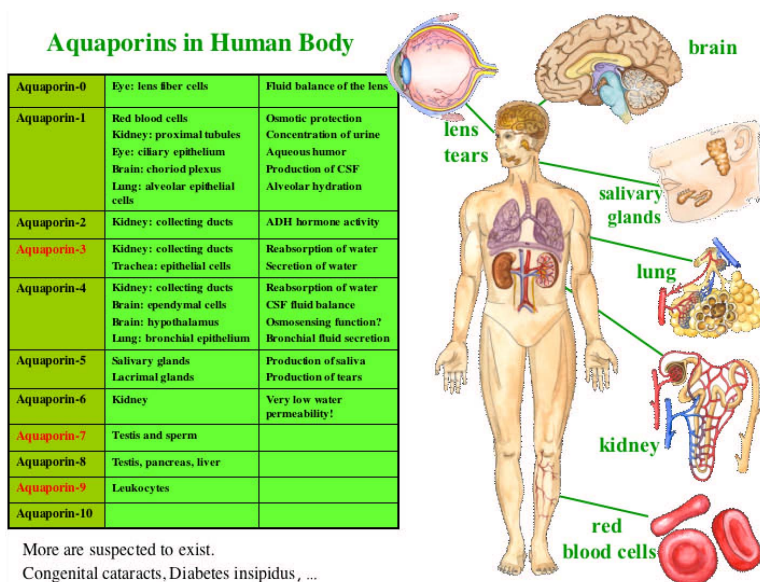


University of Illinois at Urbana-Champaign
 NIH Resource for Macromolecular Modeling and Bioinformatics
 Beckman Institute
 Computational Biophysics Workshop

Aquaporins



VMD Developers:

John Stone

Dan Wright

John Eargle

Fatemeh Khalili

Elizabeth Villa

Yi Wang

Kwok-Yan Chan

Emad Tajkhorshid

Brijeet Dhaliwal

Zan Luthey-Schulten

January 2012

A current version of this tutorial is available at
<http://www.ks.uiuc.edu/Training/Tutorials/>

Contents

1	Introduction to Aquaporin Structure	7
1.1	Structure and Function of Aquaporins	7
1.2	Loading other aquaporin structures	13
2	Structural Alignment of Aquaporins	15
2.1	Starting Multiple Sequence Alignment	15
2.2	Setting parameters for Multiple Structure Alignments	15
2.3	Aligning the molecules	16
3	Comparing Protein Sequence and Structure	19
3.1	Protein Structure	19
3.2	Protein Sequence	20
4	Residue Selection	23
4.1	Structure conservation	23
4.2	Sequence conservation	26
5	Investigating Structural Alignment	28
5.1	RMSD per Residue	28
6	Examining the Aquaporin Tetramer	31
6.1	Loading Tetramer	31
6.2	Examining channel lining and tetrameric contacts	32
7	Phylogenetic Tree	34
8	Evolutionary Profile of AQPs	36
8.1	Configure BLAST for Multiseq	36
8.2	Load Structures for AQPs in all three domains of life	37
8.3	Load Sequences for AQPs in all three domains of life	38
8.4	Align Sequences Using a Structural Profile	40
8.5	Construct an Evolutionary Profile for AQPs	40

Introduction

The Multiple Sequence Alignment extension to VMD

The Multiple Sequence Alignment (Multiseq) extension to VMD links protein structures to protein sequences and allows you to compare proteins in terms of structure and sequence.

Multiseq is designed to study protein mechanism through available genetic information, such as evolutionary conservation, offering biomedical researchers a new tool to examine protein structure and function.

This tutorial can be used by both new and previous users of VMD. However, it is recommended that new users go through the “VMD Molecular Graphics” tutorial, in order to gain further knowledge about the overall program.¹

The present tutorial has been designed specifically for VMD with Multiseq and should take about an hour to complete in its entirety.

Aquaporins

We will use the family of aquaporins for a case study in the applications of the Multiple Sequence Alignment tool. Aquaporins (AQPs) are membrane channel proteins found in a wide range of organisms, from archaea and bacteria to plants and animals. AQPs facilitate the rapid transport of water across cellular membranes and are of fundamental importance to the control of cell volume and transcellular water traffic.

In humans, there are at least 11 different aquaporin types (see cover). The kidney alone contains AQP1, AQP2 and AQP3. These proteins are responsible for filtering hundreds of liter of water per day in the human body. In addition to water, a subfamily of AQPs, named aquaglyceroporins, also selectively transport small molecules such as glycerol.

Here we use Multiseq to conduct a comparative study of the structure and sequence of four aquaporins from different species: human AQP1, bovine AQP1, AqpZ from *E.coli*, and GlpF (*E.coli* glycerol facilitator). The aquaporin AqpM from Archaea will be introduced in the last section.

¹URL: <http://www.ks.uiuc.edu/Training/Tutorials/>

Getting Started

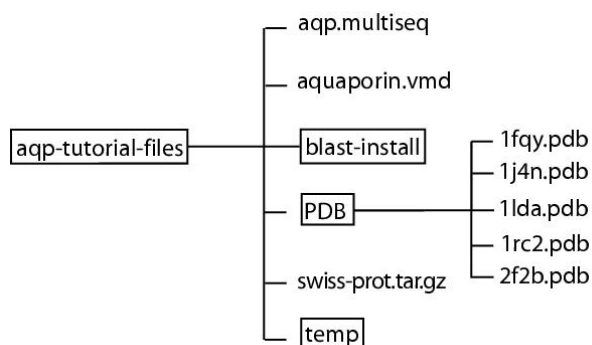
Multiseq, in its current release, is operable on the following platforms:

- Macintosh OS X
- Solaris
- Linux
- Windows

This tutorial requires certain files, located in

`/Desktop/Workshop/aqp-tutorial/aqp-tutorial-files/`

The file structure of this directory appears in the following figure:



Needed Software

VMD, containing the MultiSeq plugin (version of VMD 1.8.7 beta 6 and later), can be obtained from <http://www.ks.uiuc.edu/Research/vmd>. Once VMD is installed, you can start the tutorial by starting VMD. To start VMD:

- **Mac OS X:** Double click on the VMD application icon in the Applications directory.
- **Linux and SUN:** Type `vmd` in a terminal window.
- **Windows:** Select `Start → Programs → VMD`.

Getting Help

We have set up a mailing list, Tutorials-L, which will act as a forum for discussions and questions about the tutorials offered by the Theoretical and Computational Biophysics Group (<http://www.ks.uiuc.edu/Training/Tutorials/>). If you have questions, comments, etc, please suscribe to the list. Instructions can be found at http://www.ks.uiuc.edu/Training/Tutorials/mailling_list.

1 Introduction to Aquaporin Structure

In this unit you will use the conventional molecular graphics tools of VMD to become familiar with the key structural features of aquaporins, and how these are related to aquaporin function.

1.1 Structure and Function of Aquaporins

You will first consider bovine aquaporin (PDB code 1J4N).

In order to learn about the structural features of aquaporins, you will create several graphical representations of the molecule. Here, we will only guide you through the steps to create the necessary representations. If you want to know more about graphical representations, please look at the main VMD tutorial².



Using a VMD saved state. If you are familiar with VMD and don't want to go through all the steps of creating representations, you can use the VMD saved state `aquaporin.vmd` located in your tutorial directory to load the molecule with all the necessary representation. However, we recommend that you take the time and go through the suggested steps. If you choose to use the saved state, you still need to read through this section to learn about important structural features of aquaporins.

We also note that in case you need to interrupt the tutorial, you can save the current state of the VMD session (see VMD tutorial) and resume your study at a later time.

- 1 Choose the File → New Molecule... menu item in the VMD Main window. Another window, the Molecule File Browser, will appear on your screen.
- 2 Use the Browse... button to find the file `1j4n.pdb` in the directory `aqp-tutorial-files` → PDB in the tutorial directory. Note that when you select the file, you will be back in the Molecule File Browser window. In order to actually load the file you have to press Load. Do not forget to do this!



Getting a PDB file over the Internet. In case your computer is connected to the Internet, you can use VMD to get the file `1j4n.pdb` automatically from the Protein Data Bank³(see VMD tutorial).

Now, the aquaporin is shown on your screen in the OpenGL Display window. You may close the Molecule File Browser window at any time.

²URL: <http://www.ks.uiuc.edu/Training/Tutorials/>

³URL: <http://www.pdb.org>

- 3 Choose the **Graphics → Representations...** menu item. A window called **Graphical Representations** will appear and you will see the current graphical representation used to display your molecule highlighted in light green.
- 4 In the **Draw Style** tab you can change the style and color of the representation. In the **Selected Atoms** text entry of the **Graphical Representations** window, delete the word **all**. In the place of **all**, type **protein** and press the **Apply** button in the bottom right-hand corner of the window, or hit **Enter** or **Return** key on your keyboard. It is important that you do this every time you type something in **Selected Atoms**. From **Drawing Method**, choose the **Tube** menu item. The representation draws a tube along the backbone of the protein (Fig. 1). Now, change the color of the molecule by choosing the **Molecule** menu item from the **Coloring Method** menu.

