

BIO 152 Principles of Biology III: Molecules & Cells

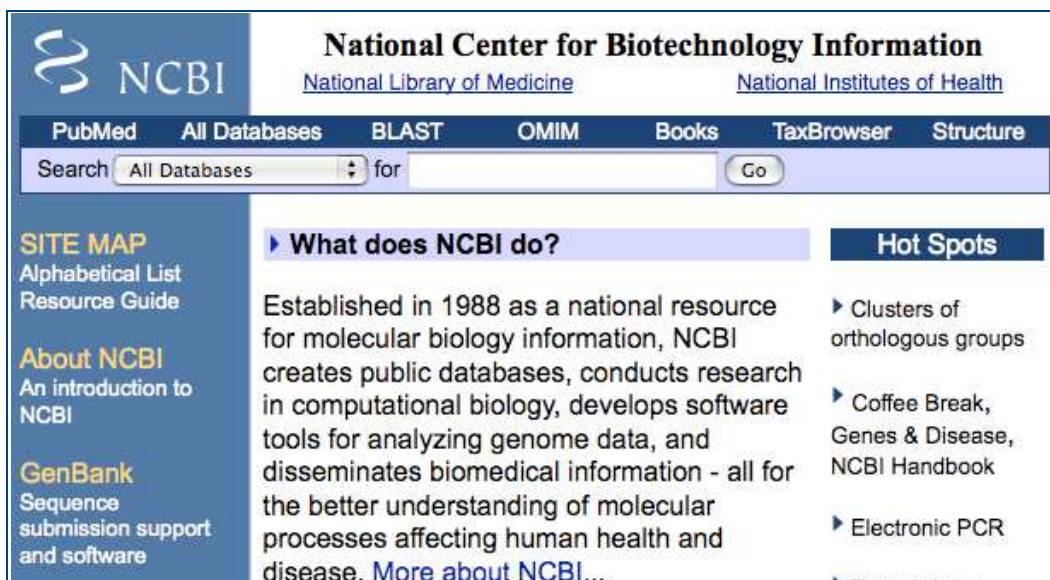
Acquiring information from NCBI (PubMed/Bookshelf/OMIM)

Note: This material is adapted from *Web-based Bioinformatics Tutorials: Exploring Genomes* by Paul Young.

It is the responsibility of scientists to communicate their findings with their peers and also with the world at large. Researchers must relate appropriate publications to the problem at hand and use the findings of others to help direct their own progress. One of the most difficult challenges for today's researcher is to keep up with the current literature. In this tutorial, we will take a closer look at the primary and secondary literature databases at National Center for Biotechnology Information (NCBI). Within the Entrez program at NCBI is the PubMed database, which is produced in collaboration with the National Library of Medicine MEDLINE database. These databases are also cross-referenced to textbooks in the form of key word searches. We will also explore the reference information that is available in the Online Mendelian Inheritance in Man (OMIM) database at NCBI.

The goal of this material and the accompanying bioinformatics assignment is to provide you with practice in accessing information in both the primary literature and textbook databases housed at NCBI. These databases provide a great resource for information about many scientific topics. Typically, we will focus on genetic recombination in eukaryotes: we have touched on in relation to inheritance during Unit One, but many of you may have further questions about the mechanisms involved. Knowledge of both the bioinformatics tools and topical material related to these explorations will not be assessed on Exam I, but may be tested on the Final Exam.

All the databases we will be working with can be freely accessed from the main NCBI site: <http://www.ncbi.nlm.nih.gov/>. Click on the links in the dark blue bar to reach specific database home pages.



The screenshot displays the NCBI website interface. At the top left is the NCBI logo. The main header reads "National Center for Biotechnology Information" with sub-links for "National Library of Medicine" and "National Institutes of Health". Below this is a dark blue navigation bar with tabs for "PubMed", "All Databases", "BLAST", "OMIM", "Books", "TaxBrowser", and "Structure". A search bar is located below the navigation bar, with a dropdown menu set to "All Databases" and a "Go" button. On the left side, there is a sidebar with sections: "SITE MAP" (Alphabetical List, Resource Guide), "About NCBI" (An introduction to NCBI), and "GenBank" (Sequence submission support and software). The main content area features a "What does NCBI do?" section, which states: "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 the better understanding of molecular processes affecting human health and disease. More about NCBI...". To the right of this section is a "Hot Spots" section with links to "Clusters of orthologous groups", "Coffee Break, Genes & Disease, NCBI Handbook", and "Electronic PCR". A blue arrow on the right side of the screenshot points from the text above towards the "Hot Spots" section.

PubMed

To start exploring, type “recombination” into the search window. How can we restrict our search? Try “genetic recombination”, “genetic recombination in eukaryotes”, or “genetic recombination in mammals”. In every case, your search should retrieve over 100,000 entries, some of which don’t seem to have much to do with the kind of recombination we studied in class so far.

Let's have a look at how our terms might be related in the formal subject headings within the National Library of Medicine. These Medical Subject Headings (MeSH) are used to classify publications. Click on MeSH Database under PubMed Services on the left panel. At the new screen, type “recombination” in the search window and press Go. You should get a result like this, with a definition and key terms highlighted near the top:

The screenshot shows the MeSH database search interface. The search bar contains "genetic recombination" and the results are displayed in a list. The first result is "Recombination, Genetic", which is expanded to show its definition and related terms. The definition is: "Production of new arrangements of DNA by various mechanisms such as assortment and segregation, CROSSING OVER; GENE CONVERSION; GENETIC TRANSFORMATION; GENETIC CONJUGATION; GENETIC TRANSDUCTION; or mixed infection of viruses. Year introduced: 1968". Below the definition, there is a list of related terms: "Conjugation, Genetic", "Crossing Over, Genetic", "Gene Conversion", "Gene Fusion", "Oncogene Fusion", "Gene Transfer, Horizontal", "Sister Chromatid Exchange", "Transduction, Genetic", "Transfection", "Transformation, Bacterial", "Transformation, Genetic", and "Transformation, Bacterial".

Clicking on the first entry (“Recombination, Genetic”) will take you to an entry-specific screen, where at the very bottom of the screen, you can see various subheadings:

[All MeSH Categories](#)

[Biological Sciences Category](#)

[Genetic Processes](#)

Recombination, Genetic

[Conjugation, Genetic](#)

[Crossing Over, Genetic](#)

[Gene Conversion](#)

[Gene Fusion](#)

[Oncogene Fusion](#)

[Gene Transfer, Horizontal](#)

[Sister Chromatid Exchange](#)

[Transduction, Genetic](#)

[Transfection](#)

[Transformation, Bacterial](#)

