Biol5705
Module: Gene Sequence Analysis

Lecture 1
Homology Searching

Dr. Morgan Langille

Outline

- What is homology?
- orthologs, paralogs, etc.
- local vs global alignment
- e values, bit scores, "coverage", identity vs similarity
- different blast flavours (blastn, blastp, tblastn, etc.)
- Blast (Web)

What is homology?

- Homology refers to shared ancestry
- Two sequences are homologous if they are derived from a common ancestral sequence
- 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.

Types of homologs

- Orthologs
  - Think same gene in different organism
  - Often thought to have similar function

- Paralogs
  - Think gene duplication
  - Less likely to have similar function

What is similarity?

- Similarity is a measure of the likeness between sequences.

- Gene searching tools calculate the similarity between sequences and rank more similar sequences higher.

- Sequences can NOT be partially homologous
  - WRONG: Gene X is 80% homologous to Gene Y

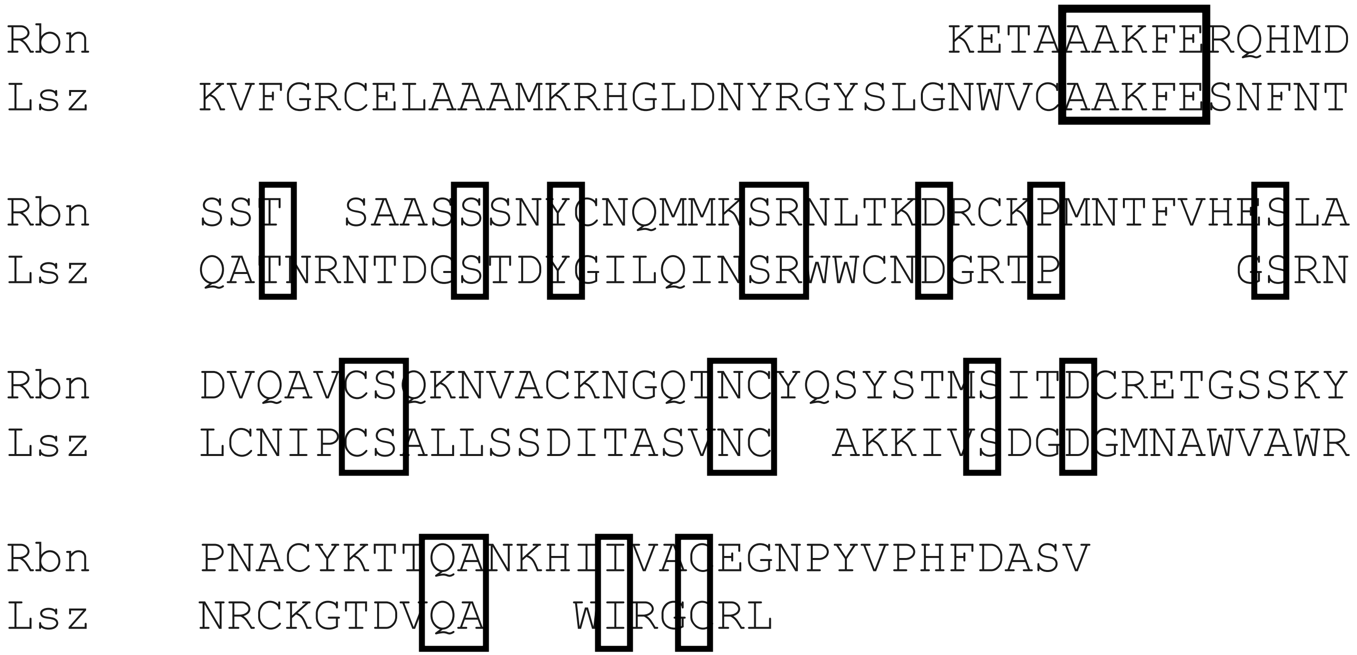
- Sequences can be partially similar
  - CORRECT: Gene X has 80% identity to Gene Y

Identity vs Similarity

- Identity is a percentage measurement that states how many characters in the sequence are identical

- Similarity can also be used as a metric which means how many characters are “positive scoring”

