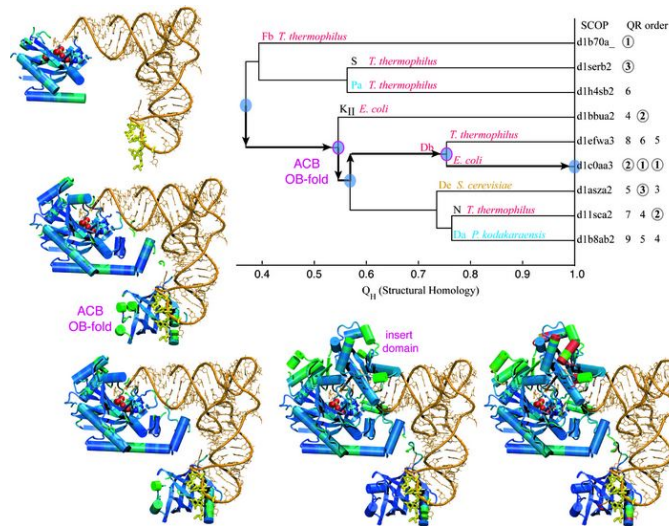


University of Illinois at Urbana-Champaign  
 Luthey-Schulten Group  
 Theoretical and Computational Biophysics Group

# Evolution of Biomolecular Structure

## Class II tRNA-Synthetases and tRNA



MultiSeq Developers:

Elijah Roberts

John Eargle

Dan Wright

Prof. Zan Luthey-Schulten

Patrick O'Donoghue

Anurag Sethi

Brijet Dhaliwal

September 25, 2006.

A current version of this tutorial is available at  
<http://www.scs.uiuc.edu/~schulten/tutorials/evolution/>

## Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	The MultiSeq Bioinformatic Analysis Environment . . . . .	4
1.2	Aminoacyl-tRNA Synthetases: Role in translation . . . . .	4
1.3	Getting Started . . . . .	7
1.3.1	Requirements . . . . .	7
1.3.2	Copying the tutorial files . . . . .	7
1.3.3	Configuring MultiSeq . . . . .	7
1.3.4	Configuring BLAST for MultiSeq . . . . .	10
1.4	The Aspartyl-tRNA Synthetase/tRNA Complex . . . . .	12
1.4.1	Loading the structure into MultiSeq . . . . .	12
1.4.2	Selecting and highlighting residues . . . . .	13
1.4.3	Domain organization of the synthetase . . . . .	14
1.4.4	Nearest neighbor contacts . . . . .	14
<b>2</b>	<b>Evolutionary Analysis of AARS Structures</b>	<b>17</b>
2.1	Loading Molecules . . . . .	17
2.2	Multiple Structure Alignments . . . . .	18
2.3	Structural Conservation Measure: $Q_{res}$ . . . . .	19
2.4	Structure Based Phylogenetic Analysis . . . . .	21
2.4.1	Limitations of sequence data . . . . .	21
2.4.2	Structural metrics look further back in time . . . . .	23
<b>3</b>	<b>Complete Evolutionary Profile of AspRS</b>	<b>26</b>
3.1	BLASTing Sequence Databases . . . . .	26
3.1.1	Importing the archaeal sequences . . . . .	26
3.1.2	Now the other domains of life . . . . .	27
3.2	Organizing Your Data . . . . .	28
3.3	Finding a Structural Domain in a Sequence . . . . .	29
3.4	Aligning to a Structural Profile using ClustalW . . . . .	30
3.5	Eliminating Redundancy with Sequence QR . . . . .	32
3.6	Phylogenetic Tree of an Evolutionary Profile . . . . .	33
3.7	Export Data . . . . .	35
3.8	MultiSeq Sessions . . . . .	35
<b>4</b>	<b>Evolutionary Analysis of tRNA</b>	<b>36</b>
4.1	tRNA and Modified Bases . . . . .	36
4.2	Structural Alignment . . . . .	38
4.3	Alignment Editing . . . . .	39
4.4	Sequence Alignment . . . . .	39
4.5	Sequence Tree . . . . .	41

<b>5</b>	<b>Appendices</b>	<b>43</b>
5.1	Appendix A: $Q$ . . . . .	43
5.2	Appendix B: $Q_H$ . . . . .	44
5.3	Appendix C: $Q_{res}$ Structural Similarity per Residue . . . . .	46

## 1 Introduction

### 1.1 The MultiSeq Bioinformatic Analysis Environment

The MultiSeq extension to VMD allows researchers to study the evolutionary changes in sequence and structure of biomolecules across all three domains of life, from bacteria to humans. The comparative sequence and structure metrics and analysis tools introduced in the accompanying articles by the Luthey-Schulten Group<sup>123</sup> are included in MultiSeq. Of particular note is the inclusion of a recently developed structure-based measure of homology,  $Q_H$  (see Appendix B), that accounts for the effect of insertions and deletions and has been shown to produce accurate structure-based phylogenetic trees. MultiSeq also includes or allows for the easy integration of several popular bioinformatics programs, including the STAMP structural alignment tool, kindly provided by our colleagues Russell and Barton<sup>4</sup>, BLAST, and ClustalW. Our goal is to offer researchers a complete and user friendly tool for examining the changes in protein sequence and structure in the correct framework, evolution. As a result, MultiSeq is an invaluable tool for relating protein structure to function.