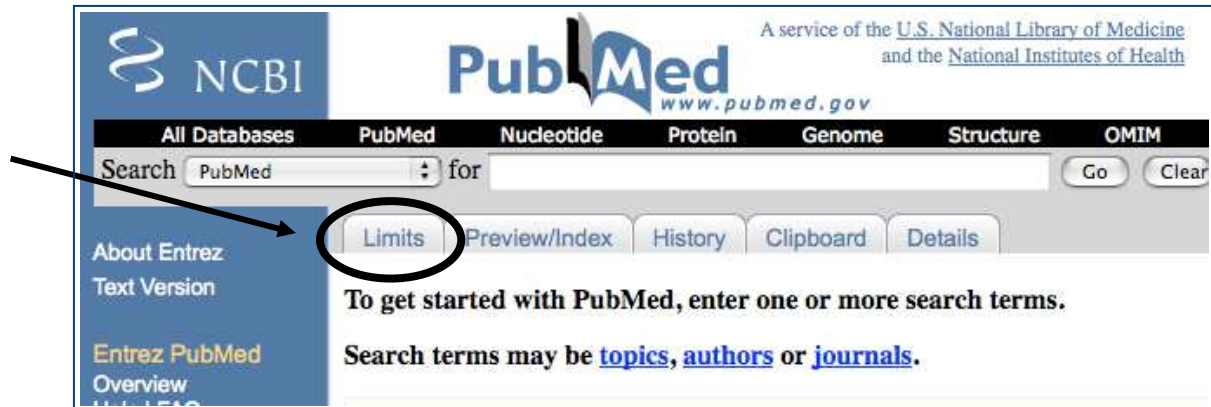
[Transformation, Genetic](#)

[Transformation, Bacterial](#)

This is more like the list of topics relevant to recombination that you would learn in a genetics course. Having a look at these terms sometimes helps narrow your search field by presenting you with some alternative (and narrower) terms. Perhaps we should try “Crossing Over

(Genetics)” in a PubMed search? This is narrower in scope and focuses on the chromosomal recombination process.

We can also restrict a search by organism (e.g., “crossing over AND yeast”) or by author (e.g., recombination AND Hartwell LH”). Another option is to click on the Limits tab, where you can restrict your search by publication date. You could also choose to look at review articles first (select under “Type of Article”); these papers are published in scholarly journals but typically provide an overview of current work in the field rather than reporting the results of a single study.

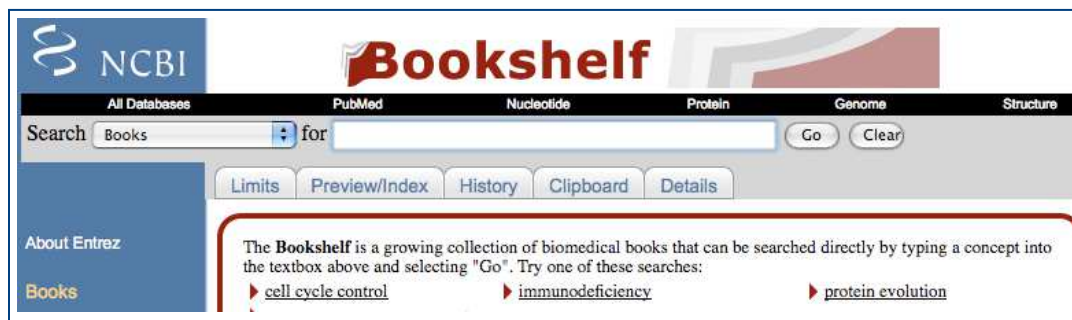


Once you have more focused search results, you can begin to assess the relevance of each article: simply clicking on each list of authors lets you see the Abstract of the paper.

Helpful hint: For each relevant paper, you can check the little box to the left of the reference. When you have found a set of papers, click Send to: Clipboard on the top menu. This will transfer these references to the Clipboard. When you have finished, click Clipboard, on the gray menu bar (see Figure above). This will display only your selections and allow you to collect them by printing or saving to your computer.

Bookshelf

You can also conduct similar searches in textbooks by clicking on the “Books” link at the far right in the dark blue bar near the top of the PubMed page. Try this to be directed to this page:



Note that you can search by topic to be directed to both text and figures from many textbooks. Again, refining your search terms may be helpful here.

OMIM

Another way to obtain reference information about a given topic's connection to human disorders is through the OMIM database, which integrates the known Mendelian genetics of human disease with the resources made available through Entrez at NCBI.



The screenshot displays the OMIM (Online Mendelian Inheritance in Man) database interface. At the top, the NCBI logo is visible on the left, and the OMIM logo with the text "Online Mendelian Inheritance in Man" and "Johns Hopkins University" is on the right. Below the logo, there is a navigation bar with tabs for "All Databases", "PubMed", "Nucleotide", "Protein", "Genome", "Structure", "PMC", and "OMIM". The search bar contains the text "OMIM" and "for", with "Go" and "Clear" buttons. Below the search bar, there are tabs for "Limits", "Preview/Index", "History", "Clipboard", and "Details". A list of instructions for using the database is displayed, including: "Enter one or more search terms.", "Use **Limits** to restrict your search by search field, chromosome, and other criteria.", "Use **Index** to browse terms found in OMIM records.", and "Use **History** to retrieve records from previous searches, or to combine searches." The page also features a sidebar with "Entrez" links and a "Help" section.

Typing a gene or disease name into the search window and clicking "Go" will reveal a list of matches with accession numbers. Clicking on one of these numbers takes you to the OMIM entry. This record is a description of the discovery of the associated gene. It includes the literature references and information about linkage and location. This is all within a brief description written and updated by a curator. The research literature is summarized under various headings such as Gene Function, Gene Structure, Inheritance, Biochemical Features, Clinical Features, etc. Clicking on the various headings on the left will take you to the subsections of the file.

After exploring these databases, please complete Bioinformatics Assignment 1.

BIO 152

BIOINFORMATICS ASSIGNMENT ONE

NAME _____

I pledge to abide by the ABC. _____

This assignment is due in Recitation at your regularly-scheduled time, either Monday September 20 or Tuesday September 21. No late assignments will be accepted.

Choose a molecular biology or genetics topic that interests you, such as a specific disease, type of cancer, or biological process. Search for articles using OMIM and PubMed. Click on the REVIEW tab to explore reviews of your topic. Use the LIMIT function to cut down on the number of positive hits. In the space below, list the specific topic and the term or parameter you used to limit the number of positive hits.

PASTE the citation for the ONE article you found via PubMed that you would READ FIRST.

Why would you read this article first as opposed to the others?

Use PubMed to search for published papers by two different Muhlenberg College biology faculty members. You will need to know their last name, initials, and have some idea of what they work on. Note that there are probably many individuals in science with a given surname. Searching for surname and Muhlenberg or the organism, gene or problem they investigate may help. Paste the TWO CITATIONS you found HERE.

Go to the Trexler Library Home Page. Under the Catalog tab, switch to “Journal Title begins:” and type “Nature Genetics” in the window. Click on the Full Text link for Academic Search Premier ([Full text available from Academic Search Premier: 06/01/1998 to 1 year ago](#)). In the second window at EBSCOhost type in the name of a genetic disease or gene that interests you. Hit search. Examine your results. Note that some of them have links to PDF full text. Click on a link to retrieve the journal article. Paste a copy of ONE CITATION you found HERE.

Go to the Trexler Library Home Page. Choose the tab that lists “Study Guides.” Choose Biology, then find and click on the link for [ScienceDirect College Edition Health and Life Sciences](#). Repeat the process using a key author or authors from your previous searches. Paste a copy of ONE CITATION you found HERE.

BIO 152 Principles of Biology III: Molecules & Cells

Working with nucleotide and protein sequences

Note: This material is adapted from *Web-based Bioinformatics Tutorials: Exploring Genomes* by Paul Young and from the Bioinformatics Online Laboratory found at <http://www.muhsenberg.edu/depts/biology/courses/bio152/BioinformaticsLab/index.html>

The goal of this material and the accompanying bioinformatics assignment is to provide you with practice in accessing and using nucleotide and protein sequence information in databases housed at NCBI. Accessing information already provided by other scientists can save months of work. After isolating a sequence, scientists can instantly compare it to the database of known sequences to gain clues about the identity or function of their gene or protein. Using this information, scientists can determine the best way to proceed with their research. The Entrez retrieval system is the starting point for searching most material at NCBI and, as you are discovering, provides links between related types of information.

