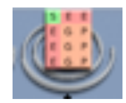
Hands-On UTOPIA

The purpose of this tutorial is to illustrate some of the issues raised during the presentation and to provide first-hand experience of using standard analysis tools in (what I hope is) an intuitive, easy-to-use software environment – UTOPIA. To this end, you will need to have UTOPIA installed on your desktop or laptop (if the software is not already installed on your machine, it is available for download at **http://utopia.cs.man.ac.uk/**).

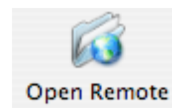
Once you have UTOPIA installed, you can begin. Broadly, the tasks are to: 1) locate a target sequence whose structural and functional characteristics we wish to explore (here, you may use your own protein, or simply follow the tutorial exemplars); 2) search a sequence database for homologues; 3) align a subset of those sequences; 4) select a number of conserved motifs – *e.g.*, those regions you suspect might be transmembrane (TM) domains; 5) computationally predict sequence properties, such as the locations of TM domains; 6) visualise and compare the locations of these motifs and domains in 3 dimensions and; 7) gather general functional annotation about this protein family.

1. Locate a target sequence

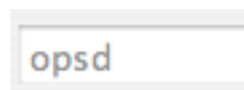
i) Click on the **CINEMA** icon: this may appear either on your desktop, on your toolbar or in your Applications menu;



ii) Click on the **Open Remote** icon at the top of the CINEMA window – this invokes the **Find-O-Matic** tool;



iii) We now want to locate a candidate sequence. Here, you may either search for your own protein or interest, of follow the tutorial using rhodopsin as an example. At this point, we could perform a free-text search using 'rho-dopsin' as a keyword, but it is quicker to locate known database identifiers (IDs). Fortunately, we know that rhodopsin IDs in UniProt:Swiss-Prot take the form opsd_X (X being an acronym for the species). So, in the **Find-O-Matic** search box, simply type **opsd** and hit **Return**;



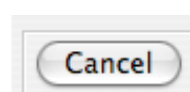
iv) Within the returned hit-list, we first want to locate the sequence and structure of bovine rhodopsin. The PDB ID is 1f88. Locate this in the **Find-o-Matic** output (you can use the **filter** option should you have problems locating 1f88). Click on **1f88** to select it and **drag-and-drop** it into the CINEMA window (you will be asked which chain to use – chain A will be highlighted, so just click on **OK**). This will invoke an alignment-editor window containing the sequence of 1f88.

2. Search a sequence database

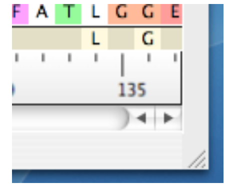
i) We now want to isolate a set of proteins whose sequences are similar to bovine rhodopsin. Locate **opsd_bovin** in the hit-list, click on it to select it, and invoke the context-sensitive menu to obtain a list of Web services available for this sequence. You should find that only a **BLAST service** is available. Click on this option – bear in mind that BLAST may take a while to run;

ii) The list of returned sequences is different from the previous list, and we want to see what some of the matches look like. From the list of matches, click on the top hit, press the shift key, then click on another sequence further down the list to select all those in between; **drag-and-drop** the selected set into **the editor window** (which should already contain 1f88) – this fetches them one-by-one from UniProt, which may take a few moments;

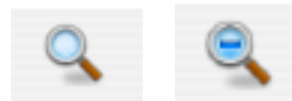
iii) **Note:** only if your previous search didn't retrieve 1f88, return to **Find-O-Matic**, type **1f88** into the search box and hit **Return** – once the match has been retrieved, stop the search using the **Cancel** button at the bottom of the window; select **1f88** from the hit-list by clicking on it, then **drag-and-drop** it into **the editor window** – this will fetch it from the PDB (you will be asked which chain to use – chain A will be highlighted, so just click on **OK**).



