Bioinformatics Database Worksheet

(based on http://www.usm.maine.edu/~rhodes/Goodies/Matics.html)

Where are the opsin genes in the human genome?

Point your browser to the **NCBI Map Viewer** at **http://www.ncbi.nlm.nih.gov/mapview/**. You find a list of species for which genome information is available.

Find **Homo sapiens (human)**, and click on the Build 35.1. (We use the oldest build because sometimes not all links are hooked up to the tools for the newest one.) You see a diagram of the human chromosomes, and a search box at the top. Enter "opsin" in the box next to **Search for**. Click **Find**.

You see the diagram again, with red marks at your "hits", the locations of genes whose entries contain "opsin" as a whole or partial word. Below the diagram is a list of the indicated genes. Among them are the rhodopsin gene (RHO), and three cone pigments, short-, medium-, and long-wavelength sensitive opsins (for blue, green, and red light detection). Four hits look like visual pigments, which probably does not surprise you. To the left of each entry is the chromosome number, allowing you to tell which red mark corresponds to each entry. Note that several hits are on the X chromosome, one of the sex-determining chromosomes. You can pursue multiple hits on the same chromosome with the **all matches** link for that chromosome.

NOTE: In the human genome lists, you will often see duplicates marked "reference" or "Celera", referring to the results from two major efforts to sequence the human genome. At first, these two efforts were separate, but eventually they came together. When you have a choice, we will choose "reference."

Click **all matches** next to X. You see a very complicated display with the X chromosome on the left and red marks at the positions of the gene(s) you've followed to this page -- in our case, the two opsins, medium- and long-wave, which are located near the bottom tip of the X chromosome. To the right are various representations of the X chromosome, with listings of annotated areas. The two opsin genes are highlighted in pink.

For now, note the information provided for the first of the two highlighted opsin genes, OPN1LW (this is called the gene symbol). You see that this is the long-wavelength-sensitive (red) opsin, and that it's a gene involved in color blindness (a sex-linked trait -- no surprise).

What do scientists know about the opsins?

Click **OPN1LW**. You have entered **Entrez Gene**. Scan down the page. Some of the information is very plain and understandable, while some is very cryptic. One of the most accessible links is to **OMIM** (for Online Mendeliam Inheritance in Man), a catalog of human genes and genetic disorders. Look down the page and find **Phenotypes**, and notice the links marked **MIM**. These are links to **OMIM** entries. Click one of them.

Each **OMIM** entry tells you about this gene and types of colorblindness, genetic disorders associated with

mutations in this gene. Read as much as your interest dictates. Follow links to other information. For more information about **OMIM** itself, click the **OMIM** logo at the top of the page. Once you've satisfied your appetite, return to the **Entrez Gene** page (use the **Back** button of your browser or your browser's history list).

Next to the **Display** button at the top, pull down the menu and select **PubMed Links** (near the bottom of the menu). **PubMed** is a free database of scientific literature, and this page shows a list of articles directly associated with this gene locus. By clicking on the authors of each article, you can see abstracts of the article. **PubMed** is your entry point to a wide variety of scientific literature in the life sciences. Use the Find tool of your browser to find the name *Nathans* on this page. Read the abstract of the article by Nathans and co-workers before returning to Entrez Gene.

What is the nucleotide sequence of this gene?

We are looking at the gene for the red-sensitive opsin in human vision, and it is located near the bottom tip of the X chromosome. Back in Entrez Gene, scroll down to **NCBI Reference Seqences (RefSeq)**. In the first section, **mRNA and Protein(s)**, all of the following are available:

- the protein sequence (sequence of this gene's protein product, the red opsin), here listed as **NP_064445.1** (P for protein);
- the mRNA Sequence (sequence of nucleotide bases in the messenger RNA), here listed as **NM_020061.4** (M for mRNA);
- the source sequences (entire sequence(s) of the genome fragment(s) containing this sequence, from **GenBank**).
- the concensus coding sequences (CDS) from **CCDS (CCDS14742.1)** (The CCDS project is a collaborative effort to identify a core set of protein coding regions that are consistently annotated and of high quality.)
- the corresponding links to **UniProt (P04000)**
- the conserved domains in this sequence **(cl09500)**

Note that the two links to mRNA sequence and protein sequence are given as NM_020061.4→ NP_064445.1, the arrow implying that the sequence of the NM entry is translated (by protein synthesis) to give the sequence of the NP entry. Click the entry number for the mRNA sequence: **NM_020061.4**

This is a typical **GenBank** nucleotide file, and a lot of it is hard to read, but a few things are clear. First note, under references, there are citations to the publication of this sequence in the scientific literature. To see an abstract of the article in which this gene was described, click the **PubMed** link (a number) below the first reference and read it. Or instead, find the word *Nathans* on the page, and click the **PubMed** link below the related article. As you see, you've been here before. There are many ways to move from one database to another.