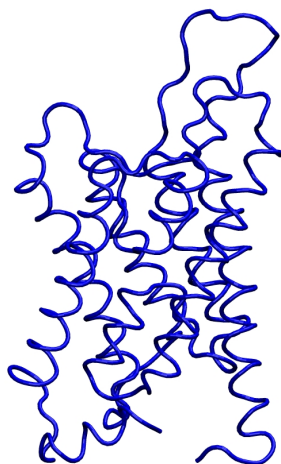


Figure 1: Tube representation of aquaporin structure.

The tube representation you created will be used throughout the tutorial to look at the structural alignment of aquaporins, but before you start using the **Multiseq** program, you will create other representations in order to learn some of the important structural features of aquaporins.

- 5 In the **Graphical Representations** window, click the **Create Rep** button. This will create a new representation, identical to the one you had before. This time, choose the drawing method **NewCartoon**. This drawing method shows the secondary structure of aquaporins (Figure 2).
- 6 In the **Graphical Representations** window, double-click on the first representation. This will make the text of the first representation red in this menu, and will hide the corresponding representation in the **OpenGL Display** window.

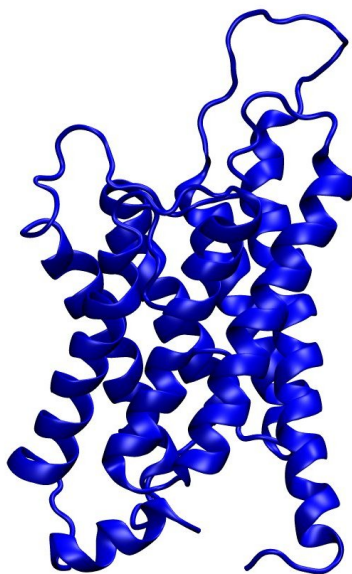
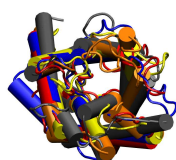


Figure 2: The NewCartoon representation of an aquaporin structure.



Secondary Structure of Aquaporins. Aquaporins are composed of six transmembrane α -helices and two reentrant loops, formed by a short helix and an extended polypeptide that returns to the surface of the protein. The two reentrant loops, which are surrounded by the six long helices, meet each other in the middle of the channel. These loops line the channel and are of great functional importance.

- 7 Click on the OpenGL window and use your mouse to rotate the molecule and look at the structure of aquaporins.

For the rest of this section, you will continue to look at the aquaporin structure in Tube representation. You can always go back and look at the NewCartoon representation, if necessary.

- 8 In the Graphical Representations window, double click on the first representation. This will show it again in the OpenGL window. Double click on the second representation (NewCartoon) to hide it.

Now look closer at the structural details of the aquaporin structural elements.

- 9 Create a new representation, clicking on the Create Rep button. This time, you will focus on the helices of the reentrant loops mentioned above (see box). In the Selected Atoms text entry of the Graphical Representations window type `resid 79 to 88 196 to 204`. These residues correspond to two helices of aquaporin that face each other in the middle of the channel. Choose the coloring method `ColorID → 1` and the drawing method `Tube`.

- 10 Repeat step 9. This time in **Selected Atoms**, type **resid 74 to 78 191 to 195**, and for the **Coloring Method**, select **ColorID → 7**. These residues correspond to the extended polypeptide regions of the reentrant loops of aquaporins (see box).

Now that you have localized the reentrant loops, the rest of the protein will appear dim.

- 11 Click (once) on the first representation. In the **Material** menu, choose **Transparent**. Your OpenGL window should look like Fig. 3.

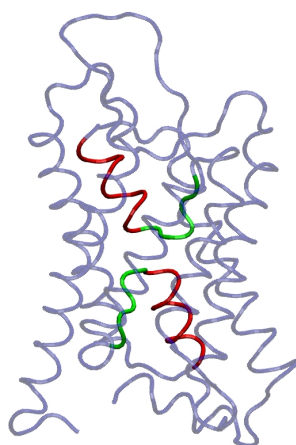
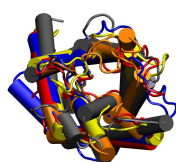


Figure 3: The reentrant loops are a key structural feature of aquaporins. Each reentrant loop is formed by a short helix, shown in red, and an extended polypeptide, shown in green.

Now that you can locate the reentrant loops, you will examine the NPA motif within the reentrant loops.



NPA motif. The NPA motifs, as the name implies, are formed by amino acids N – asparagine, P – proline, and A – alanine. They are highly conserved signature motifs in all aquaporins. They stabilize the reentrant loops through multiple hydrogen bonds, as shown in figure 4.

- 12 Create a new representation, by clicking on the **Create Rep** button. Make sure the new representation is not transparent. In the **Selected Atoms** text entry of the **Graphical Representations** window, type **resid 78 to 80 194 to 196**. These residues correspond to the two NPA motifs present in aquaporin. Choose the drawing method **Licorice** and coloring method **Type**. Look at the NPA motifs, can you see how they would be hydrogen bonded?

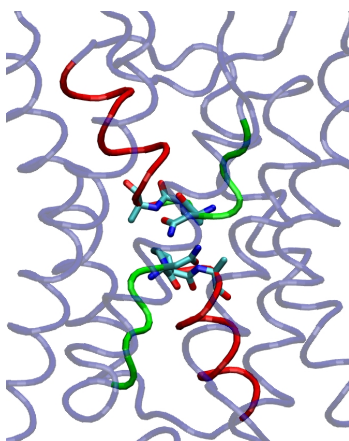


Figure 4: The NPA motifs in aquaporins.

- 13** Create another representation by clicking the Create Rep button. With the same coloring and drawing methods of the previous selection, type **resid 197** in the Selected Atoms text window. This will draw an important arginine residue in aquaporins.

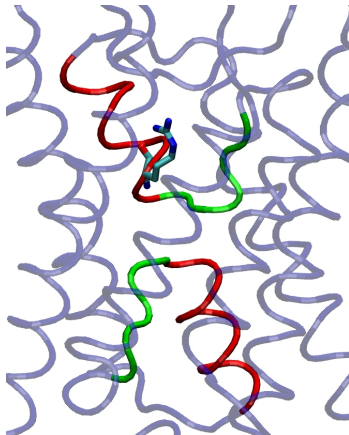
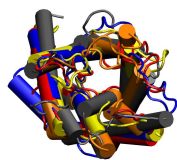
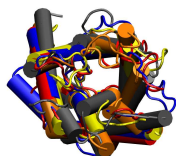


Figure 5: The conserved arginine provides H-bond sites to water molecules inside the channel.



Conserved arginine. This arginine is the only charged residue in the channel, and is conserved in all aquaporins. This arginine lines the pore and provides additional H-bond sites for water molecules permeating the channel (See Fig. 5).

- 14 Repeat step 13, but this time type `resname GLU` in the Selected Atoms text window. This will select all the glutamates in the protein. Notice that there are only two glutamate side chains in the transmembrane region that are in the helical core of the protein. If you look closely (c.f. Fig. 6) you will see that each one of these glutamates sits behind a reentrant loop.



Glutamates. The glutamates serve to stabilize the structure of the reentrant loops. They form direct hydrogen bonds with the loops and thus hold them in their place (Fig. 6). These glutamates are also highly conserved in the aquaporin family. It is interesting that mutation of one of these glutamates results in cataracts in human eyes!

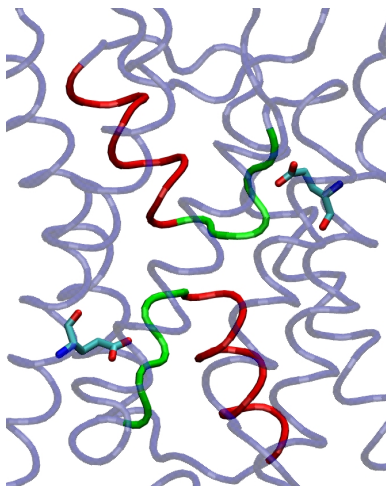
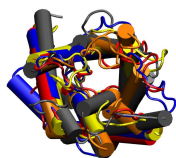


Figure 6: The glutamates help stabilize the reentrant loops.

Finally, you are going to look into the aquaporin pore.

- 15 Create a representation. In the Selected Atoms text window type `protein`. Set the drawing method to VDW (van der Waals). Each atom is now represented by a sphere. In this way you can view the volumetric distribution of the protein.
- 16 Rotate the molecule so that you can see it from the top. Look down the pore. Can you see the pore? How many water molecules you think fit there?



Pore. The pore is so narrow (Fig. 7) that it can fit water molecules only in single file. This feature is very important for the selectivity of the channel.

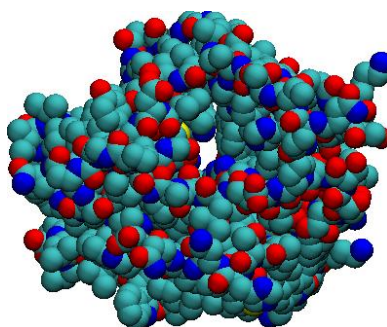


Figure 7: The pore of aquaporins can only accomodate water molecules in single file.

