Laboratory Manual

Biology 3055

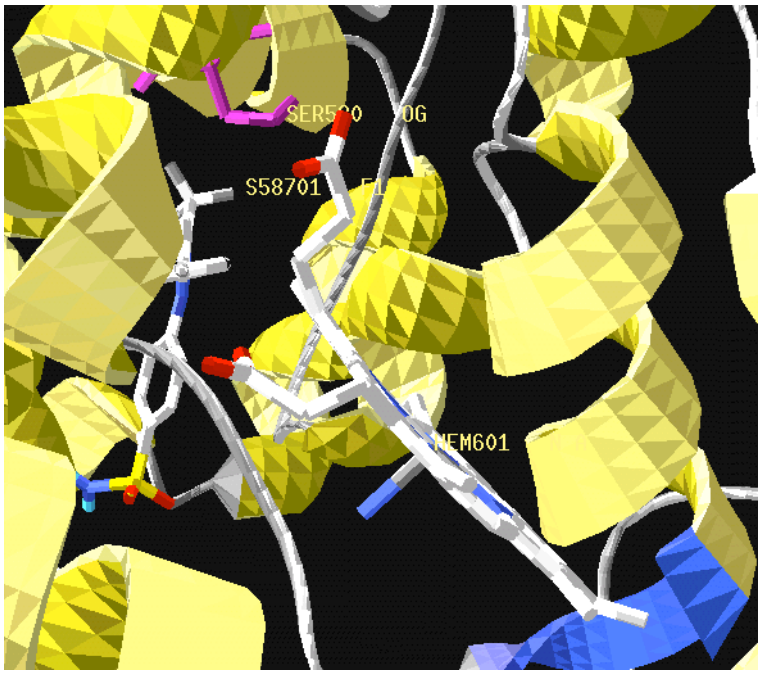


Table of Contents

Created in 2003 by:
April Bednarski
Advised by:
Professor Himadri Pakrasi

Funded by a grant from:
Howard Hughes Medical Institute to
Washington University

Biology 3055
Laboratory Syllabus

I. LAB INSTRUCTORS

Instructors: Jessica Wagenseil, Kathy Hafer, April Bednarski

II. COURSE DESCRIPTION

Welcome to the laboratory for Biology 3050. This lab is designed around bioinformatics tools that are widely used in biomedical research today. Over the past few years, the number of freely available software programs and web-based research tools has increased dramatically. Knowing how to use these tools is very important to research and health professionals in order to access and interpret the constantly growing amounts of genomic and proteomic information. These bioinformatics tools as well as genomic information are freely available on the web to anyone who knows how to access them. Biology 3055 introduces these computer-based research tools in a context that reinforces concepts presented during the Biology 3050 lecture.

III. PROJECTS

You will be given your own gene to study during the 5 lab sessions. This gene will be the focus of all your guide sheets and final report. There will be one other person working on the same gene in each lab section. You are encouraged to help each other, but work individually. All the projects are similar in their structure and format, even though they are designed around different genes. The basic process for using web-based research tools is the same for all projects.

List of Projects:

Mitochondrial Diseases

1. ATP Synthase 6
2. Cytochrome c Oxidase 1

Apoptosis

3. Caspase 1
4. Superoxide Dismutase 1

Metabolic Disease

5. Phenylalanine Hydroxylase
6. Hypoxanthine-guanine phosphoribosyltransferase

Lung Cancer

7. K-Ras
8. Cytochrome P450 1A1

Cholesterol Biosynthesis

9. HMG-CoA Reductase
10. Low Density Lipoprotein Receptor

IV. ASSIGNMENTS

Guide sheet questions:

The guide sheets are designed for your specific project. You will start working on them during the lab session, and they will be due by the next lab meeting to be graded. These sheets will help guide you through your web-based research.

Readings:

Text: Berg, J.M., Tymoczko, J.L., and Stryer, L., (2002) Biochemistry, 5th ed., W.H. Freeman.

There will be two reading assignments that are specific to your project and provide background information on the gene and disease you are studying. The

readings will be from the Berg text (above) and journal articles. The journal articles will be available on the course website.

Report:

You will be compiling printouts from each lab period into a report showing what you have found out about your gene using the web-based resources. You will also write a one-page summary of your findings. This report will be due at the last lab session. You will be presenting your report in small groups. After presenting your findings, you will turn in your report to be graded.

Report Evaluation

Content:	15 pts
Summary:	10 pts
Total	25 pts

Report format:

Final report should contain:

1. Title page with:
   - Your name and date
   - Protein name
   - Lab section number
2. Summary (1 page, single-spaced)
3. Your multiple sequence alignment
4. Three Swiss-Pdb Viewer figures

V. EVALUATION

Grades will be based on the guide sheets, reading questions, the report, and class participation. You are required to attend every lab session. There is a class participation grade based on attendance, concentrating on the assigned lab during lab meeting time, as well as helping fellow classmates during class meeting time. It is expected that the activities will take the full two hours of each lab session. Each student will receive the 10 pts unless an instructor notes lack of attendance and not correctly using lab time.

Points Breakdown:

Tutorial Sheet	5 pts
Questions Reading 1	5 pts
Questions Reading 2	5 pts
Guide Sheet 1	10 pts
Guide Sheet 2	10 pts
Guide Sheet 3	10 pts
Joint Quiz	10 pts
Structure Problem Set	10 pts
Participation	10 pts
Report	25 pts
Total	100 pts

VI. LAB CALENDAR

Week of:	Topics	Topic
Jan 17	No Lab	No Lab
Jan 23	Lab 1	Tutorial on bioinformatics tools (COX-2 or PTGS2)
Jan 30	Lab 2 Guide Sheet 1	Start individual research: Gene, ExPASy, and ClustalW Tutorial sheet due Reading questions 1 due
Feb 6	Lab 3 Guide Sheet 2	Swiss-Pdb Viewer/DeepView Structure Problem Set due Guide Sheet 1 due Reading questions 2 due
Feb 13	Exam 1 (Feb 16)	No Lab
Feb 20	Lab 4 Guide Sheet 3	OMIM and KEGG Guide Sheet 2 due
Feb 27	Lab 5 Reports and Joint Quiz	Present reports in small groups Guide Sheet 3 due Report due
Mar 6	Exam 2	No Lab
Mar 13	Spring Break – No Lab	No Lab

Laboratory 1 - Introduction
COX-2 (PTGS2) Tutorial

On the first day of lab, we will be working through a tutorial on the web-based bioinformatics programs that you will be using for your research projects. You will also be receiving your project packet, which will contain the introduction, specific directions, and reading assignments for your project. There are ten possible projects, and you will be working on a project that is the same as one of the people at the computer next to you. Although you must each complete your own research project, you are encouraged to collaborate, discuss your projects, and help each other during the lab.

The tutorial is based on the **enzyme cyclooxygenase-2 (COX-2), which also has the name prostaglandin synthase-2 (PTGS2).** You can read more about this protein on the next page. In this tutorial, the bioinformatics tools from the NCBI (National Center for Biotechnology Information) website will be introduced. NCBI is a division of the National Institute of Health (NIH). These tools include Gene, GenBank, RefSeq, and PubMed. Gene is a database of genes in which each entry contains a brief summary, the common gene symbol, information about the gene function, and links to websites, articles, and sequence information for that gene. GenBank is a historical database of gene sequences, which means it contains every sequence that was published, even if the same sequence was published more than once. Therefore, GenBank is considered a redundant database. RefSeq is a database of sequences that is edited by NCBI and is NON-redundant, meaning that it contains what NCBI determines is the strongest sequence data for each gene.