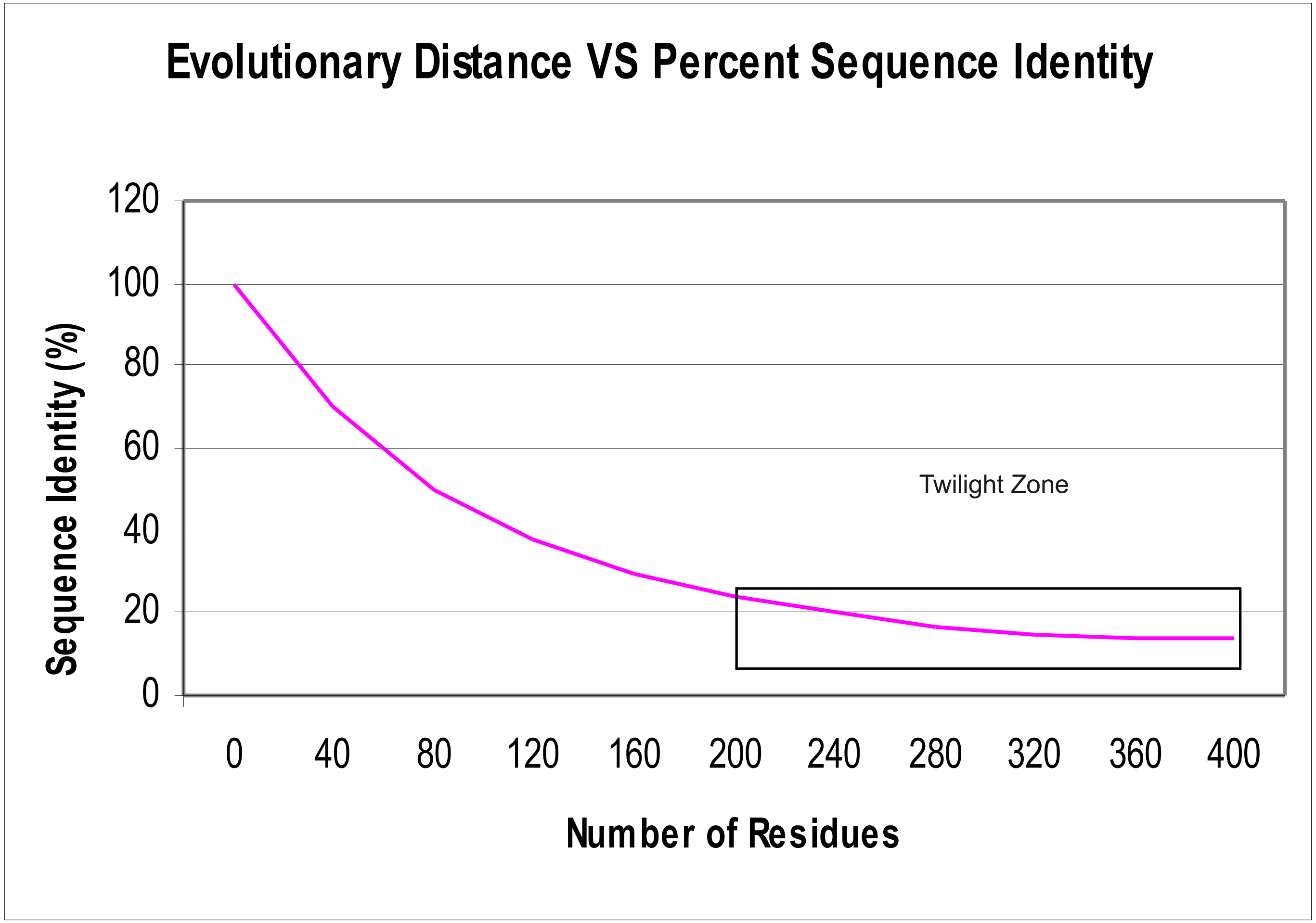
Assessing Sequence Similarity



```
Rbn                               KETAAAKFERQHMD
Lsz     KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNT

Rbn     SST  SAASSSNYCNQMMKSRNLTKDRCKPMNTFVHESLA
Lsz     QATNRNTDGSTDYGILQINSRWWCNDGRTP      GSRN

Rbn     DVQAVCSQKNVACKNGQTNCYQSYSTMSITDCRETGSSKY
Lsz     LCNIPCSALLSSDITASVNC  AKKIVSDGDGMNAWVAWR

Rbn     PNACYKTTQANKHIIVACEGNPYVPHFDASV
Lsz     NRCKGTDVQA   WIRGCRL
```

is this alignment significant?

Twilight Zone



Some Simple Suggestions

- **If two sequence are > 100 residues and > 25% identical, they are likely related**

- **If two sequences are 15-25% identical they may be related, but more tests are needed**

- **If two sequences are < 15% identical they are probably not related**

Global vs Local

- Alignments can be global or local (this is algorithm specific)
    - A global alignment is an optimal alignment that includes all characters from each sequence (Multiple Sequence Alignment)
    - A local alignment is an optimal alignment that includes only the most similar local region or regions (e.g BLAST).

Dot Plots

- **Popular freeware package is Dotter**
  http://sonnhammer.sbc.su.se/Dotter.html

The BLAST algorithm

- The BLAST programs (**B**asic **L**ocal **A**lignment **S**earch **T**ools) 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.

