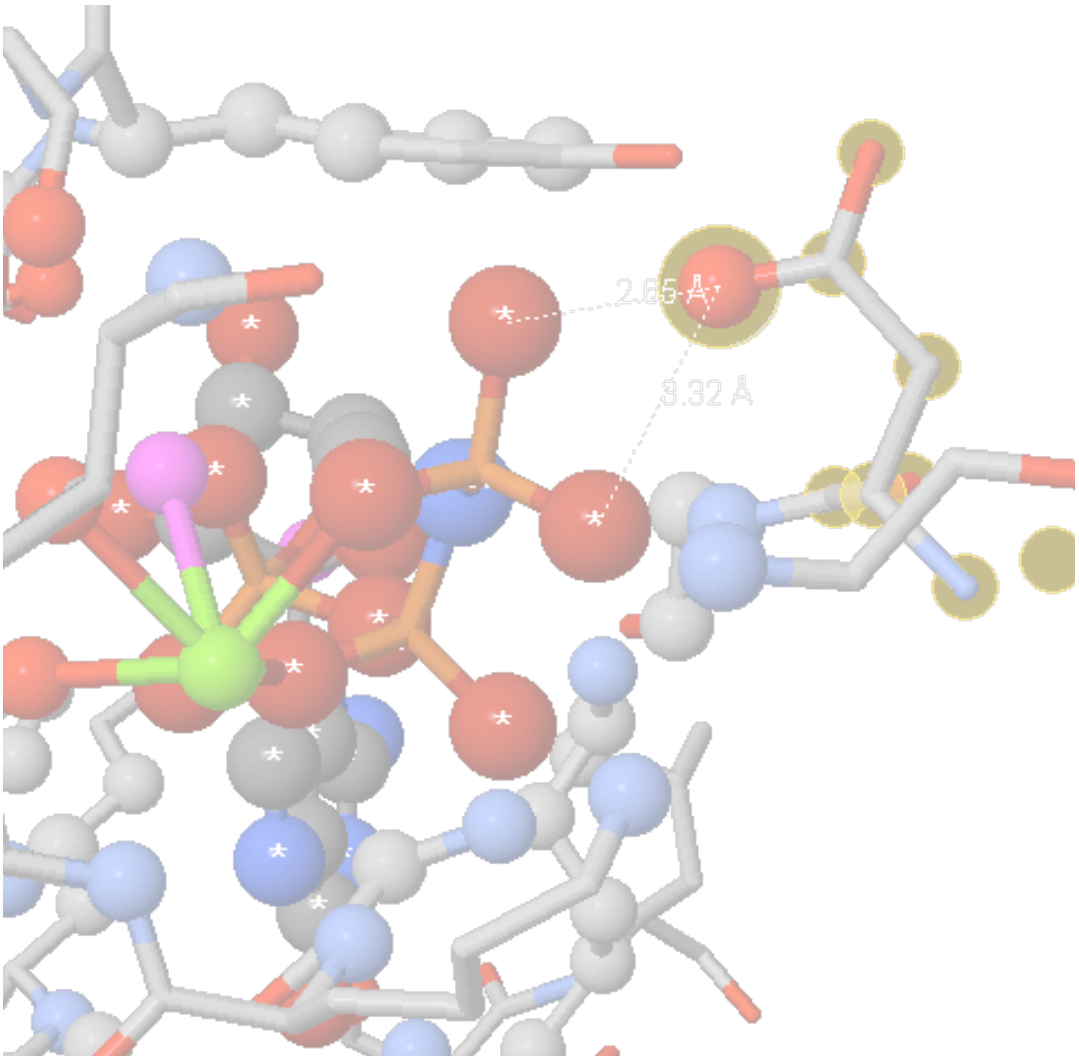
Bioinformatics Tools
Tutorial Project

Gene ID: KRas



Bednarski 2011
Original project funded by HHMI

Bioinformatics Projects
Introduction and Tutorial

Purpose of this tutorial

Illustrate the link between genotype and phenotype

The KRas tutorial guides you through a detailed look at how sequence information in the form of a known genetic mutation can lead to understanding the structure and functional changes in the mutant protein and ulitmately provide a better understanding of the genetic disease physiology – in this case, lung cancer. It is designed so that you can have the opportunity to examine some of the data yourself, observe the limits of what is known, and form your own hypothesis about the effect of the mutation on protein structure and function.

Provide a survey of genomic data and tools freely and readily available online

This tutorial was designed to use a variety of online bioinformatics tools to help you recognize what types of tools are available and how easy it can be to obtain a wealth of information about a gene, protein, or disease. The main tools highlighted by these projects are listed and described below. While working on this project, you can link to the following sites through a customized webpage, but the tools can be found at a later time easily by searching "Google" with the keywords shown in quotes for each of the following entries.