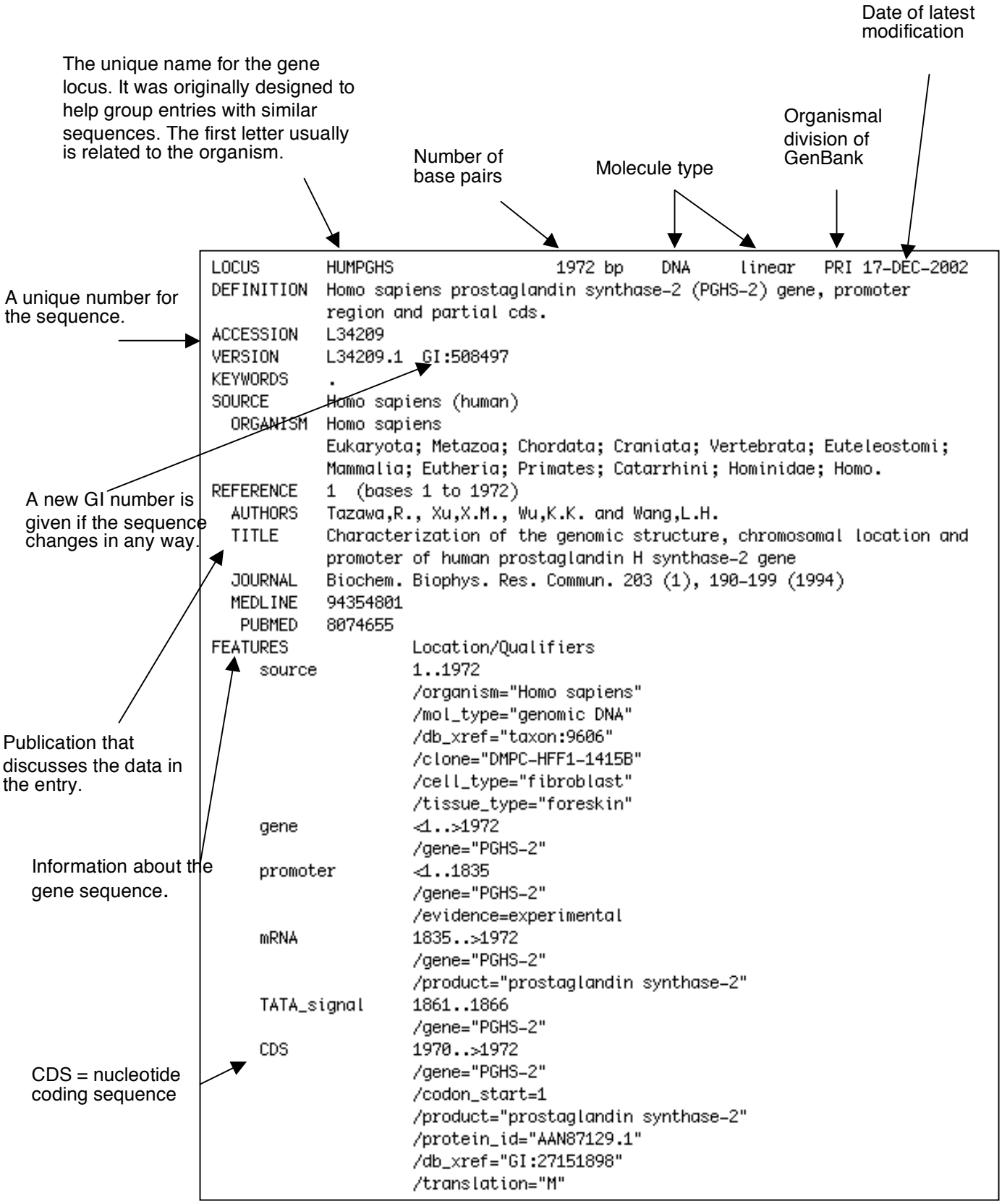
Finally, we will be learning to use ClustalW, which is a multiple sequence alignment program. It allows you to enter a series of gene or protein sequences that you believe are similar and may be evolutionarily related. These sequences are usually obtained by performing a BLAST search. ClustalW then aligns the sequences, so that the lowest number of gaps is introduced and the highest numbers of similar residues are aligned with each other. ClustalW uses a scoring matrix similar to BLOSUM-62, which is explained in your text and will be presented in lecture.

Introduction to COX-2 (PTGS2)

The enzyme we will be focusing on has two names. It is called prostaglandin H_2 synthase-2 and cyclooxygenase-2 (COX-2). COX-2 has been thoroughly studied because of its role in prostaglandin synthesis. Prostaglandins have a wide range of roles in our body from aiding in digestion to propagating pain and inflammation. Aspirin is a general inhibitor of prostaglandin synthesis and therefore, helps reduce pain. However, aspirin also inhibits the synthesis of prostaglandins that aid in digestion. Therefore, aspirin is a poor choice for pain and inflammation management for those with ulcers or other digestion problems. Recent advances in targeting specific prostaglandin-synthesizing enzymes have lead to the development of Celebrex, which is marketed as an arthritis therapy. Celebrex is a potent and specific inhibitor of COX-2. Celebrex is considered specific because it doesn't inhibit COX-1, which is involved in synthesizing prostaglandins that aid in digestion. This is a remarkable accomplishment given the great similarity between COX-1 and COX-2. This achievement has paved the way for developing new therapies that bind more specifically to their target and therefore have fewer side effects.

Understanding the enzyme structures of COX-1 and COX-2 helped researchers develop a drug that would only bind and inhibit COX-2. Many of the types of information and tools used by researchers for these types of studies are freely available on the web. In this tutorial, and throughout this lab course, you will be introduced to the databases and freely available software programs that are commonly used by professionals in research and medicine to study genes, proteins, protein structure and function, and genetic disease.

GenBank Entry



Base count (A,T,G, and C) and actual sequence follow.

SwissProt Entry

SwissProt Accession Number

Click here before printing

NiceProt View of Swiss-Prot: P35354

Printer-friendly view | Submit update | Quick BlastP search

[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	**PGH2_HUMAN**
Primary accession number	**P35354**
Secondary accession number	Q16876
Entered in Swiss-Prot in	Release 29, June 1994
Sequence was last modified in	Release 37, December 1998
Annotations were last modified in	Release 42, September 2003
Name and origin of the protein	
Protein name	**Prostaglandin G/H synthase 2 [Precursor]**
Synonyms	**EC 1.14.99.1** **Cyclooxygenase -2** **COX-2** **Prostaglandin-endoperoxide synthase 2** **Prostaglandin H2 synthase 2** **PGH synthase 2** **PGHS-2** **PHS II**
Gene name	**PTGS2** or **COX2**
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

SwissProt Entry Continued:

Guide sheet questions focus on "Comments" and "Features"

Comments

- ***FUNCTION***: MAY HAVE A ROLE AS A MAJOR MEDIATOR OF INFLAMMATION AND/OR A ROLE FOR PROSTANOID SIGNALING IN ACTIVITY-DEPENDENT PLASTICITY.
- ***CATALYTIC ACTIVITY***: Arachidonate + AH_2 + 2 O_2 = prostaglandin H2 + A + H_2O.
- ***COFACTOR***: Binds 1 heme B (iron-protoporphyrin IX) group per molecule *(By similarity)*.
- ***PATHWAY***: Prostaglandins and thromboxanes biosynthesis from arachidonate; first step.
- ***SUBUNIT***: Homodimer *(By similarity)*.
- ***SUBCELLULAR LOCATION***: Membrane-associated. Microsomal membrane.
- ***INDUCTION***: By cytokines and mitogens.
- ***DISEASE***: LIKELY TO PLAY A ROLE IN INFLAMMATORY DISEASES SUCH AS RHEUMATOID ARTHRITIS
- ***MISCELLANEOUS***: This enzyme acts both as a dioxygenase and as a peroxidase.
- ***MISCELLANEOUS***: This enzyme is the target of nonsteroidal anti-inflammatory drugs such as aspirin.
- ***SIMILARITY***: Contains 1 EGF-like domain.
- ***SIMILARITY***: Belongs to the prostaglandin G/H synthase family.

Copyright

This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See http://www.isb-sib.ch/announce/ or send an email to license@isb-sib.ch)

Cross-references

EMBL	L15326; AAA35803.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] M90100; AAA58433.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] D28235; BAA05698.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] U04636; AAA57317.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AY229989; AAO38056.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AL033533; CAB41240.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] BC013734; AAH13734.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	A46150; A46150.
HSSP	Q05769; 1CVU. [HSSP ENTRY / PDB]
Genew	HGNC:9605; PTGS2.
CleanEx	HGNC:9605; PTGS2.
GeneCards	PTGS2.
GeneLynx	PTGS2; Homo sapiens.

