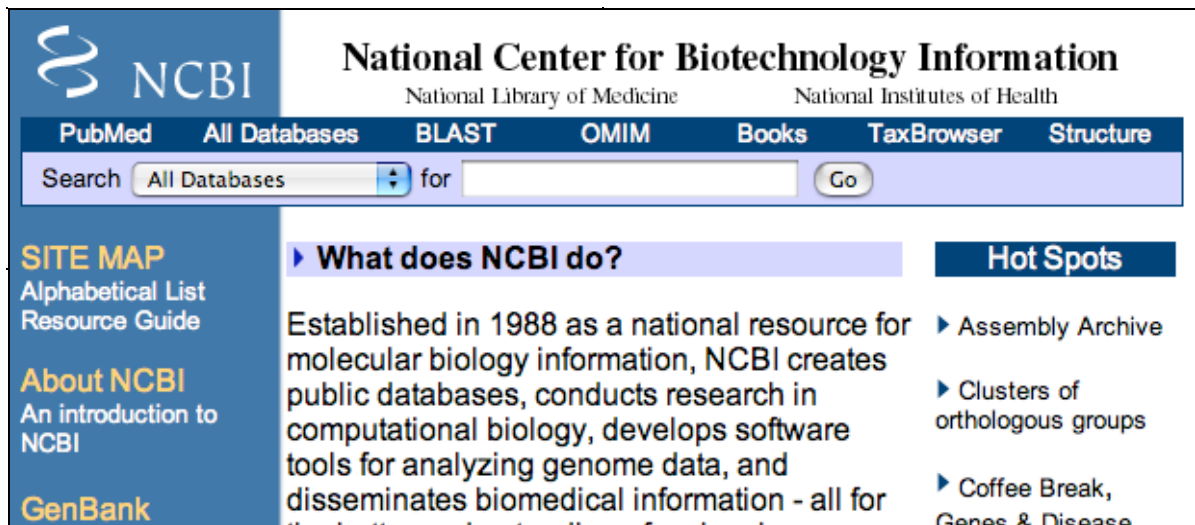


Introduction to the NCBI website: PubMed and Blast.

Biological sequence data and journal articles are collected, indexed, and made available by the National Center for Biotechnology Information (NCBI). NCBI is a unit of the National Library of Medicine (NLM) at the National Institutes of Health (NIH). Because it is a part of the NIH, the collections of sequence data and journal articles are available free to anyone at <http://www.ncbi.nlm.nih.gov/>. This is what the NCBI home page (currently) looks like....



NCBI provides tools for searching and downloading the databases it maintains through the web portal NCBI Entrez. PubMed is searched with text queries using the Entrez portal. PubMed is an index of thousands of biological journals going back as far as 1950. It also contains thousands of full-length articles in PDF format available for free download in a collection called PubMed Central.

NCBI also contains a collection of biological sequence databases. These are informally referred to as GenBank. The biological sequences are divided into DNA sequences (which includes RNA sequences) including GenBank proper, Protein sequence, and Genome Sequence. The sequence database collection is the result of collaboration between NCBI, DDBJ (Japan), and EMBL (Europe). Although the file formats and search tools may differ between the three repositories, they are essentially redundant at the data level. Most data in GenBank are in the public domain although some sequence data are patented.

GenBank sequence is usually accessed in one of two ways. A simple text search can be used to find sequences by name, authors, and other supporting information. A more sophisticated search of GenBank uses a sequence query and a collection of tools called Blast. Blast will be described in detail in the second part of this tutorial.

Other useful and interesting databases maintained by NCBI:

The Entrez portal also includes several databases that may people find useful. We will cover only one of these in this tutorial.

James Burnette and Susan Wessler, Department of Plant Biology, The University of Georgia, Athens GA. 706-542-4581. jburnette@plantbio.uga.edu

- TaxBrowser - This database provides taxonomic information on most extant and extinct organisms. This is useful to explore the relationships between organisms. An easier to understand (but less complete) taxonomy web site is the Tree of Life website.

<http://www.tolweb.org/tree/>.

- Books - Books is a virtual library of out-of-print editions of textbooks. Although out-of-print, many are still useful and all are free.

