LAB 2 — ADVANCED BLAST AND COMPARATIVE GENOMICS

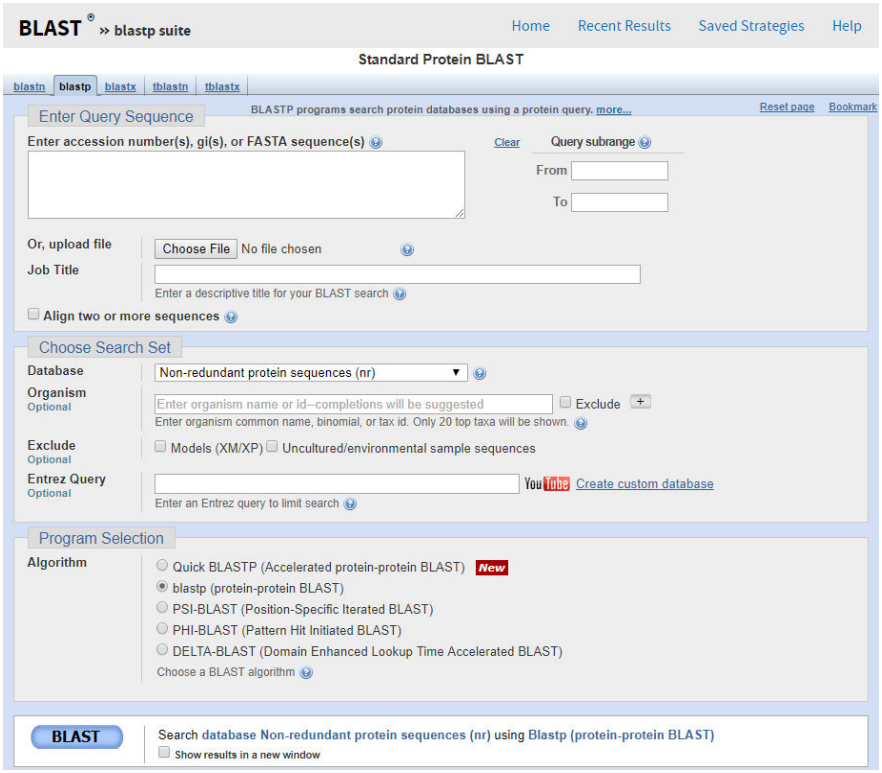
[Software needed: web access]

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.



…

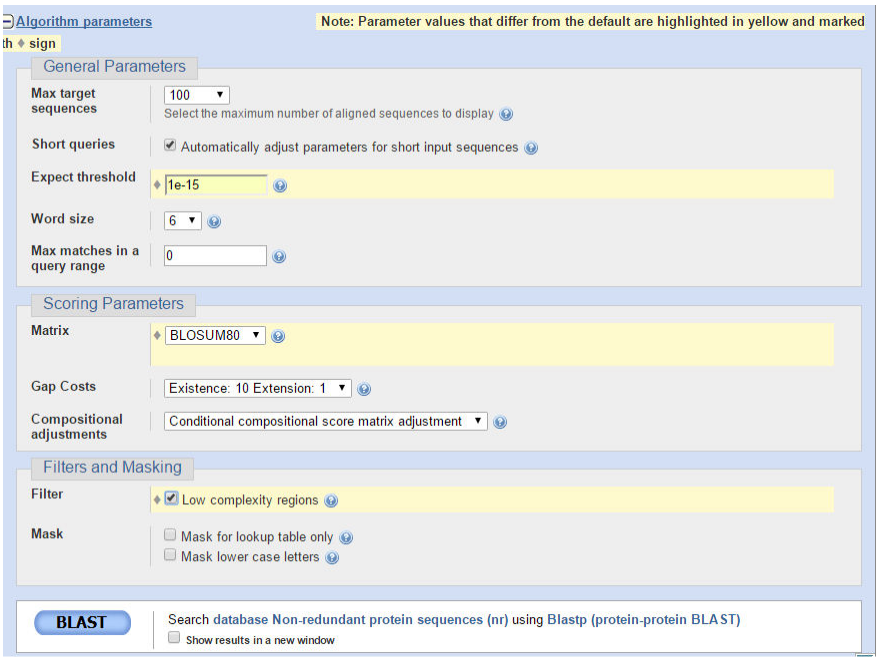


Figure 1. The blastp input page, with some default settings modified in the Parameters section.