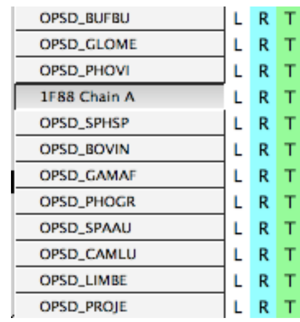
3. Align a subset of sequences



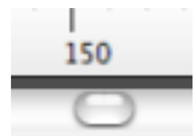
i) To increase the width of the editor window, grab the **bottom right-hand corner** and drag it out to a more convenient size;



ii) Now use the **free zoom** or **zoom out** icons on **the editor's tool bar** to see more of the sequences – you may also need to re-size the window again, as above;



iii) You may also wish to rearrange the order of sequences in which they appear in **the editor window**. To move **any sequence** within the alignment, click on the sequence to select it; in the **left-hand portion** of the window, **drag** the selected sequence to the desired position within the **list of identifiers** and release;

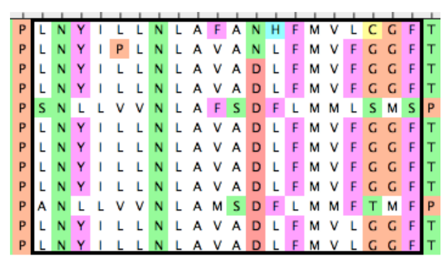


iv) Inspect the sequences using the **scroll bar** at the base of the **editor window** to navigate to & fro, noting that 1f88 (and possibly other sequences, depending on which you chose) need to be aligned;

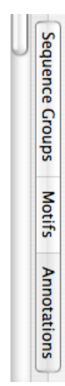
v) To align the sequences, **right click** on any sequence, select the **Alignment** option from the context-sensitive menu, and choose **ClustalW** or **Muscle** – this will send all the sequences to the chosen Web service and will return a completed alignment within seconds.

4. Select a number of conserved motifs

i) We now want to select a number of motifs corresponding to regions we presume to be TM domains; to do this, click on the **Create/edit motifs** icon in the **editor's tool bar**;



ii) Scroll to a region of interest (nb, likely TM regions will contain many white and purple residues), click on the top left-hand residue, drag out a box to the bottom right-hand residue and release – this highlights a 'motif'; any number of motifs can be selected in this way, scrolling through the alignment, and dragging out boxes over conserved white/purple regions;

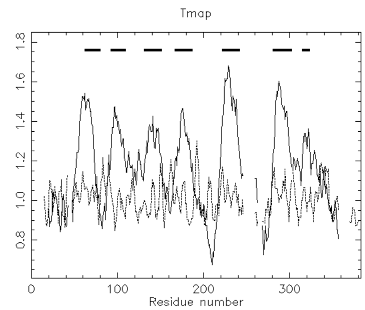


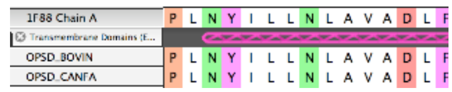
iii) Each selected motif will be listed in **CINEMA's manager window** under the **Motifs** tab – the **Motif manager** allows various manipulations to be performed on the motifs, including deleting any that you don't want, saving them, printing them, and grouping them as a fingerprint (to group motifs as a fingerprint, simply select those you wish to group, invoke the context-sensitive menu, and choose the **Group selected motifs as a fingerprint** option);

iv) once a number of such motifs have been selected, we will use them to compare both with the regions predicted by available computational tools and with the known regions from the PDB structure itself. First, let's look at the online prediction tools.

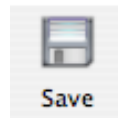
5. Computationally predict likely TM domains

i) Within the alignment, we now want to show where the TM domains are *predicted* to be. To do this, click on **1f88** to select it, invoke the context-sensitive menu, select the **Sequence 1f88** option and choose **Annotate TM domains '** (use either the **tmap** or **TMHMM** option) – **Tmap** produces a graph showing the predicted TM regions, and simultaneously highlights their locations within the sequence by means of a coloured bar beneath the relevant regions in the alignment;



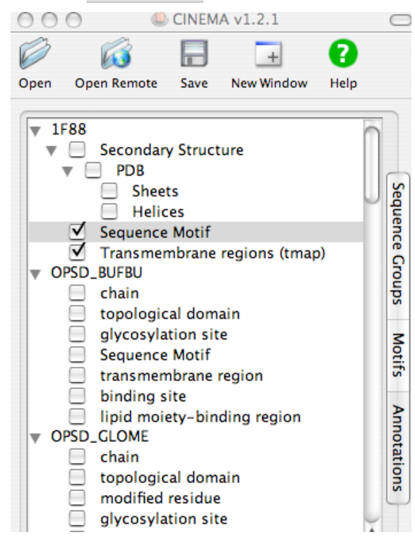


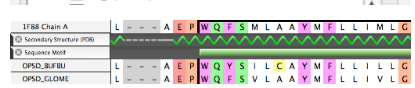
ii) Before continuing further, to safe-guard your work, it is helpful to save the alignment. To do this, click on the **Save** icon in **CINEMA's manager window**, and provide a suitable file name at the prompt.