Summary of Websites used in this tutorial (A more complete glossary is found at the end of the manual):

NCBI homepage "NCBI" – maintained by NIH and houses many powerful databases and genome sequences

- **Gene** database – contains links to DNA and protein sequences and links to entries in other databases like OMIM and KEGG
- **PubMed** database – all major biomedical research articles – abstracts available for free and some free pdf's depending on the journal
- **BLAST** search tool – use DNA or protein sequences to search for homologous sequences from other organisms
- **OMIM** database – summaries of all known genetic diseases and disease genes with links to key journal articles

EMBOSS Tools (at http://gander.wustl.edu/cgi-bin/emboss) a variety of tools for working with protein and nucleic acid sequences including "prettyplot" to format a multiple sequence alignment and "water" a pairwise sequence alignment tool.

ExPASy homepage "ExPASy" – maintained by Swiss Institute of Bioinformatics and houses many proteomics tools and databases

UniProtKB database – entries for all known proteins with information about sequence, structure, function, and expression and links to journal articles and other databases

FirstGlance in JMol – a structure-viewing program (opens .pdb files)

RCSB - Protein Data Bank "pdb" – all known protein and nucleic acid structure files (.pdb files) solved by either X-ray crystallography or NMR

KEGG database "KEGG" – biochemical pathways with links to summary pages on proteins and metabolites in the pathways

KRas

Introduction:

It is well known that smoking leads to an increased risk for lung cancer, but how does genetics play into the risk? The transformation of a normal cell into a cancerous cell can result from many causes. In one model, factors that lead to an increased rate of mutation in DNA increases the chances that a proto-oncogene will be mutated into an oncogene and help lead to a normal cell to be transformed into a cancerous cell. In this module, you will examine one proto-oncogene, K-Ras, which has been associated with many cancers, including lung cancer.

KRas is a G-protein in the signaling pathway for certain growth factors, including epidermal growth factor (EGF). KRas associates with the G-protein coupled receptor which binds the growth factor. When the growth factor binds the receptor, KRas exchanges its bound GDP for GTP and becomes an active GTPase. KRas hydolyzes GTP to GDP initiating a signal transduction cascade that ultimately leads to cellular growth. In mammalian cells, there are three isoforms, including H-, K-, and N-Ras. If KRas is mutated to a constitutively active form, it is an oncogene, meaning it transforms a normal cell into a cancerous cell.

KRas mutations have been found to be more common in smokers than non-smokers and is suggested to be a common source of formation of primary tumors in smokers with lung cancer. One hypothesis is that chemicals in cigarette smoke are converted to more dangerous chemicals by cytochrome P450 enzymes. These chemicals can form PAH-DNA adducts which increase the rate of mutation of DNA. Many mutations may have no effect, but if KRas is mutated at residue 12 from G to C, a tumor is the likely result.

In part 1, you will review database entries for KRas to learn more about this gene and its protein product. One of the databases, OMIM, will list any mutant forms of KRas that have been found in patients. In part 2, you will explore the sequence and structure conservation of this gene in different organisms using BLAST and EMBOSS tools. In Part 3 you will identify an oncogenic mutation of KRas and explore the protein structure in detail to predict

why this mutation is oncogenic. You will use EMBOSS tools and Firstglance in JMol in order to link sequence to structure to function.

Part 1

Getting started – find the entry for your gene in the NCBI-Gene database. This database is a great starting place because you can access the correct nucleic acid and protein sequences for your gene, learn about isoforms, read quickly about function, and get links to articles in PubMed about your gene. This database will also give you accession numbers and other names used for the same gene, so when you go to other databases, you can make sure you find the same gene there.

NCBI – Gene

1. Google "NCBI" to get to the NCBI website. Select the "Gene" database from the pull-down menu and search for the entry for "KRAS2". Be sure to select the *Homo sapiens* protein from the list of results.

2. Read the first paragraph on the entry page and look at the links to the right of the screen to find all the information you need here.

3. Fill in the following information from the Gene entry:
    a. Write the GeneID number here _____.
    b. What is the gene symbol?

    c. What are two other names for this gene?

    d. Where in the human genome is this gene located (what chromosome)?

    e. What is the RefSeq (see Glossary) number for the mRNA sequence? (Under "Reference Sequences" on the menu on the right) Use the RefSeq entries for the mRNA and protein sequences for K-Ras2 isoform b – also called "variant (b)."

    f. What is the RefSeq number for the protein sequence?

4. Open the entry for the RefSeq protein sequence by clicking on the name and choosing the "fasta" link on the next page. Save the sequence in

FASTA format to your desktop. Make a note here of the file name so you can find this wildtype sequence on your desktop when you need it.

OMIM:

Reminder: Read about all the databases in the Glossary before you get started. Also use the help menu and tutorials available on each of these sites as needed.

5. Go back to the NCBI homepage and select the OMIM database. Then search for "Lung Cancer". Read the first two paragraphs of the lung cancer entry.

