BLAST

(Basic Local Alignment Search Tool)

Note: This is a complete transcript to the powerpoint. It is good to read through this once to understand everything. If you ever need help and just need a quick a quick glance at something, there is a separate, much shorter help guide. It has less explanation and more directions.

Some things before we start:

What you need:

Nothing, everything is provided for you on DSAP and NCBI

Important Terms:

BLAST: Basic Local Alignment Search Tool. Essential, if you feed it a sequence, it will find an "alignment", or a match. Not all matches will be 100% identical, which is natural and normal.

*Blastn: Searches **nucleotide** sequences and matches them with other **nucleotide** sequences in the database.*

*Blastx: Searches **nucleotide** sequences and matches them with other **amino acid** sequences in the database.*

*Blastp: Searches **amino acid** sequences and matches them with other **amino acid** sequences in the database.*

Query: The DNA/amino acid sequence you entered and searched.

Subject: The result from another organism that your query was matched to.

E-Value: Shows **significance** *of search. The **lower the number** (the closer it is to zero), the **more significant it is**. The higher the number, the less significant. The e-value is a "probability" that tells you how "likely" it is to receive a totally nonsensical and irrelevant match. Therefore, a high number is a high "chance" of a nonsensical and irrelevant match- a low number is a very low "chance" of a bad result and therefore a high "chance" of a very good result. It's not exactly probability, but it's best to think of it that way.*

tldr; The closer to zero, the better the e-value.

Description: The entire description of the sequence/protein. This includes the protein, organism, and other information the scientist might include.
Accession Number: Also called sequence ID in the detailed results, this is simply the number assigned to a protein in the database (like a call number for a book).

Toolbox: A program developed by WSSP staff that helps us convert DNA to amino acid.

Reading Frame: The "frame" in which the DNA is read. There are three certain grouping that codons can be grouped into. Only one frame will give the correct amino acid and therefore functioning protein.

Conserved Sequence: A conserved sequence/region is an area of your DNA/amino acid sequence that is identical/near identical to sequences in other organisms. This is important because conservation implies importance. This is seen in DNA sequences, as well as proteins themselves, as you will see in analysis.

First, Review:

1. *Reading sequences. Good sequences should have clear, distinct peaks and minimals N's in your sequence. N's mean that the sequencing machine cannot tell what base it is since it is unclear. Usually you can determine yourself what the base is, and can edit out the N and replace it with the right base.*
    a. *Here's an easy code: A=Green, T=Red, C=Blue, G=Black*
2. *Unreadables: See the cropping guide for a list of unreadable cases and what they mean.*
3. *Your sequence starts at the base after CGGCCGGG(etc.) and ends at the last base of the poly-A tail.*

Congratulations!!! *After cropping and editing your clone, you now have successfully isolated your DNA sequence. But now what?*

Before we can do any research, we need to find out what we're researching! BLAST helps us find what protein our DNA codes for. We run our sequence through three different BLAST search engines; all three serve a different purpose. The basic idea for all three of them are the same, so if you understand the first BLAST (Blastn), you can (mostly) do the three others. We'll go through all three of them.

****Note: Utilize the accompanying ppt for figures and diagrams.***

Blastn

Blastn is the first BLAST search that we will do. ***It matches your DNA nucleotide sequence to other DNA nucleotide sequences.*** *This ensures that your DNA is of* ***good quality*** *(by returning good matches). It will also tell you which segment of your DNA is* ***"conserved"****, or which segment of your DNA has an exact/near exact copy in another organism's DNA. This is significant because* ***conserved segments implies importance.***

We run two Blastn databases, ***nr/nt and est****.*

To run the BLASTn (nr/nt):

1. *On the DSAP page, click the "retrieve sequence" button.*
2. *Copy the sequence.*
3. *Click on the "Blastn" link. Paste the sequence into the box labeled "****Enter accession number(s), gi(s), or FASTA sequence(s)****".*
4. *Scroll to the bottom, and click the big blue BLAST button.*

After letting the page refresh and load for some time, you will see the results page. Slide 7 shows the graphic organizer at the top, and Slide 8 shows the list of results. Each line in the graphic organizer corresponds with the result listed below it.

To read the graphic organizer:

*-****Red segments*** *are incredibly significant matches: they have the highest number of identical bases.*

-The order from highest to lowest significance goes: Red, purple, green, blue, black (as per the key at the top).

-The colored segment shows the ***"conserved"*** *region of the DNA. Your DNA sequence is the thick, red line next to "query". As you can see,* ***not 100% of your sequence will match with others****-many times, it will just be the first half or the second half. This is* ***normal*** *(as in this case).*

To read the listed results:

-The long blue hyperlink is the ***description*** *of the protein. You can click on it to jump to the detailed results (see Slide 9).*

-*You can locate the **E-value** (Expect), the **Query start** (number and base), **Query end** (number and base), **Accession Number** (Sequence ID) and **organism name** (located in the description). These are the answers to the chart in DSAP.*

-*To further understand your sequence, you can take a look at the **dashes/gaps** in the area between the query and subject. A **dash** means an identical match-a **gap** means the bases are not identical. A good match will have minimal gaps. Slide 13 shows examples of gaps/mismatches and their effects on the DNA sequence. **Note that this portion is not necessary for DSAP, but it is better for your understanding of how DNA research works.** If this confuses you, you may choose to ignore it.*