- 17** Create a new representation. In the Selected Atoms text window type **water**. Set the Drawing Method to VDW and the Coloring Method to **type**. This will create a representation with all the water molecules the crystallographers resolved in the crystal structure. You can now see that the pore is filled with water molecules forming a single file.

Now that you have learned about the structural features of aquaporins, you will load the other three aquaporins mentioned above and get ready to align them. Before that, you should turn off all representations created so far, leaving only the first.

- 18** In the Graphical Representations window, double click on each of the graphical representations, except the first. This will make them appear in light color in this menu, and will hide the corresponding views in the OpenGL window. Select the first representation and set the material to Opaque.

1.2 Loading other aquaporin structures

In this tutorial you will be comparing four aquaporin structures listed in Table 1.2. You have already loaded the **1j4n** molecule. Now, you will load the other three aquaporins.

PDB code	Description
1j4n	Bovine AQP1
1fqy	Human AQP1
1lda	E. coli Glycerol Facilitator (GlpF)
1rc2	E. coli AqpZ

Table 1: Aquaporin structures

- 1** The remaining aquaporins, **1fqy**, **1lda**, **1rc2**, need to be loaded into VMD and have their molecule's graphical representations individually changed

to Tube. To do this, you need to refer back to the previous section, *Structure and Function of Aquaporins*, and repeat steps 1 through 4. Make sure that each PDB is loaded into a new molecule.

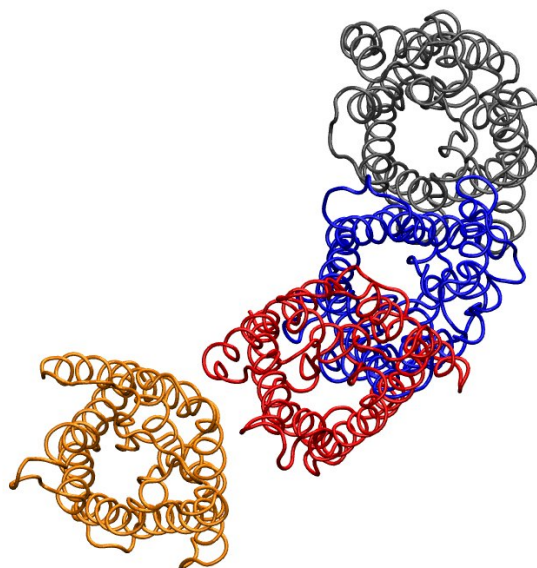


Figure 8: Aquaporin structures loaded into VMD.

You should now see the four molecules loaded in the OpenGL Display (Figure 8). If you do not see all molecules, hit **S** on the keyboard while in the OpenGL display. Then use your mouse to scale the molecules in the display, such that you can see all four. Note that the structures are not aligned, but rather placed according to the coordinates specified in their PDB files.

- 2 Close the Graphical Representations window and Molecule File Browser window.



1rc2. The original 1rc2 file contains two copies of the aquaporin molecule. This is the way the crystallographers reported this structure to the PDB since the crystal used for analysis contained two independent copies of AqpZ. For this tutorial, we include only one copy of the molecule in the provided pdb file.

You are now ready to start performing Multiple Sequence Alignment.

2 Structural Alignment of Aquaporins

In this section, you will align the aquaporin molecules loaded in VMD using the Multiseq program.

2.1 Starting Multiple Sequence Alignment

Now that you have loaded the four proteins in Table 1.2, you will use Multiseq program to align the proteins and analyze their structural and sequence relationships.

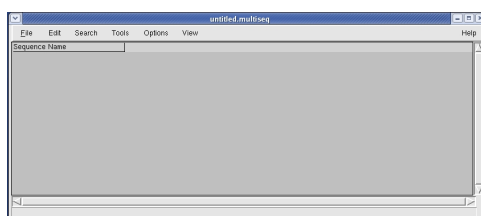


Figure 9: Multiseq Program Window

- 1 Within the VMD main window, choose the Extensions menu.
- 2 In the Extensions menu select Analysis → MultiSeq.

This is the main Multiseq program window. The rest of the tutorial and exercises will use features from this window, unless specified otherwise. You may be asked to update some databases if this is the first time you use Multiseq. If this is the case, simply click Yes and wait for Multiseq to start.

By default, Multiseq will align all four loaded molecules, unless you delete the molecule(s) in the VMD Main window. However, note that some crystal structures come with water and detergent molecules, which should not be used to align the structures. In your MultiSeq window, you will find that all the protein structures are categorized as **VMD Protein Structures** and other molecules, e.g. water, are in the **VMD Nucleic Structures** category. Keep the four protein structures and delete the two structures under **VMD Nucleic Structures** by clicking them and press the delete button on your keyboard.

2.2 Setting parameters for Multiple Structure Alignments

Multiple Sequence Alignment uses the program STAMP to align protein molecules.



STAMP. The program STAMP (Structural Alignment of Multiple Proteins) is a tool for aligning protein sequences based on a three-dimensional structure. The STAMP algorithm minimizes the C_α distance between aligned residues of each molecule by applying globally optimal rigid-body rotations and translations. Also, note that you can perform alignments on molecules that are structurally similar. If you try to align proteins that have no common structures, STAMP will have no means to align them. If you would like further information about how the alignment occurs, please refer to the STAMP manual at:
<http://www.rfcgr.mrc.ac.uk/Registered/Help/stamp/stamp.html>

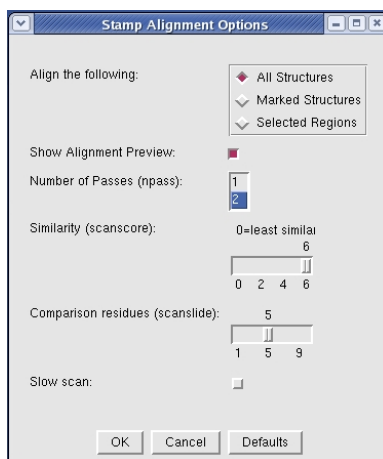


Figure 10: Stamp Alignment Options

Before you align the molecules, you may want to change the default parameters needed by STAMP.

- 1 Choose the Tools → Stamp Structural Alignment tool. The window shows several parameters that can be set (Figure 10). For this tutorial, you will use the default parameters.

You can look at the STAMP user guide (see box) for information on how to optimize the STAMP parameters to obtain a better alignment.

2.3 Aligning the molecules

Now that you have opened the MultiSeq window and made sure the STAMP parameters are correct, you can align the molecules you loaded into VMD.

- 1 In the Stamp Alignment Options window, choose Align the following: All Structures and go to the bottom of the menu and select OK.

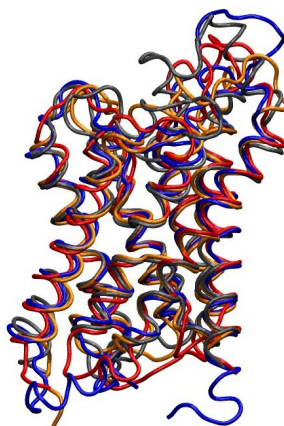


Figure 11: Structural alignment of four aquaporin molecules.

The molecules have been aligned. You can see this both in the OpenGL window (Fig. 11) and in the MultiSeq program window (Fig. 12). Take some time to look at the alignment. Note that your alignment may not immediately resemble Fig. 11. This is because MultiSeq displays the protein in “NewCartoon” representation by default. To compare your alignment result with Fig. 11, change the representations for each protein to “Tube” in your Graphical Representations window. You could also change the color for each protein.

- 2 Click on the OpenGL window (Fig. 11). Move your mouse around, rotating the molecules. Do you think this is a good alignment? Look at the top of the molecules: Can you see the pore?

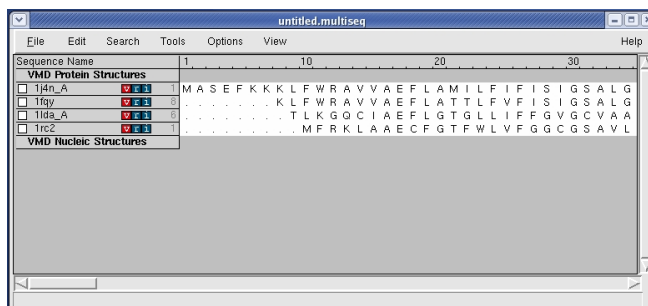


Figure 12: Sequence Alignment of four aquaporin molecules.

- 3 Now, click on the Multiseq program window (Fig. 12). Move the left-to-right scroll at the bottom of the window, and take a look at the residues in the Sequence Display. Observe that there are gaps, represented by dashes, in the sequences when they are aligned.

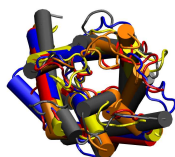
In the next section, you will start using Multiple Sequence Alignment features for the alignment you just made.