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

4. Enter the accession number (NP_001318308) that you used in Lab 1 into the BLAST search box.

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?*

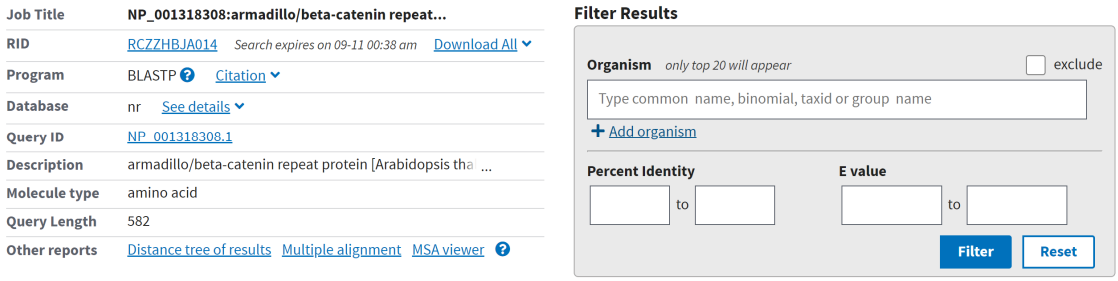
8. Check '**Low complexity regions**' filter.
   a. *Why might you want to use this filter? Don't forget to use the help icons if you need more information on any of the parameters.*

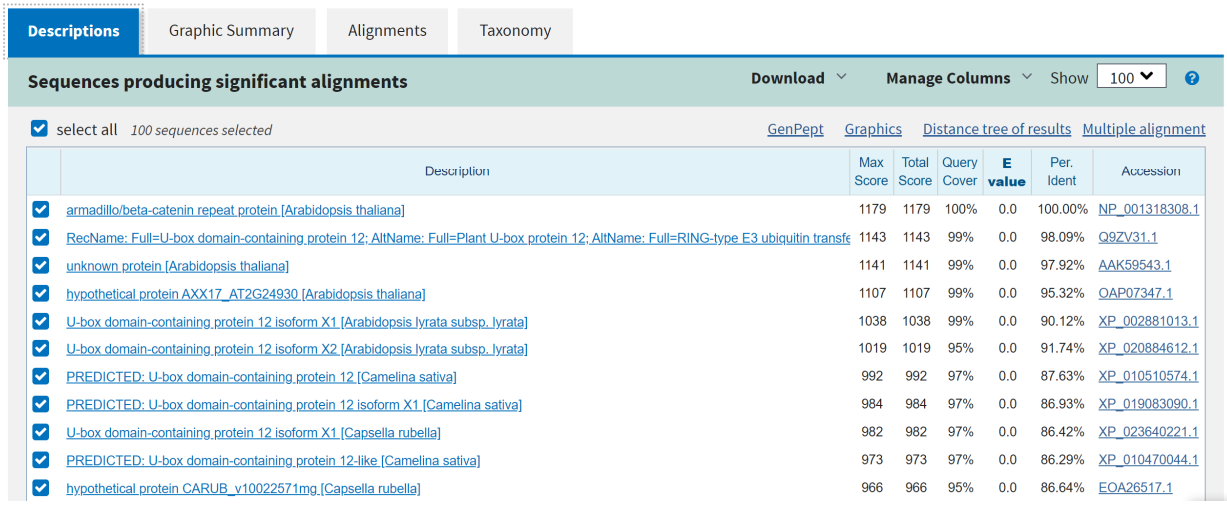
9. Check '**Show results in a new window**'.

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 (CDD). 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
    - Descriptions
    - Graphical summary containing
        - Conserved domains
        - Graphic summary
    - Alignments
    - Taxonomy