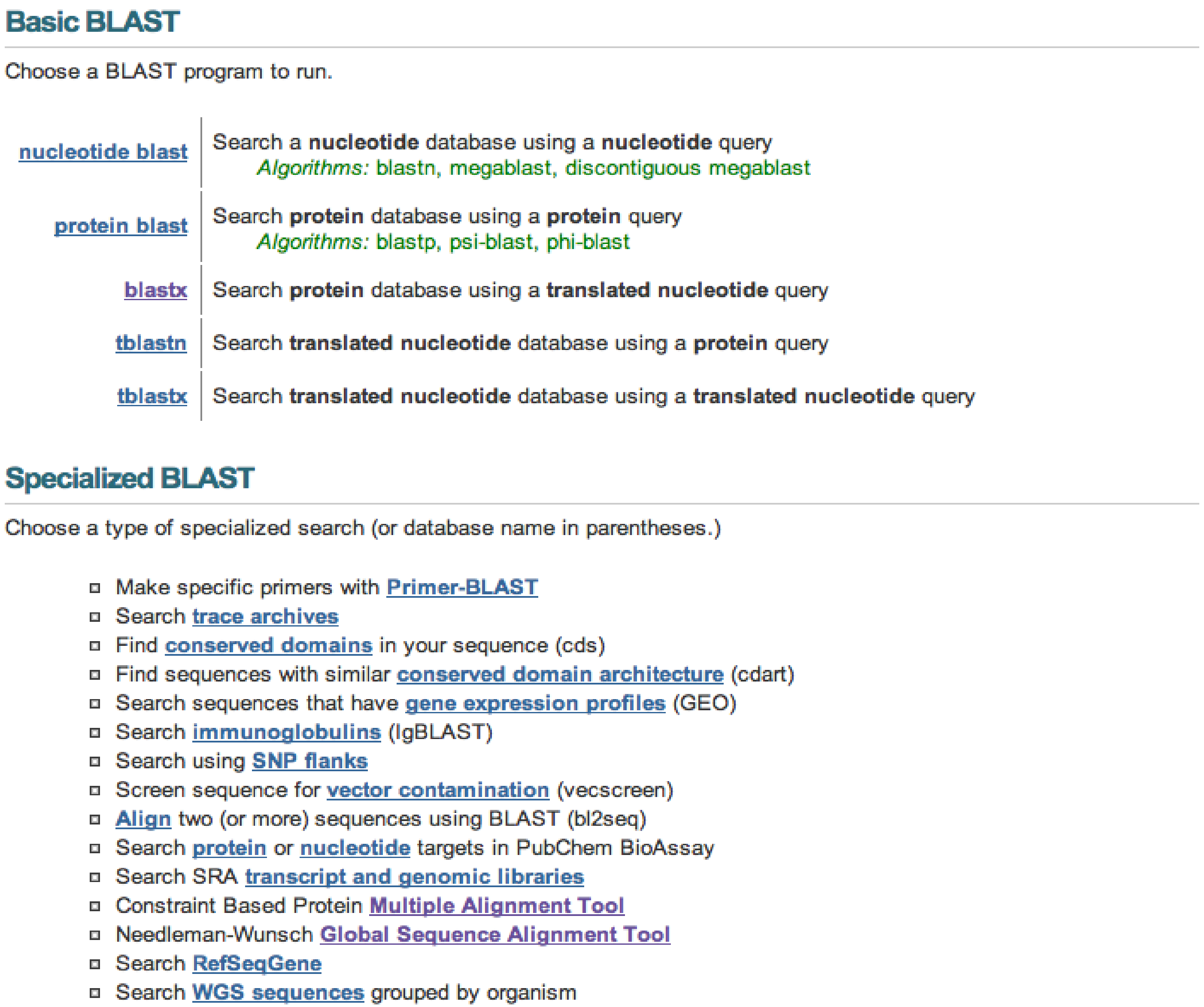
Several different BLAST programs:

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. Please note that the tblastx program cannot be used with the nr database on the BLAST Web page because it is too computationally intensive.

MegaBLAST

- megaBLAST
  - For aligning very similar sequences
  - Nucleotide only
  - Very efficient for long query sequences
  - Uses big word (k-tuple) sizes to start search
    - Very fast

http://www.ncbi.nlm.nih.gov/BLAST/



Basic BLAST

Choose a BLAST program to run.