3 Comparing Protein Sequence and Structure

In this unit you will study the structure and sequence conservation of the different aquaporins you have aligned in the Multiseq window. Conservation is a term we will use frequently in exercises from this point forward. Within the context of protein analysis, conservation refers to high levels of similarity contingent on the employed metric. Structure conservation occurs when the structural aspects (e.g. α helix, β sheets), of the aligned proteins are highly similar. Likewise, sequence conservation happens when at certain points of the aligned protein molecules the amino acid of each molecule is the same.

3.1 Protein Structure

In order to better understand the structural conservation within the aligned molecules, Multiseq provides tools that allow you to color the molecules according to their Q value (Qres), a measure of structural conservation.



Q per residue. Q is a parameter that indicates *structural identity*. Q accounts for the fraction of similar native contacts between the aligned residues in two proteins⁴. Q=1 implies that structures are identical. When Q has a low score (0.1-0.3), structures are not aligned well, i.e., only a small fraction of the C_{α} atoms superimpose. Q per residue is the contribution from each residue to the overall Q value of aligned structures.

You will now color the molecules according to the value of Q per residue obtained in the alignment.

- 1 In the Multiseq program window, choose the View $\rightarrow$ Coloring $\rightarrow$ Qres (Fig. 13).
- 2 Look at the OpenGL window to see the impact this selection has made on the coloring of the aligned molecules (Figure 14).

Rotate the molecule to see how much of it has turned blue. Notice that the transmembrane helices of the aligned molecules have turned blue. The blue areas indicate that the molecules are structurally conserved at those points. If there is no correspondence in structural proximities at these points, the points appear red. Observe how the structurally least similar segments tend to be on the periphery of the aligned molecule. Note that the loops tend to be red, while the helices are blue.

⁴Eastwood, M.P., C. Hardin, Z. Luthey-Schulten, and P.G. Wolynes. "Evaluating the protein structure-prediction schemes using energy landscape theory." IBM J. Res. Dev. 45: 475-497, 2001. URL: <http://www.research.ibm.com/journal/rd/453/eastwood.pdf>

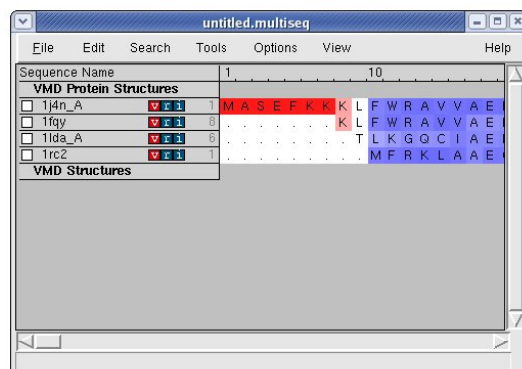


Figure 13: Molecule Coloring

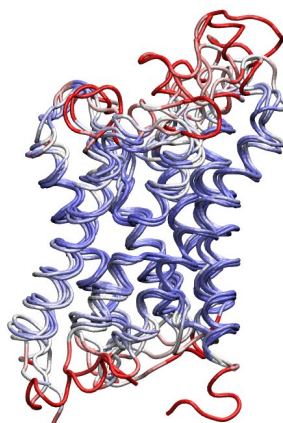


Figure 14: Molecules colored by Qres value

3.2 Protein Sequence

You have examined the structural similarity between the molecules. Now you will look at the sequence conservation. **Multiseq** has a feature to color the molecules according to how much the sequence is conserved within the aligned molecules. This tool, **Sequence identity**, colors each amino acid according to the degree of conservation within the alignment: blue means highly conserved, whereas red means very low or no conservation.

- 1 Choose View → Coloring → Sequence Identity.

Before you look at the viewer window, can you anticipate what will happen to the coloring of the molecules? Will the molecules still be blue in the trans-membrane region, as they were when **Qres** was used to determine structure conservation?

- 2 Now take a look at the viewer window. As you can see (Fig. 15), a fair portion of the molecules has turned red, indicating less sequence conservation than structural similarity.

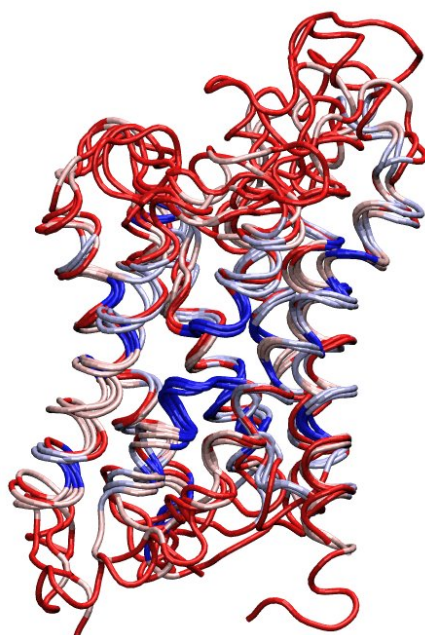


Figure 15: Molecules colored by sequence identity

- 3 Note that the residues at the site where the two short helices interact are blue. Look at the molecules from the top (c.f. Fig. 16). Do you notice the blue residues tend to be on the inside of the pore?

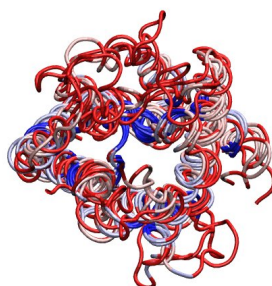


Figure 16: Conserved residues (shown in blue) are located inside the pore.

Note that the coloring of the molecules using Sequence Identity indicates that the sequence conservation is much less in comparison to the structural

conservation. Sometimes structures from the same family have less than 10% sequence identity, yet are structurally similar. A well known example is the protein myoglobin that we recommend for self-study.

To examine the relationship between sequence and structure in more detail, in the next section you will use the **Select Residues** feature.

4 Residue Selection

In this section, you will use the Select Residues tool for finding structural features of aquaporins that have been conserved throughout species and are crucial for their function.

The Select Residues features analysis of structural similarity and sequence conservation. It makes use of different measures to highlight residues in the Sequence Display and Structure Display simultaneously. It allows you to examine the conservation and similarity on a per residue basis, using two different measures, Qres value measuring structure similarity and Sequence Identity measuring sequence conservation.

4.1 Structure conservation

You can use Multiseq to look at the residues that are structurally similar above or below a certain value of Qres.

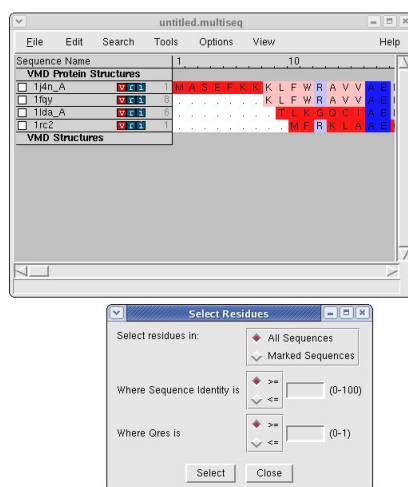


Figure 17: Residue Selection

- 1 Choose the Search → Select Residues tool. A side menu will appear (Fig. 17). Within the menu, you can select residues based on their Sequence Identity or Qres.
- 2 In box following Qres, select $\geq$ and key in 0.5 for the value.
- 3 Click on the Select button.

Note the changes in the display of the Open GL (structure) Display window (Fig. 18, left) and in the Sequence Display of the Multiseq program window (Fig. 19).

- 4 Look in the Sequence Display (Fig. 19). Since you selected the Qres value to be greater or equal to 0.5, most of the structural elements of the aligned molecules will be shown as conserved (similar), and will be displayed in yellow.

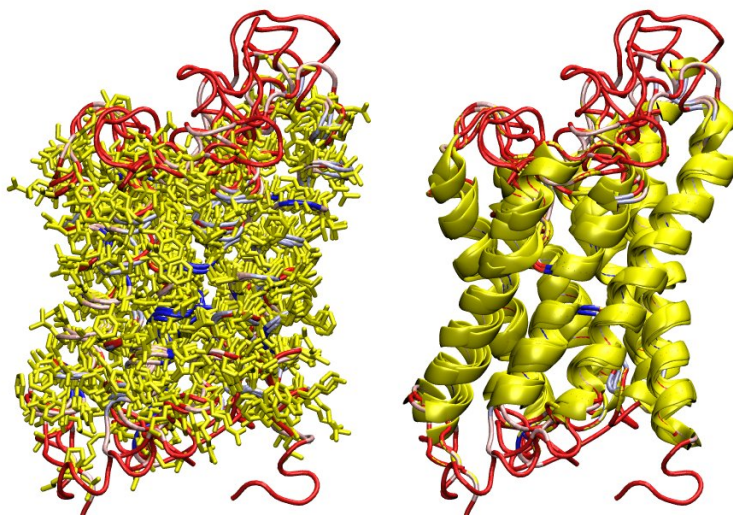


Figure 18: View of the Residue Selection for $Q_{res} \geq 0.5$ in the Open GL Display window. Highlight style set to Licorice (left) and NewCartoon (right)

- 5 Now, look at the OpenGL Display window (Fig. 18, left). In this case also most parts of the molecules are highlighted, in particular, the transmembrane segments.