This tutorial showcases the MultiSeq environment and will allow the reader to combine sequence and structure information into evolutionary profiles similar to those in the accompanying articles<sup>123</sup>. Evolutionary profiles are compact representative sets that can be used for both gene annotation and energetic analysis. The tutorial is designed such that it can be used by both new and experienced 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 should take about three hours to complete in its entirety.*

### 1.2 Aminoacyl-tRNA Synthetases: Role in translation

Before beginning the actual tutorial, a small amount of background information on the cellular translation system may be helpful. The aminoacyl-tRNA synthetases (AARSs) are key proteins involved in setting the genetic code in all living organisms and are found in all three domains of life Bacteria (B), Arachea (A), and Eukarya (E). The essential process of protein synthesis requires twenty sets of synthetases and their corresponding tRNAs for the correct transmission of the genetic information. The AARSs are responsible for loading the twenty

---

<sup>1</sup>P. O’Donoghue and Z. Luthey-Schulten. “Evolution of structure in aminoacyl-tRNA synthetases” MMBR, 67:550-73. 2003.

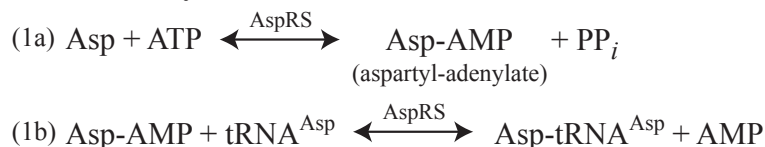
<sup>2</sup>P. O’Donoghue and Z. Luthey-Schulten. “Evolutionary profiles derived from the QR factorization of multiple structural alignments gives an economy of information” JMB, 346:875-94. 2005.

<sup>3</sup>A. Sethi, P. O’Donoghue, and Z. Luthey-Schulten. “Evolutionary profiles from the QR factorization of multiple sequence alignments” PNAS, 102:4045-50. 2005.

<sup>4</sup>R.B. Russell and G.J. Barton. “Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels.” Proteins, 14:309-23. 1992.

different amino acids onto their cognate tRNA (tRNA containing the appropriate anticodon) during protein synthesis. The synthesis (See Figure 1) of an aminoacyl-tRNA (aa-tRNA) occurs by either direct acylation or an indirect mechanism in which the amino acid or amino acid precursor in the misacylated tRNA is modified in a second step. These indirect pathways suggest interesting evolutionary links between amino acid biosynthesis and protein synthesis<sup>56</sup>.

### Direct Pathway



### Indirect Pathway

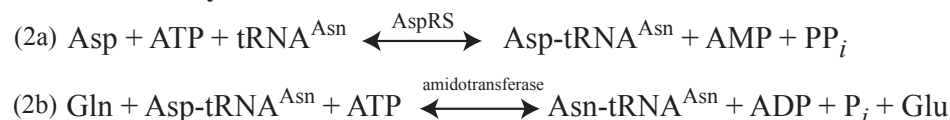


Figure 1: (1) The two step direct acylation of tRNA by aspartyl-tRNA synthetase. (a) The aspartate is first combined with an ATP molecule to form an “activated” aspartyl-adenylate and then (b) the adenylate reacts with the tRNA to form the “charged” aspartyl-tRNA. (2) The indirect mechanism for charging the tRNA. (a) The tRNA<sup>Asn</sup> is mischarged with an aspartate which is then (b) converted to an asparagine by an amidotransferase.

Each AARS is a multidomain protein consisting of (at least) a catalytic domain and an anticodon binding domain. In all known cases, the synthetases can be divided into two types: class I or class II. 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 the bacterial aspartyl-tRNA synthetase, have an “insert domain” within their catalytic domain (see Figure 2). The tRNA is charged in the catalytic domain and recognition of it takes place through interactions with the anticodon loop, acceptor stem, and D-arm of the tRNA (see Figure 17). In the first part of the tutorial we will examine the evolution of the structure and sequences of the AARSs and in the second part, provide a cursory evolutionary analysis of the tRNA and its recognition elements.

<sup>5</sup>P. O’Donoghue, A. Sethi, C. R. Woese, and Z. Luthey-Schulten. “The evolutionary history of Cys-tRNA<sup>Cys</sup> formation” PNAS, 102:4045-50. 2005.

<sup>6</sup>A. Sauerwald, W. Zhu, T. Major, H. Roy, S. Palioura, D. Jahn, W. Whitman, J. Yates, M. Ibbá, and D. Soll. “RNA-Dependent Cysteine Biosynthesis in Archaea” Science, 307:1969-72. 2005.