6. Scroll down until you get to the list of references. The first article (Ahrendt et al.) should refer to K-Ras in the title. Click on the PMID# for the first article to bring you to the PubMed entry for this article. The abstract should be displayed. Read the abstract and answer these questions.

    a. What journal was this article published in? What year was this article published?

    b. Describe who was involved in the study (how many and what categories of patients)?

    c. What did the researchers find out about K-Ras mutations?

d. What conclusion(s) did the researchers come to about K-Ras mutations based on their data? (Summarize and put into your own words)

7. Go back to the OMIM entry for lung cancer. At the very top of this entry, the genes linked to this disease are listed. Click on the link listed after KRAS2. This will take you to the OMIM entry for KRAS2.

8. Read the first paragraph of the KRAS2 entry.
   a. What does this entry say about the role of normal Ras genes and mutant Ras genes?

9. Review the major sections in this entry. An outline for the entry is provided to the right of the window. Go to the Allelic Variants section. Scroll until you see the entry for the Gly12Cys mutation.
   a. How common was a mutation at position 12 in the Ahrendt et al. (2001) study?

KEGG pathway

10. Go back to the Gene entry (at the NCBI website) for human KRAS. Scroll down the links list at the right to find "KEGG" under the "Link to other resources" and click on KEGG. KEGG stands for Kyoto Encyclopedia of

Genes and Genomes. The www.kegg.com site contains a database of metabolic maps.

11. Click on “MAPK signaling pathway.”

12. You should see a nice graphic for the proteins in the signaling pathway. Answer these questions:
    - Several types of signalling molecules that lead to Ras activation contain “GF” in their names. What does GF stand for?

    - What protein(s) is/are directly activated by activated Ras?

    - How is the cell membrane shown? Which side are the GF’s on and which side is Ras on?

    - SRF and c-Fos are transcription factors. In general, what do they act on as a result of Ras activation?

    - If Ras were mutated to be always active, what part of the pathway becomes irrelevant?

Part 2

UniProtKB

This site will provide more information about the protein's structure and function.

9. Go to "ExPasY" from the course website or google "ExPasY."
10. Search the "UniProtKB" database for "kras2". Be sure to select the human protein from the search results.
11. Scroll through and review the entry then answer the following questions:

    a. How many splice variants are there of KRas2 and what are they called?

    b. What is the difference between GDP-bound KRas and GTP-bound KRas?

    c. Where in the cell would you go to find this protein?

    d. Which region is considered "hypervariable" (under "Sequence Annotation")?

    e. Which residue(s) bind GTP (under "Sequence Annotation")?

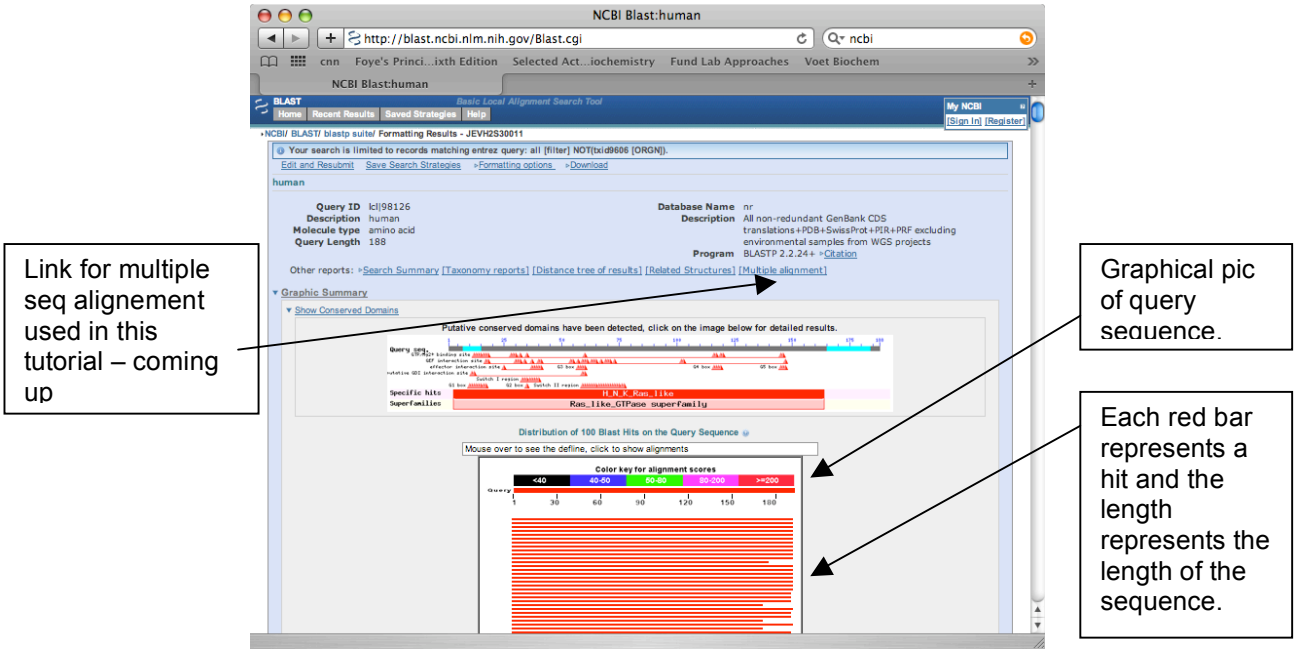
BLAST and Multiple Sequence Alignment

This site will search the protein sequence for your gene against a database of protein sequences from other organisms. Then you will choose several to compare in a multiple sequence alignment. The goal of this process is to compare the human sequence against orthologs. The parts of the sequence that has stayed the same over evolutionary time is likely to be the most important to the function of the protein. The results of this analysis give us more information about the regions of this protein that are most important to function.