Features

Feature table viewer | Feature aligner

Key	From	To	Length	Description	FTId
SIGNAL	1	17	17	*POTENTIAL*.	
CHAIN	18	604	587	PROSTAGLANDIN G/H SYNTHASE 2.	
DOMAIN	18	55	38	EGF-LIKE.	
DISULFID	21	32		*BY SIMILARITY*.	
DISULFID	26	42		*BY SIMILARITY*.	
DISULFID	44	54		*BY SIMILARITY*.	
DISULFID	22	145		*BY SIMILARITY*.	
DISULFID	555	561		*BY SIMILARITY*.	
ACT_SITE	193	193		DISTAL HISTIDINE (*BY SIMILARITY*).	
ACT_SITE	371	371		CYCLOOXYGENASE (*BY SIMILARITY*).	
METAL	374	374		IRON (HEME AXIAL LIGAND) (*BY SIMILARITY*).	
SITE	516	516	1	ASPIRIN-ACETYLATED SERINE (*BY SIMILARITY*).	
CARBOHYD	53	53		N-LINKED (GLCNAC...) (*POTENTIAL*).	
CARBOHYD	130	130		N-LINKED (GLCNAC...) (*POTENTIAL*).	
CARBOHYD	396	396		N-LINKED (GLCNAC...) (*POTENTIAL*).	
CARBOHYD	580	580		N-LINKED (GLCNAC...) (*POTENTIAL*).	
VARIANT	1	1	*	M -> I (in dbSNP:20426) [NCBI/Ensembl].	VAR_011979
VARIANT	228	228	*	R -> H.	VAR_016262
VARIANT	428	428	*	P -> A.	VAR_016263
VARIANT	488	488	*	E -> G (in dbSNP:5272) [NCBI/Ensembl].	VAR_011980
VARIANT	511	511	*	V -> A (in dbSNP:5273) [NCBI/Ensembl].	VAR_011981
VARIANT	587	587	*	G -> R.	VAR_016264
CONFLICT	165	165		E -> G (IN REF. 2).	
CONFLICT	438	438		I -> T (IN REF. 1).	

Amino acid residue numbers

Active site residues, disulfide linkages, residues bound to carbohydrates, etc.

Mutations found

Sources disagree

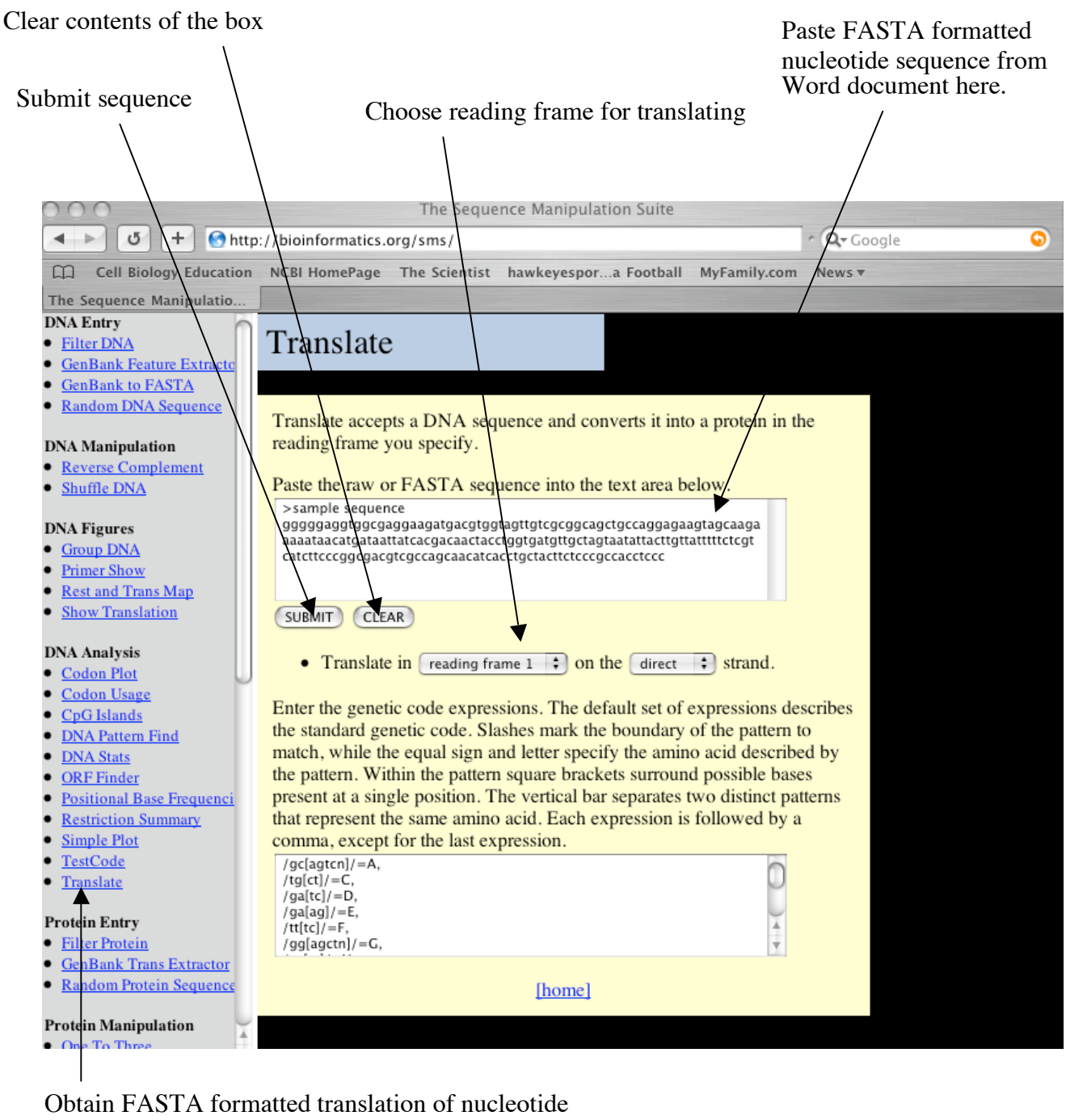
Sequence information

Length: **604 AA** [This is the length of the unprocessed precursor]	Molecular weight: **68996 Da** [This is the MW of the unprocessed precursor]	CRC64: **72FBD699F6128519** [This is a checksum on the sequence]

10 20 30 40 50 60

Best sequence known for this protein

The Sequence Manipulation Suite:



Taken from http://www.ebi.ac.uk/clustalw/index.html

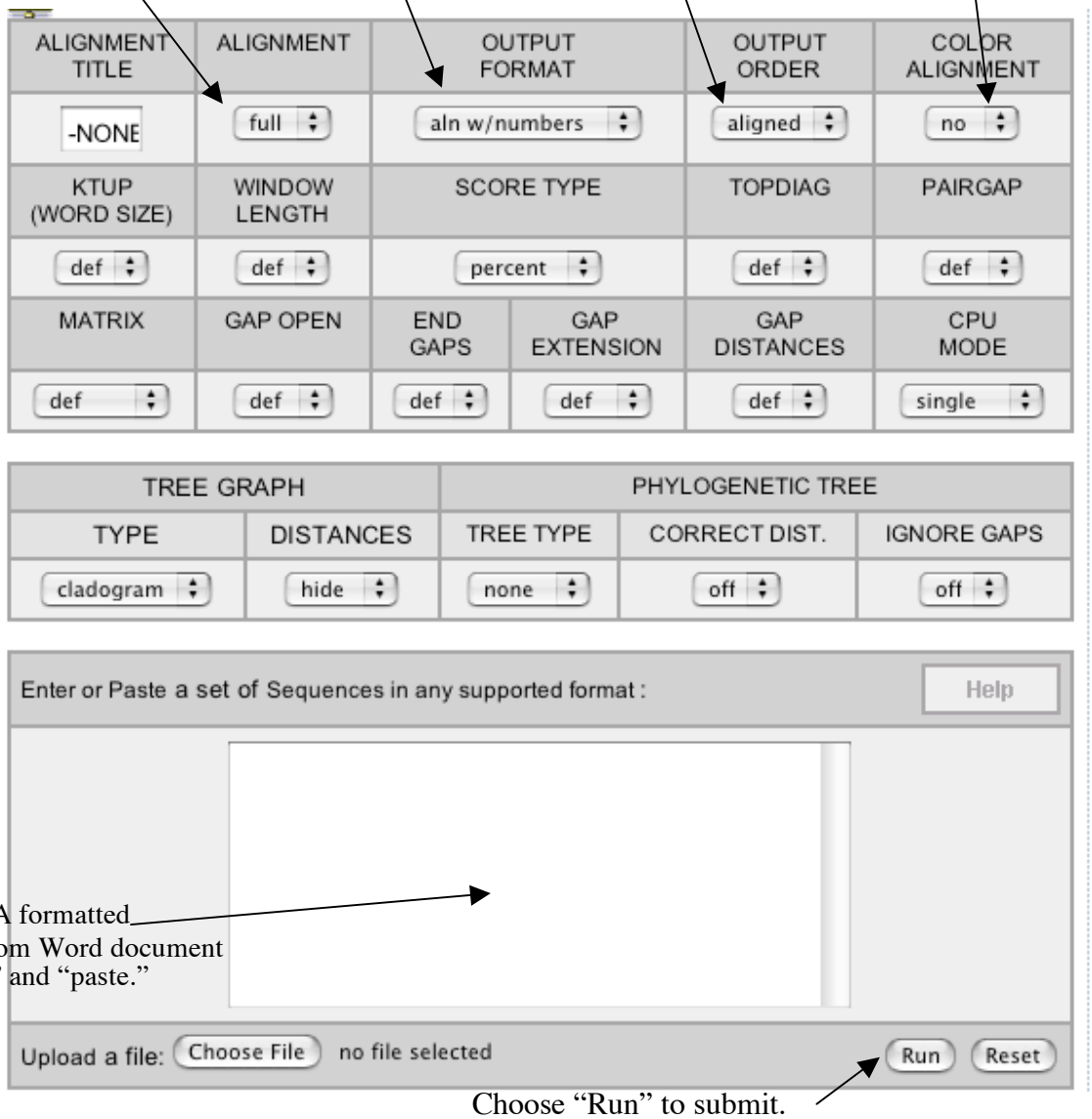
ClustalW Submission Form:

We want the human sequence on top, so output should be in order of input, not order of aligned. Change this to "input."

Choose "full"

Align with numbers, so we can find particular residues in the alignment.

No color for easier printing.



Enter FASTA formatted sequences from Word document using "copy" and "paste."

Choose "Run" to submit.

Background on PTGS1 (Cox-1) and PTGS2 (Cox-2):

Follow these directions to access the entries for PTGS1 and PTGS2 in the "Gene" database at the NCBI Website:

A. First, go to the NCBI homepage using the link on the lab webpage, or by going to: **http://www.ncbi.nlm.nih.gov**
B. Select "Gene" from the database pulldown menu. Type "PTGS" in the search box, then click "Go."
C. Scan the results for the "*Homo sapiens*" entries. There should be one called "PTGS1" and one called "PTGS2."
D. Select each entry by clicking on its name, then read the paragraph under the "Summary" section for each entry.

Read the "Summary" section for both of these genes, then answer the questions below.

1. PTGS1 and PTGS2 are isozymes. Isozymes catalyze the same reaction, but are separate genes. What types of reactions to PTGS enzymes catalyze? Also, what pathway are these enzymes a part of?

2. How is the expression of PTGS1 and PTGS2 different?

3. Which isozyme would you want to inhibit to stop inflammation?

The next two questions are not discussed in the summaries. Do your best to answer the questions, then we will discuss them as a class.

4. The drug Celebrex selectively inhibits PTGS2 while aspirin and other NSAID's inhibit both PTGS1 and PTGS2 in the same way. Why do you think researchers wanted to discover a selective inhibitor to PTGS2?

5. Describe how studying 3-D structures of PTGS1 and PTGS2 could help researchers design a drug that binds to PTGS1, but not to PTGS2.

Part 1 – Getting sequence information and viewing database entries

NCBI – Gene

1. Go back to the "Gene" entry for *Homo sapiens* PTGS2.

2. What is the gene name?

3. What is the GeneID number?

4. Where in the human genome is this gene located?

5. What is the RefSeq accession number for the mRNA sequence of *Homo sapiens* prostaglandin-endoperoxide synthase 2?_____.

6. Open the entry, then choose "FASTA" from the pull-down menu. Copy the sequence (including the title line designated by the ">" symbol) and paste it into a word document.

7. Select the "Replace" tool under the EDIT menu. In the "find" box, type "^p" to find all paragraph marks. Don't type anything into the "replace" box. Then click "Replace All." This will eliminate all the paragraph marks in the document. If you still see white spaces in the sequence, use the same procedure, but type "^w" in the "find" box to represent white spaces.

8. You now should add back a paragraph mark after the title line (that starts with ">") and before the sequence starts. Save the file as PTGS2rna.doc on your desktop.

Note: Please review the entry for "FASTA" in the Glossary (at the end of this manual). Understanding this definition will be very important for working with the bioinformatics programs.

9. What is the RefSeq accession number for the *Homo sapiens* PTGS2 protein sequence?_____. Open the entry. Follow the steps given above to save the sequence in FASTA format as a Word document called PTGS2prot.doc file on your desktop.

Swiss-Prot Entry

10. Go to the Expasy website and search for the Swiss-Prot entry for PTGS2. (Hint: use the gene name to search and be sure to select the HUMAN protein from the search results).

11. Write at least three alternate names for this protein.

12. Where in the cell is this protein located?

13. What types of drugs target this protein?

14. What amino acid is acetylated by aspirin (amino acid type and number)?

15. What His residue is in the active site?

Sequence Manipulation

16. Go to the Sequence Manipulation Suite (http://bioinformatics.org/sms/).

17. Click on "Translate" under "DNA Analysis" heading from the menu.

18. Clear the data entry box by hitting "Clear".

19. Copy the mRNA sequence from your Word file and Paste it into the data entry box on the Sequence Manipulation website.

20. Select "Reading Frame 3" and "direct" from the pull-down menus, then click "Submit".

21. When the Output window opens with your results, copy and past the sequence into a Word document and save it as, "translate.doc" on your desktop.

22. Compare this sequence in the "translate.doc" file with the sequence in the "Cox2rna.doc". What are the first residues that are the same in the sequences? Do the sequences look like they are the same? (Hint: protein sequences should start with a methionine.)

Part 2 – Multiple Sequence Alignment with ClustalW

23. On the course website under "COX2 Tutorial", there should be a file called "ClustalWseq". Click on this link to download the pdf file to your desktop. To open the file in Acrobat Reader, hold the curser over the file on your desktop, hold down the "Control" key on your keyboard as you click and hold with the mouse. A list of programs should appear on your desktop. Highlight "Adobe Acrobat," then release the mouse. The file should open in Adobe Acrobat. If you have the file open in "Preview," you will not be able to select multiple pages of text at the same time, so close the file and open it in Acrobat Reader as described above. Using the "Text select" tool, select all the text and copy or go under the "Edit" menu and "Select All." This file contains six FASTA formatted sequences of PTGS2 from different organisms. The top sequence is the human PTGS2 protein sequence you have been working with.

24. Go to the ClustalW website and enter (by using "copy" and "paste") all of the FASTA formatted sequences into the data entry box. Select "input" for the Output order so the human sequences will stay at the top in the alignment. Press "Run."

25. Copy the alignment and paste it into a Word document. To make this file readable, do the following things:
    a. Go to "Page Set-up" under "File" and change the page orientation to landscape.
    b. Select all text and change to "Courier" font, size 10. Courier is the best font for alignments because all the letters are the same width.
    c. Save this file to the desktop as "ClustalW.doc"

26. Review the alignment. What symbols are used for positions in the alignment that contain identical, highly homologous, homologous, and non-homologous residues? Are the residue numbers mentioned in steps 13 and 14 conserved? Why would you expect them to be conserved?