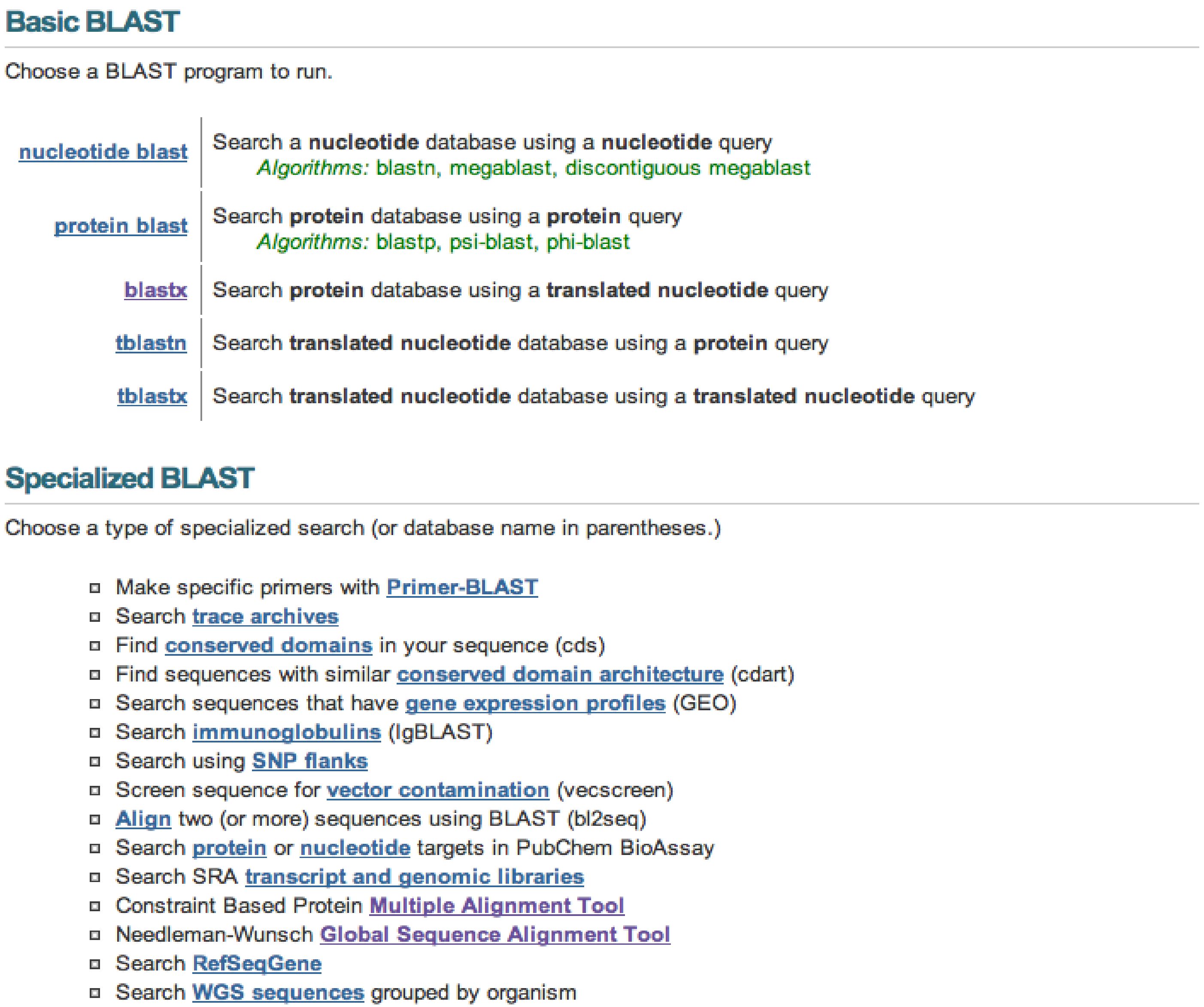
nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with Primer-BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search using SNP flanks
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA transcript and genomic libraries
- Constraint Based Protein Multiple Alignment Tool
- Needleman-Wunsch Global Sequence Alignment Tool
- Search RefSeqGene
- Search WGS sequences grouped by organism

http://www.ncbi.nlm.nih.gov/BLAST/



Basic BLAST

Choose a BLAST program to run.

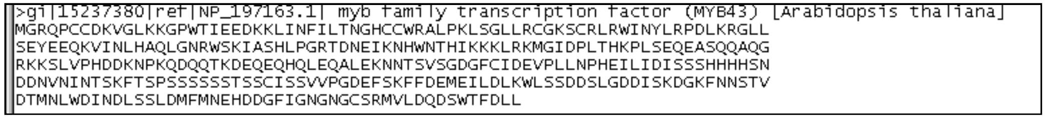
nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