Typically, we will focus on hemoglobin, which serves as a model of protein structure and function (and is a protein we have touched on already in the context of sickle cell anemia in the Genetics Unit). As you begin your searches, it is important to note that there are many members of the globin gene family; for example, functional adult human hemoglobin requires combining two β -globin proteins with two α -globin proteins.



Searching the databases at NCBI



Go to the main NCBI site: <http://www.ncbi.nlm.nih.gov/> and recall that the default allows you to search all databases. You can also restrict your search to particular databases – two that we will look in the context of this handout and accompanying assignment are ‘Protein’ and ‘Nucleotide’.


The screenshot shows the NCBI homepage. At the top, it says "National Center for Biotechnology Information" with links to "National Library of Medicine" and "National Institutes of Health". Below this is a navigation bar with tabs for "PubMed", "All Databases", "BLAST", "OMIM", "Books", "TaxBrowser", and "Structure". A search bar is present with a dropdown menu set to "All Databases" and a "Go" button. On the left, there is a vertical navigation menu with categories like "SITE Map", "Alphabet Resource", "About NCBI", "GenBank", "Literature databases", "Molecular databases", and "Genomics". The main content area features a "What's New" section with a "Hot Spots" sidebar containing links to "Clusters of orthologous groups", "Coffee Break, Genes & Disease, NCBI Handbook", "Electronic PCR", "Entrez Home", "Entrez Tools", "Gene expression omnibus (GEO)", "Human genome resources", "Influenza Virus Resource", "Map Viewer", "dbMHC", and "Mouse genome resources". There is also a "KNOW What to Do About the Flu" button.

You can search 'All Databases', essentially going through Entrez, the integrated, text-based search and retrieval system used at NCBI for the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, and others. If you search in Entrez you will get a result screen like the one below. You could then click on any specific database to see the results under that heading.

- Result counts displayed in gray indicate one or more terms not found

102523	 PubMed: biomedical literature citations and abstracts	329	 Books: online books
22982	 PubMed Central: free, full text journal articles	144	 OMIM: online Mendelian Inheritance in Man
6	 Site Search: NCBI web and FTP sites	3	 OMIA: online Mendelian Inheritance in Animals

6992	 Nucleotide: Core subset of nucleotide sequence records	none	 dbGaP: genotype and phenotype
3158	 EST: Expressed Sequence Tag records	293	 UniGene: gene-oriented clusters of transcript sequences
3	 GSS: Genome Survey Sequence records	11	 CDD: conserved protein domain database
15398	 Protein: sequence database	723	 3D Domains: domains from Entrez Structure
323	 Genome: whole genome sequences	559	 UniSTS: markers and mapping data
216	 Structure: three-dimensional macromolecular structures	77	 PopSet: population study data sets
none	 Taxonomy: organisms in GenBank	2804	 GEO Profiles: expression and molecular abundance profiles
1	 SNP: single nucleotide polymorphism	24	 GEO DataSets: experimental sets of GEO data
2188	 Gene: gene-centered information	2	 Cancer Chromosomes: cytogenetic databases
none	 SRA: Short Read Archive	5	 PubChem BioAssay: bioactivity screens of chemical substances
3	 BioSystems: Pathways and systems of interacting molecules	none	 PubChem Compound: unique small molecule chemical structures
14	 HomoloGene: eukaryotic homology groups	329	 PubChem Substance: deposited chemical substance records
none	 GENSAT: gene expression atlas of mouse central nervous system	180	 Protein Clusters: a collection of related protein sequences
609	 Probe: sequence-specific reagents	none	 Peptidome: MS/MS proteomic experiments
none	 Genome Project: genome project information		

none	 Journals: detailed information about the journals indexed in PubMed and other Entrez databases	23	 MeSH: detailed information about NLM's controlled vocabulary
688	 NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections		

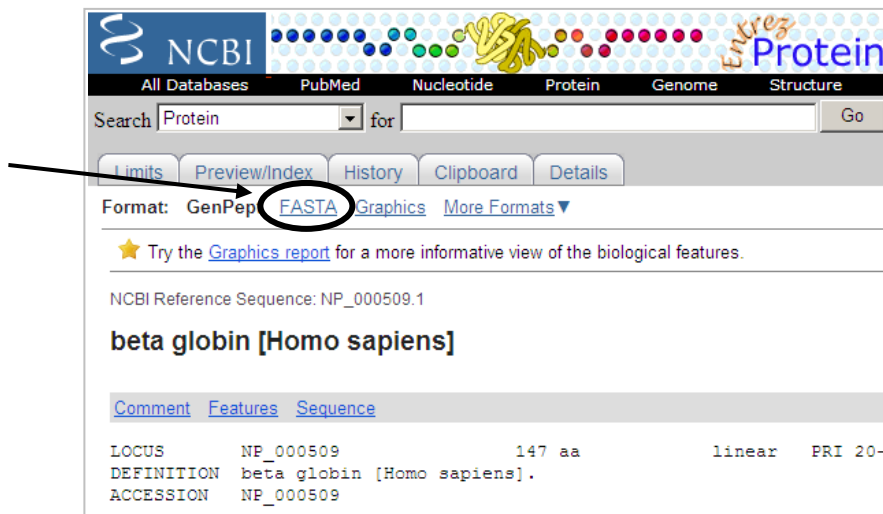
To start exploring, type "human beta globin" into the main NCBI search window and restrict this search to the PROTEIN database. You should obtain a long list of entries: closer examination will reveal entries for forms of globin arising from different alleles as well as proteins related to human beta globin. Each has an identifier number called the accession number¹ that was attached to the sequence file when it was first archived in the database. Clicking on the accession number brings you to one standard display (GenPept). Also try searching the nucleotide database. Here you will encounter slightly different information as you scroll down to the end of the display (GenBank output) for a given accession. Both of these flatfile formats are rich sources of information. Some items of note:

- the **ACCESSION** number and **DEFINITION** at the top of the record

¹ Accession numbers have different formats depending on the type of entry. For example, prefixes AAA-AAZ designate protein IDs in the GenBank database. The formats "NP_(6 digits)" and "NM_(6 digits)" refer to protein or nucleotide sequence, respectively, from the RefSeq database, which we learn more about under the BLAST section of this handout.

- the SOURCE tells us the organism from which the sequence data has been taken
- the REFERENCE section, with literature related to sequencing and characterization
- the FEATURES list, which may include conserved domains, regulatory or binding sites, etc. Features are ordered from the 5' end of a nucleotide sequence or the amino terminus of a protein sequence. The 'CDS' link may be especially helpful in defining the regions of an entry that code for functional product.

