LAB 2 — ADVANCED BLAST AND COMPARATIVE GENOMICS

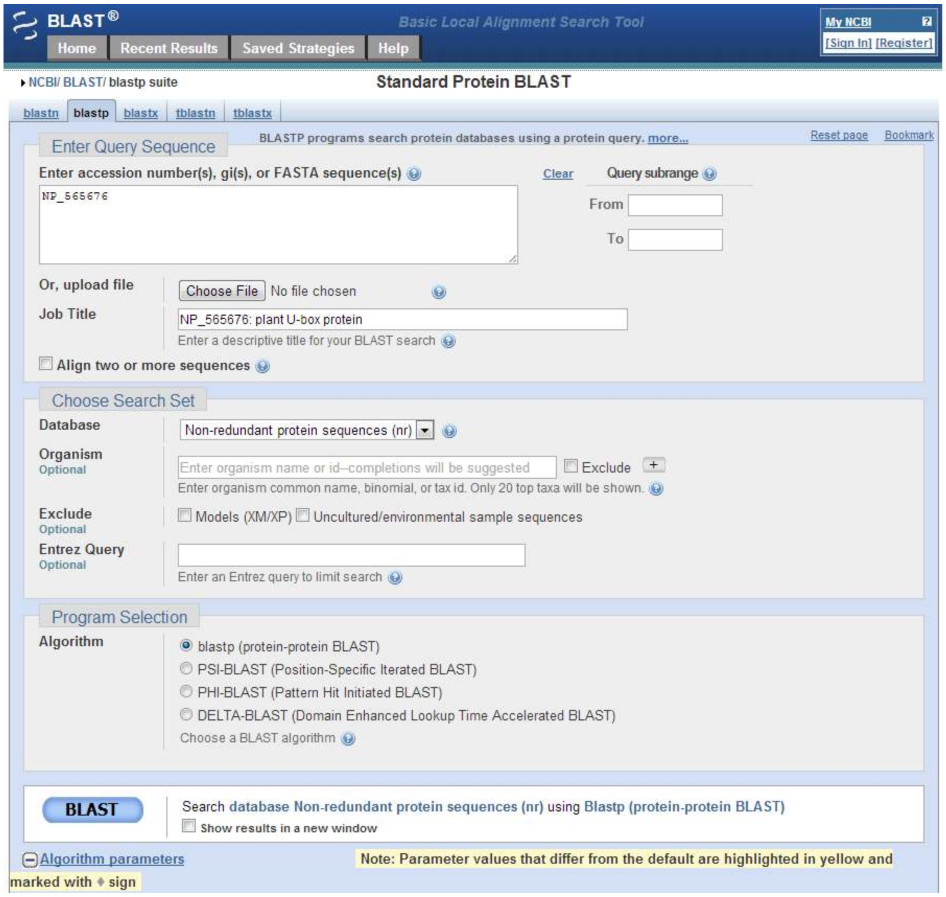
[Software needed: web access]

There are 4 sections to this lab: BlastP, PSI-Blast, Translated Blast, and Comparative Genomics. Last time we used BLAST to query a nucleotide sequence against the NCBI nr database. Now let's search using a protein sequence.

BLASTP

1. Go to NCBI (www.ncbi.nlm.nih.gov) and select **BLAST** from the Popular Resource section on the right.

2. Choose **protein blast** from the Basic BLAST section.



Standard Protein BLAST

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

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s): NP_565676

Query subrange: From ___ To ___

Or, upload file: No file chosen

Job Title: NP_565676: plant U-box protein

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

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

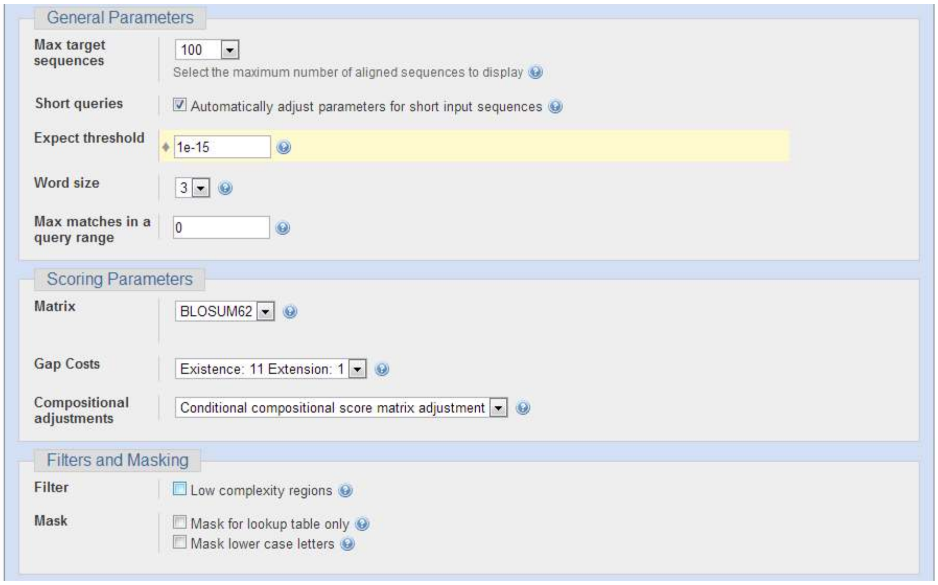
Entrez Query (Optional): Enter an Entrez query to limit search

Program Selection

Algorithm:
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
□ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign



General Parameters

Max target sequences: 100 ▾ Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 1e-15

Word size: 3 ▾

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62 ▾

Gap Costs: Existence: 11 Extension: 1 ▾

Compositional adjustments: Conditional compositional score matrix adjustment ▾

Filters and Masking

Filter: □ Low complexity regions

Mask: □ Mask for lookup table only
□ Mask lower case letters

Figure 1. The blastp input page.

3. Choose **blastp** from in the program selection tabs along the top.

4. Load your accession number (NP_565676) that you used in Lab1 into the BLAST search box. Note how BLAST automatically adjusts your default settings to search a protein database.

5. Open the **Algorithm parameters** section. Lower the expect threshold from 10 to 1e-15.
    a. *Based on what we learned in the last lab, why might we decide to do this?*

6. In the S**coring Parameters** section, the default substitution matrix is **BLOSUM62**. Change the substitution matrix to **BLOSUM80.** We will discuss substitution matrices in more detail later.

7. Look at the **Gap Costs**. **Existence** refers to creating a gap in the alignment, while **Extension** refers to extending a gap in an alignment.
    a. *Why is the former penalized more than the latter? Here it helps to think about what a gap (or insertion, from the other sequence's perspective) might mean in terms of the gene product's protein structure. If a small loop is "allowed" (structurally) to be inserted in a region, then do you think a slightly larger (extended) loop might be permissible, too?*

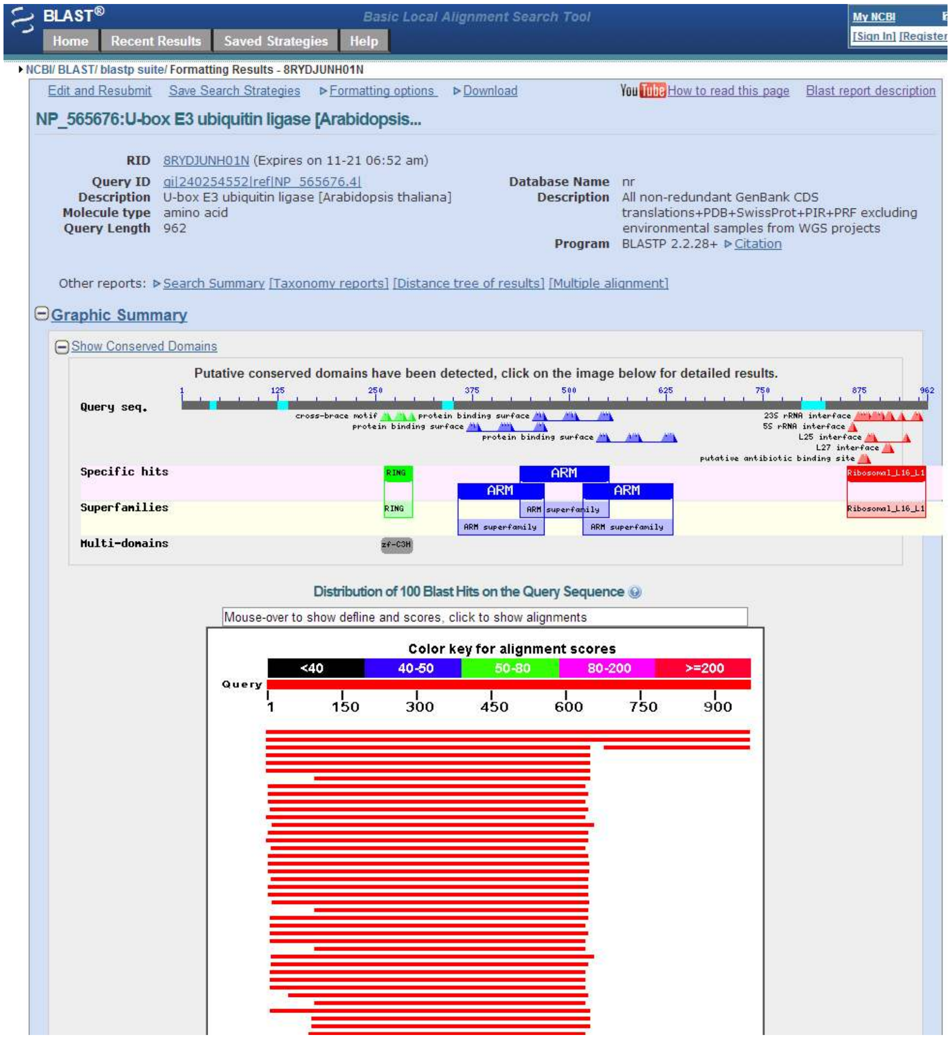
8. Check '**Low complexity regions**' filter.
    a. *Why might you want to use this filter? Don't forget to use the help icons if you need more information on any of the parameters.*

9. Check '**Show results in a new window**'.

10. Click **BLAST!**

- The first window that is likely to come up tells you the job is being processed and you should wait. You may also see information on any conserved domains found in your sequence. You can click on the schematic within the **Show Conserved Domains** box to be taken into the conserved domain database. We will examine this in more detail in a later lab.