Scroll to the bottom of the GenBank page. The last thing, labeled **ORIGIN**, is the sequence of this messenger RNA. You are seeing the actual list of As, Ts, Gs, and Cs that make up the message for synthesis of this opsin. But wait! You know that RNA contains no T. In most nucleotide databases, U from RNA is represented as T, to make for easy comparison of DNA and RNA sequences. This sequence information is not in the form that is most useful for searching in databases, say, searching for related genes. Let's display this entry in a form more useful for searching.

At the top of the page, beside the **Display** button, pull down the menu that says **GenBank** (the default display format for each entry), and select **FASTA** (note that several other display options are available). Now you will see the sequence in FASTA format. You can save this file by either selecting **File** in the pull down menu labeled **Send to** or by selecting **Text** and saving the displayed result to a file called *mrnared.txt* using your browser's "Save to file" option. In the former case, the file will be automatically called *sequences.fasta* so let's rename the saved file to *mrnared.txt*.

Click your browser's Back button until you return to the Entrez Gene page for this gene.

What is the amino acid sequence of this gene?

Under **NCBI Reference Sequences (RefSeq)**, click the entry number NP_064445.1 for the protein sequence.

Things look a lot like before, but this is a protein entry (the classical view is that gene products are proteins, but not all of them are), containing the amino-acid sequence in one-letter abbreviations. Just as with the mRNA entry, turn this into a FASTA display, and copy it into a new text file called **protred.txt**. Return to **Entrez Gene**.

What proteins in human are similar to the red opsin?

Now return to the **NCBI Map Viewer** at **http://www.ncbi.nlm.nih.gov/mapview/**. We will search the human genome for sequences similar to the red opsin.

Click the **Blast** symbol (circled **B**) next to **Homo sapiens (human)** Build 35.1. This is the NCBI's BLAST search tool. BLAST is a widely used program for finding sequences similar to a "query" sequence that you're interest in. Pick these options from the various menus:

- Database: Build Protein for previous build 35.1. (This means that you will search the protein sequences in this build of the database.)
- Program: BLASTP (Use the version of BLAST that compares protein sequences, unlike BLASTN, which compares nucleotide sequences.)
- Other Parameters, Expect: 10 (The higher the number, the less stringent the matching, and the more hits you'll get)

Next, enter the query sequence. This can be done in one of two ways: (1) copy the FASTA data from your file **protred.txt** to your clipboard, and paste it into the BLAST search box under "Enter an accession..." Check to be sure that the first character in the box is the ">" at the beginning of the FASTA data. (2) Check the button for "choose a file to upload" and specify the file **protred.txt**.

Then click **Begin Search**.

The next page is for formatting your search results. We will take all defaults, and just click the **View Report** button. When your results are ready, the **results of BLAST** page appears. Look down the page to the graphical display, a box containing lots of colored lines. Each line represents a hit from your blast search. If you pass your mouse cursor over a red line, the narrow box just above the box gives a brief description of the hit. You'll find that the first hit is your red opsin. That's encouraging, because the best match should be to the query sequence itself, and you got this sequence from that gene entry. The second hit is the green opsin -- remember that the PubMed entry reported that the red and green pigments are the most similar. The third and fourth hits are the blue opsin and the rod-cell pigment rhodopsin. Other hits have lower numbers of matching residues, and are color coded according to a score of matches. If you click on any of the colored lines, you'll skip down to more information about that hit, and you can see how much similarity each one has to the red opsin, your original query sequence. As you go down the list, each succeeding sequence has less in common with red opsin. Each sequence is shown in comparison with red opsin in what is called a **pairwise sequence alignment**. Later, you'll make **multiple sequence alignments** from which you can discern relationships among genes.

See what you can figure out about what the scores mean. Identities are residues that are identical in the hit and the query (red opsin), when the two are optimally aligned. Positives are residues that are very similar to each other (see residue number 1 in the blue opsin -- it's threonine in red opsin, and the very similar serine in the blue). Gaps are sometimes introduced into a hit to improve its alignment with the query. The more identities and positives, and the fewer gaps, the higher the score. Note that blue opsin and rhodopsin are only about 45% identical to the red opsin. Other proteins, which are apparently not visual pigments, have even lower scores. Now let's take a look at where all these hits are in the human genome.

Where are all the genes for these other proteins?