The sequence itself may be found at the bottom of the entry. You may also access sequence information by clicking on the FASTA (pronounced "FAST-A") link near the top of each entry (see screen shot below). This format has a first line of information beginning with the > sign, which can be followed by identifying information such as the accession number and species. Starting on the next line is the sequence. Note that there are no line numbers or punctuation: it is perfect for cutting and pasting.



```

NCBI Reference Sequence: NP_000509.1

beta globin [Homo sapiens]

Comment Features Sequence

LOCUS      NP_000509                147 aa                linear  PRI 20-
DEFINITION beta globin [Homo sapiens].
ACCESSION  NP_000509

>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPEEKSAVIALWGKVNVDVGGGEALGRLLVVYFWTQRFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLNLRKGTFTLSELHCDKLVHPDENFRLLGNVLVLCVLAHHEGKEFTFPVQAAYQKVVAGVAN
ALAHKYH

```

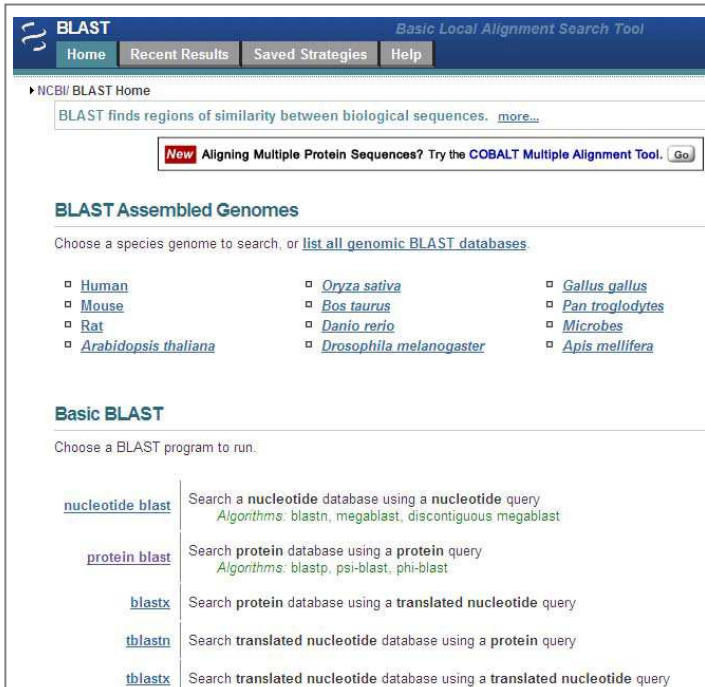
FASTA format

Many times a researcher may wish to compare one sequence to another. One way to do this is described below.

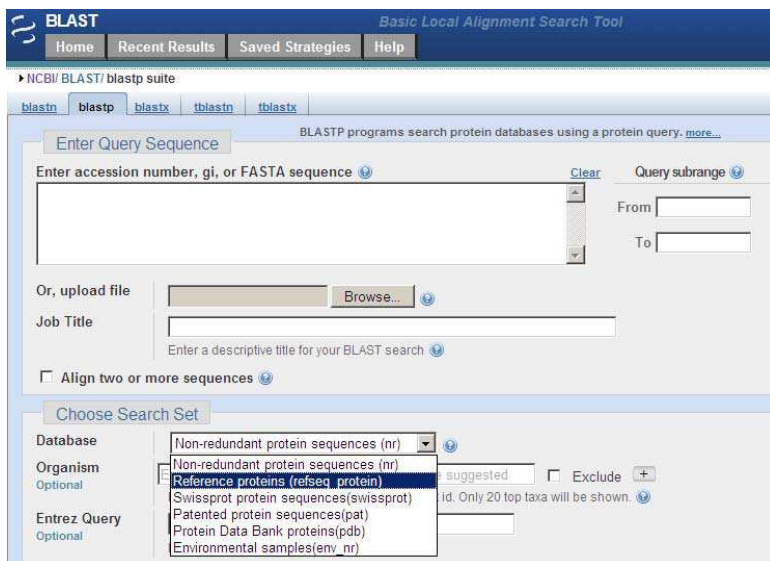
BLAST

BLAST (Basic Local Alignment Search Tool) is a powerful nucleic acid or protein alignment algorithm. It allows us to dynamically search the sequence databases to find similar sequences in different organisms. It is extremely versatile and comes in many different forms for doing different types of searches; however, the underlying method is the same in each case.

To get to the BLAST homepage from NCBI, click on 'BLAST' at the top menu. This page (screen shot below) is the starting point for several BLAST programs. You will note that you can search within a species or select a specific type of BLAST search .



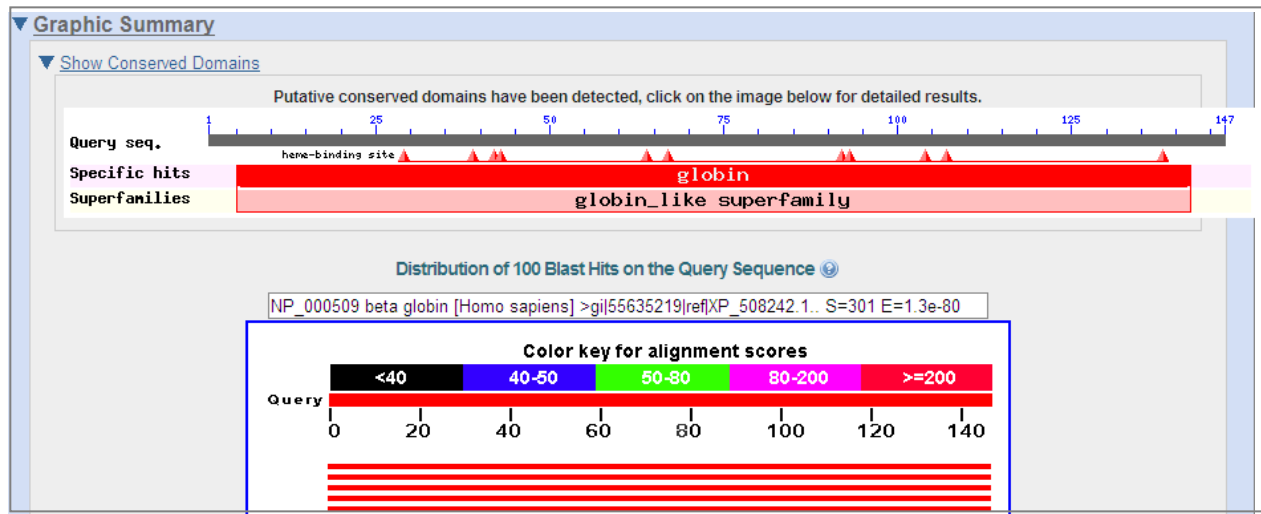
After selecting protein blast, you should reach a screen like the one shown below. Here you can copy and paste an amino acid sequence into the Query Sequence box. Alternatively, you could do a search by typing the accession number into the window; BLAST will recognize it and retrieve the sequence from the database. There is a drop-down menu that allows you to choose which part of the database you want to search; for example, 'RefSeq' is curated to be a non-redundant dataset (including one copy of each gene or protein and excluding multiple copies of each record). You can also limit your search to specific species on this screen.



Pushing the BLAST! button at the bottom of the page sends the file to the NCBI databases for the search. The page that appears next tells you that the sequence has been submitted and gives you a Search ID number. The next screen should display the result of the Conserved Domains search. Conserved domains are the functional modules of proteins. They might include a pattern of amino acids typical of a

particular catalytic site, or perhaps the binding site for a regulator of a protein. You can learn more about the identified domains by clicking on the colored bars.

