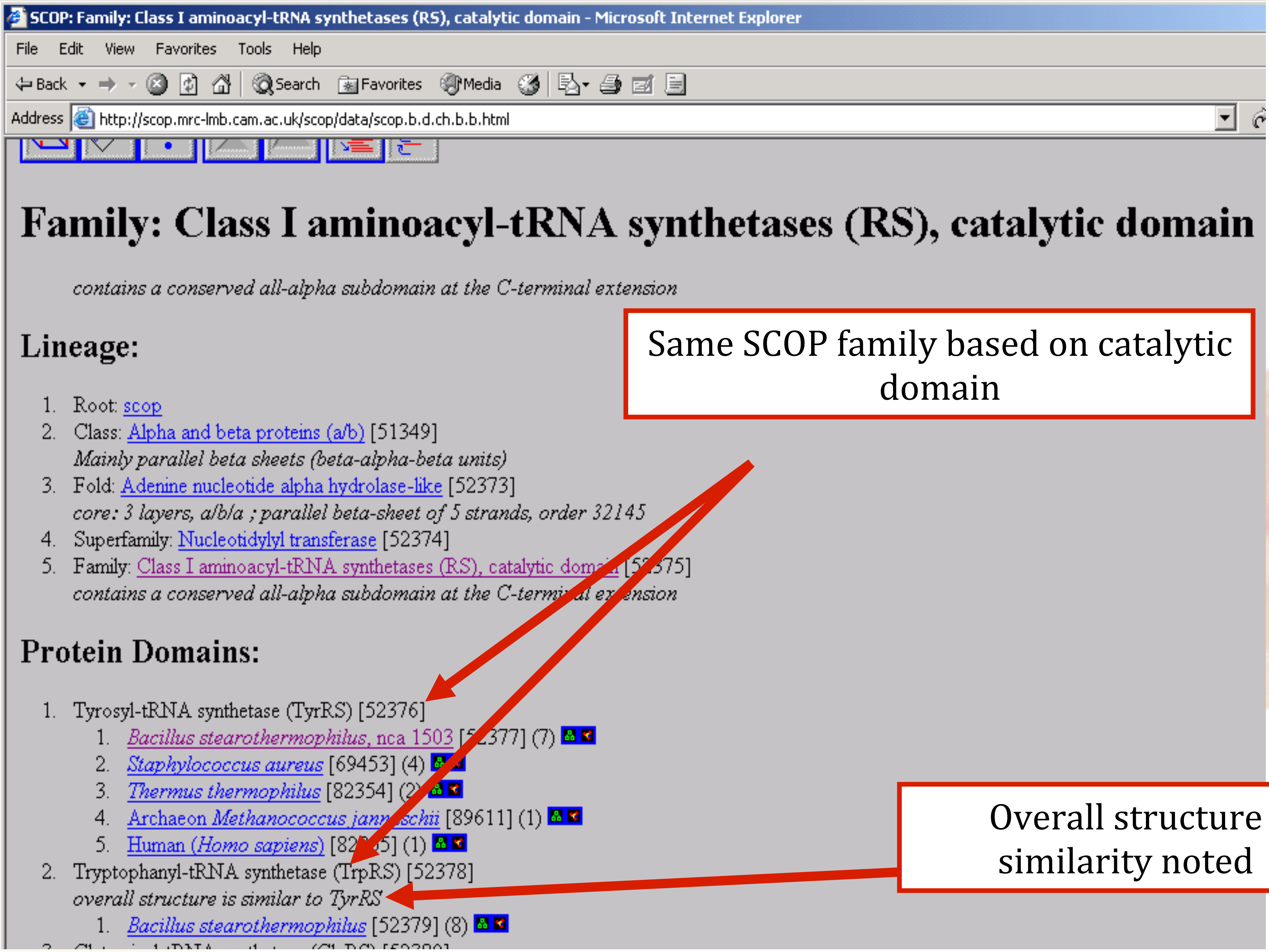
EEFGSVDGLVNNA	**1.2**
QKYGRLDVMINNA	1.2
RRLGTLNVLVNNA	0.8
GGIGPVDLLVNNA	0.8
KALGGFNVIVNNA	1.1
ARFGKIDTLIPNA	0.9
FEPEGMWGLVNNA	1.1
AQLKTVDVLINGA	1.3

PSI-BLAST



Example PSI-BLAST
Aminoacyl tRNA Synthetases

- 20 enzymes for 20 amino acids
- Each is very different
  - Big, small, monomers, tetramers…
  - All bind to their appropriate tRNAs and amino acids, with high specificity
- **Tryptophanyl-tRNA** (TrpRS) and **Tyrosyl-tRNA** (TyrRS) share only 13% sequence identity
  - But, overall structures of TrpTRS and TyrTRS are **similar**
  - **Structure ⇔ Function relationship**


SCOP: Family: Class I aminoacyl-tRNA synthetases (RS), catalytic domain - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Media
Address
http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.d.ch.b.b.html

Family: Class I aminoacyl-tRNA synthetases (RS), catalytic domain

contains a conserved all-alpha subdomain at the C-terminal extension

Lineage:

1. Root: scop
2. Class: Alpha and beta proteins (a/b) [51349]
   Mainly parallel beta sheets (beta-alpha-beta units)
3. Fold: Adenine nucleotide alpha hydrolase-like [52373]
   core: 3 layers, a/b/a ; parallel beta-sheet of 5 strands, order 32145
4. Superfamily: Nucleotidylyl transferase [52374]
5. Family: Class I aminoacyl-tRNA synthetases (RS), catalytic domain [52375]
   contains a conserved all-alpha subdomain at the C-terminal extension

Protein Domains:

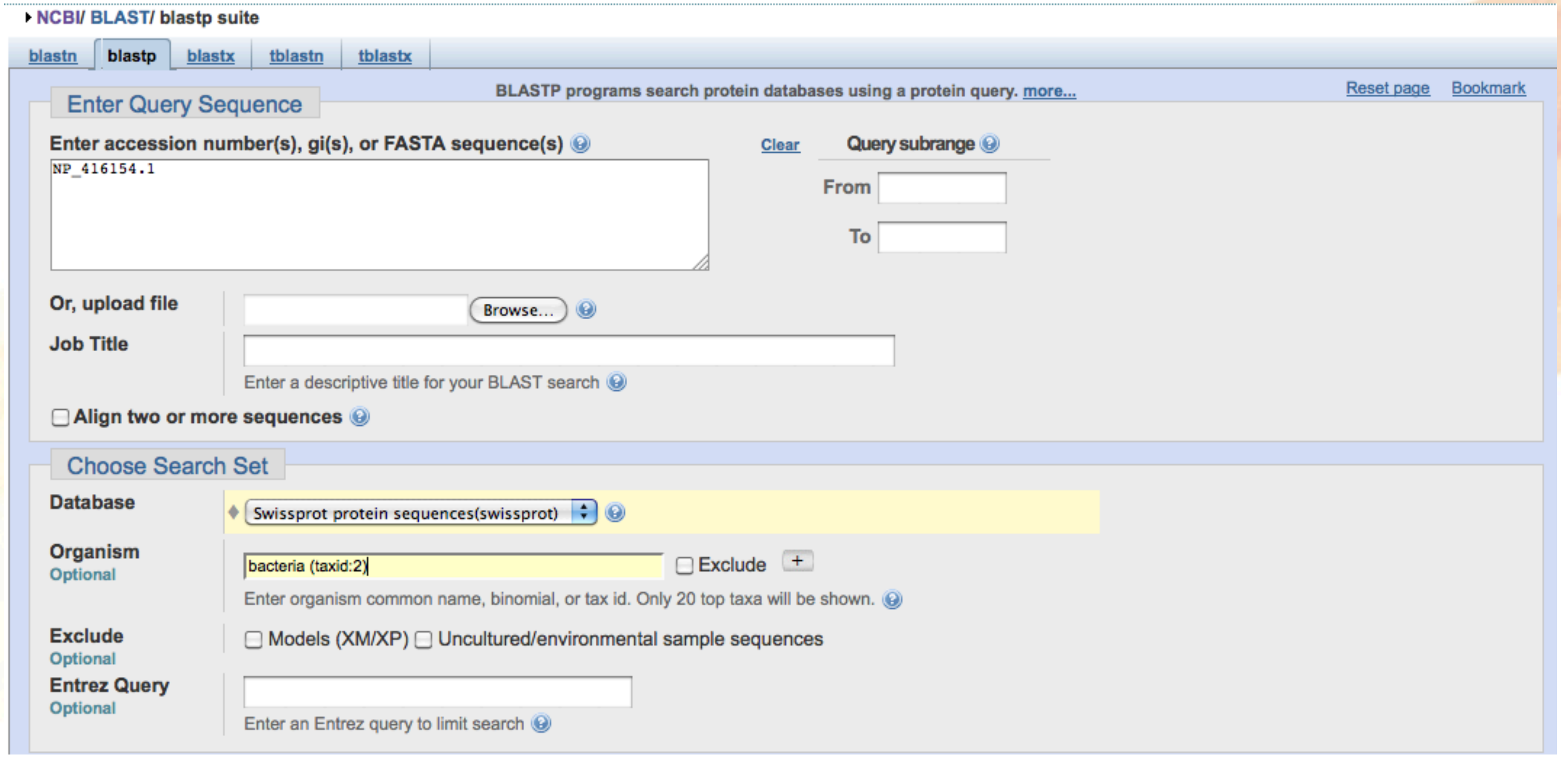
1. Tyrosyl-tRNA synthetase (TyrRS) [52376]
   1. *Bacillus stearothermophilus*, nca 1503 [52377] (7)
   2. *Staphylococcus aureus* [69453] (4)
   3. *Thermus thermophilus* [82354] (2)
   4. Archaeon *Methanococcus jannaschii* [89611] (1)
   5. Human (*Homo sapiens*) [82[illegible]5] (1)
2. Tryptophanyl-tRNA synthetase (TrpRS) [52378]
   overall structure is similar to TyrRS
   1. *Bacillus stearothermophilus* [52379] (8)
3. Glutaminyl-tRNA synthetase (GlnRS) [52380]