…image trimmed…

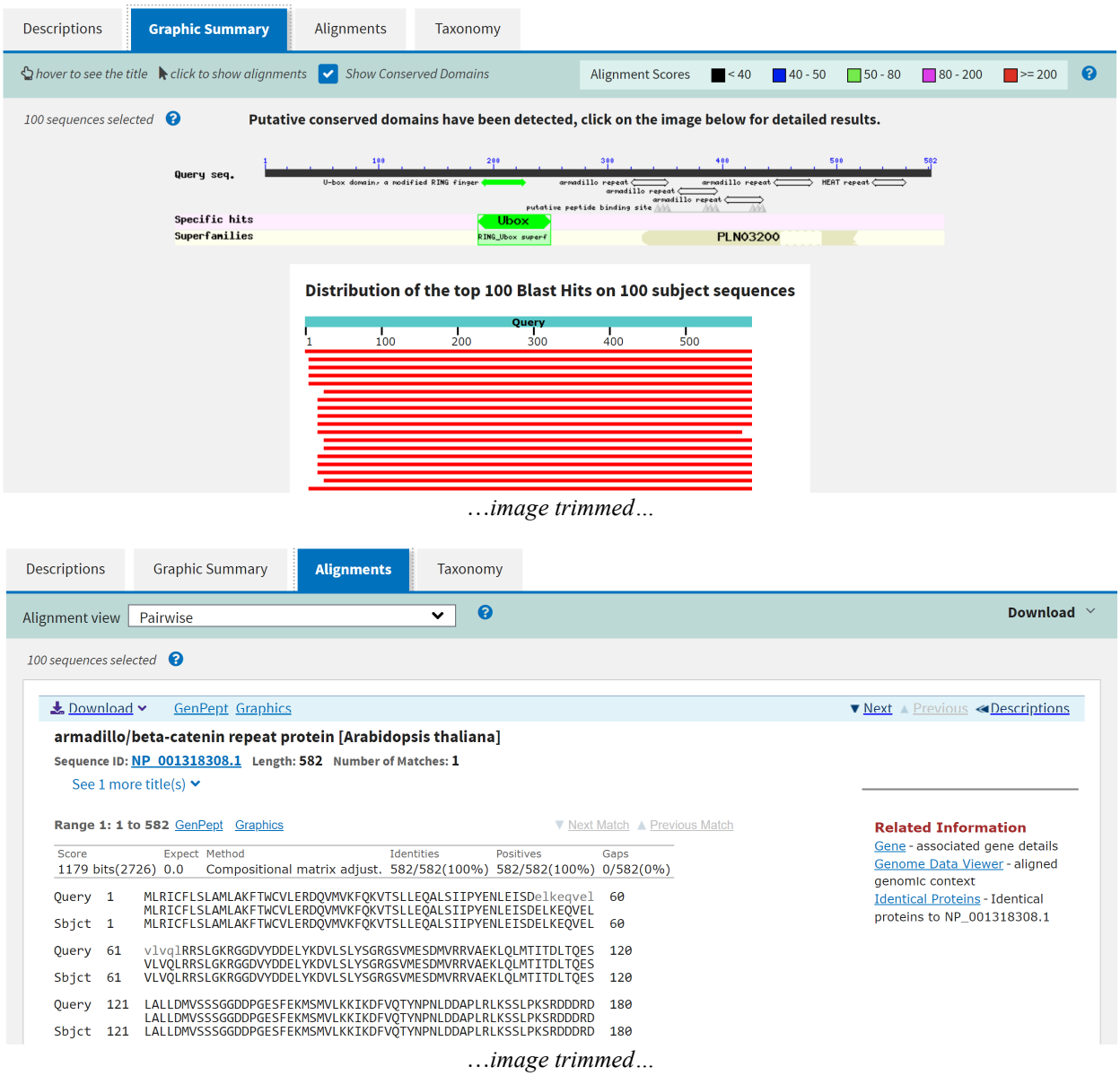


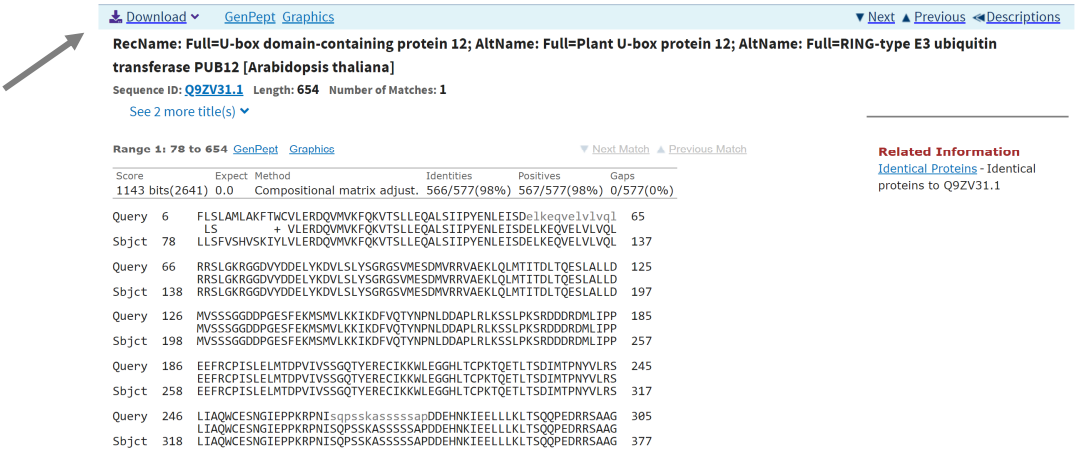
Figure 2. BLASTP output for NP_001318308 search

12. Let's try a different format for our results. Click on **Alignment View** options in the Alignments part of the page, and pick **Pairwise with dots for identities** in the **Alignment View** box to change the alignments representation (see **Figure 3**).