6. Visualise the 3D structure & motif locations

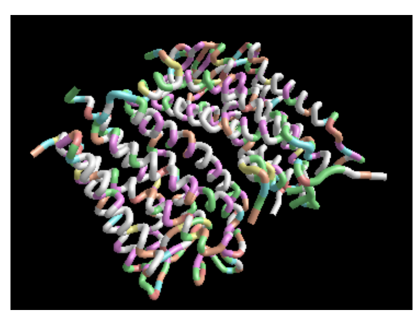
i) We will now compare the motifs you selected in step 4 with other annotated features. To do this, choose the **Annotations** tab in **CINEMA's manager window** – this lists the range of annotations available to display for each sequence in the alignment; for **1f88**, click on **Sequence Motif** – in **the editor window**, a **Sequence Motif annotation bar** will appear beneath **1f88**. Now compare your motifs with the locations of the helices defined in the 3D structure by clicking on **1f88's PDB helices** option in the **Annotations manager window** – within **the editor window**, this will invoke a **Secondary Structure (PDB) annotation bar** beneath **1f88**. Similarly, you can compare these annotations with those, say, for **opsd_bovin** – in the **Annotations manager**, scroll down to **opsd_bovin** and, from the display options, under **Secondary structures, UniProt** click on the **Helices** option; further up the list, click on the **transmembrane region** option – compare these with those already displayed for 1f88;



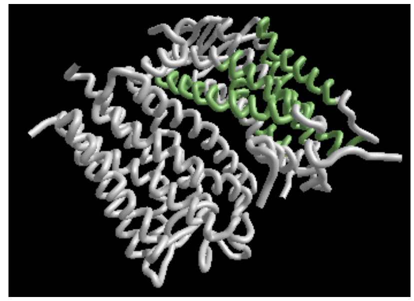


ii) It is now useful to visualise these annotations in 3D; to do this, invoke **Ambrosia** by clicking on its icon, either on your desktop, in your toolbar or on your Applications menu;

iii) To display the structure of 1f88, click on 1f88 in **the editor window**, invoke the context-sensitive menu, and choose **Show structure of 1f88.** This will display the 3D structure using default parameters – *i.e.*, a backbone coloured according to residue properties. Experiment with options to interact with the structure (*e.g.*, zoom in and out, rotate the molecule, change the default display via the context-sensitive menu);

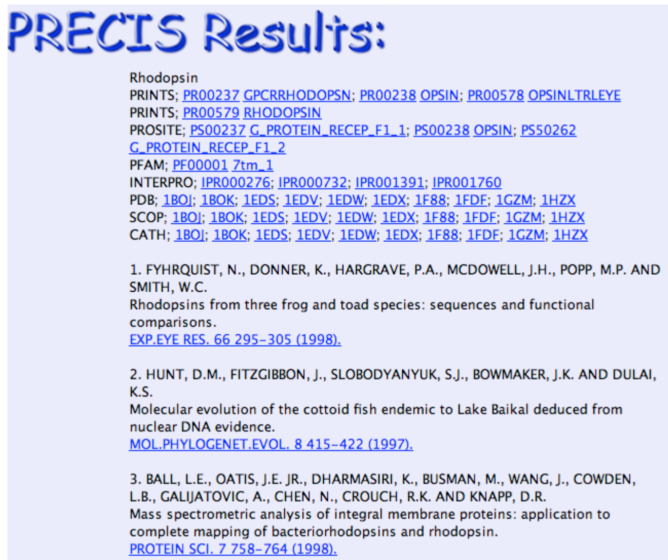


iv) To visualise the 3D locations of the motifs you selected earlier, click on the **Sequence Motif annotation bar** in **the editor window** – this will colour the motifs in **Ambrosia** so that you can see their 3D positions instantly; now click on the **tmap annotation bar** in **the editor window** and compare the locations of the predicted regions with the actual positions of the helices in **Ambrosia**;



7. Gather functional annotation

i) Finally, it is useful to learn more about this group of proteins in terms of their functions, disease associations, family relationships, and so on. To do this, click on any sequence in **the editor window** – invoke the context-sensitive menu, select the **Alignment** option and choose **PRECIS protein report** to get an overview of the function, structure and disease associations of your selected sequences; explore the hyperlinked options within the protein family report.