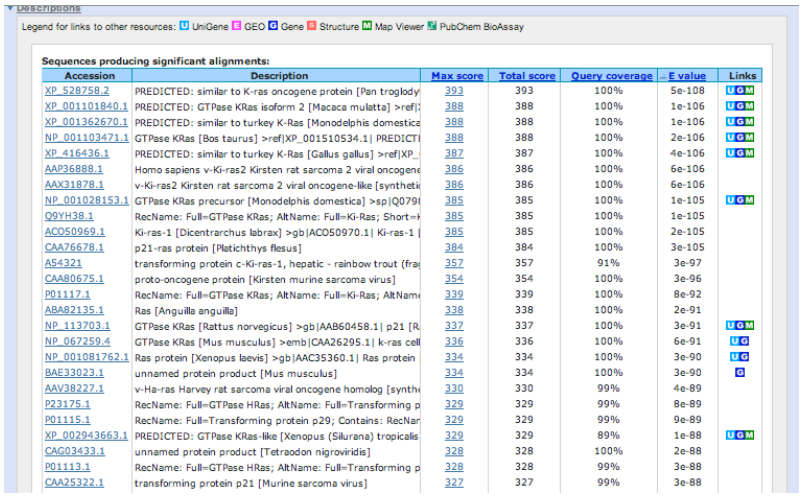
12. Perform a "Blast" search for KRAS2 using the RefSeq protein sequence by following the set of instructions below. You should already have the KRas2 RefSeq protein sequence file on your desktop. If not, go back to the section in this manual on NCBI Gene to see how to get it.
    a. Go to the "NCBI" home page and choose "BLAST" at the menu to the right for "Popular Resources."
    b. On the next page, Select "protein BLAST " (same as blastp).
    c. Paste the FASTA formatted RefSeq protein sequence in the search box.
    d. Select the "nr" protein database.
    e. Type "homo sapiens" and check the box next to "Exclude" to exclude any other human protein sequences from the search results.
    f. Click "BLAST" to begin.
    g. Several pages pop up to entertain you while you wait for the results. When the actual results page pops up, it will look somewhat like the images shown below.

SAVE the BLAST results to your desktop using "Save as" under "File." Save them in web archive format.

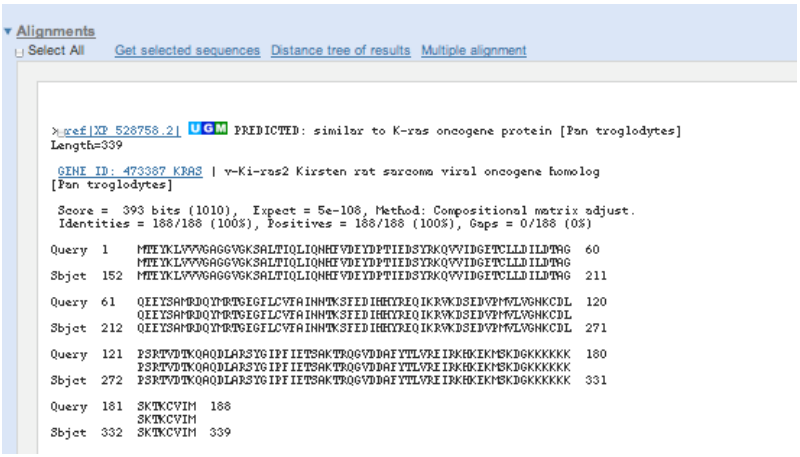
Blast Results Figure 1: The Blast results window looks like the image shown below.



Blast Results Figure 2: Scrolling down in the Blast results window shows you this:



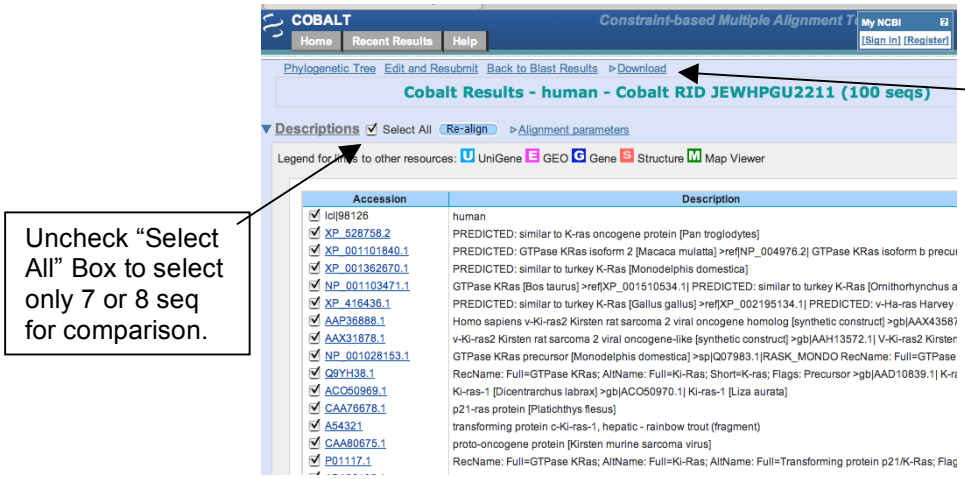
Blast Results Figure 3: Further scrolling shows you this:



The pairwise alignment of each hit with the query seq. Consensus seq is shown in the middle. In this example, the query and hit sequences are identical.

13. Choose the “multiple sequence alignment” option from the menu at the top that says, “Other reports.”

14. When the alignment results window opens, you should see something that looks like this:



15. Next, uncheck the box for “Select All” then go through and select seven or eight sequences to align. The goal in this process is to choose evolutionarily diverse organisms for the multiple sequence alignment. This type of alignment is useful because it allows us to see the parts of the sequence that has NOT changed, and therefore is likely to be important to the protein’s function. Make sure all the sequences you leave selected are from different organisms and try to select sequences that begin with NP_ since these are RefSeq sequences.

16. After you have selected all your sequences, click “Re-align” at the top of the window.

17. When the new alignment results window opens, you can scroll down to view the alignment.

18. Next download the alignment file by clicking “Download” near the top of the window. Choose “fasta plus gaps” as the format.

19. To edit the titles of each sequence in the alignment, open the alignment file you just saved in word. (Review the entry for FASTA formatting in the glossary if necessary.) Change the title line of each sequence to just contain the organism name as all one word. For example, the cow title would read “>Bostaurus”. Make sure there is a paragraph mark separating each title from the sequence.

EMBOSS Tools

This suite of tools will allow us easily format the alignment we just made to make it easier to view important features. This suite of tools is available for download to any computer as freeware. You can also access this suite through the website below:

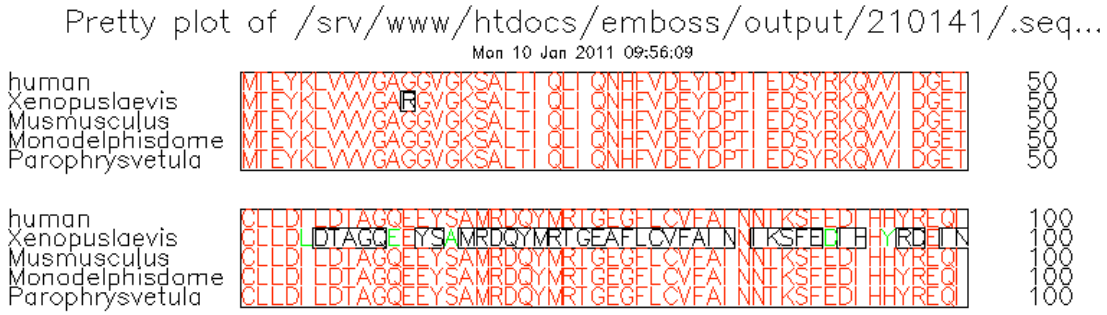
20. Go to http://gep.wustl.edu/
21. Go to “Projects” then “Annotation Resources” to select “ “EMBOSS Tools.”
22. When the EMBOSS Tools page opens, scroll down until you find “prettyplot” from the list at the left.
23. Open prettyplot and click “choose file.” Then select your editted alignment file.
24. Scroll past all the options to the bottom and click “Run prettyplot”. In the future, you can play with these ouput options before submitting to make your alignment even prettier.
25. Save the output alignment picture by right-clicking (PC) or control-click (Mac) and choosing “Save image as.”

26. Answer the following questions about your alignment:
    a. Which sequences seem the most diverged?

    b. What areas of the sequence are most highly diverged? Does this correspond with the “hypervariable” region mentioned in the SwissProt entry?

c. Which sequences show the active site residues mentioned in the SwissProt entry conserved with the human sequence?

The figure below shows a cropped image of the alignment. The titles were edited in the fasta sequence file (.fa) in Word before submitting to prettyplot.



Part 3

EMBOSS Tools

Translating the sequence

13. Obtain the mutant cDNA sequence from the course website (file name Krasmutseq.doc). Open the file and copy the sequence. Be sure to get the whole sequence. It could be a long sequence on more than one page.

14. Go to "sixpack" under the EMBOSS Tools, Nucleic Acid Translation at the left. This program will translate the nucleic acid sequence in every possible frame, including the reverse strand if you like.

15. Copy and paste the nucleic acid seqeunce in the search box. Answer "Yes" for "ORF start with an M?" and set the "minimum size of ORFs" to 50. Then hit "Run sixpack" at the bottom of the screen.