13. Move to the graphic summary section and mouse over the summaries. The information for each hit will be displayed in the tool tip. Scroll to the Descriptions section. 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



Pairwise with dots for identities format

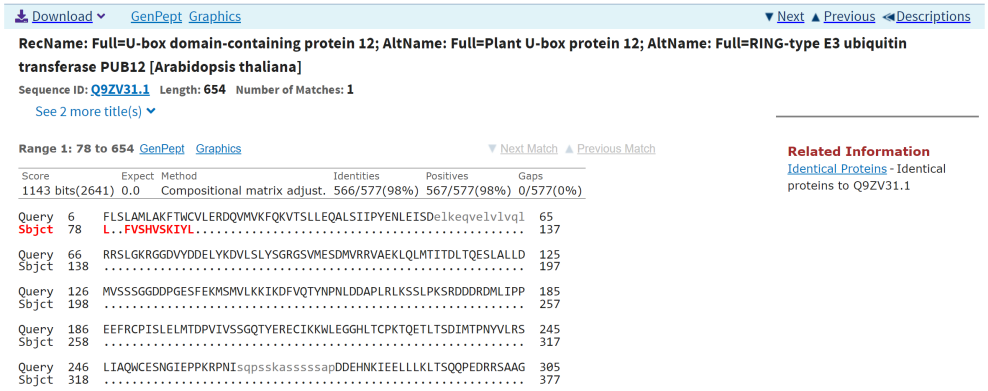
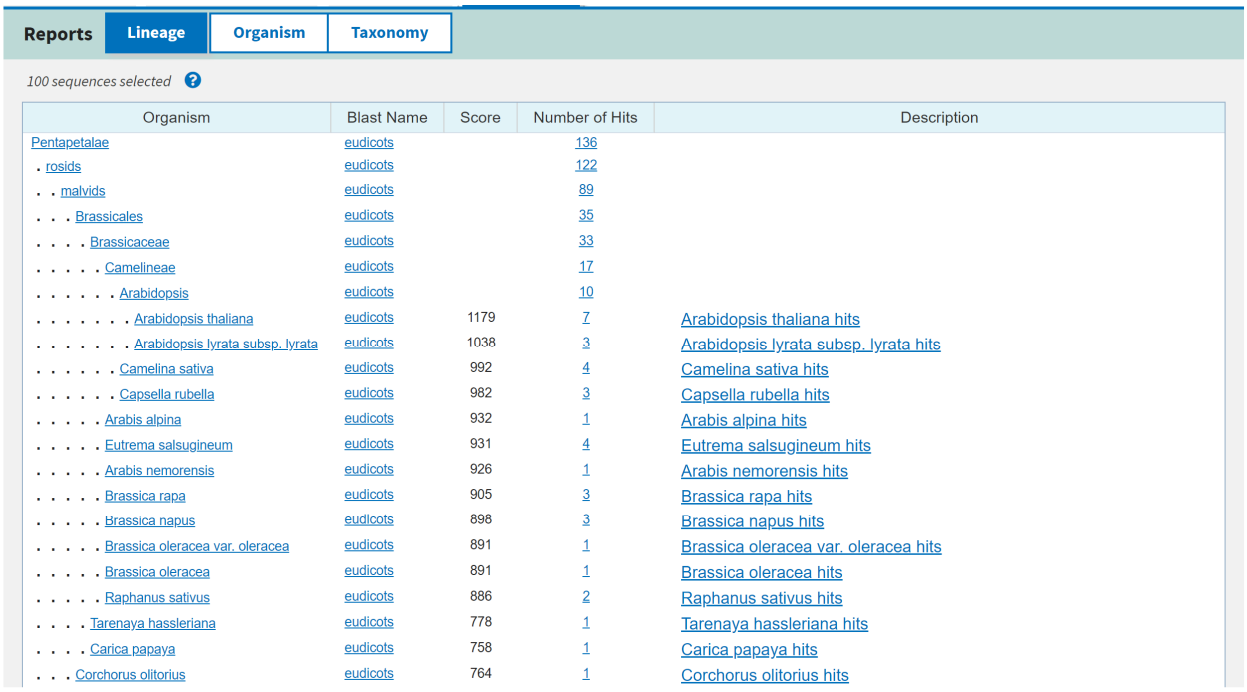


Figure 3. Two different formats of BLASTP alignments from an NP_001318308 search. Arrows show the useful Download feature, for downloading either the aligned part of the Subject sequence, or the entire Subject sequence. Downloads of multiple sequences can be done from the Descriptions Table too. This 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 **Taxonomy** tab. 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. 1179 in the case of Arabidopsis thaliana in Figure 4)? A couple of clicks and you should be able to figure this out – click on the "Arabidopsis thaliana hits" link!*