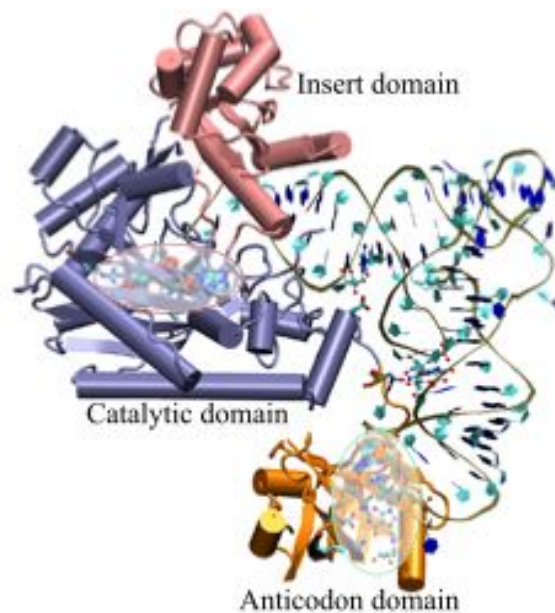


Figure 2: A snapshot of AspRS/tRNA aspartyl-adenylate complex (from *E. coli*; PDB code 1c0a) in the active form. Note the anticodon binding domain (orange), the insertion domain (pink), and the catalytic domain (blue). The tRNA is docked to AspRS, and the catalytic active site is highlighted within the catalytic domain (red oval); 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 oval). Note that specific contacts between the tRNA and Asp/RS allow for strategic positioning of the tRNA relative to the enzyme.

## 1.3 Getting Started

### 1.3.1 Requirements

MultiSeq must be correctly installed and configured before you can begin using it to analyze the evolution of protein structure. This section walks you through the process of doing so, but there are a few prerequisites that must be met before this section can be started:

- VMD 1.8.5 or later must be installed. The latest version of VMD can be obtained from <http://www.ks.uiuc.edu/Research/vmd/>
- This tutorial requires approximately 340 MB of free space on your local hard disk. MultiSeq requires about 200 MB of free space for metadata databases.

### 1.3.2 Copying the tutorial files

This tutorial requires certain files, which are available in the following directory on the tutorial CD:

```
/Tutorials/Evolution_of_Protein_Structure/tutorial-files/
```

or in the compressed file available for download from the tutorial website.

You should copy this entire directory to a location on your local hard disk. The path to the directory *must not* contain any spaces. For the remainder of this tutorial, this directory on your local drive will be referred to as `TUTORIAL_DIR`.

### 1.3.3 Configuring MultiSeq

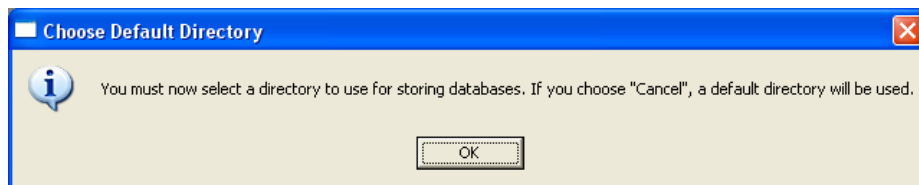
MultiSeq saves user preferences in a file named `.multiseqrc` located in your home directory. The preferences saved include the location of any local databases, previous search options, and others. When you start MultiSeq for the first time, it will ask you if it is ok to create this file and to specify the directory in which to look for any metadata databases.



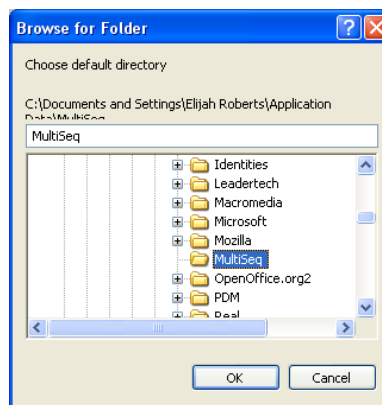
**What is metadata?** Metadata is a term meaning “data about data”. In MultiSeq the word metadata refers to information about the sequences or structures loaded into the program. MultiSeq knows how to find certain types of sequences or structures in the public metadata databases and can display information from them such as the species from which the protein originated, the taxonomic lineage of the organism, the protein’s enzymatic properties, and even how to find the protein in other databases. You’ll learn more about how this can be helpful later in the tutorial.

Follow these steps to configure MultiSeq:

1. Launch VMD.
2. Within the VMD main window, choose the Extensions menu.
3. In the Extensions menu select Analysis → MultiSeq.
4. MultiSeq will notify you that you must select a directory in which to store metadata databases. Press the OK button.