11. You will see that the format of the BLASTP output is very similar to that seen with BLASTN. The page is broken up into:
    - Job summary
    - Graphical summary containing
        - Conserved domains
        - Graphic summary
        - Descriptions
        - Alignments



▸ NCBI/ BLAST/ blastp suite/ Formatting Results - 8RYDJUNH01N

Edit and Resubmit   Save Search Strategies   ▷ Formatting options   ▷ Download   YouTube How to read this page   Blast report description

NP_565676:U-box E3 ubiquitin ligase [Arabidopsis...

RID	8RYDJUNH01N (Expires on 11-21 06:52 am)		
Query ID	gi|240254552|ref|NP_565676.4|	Database Name	nr
Description	U-box E3 ubiquitin ligase [Arabidopsis thaliana]	Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	amino acid		
Query Length	962		
		Program	BLASTP 2.2.28+ ▷ Citation

Other reports: ▷ Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

⊟ Graphic Summary

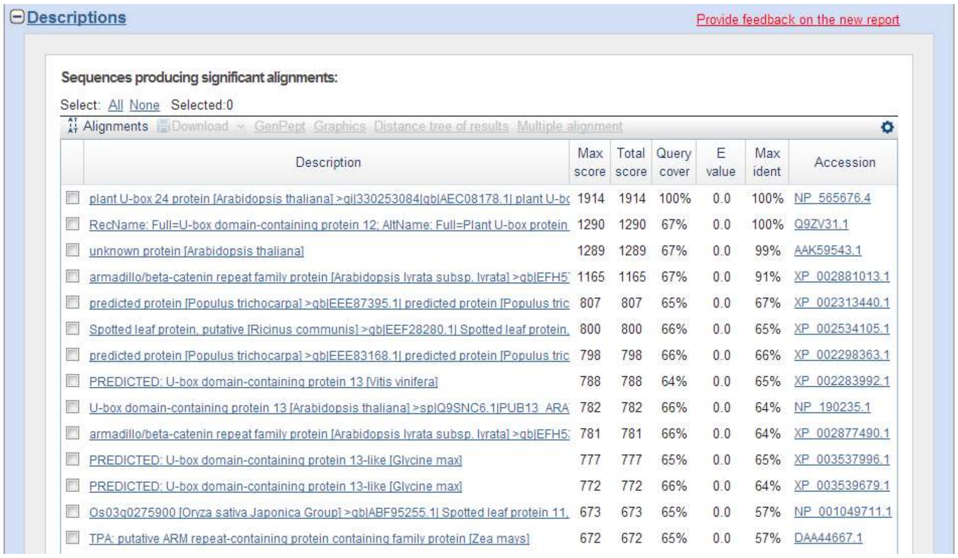
⊟ Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

…image trimmed…



Descriptions

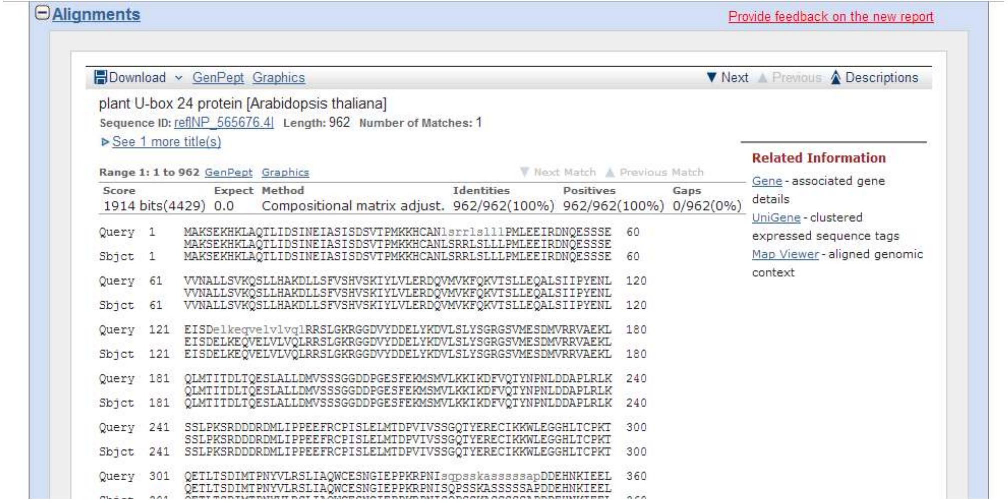
Provide feedback on the new report

Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
□	plant U-box 24 protein [Arabidopsis thaliana] >gi|330253084|gb|AEC08178.1| plant U-b(	1914	1914	100%	0.0	100%	NP_565676.4
□	RecName: Full=U-box domain-containing protein 12; AltName: Full=Plant U-box protein	1290	1290	67%	0.0	100%	Q9ZV31.1
□	unknown protein [Arabidopsis thaliana]	1289	1289	67%	0.0	99%	AAK59543.1
□	armadillo/beta-catenin repeat family protein [Arabidopsis lyrata subsp. lyrata] >gb|EFH5	1165	1165	67%	0.0	91%	XP_002881013.1
□	predicted protein [Populus trichocarpa] >gb|EEE87395.1| predicted protein [Populus tric	807	807	65%	0.0	67%	XP_002313440.1
□	Spotted leaf protein, putative [Ricinus communis] >gb|EEF28280.1| Spotted leaf protein,	800	800	66%	0.0	65%	XP_002534105.1
□	predicted protein [Populus trichocarpa] >gb|EEE83168.1| predicted protein [Populus tric	798	798	66%	0.0	66%	XP_002298363.1
□	PREDICTED: U-box domain-containing protein 13 [Vitis vinifera]	788	788	64%	0.0	65%	XP_002283992.1
□	U-box domain-containing protein 13 [Arabidopsis thaliana] >sp|Q9SNC6.1|PUB13_ARA	782	782	66%	0.0	64%	NP_190235.1
□	armadillo/beta-catenin repeat family protein [Arabidopsis lyrata subsp. lyrata] >gb|EFH5	781	781	66%	0.0	64%	XP_002877490.1
□	PREDICTED: U-box domain-containing protein 13-like [Glycine max]	777	777	65%	0.0	65%	XP_003537996.1
□	PREDICTED: U-box domain-containing protein 13-like [Glycine max]	772	772	66%	0.0	64%	XP_003539679.1
□	Os03g0275900 [Oryza sativa Japonica Group] >gb|ABF95255.1| Spotted leaf protein 11,	673	673	65%	0.0	57%	NP_001049711.1
□	TPA: putative ARM repeat-containing protein containing family protein [Zea mays]	672	672	65%	0.0	57%	DAA44667.1

…image trimmed…



Alignments

Provide feedback on the new report

plant U-box 24 protein [Arabidopsis thaliana]

Sequence ID: ref|NP_565676.4| Length: 962 Number of Matches: 1

▷ See 1 more title(s)

Related Information

Gene - associated gene details

UniGene - clustered expressed sequence tags

Map Viewer - aligned genomic context

Range 1: 1 to 962 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
1914 bits(4429)	0.0	Compositional matrix adjust.	962/962(100%)	962/962(100%)	0/962(0%)

```
Query  1    MAKSEKHKLAQTLIDSINEIASISDSVTPMKKHCANlsrrlslllPMLEEIRDNQESSSE  60
            MAKSEKHKLAQTLIDSINEIASISDSVTPMKKHCANLSRRLSLLLPMLEEIRDNQESSSE
Sbjct  1    MAKSEKHKLAQTLIDSINEIASISDSVTPMKKHCANLSRRLSLLLPMLEEIRDNQESSSE  60

Query  61   VVNALLSVKQSLLHAKDLLSFVSHVSKIYLVLERDQVMVKFQKVTSLLEQALSIIPYENL  120
            VVNALLSVKQSLLHAKDLLSFVSHVSKIYLVLERDQVMVKFQKVTSLLEQALSIIPYENL
Sbjct  61   VVNALLSVKQSLLHAKDLLSFVSHVSKIYLVLERDQVMVKFQKVTSLLEQALSIIPYENL  120

Query  121  EISDelkeqvelvlvqlRRSLGKRGGDVYDDELYKDVLSLYSGRGSVMESDMVRRVAEKL  180
            EISDELKEQVELVLVQLRRSLGKRGGDVYDDELYKDVLSLYSGRGSVMESDMVRRVAEKL
Sbjct  121  EISDELKEQVELVLVQLRRSLGKRGGDVYDDELYKDVLSLYSGRGSVMESDMVRRVAEKL  180

Query  181  QLMTITDLTQESLALLDMVSSSGGDDPGESFEKMSMVLKKIKDFVQTYNPNLDDAPLRLK  240
            QLMTITDLTQESLALLDMVSSSGGDDPGESFEKMSMVLKKIKDFVQTYNPNLDDAPLRLK
Sbjct  181  QLMTITDLTQESLALLDMVSSSGGDDPGESFEKMSMVLKKIKDFVQTYNPNLDDAPLRLK  240

Query  241  SSLPKSRDDDRDMLIPPEEFRCPISLELMTDPVIVSSGQTYERECIKKWLEGGHLTCPKT  300
            SSLPKSRDDDRDMLIPPEEFRCPISLELMTDPVIVSSGQTYERECIKKWLEGGHLTCPKT
Sbjct  241  SSLPKSRDDDRDMLIPPEEFRCPISLELMTDPVIVSSGQTYERECIKKWLEGGHLTCPKT  300

Query  301  QETLTSDIMTPNYVLRSLIAQWCESNGIEPPKRPNIsqpsskasssssapDDEHNKIEEL  360
            QETLTSDIMTPNYVLRSLIAQWCESNGIEPPKRPNISQPSSKASSSSSAPDDEHNKIEEL
```

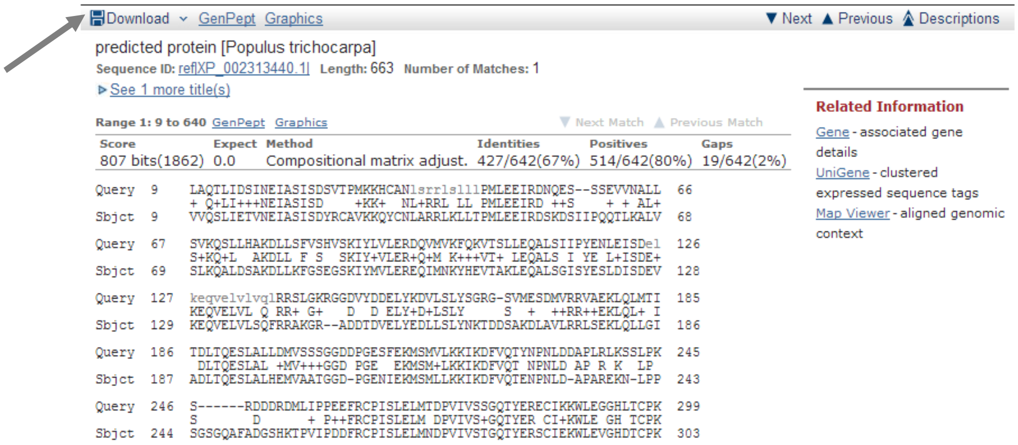
Figure 2. BLASTP output for NP_565676 search

12. Let’s try a different format for our results. Click on **Formatting options** from near the very top of the page, and pick **Pairwise with dots for identities** in the **Alignment View** box. Click **Reformat**.