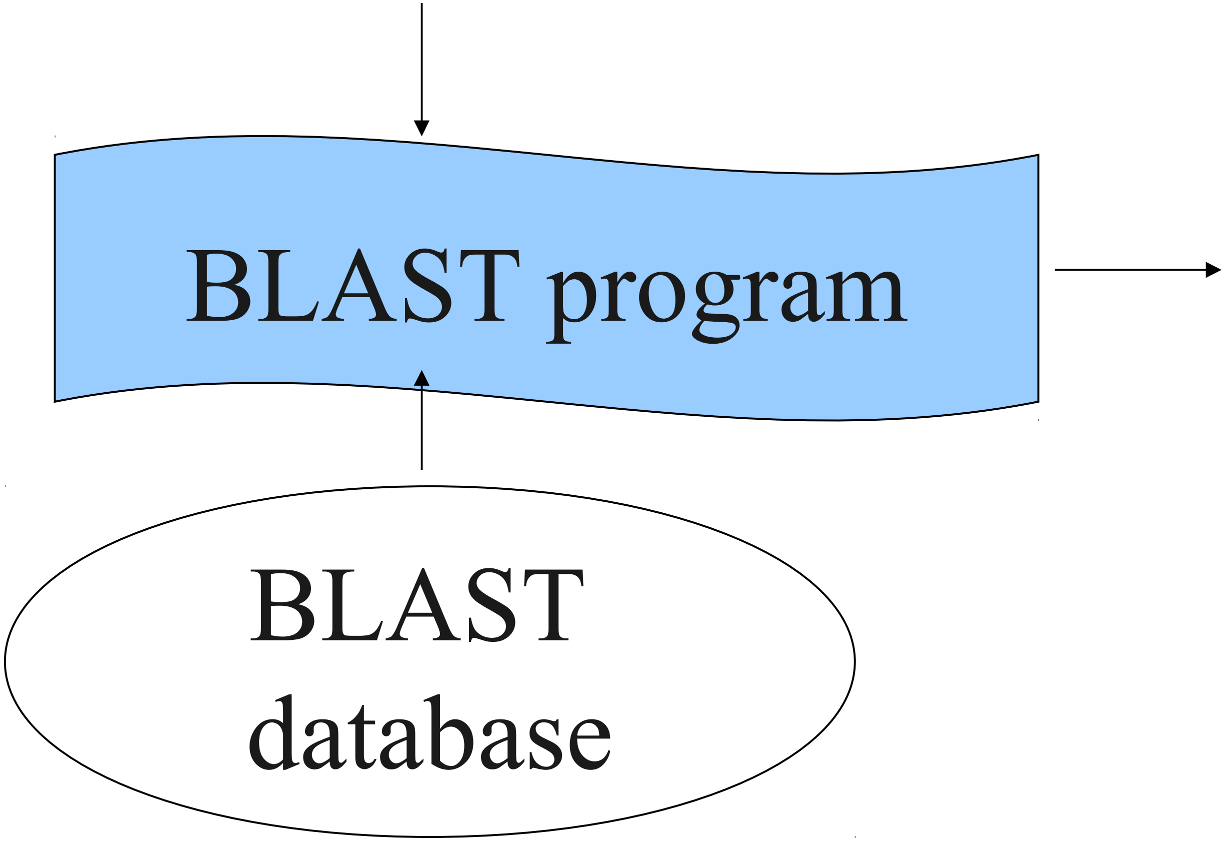
Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with Primer-BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search using SNP flanks
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA transcript and genomic libraries
- Constraint Based Protein Multiple Alignment Tool
- Needleman-Wunsch Global Sequence Alignment Tool
- Search RefSeqGene
- Search WGS sequences grouped by organism

QUERY sequence(s)



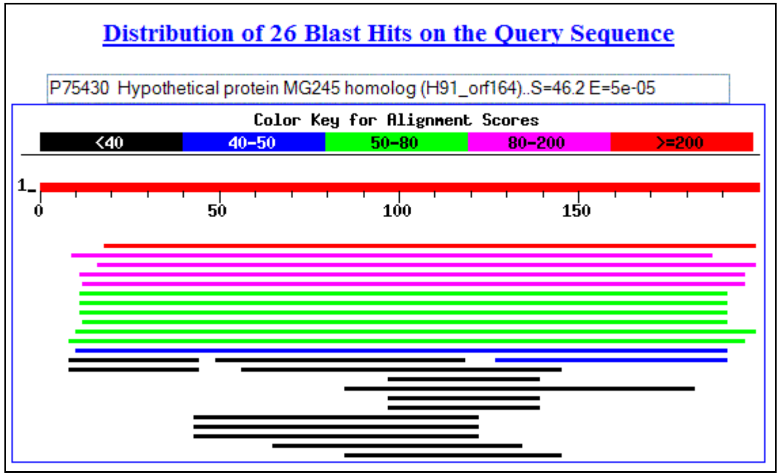
```
>gi|15237380|ref|NP_197163.1| myb family transcription factor (MYB43) [Arabidopsis thaliana]
MGRQPCCDKVGLKKGPWTIEEDKKLINFILTNGHCCWRALPKLSGLLRCGKSCRLRWINYLRPDLKRGLL
SEYEEQKVINLHAQLGNRWSKIASHLPGRTDNEIKNHWNTHIKKKLRKMGIDPLTHKPLSEQEASQQAQG
RKKSLVPHDDKNPKQDQQTKDEQEQHQLEQALEKNNTSVSGDGFCIDEVPLLNPHEILIDISSSHHHHSN
DDNVNINTSKFTSPSSSSSSTSSCISSVVPGDEFSKFFDEMEILDLKWLSSDDSLGDDISKDGKFNNSTV
DTMNLWDINDLSSLDMFMNEHDDGFIGNGNGCSRMVLDQDSWTFDLL
```