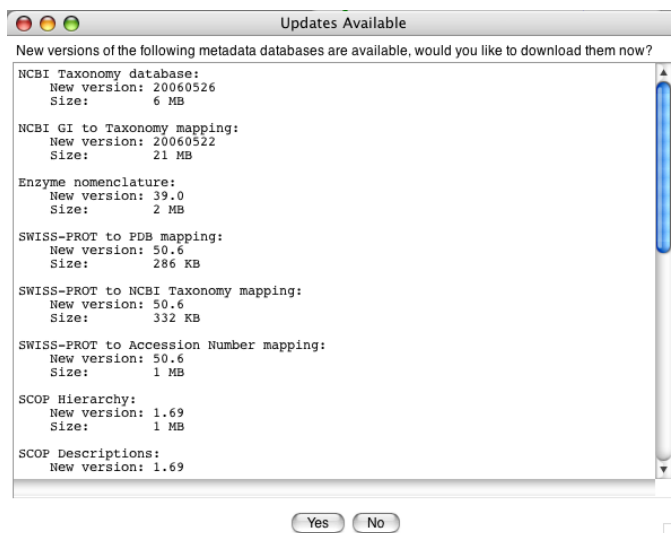


5. You will then be prompted to select the metadata directory. If the directory already contains the metadata databases, MultiSeq will use them. If not, MultiSeq will download them into the directory. If you are following this tutorial from a CD, choose the TUTORIAL\_DIR/multiseqdb directory in the dialog and press the OK button. If you are following from the Internet, select the directory where you would like MultiSeq to store the databases and press the OK button.

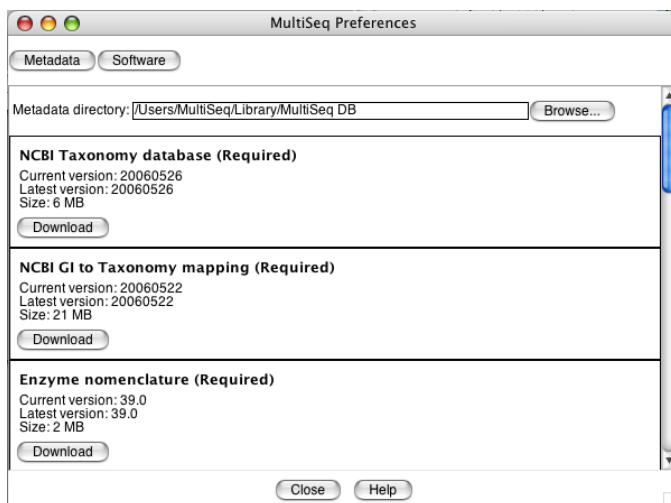




6. If updates to the metadata databases are available, MultiSeq will present a dialog showing the available updates and give you the option of downloading them. Press the **Yes** button to download the updates. MultiSeq will ask you to wait while the updates are downloaded, which may take a few minutes depending on the size of the updates and the speed of your connection.



7. The MultiSeq Preferences dialog will then appear showing the metadata directory and the currently installed databases. Press the **Close** button to save these preferences.



8. The MultiSeq program window will then appear on the screen. The rest of the tutorial and exercises will use features from this window, unless otherwise specified.

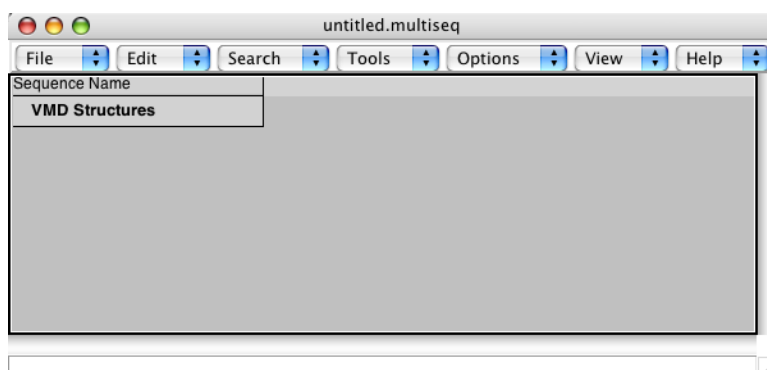



Figure 3: The MultiSeq program window

### 1.3.4 Configuring BLAST for MultiSeq

MultiSeq is now minimally configured. For the purposes of this tutorial, however, some additional functionality is needed. Specifically, the tutorial uses BLAST to perform sequences searches, requiring that a local version of BLAST be installed.



**What is BLAST and why do I need to install it?**  
BLAST is a software tool available from the NCBI (<http://www.ncbi.nlm.nih.gov/BLAST/>) that allows you to search through a database of sequences and find those that are similar to a query sequence or profile of sequences. BLAST allows for very rapid searching through a large number of sequences and is widely used in the bioinformatics community. BLAST is typically used via one of two methods: online search or local installation. An online search is very simple and requires nothing more than for a user to paste a query sequence into a web page, but the utility of such a search is somewhat limited. MultiSeq requires a local BLAST installation because it provides additional functionality to the user not available through an online search.