13. Move to the graphic summary section and mouse over the summaries. The information for each hit will be displayed in the box just above the graphical summary. Note the identities and E-value scores of the different hits. The top ~100 hits will be summarized in this section. Scroll to the last hit in the display.
    a. *What is its E-value?*
    b. *Do you consider this a good hit (based on the lecture)?*

Lab Quiz Question 1

Pairwise format



predicted protein [Populus trichocarpa]

Sequence ID: ref|XP_002313440.1| Length: 663 Number of Matches: 1

▷ See 1 more title(s)

Range 1: 9 to 640 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
807 bits(1862)	0.0	Compositional matrix adjust.	427/642(67%)	514/642(80%)	19/642(2%)

```
Query  9    LAQTLIDSINEIASISDSVTPMKKHCANlsrrlslllPMLEEIRDNQES--SSEVVNALL  66
            + Q+LI+++NEIASISD    +KK+  NL+RRL LL PMLEEIRD ++S    + + AL+
Sbjct  9    VVQSLIETVNEIASISDYRCAVKKQYCNLARRLKLLTPMLEEIRDSKDSIIPQQTLKALV  68

Query  67   SVKQSLLHAKDLLSFVSHVSKIYLVLERDQVMVKFQKVTSLLEQALSIIPYENLEISDel  126
            S+KQ+L  AKDLL F S  SKIY+VLER+Q+M K+++VT+ LEQALS I YE L+ISDE+
Sbjct  69   SLKQALDSAKDLLKFGSEGSKIYMVLEREQIMNKYHEVTAKLEQALSGISYESLDISDEV  128

Query  127  keqvelvlvqlRRSLGKRGGDVYDDELYKDVLSLYSGRG-SVMESDMVRRVAEKLQLMTI  185
            KEQVELVL Q RR+ G+   D  D ELY+D+LSLY     S  +  ++RR++EKLQL+ I
Sbjct  129  KEQVELVLSQFRRAKGR--ADDTDVELYEDLLSLYNKTDDSAKDLAVLRRLSEKLQLLGI  186

Query  186  TDLTQESLALLDMVSSSGGDDPGESFEKMSMVLKKIKDFVQTYNPNLDDAPLRLKSSLPK  245
             DLTQESLAL +MV+++GGD PGE  EKMSM+LKKIKDFVQT NPNLD AP R K  LP
Sbjct  187  ADLTQESLALHEMVAATGGD-PGENIEKMSMLLKKIKDFVQTENPNLD-APAREKN-LPP  243

Query  246  S------RDDDRDMLIPPEEFRCPISLELMTDPVIVSSGQTYERECIKKWLEGGHLTCPK  299
            S       D      + P++FRCPISLELM DPVIVS+GQTYER CI+KWLE GH TCPK
Sbjct  244  SGSGQAFADGSHKTPVIPDDFRCPISLELMNDPVIVSTGQTYERSCIEKWLEVGHDTCPK  303
```

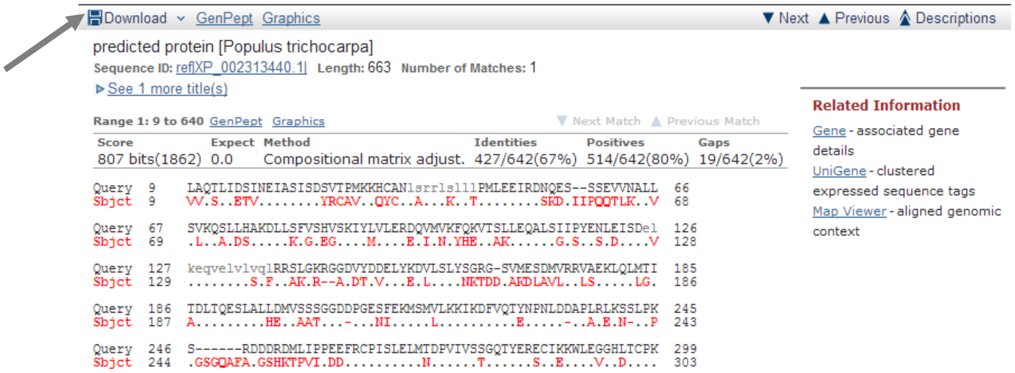
Related Information

Gene - associated gene details

UniGene - clustered expressed sequence tags

Map Viewer - aligned genomic context

Pairwise with dots for identities format



predicted protein [Populus trichocarpa]

Sequence ID: ref|XP_002313440.1| Length: 663 Number of Matches: 1

▷ See 1 more title(s)

Range 1: 9 to 640 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
807 bits(1862)	0.0	Compositional matrix adjust.	427/642(67%)	514/642(80%)	19/642(2%)

```
Query  9    LAQTLIDSINEIASISDSVTPMKKHCANlsrrlslllPMLEEIRDNQES--SSEVVNALL  66
Sbjct  9    VV.S..ETV........YRCAV..QYC..A...K..T........SKD.IIPQQTLK..V  68

Query  67   SVKQSLLHAKDLLSFVSHVSKIYLVLERDQVMVKFQKVTSLLEQALSIIPYENLEISDel  126
Sbjct  69   .L..A.DS.....K.G.EG....M....E.I.N.YHE..AK......G.S..S.D....V  128

Query  127  keqvelvlvqlRRSLGKRGGDVYDDELYKDVLSLYSGRG-SVMESDMVRRVAEKLQLMTI  185
Sbjct  129  ........S.F..AK.R--A.DT.V...E.L....NKTDD.AKDLAVL..LS.....LG.  186

Query  186  TDLTQESLALLDMVSSSGGDDPGESFEKMSMVLKKIKDFVQTYNPNLDDAPLRLKSSLPK  245
Sbjct  187  A.........HE..AAT...-...NI.....L..........E.....-..A.E.N-..P  243

Query  246  S------RDDDRDMLIPPEEFRCPISLELMTDPVIVSSGQTYERECIKKWLEGGHLTCPK  299
Sbjct  244  .GSGQAFA.GSHKTPVI.DD..........N......T......S..E....V..D....  303
```

Related Information

Gene - associated gene details

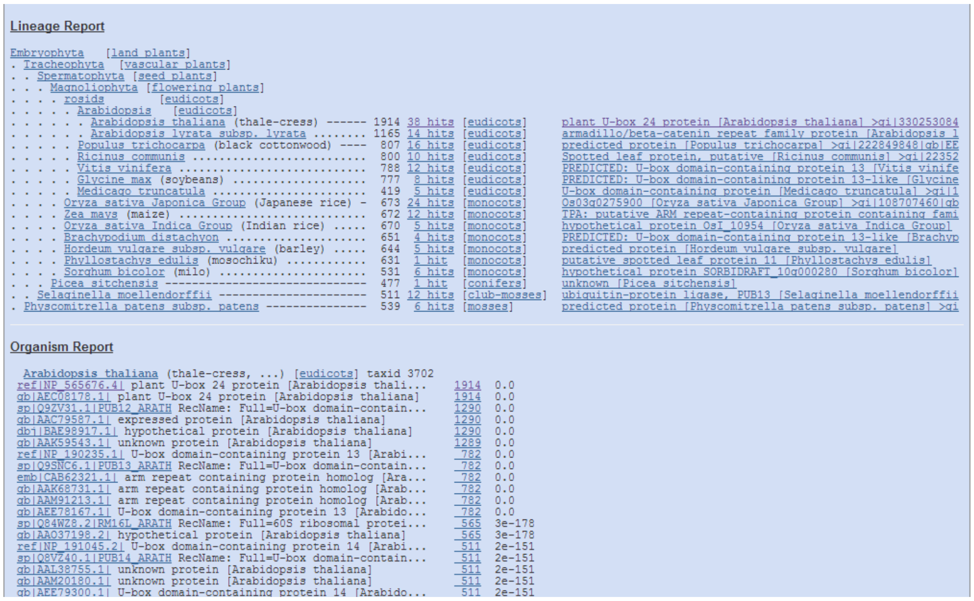
UniGene - clustered expressed sequence tags

Map Viewer - aligned genomic context

Figure 3. Two different formats of BLASTP alignments from NP_565676 search. Arrows show the useful Download feature, for downloading either the aligned part of the Subject sequence, or the entire Subject sequence. Downloads of multiple sequences can be done from the Descriptions Table too. This will be useful for the Assignment!

14. Clicking on the last hit in the graphical summary will take you to the appropriate HSP alignment. Note that there's a substantial amount of variation between your query and this database sequence. The **Pairwise with dots for identities** format makes this particularly clear as all the variable sites are noted in red letters.