- OMIM - Online Mendelian Inheritance in Man. - OMIM is a database of articles on human genes associated with diseases and medical conditions. Each article is hand curated by people who read and summarize journal articles. This database is incredibly rich with information on human genes, diseases, and population genetics. An OMIM example will be covered later in this tutorial.

Let's begin our tour by visiting the PubMed site and then move on to the BLAST site where we will be spending a great deal of our time today and in the rest of the course.

I. **PubMed:** Literature searches about a biological problem are very easy. PubMed makes the index available on its website with no access limitations. You can use PubMed (and Blast) from any computer and with any internet connection. (There is even a special access page for mobile phones.)

While searching the database and reading abstracts is free, accessing a full length article may require a subscription with the article's publisher. Many universities will have subscriptions and access is easy if you are on an university network. Many articles will be freely available either directly from the publisher, or through PubMed Central.

Steps for a PubMed search:

1. Open NCBI in a web browser by going to the NCBI home page and click on PubMed in the bar. To get to the PubMed home page click on PubMed which is part of the main menu at the top.

NCBI National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search All Databases for [] Go

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to
NCBI

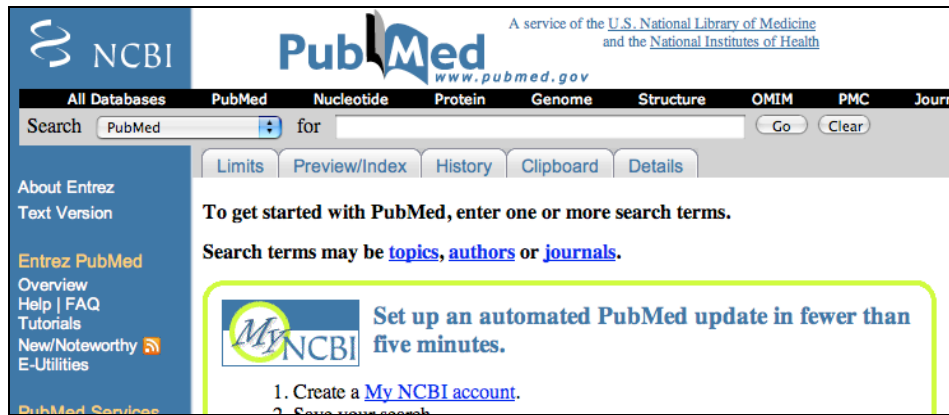
GenBank

What does NCBI do?
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for

Hot Spots

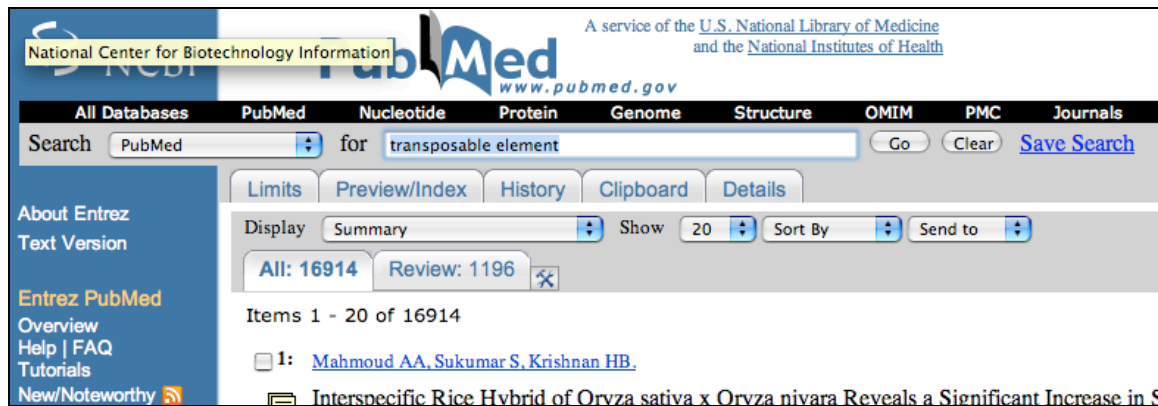
- ▶ Assembly Archive
- ▶ Clusters of orthologous groups
- ▶ Coffee Break, Genes & Disease

This is the PubMed homepage...



2. Think about the topic you want to search. You can use keywords, author last names, journal titles, publication year, or institution.