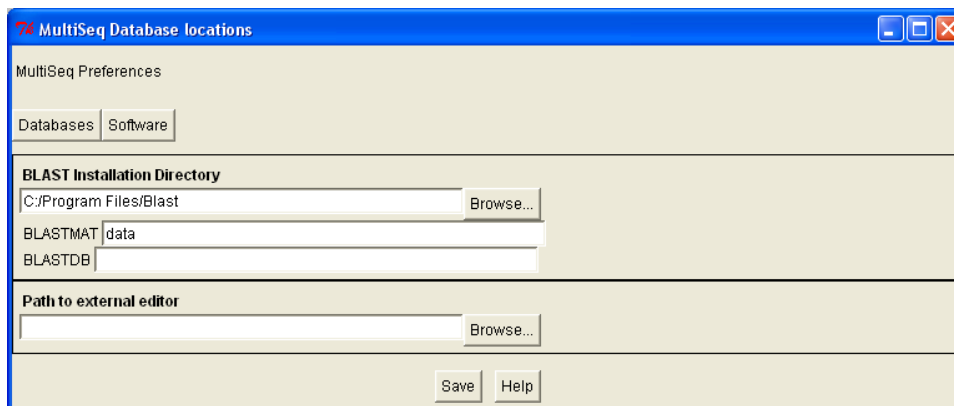
Follow these steps to install a local copy of BLAST:

1. Create a directory on your local hard disk into which BLAST will be installed. Recommended directories are:
  - Unix/Linux: `/usr/local/blast`

- Mac OS X: /Applications/Blast
  - Windows: C:\Blast
2. Archives of the BLAST installation for each of the supported platforms are located on the tutorial CD in the directory:  

```
/Tutorials/Evolution_of_Protein_Structure/blast-install/
```

or in the compressed file available for download from the tutorial website.  
  
Copy the BLAST archive file corresponding to your platform into the directory created in the previous step.
  3. Extract the archive. On Unix/Linux, use a command such as `tar zxvf filename`. On Mac OS X and Windows, the archive is a self-extracting executable, so just double-click on it.
  4. Next, you must set the BLAST installation location in MultiSeq. From the MultiSeq program window, choose `File → Preferences...` to bring up the preferences dialog.
  5. Click on the **Software** button in the upper left portion of the dialog to show the software preferences.
  6. Click on the **Browse...** button in the **BLAST Installation Directory** section and select the directory into which you installed BLAST. *Note: on Linux and Mac OS X you may have a directory called `blast-2.2.13` underneath your installation directory. If so, pick this directory in the browse dialog.*



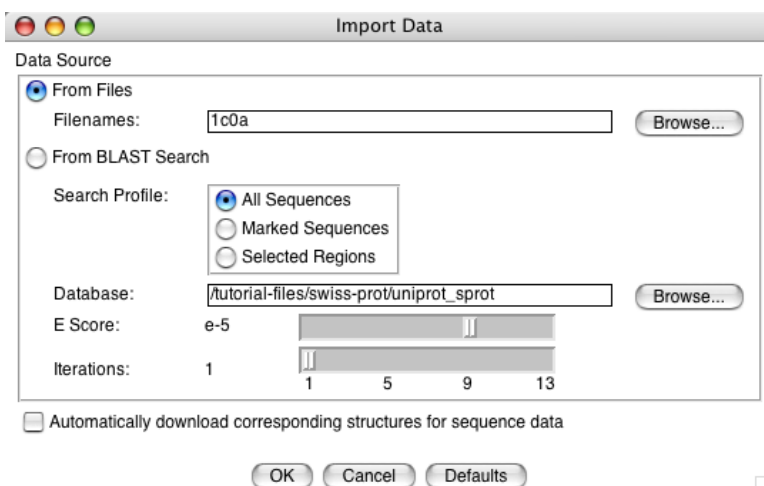
7. Press the **Close** button to save these changes. MultiSeq is now configured to use your local installation of BLAST.

## 1.4 The Aspartyl-tRNA Synthetase/tRNA Complex

### 1.4.1 Loading the structure into MultiSeq

In order to become familiar with the structural and functional features of the AARSs, we will first explore the aspartyl-tRNA synthetase (AspRS) as complexed with aspartyl-adenylate and tRNA (PDB code: 1c0a). To do this:

1. If MultiSeq is not running, start it from within VMD by choosing the Extensions menu and then selecting Analysis → MultiSeq. The MultiSeq program window will appear on your screen.
2. Choose the File menu and select Import Data.... The Import Data dialog will appear.
3. Make sure the From Files radio button is marked and in the Filenames field enter the PDB code “1c0a”. Click the OK button to have MultiSeq download the structure from the PDB website. If you do not have Internet access, you can also click on the Browse... button and select the file from your local tutorial directory at TUTORIAL\_DIR/1c0a.pdb.