Sequence Name	0.0	0.5	1.0
VMD Protein Structures			
144n.A	G	H	I S G A H L N P A V T L G L L S C I S V L A I M Y I
144p.A	G	A	H L N P A V T L G L L S C I S V L A I M Y I
144s.A	A	G V S G A H L N P A V T I A L W L F F D K R V I P F I	
144t.C	G	H I S G G H F N P A V T I G L W A G R F P A K E V V G V Y	
VMD Nucleic Structures			

Figure 19: View of the Residue Selection for $Q \geq 0.5$ in the Sequence Display window.

- 6 Due to the high level of conservation, the Licorice highlighting style is not very informative. To improve the view, chose the View $\rightarrow$ Highlight Style menu item, and click on the New Cartoon style (Fig. 18, right). You can recognize immediately that high Qres values indicating conservation of neighborhood relationships for amino acids are seen in the transmembrane region throughout.

Now you will investigate what happens for higher values of Qres.

- 7 In the Select Residues window, select $\geq$ after the Qres and key in 0.75. Click on the Select button.
- 8 Look at the OpenGL Display window (Fig. 20). Notice that most of the structurally similar regions are still in the transmembrane region, but the highlighted regions are more confined than before.

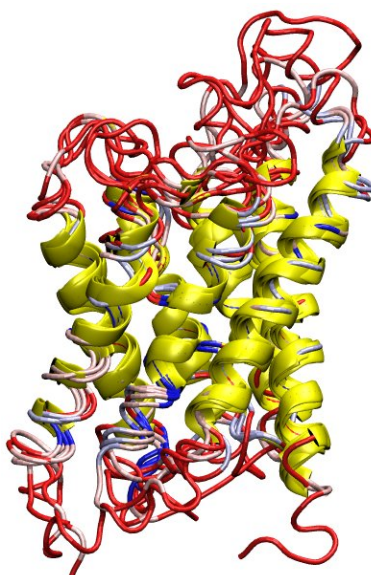


Figure 20: View of the Residue Selection for $Q_{res} \geq 0.75$ in the OpenGL Display window

- 9 Look in the Sequence Display window (Fig. 21). You can still distinguish blocks of structurally similar residues, which correspond to the helices. They are not as well defined as before, but the structural similarity is still noticeable when probing for high values of Qres.

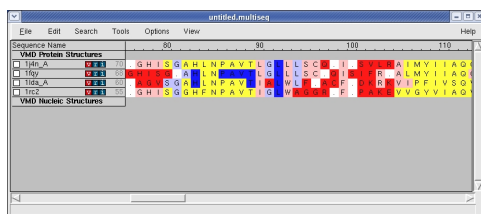


Figure 21: View of the Residue Selection for $Q_{res} \geq 0.75$ in the Sequence Display window

- 10 Repeat steps 7 to 9 and set the Q_{res} to ≥ 0.9 . You will notice that almost all the structure conservation is lost.

What does this mean? The aligned aquaporins have a moderate to high structure conservation. Structurally, the most similar areas are the transmembrane helices.

4.2 Sequence conservation

You can also use **Select Residues** to look at the residues that are conserved on the sequence level.

- 1 Go back to the main Multiseq program window and select **Search** → **Select Residues**.

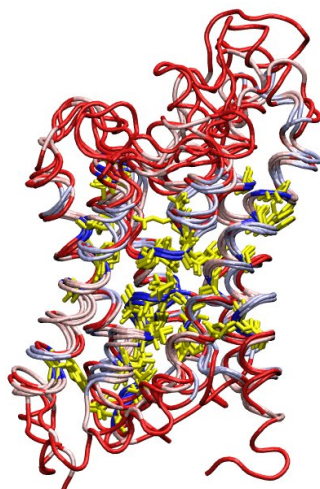


Figure 22: View of the Residue Selection by Sequence Identity in the OpenGL Display window.

- 2 In the **Select Residues** window, following **Sequence Identity** select $\geq$.
- 3 You can then key in a value between 0 and 100. Try 70, which will select residues that were assigned a sequence identity measure of 70% or more.
- 4 Click on the **Select** button.
- 5 Look at the **Sequence Display** in the Multiseq program window. You will notice very few residues are selected. If you look carefully, you will locate some of the residues that correspond to key structural features of aquaporins, for which you made representations in section 1 of this tutorial. Search for the NPA motifs, highlighted in yellow (Fig. 23).

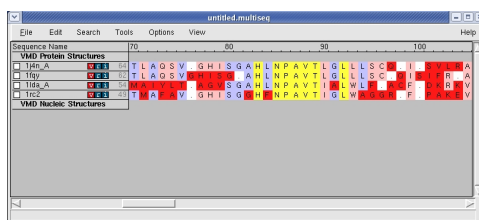


Figure 23: View of the Residue Selection by Sequence Identity in the Sequence Display window

- 6 In the OpenGL window, look at the areas highlighted in yellow. Notice that many residues in the reentrant loops (Fig. 3) conserved their sequence over the course of evolution. This is a clear indication that the reentrant loops are functionally significant.

The conserved amino acids in the sequence are mostly located in the pore, and in the reentrant loops. You will now compare these residues to the representations you created in section 1 of this tutorial, which correspond to residues that are relevant to the function of aquaporins.

- 7 In the Multiseq program window, select **View** → **Highlight Style** → **Licorice**.
- 8 In the VMD Main window, select **Graphics** → **Representations....**
- 9 In the Graphical Representations window, in the **Selected Molecules** pull-down menu, select the molecule **1j4n**.
- 10 In the Graphical Representations window, double click on the representations you created for the NPA motif, the glutamates and the arginine. What is the sequence identity for these amino acids? Why do you think the NPA motif, the lysine and the glutamates are conserved throughout species?

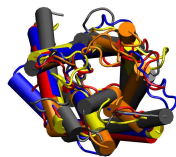
When you are done examining the NPA motif, close the Graphical Representations window to proceed to the next section.

5 Investigating Structural Alignment

In this section you will use RMSD values between molecules as a measure to see how close structures of molecules are after alignment.

5.1 RMSD per Residue

RMSD per residue values compare two proteins and show how well they align. Here we use RMSD values that compare solely the C_α atoms of the proteins.



RMSD values. The Root Mean Squared Deviation (RMSD) is a numerical measure of the difference between two structures. It is defined as :

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N_{atoms}} (r_i(1) - r_i(2))^2}{N_{atoms}}} \quad (1)$$

where N_{atoms} is the number of atoms whose positions are being compared, and $r_i(1)$ and $r_i(2)$ are the position of atom i in each molecule.

- 1 In the Multiseq program window choose Tools → Plot Data. A window entitled Data Plotting will appear on your screen. (Fig. 24)
- 2 Select the All sequences option.
- 3 Choose the RMSD checkbox, for the Per-residue data to plot.

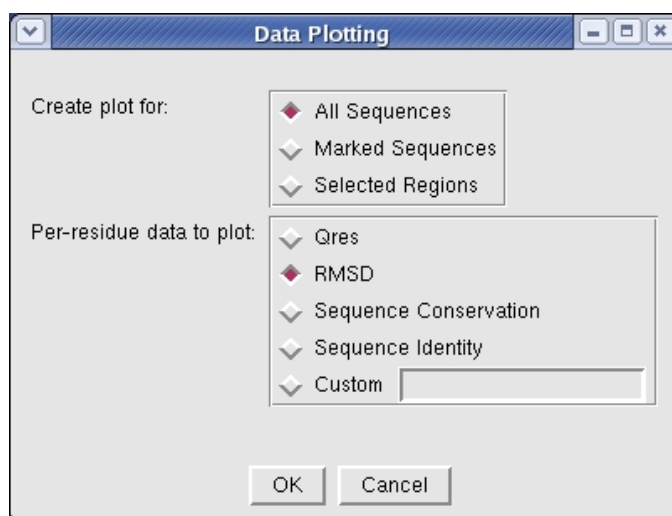


Figure 24: Data Plotting window.

- 4 Click on the OK button to produce a plot of the RMSD values between the first molecule in the Sequence Alignment window and any of the aligned structures (Fig. 25). The RMSD values between each pair of the aligned residues are shown along the sequence of the protein, exhibiting regions of close alignment (small RMSD values) and of poor alignment (large RMSD values).

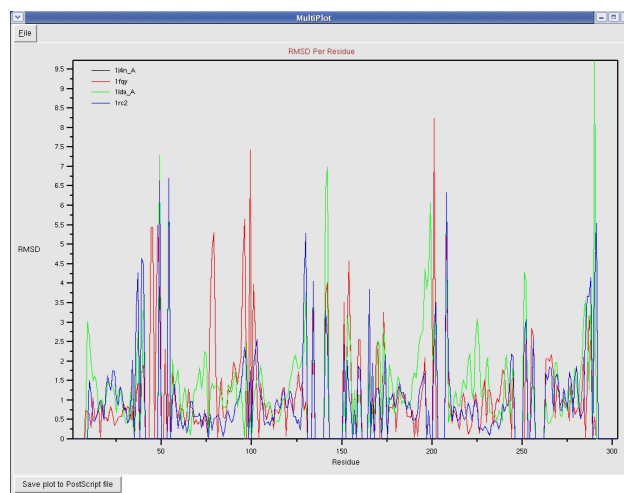


Figure 25: RMSD values between bovine AQP1 (pdb code:1j4n) and human AQP1 (1f1y), *E.Coli* GlpF (1lda), and *E.Coli* AqpZ (1rc2)

- 5 Go back to the Multiseq program window. Compare the RMSD graph with the sequence alignment, and identify the residues with large RMSD values. What structural elements do they correspond to? As you will see later, they correspond to the loops.

