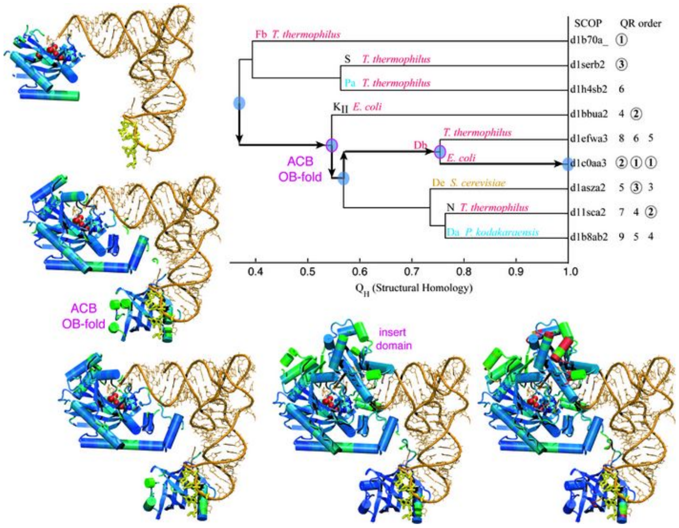
University of Illinois at Urbana-Champaign
Luthey-Schulten Group
Theoretical and Computational Biophysics Group
Computational Biophysics Workshop

Evolution of Protein Structure

Aspartyl-tRNA Synthetase



VMD Developers:

Dan Wright

John Eargle

John Stone

Dr. Zan Luthey-Schulten

Brijeet Dhaliwal

Patrick O'Donoghue

Rommie Amaro

September 2005.

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

Contents

1 Introduction

1.1 The Multiple Alignment Extension to VMD

The new Multiple Alignment version of VMD that is currently under development was originally created to allow biomedical researchers to study the evolutionary changes in sequence and structure of proteins across all three domains of life, from bacteria to humans. The comparative sequence and structure metrics, and analysis tools introduced in the accompanying article by O'Donoghue and Luthey-Schulten [1] are now part of this new version. In particular, the Luthey-Schulten group has included a recently developed structure-based measure of homology Q_H (see Appendix B), that takes into account the effect of insertions and deletions and has been shown to produce accurate structure-based phylogenetic trees. The STAMP structural alignment algorithm, kindly provided by our colleagues Russell and Barton, is included in our alpha release [2] . We plan to offer biomedical researchers a tool to examine the changes in protein structure in the correct statistical framework. As a result, Multiple Alignment is an invaluable tool for relating protein structure to its function or misfunction. Since the accompanying tutorials were created for a program that is truly a work in progress, we limit our demonstrations to the examination only of the correlation of sequence and structure changes and represent these changes in terms of structural phylogenetic trees.

This tutorial showcases the new software tools in Multiple Alignment and will allow the reader to reconstruct the figures in the accompanying review article entitled "The Evolution of Structure in Aminoacyl-tRNA Synthetases." It is designed such that it can be used by both new and previous users of VMD, however, it is highly recommended that new users go through the "VMD Molecular Graphics" tutorial in order to gain a working knowledge of the program. *This tutorial has been designed specifically for VMD with Multiple Alignment and should take about an hour to complete in its entirety.*

[1] P. O'Donoghue and Z. Luthey-Schulten. "Evolution of Structure in Aminoacyl-tRNA Synthetases" MMBR, 67(4):550-73. December, 2003.

[2] R.B. Russell and G.J. Barton. "Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels." Proteins: Struct. Func. Genet., 14:309-323. 1992.

1.2 tRNA Synthetases: Precise translation machines

The aminoacyl-tRNA synthetases (AARSs) are key proteins involved in the translation machinery in living organisms; it is not surprising, therefore, that these enzymes are found in all three domains of life. There are twenty specific tRNA synthetases (one for each amino acid), although not all organisms contain the full set. Studying the function, structure, and evolution of these proteins remains an area of intense interest as, in addition to being a major constituent of the translation process, these proteins are also believed to contain vital information spanning the evolution of life from the ancient "RNA world" to the modern form of life.

$$\text{aa} + \text{ATP} + \text{tRNA} \xleftrightarrow{\text{AARS}} \text{aa-tRNA} + \text{AMP} + \text{PP}_i$$

Figure 1: The reaction catalyzed by the aminoacyl tRNA synthetases (aa could be any amino acid).

The AARSs are responsible for loading the twenty different amino acids onto the cognate tRNA during protein synthesis (see Figure 1). Each AARS is a multidomain protein consisting of (at least) a catalytic domain and an anticodon binding domain. In all known cases, the synthetases divide into class I or class II types; class I AARSs exemplify the basic Rossmann fold, while class II AARSs exhibit a fold that is unique to them and biotin synthetase holoenzyme. Additionally, some of the AARSs, for example aspartyl-tRNA synthetase, have an "insert domain" within their catalytic domain (see Figure 2). Recognition of the tRNA molecule is typically performed by the anticodon domain, however residues that have degenerate codons (*e.g.* serine has six different codons) have been found to exploit other features in the tRNA for recognition (*e.g.* the acceptor arm or the so-called discriminator base). These molecular machines operate with remarkable precision, making only one mistake in every 10,000 translations. The intricate architecture of specific tRNA synthetases helps to discriminate against mis-coding.