Laboratory 2 - Project Research
Guide Sheet 1

Working with Primary Protein Structure Information

The research projects begin by using some of the same tools you used last week in the COX-2 tutorial. Guide sheet 1 will provide directions for the web-based research you will be doing today. First you will search for your gene in the Gene database. Gene will contain the RefSeq sequence for your protein, which you will download in FASTA format. FASTA format is defined in your Glossary. Be sure to review this definition before begin your on Guide sheet 1. You will continue to learn about your protein using the SwissProt database. The SwissProt database is a database maintained by the Swiss Bioinformatics Institute and contains entries for thousands of proteins. You can search for the protein you are studying by using the gene name given in Gene. The SwissProt entry contains some of the same information that you found in Gene, but also contains a lot information about the protein sequence, structure, and function that is summarized in a short, easy-to-read format.

The ultimate goal for today's lab is to create a multiple sequence alignment for your protein using ClustalW. You will use this alignment to identify the protein mutation, to observe regions of high sequence conservation, and to map secondary structure predictions. The protein mutation is important to identify since it is the basis for your project and is important for understanding the link between the protein and the disease you are studying. The regions of high sequence conservation are important because they often correspond to regions in the protein that are important to the protein's function. Next week, you will be studying the crystal structure of your protein and will be able to check the accuracy of the secondary structure predictions that you mapped onto your alignment. The program used for making the secondary structure predictions is called PSIPRED. If your protein is a membrane protein, you will also have MEMSTAT predictions, which predict which regions of your protein, are imbedded in a membrane.

Guide Sheet 1

Part 1 – Obtaining the basics: Getting sequence information and viewing the SwissProt and GenBank entries for your protein

Directions: Follow this guide sheet and answer the questions in your project packet that accompany this guide as you work through each section. Be sure to refer to hints specific to your project in your project packet.

Translating your patient's cDNA

1. Obtain the mutant cDNA sequence from the course website. Open the file and copy the sequence.

2. Go to the Sequence Manipulation Site (http://bioinformatics.org/sms/).

3. In the menu to the left, Click on "Translate" found under the heading "DNA analysis".

4. Clear the search box, then paste your patient's cDNA sequence into the search box. Choose a reading frame from the pull-down menu. Click "Submit."

5. You should be able to find the sequence of your protein by finding the first methionine (M), then continuing until you see the first "*" which is a stop codon. Copy the protein sequence in that region, starting with the first "M" and paste it into a word document. Save the document in your folder on the desktop. Now you have saved the file of the mutant protein sequence.

NCBI – Gene

6. Using Gene on the NCBI website, find the entry for the protein you are studying by searching with the protein name. Be sure to select the *Homo sapiens* protein from the list of results. Answer question 1.

7. Open the entry for the RefSeq protein sequence and save the sequence in FASTA format to your desktop. Name this file "wildtypeprot.doc" so that you will know it is the un-mutated protein sequence.

Swiss-Prot Entry

8. Go to the ExPASy website and search for the SwissProt entry for your protein using either the protein name the gene name. Be sure to select

the human protein from the list of results. Make sure the information in the entry is the same as you saw in the Gene entry. If your protein is an enzyme, the EC number is a good way to double-check. You may want to record the SwissProt entry number in case you want to find this entry again.

Part 2: BLAST sequence: Finding homologous proteins

Protein-protein BLAST

9. Perform a BLAST search using the RefSeq protein sequence (the un-mutated protein sequence). First, select PSI-PHI BLAST. Then paste the FASTA formatted protein sequence in the search box. Select the nr-protein database. Click "BLAST" to begin. On the next page that appears, select "Format." You may need to wait a few minutes before the results page opens. After obtaining the results, choose 5 sequences from various positions in the results. The goal is to choose a variety of sequences that differ in evolutionary distance from the human protein. Be sure not to choose any sequences that are human, since they are the same as your search sequence. For each of the five sequences, click on the sequence name to view the GenBank entry for the sequence. Then view the sequence in FASTA format. Copy and paste all the FASTA formatted sequences into the same Word file and save it to the desktop. At the beginning of this file, add your mutant protein sequence, also in FASTA format.

10. This Word file will be used to create the multiple sequence alignment, so the formatting is very important. The format for this file should be like the example used in the tutorial for COX-2. Review the entry for FASTA format in the Glossary. You should end up with a Word file that contains the 5 sequences from the BLAST search plus the un-mutated human protein sequence and your mutant sequence for a total of 7 sequences. Each sequence should be in FASTA format and contain a title line (starting with >, then text, then a return). Shorten the text to contain JUST the species information so it will fit in the alignment (next step). For example, you should erase the "gi" line and add in something simpler like "pig," "cow," etc. Your mutant sequence should read ">mutant". At the end of each title, be sure to **press return** to separate it from the rest of the sequence.

Part 3 – Multiple Sequence Alignment

11. Go to the ClustalW website and enter (by using "copy" and "paste") all your FASTA formatted sequence into the data entry box. The default parameters will work for us, except for the output order.

a. Select “input” for the Output order
b. Press “run”

12. Save the alignment by copying and pasting the alignment into a Word document. At first the alignment will look broken up. Follow these steps to make it readable again.
    a. Select all
    b. Change the font to size 10 and Courier
    c. Change the page set-up to landscape
    d. Save the file to your desktop

13. Scroll through the alignment and make sure none of the blocks of sequences are separated by a page break. Save and print the alignment. We will be working with this alignment next week and it will be part of your final report.

14. On your alignment, compare the wild-type protein sequence with your mutant sequence and mark any differences that you observe.

Secondary Structure Predictions – using PSIPRED

15. To save time, the PSIPRED and MEMSTAT predictions for your protein are already available for you in the same format as we received them. Obtain these results from the course website under your project name and print them out. The Glossary provides more information on how to submit requests for PSIPRED and MEMSTAT predictions if you want to use these tools outside this course.

16. On your Multiple Sequence Alignment, draw the regions of secondary structure and transmembrane predictions on top the corresponding sequence. Use the following symbols to represent predictions:
    h = helix
    b = beta sheet
    Mark nothing for coil predictions. Use the symbol above and an arrow to show how long each region of secondary structure spans on the alignment. See an instructor for examples if these directions are unclear.

From this lab, you should save the multiple sequence alignment for your Final Report. The multiple sequence alignment should have the secondary structure predictions and the mutation mapped.

NEXT WEEK: Do the structure problem set for next week. It’s important to view the problem set online from the lab website in order to see the structures in color!

Laboratory 3 - Project Research
Guide Sheet 2

Investigation of the Crystal Structure

The goal of this lab is to analyze the crystal structure for your protein in order find the functionally important areas of your protein and predict the effect of the mutation on protein function. You may determine that the mutation makes the protein more active, less active, or that the mutation is likely to have no effect of protein function.

To do these investigations, we will be using the protein structure-viewing program called DeepView/Swiss-Pdb Viewer. This program has some similarities to programs you have used in previous course, like Chime, and Protein Explorer. DeepView can easily model mutations and is easy to learn to use in one day. It is not a web-based program, but both the Mac and PC versions can be downloaded for free from the ExPASy website. There are also DeepView tutorials available for free from the ExPASy website. At the beginning of lab, a brief tutorial on how to use DeepView will be given using the COX-2 protein.

The first step today is to obtain the data file of the crystal structure for your protein. In some cases, the crystal structure has not been solved for your exact protein, so you will analyze the structure and model the mutation in a homologous protein. In this case, the amino acid numbering will be slightly different, but details are given in your project packets. The data files for crystal structures are called pdb files and are all stored in the Protein Data Bank. You can search the Protein Data Bank website for your protein and download the pdb file to your desktop.

In analyzing the crystal structure, the focus is on identifying the noncovalent interactions (Van der Waals, H-bonds, and ionic bonds) of the amino acid side chain before and after it is mutated. How do the interactions change or how are they maintained when the amino acid is mutated? Be sure to review both the distances and nature of these non-covalent interactions before coming to this lab.