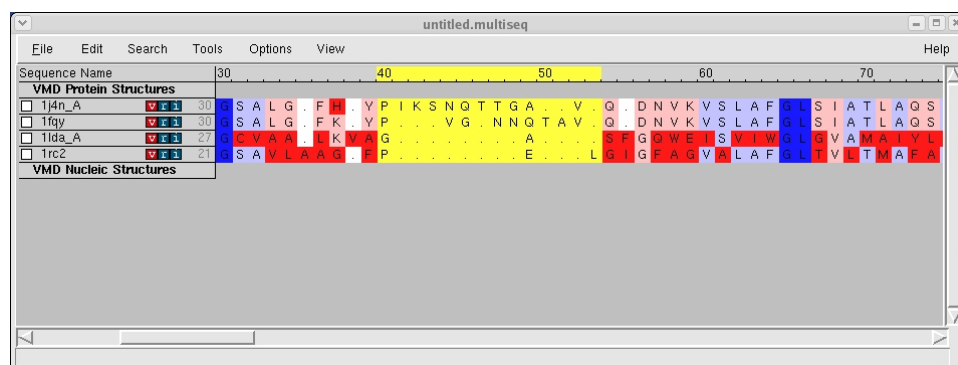


Figure 26: Multiseq program window. Residues corresponding to the first peak in the RMSD graph are highlighted in yellow.

- 6 Select the residues corresponding to the first peak of the RMSD graph in the Multiseq program window, by clicking on them. They will be highlighted as you click on them (Fig. 26). If you want to select residues in different regions, hold down the shift key while selecting them.

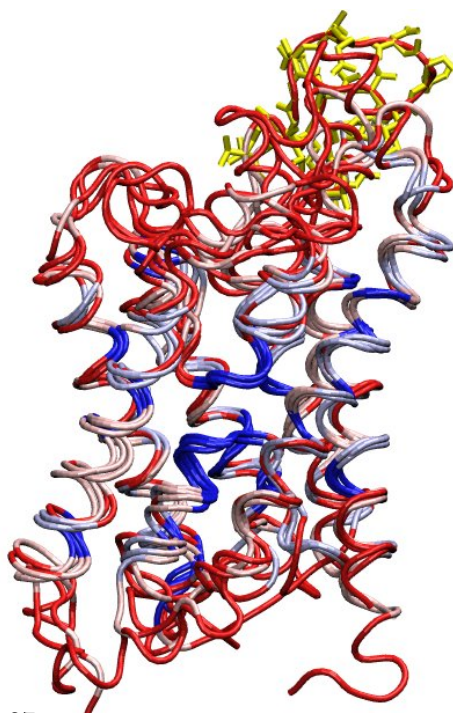


Figure 27: Highlighted residues shown in bond representation

- 7 Go back to the VMD Main window. Choose Graphics $\rightarrow$ Representations... In the Graphical Representations window, (Fig. 28) you should find that a new representation has been created for each molecule, with the residues you have selected as shown in Fig. 26. Look at the representation for molecule 1fqy and compare the selection with the RMSD graph (Fig. 25)
- 8 To proceed to the next section, close the Graphical Representations and the RMSD per Residue windows.

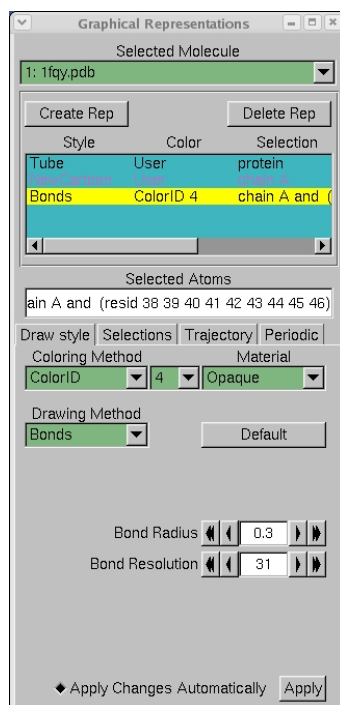


Figure 28: Graphical Representations Window

6 Examining the Aquaporin Tetramer

Aquaporins form homotetramers in cell membranes. The monomers are arranged side by side in a tight cluster. Each monomer conducts water. A fifth pore is formed in the center of the tetramer, with a yet unknown function. In the following we have a closer look at the aqua-glyceroporin GlpF (PDB code: 1lda), an aquaporin that conducts both water and glycerol.

6.1 Loading Tetramer

In this section we will load the tetramer structure of GlpF and align it with the other four AQP molecules. A pdb file with the atom coordinates of the tetramer is provided with the tutorial (see *Getting Started*).

- 1 Go to the VMD Main Window
- 2 Choose File → New Molecule. Another window, the Molecule File Browser, will appear.
- 3 Use the Browse button to find the file `glpf_tetramer.pdb` in the directory `aqp_tutorial_files` → PDB, in the tutorial directory.

- 4 Once you have selected the pdb file, press the **Load** button in the Molecule File Browser window to load the molecule. It will take a while for the tetramer to load
- 5 The GlpF tetramer is loaded in **Lines** representation. To have a better view of the molecule after alignment, you need to change the molecule's graphical representation to **Tube**. From the VMD Main window, choose **Graphics** → **Representations**. In the Graphical Representation window, change the molecule's **Coloring Method** to **SegName**, and its **Drawing Method** to **Tube**.

Now that you have the tetramer loaded, go back to the Multiseq program window and select **Tools** → **STAMP structural alignment** to align this molecule with the other four. Note that for this section you have to have the aquaporins 1j4n, 1fqy, 1lda and 1rc2 (Table 1.2) loaded. If you don't have them, go back and load the four pdb files, as explained in section 1.

6.2 Examining channel lining and tetrameric contacts

In the previous section, you have identified the conserved residues among the aquaporin molecules. Here we look where these residues are located in the tetrameric structure.

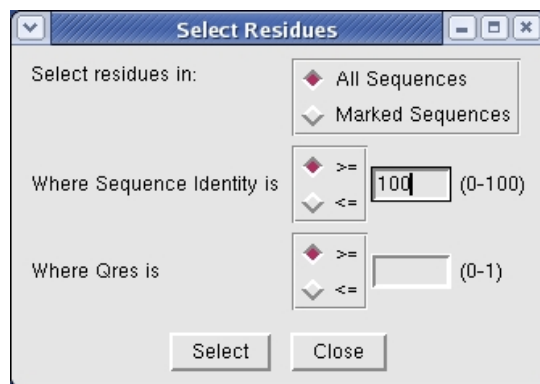


Figure 29: Select Residue Dialog

- 1 Go to the Multiseq program window.
- 2 Choose **Search** → **Select Residues**.
- 3 The Select Residue dialog will appear (Fig. 29). Select **All Sequences**.
- 4 In the following box, select **Where Sequence Identity is**, $\geq$, and type in 100 for the value.

- 5 Click on the **Select** button. Now you will see the conserved residues highlighted in the OpenGL Display window as well as in the Multiseq program window.
- 6 Take some time to examine the highlighted parts as shown in Figure 30. You recognize that most conserved residues line the channel interior. Locate the conserved residues that are outside the water conducting pore. Note that these residues are indeed located at the subunit-subunit interfaces of the tetrameric GlpF.

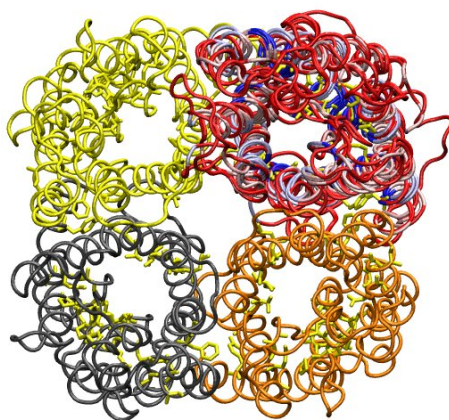


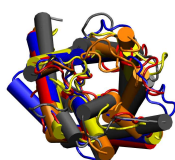
Figure 30: GlpF tetramer aligned with other aquaporin molecules. Conserved residues are highlighted in bond representation (yellow).

- 7 Before starting the next section, go to the VMD Multiseq program window, choose the tetramer molecule by clicking on it (It will turn yellow), and delete the molecule by using the **Delete** key on your keyboard.

7 Phylogenetic Tree

The Phylogenetic Tree feature, in the Multiseq program, elucidates the structure-based relationships between the four aquaporins considered.

Structure based phylogenetic trees can be constructed according to the RMSD or Q values between the molecules after alignment. Here we use the Q values to obtain the phylogenetic tree of four aquaporins.