BLAST program

BLAST database

BLAST results



Distribution of 26 Blast Hits on the Query Sequence

P75430 Hypothetical protein MG245 homolog (H91_orf164)..S=46.2 E=5e-05

Color Key for Alignment Scores

Considerations for choosing a BLAST database

- First consider your research question:
  - Are you looking for an particular gene in a particular species?
    - BLAST against the genome of that species.
  - Are you looking for additional members of a protein family across all species?
    - BLAST against the non-redudant database (nr), if you can’t find hits check wgs, htgs, and the trace archives.
  - Are you looking to annotate genes in your species of interest?
    - BLAST against known genes (RefSeq) and/or ESTs from a closely related species.

When choosing a database for BLAST…

- Changing your choice of database is changing your search space

- Database size affects the BLAST statistics

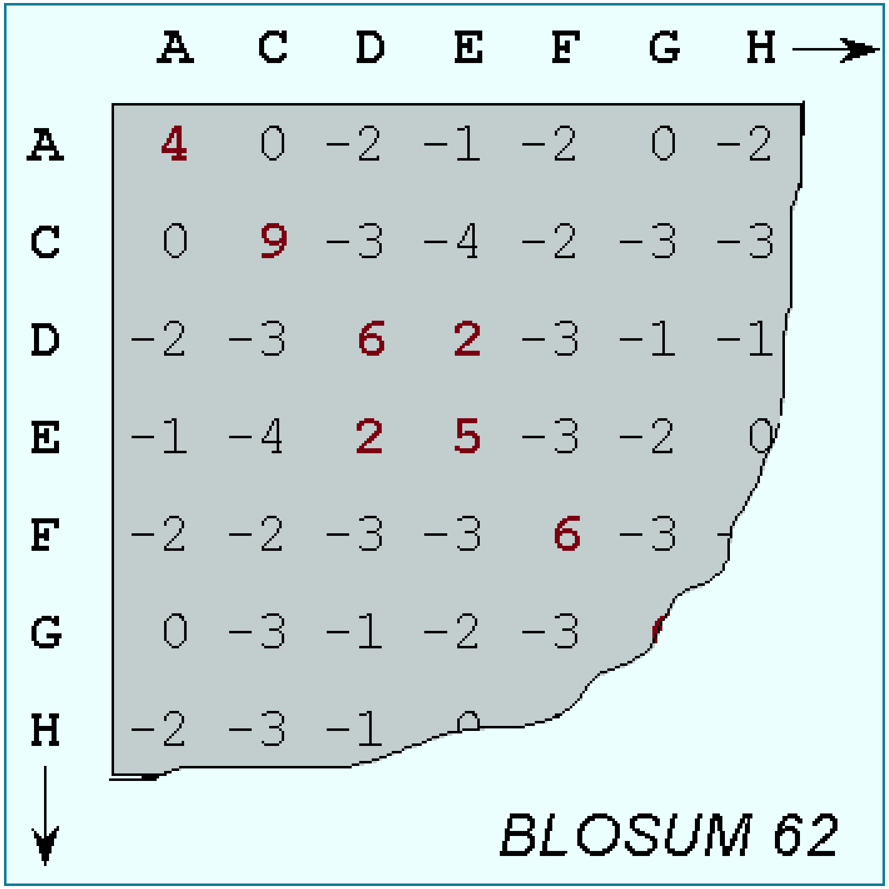
- Databases change rapidly and are updated frequently

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?

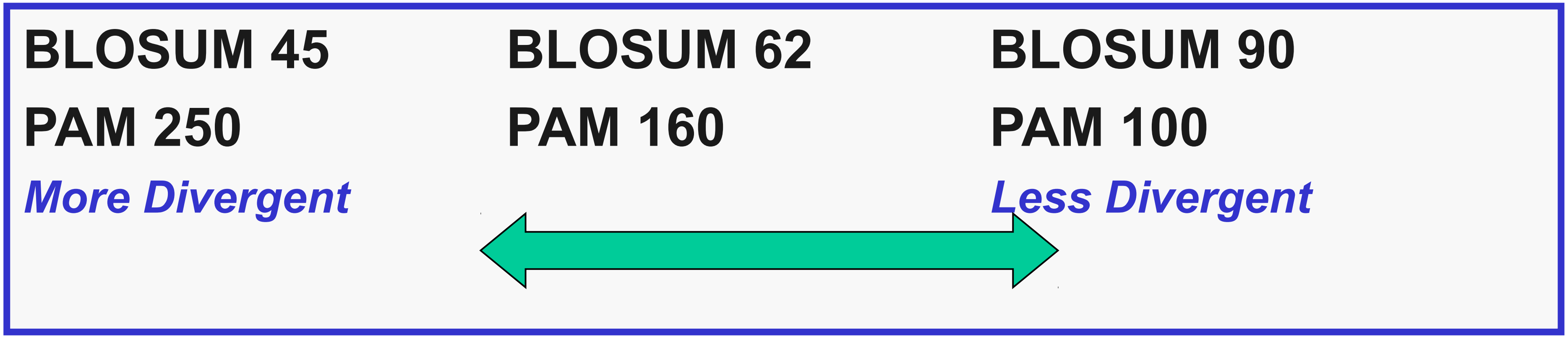
- Substitution matrices are used for amino acid alignments.
  - each possible residue substitution is given a score
- A simpler unitary matrix is used for DNA pairs
  - each position can be given a score of +1 if it matches and a score of -1 if it does not.