Crystal structures are the results of experiments, and so it is important to consider experimental error when analyzing a crystal structure. One way experimental error is reported in a crystal structure is by giving the resolution. Resolution of a crystal structure is the accuracy of the prediction of each atom location. For example, if the resolution is 2 angstroms, you can be confident that the atom is located within a 2 angstrom radius of where it is shown in the pdb file. This is important to consider when measuring distances, since each distance measured in a crystal structure is actually ± the resolution of the crystal structure.

Tool Bar

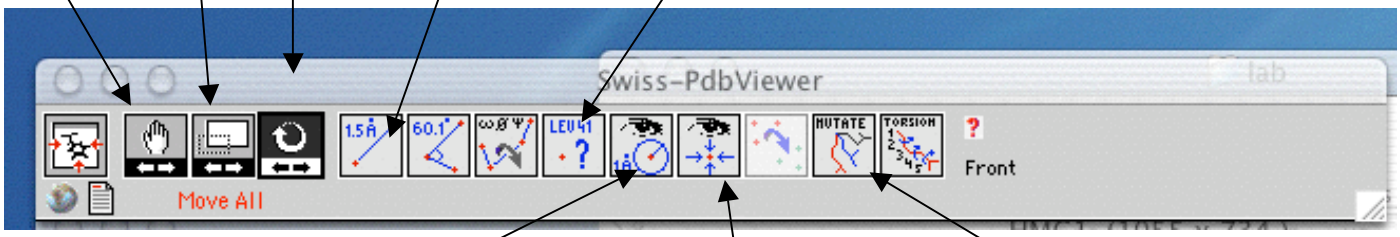
Transverse Movement (along plane of screen)

Zoom

Rotate

Distance Tool - measure the distance between two atoms in Angstroms

Identity Tool - the residue type and number will appear in the display for the atom you select



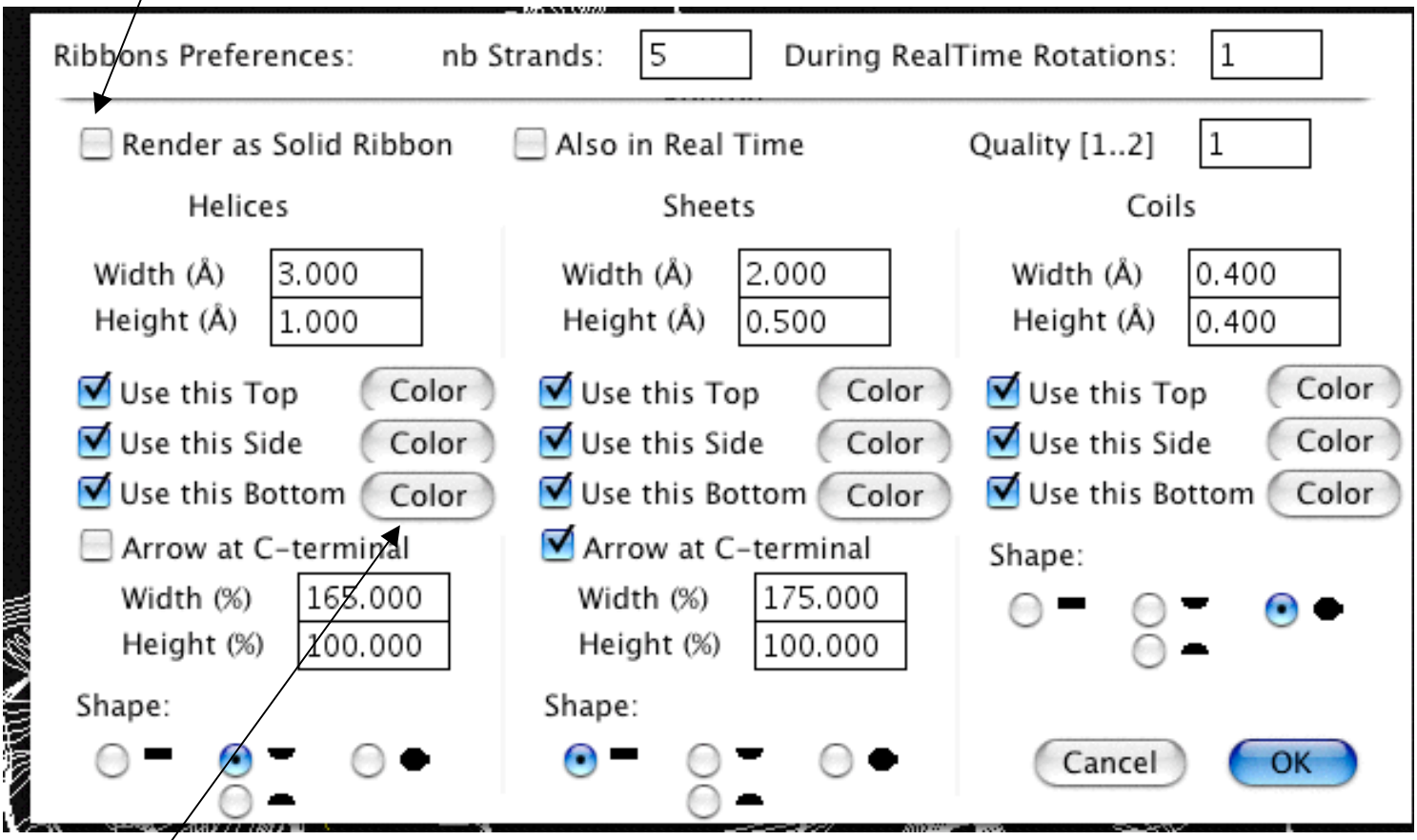
Radius Tool - the side chains of residues within a certain distance (you specify) will appear around the atom you select.

Re-center Tool - the display will re-center around the atom you select and will rotate around that atom when you use the rotate tool.

Mutate Tool - When you click on this tool, then an atom in the display, you can mutate a side chain to another type and it will appear in the display in the lowest energy conformation

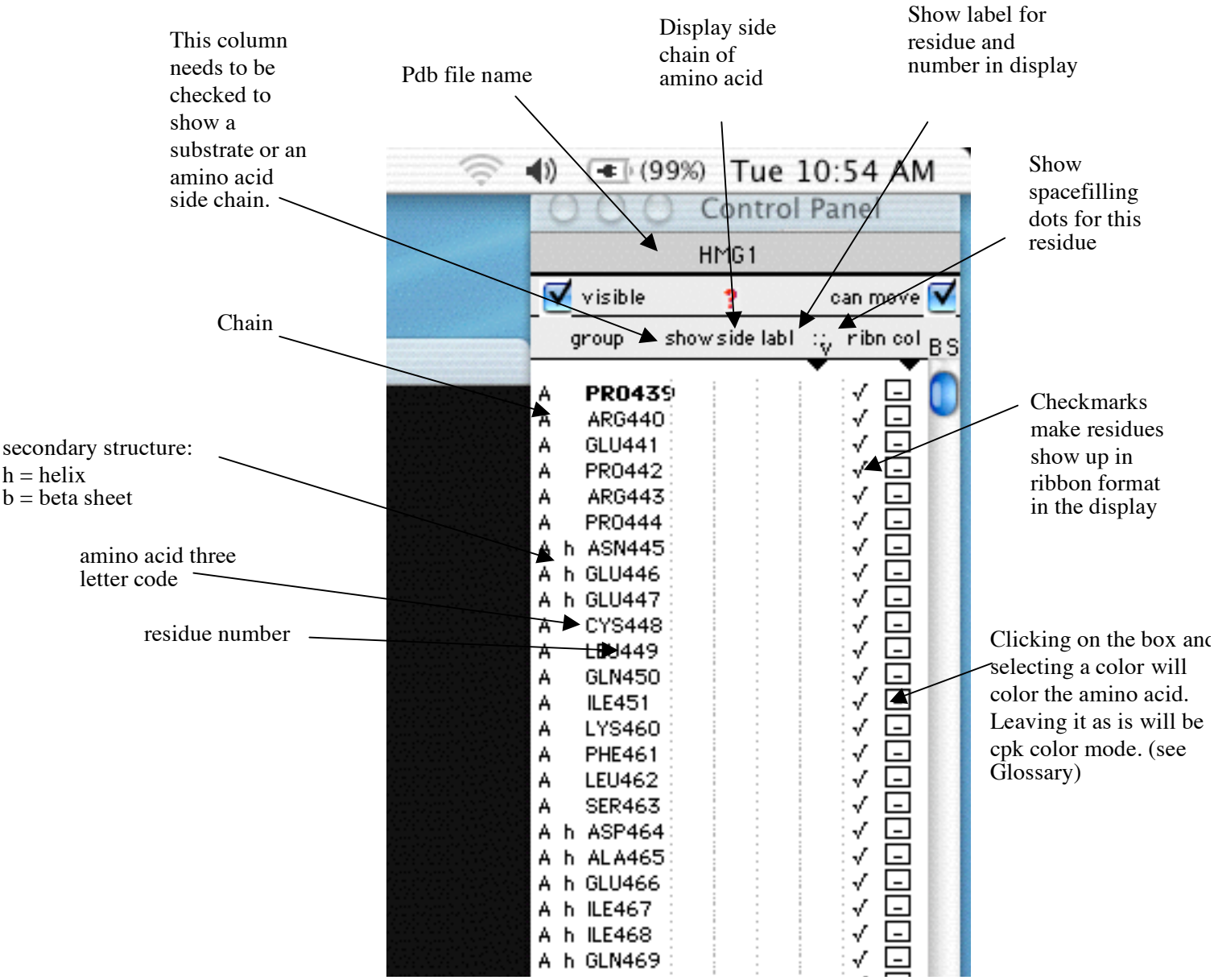
Panel under : “Prefs” ⟶ “Ribbons”