Same SCOP family based on catalytic domain

Overall structure similarity noted

So is There Sequence Similarity between TyrRS and TrpRS?

- Given structural similarities, we would expect to find sequence similarity…
- BUT!
  - BLASTp of E.coli TyrRS (NP_416154.1) against bacterial sequences in SwissProt does NOT show similarity with TrpRS at e-value cutoff of 10



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s): NP_416154.1

Query subrange — From: — To:

Or, upload file

Job Title — Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

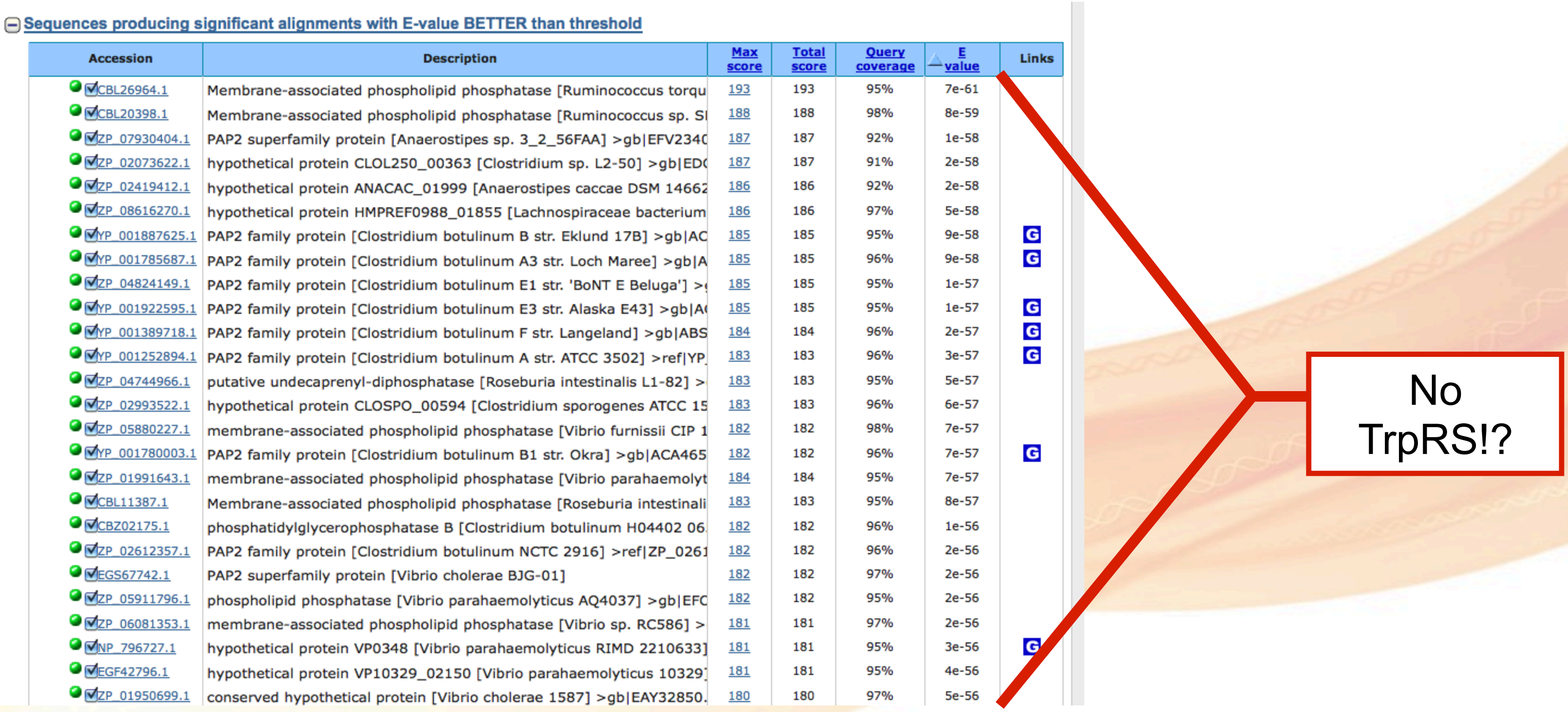
Database: Swissprot protein sequences(swissprot)

Organism (Optional): bacteria (taxid:2) □ Exclude — Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query (Optional): — Enter an Entrez query to limit search