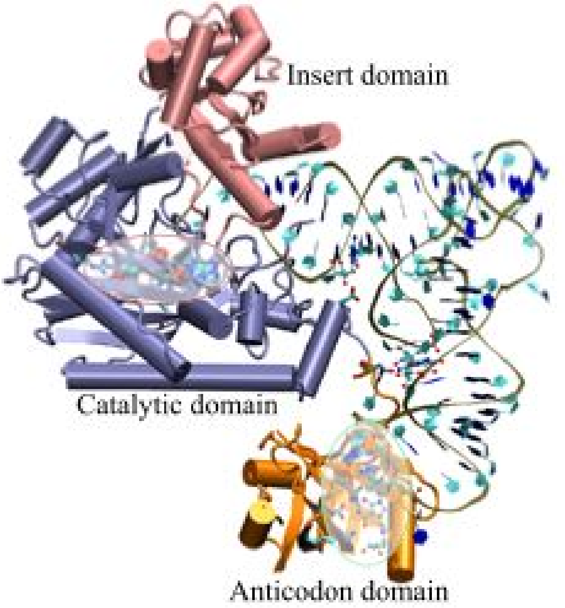


Figure 2: A snapshot of AspRS-tRNA aspartyl-adenylate complex (from *E. Coli*) in the active form. Note the anticodon binding domain (orange), the insertion domain (pink), and the catalytic domain (blue). tRNA is docked to AspRS, and the catalytic active site is highlighted within the catalytic domain (red bubble); the aspartyl-adenylate substrate is shown in space-filling representation. The residues involved in specific base recognition on the tRNA are also highlighted within the anticodon binding domain (green bubble). Note that specific contacts between the tRNA and Asp-RS allow for strategic positioning of the tRNA relative to the enzyme.

2 Getting Started

2.1 Downloading Tutorial Files

Multiple Alignment with VMD, in its current release, is operable on the following platforms:

- Macintosh OS X
- Solaris
- Linux
- Windows

The tutorial requires VMD 1.8.3 with Multiple Alignment. Additionally, certain requisite files are available at:

http://www.ks.uiuc.edu/Training/Tutorials/

You are now prepared to begin the tutorial. To start using VMD, double-click on the VMD icon in the Applications Folder.

2.2 The Aspartyl-tRNA Synthetase Aspartyl-adenylate Complex

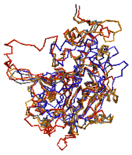
In order to become familiar with the structural and functional features of the AARSs, we will first explore the aspartyl-tRNA synthetase as complexed with aspartyl-adenylate and tRNA (PDB code: 1C0A). To do this:

1 Go to the TkConsole window.

2 Using the cd command, find the tutorial_aars file directory.

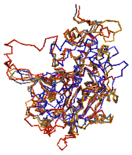
3 At the prompt type: `> source trna.vmd`

You should now have the AspRS-tRNA aspartyl-adenylate complex loaded in VMD. Take some time to explore the complex in the Open GL display; rotate the molecule; investigate the different features and components of the complex, including the location of substrates and the way tRNA is positioned in complex with the AspRS. Note that the tRNA makes contact with the synthetase in several locations.



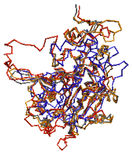
The structural domains of AspRS. Rotate the molecule in VMD and examine the different domains of the enzyme. How many domains does the AspRS have? Which domain is the catalytic domain? Which might be the anticodon binding domain? Which domain is the insertion domain? What domains does the tRNA interact with? *Hint: see figure 2.*

All of the AARSs are multidomain proteins, but the exact number and fold of each domain is specific to each aminoacyl-tRNA synthetase. AspRS has a catalytic domain (shown in blue), an anticodon binding domain (orange, sometimes also referred to as the N-terminal domain), and an insertion domain (shown in pink). Curiously, the insertion domain (residues 288 to 420) literally interrupts the sequence of the catalytic domain (comprised of residues 113 to 287 and 421 to 585) and only appears in the bacterial AspRS; archaea and eukarya AspRSs do not contain this insertion.



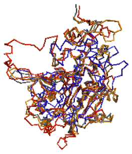
How does AspRS recognize the tRNA?. The N-terminal domain of AspRS is in close contact with which region of the tRNA? What bases make up the anticodon for aspartate?

Note how the N-terminal domain (colored orange) of the enzyme attaches itself to the anticodon in the tRNA; zoom in on the anticodon. The anticodon for aspartate is comprised of Q34, U35, and C36. Q stands for queuine and is a hypermodified base that marks the first position of the anticodon in the AARSs that code for Asp, Asn, His, and Tyr.



The chemistry of AARSs. Explore the active site of the AspRS-tRNA aspartyl-adenylate complex to answer the following questions: What reaction is catalyzed in the active site? What are the substrates? What products are made by this reaction? What part of the tRNA is involved in this reaction?

Use VMD to zoom in on the active site within the catalytic domain; you may want to rotate the molecule to get the best view possible. Note how the acceptor end of the tRNA sticks into the active site of the aspartyl synthetase. The substrate, aspartyl-adenylate, is shown in space-filling representation. The formation of the aspartyl-adenylate comes from one aspartate molecule and ATP; this adenylated species is "activated" and from here can easily be linked to the cognate tRNA with energy provided from the hydrolysis of ADP to AMP. Also note how the architecture of the active site prohibits the diffusion of this activated amino acid outside of the active site; the aspartyl-adenylate is trapped between the catalytic domain and the tRNA.



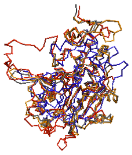
Where does the tRNA go once it is "charged" with its amino acid?. At the ribosome, the anticodon of the charged tRNA is matched to the mRNA codon. Then the tRNA is *deacylated* with the amino acid being added as the next residue to a nascent protein chain.