Reports | Lineage | Organism | Taxonomy

100 sequences selected

Organism	Blast Name	Score	Number of Hits	Description
Pentapetalae	eudicots		136	
. rosids	eudicots		122	
. . malvids	eudicots		89	
. . . Brassicales	eudicots		35	
. . . . Brassicaceae	eudicots		33	
. . . . . Camelineae	eudicots		17	
. . . . . . Arabidopsis	eudicots		10	
. . . . Arabidopsis thaliana	eudicots	1179	7	Arabidopsis thaliana hits
. . . . Arabidopsis lyrata subsp. lyrata	eudicots	1038	3	Arabidopsis lyrata subsp. lyrata hits
. . . . . . Camelina sativa	eudicots	992	4	Camelina sativa hits
. . . . . . Capsella rubella	eudicots	982	3	Capsella rubella hits
. . . . . Arabis alpina	eudicots	932	1	Arabis alpina hits
. . . . . Eutrema salsugineum	eudicots	931	4	Eutrema salsugineum hits
. . . . . Arabis nemorensis	eudicots	926	1	Arabis nemorensis hits
. . . . . Brassica rapa	eudicots	905	3	Brassica rapa hits
. . . . . Brassica napus	eudicots	898	3	Brassica napus hits
. . . . . Brassica oleracea var. oleracea	eudicots	891	1	Brassica oleracea var. oleracea hits
. . . . . Brassica oleracea	eudicots	891	1	Brassica oleracea hits
. . . . . Raphanus sativus	eudicots	886	2	Raphanus sativus hits
. . . . Tarenaya hassleriana	eudicots	778	1	Tarenaya hassleriana hits
. . . . Carica papaya	eudicots	758	1	Carica papaya hits
. . . Corchorus olitorius	eudicots	764	1	Corchorus olitorius hits

Figure 4. Lineage Report for NP_001318308 (partial), under the Taxonomy tab.

16. Go back to the main results page, and again select **Formatting options.** In the **Alignments** tab, select **Query-anchored with dots for identities**. 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 "61" in the output). Why do you think this is? Think about evolutionary trajectories!*

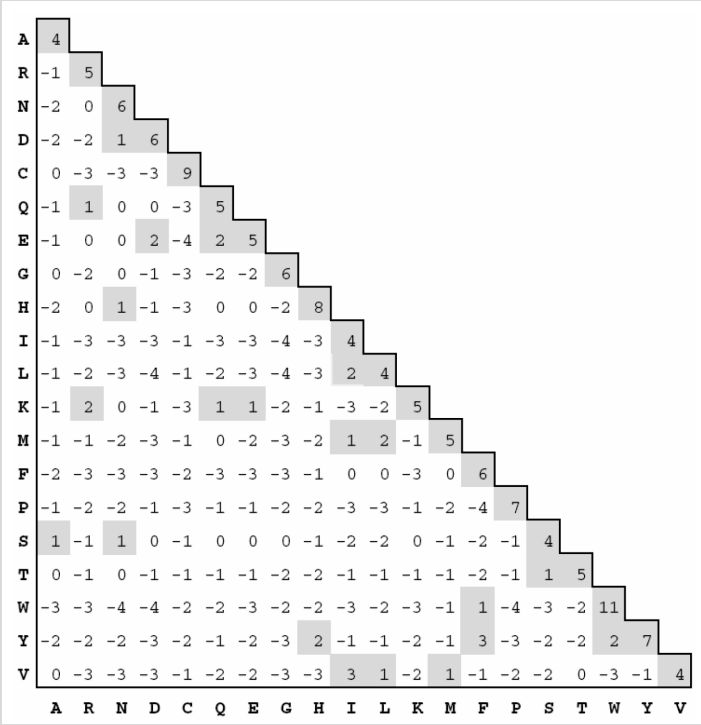
PSI-BLAST

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 off the diagonal indicate a higher probability for one amino acid to change to another, 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	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
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

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 (see lecture material).