Not Found by BLASTP



Sequences producing significant alignments with E-value BETTER than threshold

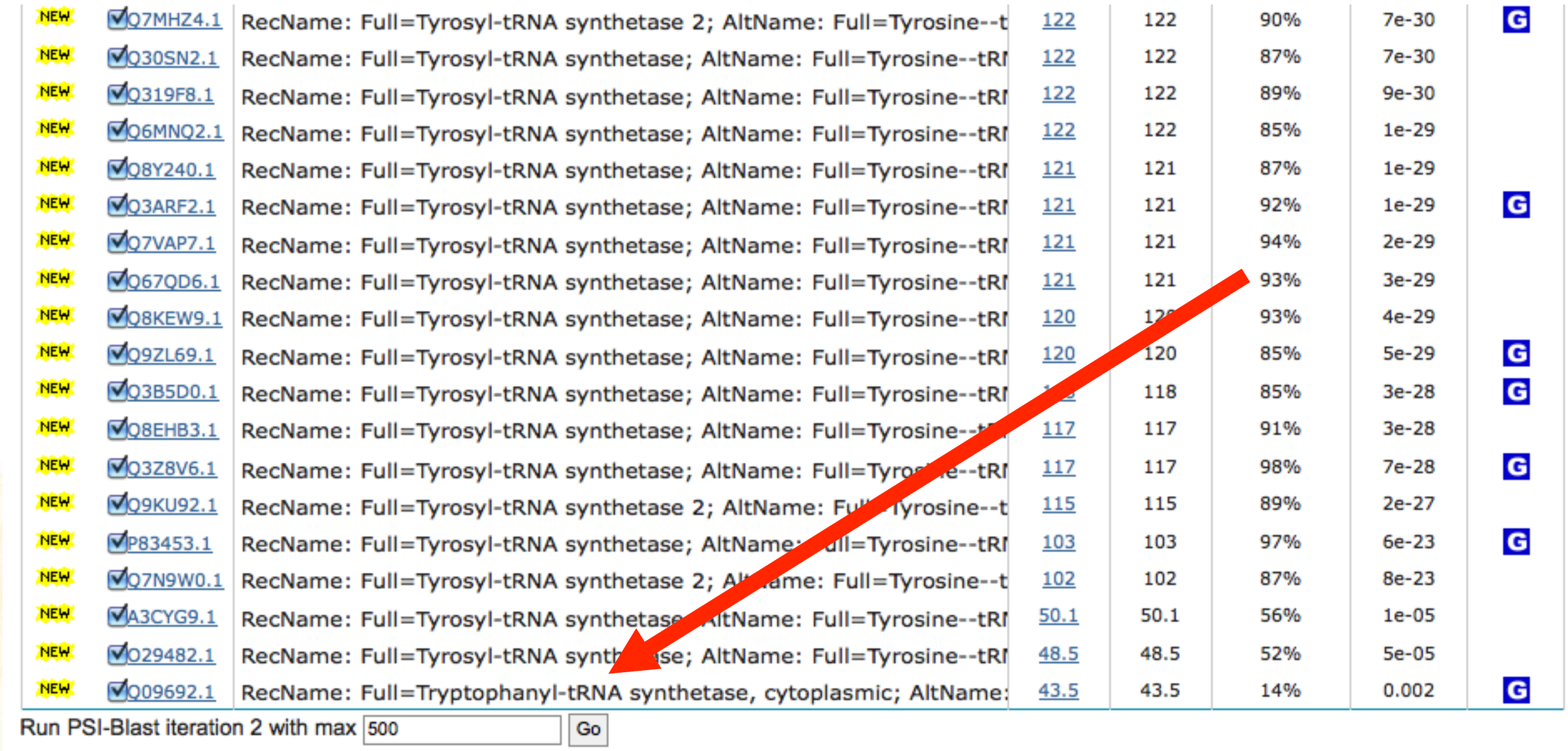
Accession	Description	Max score	Total score	Query coverage	E value	Links
☑CBL26964.1	Membrane-associated phospholipid phosphatase [Ruminococcus torqu	193	193	95%	7e-61	
☑CBL20398.1	Membrane-associated phospholipid phosphatase [Ruminococcus sp. SI	188	188	98%	8e-59	
☑ZP_07930404.1	PAP2 superfamily protein [Anaerostipes sp. 3_2_56FAA] >gb|EFV234C	187	187	92%	1e-58	
☑ZP_02073622.1	hypothetical protein CLOL250_00363 [Clostridium sp. L2-50] >gb|EDC	187	187	91%	2e-58	
☑ZP_02419412.1	hypothetical protein ANACAC_01999 [Anaerostipes caccae DSM 14662	186	186	92%	2e-58	
☑ZP_08616270.1	hypothetical protein HMPREF0988_01855 [Lachnospiraceae bacterium	186	186	97%	5e-58	
☑YP_001887625.1	PAP2 family protein [Clostridium botulinum B str. Eklund 17B] >gb|AC	185	185	95%	9e-58	G
☑YP_001785687.1	PAP2 family protein [Clostridium botulinum A3 str. Loch Maree] >gb|A	185	185	96%	9e-58	G
☑ZP_04824149.1	PAP2 family protein [Clostridium botulinum E1 str. 'BoNT E Beluga'] >	185	185	95%	1e-57	
☑YP_001922595.1	PAP2 family protein [Clostridium botulinum E3 str. Alaska E43] >gb|A	185	185	95%	1e-57	G
☑YP_001389718.1	PAP2 family protein [Clostridium botulinum F str. Langeland] >gb|ABS	184	184	96%	2e-57	G
☑YP_001252894.1	PAP2 family protein [Clostridium botulinum A str. ATCC 3502] >ref|YP	183	183	96%	3e-57	G
☑ZP_04744966.1	putative undecaprenyl-diphosphatase [Roseburia intestinalis L1-82] >	183	183	95%	5e-57	
☑ZP_02993522.1	hypothetical protein CLOSPO_00594 [Clostridium sporogenes ATCC 15	183	183	96%	6e-57	
☑ZP_05880227.1	membrane-associated phospholipid phosphatase [Vibrio furnissii CIP 1	182	182	98%	7e-57	
☑YP_001780003.1	PAP2 family protein [Clostridium botulinum B1 str. Okra] >gb|ACA465	182	182	96%	7e-57	G
☑ZP_01991643.1	membrane-associated phospholipid phosphatase [Vibrio parahaemolyt	184	184	95%	7e-57	
☑CBL11387.1	Membrane-associated phospholipid phosphatase [Roseburia intestinali	183	183	95%	8e-57	
☑CBZ02175.1	phosphatidylglycerophosphatase B [Clostridium botulinum H04402 06.	182	182	96%	1e-56	
☑ZP_02612357.1	PAP2 family protein [Clostridium botulinum NCTC 2916] >ref|ZP_0261	182	182	96%	2e-56	
☑EGS67742.1	PAP2 superfamily protein [Vibrio cholerae BJG-01]	182	182	97%	2e-56	
☑ZP_05911796.1	phospholipid phosphatase [Vibrio parahaemolyticus AQ4037] >gb|EFC	182	182	95%	2e-56	
☑ZP_06081353.1	membrane-associated phospholipid phosphatase [Vibrio sp. RC586] >	181	181	97%	2e-56	
☑NP_796727.1	hypothetical protein VP0348 [Vibrio parahaemolyticus RIMD 2210633]	181	181	95%	3e-56	G
☑EGF42796.1	hypothetical protein VP10329_02150 [Vibrio parahaemolyticus 10329]	181	181	95%	4e-56	
☑ZP_01950699.1	conserved hypothetical protein [Vibrio cholerae 1587] >gb|EAY32850.	180	180	97%	5e-56	

No TrpRS!?

Try Using PSI-BLAST…

- PSI-BLAST available from BLAST main page
- Query form just like for BLASTp
  - BUT: one extra formatting option must be used
  - "Program Selection" – activate the check box
    - PSI-BLAST (Position-Specific Iterated BLAST)"
  - Second e-value cutoff used to determine which alignments will be used for PSSM build… "Threshold for inclusion"
- First search using BLASTp of E.coli TyrRS as query
  - Threshold for inclusion = 0.005