15. At the bottom on the **Job Summary** section (top of the page) you will see a link to **Taxonomy reports.** Click on this. Look through the list of hits, for each subsection.
    a. *How is this list for the first subsection (Lineage) organized?*
    b. *What do the numbers between the name of the species and the number of hits mean (e.g. 1914 in the case of Arabidopsis thaliana in Figure 4)? A couple of clicks and you should be able to figure this out!*



Lineage Report

```
Embryophyta   [land plants]
. Tracheophyta  [vascular plants]
. . Spermatophyta [seed plants]
. . . Magnoliophyta [flowering plants]
. . . . rosids        [eudicots]
. . . . . Arabidopsis   [eudicots]
. . . . . . Arabidopsis thaliana (thale-cress) ------ 1914 38 hits [eudicots]     plant U-box 24 protein [Arabidopsis thaliana] >gi|330253084
. . . . . . Arabidopsis lyrata subsp. lyrata ........ 1165 14 hits [eudicots]     armadillo/beta-catenin repeat family protein [Arabidopsis l
. . . . . Populus trichocarpa (black cottonwood) ----  807 16 hits [eudicots]     predicted protein [Populus trichocarpa] >gi|222849848|gb|EE
. . . . . Ricinus communis ..........................  800 10 hits [eudicots]     Spotted leaf protein, putative [Ricinus communis] >gi|22352
. . . . . Vitis vinifera ............................  788 12 hits [eudicots]     PREDICTED: U-box domain-containing protein 13 [Vitis vinife
. . . . . Glycine max (soybeans) ....................  777  8 hits [eudicots]     PREDICTED: U-box domain-containing protein 13-like [Glycine
. . . . . Medicago truncatula .......................  419  5 hits [eudicots]     U-box domain-containing protein [Medicago truncatula] >gi|1
. . . . Oryza sativa Japonica Group (Japanese rice) -  673 24 hits [monocots]     Os03g0275900 [Oryza sativa Japonica Group] >gi|108707460|gb
. . . . Zea mays (maize) ............................  672 12 hits [monocots]     TPA: putative ARM repeat-containing protein containing fami
. . . . Oryza sativa Indica Group (Indian rice) .....  670  5 hits [monocots]     hypothetical protein OsI_10954 [Oryza sativa Indica Group]
. . . . Brachypodium distachyon .....................  651  4 hits [monocots]     PREDICTED: U-box domain-containing protein 13-like [Brachyp
. . . . Hordeum vulgare subsp. vulgare (barley) .....  644  5 hits [monocots]     predicted protein [Hordeum vulgare subsp. vulgare]
. . . . Phyllostachys edulis (mosochiku) ............  631  1 hit  [monocots]     putative spotted leaf protein 11 [Phyllostachys edulis]
. . . . Sorghum bicolor (milo) ......................  531  6 hits [monocots]     hypothetical protein SORBIDRAFT_10g000280 [Sorghum bicolor]
. . . Picea sitchensis ------------------------------  477  1 hit  [conifers]     unknown [Picea sitchensis]
. . Selaginella moellendorffii ----------------------  511 12 hits [club-mosses]  ubiquitin-protein ligase, PUB13 [Selaginella moellendorffii
. Physcomitrella patens subsp. patens ---------------  539  6 hits [mosses]       predicted protein [Physcomitrella patens subsp. patens] >gi
```

Organism Report

```
  Arabidopsis thaliana (thale-cress, ...) [eudicots] taxid 3702
 ref|NP_565676.4| plant U-box 24 protein [Arabidopsis thali...    1914  0.0
 gb|AEC08178.1| plant U-box 24 protein [Arabidopsis thaliana]     1914  0.0
 sp|Q9ZV31.1|PUB12_ARATH RecName: Full=U-box domain-contain...    1290  0.0
 gb|AAC79587.1| expressed protein [Arabidopsis thaliana]          1290  0.0
 dbj|BAE98917.1| hypothetical protein [Arabidopsis thaliana]      1290  0.0
 gb|AAK59543.1| unknown protein [Arabidopsis thaliana]            1289  0.0
 ref|NP_190235.1| U-box domain-containing protein 13 [Arabi...     782  0.0
 sp|Q9SNC6.1|PUB13_ARATH RecName: Full=U-box domain-contain...     782  0.0
 emb|CAB62321.1| arm repeat containing protein homolog [Ara...     782  0.0
 gb|AAK68731.1| arm repeat containing protein homolog [Arab...     782  0.0
 gb|AAM91213.1| arm repeat containing protein homolog [Arab...     782  0.0
 gb|AEE78167.1| U-box domain-containing protein 13 [Arabido...     782  0.0
 sp|Q84WZ8.2|RM16L_ARATH RecName: Full=60S ribosomal protei...     565  3e-178
 gb|AAO37198.2| hypothetical protein [Arabidopsis thaliana]        565  3e-178
 ref|NP_191045.2| U-box domain-containing protein 14 [Arabi...     511  2e-151
 sp|Q8VZ40.1|PUB14_ARATH RecName: Full=U-box domain-contain...     511  2e-151
 gb|AAL38755.1| unknown protein [Arabidopsis thaliana]             511  2e-151
 gb|AAM20180.1| unknown protein [Arabidopsis thaliana]             511  2e-151
 gb|AEE79300.1| U-box domain-containing protein 14 [Arabido...     511  2e-151
```

Figure 4. Lineage Report for NP_565676

16. Go back to the main results page, and again select **Formatting options.** In the **Alignment view** box, select **Query-anchored with dots for identities**. Click **Reformat** and scroll down to the alignments section.
    a. *How are the alignments organized now?*
    b. *Do the substitutions appear to occur randomly between sequences, or do patterns emerge? (You might be able to get a better feel for this by scrolling towards the middle of the sequence: look for "Query" and amino acid "181" in the output). Why do you think this is? Think about evolutionary trajectories!*

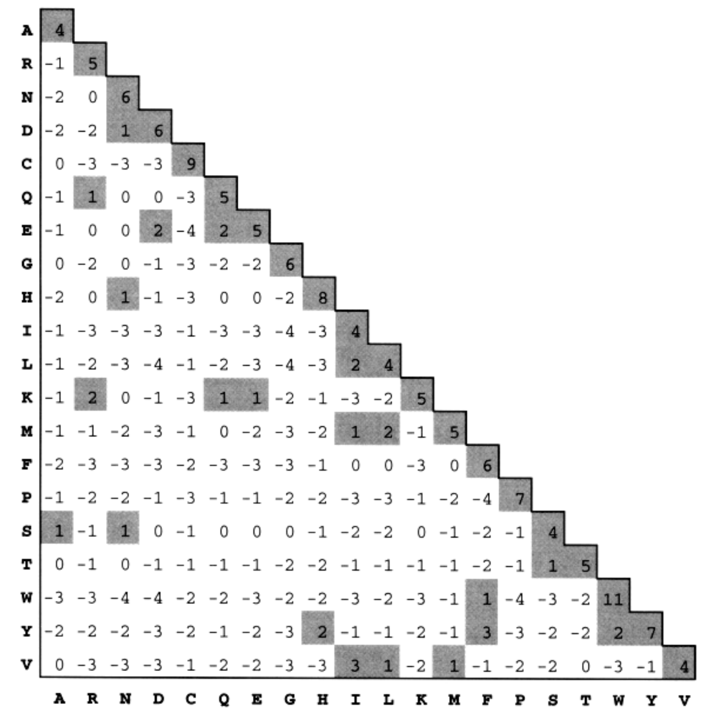
PSI-BLAST

We're now going use a new protein search algorithm: Position-Specific Iterated (PSI)-BLAST. PSI-BLAST is a highly sensitive BLAST program that is extremely useful for finding distantly related proteins or new members of a protein family. Beyond tracking down protein family members, you can use PSI-BLAST when your standard protein-protein BLAST search either failed to find significant hits, or returned hits with descriptions such as "hypothetical protein" or "similar to...".

In a very general sense, PSI-BLAST starts with a standard protein-protein BLAST, and then uses these results to build up a more refined search that is tailored to your query over successive iterations of the search. It does this by building a *position-specific scoring matrix (PSSM),* which identifies the specific amino acid changes that are most likely to be present between your query sequence and similar database sequences. Position-specific scoring matrices are essentially substitution matrices tailored to your query of interest.

Box 1. Substitution Matrices

Substitution matrices describe the likelihood that a residue (whether nucleotide or amino acid) will change over evolutionary time. They are scoring systems for comparing nucleotide or protein sequences that take into account constraints on the evolution of the sequences. The most well-known protein substitution matrices are the PAM and BLOSUM matrices developed to identify which amino acids changes are more likely to occur over specific amounts of evolutionary time. The matrix below is the BLOSUM62. The letters along the X and Y axes represent the 20 amino acids. Positive numbers indicate a high probability for going from one amino acid to the other, while low numbers indicate a low probability. You will notice that the numbers on the diagonal are all very strongly positive, indicating that the most likely thing for an amino acid residue to do is to stay the same.



A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

All substitution matrices are indexed by numbers (e.g. PAM 120, PAM250, BLOSUM62, BLOSUM80). The numbers mean different things depending on the particular matrix. For the PAM matrix, the high number matrices are better for more divergent alignments, while the opposite is true for the BLOSUM matrices.

Position-specific scoring matrices (PSSMs) are essentially substitution matrices that have been developed specifically for the protein family of interest, as opposed to PAM and BLOSUM matrices that have been developed to be generally useful for a very wide range of protein sequences.

1. Go back to the **blastp** page from the BlastP part of the lab, load the same accession or sequence (NP_565676) and parameters as in the first part of the lab (Blosum 80 Matrix and

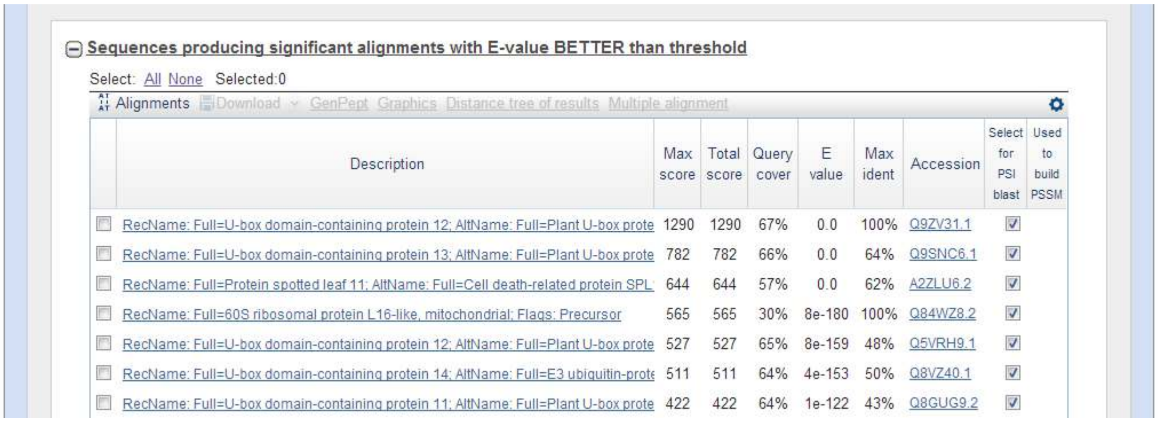
1e-15 as the Expect Threshold). Under ‘Database,’ select ‘**UniProtKB/Swiss-Prot**’. The Swiss-Prot database is well-curated database that includes only the most best annotated (characterized) protein sequences. The trade-off of using this database is that it’s less comprehensive than the default nr option.