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_001318308) and parameters as in the first part of the lab (Blosum 80 Matrix and 1e-15 as the Expect Threshold, low complexity filter). 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 Description summary section – one that has sequences with **E-values BETTER than [PSI-BLAST] threshold**, and another that has 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 sequences are better than the threshold? (Tip: select <u>All</u> 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 Q9C7G1.1).*

Lab Quiz Question 2

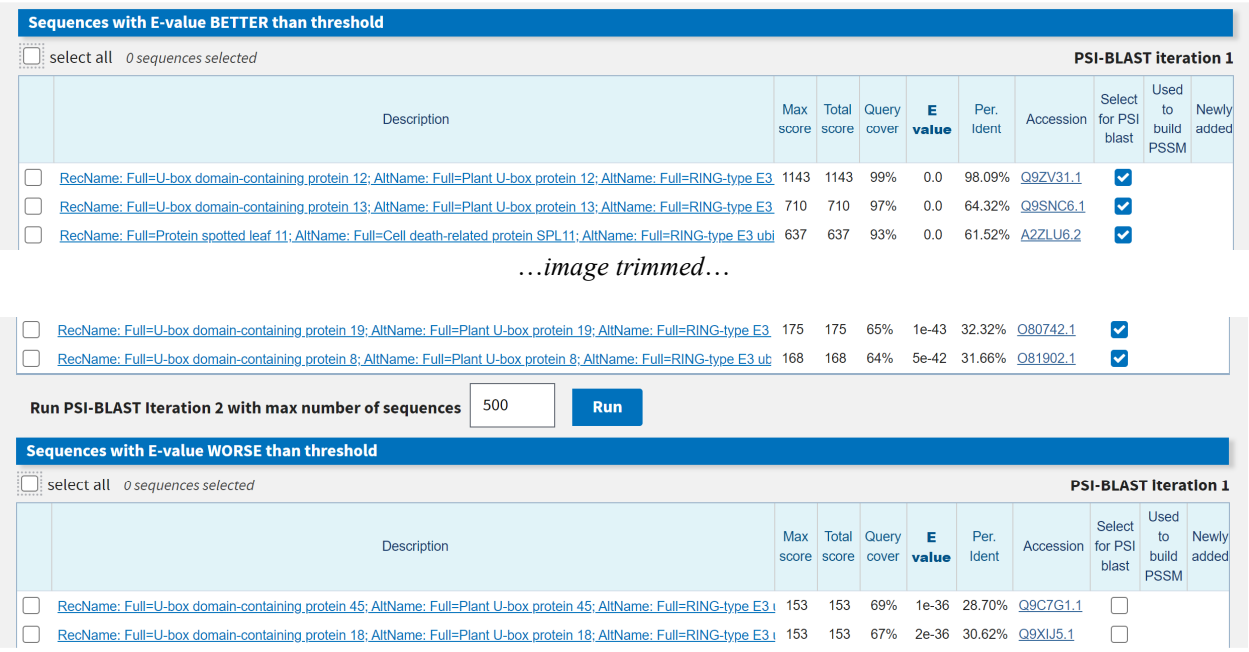


Figure 5. PSI-BLAST results for NP_001318308, first iteration

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 run if you have the coding sequence of a gene and want to find homologous proteins in the database?*

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?*

… image trimmed …

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics

Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=Ice nucleation protein	516	2729	99%	2e-170	100%	O30611.1
Ice nucleation protein [Pseudomonas syringae]	503	2396	99%	5e-166	96%	WP_122265877.1
Ice nucleation protein [Pseudomonas syringae pv. aceris]	503	2395	99%	7e-166	96%	RMS60456.1

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?*

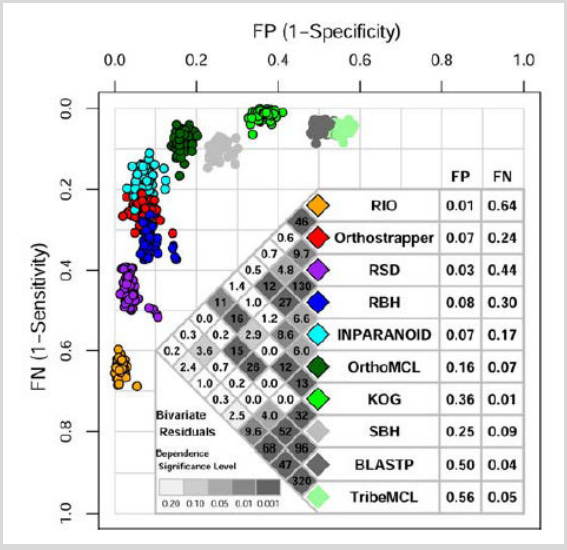
Comparative Genomics

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, fruit flies, 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.