Click the **Human genome view** button near just below the introductory information at the top of this result page.

You have come full circle. You are back that the human chromosome diagram, and all the hits of your search, in the colors that signify their BLAST scores, are located for you on the diagram. Notice that there are about 100 proteins (discovered so far, that is) that have 40% or more positives in alignment with red opsin. The opsins are members of the very large family of G protein-coupled receptors, key players in signal transduction.

How are the opsin genes related to each other?

Answering this question requires making a **multiple sequence alignment** and then using it to make a **phylogenetic tree**. For these tasks, we move to another database where it's a little easier to gather a bunch of sequences into a single FASTA file. Point your browser to the home page of **UniProtKB**, at **http://beta.uniprot.org/**. Set up a search for human opsins, as follows:

- Search **UniProtKB**.
- Enter query: opsin
- Click the "Fields >>" link, and choose "Organism" for the Field
- Term: Enter "Human." A list of matching terms will appear, from which Human can be selected.

Click **Add & Search**. Look over the results. There should be 24 hits, including the rod pigment rhodopsin (**OPSD**), along with the three cone pigments (**OPSB, OPSG, OPSR**). There is also a "visual pigment-like receptor peropsin", **OPSX**. Sound mysterious. Let's find out more about it, and in the process, see a typical UniProt entry. Next to **OPSX_Human**, click on the number (014718) in the column headed Accession. You see the **UniProtKB/Swiss-Prot** View of entry O14718. Peruse this entry and try to find out just what this rhodopsin-like protein is thought to do. Under **General Annotation (Comments)**, you'll learn that it's found in the retina (the RPE or retinal pigment epithelium), and that it may detect light, or perhaps monitors levels of retinoids, the general class of compounds that are the actual light absorbers in opsins. Also under **Comments – Sequence Similarities**, you see, as mentioned earlier, that this protein is a member of the large family of G protein-coupled receptors (GPCRs). If you click on either of the links for **G-protein coupled receptor 1 family** or **Opsin subfamily**, you find a list of all purported members of this subfamily in UniProt.

Now back up to the **UniProtKB/Swiss-Prot** entry page for 014718, **OPSX_HUMAN**. Under **References** click the journal citation, "Proc. Natl. Acad. Sci. U.S.A. 94:9893-9898(1997). From the resulting page, you can read a full article in the **Journal of the National Academy of Sciences (PNAS)** about this protein. Like many journals, PNAS puts full articles online just 6 to 12 months after publication.

Return from the PNAS reference, and look further down the entry page, where you find cross-references to this protein or its gene in other databases, predicted structural features of the protein, the sequence. Now let's compare the sequences with each other. We'll use the program **ClustalW** to make a multiple sequence alignment. Return to the search result page, for human opsin containing 24 hits. Scroll down the result page and check the boxes at the left of these entries

- **OPSB** (blue-sensitive opsin)
- **OPSD** (rhodopsin)
- **OPSG** (green-sensitive opsin)
- **OPSR** (red-sensitive opsin)
- **OPSX** (visual pigment-like receptor opsin)

At the bottom of the page, a green area indicating the selected genes should be displayed. To the right, there are four buttons to perform various actions on these genes. Select **Align** to perform a multiple sequence alignment using **ClustalW**.

You see the typical **ClustalW** alignment, showing our five protein sequences aligned to maximize identical and similar residues. Below each line of five sequences are symbols to show the extent of similarity among the sequences. An asterisk (*) means that the same residue is always (that is, for all of these sequences) found at that location; for example, the first asterisk marks a location where only N (asparagine) is found. Colon (:) means that all residues at this location are very similar; for example, the first colon is where only M (methionine), and L (leucine) -- residues with large, nonpolar sidechains -- occur. Period (.) means somewhat similar residues; for example, at the first period, serine, threonine, and asparagine occur -- all polar, but varied in size. If there is no mark then the residues at that location display no predominant common properties.

The same alignment can be viewed with **KEGG**. Copy the FASTA format at the top of this page, and open the KEGG multiple alignment tool at **http://align.genome.jp/**. Paste the sequences into the text box. Let's take this opportunity to modify the sequence names by replacing the accession numbers with the names of the genes, as follows.

- **P03999 ->** blue-sensitive opsin
- **P08100 ->** rhodopsin
- **P04001 ->** green-sensitive opsin
- **P04000 ->** red-sensitive opsin
- **O14718** -> visual pigment-like receptor opsin

Click the **Execute Multiple Alignment** button. This will display a less-graphical view of the same alignment. At the bottom of this page, a pull down menu for building a phylogenetic tree is available. Select any option that includes branch length and click the **Exec** button. Again, we should see that the most similar, and most recently diverged, proteins are the red- and green-sensitive opsins.