The Power of a Single PSSM

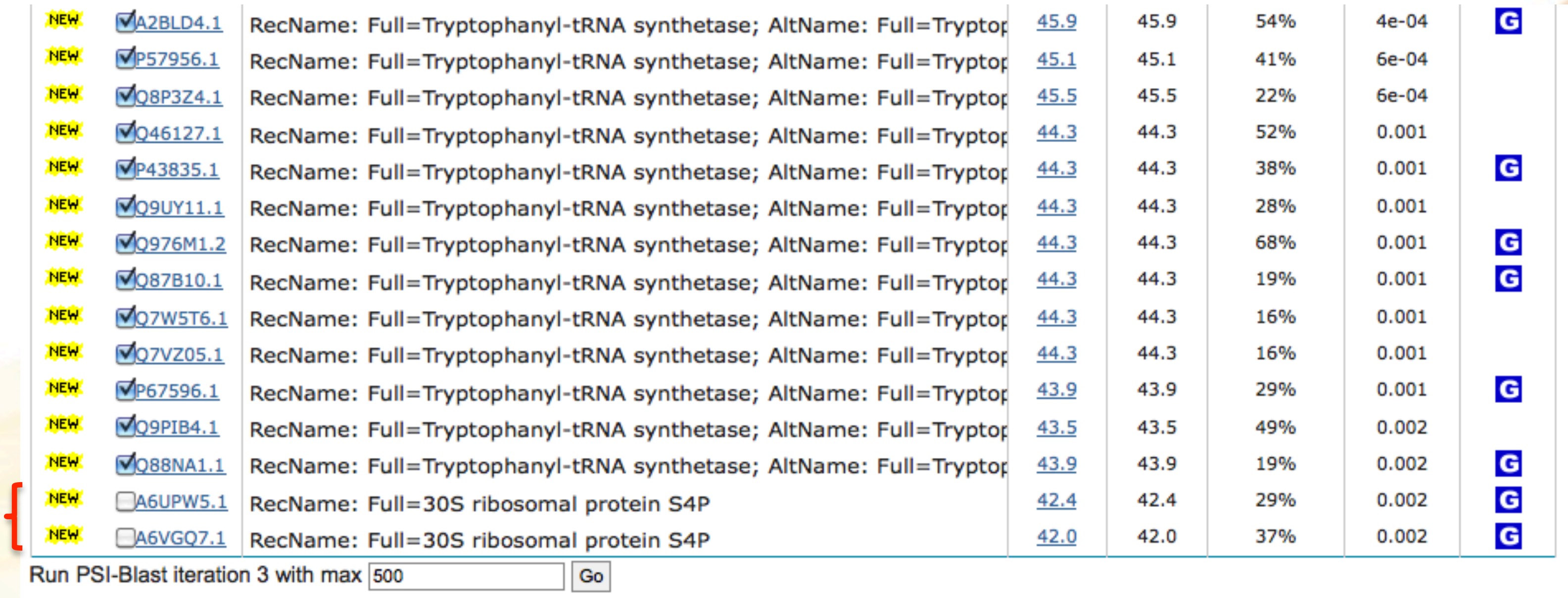


NEW	☑Q7MHZ4.1	RecName: Full=Tyrosyl-tRNA synthetase 2; AltName: Full=Tyrosine--t	122	122	90%	7e-30	G
NEW	☑Q30SN2.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	122	122	87%	7e-30	
NEW	☑Q319F8.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	122	122	89%	9e-30	
NEW	☑Q6MNQ2.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	122	122	85%	1e-29	
NEW	☑Q8Y240.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	121	121	87%	1e-29	
NEW	☑Q3ARF2.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	121	121	92%	1e-29	G
NEW	☑Q7VAP7.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	121	121	94%	2e-29	
NEW	☑Q67QD6.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	121	121	93%	3e-29	
NEW	☑Q8KEW9.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	120	120	93%	4e-29	
NEW	☑Q9ZL69.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	120	120	85%	5e-29	G
NEW	☑Q3B5D0.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	118	118	85%	3e-28	G
NEW	☑Q8EHB3.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	117	117	91%	3e-28	
NEW	☑Q3Z8V6.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	117	117	98%	7e-28	G
NEW	☑Q9KU92.1	RecName: Full=Tyrosyl-tRNA synthetase 2; AltName: Full=Tyrosine--t	115	115	89%	2e-27	
NEW	☑P83453.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	103	103	97%	6e-23	G
NEW	☑Q7N9W0.1	RecName: Full=Tyrosyl-tRNA synthetase 2; AltName: Full=Tyrosine--t	102	102	87%	8e-23	
NEW	☑A3CYG9.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	50.1	50.1	56%	1e-05	
NEW	☑O29482.1	RecName: Full=Tyrosyl-tRNA synthetase; AltName: Full=Tyrosine--tRI	48.5	48.5	52%	5e-05	
NEW	☑Q09692.1	RecName: Full=Tryptophanyl-tRNA synthetase, cytoplasmic; AltName:	43.5	43.5	14%	0.002	G

Run PSI-Blast iteration 2 with max 500 Go

The last hit was a Trp-tRNA – Remember a PSSM

The Power of Iterations



NEW	☑A2BLD4.1	RecName: Full=Tryptophanyl-tRNA synthetase; AltName: Full=Tryptop	45.9	45.9	54%	4e-04	G
NEW	☑P57956.1	RecName: Full=Tryptophanyl-tRNA synthetase; AltName: Full=Tryptop	45.1	45.1	41%	6e-04	
NEW	☑Q8P3Z4.1	RecName: Full=Tryptophanyl-tRNA synthetase; AltName: Full=Tryptop	45.5	45.5	22%	6e-04	
NEW	☑Q46127.1	RecName: Full=Tryptophanyl-tRNA synthetase; AltName: Full=Tryptop	44.3	44.3	52%	0.001	
NEW	☑P43835.1	RecName: Full=Tryptophanyl-tRNA synthetase; AltName: Full=Tryptop	44.3	44.3	38%	0.001	G
NEW	☑Q9UY11.1	RecName: Full=Tryptophanyl-tRNA synthetase; AltName: Full=Tryptop	44.3	44.3	28%	0.001	
NEW	☑Q976M1.2	RecName: Full=Tryptophanyl-tRNA synthetase; AltName: Full=Tryptop	44.3	44.3	68%	0.001	G
NEW	☑Q87B10.1	RecName: Full=Tryptophanyl-tRNA synthetase; AltName: Full=Tryptop	44.3	44.3	19%	0.001	G
NEW	☑Q7W5T6.1	RecName: Full=Tryptophanyl-tRNA synthetase; AltName: Full=Tryptop	44.3	44.3	16%	0.001	
NEW	☑Q7VZ05.1	RecName: Full=Tryptophanyl-tRNA synthetase; AltName: Full=Tryptop	44.3	44.3	16%	0.001	
NEW	☑P67596.1	RecName: Full=Tryptophanyl-tRNA synthetase; AltName: Full=Tryptop	43.9	43.9	29%	0.001	G
NEW	☑Q9PIB4.1	RecName: Full=Tryptophanyl-tRNA synthetase; AltName: Full=Tryptop	43.5	43.5	49%	0.002	
NEW	☑Q88NA1.1	RecName: Full=Tryptophanyl-tRNA synthetase; AltName: Full=Tryptop	43.9	43.9	19%	0.002	G
NEW	□A6UPW5.1	RecName: Full=30S ribosomal protein S4P	42.4	42.4	29%	0.002	G
NEW	□A6VGQ7.1	RecName: Full=30S ribosomal protein S4P	42.0	42.0	37%	0.002	G

Run PSI-Blast iteration 3 with max 500 Go

Uncheck why?

Iterations with a new PSSM bring about more TrpRS

Why (not) PSI-BLAST

- If the sequences used to construct the Position Specific Scoring Matrices (PSSMs) are all homologous, the sensitivity at a given specificity improves significantly
- However, if non-homologous sequences are included in the PSSMs, they are "**corrupted**"
  - Then they pull in more non-homologous sequences, and become worse than generic

Power of PSI-BLAST

- We knew TyrRS and TrpRS were similar
  - Functionally and structurally
- BLASTP gave no indication
  - PSI-BLAST able to detect their weak sequence similarity
- Words of caution:
  - Be sure to inspect and think about the results included in the PSSM build
  - Include/exclude sequences on basis of biological knowledge: you are in the driving seat!
  - PSI-BLAST performance varies according to choice of matrix, filter, statistics etc just like BLASTP

FYI - PSI-BLAST caveats

- Increased ability to find distant homologues
- Cost of additional required care to prevent non-homologous sequences from being included in the PSSM calculation
  - When in doubt, leave it out!
  - Examine sequences with moderate similarity carefully
- Be particularly cautious about matches to sequences with highly biased amino acid content
  - Low complexity regions, transmembrane regions and coiled-coil regions often display significant similarity without homology
  - Screen them out of your query sequences!