2. Under program selection, select ‘**PSI-BLAST**’. Near the bottom of the page under **PSI/PHI BLAST** options, note the **PSI-BLAST Threshold** of 0.005. Let’s lower it to 1e-40.
    a. *How will this affect our search results?*

3. **BLAST!**

4. Examine the PSI- BLAST output. You will notice one very significant difference. There are now two sections in the HSP summary section – one that have sequences with **E-values BETTER than [PSI-BLAST] threshold**, and another that have sequences with **E-values WORSE than [PSI-BLAST] threshold**.
    a. *Where is this cutoff (what E-value)?*
    b. *How was it determined?*
    c. *How many HSPs are better than the threshold? (Tip: select All and look at the number selected).*
    d. *Record the accession number, bit score and E-value for one of the top hits in the group of sequences that **did not** reach the threshold (here it’s Q2W2J8.1).*

Lab Quiz Question 2

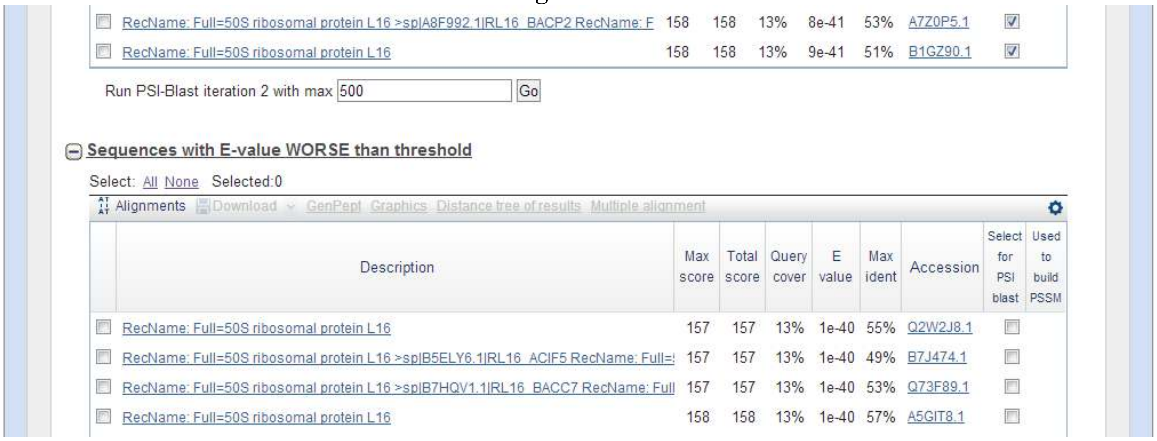


⊖ Sequences producing significant alignments with E-value BETTER than threshold

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Max ident	Accession	Select for PSI blast	Used to build PSSM
□	RecName: Full=U-box domain-containing protein 12; AltName: Full=Plant U-box prote	1290	1290	67%	0.0	100%	Q9ZV31.1	☑	
□	RecName: Full=U-box domain-containing protein 13; AltName: Full=Plant U-box prote	782	782	66%	0.0	64%	Q9SNC6.1	☑	
□	RecName: Full=Protein spotted leaf 11; AltName: Full=Cell death-related protein SPL	644	644	57%	0.0	62%	A2ZLU6.2	☑	
□	RecName: Full=60S ribosomal protein L16-like, mitochondrial; Flags: Precursor	565	565	30%	8e-180	100%	Q84WZ8.2	☑	
□	RecName: Full=U-box domain-containing protein 12; AltName: Full=Plant U-box prote	527	527	65%	8e-159	48%	Q5VRH9.1	☑	
□	RecName: Full=U-box domain-containing protein 14; AltName: Full=E3 ubiquitin-prote	511	511	64%	4e-153	50%	Q8VZ40.1	☑	
□	RecName: Full=U-box domain-containing protein 11; AltName: Full=Plant U-box prote	422	422	64%	1e-122	43%	Q8GUG9.2	☑	

…image trimmed…



	Description	Max score	Total score	Query cover	E value	Max ident	Accession	Select for PSI blast	Used to build PSSM
□	RecName: Full=50S ribosomal protein L16 >sp|A8F992.1|RL16_BACP2 RecName: F	158	158	13%	8e-41	53%	A7Z0P5.1	☑	
□	RecName: Full=50S ribosomal protein L16	158	158	13%	9e-41	51%	B1GZ90.1	☑	

Run PSI-Blast iteration 2 with max 500 Go

⊖ Sequences with E-value WORSE than threshold

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Max ident	Accession	Select for PSI blast	Used to build PSSM
□	RecName: Full=50S ribosomal protein L16	157	157	13%	1e-40	55%	Q2W2J8.1	□	
□	RecName: Full=50S ribosomal protein L16 >sp|B5ELY6.1|RL16_ACIF5 RecName: Full=!	157	157	13%	1e-40	49%	B7J474.1	□	
□	RecName: Full=50S ribosomal protein L16 >sp|B7HQV1.1|RL16_BACC7 RecName: Full	157	157	13%	1e-40	53%	Q73F89.1	□	
□	RecName: Full=50S ribosomal protein L16	158	158	13%	1e-40	57%	A5GIT8.1	□	

Figure 5. PSI-BLAST results for NP_565676, first iteration

5. This first round of PSI-BLAST was just a standard BLASTP. Now we will run another iteration to refine our search. The successive iterations use all sequences better than the cut-off to make a new position-specific-substitution matrix, which replaces the BLOSUM80 matrix used in the original search. This matrix scores based on the non-random patterns of residue conservation occurring at *each site* in each pairwise alignment BLAST performs – these are patterns you may have noticed when examining the blastp alignment from the first part of the lab, under point 16 for example.

6. Click '**Run PSI-BLAST iteration 2**'. Scroll down the sequence list. Note how beside some of the sequences there's either a green check mark, or they're highlighted in yellow – these are new sequences are those that weren't significant in the previous iteration, but scored significantly with the refined PSSM.
   a. *How many new sequences better than the cutoff do you get now?*
   b. *Search for the accession number you saved in step 4. Is it better than the threshold now? Do the bit score and E-value change for this accession? If so, why?*

7. Iterate the search at least another five times.
   a. *What do you notice about the number of new sequences in each iteration?*
   b. *What qualities should high-scoring PSI-BLAST hits theoretically share that would be of interest to an inquiring geneticist?*
   c. *Can you guess how you would include sequences that were of interest to you, but which originally did not score better than the threshold in future search?*
   d. *Likewise, how would you remove sequences that scored above the threshold, but were not of interest?*

Box 2. Potential PSI-BLAST issues

While PSI-BLAST is extremely powerful, it does suffer from some potential problems.

- We must assume that the database sequences are independent and that the sample space is large enough to represent the true underlying diversity of the family. If they are not then the PSSM will be equally biased. For example, if you are searching a database containing only proteins from Proteobacteria then your PSSM will be appropriate only for this taxonomic group.
- You may see false conservation if your database contains a large number of closely related proteins. In this case, some residues appear functionally conserved, but in fact they simply are so closely related that they haven't had time to diverge.

Translated BLAST

In addition to *nucleotide and protein blast*, there are three other searches called ***blastx***, ***tblastn***, and ***tblastx***. These three flavours of BLAST can be grouped together under the general category of translated BLAST searches. Translated searches allow you to move back and forth between the nucleotide and protein levels. They are often used to link protein and nucleotide queries to homologous DNA sequences and protein outputs in unannotated databases. Because they use either queries and/or databases translated along all six frames, they maintain robustness even in the presence of sequencing errors and frameshift mutations.

Table 1 describes the basic and translated BLAST programs. The alignment column in the table describes at what level the query and database sequences will be compared. So, for example, BLASTX will translate your DNA query to protein, and align it against the protein database. You will notice that the translated BLAST programs perform multiple searches – one search for each reading frame of either the query and / or the database.

1. *Why do tblastn and blastx perform 6 searches, while tblastx performs 36 searches?*
2. *What program should you do if you have the coding sequence of a gene and want to find homologous <u>proteins</u> in the db?*

Lab Quiz Question 3

Table 1: Basic and Translated BLAST Programs

Program	Query	Database	Alignment	N searches	Uses
blastn	DNA	DNA	DNA	1	find homologous DNA sequences
tblastx	DNA	DNA	protein	36	find homologous proteins from unannotated query and db sequences
blastx	DNA	protein	protein	6	identify coding sequences in query DNA sequence
tblastn	protein	DNA	protein	6	find homologous proteins in unannotated DNA db
blastp	protein	protein	protein	1	find homologous proteins

Example: Using blastx

Suppose you have a mystery prokaryotic ***nucleotide*** sequence below, which you obtained by sequencing a random genomic library clone from a bacterial genome. You want to know if it codes for a protein, and if so, the putative function of that protein.