Send the tRNA off to the ribosome yourself by deleting the molecule before you begin the next part of the tutorial. You can do this by highlighting the molecule with your mouse in the VMD main window. Then selecting Molecule → Delete Molecule*.*

In the subsequent parts of this tutorial, we will use Multiple Alignment to align the catalytic domains of three AspRS molecules, one from each of the domains of life, as well as one serine tRNA synthetase. The catalytic domain of each species has been directly extracted from the ASTRAL database, which contains the structures of each of the proteins' domains. This tutorial will emphasize both structural and sequence-based analyses of the AARSs and ultimately create a phylogenetic tree illustrating the evolution of the proteins with respect to one another. For a more thorough explanation of the evolutionary considerations, as well as the computational methods involved, please see Ref. 1.

2.3 Loading Molecules

To further explore AARSs we will now examine three AspRS molecules alongside one SerRS molecule in Multiple Alignment. Before we begin, make sure you have deleted the aminoacyl-adenylate complex in the VMD Main window.



The Four AARSs molecules.

1asza2.pdb AspRS from yeast.

1b8aa2.pdb AspRS from *P.kodakaranesis*.

1efwa3.pdb AspRS from *T.thermophilus* complexed to tRNA.

1sera2.pdb SerRS from *T.thermophilus*.

To load our molecules we need to source a tcl script. The tcl script loads ASTRAL files that contain the atom coordinates of the four AARSs. In the TkConsole window, after you made sure you are in the tutorial_aars directory, at the prompt type:

```
> source msdemo.tcl.
```

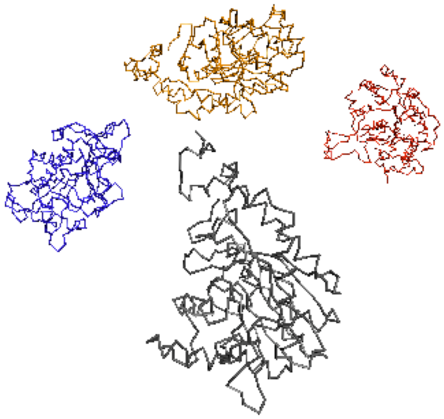


Figure 3: Four unaligned AARSs molecules

Within the OpenGL display window, the four molecules will appear randomly. We will now walk through the steps for aligning these molecules.

What is the ASTRAL database?. The ASTRAL database (http://astral.stanford.edu) is a compendium of protein domain structures derived from the PDB database. It divides each protein structure into its domain components. For example, AspRS is divided into three separate PDB files: one containing the catalytic domain, one with the insertion domain, and one for the anticodon binding domain. The names of the files contain the PDB extension, the letter *a* for ASTRAL, and a number, which corresponds to which domain it is in the original PDB file. For example, the anticodon binding domain for the AspRS-tRNA complex we have been investigating is: 1c0aa1.pdb.

2.4 Starting Multiple Alignment

From this point forward we will use Multiple Alignment extensively. In order to align and analyze the structural relationships of the four loaded molecules, we need to open Multiple Alignment.

1 Within the VMD main window, choose the Extensions menu.

2 In the Extensions menu select Analysis → Multiple Alignment.

A window entitled Multiple Alignment will appear on your screen. This is the main Multiple Alignment program window. The rest of the tutorial and exercises will use features from this window, unless otherwise specified.

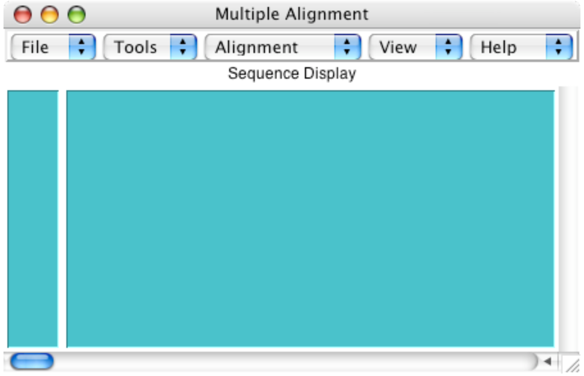


Figure 4: Multiple Alignment Window

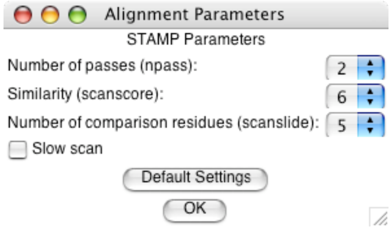
Multiple Alignment will align all four loaded molecules, unless you delete the molecule(s) in the VMD Main window.

2.5 Setting parameters for Multiple Structure Alignments

Before you align the molecules you may want to set certain parameters for the alignment.

1 Go to the Multiple Alignment window and select Alignment in the top pull-down menu.

2 Then click on Alignment Parameters. A new window entitled Alignment Parameters will appear with four setting options (see Figure 5).



STAMP Parameters

Number of passes (npass): 2

Similarity (scanscore): 6

Number of comparison residues (scanslide): 5

□ Slow scan

Figure 5: STAMP Parameters

This tutorial will use these settings. If you would like more information about STAMP parameters, please refer to the STAMP manual.[3] To proceed with the tutorial, close the Alignment Parameters window.

[3] The STAMP manual is available at http://www.rfcgr.mrc.ac.uk/Registered/Help/stamp/stamp.html

2.6 Aligning the molecules

Now that we have opened Multiple Alignment and made sure the STAMP parameters are correct, we can align the four molecules loaded into VMD.

How molecules are aligned in Multiple Alignment. Multiple Alignment uses the program STAMP to align protein molecules. 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(See Ref. 3).

To align the molecules:

1 In the main Multiple Alignment window go to the top pull-down menu and select Alignment.

2 Then select Run Structural Alignment.

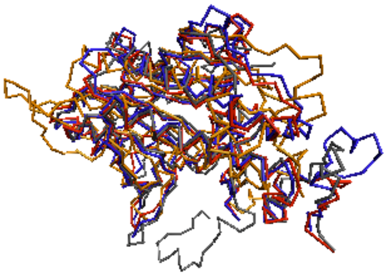


Figure 6: Structural alignment showing superposed AARSs.