Make sure there is a check in this box to show a cartoon ribbon in the display



In this panel, you can choose a different color for each side of the 3-dimensional ribbon as well as different colors for the helices, sheets, and coils (loops). There should be checkmarks as shown here for each color box. When you click on the “Color” button, a panel appears to choose a color. It is a good idea to choose the same color for each of the top, bottom and sides of all helices. You don’t need to change any of the other settings in this panel for this course. Click “OK” when finished.

Control Panel



Guide Sheet 2
Obtaining 3D Structure Information:

Searching for Structure Files (pdb files):

Data files that contain the three-dimensional coordinates for protein structures are called “pdb” files. Pdb files are named in 4 characters (numbers and letters). These files are stored in the Protein Data Bank (www.rcsb.org/pdb). You can perform a search to find the structure you are looking for either from the NCBI website using their “Structure” database, or search directly at the Protein Data Bank website. The search results are a little easier to understand from the NCBI website. If you know exactly what you are looking for, you can directly search the Protein Data Bank.

Follow the directions in your project manuals to get directions for finding the pdb file and analyzing the structure in Swiss-Pdb Viewer/DeepView.

Laboratory 4 - Project Research
Guide Sheet 3

Investigating the biological impact of the mutation - using the OMIM and KEGG databases

For this lab session, your research will focus on investigating the biological impact of the mutation you are studying. To do this, you will use the OMIM and KEGG websites. OMIM stands for the Online Mendelian Inheritance in Man database. The OMIM database was started at John Hopkins University and is now maintained by NCBI and can be found through a link on the NCBI homepage. The OMIM database contains entries for both diseases with known genetic links and entries for the genes that have been linked to a disease. Each OMIM entry is a summary of the research that has been performed on the disease or gene and contains links to the research articles that it summarizes. You will be able to read about the clinical and biochemical research that has been performed related to the mutation you are studying. Each link in the OMIM entry will open an abstract from the PubMed database. PubMed is a literature database, and is also maintained by NCBI. PubMed is a searchable database of medical and life science journal articles. Most of the abstracts for these articles can be accessed through PubMed, but in order to access the entire article, you need to go to each individual journal website and have a subscription to the journal. The WashU library system has subscriptions to electronic versions of many of these journals that you can access through the E-journal link on the WashU library home page. Most journals have their articles available online as .pdf files for articles published between 1995 to present. However, the older articles must still be accessed through the paper versions stored in libraries.

The other database that you will be using is the KEGG database. KEGG stands for the Kyoto Encyclopedia of Genes and Genomes. It is a database of metabolic pathways that is maintained by a research institute in Japan. It contains all the known metabolic and signaling pathways. Each protein in the pathway and each small molecule metabolite (ex. ATP) has its own entry in the database that can be accessed by clicking on the protein or metabolite in the pathway figure. By using this website, you can make predictions about what would happen to downstream events in the pathway if the protein you are studying is either less active or more active.

In order to write the summary for your final report, you will need to use the information from this lab's research as well as what you learned from studying the crystal structure of your protein in order to draw a conclusion about what the biological impact is of the mutation you are studying. This conclusion should be supported by the web-based research you have performed, but will most likely require further clinical and biochemical research in order to be proven or disproved.

Guide Sheet 3

Using Web-based resources to investigate the biological impact of the mutation and its possible role in human disease

OMIM search: The OMIM (Online Mendelian Inheritance of Man) database contains short, referenced reviews about genetic loci and genetic diseases. It can be a very useful resource for finding out what type of research has been done on a gene or a disease.

KEGG search: The Gene entry for your gene that you viewed during the first day of your research (Guide Sheet 1) will contain a link for the KEGG pathway(s) related to your protein. Scroll down to the "Additional Links" and select the "KEGG pathway" link. KEGG stands for Kyoto Encyclopedia of Genes and Genomes. The www.kegg.com site contains a database of metabolic maps.

Follow the directions in your project manuals to answer Guide Sheet 3 questions.

Laboratory 5 – Final Reports and Joint Quiz

On the final lab day, you should come to class with your reports ready to hand in. The format for the report is provided in the laboratory syllabus. You can use your report as a reference while giving your oral reports in the small group and for the joint quiz. The format of the class will be to first meet in your small groups of 4 – 6 and present your reports. You will need to plan with your partner which part of your research project you will each be presenting. For example, one partner should present the introduction to the protein and disease and the multiple sequence alignment and the other partner should present the findings from the structure analysis, OMIM, and KEGG research. Each person can explain his or her final conclusions. It is possible that the two partners have reached different final conclusions about the biological impact of the mutation, and both conclusions should be presented. Next, the people working on the other project in the group will present. Finally, the group will collaborate to provide answers to the joint quiz. Each member of the group will write their own answers to the quiz questions, but the group can discuss the answers before writing them down. You will also be able to use your notes, guide sheets, and reading answers. There will be questions from both projects on the joint quiz.

Glossary for Bio3055

BLAST – Basic Local Alignment Search Tool – A program that compares a sequence (input) to all the sequences in a database (that you choose). This program aligns the most similar segments between sequences. BLAST aligns sequences using a scoring matrix similar to BLOSUM (see entry). This scoring method gives penalties for gaps and gives the highest score for identical residues. Substitutions are scored based on how conservative the changes are. The output shows a list of sequences, with the highest scoring sequence at the top. The scoring output is given as an E-value. The lower the E-value, the higher scoring the sequence is. E-values in the range of 1^-100 to 1^-50 are very similar (or even identical) sequences. Sequences with E-values 1^-10 and higher need to be examined based on other methods to determine homology. An E-value of 1^-10 for a sequence can be interpreted as, "a 1 in 1^10 chance that the sequence was pulled from the database by chance alone (has no homology to the query sequence)."

This program can be accessed from the NCBI homepage or:
http://www.ncbi.nlm.nih.gov/BLAST

Reference: Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

BLOSUM – Block Scoring Matrix - A type of substitution matrix that is used by programs like BLAST to give sequences a score based on similarity to another sequence. The scoring matrix gives a score to conservative substitutions of amino acids. A conservative substitution is a substitution of an amino acid similar in size and chemical properties to the amino acid in the query sequence. Discussed in the Berg text, p.175 – 178.

Bioinformatics - Bioinformatics is a field of study that merges math, biology, and computer science. Researchers in this field have developed a wide range of tools to help biomedical researchers work with genomic, biochemical, and medical information. Some types of bioinformatics tools include data base storage and search programs as well as software programs for analyzing genomic and proteomic data.