Once the search is complete, you will see the reference for the algorithm (Altschul, et al.) at the top of the page. Scroll down. The first major section is a graphical display of the strongest matches to the query sequence. They are color-coded according to the alignment score. If you roll your mouse over the various lines, the identification information, alignment score (S), and E (expect) value appear in the text box above the graphical display of alignments. An alignment score (S) indicates how strong the match was (higher is better). A statistical measure of the significance of the match is given as (E); the E value is the expectation that the match would have been found in the database by chance alone (lower is better).



The second section of the results is a detailed list of hits ordered by their alignment scores. They correspond to the ones displayed graphically. Note that each line gives the identification information for the protein followed by the alignment score and the E value. Clicking on any of the scores takes you to the alignment itself, and example of which is shown on the next page.

Sequences producing significant alignments:		Score (Bits)	E Value
gb AAX37051.1 	hemoglobin beta [synthetic construct]	<u>301</u>	9e-81
gb AAX29557.1 	hemoglobin beta [synthetic construct]	<u>301</u>	1e-80
ref NP_000509.1 	beta globin [Homo sapiens] >gi 55635219 ref ...	<u>301</u>	1e-80
sp P02024 HBB GORGO	Hemoglobin subunit beta (Hemoglobin beta ...	<u>300</u>	3e-80
gb AAN84548.1 	beta globin chain variant [Homo sapiens]	<u>299</u>	3e-80
gb AAZ39780.1 	beta globin [Homo sapiens] >gb AAZ39781.1 bet...	<u>299</u>	3e-80
gb AAD19696.1 	hemoglobin beta chain [Homo sapiens]	<u>299</u>	4e-80

Further down the list, the output gives the actual alignments for the various proteins (hits) in the list. In the example below, note that the sequences are lined up ("aligned"), one above the other, so that each amino acid residue (represented by the one letter code) of one sequence can be compared to the corresponding residue of the other sequence. The 'query' is the sequence you performed BLAST with, while the 'sbjct' (subject) represent the sequence of this specific hit. The middle line compares the two sequences: empty spaces indicate mismatches and a + sign indicates similarity between the two different

amino acids compared. Sometimes one sequence must be “cut” and a gap introduced (denoted by -), in order to make this sequence align in the optimal way with the other sequence. More information about the alignment is displayed beneath the score and expect values: identities reflect the number of identical amino acids at given positions, while ‘positives’ will typically be higher as amino acids with similar chemistry will also be scored as a match here.

```
>[ref|NP_001028265.1] UG alpha globin-like [Danio rerio]
Length=143

GENE ID: 497162 si:xx-by187q17.1 | si:xx-by187g17.1 [Danio rerio]

Score = 117 bits (293), Expect = 3e-25, Method: Compositional matrix adjust.
Identities = 60/148 (40%), Positives = 91/148 (61%), Gaps = 7/148 (4%)

Query 2  VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRRFFESFGDLSTPDAVMGNP  59
      + L+ ++K+ V A+W K++  DE+G EAL R+L VYP T+ +F  + DLS      G+
Sbjct 1  MSLSDKDKAVVKAIWAKISPKADEIGAEALARMLTVYPQTKTYFSHWSDLSP-----GSG  55

Query 60  KVKAHGKKVLGAFSDGLAHLNLDNLKGTTFATLSELHCDKLHVDPENFRLLGNVLVCLAHHF  119
      VK HGK ++GA  + ++ +D+L G  A LSELH  KL VDP NF++L + ++ V+A  F
Sbjct 56  PVKKHGKTIMGAVGEAISKIDDLVGGGLAALSELHAFKLRVDPANFKILSHNVIVVIAMLF  115

Query 120 GKEFTFPVQAAAYQKVVAGVANALAHKYH  147
      +FTP V  +  K      +A AL+ KY
Sbjct 116 PADFTPEVHVSVDKFFNNLALALSEKYR  143
```

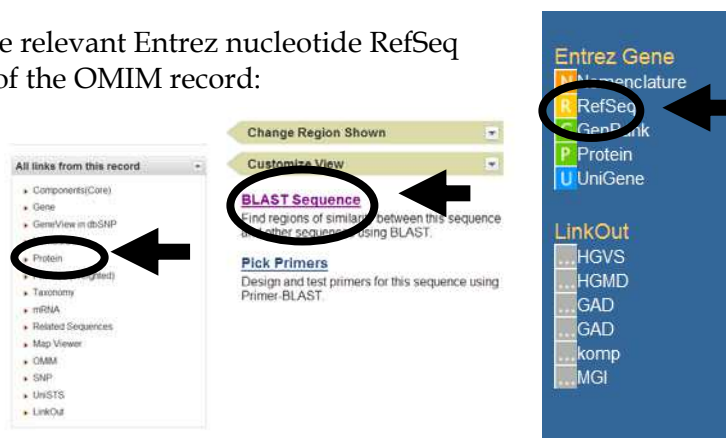
Note that only portions of the query sequence as well as portions of other accessions in the database are aligned by the BLAST algorithm, which is a local alignment tool, displaying the most strongly matching regions of the overall comparison. BLAST is able to return matches relatively quickly compared to other programs (e.g., Clustal) that you may encounter, which use algorithms that perform global alignments, comparing sequences in their entirety.

After exploring these databases, please complete **Bioinformatics Assignment 2**. In this assignment you will need to move between OMIM, Entrez, BLASTN, back to sequence records, and then to BLASTP.

To move directly from an OMIM record to the relevant Entrez nucleotide RefSeq sequence file, use the link on the LEFT SIDE of the OMIM record:

To move from a sequence file to BLASTN or BLASTP automatically, choose the “BLAST Sequence” link on the RIGHT SIDE of a sequence record:

To move from a nucleotide sequence file use the “Protein” link on the Lower sequence record:



BIO 152

BIOINFORMATICS ASSIGNMENT TWO

NAME _____

I pledge to abide by the ABC. _____

This assignment is due in Recitation at your regularly-scheduled time, either Monday October 4 or Tuesday October 5. **No late assignments will be accepted.**

*Use ENTREZ to retrieve a PROTEIN sequence for the rat (*Rattus norvegicus*) Huntington Disease protein (called Huntingtin). You will probably have to think about what search terms to include in order to ensure that you do not get too many hits. Paste the ACCESSION NUMBER and the first 5-10 amino acids of the record you found HERE:*

Use ENTREZ to retrieve a NUCLEOTIDE sequence for the rat Huntington Disease gene. Be sure to search the mRNA RefSeq database (recall that the record will be in DNA form however). Paste the ACCESSION NUMBER here:

By closing examining the annotation of the nucleotide record you retrieved, identify the START CODON and STOP CODON in the sequence. Paste the sequence of the start codon plus the following 5-10 bases and the stop codon plus the following 5-10 bases here:

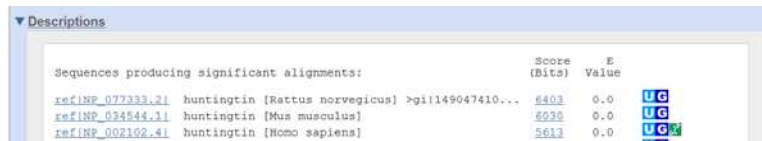
Use OMIM to retrieve the record for a well-characterized human genetic disease (OTHER THAN Huntington Disease). In the space below, insert the name and synonyms for the disease, whether it is dominant/sd/recessive, and its chromosomal location. If you cannot find that information, choose a different disease.