Once this step is completed, you will be able to view the aligned molecule in the OpenGL Display window.

3 Comparing Protein Structure And Sequence

3.1 Protein Structure

In order to better understand the structure conservation between aligned molecules, certain tools involving the coloring of molecules are used within Multiple Alignment. In particular, Q per residue measures structure conservation. Structure conservation occurs when the structures between aligned proteins are similar.

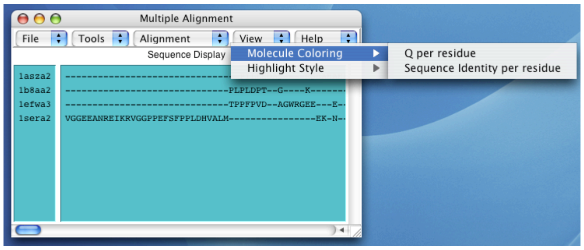
What is Q per residue?. To answer this question we first must consider "What is Q?" Q is a parameter borrowed from protein folding that indicates *structural identity*. Traditionally, Q has meant "the fraction of similar native contacts" between the aligned residues in two proteins[a], or in two different conformational states of the same protein. When Q = 1, it indicates that the structures are identical. When Q has a low score (*0.1*), it means the structures do not align well, or, in other words, only a small fraction of the C-alpha atoms superimpose. You will discover that homologs typically have Q$\geq$0.4. Q per residue is the contribution from each residue to the overall average Q value. For more information see Appendix A.

[a] Eastwood, M.P., C. Hardin, Z. Luthey-Schulten, and P.G. Wolynes. "Evaluating protein structure-prediction schemes using energy landscape theory." IBM J . Res. Dev. 45: 475-497. 2001

Q per residue, is accessed by:

1 Click on the View menu on the Multiple Alignment window.

2 Select Molecule Coloring.



```
1asza2  ------------------------------
1b8aa2  ------------------------------PLPLDPT--G----K------
1efwa3  ------------------------------TPPFPVD--AGWRGEE---E--
1sera2  VGGEEANREIKRVGGPPEFSFPPLDHVALM----------------EK-N--
```

Figure 7: Molecule Coloring

When you select Molecule Coloring another side menu should appear with the following options: Q per residue and Sequence Identity per residue.
Select Q per residue to visualize structural conservation. Look at the OpenGL Display window to see the impact this selection has made on the molecules.

Is there structure conservation?. Will there be a significant amount of structure conservation? Has there been a change in coloring for the aligned molecule? What does this mean in terms of the evolution of the AARSs?

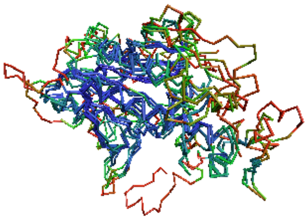


Figure 8: Structure Conservation

You will probably notice that several portions within the interior of the aligned molecules have turned blue. Rotate the molecule to see how much of it has turned blue. The blue areas indicate that the molecules are structurally conserved at those points. The red regions are not structurally conserved, and these correspond to insertions, which are typically found on the periphery of the molecule as shown in Figure 7 of O'Donoghue et al.

Exploring Insertions Further.. You can explore insertions by doing the following:

1. Go to the Sequence Display in the Multiple Alignment window.
2. Using your mouse, highlight an area of the aligned sequence that appears mostly as dashes.

In the Open GL display, a red region should appear in the Bonds Highlight Style. The yellow highlighted area corresponds with what you highlighted in the Sequence Display.

3.2 Protein Sequence

Now that we have examined the structural conservation between the molecules, it is important to examine the sequence conservation. Sequence conservation occurs when amino acid identity of the aligned residues match.
In order to access Sequence Identity per residue:

1 Click on the View menu.

2 Select Molecule Coloring → Sequence Identity per residue.

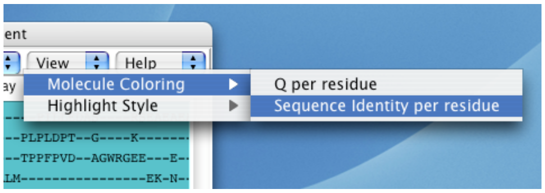
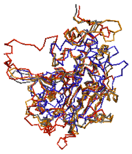


Figure 9: Sequence Identity per residue



How similar are sequence and structure conservation?. Before you look at the OpenGL Display window, can you anticipate what has happened to the molecules? Will the molecules still be blue in the center, as they were when Q per residue was used to determine structure conservation?

Now take a look at the OpenGL Display window. As you can see, the majority of the aligned molecules have turned red. Notice that only 10 residues are strictly conserved (blue), and these areas are important for catalysis and dimerization.

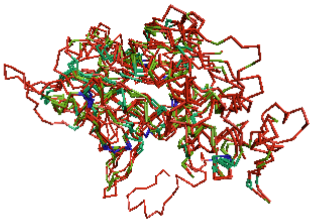
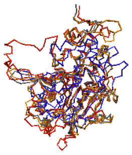


Figure 10: Sequence Conservation

The coloring of the molecules using Sequence Identity per residue indicates that the sequence conservation is much less in comparison to the structural conservation. This may be difficult to see right now. However, in Residue Selection sequence conservation is easier to observe, using the Sequence Display.



What does this all mean?. What impact does this comparison have on the evolution of AspRS? Does this mean that the structure of the protein molecule is more conserved than the sequence of amino acids? What does this comparison reveal about sequence and structure evolution in this group of the AARSs?

To examine the relationship between sequence and structure in more detail, we will use the Residue Selection feature.