Another Sequence for PSI-BLAST

```
>gi|294497031|ref|YP_003560731.1| hypothetical protein BMQ_0196 [Bacillus megaterium QM
B1551]
MDKLMNRSWVMKIIALLLAFMLYLSVNLDDGASSSNKILNRSSSANTGVETLTDVPVQVSYNEKNRIVRG
VPDTVIMTLEGPKNILAQTKLQKDYQAYIDLDNLSLGQHRVKVQYRNISDNLNVVVKPDIVNVTIEERDS
KQFSVEASYDKNKVKNGYEAGEATVSPRAVTVTGASSQLDQVAYVKAIIDLDNASKTVTKQATVVALDKN
LNKLNVTVQPETVNVTIPVRNISKKVPIDVIQEGTPGDGVNITKLEPKTDTVKIIGPSDSLEKIDKIDNI
PVDVTGITKSKDIKVNVPVPDGIDSVSPKQITVHVEVDKQGDEKDAEETDASAAETKSFKNLPVSLTGQS
SKYTYELLSPTSVDADVKGPKSDLDKLTKSGISLSANVGNLSAGEHTVPIIINSPDSVTSTLSTKQAKVR
VTAKKQSGTNDEQTDDKETSGSTSDKETSGSTSDKETKPDTGTGSGTNPGTGNSGDSADKPSEETDTPED
NTDTPTDSTETGDDSSNQSDENSTPVDGQTDNTSGN
```

Exercise: Try is yourself!

You don’t need to cut & paste, remember all you need is the accession number

PHI-BLAST

PHI-BLAST

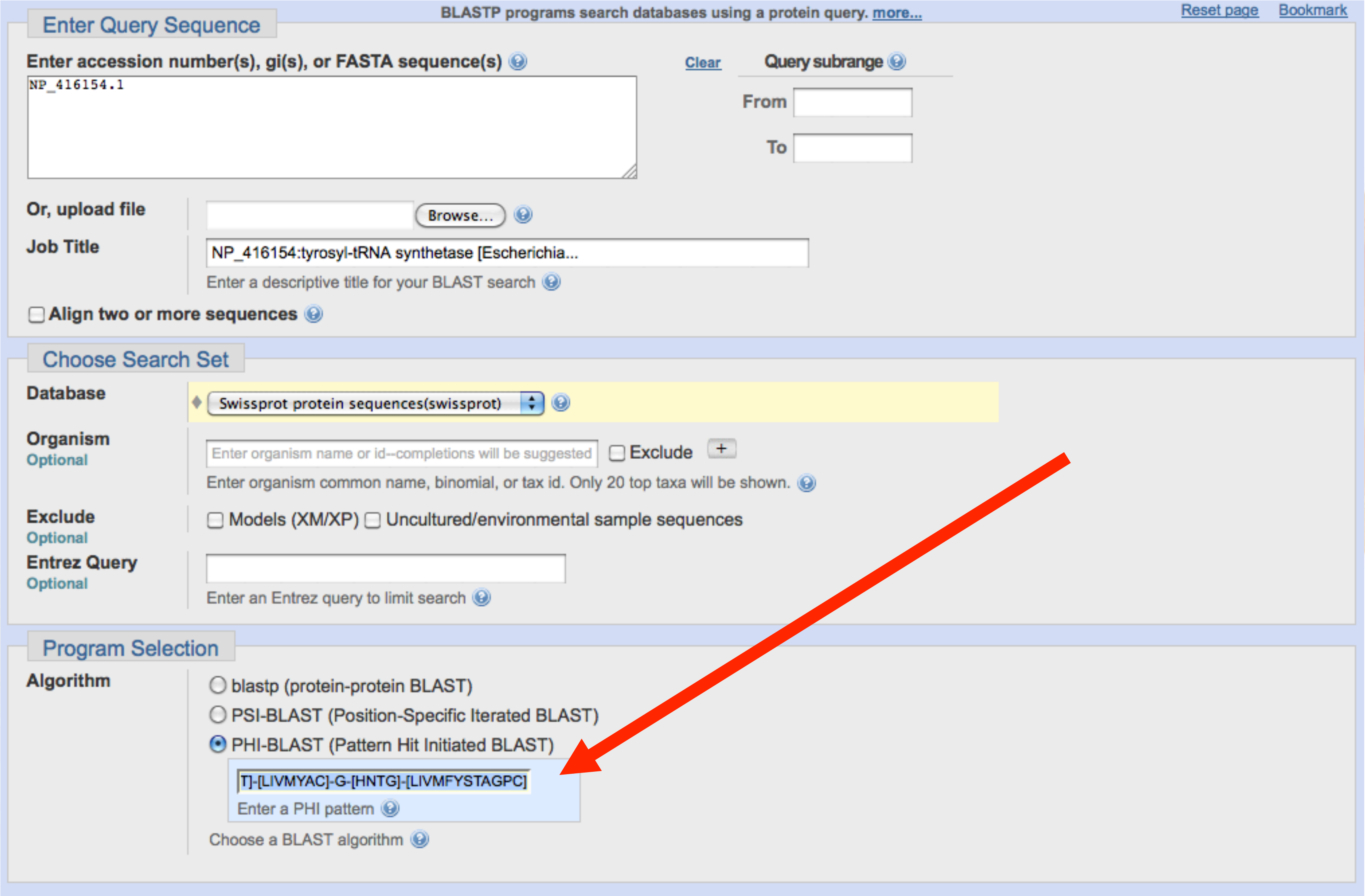
- Pattern Hit Initiated – BLAST
- PHI-BLAST principle:
  - Same method as PSI-BLAST
  - Starts first search with query sequence + pattern for a motif in the query
- PHI-BLAST finds sequences containing the motif and having significant sequence similarity in the vicinity of the motif occurrence
  - Highly specific

Example: TyrRS

- TyrRS contains the aaRS class-I signature
- Want to find sequences containing that motif, and regional similarity to TyrRS
- First: get the pattern for the class-I signature
  - Where can I get this?
  - PROSITE = Database of protein domains, families and functional sites
    - **Aminoacyl-transfer RNA synthetases class-I signature**

P-x(0,2)-[GSTAN]-[DENQGAPK]-x-[LIVMFP]-[HT]-[LIVMYAC]-G-[HNTG]-[LIVMFYSTAGPC]

Same as Before, but Use a Pattern



Enter Query Sequence

BLASTP programs search databases using a protein query. more...

Reset page Bookmark

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

NP_416154.1

From

To

Or, upload file: Browse...

Job Title: NP_416154:tyrosyl-tRNA synthetase [Escherichia...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Swissprot protein sequences(swissprot)

Organism Optional: Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

○ blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

● PHI-BLAST (Pattern Hit Initiated BLAST)

T]-[LIVMYAC]-G-[HNTG]-[LIVMFYSTAGPC]

Enter a PHI pattern

Choose a BLAST algorithm

PHI-BLAST Results

- After first search, PHI-BLAST functions same as PSI-BLAST
- Result page is the same
- Can iterate in same way