16. The longest amino acid sequence in your results is your mutant protein sequence in Fasta format. Copy the protein sequence and paste it into a word document. Save the document in your folder on the desktop. This is the **mutant** protein sequence.

Align mutant and wildtype KRas2 to identify the mutation

27. Go to "water" under "Alignment Local" in the EMBOSS Tools.

28. Paste the wildtype human sequence in Fasta format in the top search box and the mutant human sequence in Fasta format the bottom search box.

29. Click "run water" at the bottom of the screen. (Note: if you get an error message instead of alignment results, review the Fasta entry in the Glossary, check the Fasta formatting of your sequences, and remove any spaces or other formatting that may be present and try again.)

30. Scroll to see the alignment of these two sequences and answer the questions below.

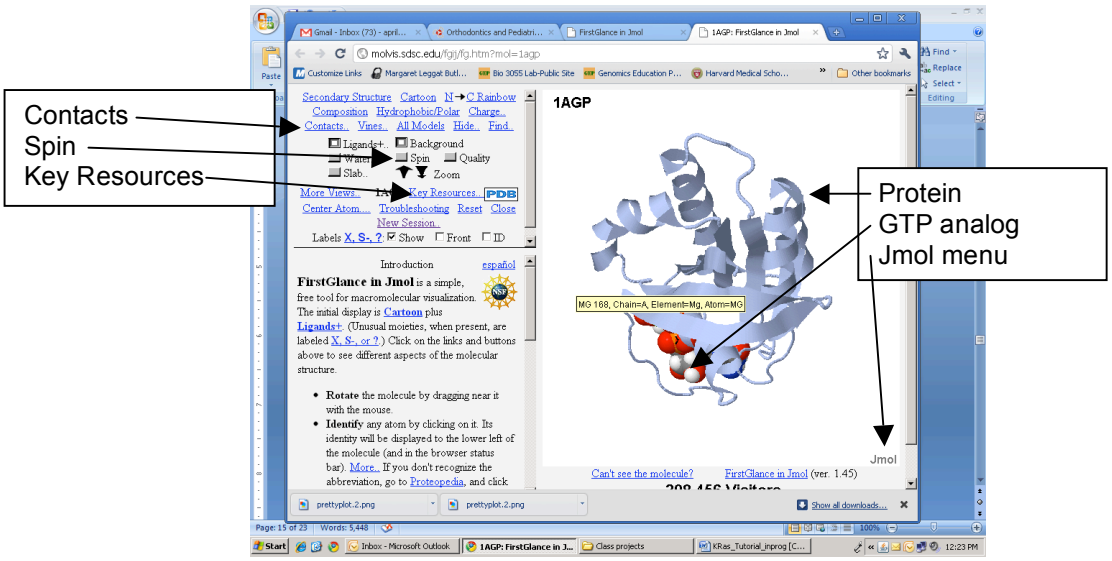
    a. What is/are the mutation(s)? Write each in the following format "Res123Res" where the first Res is the three-letter code for the amino acid in the un-mutated (wild type) protein and the second Res is the amino acid in the mutated protein. In place of "123" put the amino acid residue number of the mutation.

b. What does the UniProtKB entry (under “Sequence Annotation”) mention about the function of this region of the sequence?

Firstglance in JMol

For this project, you will analyze the crystal structure of a closely related protein, H-Ras. The crystal structure of H-Ras complexed with GTP bound (active conformation) has been solved. This is the structure you read about in the reading assignment due for this lab. The pdb file name is 1AGP.

1. Go to “Firstglance in Jmol” by searching in google.
2. Enter 1AGP in the box that asks for the PDB Identification code and hit “submit.”
3. A window like the one shown below should open. The links at the left will change the view in the molecule window. The molecule window shows the protein in ribbon form to show the alpha helices and beta sheets. Any bound small molecules or atoms will be in space-filling mode. In this case, the GTP analog is in space-filling mode and colored in CPK (in glossary).



4. First, to learn more about the protein and conditions in which it was crystallized, go to the PDB entry by following these steps. Click on the “Key Resources” and go to the PDB link.

5. Record the following information from the PDB database entry for 1AGP by looking through the different data tables there:
    a. Resolution (see Glossary)

    b. Organism of the protein

    c. Ligand(s) bound to the protein in the structure

IMPORTANT:

The 1AGP crystal structure was solved with GTP bound in the active site. The H-Ras in this crystal structure is a mutant protein, with Gly12 replaced with an aspartate residue. The GTP is labelled, "GNP 167" in the control panel.