For the first search enter 'transposable element' and click 'Go.' The result of the search is shown below.



3. Before we discuss the results, click on the 'Details' tab. This will show you the details of the search that was performed by PubMed's search engine.

The screenshot shows the NCBI PubMed search interface. The search bar contains the query 'transposable element'. The results page displays the following information:

- Query Translation:** ("dna transposable elements"[TIAB] NOT Medline[SB]) OR "dna transposable elements"[MeSH Terms] OR transposable element[Text Word]
- Result:** 16914
- Translations:**

transposable element	("dna transposable elements"[TIAB] NOT Medline[SB]) OR "dna transposable elements"[MeSH Terms] OR transposable element[Text Word]
----------------------	---
- Database:** PubMed
- User query:** transposable element

While you thought you were just searching "transposable element" PubMed was actually using these search terms (with added comments):

("dna transposable elements"[TIAB] NOT Medline[SB]) •Search titles and abstracts, not the medline subset
 OR "dna transposable elements"[MeSH Terms] •Search Medline Subject Headings
 OR transposable element[Text Word] •Search all text

As you can see, the simple query 'transposable element' is expanded into a more structured query by PubMed. MeSH is a controlled vocabulary for indexing PubMed. Curators at NCBI and journal editors assign these keywords based on suggestions by authors. Because of this query expansion it is a good idea to check the 'Details' tab whenever a search gives no results or unexpected results.

4. Click on the Browser's Back Button. The icons and other details of the results list like the one shown below will be discussed in class.

Display Summary Show 20 Sort By Send to

All: 16914 Review: 1196

Items 1 - 20 of 16914 Page 1 of 846 Next

1: [Mahmoud AA, Sukumar S, Krishnan HB.](#) Related Articles, Links
 Interspecific Rice Hybrid of *Oryza sativa* x *Oryza nivara* Reveals a Significant Increase in Seed Protein Content.
 J Agric Food Chem. 2007 Dec 29; [Epub ahead of print]
 PMID: 18163552 [PubMed - as supplied by publisher]

2: [Van K, Onoda S, Kim MY, Kim KD, Lee SH.](#) Related Articles, Links
 Allelic variation of the Waxy gene in foxtail millet [*Setaria italica* (L.) P. Beauv.] by single nucleotide polymorphisms.
 Mol Genet Genomics. 2007 Dec 19; [Epub ahead of print]
 PMID: 18157676 [PubMed - as supplied by publisher]

3: [Mukherjee S, Chakraborty R.](#) Related Articles, Links
 Conjugation potential and class 1 integron carriage of resident plasmids in river water copiotrophs.
 Acta Microbiol Immunol Hung. 2007 Dec;54(4):379-97.
 PMID: 18088011 [PubMed - indexed for MEDLINE]

4: [Fontanillas P, Hartl DL, Reuter M.](#) Related Articles, Links
 Genome organization and gene expression shape the transposable element distribution in the *Drosophila melanogaster* euchromatin.
 PLoS Genet. 2007 Nov 30;3(11):e210. Epub 2007 Oct 10.
 PMID: 18081425 [PubMed - in process]

5: [Metcalfe CJ, Bulazel KV, Ferreri GC, Schroeder-Reiter E, Wanner G, Rens W, Obergfell C, Eldridge MD, O'Neill RJ.](#) Related Articles, Links
 Genomic instability within centromeres of interspecific marsupial hybrids.
 Genetics. 2007 Dec;177(4):2507-17.
 PMID: 18073443 [PubMed - in process]

5. Click on one of the underlined authors (in blue). This will give you detailed information about the article including the abstract. The abstract provides a detailed summary of the paper. On the right are two icons. Clicking on either of those will take you to a download page for the full article. The Related Links section is also a useful area to help refine searches and will be discussed in class.

Display AbstractPlus Show 20 Sort By Send to

All: 1 Review: 0

1: [BMC Evol Biol.](#) 2007 Aug 29;7:152.

Full text free on... **BioMed Central** **FREE full text article** in PubMed Central Links

Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*.