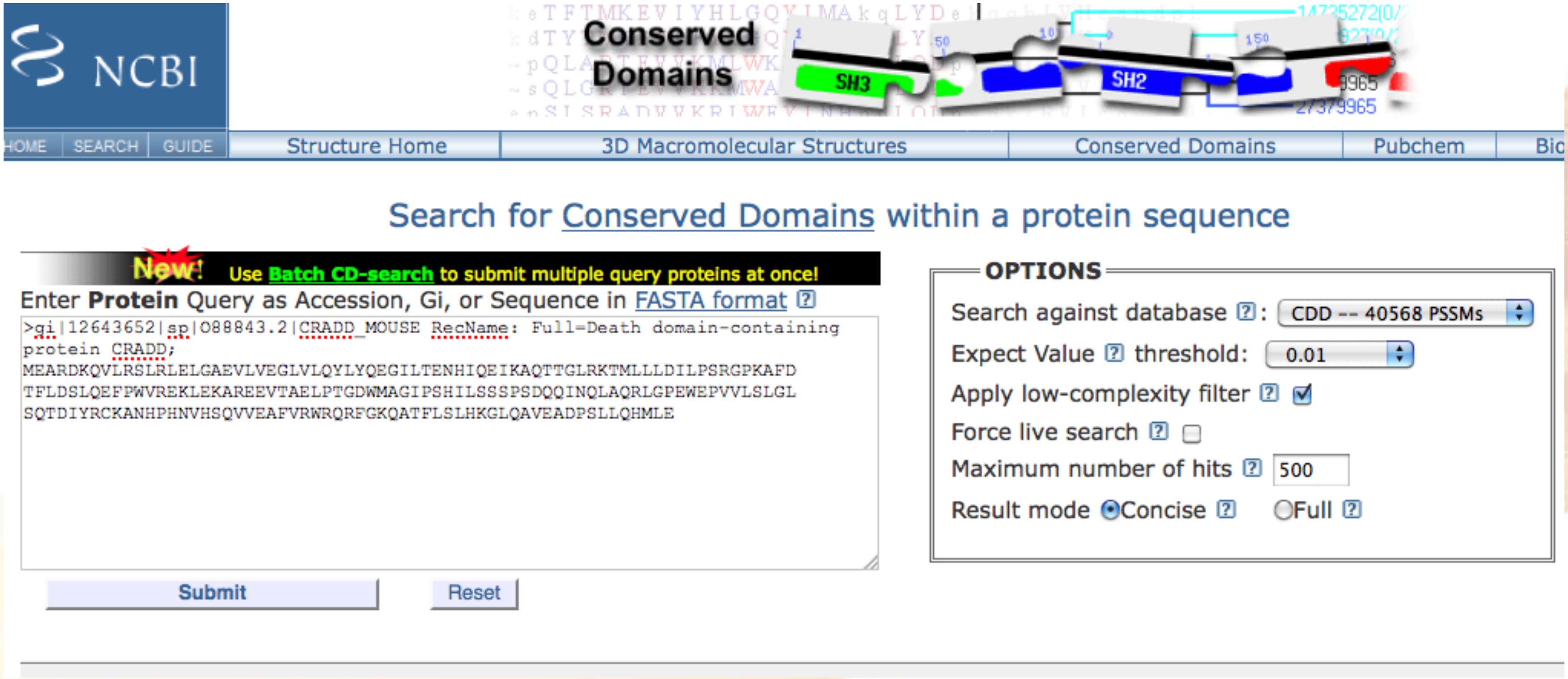
- Try it later if you like…

RPS-BLAST

- Annotated collections of multiple sequence alignments defining where domains exist
  - Conserved domain database (CDD)
  - Contains 40,568 PSSMs (2011)
    - Contains 43,334 PSSMs (2012)
- Can search the CDD using CD search
  - Uses RPS-BLAST
    - Reverse Position Specific–BLAST
  - Opposite of PSI-BLAST
- CDD multiple alignments converted to PSSMs
- PSSMs are processed and turned into a searchable database
- Queries are searched against PSSMs using RPS-BLAST
- Output indicates conserved domains within the query sequence

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

Example: CRADD protein



Search for Conserved Domains within a protein sequence

New! Use Batch CD-search to submit multiple query proteins at once!

Enter **Protein** Query as Accession, Gi, or Sequence in FASTA format ?

```
>gi|12643652|sp|O88843.2|CRADD_MOUSE RecName: Full=Death domain-containing
protein CRADD;
MEARDKQVLRSLRLELGAEVLVEGLVLQYLYQEGILTENHIQEIKAQTTGLRKTMLLLDILPSRGPKAFD
TFLDSLQEFPWVREKLEKAREEVTAELPTGDWMAGIPSHILSSSPSDQQINQLAQRLGPEWEPVVLSLGL
SQTDIYRCKANHPHNVHSQVVEAFVRWRQRFGKQATFLSLHKGLQAVEADPSLLQHMLE
```

OPTIONS

Search against database ?: CDD -- 40568 PSSMs

Expect Value ? threshold: 0.01

Apply low-complexity filter ? ☑

Force live search ? □

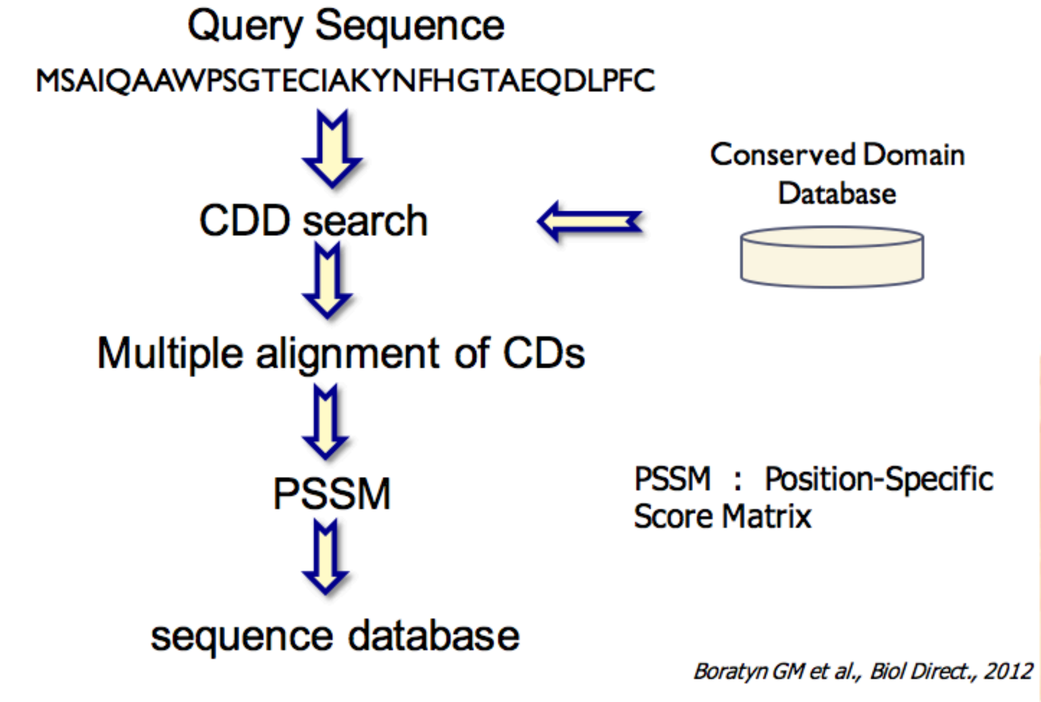
Maximum number of hits ? 500

Result mode ● Concise ? ○ Full ?

DELTA-BLAST

- **DELTA-BLAST (Domain Enhanced Look-up Time Accelerated BLAST) is a new application to perform protein-protein queries for the detection of distant protein homologs**
- Generally, BLAST is used to compare a query sequence to a database of known sequences
- Position-specific-iterated BLAST (PSI-BLAST) iteratively searches a protein sequence database, using the matches in round A to construct a position-specific score matrix (PSSM) for searching the database in round A + 1 and so on
- **DELTA-BLAST searches a database of pre-constructed PSSMs before searching a protein-sequence database**
  - **Performs a multiple sequence alignment of the query sequence with domains described in the CDD (Conserved Domain Database from NCBI) database and then uses a PSSM derived from this alignment to search a sequence database**

DELTA-BLAST



Query Sequence

MSAIQAAWPSGTECIAKYNFHGTAEQDLPFC

CDD search

Conserved Domain Database

Multiple alignment of CDs

PSSM

sequence database

PSSM : Position-Specific Score Matrix

Boratyn GM et al., Biol Direct., 2012

The difference with PSI-BLAST is that PSI-BLAST uses the results of a first blastp iteration to construct a PSSM and then uses it to search the sequence database

DELTA-BLAST uses PSSM derived from the CDD database, so the initial PSSM construction is much more quicker than PSI-BLAST