ClustalW – A program for making multiple sequence alignments.
http://www.ebi.ac.uk/clustalw/index.html

W. R. Pearson (1990) "Rapid and Sensitive Sequence Comparison with FASTP and FASTA" Methods in Enzymology 183:63 - 98.

Conserved – when talking about a position in a multiple sequence alignment, "conserved" means the amino acid residues at that position are identical throughout the alignment.

Conservative residue change – when talking about a position in a mulitple sequence alignment, a "conservative change" is when there is a change to a homologous amino acid residue.

cpk coloring mode - This coloring mode colors based on atom identity:
red = oxygen
blue = nitrogen
orange = phosphorous
yellow = sulfur
gray = carbon

DeepView/Swiss-Pdb Viewer – a program for viewing 3-D structures. It loads ".pdb" files, which contain the 3-D coordinates for molecular structures. Swiss-Pdb Viewer is easy and free to download on any computer (Mac of PC) and can be used no matter what Browser you are using. It is fairly easy to learn to use at the basic level, however, it also has very advanced capabilities that can be useful in research. It is also a nice program to use with PovRay, which allows you to make graphic files from pdb information. This is important when making figures for a presentation, report, or journal article. If you would like to download Swiss-Pdb Viewer for your own computer, the program is available for free and is easy to download from the website, "us.expasy.org/spdbv". A help manual is also available here if you have further questions that aren't addressed in this course.

http://us.expasy.org/

To run this program with Mac OSX, you must first change the monitor settings.

a. Open "System Preferences" on your computer.
b. Double click on the "Displays" icon.
c. On the right-hand side of the panel, choose "thousands" of colors from the list (changing it to "thousands" from "millions").
d. Then close System Preferences and then open Swiss-Pdb Viewer.

Names of some other structure viewing programs:
RasMol (www.openrasmol.org)
Kinemage (www.kinemage.biochem.duke.edu)
Protein Explorer (www.proteinexplorer.org)

EC number - Enzyme Committee number - Given by the IUBMB (International Union of Biochemistry and Molecular Biology) classifies enzymes according to the reaction catalyzed. An EC Number is composed of four numbers divided by

a dot. For example the alcohol dehydrogenase has the EC Number 1.1.1.1

ExPASy – Expert Protein Analysis System - A server maintained by the Swiss Institute of Bioinformatics. Home of SWISS-PROT, the most extensive and annotated protein database. The Swiss-Pdb Viewer protein-viewing program is also available at this site for free download.

http://us.expasy.org/

FASTA – A way of formatting a nuleic acid or protein sequence. It is important because many bioinformatics programs require that the sequence be in FASTA format. **The FASTA format has a title line for each sequence that begins with a ">" followed by any needed text to name the sequence. The end of the title line is signified by a paragraph mark (hit the return key).** Bioinformatics programs will know that the title line isn't part of the sequence if you have it formatted correctly. The sequence itself does NOT have any returns, spaces, or formatting of any kind. The sequence is given in one-letter code. An example of a protein in correct FASTA format is shown below:

>K-Ras protein *Homo sapiens*
MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDI
LDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVP
MVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIETSAKTRQGVDDAFYTLVREIRK
HKEKMSKDGKKKKKKSKTKCVIM

GenBank - a database of nucleotide sequences from >130,000 organisms. This is the main database for nucleotide sequences. It is a historical database, meaning it is redundant. When new or updated information is entered into GenBank, it is given a new entry, but the older sequence information is also kept in the database. GenBank belongs to an international collaboration of sequence databases, which also includes EMBL (European Molecular Biology Laboratory) and DDBJ (DNA Data Bank of Japan). In contrast, the RefSeq database (see entry) is non-redundant and contains only the most current sequence information for genetic loci. The GenBank database can be searched at the NCBI homepage:

http://www.ncbi.nlm.nih.gov/

Gene – an NCBI database of genetic loci. This database used to be called LocusLink. Entries provide links to RefSeqs, articles in PubMed, and other descriptive information about genetic loci. The database also provides information on official nomenclature, aliases, sequence accession numbers, phenotypes, EC numbers, OMIM numbers, UniGene clusters, map information, and relevant web sites. Access through the NCBI homepage by selecting "Gene" from the Search pulldown menu.

Genome – The entire amount of genetic information for an organism. The human genome is the set of 46 chromosomes.

Homologous – When referring to amino acids, a homologous amino acid is similar to the reference amino acid in chemical properties and size. For example, glutamate can be considered homologous to aspartate because both residues are roughly similar in size and both residues contain a carboxylic acid moiety which gives them similar chemical properties.

KEGG – Kyoto Encyclopedia of Genes and Genomes – This website is used for accessing metabolic pathways. At this website, you can search a process, gene, protein, or metabolite and obtain diagrams of all the metabolic pathways associated with your query. You will see a link to the KEGG entry at the end of the Gene entry for a gene.

http://www.genome.ad.jp/kegg/

NCBI – National Center for Biotechnology Information – This center was formed in 1988 as a division of the NLM (National Library of Medicine) at the NIH (National Institute of Health). As part of the NIH, NCBI is funded by the US government. The main goal of the center is to provide resources for biomedical researchers as well as the general public. The center is continually developing new materials and updating databases. The entire human genome is freely available on this website and is updated daily as new and better data becomes available. The NCBI homepage:

http://www.ncbi.nlm.nih.gov

NCBI also maintains an extensive education site, which offers online tutorials of its databases and programs:

http://www.ncbi.nlm.nih.gov/About/outreach/courses.html

OMIM - Online Mendelian Inheritance in Man – a continuously updated catalog of human genes and genetic disorders, with links to associated literature references, sequence records, maps, and related databases. Access through the NCBI homepage or:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

Protein Data Bank – (PDB) – A database that contains every published 3-D structure of biological macromolecules. It contains mostly proteins, but also DNA and RNA structures. Also see RCSB.

http://www.rcsb.org/pdb/

A pdb file is a file containing the three-dimensional coordinates (x,y,z) for each of the atoms in the protein. This type of file is made using the data obtained from

either an X-ray crystallography experiment or an NMR experiment. Once you have pdb file of a protein, you can open the file in various structure viewing programs to view the protein structure.

Proteome – the entire set of expressed proteins for an organism. This term is commonly used to discuss the set of proteins that are expressed in a certain cell type or tissue under specific conditions.

PSIPRED – a server for predicting secondary structure from protein sequences. The predictions are made based on a database of known secondary structures for protein sequences. These predictions are estimated to be correct ~80% of the time. This server can also be used to predict transmembrane segments.

http://bioinf.cs.ucl.ac.uk/psipred/

McGuffin LJ, Bryson K, Jones DT. (2000) The PSIPRED protein structure prediction server. Bioinformatics. 16, 404-405.
Jones DT. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292: 195-202.

PubMed – a retrieval system containing citations, abstracts, and indexing terms for journal articles in the biomedical sciences. PubMed contains the complete contents of the MEDLINE and PREMEDLINE databases. It also contains some articles and journals considered out of scope for MEDLINE, based on either content or on a period of time when the journal was not indexed, and therefore is a superset of MEDLINE.

http://www.ncbi.nlm.nih.gov/

RCSB – Research Collaborative for Structural Bioinformatics – A non-profit consortium that works to provide free public resources and publication to assist others and further the fields of bioinformatics and biology dedicated to study of 3-D biological macromolecules. Members include Rutgers, San Diego Supercomputer Center, University of Wisconsin, and CARB-NIST (at NIH).

RefSeq - NCBI database of Reference Sequences. Curated, non-redundant set including genomic DNA contigs, mRNAs, proteins, and entire chromosomes. Accession numbers have the format of two letters, an underscore bar, and six digits. Example: NT_123456. Code: NT, NC, NG = genomic; NM = mRNA; NP = protein (See NCBI site map for more of the two letter codes).

Sequence Manipulation Suite – a website that contains a collection of web-based programs for analyzing and formatting DNA and protein sequences.
http://bioinformatics.org/sms/

SNP = Single Nucleotide Polymorphism.

synonymous change– The nucleotide change results in NO change in amino acid.

non-synonymous change – The nucleotide change DOES result in a change in amino acid.

heterozygosity – A measure of the genetic variation in a population with respect to one locus. Stated as the frequency of heterozygotes for that locus.