4 Residue Selection

4.1 Starting up Residue Selection

The Residue Selection feature lets you analyze conservation, using different measures, and highlight residues in the Sequence Display and Structure Display simultaneously. Residue Selection allows you to examine the conservation on a per residue basis.

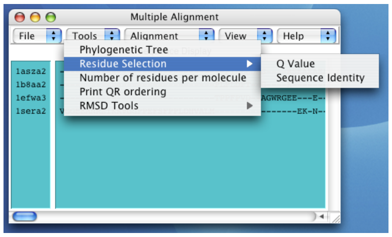


Figure 11: To access Residue Selection:

1 Click on Tools menu in the main Multiple Alignment window top pull-down menu.

2 Select Residue Selection.

From this pull-down menu item you can select Q Value or Sequence Identity. Each tool has its own window that allows you to select a a greater than or less than value between 0.0 to 1.0.

4.2 Exercise 1: Investigating Structure Conservation

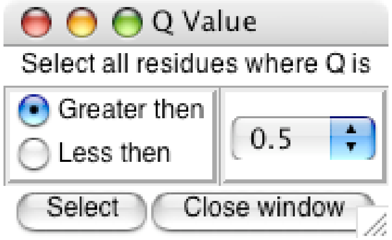
For this exercise, we will first examine structure conservation. Make sure Molecule Coloring is set to Q per residue.

1 In the main Multiple Alignment window top pull-down, select Tools → Residue Selection → Q Value.

2 A new window entitled Q Value will appear. Select Greater than

3 To the right, you can select a value between 0.0 to 1.0. Select 0.5 for a value.

4 Then, click on the Select button.


Select all residues where Q is

● Greater then
○ Less then

0.5

Figure 12: Exercise 1

Look in the OpenGL Display window and see what happens to the molecule. Since you selected the Q Value to be greater than or equal to 0.5, the majority of the molecule will be yellow. Also, the Sequence Display, in the Multiple Alignment window, will have the majority of the molecules highlighted. Why is this? By selecting the Q Value to be greater than or equal to 0.5, you have selected structural conservation at Q=0.5 or greater. In other words, you have demonstrated where the Q Value, structural conservation, is at a moderate to high level in all 4 aligned molecules - almost obscuring where high levels of conservation occur. Due to the high level of conservation, the Bond highlight style is not very informative.

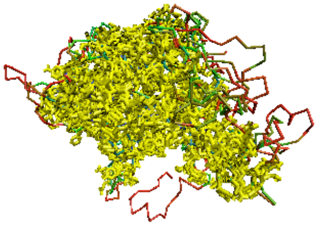


Figure 13: Q≥0.5 with Bonds Highlight Style

You can change this by accessing the main Multiple Alignment window top pull-down menu:

5 Clicking on View → Highlight Style.

6 A side menu will appear alongside Highlight Style. Select Trace for the highlight style.

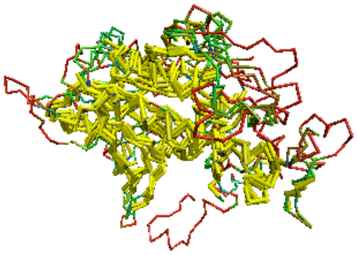


Figure 14: Q≥0.5 with Trace Highlight Style

As you can see, the areas with high levels of conservation are easily discerned using the Trace style. Now go back to the Q Value window and click on the Close Window button.

4.3 Exercise 2: Sequence Conservation

Before beginning this exercise, change the highlight style back to Bonds and the Molecule Coloring to Sequence Identity per residue. In order to display sequence conservation, follow these steps.

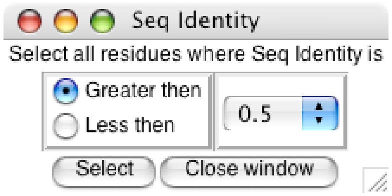
1 Go back to the main Multiple Alignment window and select Tools → Residue Selection.

2 Select Sequence Identity.

3 In the Sequence Identity window, set the measure to Greater than.

4 Select the value 0.5.

5 Click the Select button.


Seq Identity

Select all residues where Seq Identity is

● Greater then
○ Less then

0.5

Figure 15: Exercise 2

Do you notice a difference in the highlighted areas, in comparing the structure (Q) to sequence? Upon examining the residues in the Sequence display, notice how the majority of the sequence conservation occurs among the top 3 molecules. Since the top 3 molecules are AspRS and the bottom is SerRS, this pattern of sequence conservation makes sense.

Go back to the Sequence Identity window. Decrease the value to 0.2 and click Select. Take note of how many conserved residues are in the OpenGL Display and Sequence Display. Now, increase the value back to 0.5. You will see, as you progress, the yellow color diminishing among the molecules.

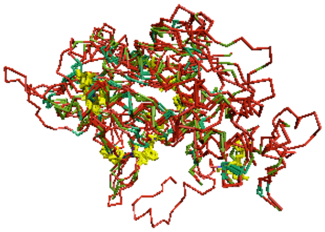
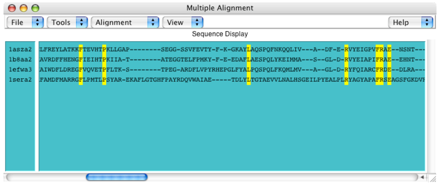