The Phylogenetic Tree. A phylogenetic tree is a dendrogram, representing the succession of biological form by similarity-based clustering. Classical taxonomists use these methods to infer evolutionary relationships of multicellular organisms based on morphology. Molecular evolutionary studies use DNA, RNA, protein sequences or protein structures to depict the evolutionary relationships of genes and gene products. In this tutorial we employ Q_H and RMSD to depict evolution of protein structure. For a comprehensive explanation of phylogenetic trees, see *Inferring Phylogenies* by Joseph Felsenstein⁵.

To utilize this feature:

- 1 Align the structures again, by going to the Multiseq program window and selecting Tools → Stamp Structural Alignment.
- 2 In the Stamp Structural Alignment window, select All Structures, and keep the default values for the rest of the parameters. Press the OK button to align the structures.
- 3 In the Multiseq program window choose Tools → Phylogenetic Tree. The Create Phylogenetic Tree dialog will appear (Fig. 31).

Here you can choose to construct a structure based phylogenetic tree according to the Q_H values or RMSD. Q_H is a variation of Q (see below) that accounts for both gapped and aligned regions.



What is Q_H ? Q_H is an adaptation of the Q value, and is essentially a *metric for structural homology*. Q_H is comprised of two terms: Qalign and Qgap. Each time you align two structures, unless they align perfectly, the structure can be divided into two parts: the part that aligns (Qalign) and the part that due to insertion and deletion of the protein sequence, does *not* align (Qgap). Qgap accounts for how much the insertion perturbs the core structure of the protein and for the size of the insertion in sequence and structure.

⁵J. Felsenstein *Inferring Phylogenies*. Sinauer Associates, Inc.: 2004.

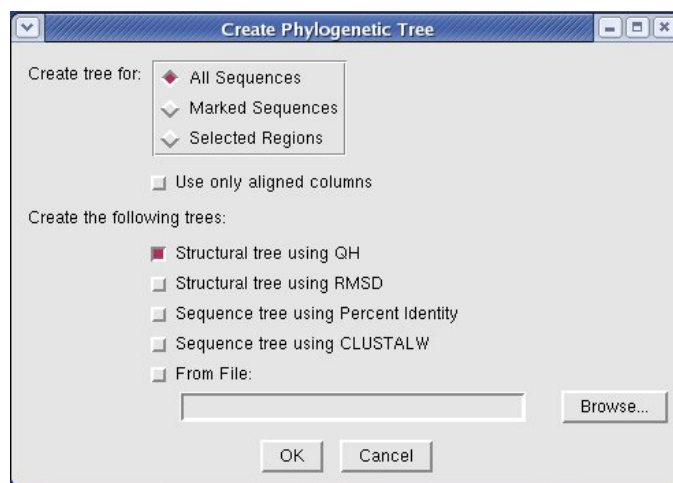


Figure 31: Create Phylogenetic Tree window.

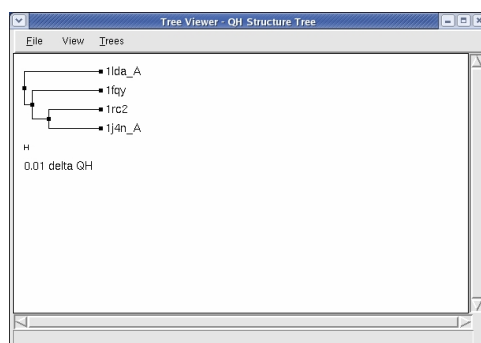
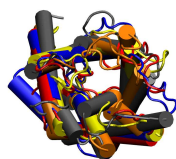


Figure 32: Structure-based phylogenetic tree, for the aquaporins in Table 1.2.

- 4 Select Structural tree using Q_H , and press the OK button. A phylogenetic tree based on the Q_H values will be calculated and drawn as shown in Figure 32. Here you can see the relationship between the four aquaporin, e.g., how the *E.Coli* AqpZ (1r2c) is related to human AQP1 (1fqy). Identify bovine and human aquaporins in the Tree.
- 5 You can choose to construct the phylogenetic tree of the four aquaporins based on their sequence information, and will be discussed further in section 8.
- 6 You can close the Phylogenetic tree window.
- 7 Quit VMD.

8 Evolutionary Profile of AQPs

So far you have learned how to construct a phylogenetic tree using the structural alignment of AQPs. The sequence alignment can also be used to build a phylogenetic tree, especially when protein structures are not available. To make use of both the structural and sequence information, Multiseq now allows you to merge the two types of alignments and construct a complete evolutionary profile (EP) for the proteins being studied. In this section, you will learn how to obtain the EP for AQPs. For more information on using Multiseq program to perform evolutionary analysis, please refer to the Evolution of Biomolecular Structure tutorial.



Evolutionary Profile. Due to practical reasons such as the amenability of the system, protein databases are often biased towards one domain of life, *e.g.* bacteria. To characterize proteins based on their patterns in evolution, rather than their occupancy in the current databases, Multiseq utilizes the QR factorization algorithm to allow for the selection of a minimum non-redundant set of sequences and structures. The derived non-redundant representative multiple alignments and statistical representations are termed evolutionary profiles.

8.1 Configure BLAST for Multiseq

For the following section you will need to install BLAST in your computer. BLAST is a software that searches through sequence databases and locate those sequences that are similar to a query sequence. It is available online at <http://www.ncbi.nlm.nih.gov/BLAST/>. Here we will install a local copy of BLAST for Multiseq.

- 1 Create a directory into which BLAST will be installed.

Examples:

Unix/Linux: /usr/local/blast;

Mac OS X: /Applications/Blast;

Windows: C:\ Blast

- 2 Extract the archives of BLAST.

Copy the blast installation file for your platform from the aqp-tutorial-files → blast-install directory to the directory you've made. In Unix or Linux, extract the files by using the command `tar zxvf filename`. On Mac OS X or Windows, double-click the file.

- 3 Do the same for swiss-prot.tar.gz.

Repeat the above two steps: create a directory for swiss-prot, copy the file swiss-prot.tar.gz from aqp-tutorial-files to the directory you've created, and extract it.

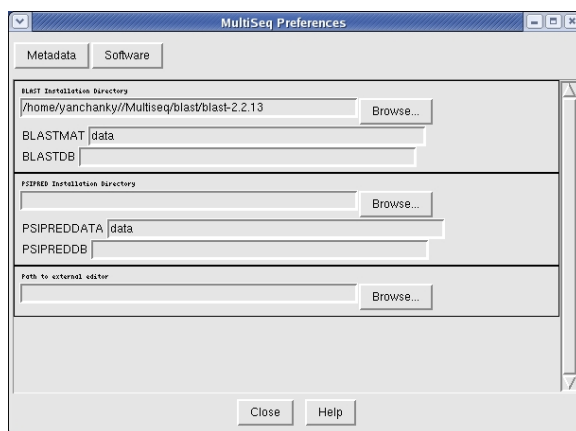


Figure 33: Choose the directory for BLAST.

4 Set the BLAST installation location in Multiseq.

In the Multiseq program window, choose **File** → **Preferences**. Click on the **Software** button in the new dialog to bring up the software preferences. Click on the **Browse** button in the **BLAST Installation Directory** section and select the directory into which you installed BLAST (Fig. 33). *Note: You may be asked by Multiseq to update certain databases before you could continue, if so, click Yes and wait for Multiseq to finish the update. In Linux or Mac OS X, you may have a directory named blast-2.2.13 in your installation directory. Pick this directory if you have it.*

8.2 Load Structures for AQPs in all three domains of life

AQPs are present in all three domains of life. To build a complete EP for AQPs, we will first perform a structural alignment for AQPs in all three domains of life: Eukaryota, Bacteria, and Archaea. In the previous sections, you have seen the structures of human AQP1 (1fqy) and *E. coli* AqpZ (1rc2). Here you will also need AqpM (2f2b) from Archaea to construct the EP. We have provided these pdb files in the tutorial files for you.

- 1 Open a new VMD and load the pdb files 1fqy, 1rc2, and 2f2b one by one.
- 2 Open the Multiseq program by clicking **Extensions** → **Analysis** → **Multiseq**.
- 3 In the Multiseq program window, keep the protein structures under **VMD Protein Structures** and delete all structures under **VMD Nucleic Structures**.

If you loaded your structures by giving the pdb code to VMD (not with the files we prepared for you), you may have more than one structure for each of the pdb code you entered, *i.e.*, besides 1rc2_A, you may also have 1rc2_B. This

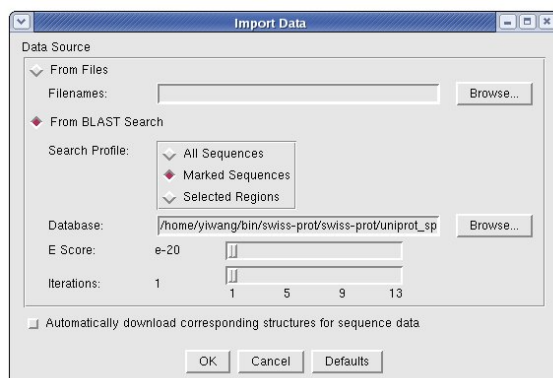


Figure 34: Import structures for AQPs in Multiseq.

indicates that in the original pdb file, there are two different structures for the protein. The difference between them is usually very small, and does not affect the alignment we are going to perform. Therefore, simply delete 1rc2_B and keep 1rc2_A.