	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	-
G	0	-3	-1	-2	-3	6	
H	-2	-3	-1	0			

BLOSUM 62

BLOSUM vs. PAM



BLOSUM 45	**BLOSUM 62**	**BLOSUM 90**
PAM 250	**PAM 160**	**PAM 100**
More Divergent		***Less Divergent***

- BLOSUM 62 is the default matrix in BLAST. 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.

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.

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

- The quality of the alignment is represented by the Score.
  - **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 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.

I’m confused! What does the E-value mean again?

- **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.
- When *E* < 0.01, *P*-values and *E*-value are nearly identical.
  - So, the E-value is the number of times you expect to see your hit occur in the database (with as good as or better score) due to random chance alone.

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

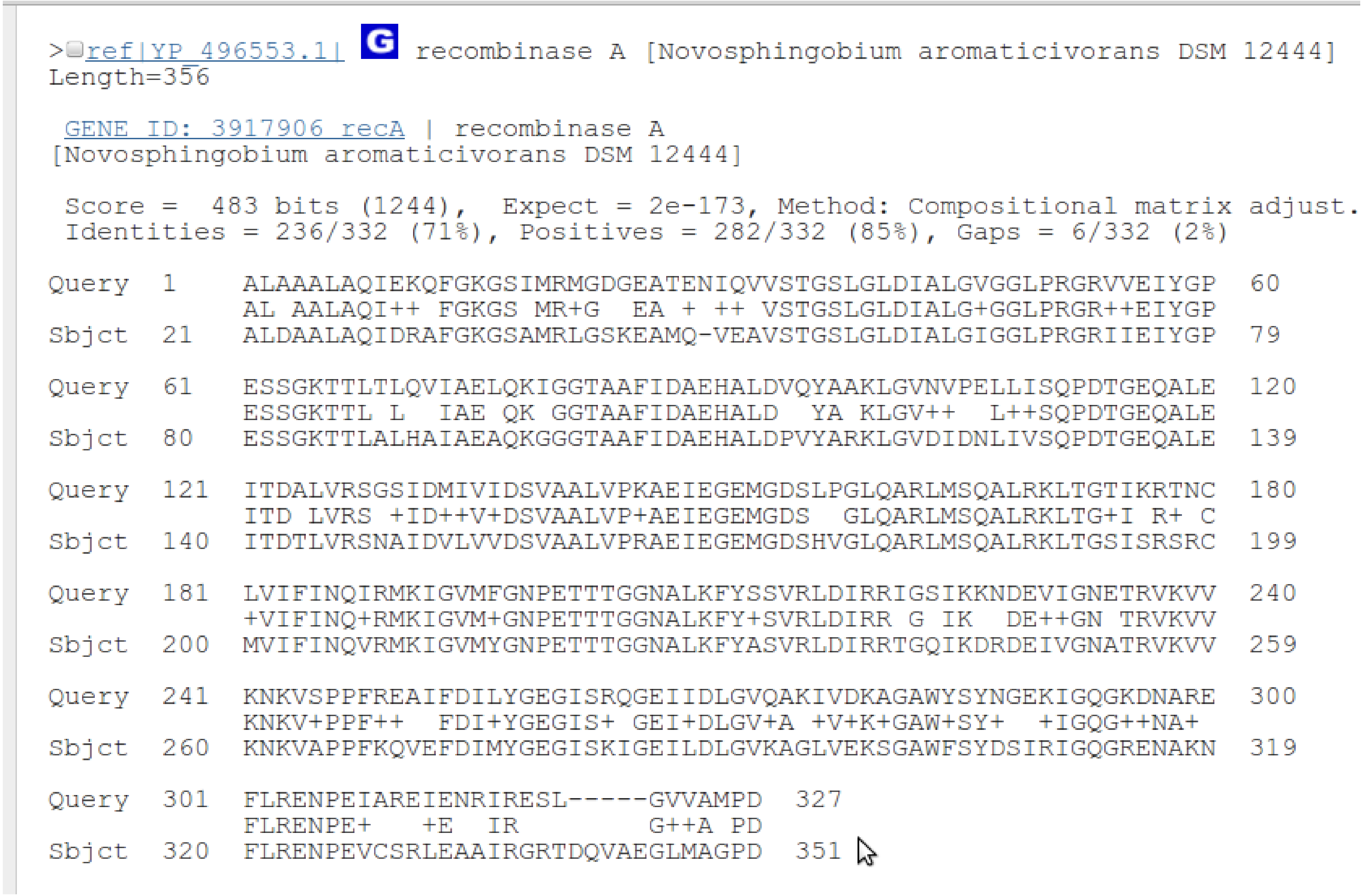
Coverage

- Coverage: The proportion of the aligned length with respect to the length of the query or subject.
- Example
  - Your gene is 1000bp, and you have a Blast alignment from 250-500. What is the query coverage?

FASTA File Format

- Plain text file (e.g. don't open with Word!)
- Each sequence has 2 parts.
  - One header line starts with “>”
    - e.g. “>This is a fasta header. Any text goes here.”
  - One or more sequence lines:
    - e.g. “ATTCTCGCTCGAATCGATCGCATAGTAGCA”
- Each file can contain multiple sequences
- Sequences can be DNA or protein (not a mixture)
  -

Alignments



```
>ref|YP_496553.1| G recombinase A [Novosphingobium aromaticivorans DSM 12444]
Length=356

 GENE ID: 3917906 recA | recombinase A
[Novosphingobium aromaticivorans DSM 12444]

 Score =  483 bits (1244),  Expect = 2e-173, Method: Compositional matrix adjust.