Figure 16: Sequence Identity≥0.7

The yellow coloring remains towards the core of the molecule and at the dimerization site. Increase the value to 0.7 and click Select; you should notice that the only highlighted sequences in the Sequence Display are those that match exactly. Finally, change the value to 1.0. Between Sequence Identity ≥0.7 and Sequence Identity ≤1.0, there are only 10 residues which are identical and span all four molecules. Go back to the Sequence Identity window and click on the Close Window button.



```
1asza2  LFREYLATKKFTEVHTPKLLGAP--------SEGG-SSVFEVTY-F-K-GKAYLAQSPQFNKQQLIV---A--DF-E-RVYEIGPVFRAE--NSNT---
1b8aa2  AVRDFFHENGFIEIHTPKIIA-T--------ATEGGTELFPMKY-F-E-EDAFLAESPQLYKEIMMA---S--GL-D-RVYEIAPIFRAE--EHNT---
1efwa3  AIWDFLDREGFVQVETPFLTK-S--------TPEG-ARDFLVPYRHEPGLFYALPQSPQLFKQMLMV---A--GL-D-RYFQIARCFRDE--DLRA---
1sera2  FAMDFMARRGFLPMTLPSYAR-EKAFLGTGHFPAYRDQVWAIAE-----TDLYLTGTAEVVLNALHSGEILPYEALPLRYAGYAPAFRSEAGSFGKDVF
```

Figure 17: Highlighted residues in Sequence Display

4.4 Exercise 3: The physical meaning of Q

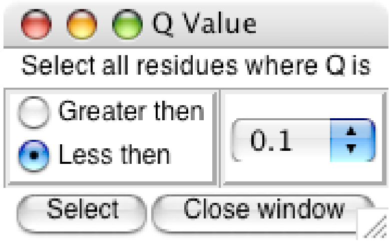
This exercise will further examine structure conservation. Make sure to change Molecule Coloring back to Q per residue.

1 Go back to the main Multiple Alignment window top pull-down menu and select Tools → Residue Selection.

2 Select Q Value.

3 Select Less than and 0.1 value.

4 Then click on the Select button.


Q Value

Select all residues where Q is

○ Greater then
● Less then

0.1

Figure 18: Exercise 3

If you had set Molecule Coloring to Q per residue you will notice that the majority of the red areas have turned yellow, and the blue is visible. If you look in the Sequence Display, you will see that the majority of SerRS residues have been highlighted. By using these settings in Q value, you have demonstrated where the aligned molecules have the least structural conservation,$Q \leq 0.1$.

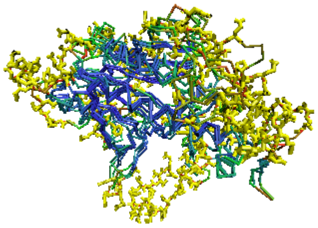


Figure 19: $Q \leq 0.1$

Now select Greater than or equal to and 0.6 in the Q Value window , and click on Select. Notice, in the OpenGL Display that the majority of the blue area inside the molecules has turned yellow. Also, note what has occurred in the sequence display. Many areas of the AspRS molecules are highlighted. Occasionally, an amino acid in SerRS is highlighted along with those of the AspRS. As you increase the value to 0.7, there are no highlighted residues from SerRS. Now that we have completed the exercises in Residue Selection, close the Q Value window.

4.5 Summary of Results

How does this analysis indicate which molecules are more closely related to each other? In Exercise 1, we evaluated structure conservation on a moderate to high level and noticed when we changed highlight style that the structural conservation is highest in the core. Sequence conservation, in Exercise 2, was also highest in the core of the molecule and at the dimerization site. We then further analyzed structure conservation during Exercise 3. We noticed that upon increasing the value of structure conservation, the highest structural conservation is among the three AspRSs. When the Q≥0.7, no residues is SerRS had Q≥0.7 with the three AspRS, indicating poor structure conservation. The three AspRS molecules are more closely related to each other in comparison to SerRS.

Having examined the relationship between the four AARSs using structure and sequence conservation, we will now use phylogenetic trees to display the evolutionary relationships graphically.

5 Phylogenetic Tree

5.1 Determining Structure-based Relationships

The Phylogenetic Tree feature in Multiple Alignment helps in determining the structure-based relationships between the four AARSs. To do this, it uses a modification of Q that accounts for both gapped and aligned regions. This new metric, Q_H, creates a structure-based phylogeny that is congruent to the sequence-based phylogenies of AARSs previously reported by Woese et al.[4] (also see Ref. 1).

What is Q_H?. Q_H is an adaptation of the traditional Q, and is essentially a *metric for structural homology*. Q_H is comprised of two terms: Qalign and Qgap (Ref. 1). Each time you align two structures, unless they align perfectly, the structure can be divided into 2 parts: the part that aligns (Qalign) and the part that does *not* align (Qgap), due to insertions and deletions. Qgap accounts for how much the insertion perturbs the core structure of the protein and the size of the insertion in sequence and structure. See Appendix B for more information.

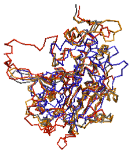
To utilize this feature:

1 Go to Tools → Phylogenetic Tree. A new window entitled Phylogenetic Tree will open.

[4] C.R. Woese, G. Olsen, M. Ibba, D. Soll. “Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process.” MMBR 64:202-236, 2000.

2 To display the structure-based Phylogenetic Tree with appropriate labelling, refer to figure 20.

3 To compute and display the tree, select by which measure - RMSD or QH - you would like to view the tree.



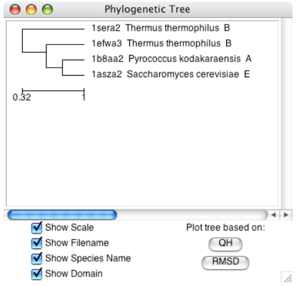
The Phylogenetic Tree. A phylogenetic tree is a dendogram, representing the succession of biological form by similarity-based clustering. Classical taxonomists use these methods to infer evolutionary relationships of multicellular organisms based on morphology. Molecular evolutionary studies use DNA, 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 Felsenstein.[a]

[a] J. Felsenstein *Inferring Phylogenies*. Sinauer Associates, Inc.: 2004.