[Zuccolo A, Sebastian A, Talaq J, Yu Y, Kim H, Collura K, Kudrna D, Wing RA.](#)

Arizona Genomics Institute, Department of Plant Sciences, BIOS Institute, University of Arizona, Tucson, AZ 85721, USA. azuccolo@ag.arizona.edu

BACKGROUND: The genus *Oryza* is composed of 10 distinct genome types, 6 diploid and 4 polyploid, and includes the world's most important food crop - rice (*Oryza sativa* [AA]). Genome size variation in the *Oryza* is more than 3-fold and ranges from 357 Mbp in *Oryza glaberrima* [AA] to 1283 Mbp in the polyploid *Oryza ridleyi* [HHJJ]. Because repetitive elements are known to play a significant role in genome size variation, we constructed random sheared small insert genomic libraries from 12 representative *Oryza* species and conducted a comprehensive study of the repetitive element composition, distribution and phylogeny in this genus. Particular attention was paid to the role played by the most important classes of transposable elements (Long Terminal Repeats Retrotransposons, Long interspersed Nuclear Elements, helitrons, DNA transposable elements) in shaping these genomes and in their contributing to genome size variation. RESULTS: We identified the elements primarily responsible for the most strikingly genome size variation in *Oryza*. We demonstrated how

Related Links

- Evolutionary history of *Oryza sativa* LTR retrotransposons: a preliminary survey of the rice genome sequ [BMC Genomics. 2004]
- Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in th [Plant J. 2007]
- Differential lineage-specific amplification of transposable elements is responsible for genome size variat [Genome Res. 2006]
- Long terminal repeat retrotransposons of *Oryza sativa*. [Genome Biol. 2002]
- Dasheng and RIRE2. A nonautonomous long terminal repeat element and its putative autonomous partner i [Plant Physiol. 2002]

See all Related Articles...

6. Learning to search PubMed and all of the features takes time and practice. Research a topic that is interesting to you.

II. OMIM

This short introduction will get you acquainted with OMIM. With OMIM you can learn about a genetic disease, find examples for class and tests.

1. Click "OMIM" on the Entez Portal.

The screenshot shows the NCBI homepage. The navigation bar includes links for PubMed, All Databases, BLAST, OMIM (circled in red), Books, TaxBrowser, and Structure. Below the navigation bar is a search box with the text "Search All Databases for" and a "Go" button. The main content area is divided into several sections: "What does NCBI do?", "Hot Spots", "GenBank vs. RefSeq", and "Site Map".

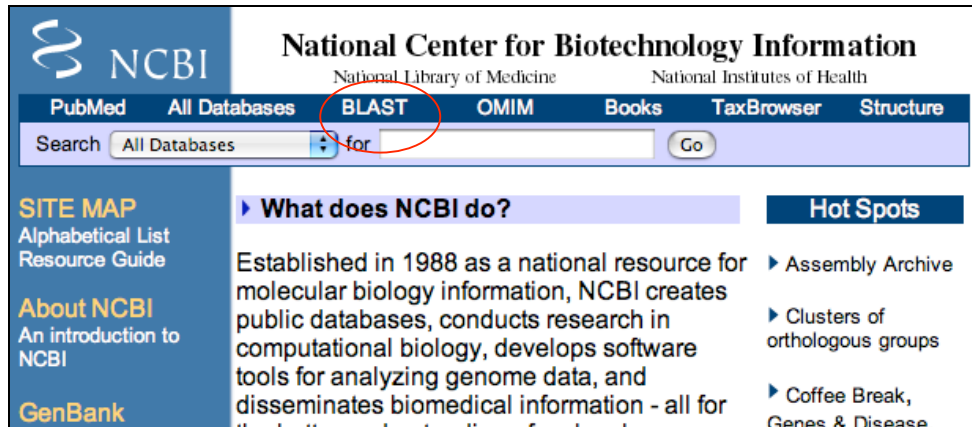
2. Enter "transposable element" in the text box and Click "Go". You search OMIM with text queries similar to PubMed. If you wanted to search on a specific disease you would enter the disease name.

The screenshot shows the OMIM search interface. The search bar contains the text "OMIM" and "transposable element" (circled in red). The "Go" button is also circled in red. Below the search bar are buttons for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The main content area displays a list of search results and a section titled "OMIM™ - Online Mendelian Inheritance in Man™".