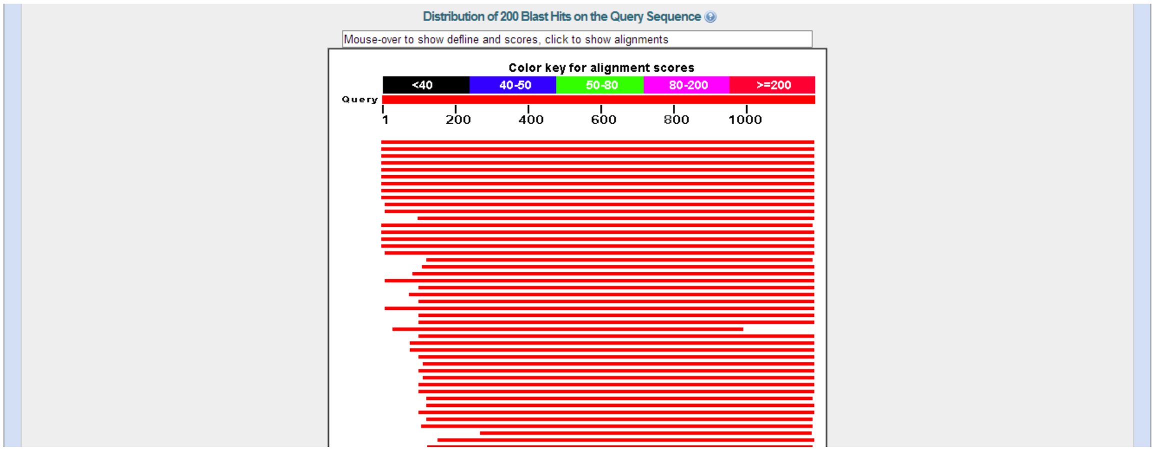
```
>mystery_sequence
GTCACGTTACCGGTGGCCGAACAGGCCCGTCATGAAGTGTTCGATGTCGCGTCGGTCAGCGCGGCTGCCGCCCCAGTAAACA
CCCTGCCGGTGACGACGCCGCAGAATTTGCAGACCGCCACTTACGGCAGCACGTTGAGTGGCGACAATCACAGTCGTCTGAT
TGCCGGTTATGGCAGTAACGAGACCGCTGGCAACCACAGTGATCTAATTGCCGGTTATGGAAGTACAGGCACCGCCGGCTAC
GGCAGTACCCAGACTTCCGGAGAAGACAGCTCGCTCACAGCGGGTTACGGCAGCACGCAAACGGCTCAGGAAGGCAGCAATC
TCACCGCTGGGTATGGCAGCACCGGCACGGCAGGCTCGGACAGCTCGTTGATCGCCGGTTATGGCAGTACACAAACCTCGGG
AGGCGACAGTTCGCTGACCGCGGGCTACGGCAGTACGCAGACGGCCCAGGAGGGCAGCAATCTGACGGCGGGGTACGGCAGC
ACGGGTACAGCAGGTGTCGACAGCTCTCTGATCGCGGGATACGGCAGCACGCAGACCTCGGGAAGTGACAGCGCCCTGACCG
CAGGCTATGGCAGCACGCAAACGGCCCAGGAAGGCAGCAATCTCACTGCTGGGTATGGCAGCACCGGCACGGCAGGTTCCGA
CAGCTCGCTGATCGCCGGTTACGGCAGCACGCAAACCTCGGGCAGTGACAGCTCGCTCACGGCGGGGTACGGCAGTACGCAG
ACGGCTCAGGAAGGCAGCAATCTGACGGCGGGGTACGGCAGCACGGGTACAGCAGGTGTCGACAGTTCGTTGATCGCCGGAT
ATGGCAGCACGCAGACCTCGGGAAGTGACAGTGCGCTGACAGCGGGTTACGGCAGCACGCAAACGGCCCAGGAAGGCAGCAA
CCTGACGGCGGGCTACGGCAGCACTGGCACGGCAGGTGCCGACAGTTCGTTGATCGCCGGATATGGCAGCACGCAGACGTCA
GGCAGCGAAAGTTCGCTTACCGCAGGCTATGGCAGTACCCAGACTGCCCGTGAGGGCAGCACCCTGACGGCCGGATATGGCA
GTACCGGAACAGCTGGCGCTGACAGCTCGCTGATCGCCGGTTACGGCAGCACGCAAACCTCGGGCAGTGAAAGCTCGCTCAC
GGCAGGTTATGGCAGTACCCAGACCGCACAGC
```

3. Go to the BLAST main page.

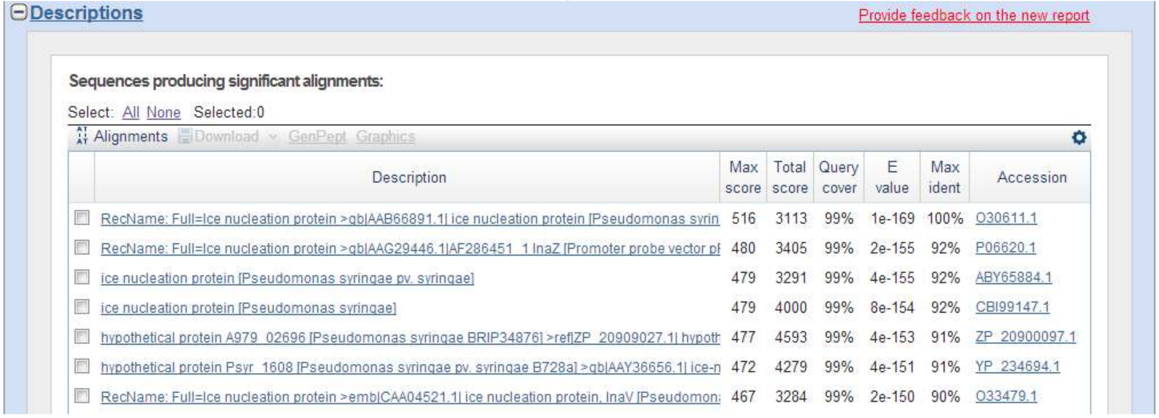
4. Select **blastx**, and copy and paste the sequence into the search box. Make sure that the **Non-redundant protein sequences (nr)** is selected as the **Database**. You can use the rest of the default parameters.

5. BLAST!

6. Scroll over your results.
   a. *What species do the top-scoring* ***protein*** *sequences belong to?*
   b. *What would you guess is the function of this gene?*
   c. *What organism do you think it came from?*


Distribution of 200 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
200
400
600
800
1000

---------------- *image trimmed* ----------------



Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
□	RecName: Full=Ice nucleation protein >gb|AAB66891.1| ice nucleation protein [Pseudomonas syrin	516	3113	99%	1e-169	100%	O30611.1
□	RecName: Full=Ice nucleation protein >gb|AAG29446.1|AF286451_1 InaZ [Promoter probe vector pI	480	3405	99%	2e-155	92%	P06620.1
□	ice nucleation protein [Pseudomonas syringae pv. syringae]	479	3291	99%	4e-155	92%	ABY65884.1
□	ice nucleation protein [Pseudomonas syringae]	479	4000	99%	8e-154	92%	CBI99147.1
□	hypothetical protein A979_02696 [Pseudomonas syringae BRIP34876] >ref|ZP_20909027.1| hypoth	477	4593	99%	4e-153	91%	ZP_20900097.1
□	hypothetical protein Psyr_1608 [Pseudomonas syringae pv. syringae B728a] >gb|AAY36656.1| ice-n	472	4279	99%	4e-151	91%	YP_234694.1
□	RecName: Full=Ice nucleation protein >emb|CAA04521.1| ice nucleation protein, InaV [Pseudomon:	467	3284	99%	2e-150	90%	O33479.1

Figure 6. BLASTX output

7. Scroll down to the first alignment.
   a. *You may notice that some of the query residues are in lower case gray letters (as opposed to upper case black). Can you guess what these may signify? Try going back to the query page and change the Filters and Masking options. How does this affect your results?*

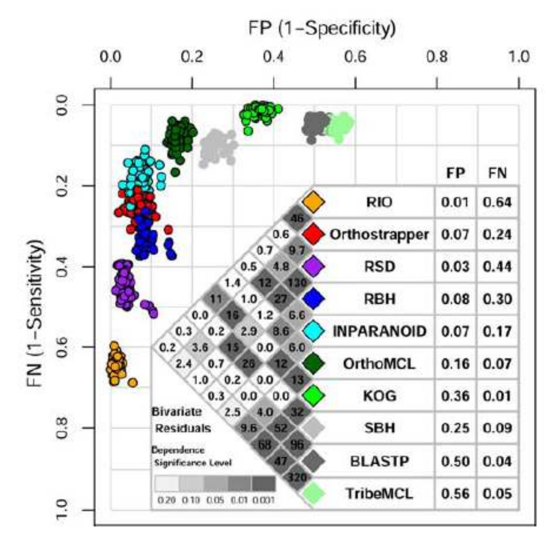
Comparative Genomics

High-throughput sequencing initiatives, proteomic efforts, transcriptomics and other high-throughput genomic technologies, in conjunction with molecular characterization and literature curation, have resulted in large data collections. In addition to one for the human genome itself, repositories for the model organisms of worms, fruitflies, mice, *Arabidopsis thaliana* and others exist. However, the representation of genomic data is challenging due to the sheer scope, complexity and volume of these data. Tools and the means for effectively displaying such complex data are continually being developed and improved. We will look at several applications that attempt to deal with this problem, and that permit comparisons of genomic regions across related species. Importantly, orthologous genes studied in one species can be assumed to have similar functions in another species, and residues within orthologs that are highly conserved can be assumed to be critical for that protein's function. Therein lies the power of comparative genomics.

Box 3. Comparative Genomics

As we saw in Lab 1, in order to be able to do comparative genomics we need to be able to determine orthologs and paralogs. The conceptually simplest method is to Blast regions (genes) of one genome against another genome and identify the regions (genes) that have the lowest e-value in the 2nd genome. This is problematic because what happens if the region identified in the 2nd genome actually has a better match elsewhere in the first genome? A variation of this method involves Blasting in both directions, i.e. from a gene on one genome to the other genome and then doing the Blast in the opposite direction, and identifying the "reciprocal best hit" or RBH. This reduces the number of false positive orthologs, but increases the number of false negatives.

A variety of methods have been developed to address this issue and involve using either phylogenetic methods (which are computationally more "expensive") or BLASTP followed by clustering methods to identify orthologous genes. These methods are RIO and Orthostrapper in the first case, or InParanoid and OrthoMCL in the latter. A summary of the performance of these methods in terms of false positives and false negatives is shown to the right. See Chen et al., 2007, PLoS One 2(4): e383 for further details.