Use the RefSeq link at the bottom left of the OMIM record for your disease to identify the corresponding nucleotide sequence record. There may be more than one, but you will need to select one of the records for further analysis. Click on the nucleotide sequence record. Record the accession number here:

Choosing "BLAST sequence" using the link at RIGHT will take you directly to BLASTN, allowing you to perform a search for similar nucleotide sequences. The accession

number will automatically appear in the window, but note that you could also simply copy the entire sequence and paste it into the window in "FASTA" format. Under "DATABASE" on the BLASTN page, change the window to read "Reference mRNA sequences". Run BLASTN.

Examine your hits. Paste in the descriptions (as shown in the example at right) of the top three hits here:



The screenshot shows a window titled "Descriptions" with a table of sequences producing significant alignments. The table has three columns: the sequence identifier and description, the Score (Bits), and the E Value. To the right of the E Value are three blue links labeled "UG".

Sequences producing significant alignments:	Score (Bits)	E Value	
ref NP_073333.2 huntingtin [Rattus norvegicus] >gi 149047410...	4403	0.0	UG
ref NP_034544.1 huntingtin [Mus musculus]	4028	0.0	UG
ref NP_002102.4 huntingtin [Homo sapiens]	5613	0.0	UG

Examine your E values. How significant are the matches between your query and each database record?

Return to your nucleotide sequence record. Use the "Protein" link at the right side of the sequence record to jump over to the protein sequence record. Examine the protein sequence and record the accession number here:

Once again, choose "BLAST sequence" and note that you are now automatically transported to BLASTP. Why do you get to BLASTP instead of BLASTN this time?

Run BLASTP and examine your hits. Paste in the description of the top three hits again.

Why are the BLASTP results not in exactly the same order as the BLASTN results, even though you were using the same gene for both searches?

BIO 152 Principles of Biology III: Molecules & Cells

DNA microarrays

As mentioned in lecture, DNA microarrays can be used to measure the “transcriptome”; more specifically, the two-color labeling method discussed here can provide information about the relative expression of genes between two experimental conditions. Unlike other existing technologies used to detect and quantitate mRNA, microarrays are one example of a newer approach that allows investigators to obtain genome-wide data in a single experiment. With this large amount of data comes questions of how best to best organize, interpret, and present these results, some of which we will consider here.

The goal of this material and the accompanying bioinformatics assignment is to increase your understanding of the principles underlying microarray technology (including reverse transcription and nucleic acid hybridization) and to provide you with practice in interpreting data for DNA microarray experiments. Topically, part II of the assignment will focus on analysis of yeast strains grown under two different metabolic conditions (which can connect to information covered in the upcoming Biochemistry Unit), with data taken from a now classic paper by DeRisi et al.¹. Background information will be provided through the use of an animation developed by GCAT (Genome Consortium for Active Teaching), which can be found at <http://gcat.davidson.edu/Pirelli/index.htm> (click “Enter MediaBook” at the bottom of the homepage). Guidelines to help you navigate through the site are found below.

How microarrays work

Please review the entire Method section of the animation (19 “bubbles”). Note that by convention the “control” condition is labeled green and the experimental is labeled red, so the tutorial discusses green signals as representing “repression” (expressed in control, but not expressed in experimental) and red signals as indicating “induction” (expressed at higher levels in experimental sample).

How to analyze and present microarray data

Please review the beginning of the Data Interpretation section (“bubbles” 1- 11). Make sure you understand the meaning of different colored “spots” (red, green, yellow, and black) and can give examples of expression ratios or fold change that would correspond to each color. While you can qualify fold change by referring to fold induction or fold repression, you should also consider the advantages to log transforming expression ratios as described in the animation.

After exploring these databases and following our discussion in recitation, please complete Bioinformatics Assignment 3.

¹ DeRisi, et al. 1997. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science* 278: 680-686.

BIO 152 Principles of Biology III: Molecules & Cells Bioinformatics Assignment 3

This assignment is due the week of November 1 at the beginning of your regularly scheduled recitation time. **Late assignments will not be accepted.** Please confine your answers to a maximum of two pages. While you may explore the databases together, submitted work should include a signed ABC pledge indicating that you answered the following questions individually. In contrast to the previous assignments, please compose your responses to the specific numbered questions and hand-in double-spaced, stapled, etc. with ABC pledge signed at the top.

Part I: Learning to analyze microarray data

Imagine that you have conducted a microarray experiment comparing gene expression profiles in both HER-2 positive breast cancer cells and non-cancerous breast tissue in which cDNA from the cancerous cells is labeled red and control cDNA is labeled green. A section of a representation array and a data table with corresponding signal intensities are shown below.



Spot # (begin top left, read left to right, top to bottom)	Green signal intensity	Red signal intensity
1	6400	4900
2	3400	3500
3	1500	15500
4	10400	11000
5	14000	18400
6	500	100
7	10800	13000
8	9100	1400

1. Theoretically, what range of expression ratios could indicate that a gene was not expressed in cancerous tissue, but was expressed in healthy tissue¹? Similarly, what range of ratios could indicate that a gene was not expressed in healthy tissue, but was expressed in cancerous tissue? What range of expression ratios could indicate that a

¹ Recall the standard way of creating an expression ratio as described in the Pirelli animation (<http://gcat.davidson.edu/Pirelli/index.htm>)

gene was **equally** expressed in both cancerous and healthy tissue? What range of ratios could indicate that a gene was **not** expressed in either cancerous or healthy tissue?

2. Explain how a single gene expression ratio (*e.g.*, 4) can correspond to many different *absolute* levels of gene expression in the cancerous and healthy tissue samples.

3. Provide your interpretation of relative levels of gene expression for each of the 8 numbered features (genes) shown here (*i.e.*, approximately what fold induction or repression is observed?).

4. Typically expression ratios are log transformed. Explain why this is done and illustrate this using an example from this sample array.

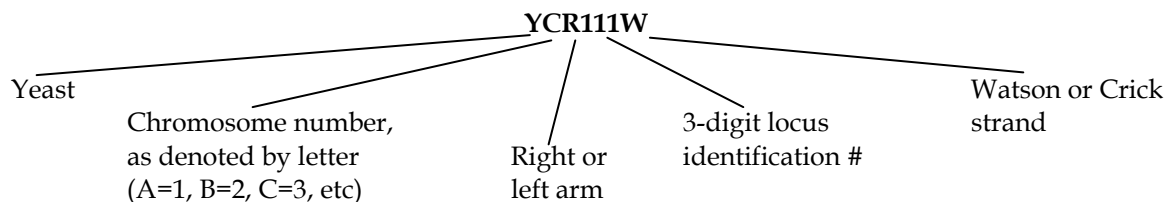
Part II: Dealing with larger data sets

Examine the Excel file posted on Blackboard with data taken from an experiment by DeRisi et al.² examining yeast grown under different nutrient conditions. As stated in the introduction to their paper “Inoculation of yeast into a medium rich in [glucose] is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source...”. Thus, the goal of this experiment was to look at potentially widespread changes in expression of genes involved in fundamental metabolic processes. For the particular experiment we are focusing on, cDNA produced from yeast cells harvested soon after inoculation into glucose-rich media was labeled green, while cDNA obtained from cells harvested well after the switch to ethanol as a carbon source was labeled red. Signal intensities from scans of one section of the microarray at both these wavelengths are given in the Excel file (includes 1600 features, mostly yeast genes³, with some control spots included as well).