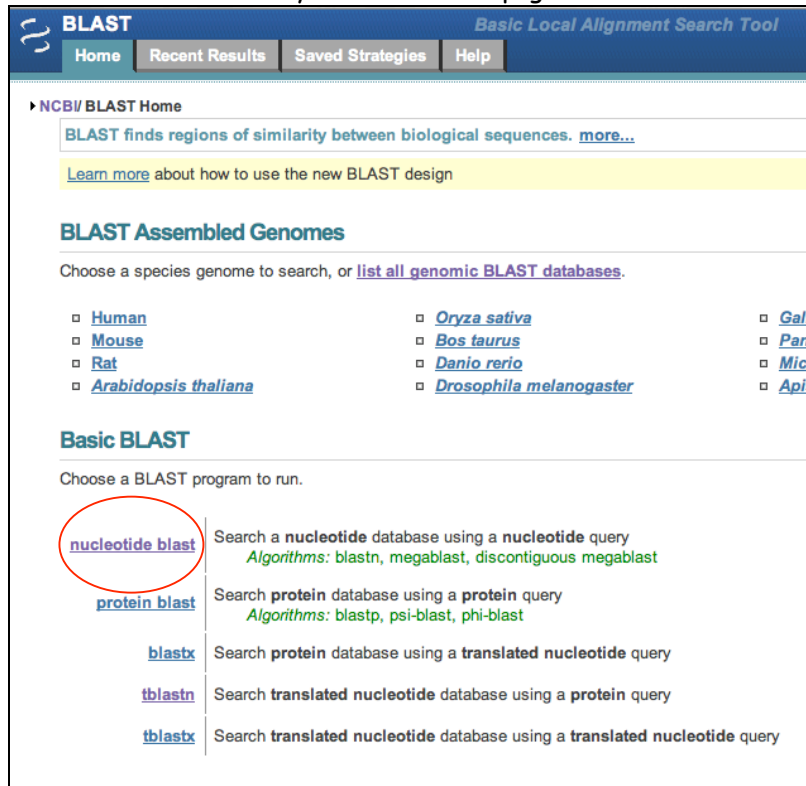
III. Introduction to **Blast**:

You will use Blast a lot in the workshop. It is the major biological sequence search tool for DNA, RNA, and protein databases. Whole genomes can be searched using Blast.

Access Blast by clicking on the Blast link on the NCBI home page.



The Blast link will take you to the Blast page and to the Basic Blast Menu.



There are six different versions of BLAST because you can use a nucleotide sequence or protein sequence to query nucleotide or protein sequence databases, This is summarized in the screenshot above. Today, we will give three of these search tools a test drive: nucleotide blast, protein blast, and tblastn.

- A. **Nucleotide Blast:** This is the most straightforward type of search. You begin with a nucleotide sequence you want to know more about (the query) and "blast" it against a nucleotide database (the subject).

You can learn a lot about your query sequence with a blast including:

- Are there publications that already report information about this sequence (have you been "scooped")?
 - Where is the sequence located in the genome (more on location in class)?
 - Is the sequence found in genomes of closely related organisms?
 - Does it code for an RNA and/or a protein? If so is anything known about its function?
1. Select 'nucleotide blast.' Copy and paste the following sequence in the Query text window (Enter accession number...):

>mPing

```
ggccagtcac aatgggggtt tcaactggtgt gtcatgcaca ttaataggg gtaagactga
ataaaaaatg attatttgca tgaatgggg atgagagaga aggaaagagt ttcacctgg
tgaactcgt cagcgtcgtt tccaagtctt cggtaacaga gtgaaacccc cgttgaggcc
gattcgttc attaccgga tctcttgcgt ccgcctccgc cgtgcgacct ccgcatttc
ccgcgccgcg ccggattttg ggtacaaatg atcccagcaa ctgtatcaa ttaaagtctt
tgcttagtct tggaaacgtc aaagtgaac ccctccactg tggggattgt ttcataaaag
atctcatttg agagaagatg gtataatatt ttgggtagcc gtgcaatgac actagccatt
gtgactggcc
```