		FP	FN
🟧	RIO	0.01	0.64
🟥	Orthostrapper	0.07	0.24
🟪	RSD	0.03	0.44
🟦	RBH	0.08	0.30
◆	INPARANOID	0.07	0.17
🟩	OrthoMCL	0.16	0.07
🟩	KOG	0.36	0.01
◆	SBH	0.25	0.09
◆	BLASTP	0.50	0.04
◆	TribeMCL	0.56	0.05

Exploring Genomes with Genome Browsers

Each "model" organism (these organisms are so designated because of a long history of being studied in a medical or agricultural context due to their ease of manipulation, space requirements, good genetics, among other reasons) has its own genome database that permits the exploration of genomic regions. Such regions often have other molecules – homologous genes, ESTs etc. – associated with them. These genomic regions can be explored with Genome Browsers. We will only look at the Mouse genome browser in this part of the lab, but here are some others portals that may be of use in your future studies:

FlyBase

FlyBase is the genomic repository for information for the model organism *Drosophila melanogaster* and many other related drosophilid species. Connect to FlyBase at http://flybase.org and click on the GBrowse icon to access the Genome Browser.

WormBase

WormBase is the repository for *Caenorhabditis elegans* and other worm model species genomic data. Connect to wormbase at http://www.wormbase.org. Click on the "GBrowse" link in the Tool section at the top of the page to access the Genome Browser for *C. elegans*.

The Arabidopsis Information Resource

As we've seen, TAIR is the repository for Arabidopsis genomic information. Connect to TAIR at http://www.arabidopsis.org and under the Tools tab, click on GBrowse.

NCBI

One can also examine the human genome in a comparative manner using the NCBI map viewer application. Connect to the genomes section of the NCBI website at: http://www.ncbi.nlm.nih.gov/Genomes/ and select the "*Human Genome*" link under the Custom Resources, then on the icons of the chromosomes for the most recent Map Viewer release (you used this in the first lab, too). Under Maps and Options, it is possible to select sequences from other species to display on the human Map Viewer.

Mouse Genome Informatics – Comparative Genomics Example

Connect to the Mouse Genome Informatics site at: http://www.informatics.jax.org/. Enter *Pax6* into the Quick Search box at the top right. *Pax6* is a gene important for proper eye development. Click on the first link in the results list, labeled Pax6. You'll be taken to the Gene Detail page for this gene. In the fifth row, called **Sequence Map**, click on the Ensembl Genome Browser link. Ensembl is run by the European Bioinformatics Institute, the European counterpart of the NCBI. This browser is in some regards more powerful than the NCBI map viewer, but because it contains so much information and so many options, it is a bit more confusing. We can use it to examine similar human and mouse genomic regions, and to view orthologs.

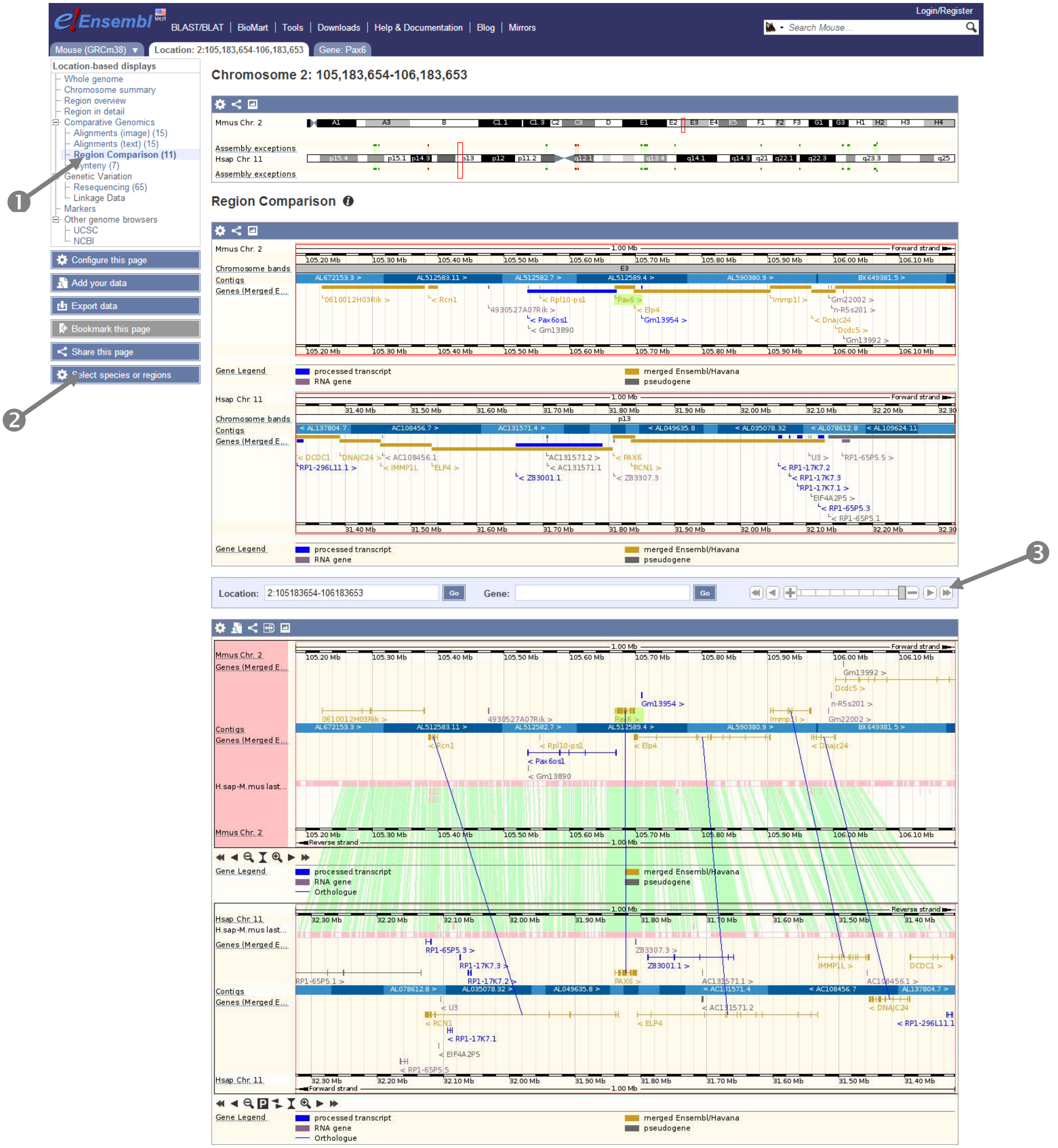


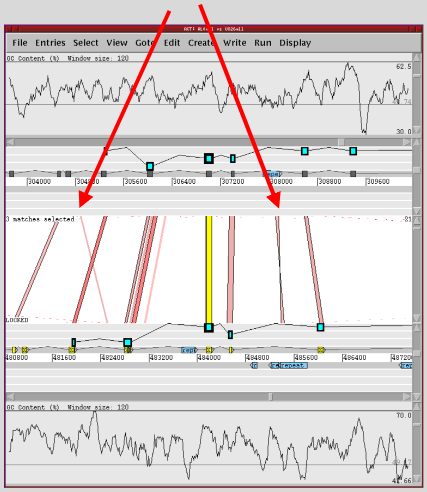
Figure 7: Enseml Genome Browser showing mouse and human Pax6 genomic regions. Click on ❶ Region Comparison then ❷ Select Species or Regions to add e.g. the corresponding human region. Use ❸ zoom slider to zoom out to view 1000000 base pairs at once. Thin diagonal blue lines in the bottom panel represent orthologous genes between mouse and human.

1. In the Ensembl Genome Browser showing Pax6 in its genomic context, use the menu on the left to select Region Comparison. Use the Select Species or Regions tab on the bottom left to add the corresponding human region. Use the zoom slider to zoom out to view 1000000 base pairs at once (see **Figure 7**). Thin diagonal blue lines in the bottom panel represent orthologous genes between mouse and human.
   a. *Is there a human ortholog for this gene? What chromosome is it on? (Chromosome numbers are listed to the left of the chromosome pictographs)*
2. Look at the other human orthologs surrounding PAX6 and their chromosomal locations.
   a. *Would you say that there is synteny between the chromosomal region containing mouse* Pax6 *and human* PAX6*? Why?*

Box 4. Comparative Genomics and Synteny

Gene order is often conserved between closely related species, and even between species that are less than closely related, such as human and mouse. Synteny can be observed by using several visualization tools that have been developed, e.g. the Artemis Comparison Tool, ACT.

An example ACT output of the comparison of a region of the *Homo sapiens* X chromosome versus a region of the *Mus musculus* X chromosome is shown to the right. Such visualization tools are useful for identifying insertions – the greater divergence of the the slopes of two blocks connecting the horizontally-displayed genomes, the greater the insertion in one or the other of them, and translocations and inversions – these show up as blocks which cross other blocks, and as "X" shaped figures. These are readily visible in the figure to the right. The ACT is published in Carver *et al.*, 2005, Bioinformatics 21(16):3422-3423.


File Entries Select View Goto Edit Create Write Run Display
GC Content (%) Window size: 120
3 matches selected
LOCKED

Potential translocations/inversions

WebACT

3. You can check out the similarities and differences of several precomputed genomic comparisons, using the http://www.webact.org/WebACT/prebuilt tool from the Sanger Institute. For example, examine the *Agrobacterium tumefaciens* strain C58 / ATCC 33970, sub_strain Cereon circular chromosome versus the *Agrobacterium tumefaciens* strain C58 / ATCC 33970, sub_strain Dupont genomes. Use the defaults. You may need to download the .jnlp file that WebACT uses and run it on your computer with Java, if Java Web Start isn't active in your browser.
   a. *How similar are these genomes?*
   b. *Even though these genomes are ostensibly the same ATCC (American Type Culture Collection) identifiers, what can you say by scanning along the length of the genomes?*

End of lab!