5. Create standard and log₂ transformed expression ratios for each feature in the Excel spreadsheet. Sort these ratios so that you can easily identify the three genes with highest fold induction and the three genes with the highest fold repression. Provide the names of these 6 genes and the log transformed expression ratios as a part of your

² DeRisi, J.L., V.R. Iyer, and P.O. Brown. 1997. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science* 278: 680-686.

³ Yeast genes are named according to the following nomenclature:



answer to this question. Generally, how do the levels of induction and repression observed here compared to your theoretical predictions given in answer to question #1?

6. How many genes (percentage) show more than 4-fold induction in the later growth stages where ethanol is used as a carbon source? How many genes are repressed 4-fold or more? How many genes show less than a 2-fold change in expression between the two conditions? Are these percentages surprising to you? Why or why not?

7. Identify the expression ratio for the gene denoted as YMR170C (to find this feature in your spreadsheet, it may be easiest to re-sort your data by grid and then column number, as then the yeast genes appear "alphabetically" by their identifiers"). Go to *Saccharomyces* Genome Database (www.yeastgenome.org) and type in this identifier in the top search box to learn more about this gene. What is its protein product? Can you offer an explanation as to why its expression is affected the way it is under these growth conditions?

BIO 152 Principles of Biology III: Molecules & Cells

Examining protein structures

Note: This material is adapted from *Web-based Bioinformatics Tutorials: Exploring Genomes* by Paul Young and from Dr. Edwards.

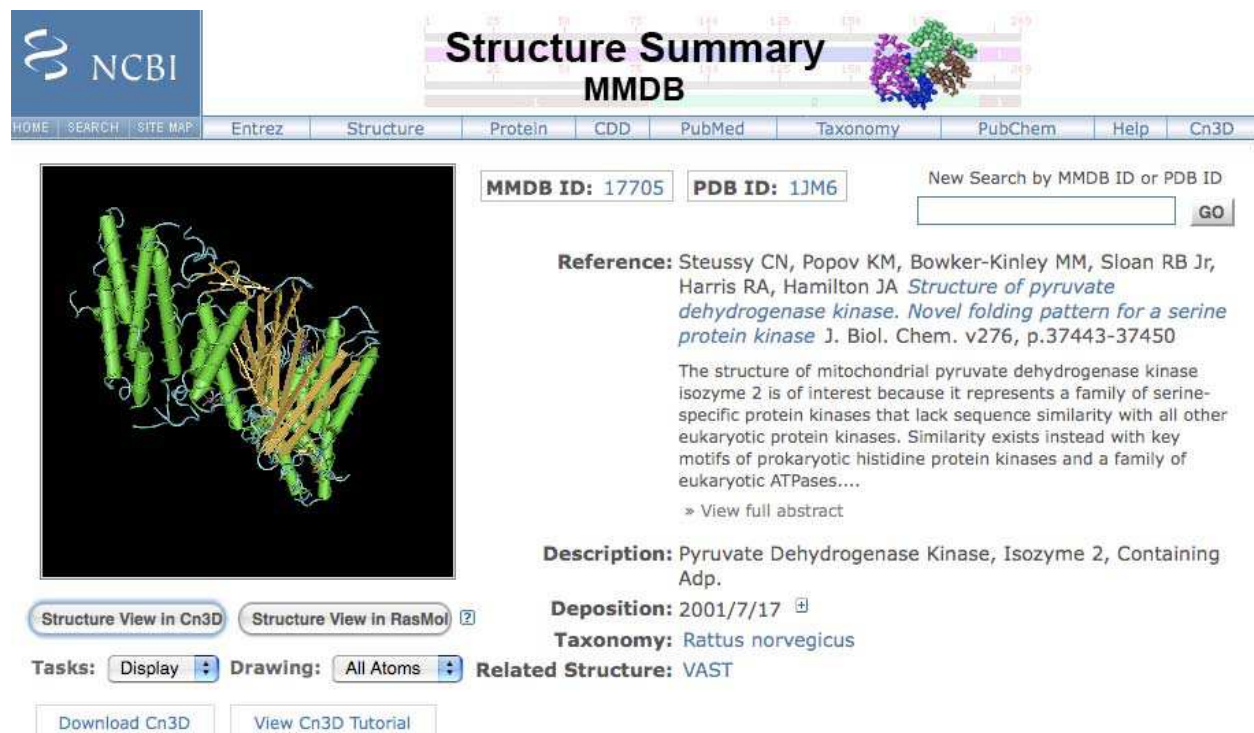
Tools such as BLAST make it relatively easy to search the databases and align primary sequences, which allows us to ask questions about similar proteins and gene families. Structural databases also exist which contain details of three-dimensional protein structure based X-ray crystallography and NMR studies; one example of a structure database that we will use is also housed at NCBI:

<http://www.ncbi.nlm.nih.gov/sites/entrez?db=Structure>

A viewer called Cn3D is required which is available for download here:

<http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml>

Let's look at the structure for a particular regulatory enzyme, pyruvate dehydrogenase kinase 2 (PDHK2 is a mitochondrial protein kinase that phosphorylates the pyruvate dehydrogenase complex, thereby down-regulating the oxidation of pyruvate – don't worry, this should make more sense after our lectures on respiration). The identification code is 1JM6; you can enter this or use "rat pyruvate dehydrogenase kinase" as a search term on the NCBI Structure page to get to the structure. A small image along with the identifier and other information will appear as the search result. Click on the image, which should bring you to a Structure Summary page (shown below).



NCBI

Structure Summary

MMDB

HOME | SEARCH | SITE MAP | Entrez | Structure | Protein | CDD | PubMed | Taxonomy | PubChem | Help | Cn3D

MMDB ID: 17705 PDB ID: 1JM6

New Search by MMDB ID or PDB ID GO

Reference: Steussy CN, Popov KM, Bowker-Kinley MM, Sloan RB Jr, Harris RA, Hamilton JA *Structure of pyruvate dehydrogenase kinase. Novel folding pattern for a serine protein kinase* J. Biol. Chem. v276, p.37443-37450

The structure of mitochondrial pyruvate dehydrogenase kinase isozyme 2 is of interest because it represents a family of serine-specific protein kinases that lack sequence similarity with all other eukaryotic protein kinases. Similarity exists instead with key motifs of prokaryotic histidine protein kinases and a family of eukaryotic ATPases....

» View full abstract

Description: Pyruvate Dehydrogenase Kinase, Isozyme 2, Containing Adp.