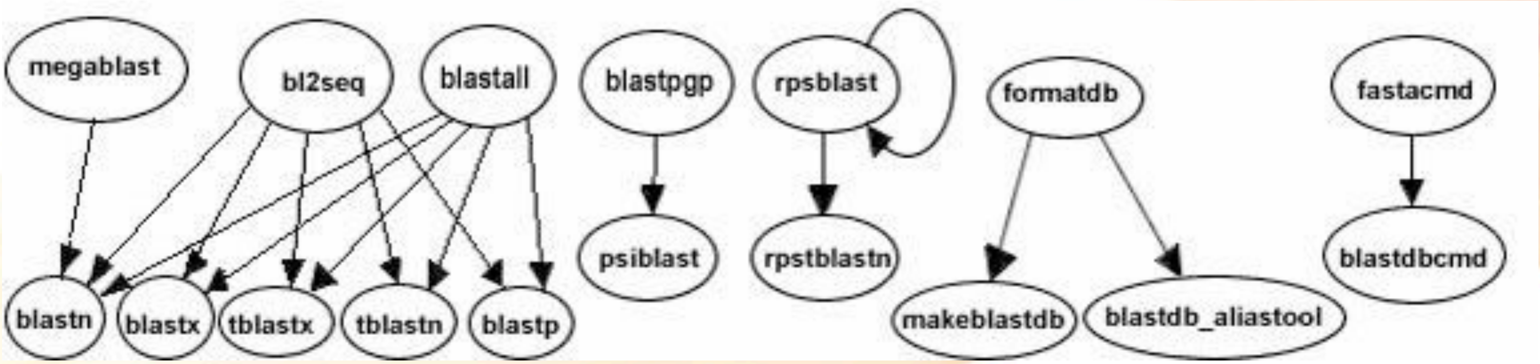
http://www.biologydirect.com/content/7/1/12

5 mins…

Installing and Using the New BLAST+

Functionality Offered by BLAST+

- BLAST+ applications has been organized by program type
- Resembles Web BLAST
- Following graph depicts a correspondence between NCBI C Toolkit BLAST CLI and the BLAST+ applications



J. Peter Gogarten

Step 1 – Installing BLAST+ tools

- http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download
- **Windows 32/64:** .exe installer
  - **Linux 32/64:** compiled binaries (RPM or tar.gz)
  - **Other Unix:** compiled binaries (in tar.gz)
- Apply platform-specific configuration details for your operating system
- Read the good documentation:

  http://www.ncbi.nlm.nih.gov/books/NBK1763/

Richard Bruskiewich

Installing From Linux .tar.gz Archive

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

```
$ ##DO NOT DO THIS ON FISHER
$ ##THIS IS ON YOUR OWN LINBOX or MAC
$ su
$ wget ftp://ftp.ncbi.nlm.nih.gov/blast/executables/ \
blast+/LATEST/ncbi-blast-2.2.28+-x64-linux.tar.gz
$ tar -zxvf ncbi-blast-2.2.28+-x64-linux.tar.gz
$ mv ncbi-blast-2.2.28+ /usr/local
$ cd /usr/local/
$ ln -s ncbi-blast-2.2.28+ ncbi
$ exit
```

Adopted from Richard Bruskiewich

Edit Your .bashrc File in /home/username

```
$ vi .bashrc

#added to my for blast+
export BLAST=/usr/local/ncbi
export PATH=$BLAST/bin:$PATH

#exit vi

$ source .bashrc
```

if it worked, then :

```
$ echo $BLAST
/usr/local/ncbi
```

Adopted from Richard Bruskiewich

Option A: Obtain a Existing NCBI Database… Manual Way (Automatic to Come)

```
$ su

$ mkdir $BLAST/db

$ wget ftp://ftp.ncbi.nlm.nih.gov/blast/db/swissprot.tar.gz

$ tar -zxvf swissprot.tar.gz $BLAST/db

$ exit

$ cd
```

Adopted from Richard Bruskiewich

Option B: Create a simple BLAST database from a local FASTA sequence file

- **Makeblastdb (formerly formatdb)** produces BLAST databases from FASTA files
  - In the simplest case the FASTA definition lines are not parsed by makeblastdb and may be completely unstructured (but can only be BLAST'ed and not be directly retrieved)

```
$ makeblastdb -in mydb.fsa -dbtype nucl
Legacy
$ formatdb –i mydb.fasta –p F –o F
```

- Creates a BLAST database from a nucleotide FASTA sequences which can be put into the "db" directory for searching
- Of course, like all blast programs, there are a rich set of parameters which can be used to customize the generation of the database (see the BLAST manual)

Adopted from Richard Bruskiewich

With Either Option, Still Need Some Configuration Details

```
$ cd  # back to home directory
# need to point to the database…
$ cat >.ncbirc <<EOF
[BLAST]
BLASTDB=/usr/local/ncbi/db
EOF
```

Do not change your `.ncbirc` file on fisher!!!!

Adopted from Richard Bruskiewich

Step 3 – Executing a BLAST Operation

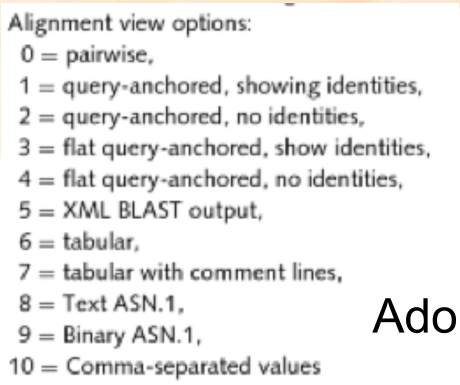
- Command line programs (only) but parameters are generally equivalent to (or a superset of) the NCBI web BLAST application
- Sample run:
  - Retrieve a sequence from the database and put it in test_query.txt:

```
$ blastdbcmd -db nr -entry Q9MAH0 -outfmt "%f" -out
test_query.txt
```

```
%f = fasta
%l = length
%a = accession
```

  - Blast it back against the same database:

```
$ blastp -query test_query.txt -db swissprot -out
output.txt -outfmt 5 -seg yes
```



Alignment view options:
0 = pairwise,
1 = query-anchored, showing identities,
2 = query-anchored, no identities,
3 = flat query-anchored, show identities,
4 = flat query-anchored, no identities,
5 = XML BLAST output,
6 = tabular,
7 = tabular with comment lines,
8 = Text ASN.1,
9 = Binary ASN.1,
10 = Comma-separated values

Adopted Richard Bruskiewich

(Step 3) – Legacy - Executing a BLAST Operation

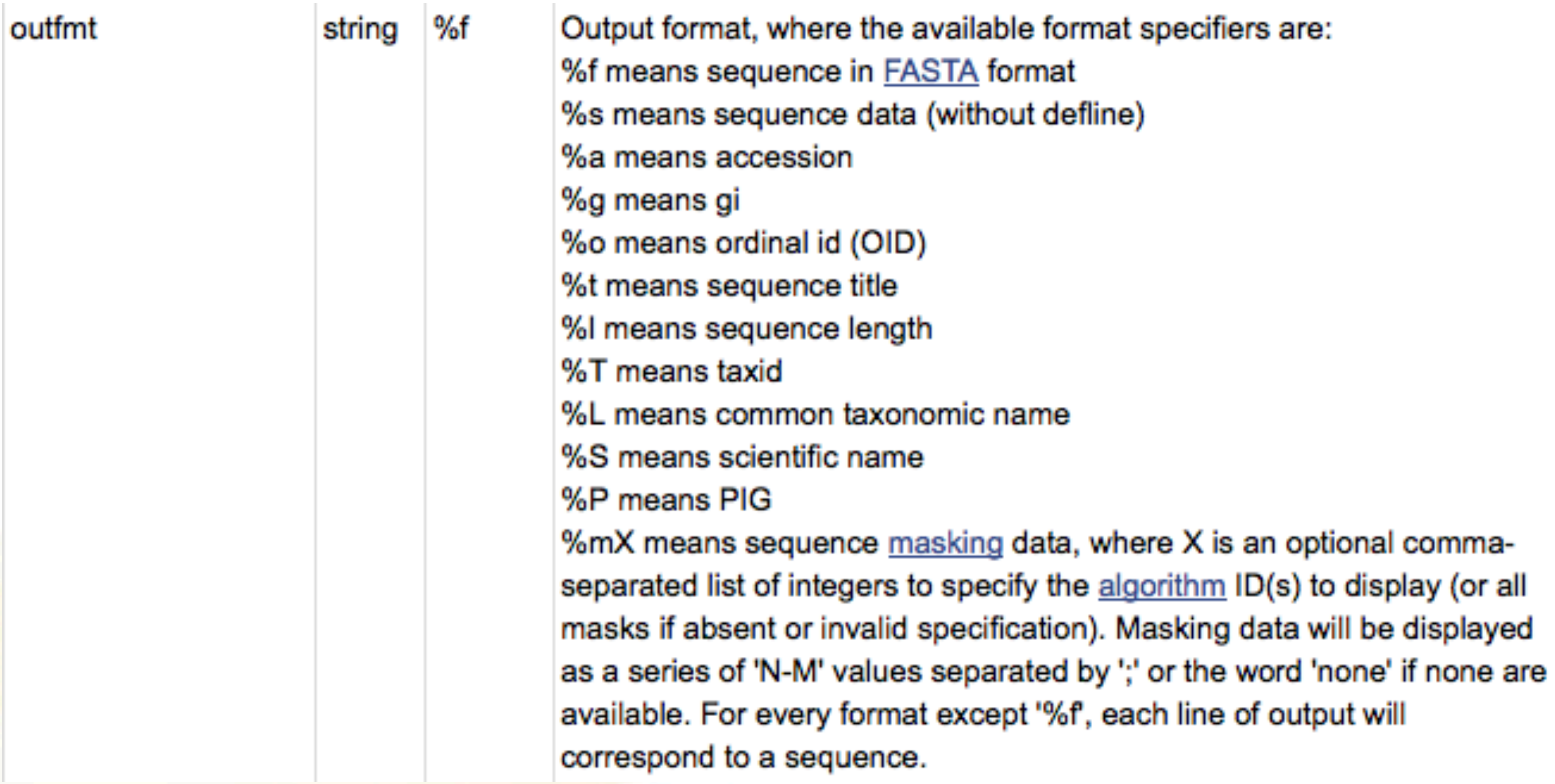
- Command line programs (only) but parameters are generally equivalent to (or a superset of) the NCBI web BLAST application
- Sample run:
  - Retrieve a sequence from the database and put it in test_query.txt:

```
$ fastacmd -s Q9MAH0 -d nr -o test_query.txt
```