Lab 2 Objectives

By the end of Lab 2 (comprising the lab including its boxes, and the lecture), you should:

- understand the general concept underlying substitution matrices used for scoring protein similarity;
- know which type of matrix to use to identify more distantly related sequences or those that are more closely related;
- be able to interpret a dot matrix alignment;
- be familiar with the theory behind the working of the BLAST algorithm;
- know which flavour of BLAST (blastn, blastp, tblastx etc.) to use when – the key is to know whether your query sequence is nucleotide or protein, and whether your database is nucleotide or protein;
- be able to use the appropriate GenBank database to find what you're looking for when you BLAST;
- know how to reduce the number of hits returned in a given BLAST output by decreasing the e-value threshold, and how to reformat the BLAST output;
- appreciate the value of PSI-BLAST for identifying distantly related sequences;
- be familiar with genome browsers for identifying orthologous genes;
- know what is meant by synteny.

Do not hesitate to check the Coursera forums if you have any questions after reading the relevant material.

Further Reading

Chapter 4 "Database Similarity Searching" in *Essential Bioinformatics* by Jin Xiong, Cambridge University Press, 2006. pp 52-57.

SF Altschul , TL Madden , AA Schaffer , J Zhang , Z Zhang , W Miller , and DJ Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25: 3389-3402.

S Henikoff, JG Henikoff (1992). Amino Acid Substitution Matrices from Protein Blocks. Proceedings of the National Academy of Sciences U.S.A. 89:10915-10919.

Section 9.8 "Large Genome Comparisons" in Chapter 9 "Revealing Genome Features" in *Understanding Bioinformatics* by Marketa Zvelebil and Jeremy Baum, Garland Science, 2008. pp. 352-354.

Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J (2005). ACT: the Artemis Comparison Tool. Bioinformatics 21(16):3422-3.

Chen F, Mackey AJ, Vermunt JK, Roos DS (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE 2(4):e383.

Appendix 1: BLASTable Databases

Protein Databases	
nr	Non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF
swissprot	Last major release of the SWISS-PROT protein sequence database
pat	Proteins from the Patent division of GenBank.
month	All new or revised GenBank CDS translations + PDB + SwissProt + PIR + PRF released in the last 30 days.
pdb	Sequences derived from the 3-dimensional structure records from the Protein Data Bank
Nucleotide Databases	
nr	All GenBank + EMBL + DDBJ + PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences).
est	Database of GenBank + EMBL + DDBJ sequences from EST division
est_human	Human subset of est
est_mouse	Mouse subset of est.
est_others	Subset of est other than human or mouse.
gss	Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
htgs	Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2. Finished, phase 3 HTG sequences are in nr.
pat	Nucleotides from the Patent division of GenBank.
pdb	Sequences derived from the 3-dimensional structure records from Protein Data Bank.
month	All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
alu_repeats	Select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences
dbsts	Database of Sequence Tag Site entries from the STS division of GenBank + EMBL + DDBJ.
chromosome	Complete genomes and complete chromosomes from the NCBI Reference Sequence project.
wgs	Assemblies of Whole Genome Shotgun sequences

Appendix 2: The Command Line

File-Naming Conventions

Careful file naming can save time and frustration. Always choose names which provide a clue to the file's contents. If you are working with a series of related files, use a number somewhere in the name to indicate which version you have created.

- You can use files with spaces in their names when working through the command prompt. Simply enclose the name in single quotes.
- When working on a Windows machine it is always a good idea to end your file with a characteristic file extension. The file extension is usually a 3 letter names that follows the last period in the file name. Examples:

.ali	clustal alignment files
.fas	fasta data files
.meg	MEGA data files
.nex	NEXUS data files
.phy	PHYLIP data files
.tre	phylogenetic tree files – typically in Newick format
.txt	text files

Change the Default Drive

To change the default drive, simply type the letter of the your choice. The new default will be listed in subsequent DOS prompts.

```
C> A: [enter]          Changes the default drive from C to A.
A> C: [enter]          Changes the default drive from A to C.
```

[*enter*] means that you must press the Enter Key before the format command will execute. [Enter] is required after any DOS command, it is assumed in all commands found below.

DIR (Directory)

The DIRECTORY command lists the names and sizes of all files located on a particular disk.

```
C> dir        Shows the contents of the current directory
```

Two characters, '*' and '?', will make your life much easier. The '*' is a wild-card character which allows you to enter only a limited part of a file specification to find a file. It is useful when you wish to locate a group of files with the same filename or the same extension, or if you forgot part of the file name. The '?' permits wild-card searches for single characters.

```
C> dir *.fas  Lists all files with the file extension of 'fas'.
```

CHDIR (CD) Change Directory

Once you have located the directory you want, you may move from directory to directory using the CD command (change directory)

```
C> cd NCBI             Moves you to the directory called 'NCBI'
C> cd \NCBI\blast      Moves you to the subdirectory called 'blast' under the 'NCBI' directory.
C> cd ..               Moves you up one level in the path.
C> cd \                Takes you back to the root directory (C: in this case).
```

COPY

The COPY command can be used both to copy files from disk to disk or to create a second copy of a file on a single disk.

```
C> copy c:seq.fas a:       Copies the file 'seq.fas' from the C drive to the A drive and gives it the same name.
C> copy c:seq.fas d:\data\seq1.fas          Creates a copy of 'seq.fas' from drive C on drive D, putting it in the
'data' subdirectory and renaming it 'seq1.fas'.
```

Remember that the first file specified after the COPY command is the source file, the second is the target file. The source is the file to be copied. The target will be the location and name of the new file. If the file name and extension are omitted after the target's drive specification, the new file will have exactly the same name as the source file.

RENAME (REN)

The RENAME command permits users to change the name of a file without making a copy of it.

```
C> ren seq.fas seq1.fas  Changes the name of 'seq.fas’ to 'seq1.fas'.
```

ERASE

The ERASE command deletes specified files. Be careful with this command!

```
C> erase seq.fas           Erases the file ‘seq.fas’
```

MKDIR (MD) Make Directory

This command creates a new directory.

```
C> mkdir data              Creates a directory called ‘data’
```

RMDIR (RD) Remove Directory

This command removes a directory. It is only possible to execute this command if the directory you wish to remove is empty.

```
C> rd data      Removes directory called 'data'.
```

Stop Execution

Use **Ctrl-Break** to stop the computer while it is executing a command or program.

Appendix 3: Standalone BLAST

There may be times when you want to perform BLAST searches on dozens, hundreds, or even thousands of queries. While the NCBI website is extremely powerful, it can also be slow, and is definitely not appropriate for dealing with high-throughput data analysis. In these cases you should consider using standalone BLAST from your local computer.

All of the BLAST programs can be easily installed locally (on your own computer), and run on almost any computer platform. Executables and databases are freely available from NCBI at ftp://ftp.ncbi.nih.gov/blast/. The executables for a wide range of operating systems are available at ftp://ftp.ncbi.nih.gov/blast/executables/LATEST. General documentation for the BLAST programs and utilities is available at ftp://ftp.ncbi.nih.gov/blast/documents/. Help for installing the programs can be found at ftp://ftp.ncbi.nih.gov/blast/documents/blast.html.

1. The database that you want to BLAST against needs to be formatted properly for BLAST to make use of it. We use the program ***formatdb*** to do this.

   To see all the options and adjustments you can make in ***formatdb***, in your command prompt window type:

   ```
   formatdb –
   ```

   You should see a list of all the *formatdb* options.

   A typical *formatdb* command might look like:

   ```
   formatdb -i C:\yeast.aa -p T -o T
   ```

   where yeast.aa would be a file containing all yeast protein sequences (as an example) in FASTA format. Index files will be created when you run *formatdb*. You are now almost ready to standalone BLAST.

2. Your query sequence will need to be saved a text file in FASTA format (recall how to do this from Lab 1). Use a text editor (e.g. Crimson) to save a copy using a name such as <*accession number*>*.fas*. Be aware of the directory you have saved this sequence in, as you will need to specify it in your BLAST command.

   At the command prompt (which should still be open), you'd type something like this to do a standalone BLAST:

   ```
   blastall -p blastp -i C:\<accession number>.fas -d C:\yeast.aa -o C:\out.txt
   ```

   This command basically means:
   use executable file *blastall*
   program —> *blastp*
   search query —> <accession number>.fas
   database to search—>Swiss-prot
   download the results into the file —>out.txt

   See appendix 4 for some commonly used BLAST options.

3. You can now view your output by opening the out.txt file in a text editor.

4. Sometimes you need to extract specific information out of the BLAST output. This is called parsing. NCBI provides a very simple BLAST parser to extract gi numbers from your analysis. To use this simply go to http://www.ncbi.nlm.nih.gov/Class/BLAST/new_parse.html, copy – paste your BLAST results into the window, and press **Get GI list**. If you then highlight all the numbers given and paste them into Entrez's search box, it will get you the entire list of results, which you can then download as a text file and use in future analyses..

Appendix 4: Some Commonly Used BLAST Options for standalone BLAST

-p Program Name [String]

Input should be one of "blastp", "blastn", "blastx", "tblastn", or "tblastx".

-d Database [String]
default = nr

The database specified must first be formatted with formatdb. Multiple database names (bracketed by quotations) will be accepted. An example would be:

-d "nr est"

which will search both the nr and est databases, presenting the results as if one 'virtual' database consisting of all the entries from both were searched. The statistics are based on the 'virtual' database of nr and est.

-i Query File [File In]
default = stdin

The query should be in FASTA format. If multiple entries are in the input file, all queries will be searched.

-e Expectation value (E) [Real]
default = 10.0

-o BLAST report Output File [File Out] Optional
default = stdout

-F Filter query sequence (DUST with blastn, SEG with others) [String]
default = T

BLAST uses the dust low-complexity filter for blastn and seg for the other programs.

If one uses "-F T" then normal filtering by seg or dust (for blastn) occurs (likewise "-F F" means no filtering whatsoever).

-S Query strands to search against database (for blast[nx], and tblastx).
3 is both, 1 is top, 2 is bottom [Integer]
default = 3

-T Produce HTML output [T/F]
default = F

-l Restrict search of database to list of GI's [String] Optional

This option specifies that only a subset of the database should be searched, determined by the list of gi's (i.e., NCBI identifiers) in a file. One can obtain a list of gi's for a given Entrez query from http://www.ncbi.nlm.nih.gov/Entrez/batch.html. This file should be in the same directory as the database, or in the directory that BLAST is called from.

-U Use lower case filtering of FASTA sequence [T/F] Optional
default = F

This option specifies that any lower-case letters in the input FASTA file should be masked.