Investigating the environment of the amino acid at position 12 (the mutated side chain)

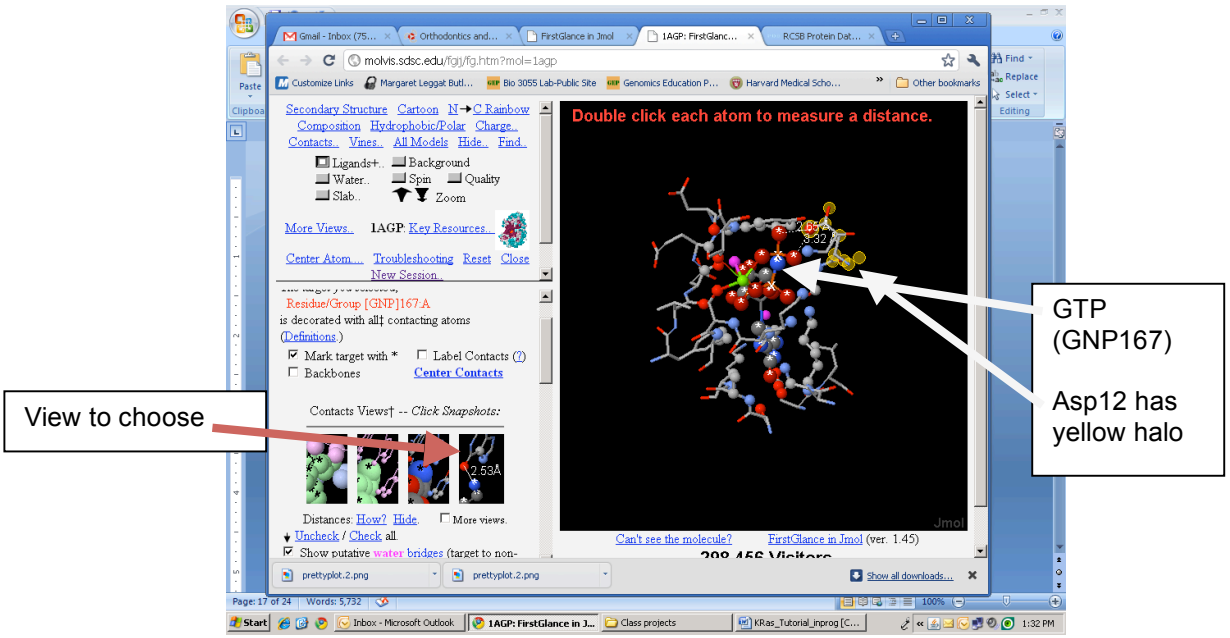
This project focuses on understanding the role of the amino acid at position 12 in order to propose a reason that mutation from a glycine (G) to a cysteine (C) at position 12 in the protein results in an oncogenic form of KRas2.

Firstglance in Jmol will allow us to view atoms and measure distances between atoms to allow prediction of non-covalent interactions that may be occurring. For example, an oxygen and a nitrogen that are less than 3 angstroms apart in a crystal structure are likely sharing a hydrogen in a hydrogen-bonding interaction. Since hydrogens do not have enough electron density to appear in a crystal structure, their locations must be estimated based on knowledge of amino acid side chain structure and pK_a data. Two carbon atoms that are within 3 angstroms in a crystal structure would be predicted to be in van der waal's contact with each other.

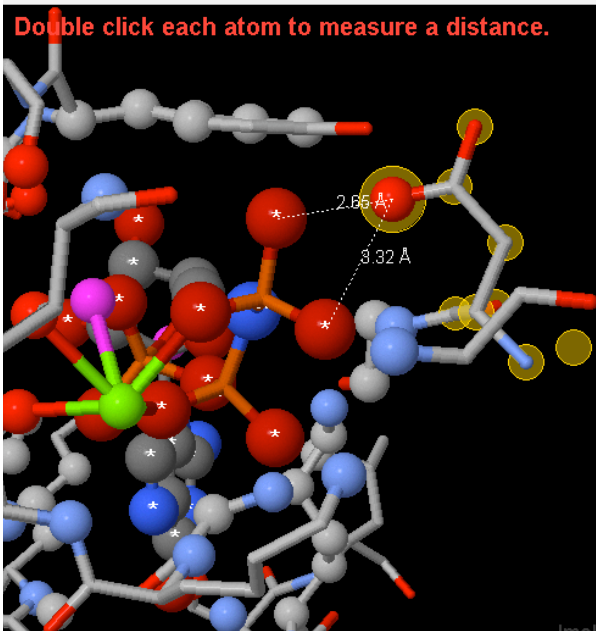
1. Click back to the Firstglance in Jmol window to view the molecule window.
2. Deselect "Spin" to stop the molecule from spinning. Now, practice moving the molecule using your mouse and zoom in on the molecule. Make sure you can do this before moving on. Use the h
3. Click on "Find" from the menu at the left. Type 12 in the search window that appears. The atoms in amino acid 12 should appear with yellow halos.
4. Now choose "Contacts" from the menu at left.

a. Where it says, “View: Spacefill or Cartoon”, choose cartoon.
b. The GTP analog (GNP 167) will appear in spacefill mode.
c. Choose “Residues/Groups.”
d. Click on any atom in the GNP 167 and all the atoms should have white stars.
e. Click on “Show atoms contacting target.”
f. Deselect “Backbones.”
g. Click on “Center contact”
h. Where there are four view choices shown, choose the view with cpk color and distances shown in white (the view on the right with red, blue, and gray colors shown).
i. Double-click on atoms to measure the distances between atoms.