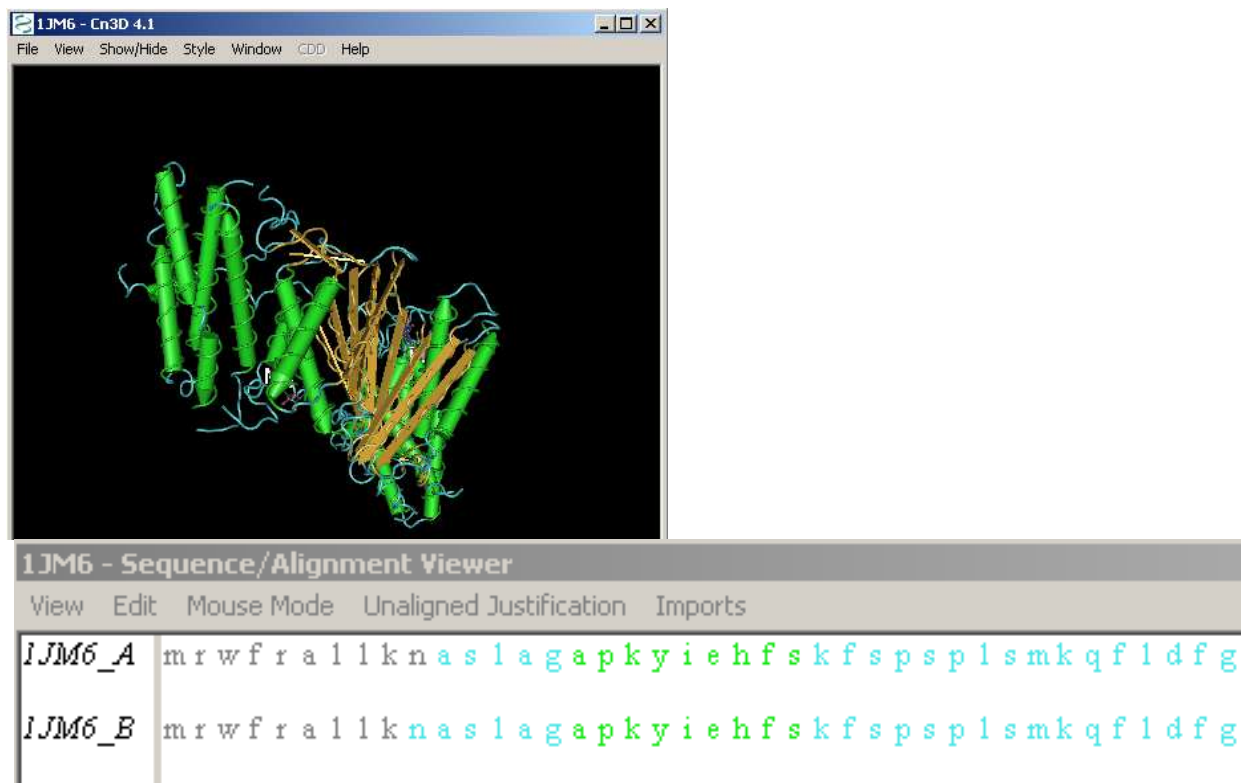
Deposition: 2001/7/17

Taxonomy: Rattus norvegicus

Tasks: Display Drawing: All Atoms Related Structure: VAST

Download Cn3D View Cn3D Tutorial

Note the reference information available on the right side on the screen. Double click on the structural image at the left and use Cn3D to open up an interactive display (shown below).



The upper pop-up window shows a 3D representation of the protein that can be rotated by holding down your computer mouse button and dragging. You can zoom in or out under the "View" menu. This representation displays α helices as green cylinders, β sheets in gold, and loops in blue. From the menu at the top, you can open the drop down menu under "Style", select "Rendering Shortcuts", and change the type of representation of the amino acid chain. You can also select "Toggle Sidechains" at the bottom of that menu to see both secondary structure and the R-groups.


The lower pop-up window shows the primary sequence of the two amino acid chains or subunits that make up this protein. The amino acid residues are indicated by their one letter abbreviation, which is color-coded to match the representations in the 3D viewer. If you click and highlight amino acid residues in the linear sequence, the amino acids will be highlighted in the structural representation as well (if you select an individual amino acid, the residue number will be shown at the bottom left of the lower pop-up window as well).

To copy and paste a structure, you can use the method outlined at the top of the next page (for a PC; for a Mac, try using the Grab program):

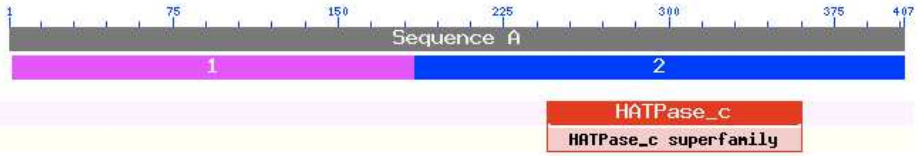
1. When you have the image you want on your screen, press **Alt+PrintScrn** at the same time.
2. Open the Paint program.
3. Use the “paste” command to place your screen shot into Paint.
4. Select regions you want to cut and paste.

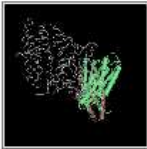
Return to the Structure Summary page and scroll down (shown below). In addition to displaying a solved structure of a protein of interest, NCBI Structure identifies conserved domains. If you refer to the “Domain Families” line on the Structure Summary page, you can see that this protein has an ATPase that is conserved in other proteins. You can click on this link to find out more about this domain or locate other proteins that share this structural and functional feature. Just below this information you can locate the ligands that are associated with the protein structure (in this case, ADP and magnesium ions; did you see these in the 3D structure?)

Molecular components in the MMDB structure are listed below and may include macromolecular chains, 3D domains, protein classifications (domain families), and ligands, as available. Mouse over each icon for more information on the component. ⓘ

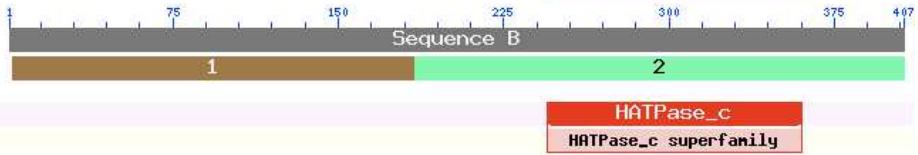


[Protein](#)
[3d Domains](#)
[Domain Families](#)
Specific hits
Superfamilies







[Protein](#)
[3d Domains](#)
[Domain Families](#)
Specific hits
Superfamilies



Ligand



ADP
2 occurrences



Mg
Magnesium
2 occurrences

After exploring these databases and following our discussion in recitation, please complete **Bioinformatics Assignment 4**.

BIO 152

BIOINFORMATICS ASSIGNMENT FOUR

NAME _____

I pledge to abide by the ABC. _____

DUE: Monday or Tuesday November 8 or 9 at the beginning of your recitation section.

1. Use the NCBI Structures database to identify the structure of an **amino-acyl tRNA synthetase**. Paste the structure and the “sequence description” text here.
2. Recover structure the structure with the PDB ID# **3IOW**. It should be the structure of the **Huntingtin N-terminus with 17-Gln**. Use the Cn3D plug-in to visualize and manipulate the structure. How many alpha helices and how many beta-sheets are there? Count each beta-sheet consisting of two or more beta strands as one beta sheet—in other words don't count the number of beta-strands. You should consider multiple parallel (or anti-parallel) beta strands as comprising a single beta sheet. Note that beta sheets do twist a bit, which becomes very evident when there are more than two beta strands in a sheet.
3. Repeat step #1 and #2 for a protein that interests you. Paste a copy of the structure and sequence description text here and count the number of alpha-helices and beta-sheets. Do you see any beta-turns?

Can't cut and paste? No problem! 1. When you have the image you want on your screen, press **Alt+PrintScrn** at the same time. 2. Go to the **START** menu and click on **Run**. 3. Type "**pbrush**" and then **OK**. (or simply double-click Paintbrush under ACCESSORIES). When the Paint program opens, then under **Edit**, select **Paste**. 4. Now you can highlight, cut and paste into other document types.

-SAVE PAPER! REDUCE YOUR IMAGES SIZES SO THEY ALL FIT ON ONE PAGE!-