  - Blast it back against the same database:

```
$ blastall –p blastp -i test_query.txt -d swissprot -o
output.txt –m 7
```

blastdbcmd outfmt



outfmt	string	%f	Output format, where the available format specifiers are: %f means sequence in FASTA format %s means sequence data (without defline) %a means accession %g means gi %o means ordinal id (OID) %t means sequence title %l means sequence length %T means taxid %L means common taxonomic name %S means scientific name %P means PIG %mX means sequence masking data, where X is an optional comma-separated list of integers to specify the algorithm ID(s) to display (or all masks if absent or invalid specification). Masking data will be displayed as a series of 'N-M' values separated by ';' or the word 'none' if none are available. For every format except '%f', each line of output will correspond to a sequence.

Command Line Parameters: Statistics

- **-evalue**: expect value, normally set to 10
- **-word_size**: “k-tuple” size; increase for speed, decrease for sensitivity
- **-gapopen**: cost to open a gap; increase for stringency
- **-gapextend**: cost to extend a gap; increase for stringency
- **-matrix**: substitution scoring matrix (default BLOSUM62); change if sequences too related or too distant
- To get more information use option “-help”

Richard Bruskiewich

Command Line Parameters: Input/Output

- **-query** in.txt: specify input file
- **-out** out.txt: specify output file
- **-db** nr: which database (created with makeblastdb)
- **-dust** yes/no: filter low complexity regions in nucleotide sequence search yes/no (default is yes)
- **-seg** yes/no: filter low complexity regions in protein sequence search yes/no (default is no)
- **-html** format output as HTML
- **-outfmt** specify output format, e.g. 5 = XML blast output
  - (use –help flag to see other options)

Richard Bruskiewich

Additional Useful Program Options

- Depending on program:
  - **-num_threads**: use multiple CPUs (speeds up search)
  - **-subject**: specify a second input sequence instead of a database (former ‘bl2seq’)
  - **-task megablast**: much faster for highly similar nucleotide sequences
  - **-task blastn_short**: find similar short sequences (e.g. primer sequences)

Richard Bruskiewich

Step 4 – Parse the Output

- If you just have one query sequence, simply view the BLAST+ text file
- If you are doing a lot of queries on the database and looking for "best hits", you may wish to use a parsing script (e.g. BioPerl to the rescue!)
- XML (`-outfmnt 5`) output is parsed easily, as we'll see, and `-outfmnt 7` is a tab delimited output

Adopted from Richard Bruskiewich

Legacy Commands & Learning the All New Commands?

- Easiest way to get stared using new commands is by means of `legacy_blast.pl`
  - PERL script which is bundled along with the BLAST+ applications
- Script is part of blast+, and translates blastall commands into the blast+ syntax. E.g.:

```
$ legacy_blast.pl megablast -i query.fsa -d nt -o mb.out
   --path /usr/local/ncbi-blast-2.2.27+/bin --print_only
```

```
/usr/local/ncbi-blast-2.2.27+/bin/blastn -query
   query.fsa -db nt -out mb.out
```

`legacy_blast.pl` comes with BLAST+

Getting the Blast Databases

- NCBI puts out a script, update_blastdb.pl to download the databases
  - http://www.ncbi.nlm.nih.gov/BLAST/docs/update_blastdb.pl
- See ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blastdb.html for details

```
$ cd $BLAST/db
$ perl update_blastdb.pl nr nt refseq pdbaa swissprot refseq_protein taxdb
cdd_delta
```

- **You will now have a bunch of *.tar.gz files for each of the dbs downloaded, in order to use you must gunzip and untar**

```
#! /usr/bin/perl
use strict;
use warnings;


my @arr = `ls *gz`;

foreach my $file (@arr){
        chomp $file;
        system("tar -zxvf $file");
        system("rm $file");
}
```

FYI - The BLAST Taxonomy Database
taxdb

- Required in order to print the scientific name, common name, blast name, or super kingdom as part of the BLAST report or in a report with blastdbcmd
- BLAST database contain only the taxid (an integer) for each entry, and the taxonomy database allow BLAST to retrieve the scientific name etc. from a taxid
- The BLAST taxonomy database consists of a pair of files (taxdb.bti and taxdb.btd) that are available as a compressed archive from the NCBI BLAST FTP site (ftp://ftp.ncbi.nlm.nih.gov/blast/db/taxdb.tar.gz)
- update_blastdb.pl script can be used to download and update this archive

```
$ blastdbcmd -entry AC147927  -outfmt "%L::%S::%T" -db nt
legacy version:
$ fastacmd -s AC147927 -d nt -T -t

$ blastdbcmd -entry CAA27203.1  -outfmt "%L::%S::%T" -db nr
legacy version:
$ fastacmd -s CAA27203.1 -d nr -T -t
```

`:: used as a delimiter`

RPS-BLAST & DELTA-BLAST

```
$ deltablast -query test.protein2.fasta -db pdbaa -out test.protein2.fasta.delta.out
```

- DETA-BLAST performs a multiple sequence alignment of the query sequence with domains described in the CDD (Conserved Domain Database from NCBI) database and then uses a PSSM derived from this alignment to search a sequence database

```
$ rpsblast -query test.protein2.fasta -db cdd_delta  -out test.protein.fasta.rps.out
```

- RPS-BLAST (Reverse PSI-BLAST) searches a query sequence against a database of profiles, here the CDD

```
$ perl update_blastdb.pl nr nt refseq pdbaa5 swissprot refseq_protein taxdb cdd_delta
```

Over the Next Couple Weeks

- Go over BLAST Command Line Applications User Manual
  - http://www.ncbi.nlm.nih.gov/books/NBK1763/
- Check out these online tutorials
  - http://evomics.org/learning/bioinformatics/blast-laboratory/
  - http://evomics.org/learning/bioinformatics/fastablast/

For Thursday

- Remember presentations begin Thursday
  - 20-25 minutes presentation
  - Practice your talk before you come
    - Get your timing down
    - Remember
      - breathe slowly when you get to your slides
      - you know more than the audience about the topic
    - 3-4 minutes (max 5) on Introduction
    - 15 Minutes on database (the algorithm, parameters, searching, etc)
    - 5 Minutes on conclusion (what it can be used for, why [**or why not**] you found the database to be useful, what type of improvements would be useful)
  - Read the two papers you are assigned to be peer-review for