For our first example, let's select all four options at the bottom of the window and use Q_H as our measure to generate the structure-based relationships.

Can you guess as to how the protein molecules will fall relatively to one another on the tree, considering where they fall on the Phylogenetic tree?
If you guessed that the SerRS would be the furthest out and that the *P.kodakaraensis* AspRS and *S.cerevisiae* AspRS would be the closest, you are correct. Note that *P.kodakaraensis* is from the Archaea branch of the phylogenetic tree and *S.Cerevisiae* is from the Eucarya. The SerRS is used as an outlier case for this alignment, as demonstrated in Residue Selection.

To continue to the next section of the tutorial, close the Phylogenetic Tree window.



1sera2 Thermus thermophilus B
1efwa3 Thermus thermophilus B
1b8aa2 Pyrococcus kodakaraensis A
1asza2 Saccharomyces cerevisiae E

☑ Show Scale
☑ Show Filename
☑ Show Species Name
☑ Show Domain

Figure 20: Example of structure-based phylogenetic tree.

6 Investigating Structural Alignment

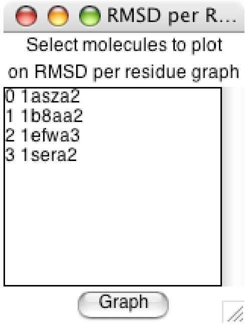
6.1 RMSD per Residue

The ability to graphically display the RMSD per Residue between two proteins is a useful feature for demonstrating how well the proteins align. The root mean square deviation (RMSD) measures the distances in angstroms between the C-alpha atoms of 2 aligned residues.
To begin RMSD per Residue:

1 Go to Tools → RMSD Tools → RMSD per Residue. A window called RMSD per Residue will pop up with the four proteins listed.

2 Highlight two, three, or four proteins you want to compare by simply clicking on each one.

3 Click the Graph button. Another window, RMSD Per Residue, should appear that displays the graph of the plots.


RMSD per R...

Select molecules to plot
on RMSD per residue graph

0 1asza2
1 1b8aa2
2 1efwa3
3 1sera2

Figure 21: Dialog

If you select two proteins, there will be one line on the graph. If you select three, there will be two lines differentiated by color on the graph.

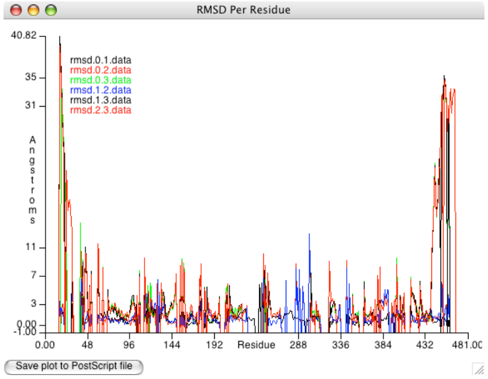


Figure 22: This graph plots the RMSD between all four molecules. Residues in gaps are assigned an RMSD value of -1.

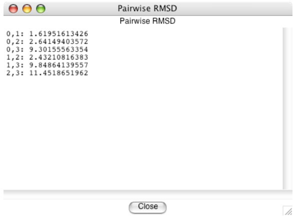
Now that we have examined RMSD per Residue, close both windows.

6.2 Pairwise RMSD

Pairwise RMSD prints the average overall RMSD for each pair of aligned proteins. To find the Pairwise RMSD of each molecule in relation to the other:

1 In the main Multiple Alignment window, select Tools → RMSD Tools.

2 Then select Pairwise RMSD in the side menu. A new window called Pairwise RMSD will appear printing the results of the RMSD calculations.



```
0,1: 1.61951613426
0,2: 2.64149403572
0,3: 9.30155563354
1,2: 2.43210816383
1,3: 9.84864139557
2,3: 11.4518651962
```

Figure 23: Printout of results.

To continue to the next section, close the Pairwise RMSD window.

7 Exporting File Formats

7.1 Exporting a PDB file from user-specified selections

As you walk through this tutorial, notice many of the new images you've created in the OpenGL Display. Often these images are generated by highlighting specific portions of the aligned protein sequences. If you would like to study your selections further, you can can do so by generating your own PDB file(s). To begin this process:

1 Highlight the portions of the sequence that you want to examine in the Sequence Display of the main Multiple Alignment window.

2 In the main Multiple Alignment window top pull-down menu, go to View → Highlight style. Multiple Highlight styles will appear to choose from. Select one and make sure it appears in the OpenGL Display.

3 Go back to the Multiple Alignment window and click on File → Write PDB from selection....

4 The PDB file(s) will be saved in a directory that can be chosen by clicking on the File → Choose Work Directory.... If you haven't selected your Work directory, you be prompted to choose a directory when you click on Write PDB from selection....

5 If you want to save the entire alignment, you can do so by going the Residue Selection and selecting Q Value, Greater than, and 0.0. Then click on the Select button in the Q Value window. The four sets of sequences will be highlighted. Go to File → Write PDB from selection... and four PDB files will be generated.

These molecules can then be loaded to VMD and Multiple Alignment for further analysis.

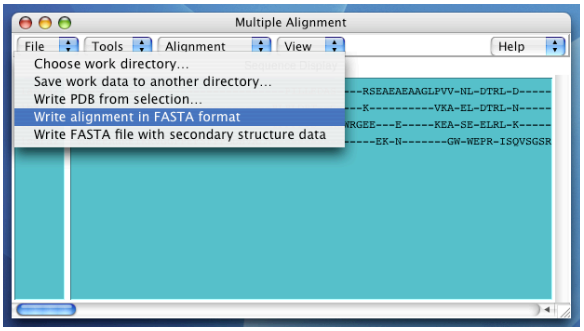
7.2 Exporting a multiple alignment in FASTA format

You can create a FASTA format file from the aligned molecules. To do this:

1 Go to the main Multiple Alignment window and select File → Write alignment in FASTA format.

2 A browser window should appear where you can save the file. Select where you want to save the FASTA file. You may also change the name of the file in the provided dialog.