**Loading multiple structures.** When performing an evolutionary analysis, it is common to load numerous structures. MultiSeq makes this easy by allowing you to select multiple files from your hard disk when using the Browse... button on the Import Data dialog. You can also have MultiSeq download multiple structures from the PDB by entering them into the Filenames field separated by commas, e.g. “1c0a,1asy,1b8a”. In addition to PDB structures, MultiSeq allows you to download structures directly from the Astral database by entering their SCOP domains. You’ll learn more about Astral and SCOP later in the tutorial.



an analysis. We'll see how to use them later on.

Now try selecting a larger region by clicking a residue and dragging the mouse in the MultiSeq program window. You can also highlight regions in MultiSeq by holding down the **Shift** and **Control** keys while clicking with the mouse, as you would in any other GUI program. These operations are called **Shift** clicking and **Control** clicking and will be useful throughout the tutorial. One additional thing to note is that you can change the color that is used to highlight your selection in the Open GL display. Try doing so by selecting the **View** → **Highlight Color** → **Name** menu option. Now each atom is colored according to its name. This coloring method can be very helpful when looking at specific atomic level interactions between residues, such as hydrogen bonds.

### 1.4.3 Domain organization of the synthetase

All of the AARSs are multidomain proteins, but the exact number and fold of each domain is specific to each synthetase. AspRS has a catalytic domain (comprised of residues 113–287 and 421–585), an anticodon binding domain (residues 1–112, also referred to as the N-terminal domain), and an insertion domain (residues 288–420). Curiously, the insertion domain interrupts the sequence of the catalytic domain and only appears in the bacterial AspRS; archaeal and eukaryal AspRSs do not contain this insertion. Try selecting each domain one at a time. You can select two non-contiguous regions in MultiSeq by clicking the first residue of the first region, **Shift** clicking the last residue of the first region, **Control** clicking the first residue of the second region, and finally **Control-Shift** clicking the last residue of the second region.

The anticodon for aspartate is comprised of Q634, U635, and C636. Note how the N-terminal domain of the enzyme attaches itself to the anticodon in the tRNA; zoom in on the anticodon. Q stands for queuosine and is a hypermodified base that marks the first position of the anticodon in the AARSs that code for Asp, Asn, His, and Tyr in Bacteria (see Genetic code in Figure 18). Select the anticodon in MultiSeq. Do you notice anything different about these bases? You will examine the tRNA in more detail in Section 4.

### 1.4.4 Nearest neighbor contacts

When analyzing protein structures, it is often desirable to know what residues are in contact with each other. Here we will identify what contacts the anticodon makes with the synthetase. To make this process easier, MultiSeq provides a function that allows you search for residues in contact with a selected region. Since we want to find residues in the synthetase, click the checkbox to the left of the name of sequence **1c0a\_A**. The sequence should appear checked as shown below.

Sequence Name	20	30	40
<b>VMD Structures</b>			
<input checked="" type="checkbox"/> 1c0a_A	20	L C G W V N R R R D L G S L I F I D M R D R E G I V Q V F	
<input type="checkbox"/> 1c0a_B	620	D D A G A A U A C C U G C C U ? U C A C G C A G G G G M U	

This is called marking a sequence; any number of sequences can be marked in MultiSeq at the same time. MultiSeq allows you to limit the scope of many operations to sequences that are marked. Now, with the three anticodon bases highlighted in MultiSeq, select the Search → Select Contact Shells menu option. The Select Contact Shells dialog will appear. Change the scope of the search to be only the marked sequences by selecting the Marked Sequences radio button, change the contact distance to be 3.0 Å, and then and press the OK button.

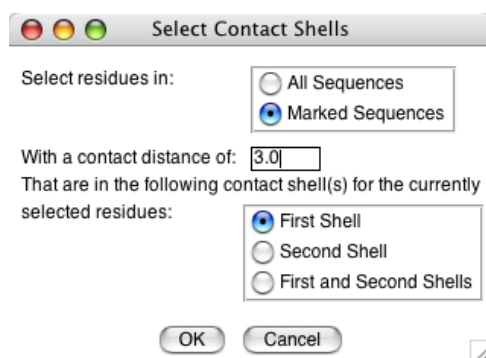
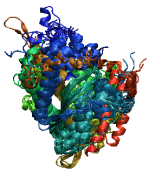


Figure 5: Select Contacts Shells dialog

The residues of the protein that are within 3.0 Å of the anticodon are selected in both the MultiSeq window and the Open GL display, as shown in Figure 6. Notice the classic base stacking interaction between residue F35 of the synthetase and bases U635 and C636 of the tRNA. What other types of interactions between the protein and tRNA can you recognize?



**The chemistry of AARSs** Explore the active site of the AspRS-tRNA complex in a similar way to what you did above for the anticodon region and answer the following questions: What step of the reaction shown in Figure 1 does this structure represent? What are the substrates? What products are synthesized by this reaction? What part of the tRNA is involved in this reaction? What part of the protein is involved?