 Identities = 236/332 (71%), Positives = 282/332 (85%), Gaps = 6/332 (2%)

Query  1    ALAAALAQIEKQFGKGSIMRMGDGEATENIQVVSTGSLGLDIALGVGGLPRGRVVEIYGP  60
            AL AALAQI++ FGKGS MR+G  EA + ++ VSTGSLGLDIALG+GGLPRGR++EIYGP
Sbjct  21   ALDAALAQIDRAFGKGSAMRLGSKEAMQ-VEAVSTGSLGLDIALGIGGLPRGRIIEIYGP  79

Query  61   ESSGKTTLTLQVIAELQKIGGTAAFIDAEHALDVQYAAKLGVNVPELLISQPDTGEQALE  120
            ESSGKTTL L  IAE QK GGTAAFIDAEHALD  YA KLGV++  L++SQPDTGEQALE
Sbjct  80   ESSGKTTLALHAIAEAQKGGGTAAFIDAEHALDPVYARKLGVDIDNLIVSQPDTGEQALE  139

Query  121  ITDALVRSGSIDMIVIDSVAALVPKAEIEGEMGDSLPGLQARLMSQALRKLTGTIKRTNC  180
            ITD LVRS +ID++V+DSVAALVP+AEIEGEMGDS  GLQARLMSQALRKLTG+I R+ C
Sbjct  140  ITDTLVRSNAIDVLVVDSVAALVPRAEIEGEMGDSHVGLQARLMSQALRKLTGSISRSRC  199

Query  181  LVIFINQIRMKIGVMFGNPETTTGGNALKFYSSVRLDIRRIGSIKKNDEVIGNETRVKVV  240
            +VIFINQ+RMKIGVM+GNPETTTGGNALKFY+SVRLDIRR G IK  DE++GN TRVKVV
Sbjct  200  MVIFINQVRMKIGVMYGNPETTTGGNALKFYASVRLDIRRTGQIKDRDEIVGNATRVKVV  259

Query  241  KNKVSPPFREAIFDILYGEGISRQGEIIDLGVQAKIVDKAGAWYSYNGEKIGQGKDNARE  300
            KNKV+PPF++  FDI+YGEGIS+ GEI+DLGV+A +V+K+GAW+SY+  +IGQG++NA+ 
Sbjct  260  KNKVAPPFKQVEFDIMYGEGISKIGEILDLGVKAGLVEKSGAWFSYDSIRIGQGRENAKN  319

Query  301  FLRENPEIAREIENRIRESL-----GVVAMPD  327
            FLRENPE+   +E  IR        G++A PD
Sbjct  320  FLRENPEVCSRLEAAIRGRTDQVAEGLMAGPD  351
```

Databases

- NR “non-redundant” database
  - Sequences from various experiments (not just completed genomes)
  - May not be that “non-redundant”
- RefSeq
  - Curated sequences by NCBI
  - Does not contain duplicates
- Swissprot
  - A manually curated sequence of proteins
- Protein Data Bank
  - Contains protein sequences that have 3D structures available

Blast Web Demo

- Assignment 1
  - http://morganlangille.com/teaching/biol5705/assignment1.pdf
- Due before next class