3 Save the file by hitting either a Save or OK button, or by hitting the Enter/Return key.


Multiple Alignment
File
Tools
Alignment
View
Help
Choose work directory...
Save work data to another directory...
Write PDB from selection...
Write alignment in FASTA format
Write FASTA file with secondary structure data
Sequence Display

```
---RSEAEAEAAGLPVV-NL-DTRL-D-----
---K----------VKA-EL-DTRL-N-----
WRGEE---E-----KEA-SE-ELRL-K-----
-----EK-N-------GW-WEPR-ISQVSGSR
```

Figure 24: FASTA menu

You also have the option of writing the secondary structure data into a FASTA file.
To do this:

4 Select File → Write FASTA file with secondary structure data.

5 A file browser window will appear. Select where you want to save the FASTA file. You also have the option of changing the file name in the browser window.

6 Click either the Save or OK button, or hit the Enter or Return key to save the FASTA files.

FASTA is a common alignment file, which enables you to study multiple alignments in many other applications.

8 Appendices

8.1 Appendix A: Q

The following equation is from the article “Evaluationg protein structure-prediction schemes using energy landscape theory” by Eastwood, et al.

$$Q = \frac{2}{(N-1)(N-2)} \sum_{i<j-1} \exp\left[-\frac{\left(r_{ij} - r_{ij}^N\right)^2}{2\sigma_{ij}^2}\right]$$

r_{ij} is the distance between a pair of C^α atoms.

r_{ij}^N is the C^α-C^α distance between residues i and j in the native state.

$\sigma_{ij}^2 = |i-j|^{0.15}$ is the standard deviation, determining the width of the Gaussian function.

N is the number of residues of the protein being considered.

8.2 Appendix B: Q_H

The following text is in the article “On the evolution of structure in aminoacyl-tRNA synthetases.” by O’Donoghue et al.

Homology Measure

We employ a structural homology measure which is based on the structural similarity measure, Q, developed by Wolynes, Luthey-Schulten, and coworkers in the field of protein folding. Our adaptation of Q is referred to as Q_H, and the measure is designed to include the effects of the gaps on the aligned portion: Q_H=$\aleph$(q_{aln}+q_{gap}), where $\aleph$ is the normalization, specifically given below. Q_H is composed of two components. q_{aln} is identical in form to the unnormalized Q measure of Eastwood et al. and accounts for the structurally aligned regions. The q_{gap} term accounts for the structural deviations induced by insertions in each protein in an aligned pair:

$$Q_H = \aleph\left[q_{aln} + q_{gap}\right]$$

$$q_{aln} = \sum_{i<j-2} \exp\left[-\frac{\left(r_{ij} - r_{i'j'}\right)^2}{2\sigma_{ij}^2}\right]$$

$$\begin{aligned} q_{gap} &= \sum_{g_a}\sum_{j}^{N_{aln}} \max\left\{\exp\left[-\frac{\left(r_{g_a j} - r_{g'_a j'}\right)^2}{2\sigma_{g_a j}^2}\right], \exp\left[-\frac{\left(r_{g_a j} - r_{g''_a j'}\right)^2}{2\sigma_{g_a j}^2}\right]\right\} \\ &+ \sum_{g_b}\sum_{j}^{N_{aln}} \max\left\{\exp\left[-\frac{\left(r_{g_b j} - r_{g'_b j'}\right)^2}{2\sigma_{g_b j}^2}\right], \exp\left[-\frac{\left(r_{g_b j} - r_{g''_b j'}\right)^2}{2\sigma_{g_b j}^2}\right]\right\} \end{aligned}$$

The first term, q_{aln}, computes the unnormalized fraction of C^α-C^α pair distances that are the same or similar between two aligned structures. r_{ij} is the spatial C^α-C^α distance between residues i and j in protein a, and $r_{i'j'}$ is the C^α-C^α distance between residues i’ and j’ in protein b. This term is restricted to aligned positions, e.g., where i is aligned to i’ and j is aligned to j’. The remaining terms account for the residues in gaps. g_a and g_b are the residues in insertions in both proteins, respectively. g'_a and g''_a are the aligned residues on either side of the insertion in protein a. The definition is analogous for g'_b and g''_b.

The normalization and the σ_{ij}^2 terms are computed as:

$$\aleph = \frac{1}{\frac{1}{2}\left(N_{aln} - 1\right)\left(N_{aln} - 2\right) + N_{aln}N_{gr} - n_{gaps} - 2n_{cgaps}}$$

$$\sigma_{ij}^2 = |i - j|^{0.15}$$

where N_{aln} is the number of aligned residues. N_{gr} is the number of residues appearing in gaps, and n_{gaps} is sum of the number of insertions in protein “a”, the number of insertions in protein “b” and the number of simultaneous insertions (referred to as bulges or c-gaps). n_{cgaps} is the number of c-gaps. Gap-to-gap contacts and intra-gap contacts do not enter into the computation, and terminal gaps are also ignored. σ_{ij}^2 is a slowly growing function of sequence separation of residues i and j, and this serves to stretch the spatial tolerance of similar contacts at larger sequence separations. Q_H ranges from 0 to 1 where $Q_H = 1$ refers to identical proteins. If there are no gaps in the alignment, then Q_H becomes $Q_{aln} = \aleph q_{aln}$, which is identical to the Q-measure described into the Q measure described before.