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 stem of the tRNA extends into the active site of the aspartyl synthetase.

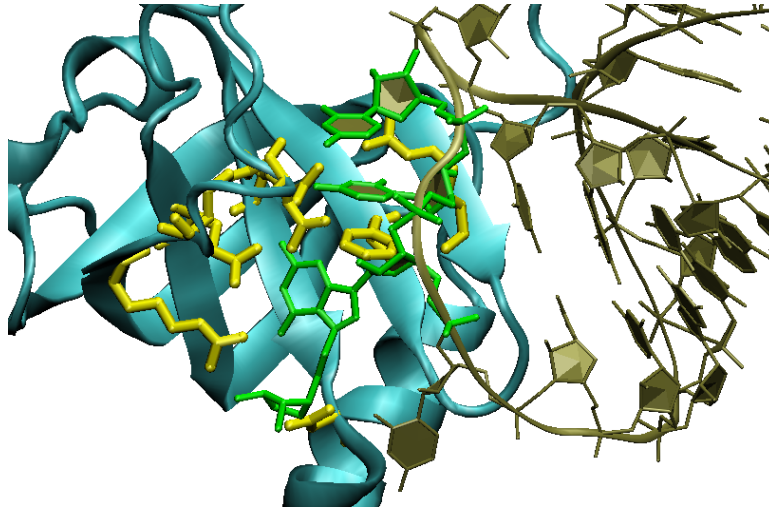
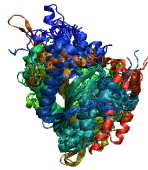


Figure 6: Residues of AspRS (yellow) within 3.0 Å of the anticodon (green)

Select the third residue in chain X. 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 the adenylate complex 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 in the 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 selecting the File → New Session menu option.*



## 2 Evolutionary Analysis of AARS Structures

In this part of the tutorial, we will use MultiSeq to align the catalytic domains of 16 class II AARS structures, representing 9 different specificities from each domain of life. The catalytic domain of each structure has been directly extracted from the ASTRAL database, which contains the structures of each of the proteins' domains. This part of the 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. A structural phylogenetic tree allows examination of more distant evolutionary events where specificity was being acquired. A sequence tree, on the other hand, addresses the more recent evolutionary history of a protein. We use as a reference for all trees the universal tree developed by Carl Woese using 16S ribosomal RNAs (Figure 7).

### Phylogenetic Tree of Life

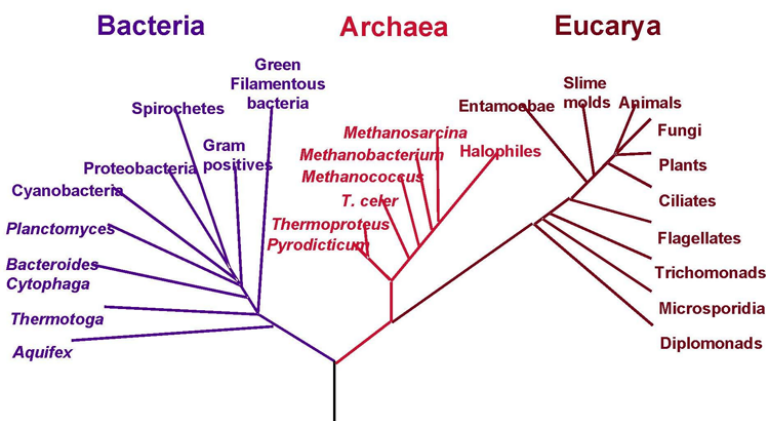


Figure 7: Universal tree of life

### 2.1 Loading Molecules

To further explore AARSs, we will now examine the catalytic domain of 16 AspRS structures in MultiSeq. Before we begin, make sure you have deleted any molecules in the MultiSeq program window.

1. Select the File→Import Data item in the Main MultiSeq window. We will be importing Data From Files. Make sure From Files is selected.
2. Hit the Browse button. A file browser window will appear. Navigate the file browser to the TUTORIAL\_DIR/class-2-synthetases directory.

3. There are 16 PDB files you want to load from the directory. Select all of the files by clicking on the first file with your mouse and holding down the shift key and then selecting the last file. You may need to change the filename filter to allow for selection of PDB files<sup>7</sup>.
4. Hit the OK button in the file browser window.
5. Notice that all of the file names will appear in the field Filenames. If this looks correct hit the OK button at the bottom of the Import Data dialog.

Since there are several files, it will take VMD about at least a minute to fully load the molecules. Once the molecules are in VMD and MultiSeq, you will see a 3D representation in the OpenGL display and sequence information in the Sequence Display of the main MultiSeq window.

Within the OpenGL display window, the 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 defined by SCOP. 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 code, the chain name, and a number, which corresponds to the structural domain. For example, the anticodon binding domain for the AspRS-tRNA complex we have been investigating is: d1c0aa1.