-*Slide 14 shows you a couple reasons why errors in matching DNA sequences come up. Other than your own mistakes (such as bad cropping), simply being unlucky and having a bad sequence may cause errors. In addition, mutations and such across organisms may cause slight differences in DNA even though they code for the same protein. One reason why you may have many gaps and mismatches in your DNA is the way we make cDNA: sometimes, Blastn may attempt to match your exon with an intron region.*

-*Remember that **if you receive incredibly short matches, they are not significant.** Slides 15 and 16 shows examples of what these matches look like, both in the detailed results and graphic organizer.*

*Blastn nr/nt matches the entire nucleotide sequence. We also want to check out Blastn est, since this database matches only the **"expressed sequence tags"**, or the **segments of DNA that are expressed**. The results may be different than in Blastn nr/nt.*

*To run Blastn est, **re-do the steps for Blastn, but before clicking BLAST, find the drop-down menu next to "database" and select expressed sequence tags (est)**. See Slide 10 for reference.*

The analysis of results for Blastn nr/nt and est are identical, as are the answers. Slide 17 shows an example of a good Blastn result.

To answer questions 3 and 4:

3. Your answer is almost always significant: that's what all the analysis above was for!
4. To check if all your results are from the Kingdom Plantae, scroll through the results and simply look for the organisms in the descriptions. It's almost 99% always just plants.

This concludes Blastn!!!! :)

Blastx

*Blastx matches your **nucleotide sequence** to an **amino acid sequence**. The steps to run this search, as well as analysis of results, are identical to Blastn. Look through slides 18-21 as an example. However, there is one major difference in analyzing the results:*

The Reading Frame.

What is a reading frame?

*Take a look at Slide 23. When we receive the DNA sequence, it's all mashed together in one continuous sequence-it hasn't been divided into triplets. **Therefore, we don't know how the codons are grouped**. Slide 23 shows different starting points for the DNA sequences. For a clearer example, Slide 24 shows examples of the different ways we can group codons given one sequence. These three ways are called the three **"reading frames".***

*-Note: We are only working with **positive** reading frames.*

Toolbox

*Toolbox is the intermediary step between Blastx and Blastp. It **converts your nucleotide sequence into a protein sequence**. If you take a look at the results in Blastx, the detailed results are listed in amino acids instead of nucleotide DNA. **However, the Blast program is still using a nucleotide sequence to search**. **We want the Blast program to use an amino acid sequence to search.***

*In addition, Toolbox finds the **ORF**, or **Open Reading Frame**. This is the region of your DNA sequence that is coding. The other two areas, the 5' and 3' areas, are non-coding and are not relevant in searching for our protein. Further details in analysis.*

To use toolbox:

1. *Copy and paste your DNA.*
2. *Select the reading frame given to you in Blastx. This is 99% correct (it is only incorrect if you have a poor sequence and Blastx contained many errors).*
3. *Slide 27 shows the converted amino acid sequence. The **bold, underlined portion is the longest possible ORF**...it's **not necessarily the actual ORF**. Always go to the underlined portion first, since there is a higher chance that is the actual ORF, but it's not 100% true (more like 98% true).*
    a. *If you receive poor Blastp results, the first thing to check is to make sure the other sections of the amino acids aren't the ORF. Simply copy and paste those segments into Blastp and search those too.*
4. ***The ORF starts at the first green M.***

a. *If there is a red asterisk is in front of this M, then you know 100% for sure that this M is the beginning.*
b. *If there is no red asterisk, as is the case in Slide 27, then it may be an internal M and not the start (the start may be cut off our sequence due to an error).*
c. ***For now, do not worry about scenario b: we will deal with these cases later on.***
d. ***The ORF ends at the red asterisk.***

5. *When you highlight the ORF, you can see the* ***nucleotide bases*** *appear. Now you have all the info to answer DSAP questions!*

Blastp

Once again, running a Blastp search is the same as a Blastn or Blastx. The results are also identical to analyze.

One important note is to ***compare the start and end bases of the query*** *(your sequence) and the subject (the matched sequence). If they both start at the same base, then you have the correct ORF. If not, then there may be a different ORF. **You will not see this in the early practice clones-we will discuss them when they do pop up.*

Likely start and end:

-Remember that the ORF started with an M. If your Blastp results also start with an M1, then your ORF codes for the beginning.

-Remember that the ORF ends with an asterisk. It is hard, of course, to memorize the many different stop codons. Therefore, if there is an asterisk in toolbox that you highlighted, it is safe to say that your sequence codes for the end as well.

-You can double-check this on the Blastp graphic organizer.

-Slide 32 shows examples to the answers.

Blastp v.s. Blastx:

This is simple and straightforward: the accession numbers should be the same (if not, the proteins should be the same), the e-values should be close to each other. This ensures that Blastx and Blastp returned the same results, a confirmation that we did find a protein that the DNA codes for.

-*Slide 33 shows examples to the answers.*

NOW YOU ARE DONE WITH BLAST!!!
You're welcome for the images. :)

Pro Tip: Sometimes, it's easier to start copying/pasting everything first without understanding everything. Then, while you're doing it, you'll start to see everything fall in place and understand what's going on.