The screenshot shows the 'Enter Query Sequence' interface. The main text area contains the sequence:


```
ggccagtcac aatgggggtt tcaactggtgt gtcatgcaca ttaataggg gtaagactga
ataaaaaatg attatttgca tgaatgggg atgagagaga aggaaagagt ttcacctgg
tgaactcgt cagcgtcgtt tccaagtctt cggtaacaga gtgaaacccc cgttgaggcc
gattcgttc attaccgga tctcttgcgt ccgcctccgc cgtgcgacct ccgcatttc
ccgcgccgcg ccggattttg ggtacaaatg atcccagcaa ctgtatcaa ttaaagtctt
tgcttagtct tggaaacgtc aaagtgaac ccctccactg tggggattgt ttcataaaag
atctcatttg agagaagatg gtataatatt ttgggtagcc gtgcaatgac actagccatt
gtgactggcc
```

 Below the text area, there are two options: 'Or, upload file' with a 'Choose File' button and 'no file selected' text, and 'Job Title' with a text input field and the prompt 'Enter a descriptive title for your BLAST search'. To the right, there is a 'Query subrange' section with 'From' and 'To' input fields.

2. Under "Choose Search Set" select "Others" and the drop down list changes to "Nucleotide Collection (nr/nt)." This is the complete non-redundant nucleotide database.

Choose Search Set

Database	<input type="radio"/> Human genomic + transcript <input type="radio"/> Mouse genomic + transcript <input checked="" type="radio"/> Others (nr etc.):
	<div style="border: 1px solid #ccc; padding: 2px; display: inline-block;">Nucleotide collection (nr/nt)</div>
Organism <small>Optional</small>	<input style="width: 100%;" type="text" value="Enter organism name or id--completions will be suggested"/>
	<small>Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.</small>
Entrez Query <small>Optional</small>	<input style="width: 100%;" type="text" value="Enter an Entrez query to limit search"/>

3. The next section gives you three options for a nucleotide blast. Choose megablast (default) for now.

Program Selection

Optimize for	<input checked="" type="radio"/> Highly similar sequences (megablast) <input type="radio"/> More dissimilar sequences (discontiguous megablast) <input type="radio"/> Somewhat similar sequences (blastn)
	<small>Choose a BLAST algorithm</small>

4. Select the "Blast" button. What you see below is called the queue page:

▶ [NCBI/BLAST/blastn/Formatting Results - S9WZE65Y013](#) [\[Formatting options\]](#)

Job Title: lcl|21740 (430 letters)

Request ID	S9WZE65Y013
Status	Searching
Submitted at	Wed Jan 9 11:18:54 2008
Current time	Wed Jan 9 11:18:56 2008
Time since submission	00:00:01

This page will be automatically updated in **10** seconds


```

Query  361  ATTCATTTGAGAGAAGATGGTATAATATTTTGGGTAGCCGTGCAATGACACTAGCCATT  420
      |||
Sbjct  3992  ATTCATTTGAGAGAAGATGGTATAATATTTTGGGTAGCCGTGCAATGACACTAGCCATT  4051

Query  421  GTGACTGGCC  430
      |||
Sbjct  4052  GTGACTGGCC  4061

```

A short discussion on how Blast works.

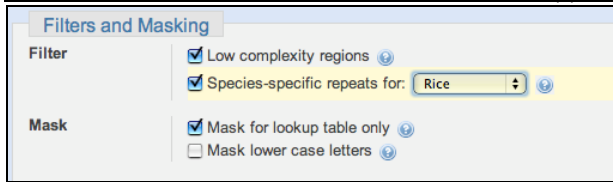
Blast takes the query sequence and divides it into "words" based on the word size parameter (the default is usually "fine"). For a megablast query the default (and minimum) is a size of 28. The algorithm then takes these "words" and runs them against a database. When an exact match occurs, the program attempts to extend the alignment in each direction. If the alignment extends then a score is calculated and as long as the score remains above a threshold the alignment continues. If a mismatch occurs the score decreases, but as long as the score remains above threshold the mismatch is allowed. Word size can be changed. Long word sizes increase stringency.

The threshold is determined by the Expect value in the "Algorithm Parameters" tab on the Blast page. The default Expect value is 10. This means that you expect to find 10 matches to your query in randomly generated sequence. Blast uses this value, the size of the query sequence, and the size of the database (called the search space) to calculate a threshold on 10 random matches and then reports only hits that score better than the random model. Lowering the Expect value increases the stringency of the search.