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

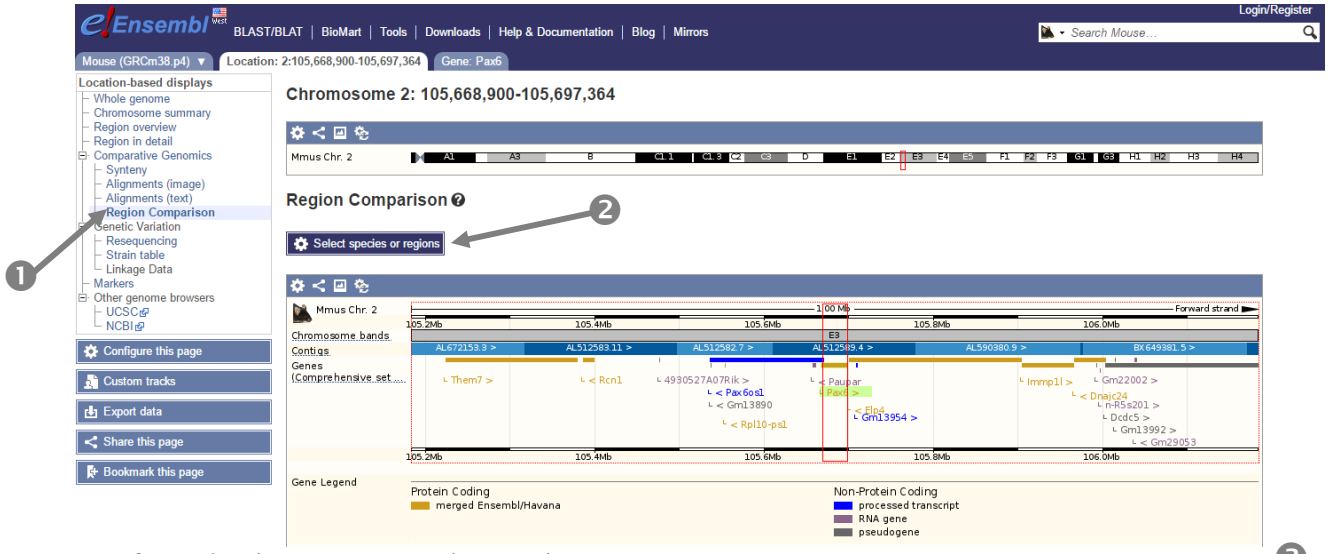
As we've seen, TAIR is the repository for Arabidopsis genomic information. Connect to TAIR at http://www.arabidopsis.org and under the Tools tab, click on GBrowse. The Arabidopsis Information Portal is a different initiative aimed at integrating all data from different source for Arabidopsis: http://araport.org. You can also check out the JBrowse instance!

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. 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 left. *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 second row, called **Location & Maps**, click on "More" to expand the section and then 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.