What is the structure of an opsin?

By now, I'm particularly curious about peropsin, but it's not likely that the structure of a recently discovered protein of unknown function has been determined. But it is likely that all opsins are similar in structure, so let's see is we can find an opsin in the database for macromolecular structures, the **Protein Data Bank (PDB)**. It will give us an idea of what kind of thing an opsin is.

Point your browser to **http://www.rcsb.org/pdb/**.

The PDB home page contains a simple search box at the top. You can search for models using simple keywords or PDB ID codes. An PDB code has four characters, like 1CYO. How would you ever know a model by its code? When a new structure is published, the authors usually give the PDB code in the last

reference of the bibiography. With that code, you can go straight to the model you want to see. But more often, your question, like ours, is more general. For such cases, PDB also provides forms for more sophisticated searches. For now, let's just see if any opsin models are available. Type "opsin" into the search box, make sure the **PDB ID or keyword** is selected, and click **Search**.

This search returns no models, which is quite puzzling, because a search for "rhodopsin" returns 64 models. So this tells me that the quicky (quirky?) search tool at the PDB still needs some work. So let's first try by searching for similar entries by sequence. In PDB terminology, the red opsin sequence is the **query**, and similar models found are called **subjects**.

First, open your query file **protred.txt** (FASTA sequence of red human opsin), and copy the sequence portion only to the clipboard; omit the comment line that begins with >. At the top left of any PDB page, click **Search**. From the list of search types, click **Sequence**. On the resulting page, paste your red opsin sequence into the box next to "Sequence." Note that the search tool is your new friend Blast, and that a E cut-off value of 10 is given as a default. This will limit the results to subjects of quite high similarity. Higher E-value cutoffs will be less restrictive, and give more hits or subjects. Click the **Evaluate Query** button near the bottom right. The search tool is now looking for PDB models whose sequences are similar to the human red opsin sequence.

21 PDB models whose sequences are homologous to the search sequence should appear. Each is listed with an E-value (not the same as the E cut-off value on the previous page), which is the probability that the sequence similarity between query and subject is a coincidence. The first result or subject is PDB model **2j4y**, a model of mammalian mutant rhodopsin. The E-value is 2.4 x 10 -76 . In other words, while the probability that a coin flip and your call will agree just by chance is 0.5, the probability that the similarity between human red opsin and bovine rhodopsin is just a chance occurence is

0.00000000000000000000000000000000000000000000000000000000000000000000000000000024,

which means, to any sane biologist, that these two molecules descended from a common ancestor. There is no chance that, in the history of the universe, two proteins could arrive at sequences this similar by chance. This also means that the structure of the rhodopsin is a sure bet to be very similar to that of the human red opsin, whose structure is unknown (if it were known, this search would have found it).

So let's click on the top result, **2j4y**. You have come to the **Structure Summary** page for this model, which is its home page at the PDB. This page is connected to just about everything you could possible do with this model. At the PDB, your first goal is always to get to the Structure Summary page for the model you are seeking.

NOTE: **Structure Summary** does not exactly jump out at you on this page. It's the tab selected over the main part of the entry. Those tabs should be more prominent—they are what distinguishes each of the important pages for a model. If you want to know where you are in the PDB, look at two sets of tabs at the

top of the page. The other tabs open LOTS of information about this model, but we will stick with structure.

In the left column of all PDB pages, you find a set of nested menus. Click **Display Molecule** to open the PDB display options. Click **Jmol Viewer**. Assuming that your computer has up-to-date Java software, your browser will load the viewer, and it will load the file 1GZM. Your should see models of two rhodopsin molecules—with backbones shown as ribbon-like cartoons, one green, one blue—and several ball-and-stick models of smaller molecules. Is rhodopsin a dimer? No, but the crystals of rhodopsin from which this model was derived contained two rhodopsin molecules per asymmetric unit (the smallest portion from which the entire crystal is constructed). PDB files usually show the full contents of the asymmetric unit. If more than one molecule is present, they are referred to as chains in the model.

NOTE ON VIEWERS: The viewer on display is the widely used Jmol, which you will find in use as a molecular viewer at many web sites. If you take time to get to know this viewer fairly well, you will get more out of the many sites that use it. Like most of the other viewers listed at PDB, Jmol is quite limited in its capacity for analysis of protein structure.