PRECIS Results:

Rhodopsin
PRINTS; PR00237 GPCRRHODOPSN; PR00238 OPSIN; PR00578 OPSINLTRLEYE
PRINTS; PR00579 RHODOPSIN
PROSITE; PS00237 G_PROTEIN_RECEP_F1_1; PS00238 OPSIN; PS50262 G_PROTEIN_RECEP_F1_2
PFAM; PF00001 7tm_1
INTERPRO; IPR000276; IPR000732; IPR001391; IPR001760
PDB; 1BOJ; 1BOK; 1EDS; 1EDV; 1EDW; 1EDX; 1F88; 1FDF; 1GZM; 1HZX
SCOP; 1BOJ; 1BOK; 1EDS; 1EDV; 1EDW; 1EDX; 1F88; 1FDF; 1GZM; 1HZX
CATH; 1BOJ; 1BOK; 1EDS; 1EDV; 1EDW; 1EDX; 1F88; 1FDF; 1GZM; 1HZX

1. FYHRQUIST, N., DONNER, K., HARGRAVE, P.A., MCDOWELL, J.H., POPP, M.P. AND SMITH, W.C.
Rhodopsins from three frog and toad species: sequences and functional comparisons.
EXP.EYE RES. 66 295-305 (1998).

2. HUNT, D.M., FITZGIBBON, J., SLOBODYANYUK, S.J., BOWMAKER, J.K. AND DULAI, K.S.
Molecular evolution of the cottoid fish endemic to Lake Baikal deduced from nuclear DNA evidence.
MOL.PHYLOGENET.EVOL. 8 415-422 (1997).

3. BALL, L.E., OATIS, J.E. JR., DHARMASIRI, K., BUSMAN, M., WANG, J., COWDEN, L.B., GALIJATOVIC, A., CHEN, N., CROUCH, R.K. AND KNAPP, D.R.
Mass spectrometric analysis of integral membrane proteins: application to complete mapping of bacteriorhodopsins and rhodopsin.
PROTEIN SCI. 7 758-764 (1998).

8. A closer look at supposed GPCRs

i) During the presentation, we looked at a sequence that was matched by Pfam's HMM for rhodopsin-like GPCRs. Let's take a closer look. Type **gp153_human** into **Find-o-Matic**'s search box; **drag-and-drop** the sequence into CINEMA's **editor window**; invoke the context-sensitive menu and select the **alignment** option – align the sequences using **Muscle** or **ClustalW**. Does the result look convincing? Try predicting the locations of TM domains using **Tmap** or **TMHMM** – again, what do you see? Does this sequence look like a rhodopsin-like GPCR?
ii) We also looked at a sequence that was published in *Science* as the second ever GPCR identified in *Arabidopsis thaliana*. Again, let's take a close look. First, delete **gp153_human** from the alignment, by invoking the context-sensitive menu and choosing the **delete sequence** option. Now retrieve **q9c929_arath** using **Find-o-Matic**; as before, drop this into the **editor window** and align the sequences using **Muscle** or **ClustalW**. Again, does the result look convincing? Predict the locations of TM domains using **Tmap** or **TMHMM** – what do you see? Does this sequence look like a GPCR? Would you have published it in *Science*?! In the original *Science* paper, it was pointed out that the sequence was a homologue of the lanthionine synthase C-like proteins, the structure of one of which has been solved – **2G0D**. Try adding this to your alignment and then visualising the locations of the predicted TM domain. Have fun!

Getting help on CINEMA's full functionality

CINEMA is a fully functional alignment editor, so has numerous functions that we've not had time to describe here. For example, you may add and delete sequences, insert and delete gaps, group sequences together to perform group edits, find particular regions or motifs, auto-align to anchor motifs, and so on. For more details, there is a comprehensive help document at http://utopia.cs.man.ac.uk/documentation