….after selecting Human as the species

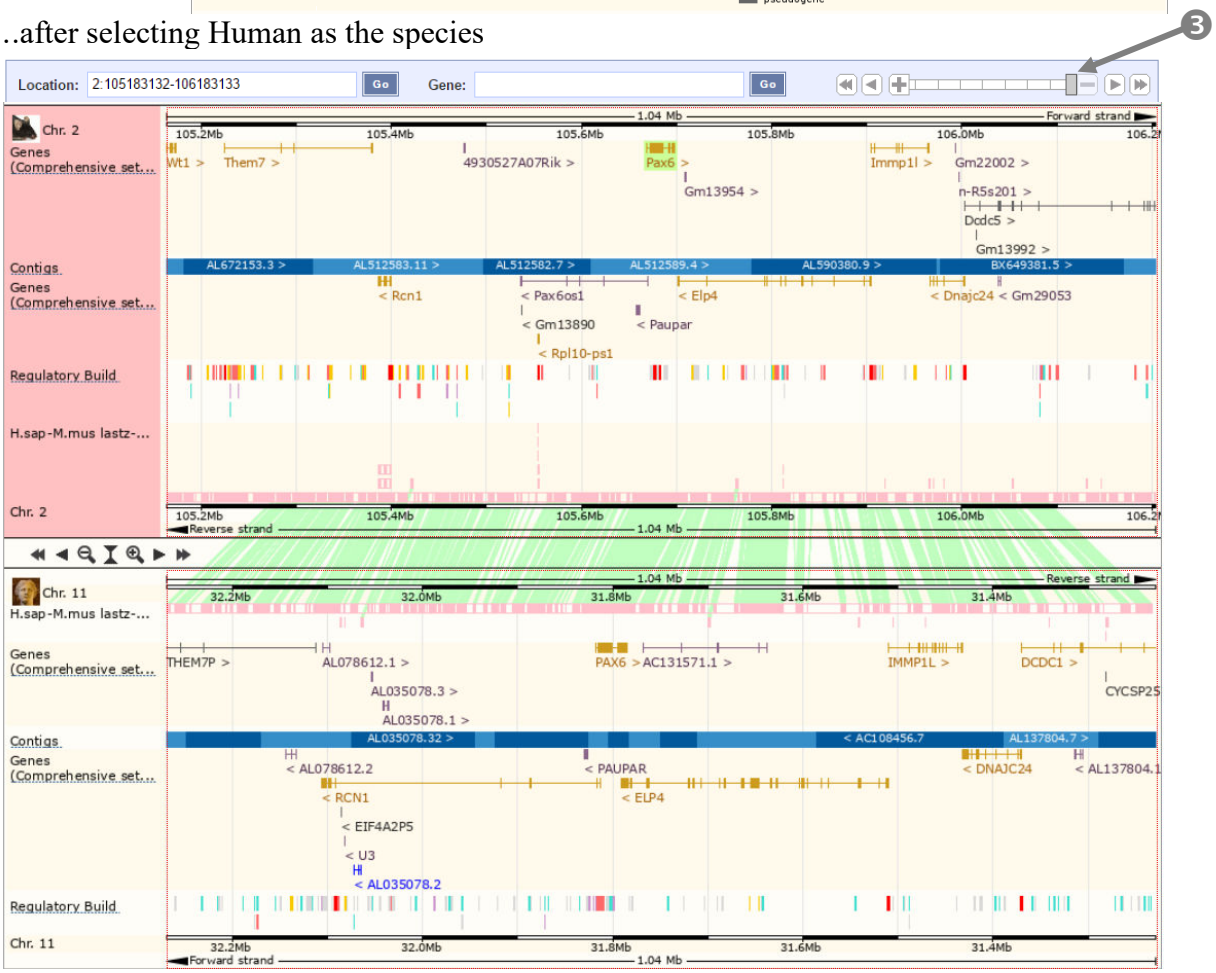


Figure 7: Enseml Genome Browser showing mouse and human Pax6 genomic regions. Click on ❶ Region Comparison then ❷ Select Species or Regions to add e.g. the corresponding human region. Use ❸ zoom slider to zoom out to view 1000000 base pairs at once. Thin diagonal green lines in the bottom panel represent similar chromosomal regions 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** – there may be slight differences depending on version). The diagonal green lines in the bottom panel represent similar regions 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.

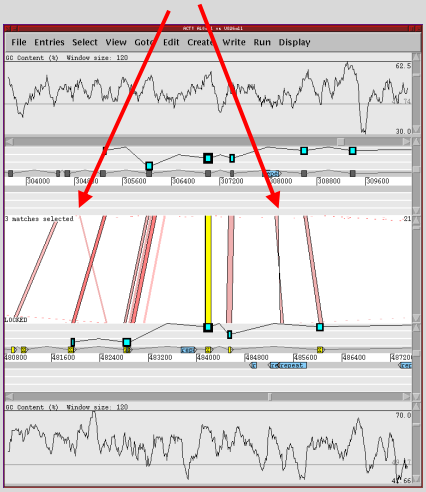
   a. *Would you say that there is synteny between the chromosomal region containing mouse* Pax6 *and human* PAX6*? Why?*

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.

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 (AE007869) versus the *Agrobacterium tumefaciens* strain C58 / ATCC 33970, sub_strain Dupont (AE008688) genomes. Use the defaults. You may need to download the .jnlp and/or .zip file that WebACT provides and run it on your computer with Java, if Java Web Start isn't active in your browser (and add http://www.webact.org:80/ to the Exception Site List in the Configure Java settings).

   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

Chapters 4 and 6 "Substitution Matrices/Blast and Multiple Sequence Alignment" in *Concepts in Bioinformatics and Genomics* by Jamil Momand and Alison McCurdy, Oxford University Press, 2017. pp. 66-88 (Substitution Matrices) and pp. 106-117 (Blast section).

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/nt	All GenBank + EMBL + DDBJ + PDB + RefSeq sequences (but no EST, dbSTS, GSS, WGS,TSA or phase 0, 1 or 2 HTGS sequences).
est	Database of GenBank + EMBL + DDBJ sequences from EST division
refseq_rna	NCBI transcript reference sequences
refseq_representative_genomes	Reference and representative genomes selected from the NCBI Refseq Genomes database
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.
tsa	Transcriptome Shotgun Assembly (TSA) database is an archive of computationally assembled mRNA sequences
sra	Search for sequences associated with a particular SRA (sequence read archive) accession, scientific name, or taxonomic identifier
dbsts	Database of Sequence Tag Site entries from the STS division of GenBank + EMBL + DDBJ.
refseq_genomes	NCBI Refseq genomes across all taxonomy groups. Contains only the top-level sequences, i.e. chromosomal sequences where available (but not the contigs used to assemble them)
wgs	Assemblies of Whole Genome Shotgun sequences