Here are some other things you can do to get to know models in a Jmol frame (to get back to the original rendition, reload the page):

- Click/drag (left button if you have more than one) on the image to rotate the structure. You should be able to tell that is has a lot of alpha helix.
- Hold down **alt** (for Windows; **option** for Macintosh) and click/drag to zoom in (drag towards you) or out (drag away) or the rotate the model in the plane of the screen (drag left or right).
- Right-click (or hold down **ctrl**) the image: up pops a set of menus, and if you browse around on them, you'll see that there is much more to Jmol. Let's try just a couple of things to give you some general ideas.
- Using the pop-up menus, **Select:Protein:All**
- Nothing appears to happen. You have selected part of the model (the protein part, but not the small molecules). Now let's change it.
- **Color:Cartoon:By Scheme:Secondary Structure**
- The cartoons become red (well, bright pink) for alpha helix, and yellow for beta sheet. I'll bet you had not noticed the beta sheet in the models before. Look one of the chains over carefully to get a feeling for its structure. How many helices are present? How many strands of beta sheet? Are the strands parallel or antiparallel?
- **Select:Protein:All** (means select both backbone and sidechains). Then **Render:Scheme:CPK Spacefill**. The protein portion is now shown as a space-filled model. In this rendition, you get a good idea of the overall shape of the protein.
- **Render:Scheme:Wireframe**. Now you see all of the protein parts of this model in wire-frame. This is not as impressive as some other schemes, but is actually the most useful when you start exploring models in detail, because the wires do not hide each other like ball and sticks or space-filled models.

To learn more about Jmol, consult the help links at PDB below the display. You can also find extensive help for all viewers listed there.

Now the browser's Back button to return to the results page. Try to answer these questions about the comparison between human red opsin and the bovine rhodopsin in PDB **1F88**:

1. How many corresponding residues, and what percent of the residues, do the two proteins have in common (exact matches)?

2. How many and what percent of corresponding residues are similar in chemical properties?

3. How many gaps did the alignment program introduce, and how many residues in each gap, to get best alignment between human red opsin and 1F88?

4. Find the longest string of exact matches between the two proteins. How many matches does it contain, and what are the beginning and ending residue numbers?

Now you know how to search the PDB for models whose sequences are similar to a target or query sequence. Structural biologists use such searches when they have a new protein sequence and want to know its structure. If the structure is known, this search would find it. If not, any hits with high sequence similarity can tell researchers the overall fold of the newly discovered protein.

Questions

1. Start at the NCBI Map Viewer. How many genes in the human genome contain the term "homeo" in their name? To be sure you find them all, search for "*homeo*". Number found: _____ .
2. Which chromosome contains the largest number of these genes? How many? Chromosome # _____ ; Number of "homeo" genes on this chromosome: _____ .
3. Among the genes found in question 1, find one that has a role in insulin action.
   Name of the gene: _____. Four-character ID: _____ .
4. What chromosome contains this gene? Chromosome # _____ .
5. According to OMIM, what is the role of the protein encoded by this gene?
   Role (limit to 25 words): _____ .
6. Obtain the protein sequence of this gene, in FASTA format. File name: HmPrt.txt
7. Go to UniProt. How many annotated human genes in UniProt contain the term "homeo"? Note that "*" is automatically used as prefix and suffix unless you specify otherwise. Number found: _____ .
8. Make a phylogenetic tree of the first 24 of these genes plus the insulin-related gene found in question 3, a total of 25 sequences. Use the insulin-related gene as the "outgroup".
9. According to your tree, what two entries in this group are the most similar?
   Entry numbers _____ and _____ .
10. What entry is most similar to the insulin-related gene? SwissProt entry number _____ .
11. What can you find out about the function of this similar gene?
12. Go to the Protein Data Bank. Search for models of human homeodomain proteins.
    1. How many models do you find?
       Number: _____ .
    2. What method of structure determination produced the first of these models?
       PDB ID code: _____ . Method: _____ .
    3. View the first model on the list with QuickPDB or your favorite molecular viewer.
       What are the main secondary structural elements (helix, sheet, coil) in this protein?
       Secondary structural elements: _____ .
    4. Give beginning and ending residue numbers of three secondary structural elements.
       1. start: _____ end: _____ .
       2. start: _____ end: _____ .
       3. start: _____ end: _____ .
13. Find a model of a human homeodomain/DNA complex.
    1. How many models do you find? Number: _____ .
    2. What method of structure determination produced the first of these models?
       PDB ID code: _____ . Method: _____ .
    3. View the first model on the list with QuickPDB or your favorite molecular viewer.
       What secondary structural element(s) (helix, sheet, coil) interact with DNA?
       Secondary structural elements: _____ .
    4. Give beginning and ending residues of main secondary structural element in contact with DNA.
       Residue start: _____ end: _____ .