While extending the alignment Blast may encounter a series of mismatched nucleotides. Blast will try to skip over the mismatch region (called opening a gap) to see if the alignment begins again. If the alignment begins again, Blast will continue. If the alignment does not begin again, the alignment process stops and Blast reports the hit. Opening a gap is penalized heavily. Extending a gap is also penalized. The process of opening gaps is necessary to allow for small insertion mutations (called indels) that occur fairly frequently in a genome.

An important point for searches involving Transposable Elements: The ubiquitous low complexity filter.

Repeat the mega blast but with the following modification. Select the "Algorithm Parameters" and go to the Filters and Masking section. Check 'Species-specific repeats for:' and select Rice. Run the Blast. What happened?



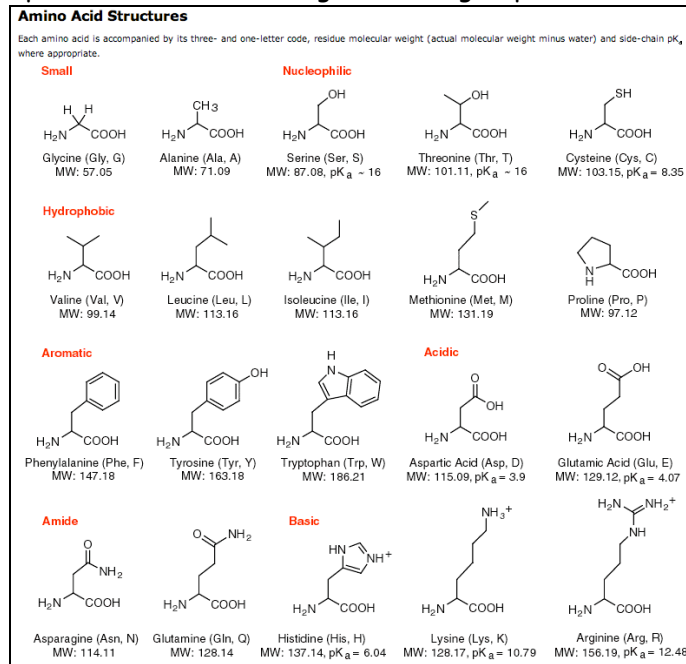
Filters and Masking	
Filter	<input checked="" type="checkbox"/> Low complexity regions
	<input checked="" type="checkbox"/> Species-specific repeats for: Rice
Mask	<input checked="" type="checkbox"/> Mask for lookup table only
	<input type="checkbox"/> Mask lower case letters

We will discuss low complexity, filtering, and masking.

B. Protein Blast:

A protein blast utilizes an amino acid sequence query from the user as the input and searches a protein database. This is often useful to determine whether the sequence already exists in the database or to predict the function of the predicted protein. The steps for submitting a query are similar to a nucleotide blast and the algorithm is essentially the same.

There is one key difference in the protein vs. nucleotide algorithm. When a nucleotide is compared to a nucleotide only matches between the same base are allowed (A→A, G→G, etc). In contrast, some amino acids have similar chemical properties. For example asparagine (asp) and glutamine (glu) have the same functional group with glutamine having a slightly longer side chain due to an extra methyl group. Asp and glu are often interchangeable without detriment to protein function. The figure below groups the amino acids by functionality.



(www.neb.com)

To score similar amino acid matches, blast uses a look-up table called a BLOSUM matrix. This table contains all possible amino acid matches and a score to use for each. The default matrix is BLOSUM62.

Common groupings of the amino acids (from

<http://www.uky.edu/Classes/BIO/520/BIO520WWW/blosum62.htm>):

G,A,V,L,I, M	aliphatic (though some would not include G)
S,T,C	hydroxyl, sulfhydryl, polar
N,Q	amide side chains
F,W,Y	aromatic
H,K,R	basic
D,E	acidic

1. Open a protein blast from the blast home page

(<http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>), and choose protein blast.