The view at this point should look something like the one below, but may be in a different orientation:



5. Take a minute and find the terminal phosphate of the GTP substrate. This is the phosphate farthest from the guanine ring and zoom in to measure distances between this phosphate and Asp12. Double click on each atom and the distance in angstroms should appear. A sample view is shown below:



Note: To create a picture from Firstglance in JMol for a Word or Powerpoint document, simply use the screen capture function for your computer then the crop tool in Word or Powerpoint to focus on what you want to show. More information can be found by clicking on "Key Resources" in Firstglance in Jmol, then scroll to the bottom of the directions window and click on "How?" where it mentions copy and paste.

6. At this point, answer the following questions about the role of the amino acid at position 12:
   a. What atom is represented in green in the figure above?

   b. What atom is represented by pink?

   c. What are the non-covalent interactions between the protein and the terminal phosphate of the GTP analog (GNP167)?

   d. If the wildtype human KRas2 has a glycine at position 12, how would these non-covalent interactions be affected? Would you predict an increase in GTP binding or a decrease in GTP binding in the wildtype KRas2 compared with the protein shown here?

e. If there were a cysteine present at position 12, would this lead to an increase or a decrease in GTP binding from wildtype KRas2? Please explain by drawing a picture of the noncovalent interaction between the cysteine side chain and the terminal phosphate.

f. Would you predict that the amino acid at position 12 would have an effect on GDP binding? Why or why not?

g. Do you predict that the G12C mutation leads to a more active, less active, or unchanged form of Ras? Please explain your reasoning.

Project summary:

Please summarize your findings from this project in a one-page document. Summarize in your own words the background information on KRas and lung cancer. Identify the mutation and describe how this change in DNA sequence leads to an oncogenic form of KRas2.

Glossary

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

EMBOSS Tools – a suite of bioinformatics tools like multiple sequence alignment programs, translating programs, drawing circular DNA programs, open reading frame searches, etc. Freeware, but must be downloaded to your computer from the website below. Could also use the web-based EMBOSS tools at the gep site at http://gander.wustl.edu/cgi-bin/emboss

http://emboss.sourceforge.net/

ExPASy – Expert Protein Analysis System - A server maintained by the Swiss Institute of Bioinformatics. Home of UniProtKB, the most extensive and annotated protein database. The Swiss-Pdb Viewer (DeepView) protein structure-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*
MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGET
CLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQI
KRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIETSAKTR
QGVDDAFYTLVREIRKHKEKMSKDGKKKKKKSKTKCVIM

To learn more, go to:
http://www.ncbi.nlm.nih.gov/blast/fasta.shtml

Firstglance in Jmol – a web-based program for viewing 3-D structures. It loads ".pdb" files, which contain the 3-D coordinates for molecular structures. The same type of program as DeepView or Rasmol, but easier to use since no knowledge of command lines are needed and tutorials and help information pop up with each click. The Jmol menu is also available to use in the molecule window. If Jmol commands are known, they can be directly typed by opening the console window under the Jmol menu. This provides multiple ways to manipulate protein structure views.
http://molvis.sdsc.edu/fgij/
Note: To create a picture from Firstglance in JMol for a Word or Powerpoint document, simply use the screen capture function for your computer then the crop tool in Word or Powerpoint to focus on what you want to show. More information can be found by clicking on "Key Resources" in Firstglance in Jmol, then scroll to the bottom of the directions window and click on "How?" where it mentions copy and paste.

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

PDB – see Protein Data Bank

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.

To learn more about pdb files, watch the animation at the following link using Firefox browser:

http://www.fivth.com/fiVthSite/web-content/NewFiles/GrahamJcom/web-content/NewFiles/gjPortfComp/gjCBanim/7PDBanatGrahamGarland.mov

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.

Proteopedia – A searchable web-based database that houses 3-D pictures and descriptions of every known protein. It is also user-editable and functions in the same way Wikipedia does. Users can create sites that are public or private and could include tutorials for students and with Jmol windows to edit and view protein structures. Students can also create wiki reports that use 3-D structures as figures and easily integrate Jmol windows into their report.
www.proteopedia.org

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. This database contains abstracts to a large number of biomedical journal articles and often links to full text versions of the articles. Look for the "full text" icon to the bottom or right of the abstract. It is also possible to have PubMed only search for the free full text articles. 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.

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).

Resolution – This term is used to describe the quality of the data obtained from an X-ray crystallography experiment. You can think of it as how fuzzy the picture of the protein is. The lower the resolution, the clearer the picture. Resolution is given in units of Angstroms and a typical resolution is 2 angstroms. Structures at this resolution do not have strong enough data to predict all the hydrogen locations, but structures at 1.5 angstroms or lower can often resolve hydrogen atoms. To learn more about resolution and other terminology associated with pdb files and crystallography, go to the RCSB homepage. Scroll down to select "Understanding pdb Data" under the "Education" heading on the left of the page. Also see the "Protein Data Bank" entry in this Glossary.

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.