8.3 Load Sequences for AQPs in all three domains of life

Now that you have the structures of AQPs loaded, we will use BLAST to find sequences of AQPs in all three domains of life. Each of the three structures will be used as a query sequence by BLAST, and sequences in the `swiss-prot` database will be compared with them, one at a time. Those sequences similar to our query sequence will be picked by BLAST and loaded in Multiseq.

- 1 In the Multiseq window, check the box in front of 1fqy. Then click File → Import Data.

You will find the same window you've seen when loading the pdb structures. This time, choose From BLAST Search under Data Source and select Marked Sequences (Fig. 34).

- 2 Click the Browse button after Databases, and go to the directory where you extracted the file `swiss-prot.tar.gz`. You should find a directory named `swiss-prot`. Go into that directory and select the file `uniprot_sprot`.
- 3 Choose e^{-20} for E Score and 1 for Iterations and then click OK.

BLAST is now searching the database with 1fqy as a query sequence. This should take a minute or two. A new window named BLAST Search Results will open once the search has finished. Note that the `swiss-prot` database provided here only contains sequence data for proteins in this session. You cannot rely on it for other proteins that you want to investigate. Moreover, the database is not an updated one, so visit the BLAST online databases if you want the latest

results. As you may have noticed, 100 sequences have been found using the query sequence 1fqy. We will only keep those sequences from the Eukaryota domain, since our query sequence is from Eukaryota. Later we will find sequences in Bacteria and Archaea using the query sequences 1rc2 and 2f2b, respectively. This should make our search more accurate.

- 4 In the BLAST Search Results window, under Domains, unselect the All list and select Eukaryota. Click Apply Filter.

You will find that only 87 sequences are left (Fig. 35).

- 5 Click Accept. The sequences will be loaded in Multiseq.

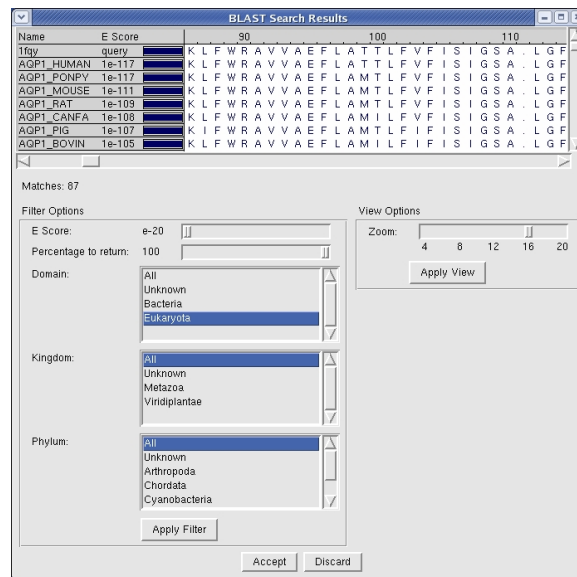


Figure 35: Search result of BLAST.

- 6 Check the box in front of 1rc2 and uncheck 1fqy in the Multiseq window.

Now you could repeat the above process and find Bacteria sequences using 1rc2 as a query sequence. You should find 28 sequences from Bacteria. Repeat this process using 2f2b as a query sequence and get 3 sequences for Archaea.

Before we continue, save your Multiseq session by clicking File → Save Session and save it as `aqp.multiseq`. You can load the session later by clicking clicking File → Load Session. There is a saved `aqp.multiseq` session in the tutorial files, in case you'd like to check with it.

8.4 Align Sequences Using a Structural Profile

In order to analyze the three structures and the 118 sequences of AQPs together, we need to first align them. What we will do is to first align the structures using the STAMP structural alignment tool mentioned in section 2, and then we will use the structural alignment to guide the sequence alignments.

- 1 Mark the three pdb structures by checking the boxes in front of them. Make sure that no other sequences are marked.
- 2 Click Tools → Stamp Structural Alignment and choose to align Marked Structures and then click OK.
- 3 Unmark structures and mark all the sequences. Remove gaps in the sequences by clicking Edit → Remove Gaps and then select Remove gaps from: Marked sequences, and Remove these types of gaps: All gaps.

You could select all the sequences at once by clicking on the first sequence, pressing the **shift** button and then clicking on the last sequence. All the sequences should appear in yellow now, which means they are highlighted. Press the **shift** button and check one box in front of any highlighted sequence. All other boxes for the highlighted sequences should be automatically checked.

- 4 Click Tools → Sequence Alignment.

A new window named Sequence Alignment Options should appear (Fig 36). Check ClustalW under Alignment Program. As we are going to align the sequences using the structural alignment, choose Profile/Sequence Alignment instead of Multiple Alignment in the window. Under Align marked sequences to group, select VMD Protein Structures, and then click OK. This should take two or three minutes.

Now you have a complete structural based alignment of the AQPs in all three domains. Try coloring it by sequence identity by clicking View → Coloring → Sequence Identity (Fig.37).

8.5 Construct an Evolutionary Profile for AQPs

Although we have obtained the structures and sequences for AQPs in all three domains and aligned them together, what we have now is not an evolutionary profile yet. We still need to get rid of the redundancy in these sequences caused by the biased databases. Multiseq provides a Sequence QR tool which can be used to select a minimum non-redundant set from the sequences, using a threshold specified by the user.

- 1 Mark all the sequences and make sure that the structures are unmarked. Click Search → Select Non-Redundant Set from the menu.

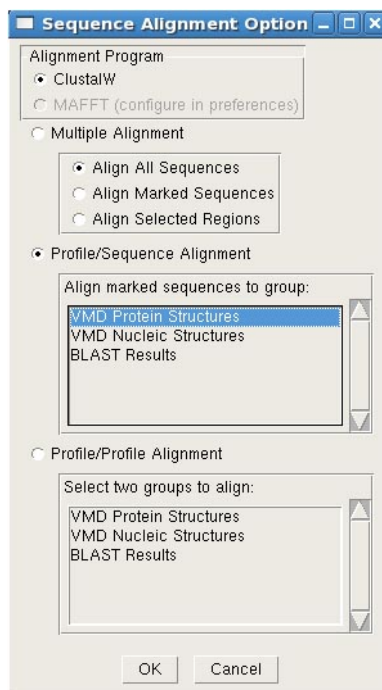


Figure 36: Sequence Alignment Options window.

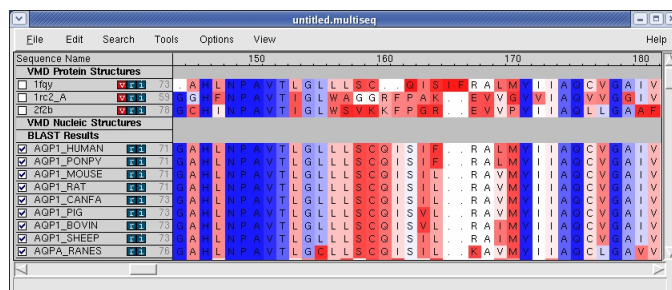


Figure 37: ClustalW alignment result using the structural profile

A new window named Select Non-Redundant Set should show up (Fig. 38). In this window, choose Select from → Marked Sequences, and choose Using Sequence QR. Set the Maximum PID to 75 and then click OK.

You should find that some of your sequences are highlighted after the program stopped calculating. These represent the non-redundant set that Multiseq selected for you. Group them together by clicking Options → Grouping → From Selection and enter “NR set” for the new group. This should put all your highlighted sequences into a group named “NR set”. This is the evolutionary profile (EP) for AQPs. You could now create the phylogenetic tree using the EP of AQPs: simply delete all the sequences except the ones in the NR set and create

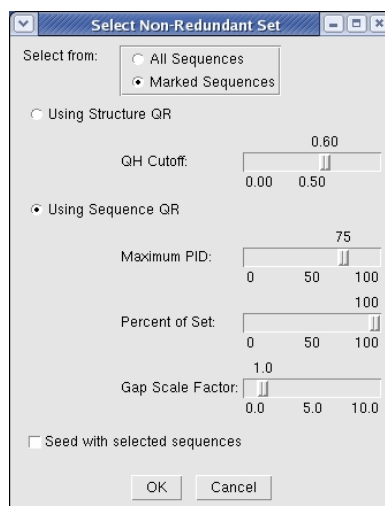


Figure 38: Select Non-Redundant Set window.

a phlogenetic tree as you did in section 7.

Evolutionary profile provides an “unbiased” view for the evolutionary relationship of the proteins in investigation. Using EP, scientists have successfully identified a new subfamily for the protein cysteinyl-tRNA synthetase. For more details on constructing EP and performing evolutionary analysis, please refer to the Evolution of Biomolecular Structure tutorial.

Acknowledgements

Development of this tutorial was supported by the National Institutes of Health (P41-RR005969 - Resource for Macromolecular Modeling and Bioinformatics).