Copy-and-paste this sequence into the window. If you also copy the top line preceding the amino acid sequence the search will be given a job title.

```
>Pong sequence
GSIDCMHWIWENGP TAWKGQYCRGDH GKPTIILEAIASQDLWIWHAF
FGVAGSNNNDINVLNQSDVFNDVLQGKAPEVQFTLNGTTYNMGYYLAD
EIYPEWATFVK TISMPQGEKRK LFAQH Q
```

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

>Pong sequence
 GSIDCMHWIWENGP TAWKGQYCRGDH GKPTIILEAIASQDLWIWHAF
 FGVAGSNNNDINVLNQSDVFNDVLQGKAPEVQFTLNGTTYNMGYYLAD
 EIYPEWATFVK TISMPQGEKRK LFAQH Q

Or, upload file no file selected

Job Title
 Enter a descriptive title for your BLAST search

2. Run the Blast with all default parameters. The queue screen will report that it found a similarity between your query sequence and the Protein Family (PFam) database. This suggests that the query sequence came from a plant TE protein. Makes sense since Pong is a TE.

Job Title: Pong sequence

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

Request ID: SA3V1FMA011
 Status: Searching
 Submitted at: Wed Jan 9 13:16:01 2008
 Current time: Wed Jan 9 13:16:06 2008
 Time since submission: 00:00:05

Query sequence: [(local sequence)|c|3749]
 Pong sequence

Concise Result Full Result Show Search Information

Click on the colored bar for a conserved domain to view your query sequence within the multiple sequence alignment for that domain. To see only the sequences used to generate the domain, click on its PSMID in the tabular summary.

PSMID	Title	Psamid	Multi-Dom	E-value
pfam04827	Plant_tran, Plant transposon protein. This family contains plant transposas...	08502	No	9e-20

3. The results page is similar in organization to the nucleotide blast results page. Here is the first alignment reported (line numbers were added for discussion). Note in this alignment that when two similar amino acids match a '+' is used.

```
1 >gb|AAx92907.1| transposon protein, putative, ping/pong/SNOOPY sub-class [Oryza
  sativa (japonica cultivar-group)]
2 Length=443
3 Score = 203 bits (517), Expect = 2e-51, Method: Composition-based stats.
4 Identities = 88/122 (72%), Positives = 103/122 (84%), Gaps = 0/122 (0%)

5 Query 1 GSIDCMHWIWENGP TAWKGQYCRGDH GKPTIILEAIASQDLWIWHAF FGVAGSNNNDINVL 60
6 GSIDCMHW WE PTAW GQ+ RGD+G PTIILEA+AS DL IWHAFFGVAGSNNNDINVL
7 Sbjct 158 GSIDCMHWRWEKCP TAWSGQFTRGDYGVPTIILEAVASYDLRIWHAFFGVAGSNNNDINVL 217

8 Query 61 NQSDVFNDVLQGKAPEVQFTLNGTTYNMGYYLADEIYPEWATFVK TISMPQGEKRK LFAQ 120
9 NQS +F DVL+G AP+V+F++NG Y+ GYYLA+ IYPEWA FVK+I +PQ EK KL+AQ
10 Sbjct 218 NQSPFLFDV LKGDAPQVKF SVNGNEYSTGYLLANGIYPEWAAFVKS IHL PQT EKH KLYAQ 277

11 Query 121 HQ 122
12 +Q
13 Sbjct 278 YQ 279
```

C. tblastn:

This type of blast takes a protein query sequence and blasts it against a nucleotide database. This is incredibly useful because:

1. it can find the location of the protein in a genome.
2. it can find similar sequences in the genome.
3. it can find similar sequences in related genomes.

To search a nucleotide database with a protein query, the database must first be translated. NCBI stores the nucleotide databases translated in 6 frames.

Why 6 frames?

1. Start at the Blast page and click on *tblastn*, the fourth choice down.

2. Enter the query sequence. Remember, this process compares a sequence of amino acids against sequences in existing genomes.

>Pong sequence
 GSIDCMHWIWENGP TAWKGQYCRGDH GKPTIILEAIASQDLWIWHAF
 FGVAGSNNDINVLNQSDVFNDVLQGKAPEVQFTLN GTTYNMGYLAD
 EIYPEWATFVKTISMPQGEKRLKFAQH Q

5. Go the bottom and click on BLAST! The Algorithm parameters are similar to the nucleotide blast and protein blast search. They serve the same functions here.

BLAST Search database **nr** using **Tblastn (search translated nucleotide database using a protein query)**

Show results in a new window

▶ [Algorithm parameters](#)

6. **Results:** Will be discussed in class, but by now you should be able to read this page yourself.