## 2.2 Multiple Structure Alignments

Next we will structurally align the molecules:

1. Go to the MultiSeq program window and select Tools in the top pull-down menu.
2. Then click on Stamp Structural Alignment. A new window entitled Stamp Alignment Options will appear with default settings (see Figure 8).

Perform the alignment by hitting the OK button. Once this step is complete, you will be able to view the structural alignment in both the OpenGL Display window and the main MultiSeq Window.

If you would like more information about STAMP parameters, please refer to the STAMP manual.<sup>8</sup>

<sup>7</sup>Note these commands for selecting all of the PDB files may differ on various operating systems. Select all of the files as appropriate for your operating system.

<sup>8</sup>The STAMP manual is available at <http://www.compbio.dundee.ac.uk/manuals/stamp.4.2/>

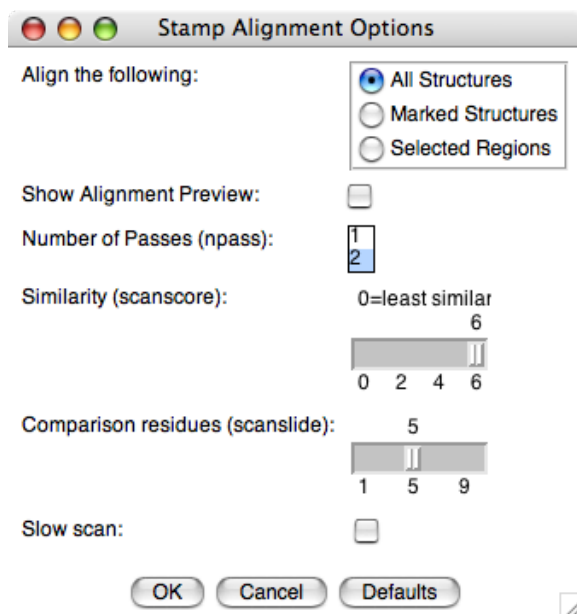



Figure 8: Stamp Alignment Options Window



**How molecules are aligned in a multiple structural alignment**

MultiSeq uses the program STAMP to align protein molecules. The STAMP algorithm minimizes the  $C_{\alpha}$  distance between aligned residues of each molecule by applying globally optimal rigid-body rotations and translations. Also, note that you can only 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).

### 2.3 Structural Conservation Measure: $Q_{res}$

MultiSeq features various coloring metrics for protein analysis. When applied to structures, the coloring is displayed in both the OpenGL display and the main MultiSeq window.  $Q_{res}$  is the coloring metric for structure similarity in multiple alignment of structures. Determining structure conservation is one method in evolutionary analysis that helps us understand what regions of a protein, or in this case what structural elements of the catalytic domain of AARSs, are conserved across all specificities.



**What is Qres?** To answer this question we first must consider “What is Q?” Q is a parameter borrowed from protein folding that indicates *structural similarity*. Traditionally, Q has meant “the fraction of similar native pairwise distances” between aligned residues in two proteins, 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 that few pair distances are similar to their native values, or, in other words, the structures do not align well. Homologs typically have  $Q \geq 0.4$ .  $Q_{res}$  is the contribution from each residue to the overall average Q value. For more information see Appendices A–C

Qres, is accessed by:

1. Click on the View menu in the MultiSeq program window.
2. Make sure Apply to all is checked and select Coloring  $\rightarrow$  Qres.

Look at the OpenGL Display window to see the impact coloring by Qres has made on the molecules.

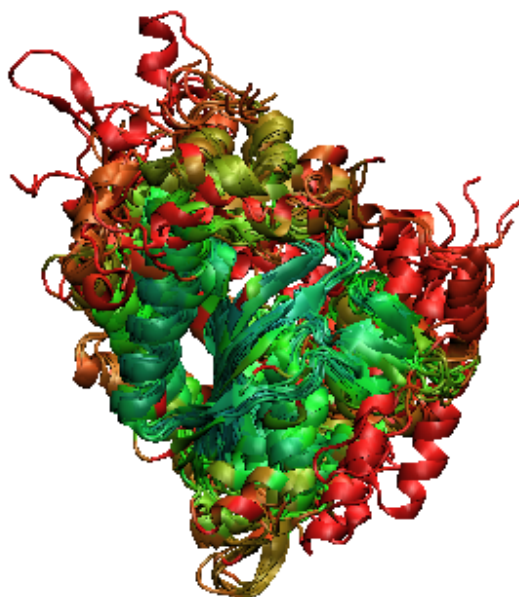
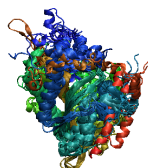


Figure 9: The catalytic domain colored by Qres.

You will probably notice that several regions within the interior of the aligned molecules have turned green. Rotate the molecule to see how much of it has

turned green. Green indicates that the molecules are somewhat structurally conserved at those points. If an area is highly structurally conserved, it would be blue. Red regions are not as highly conserved. Such regions often correspond to insertions that are unique to one specificity.



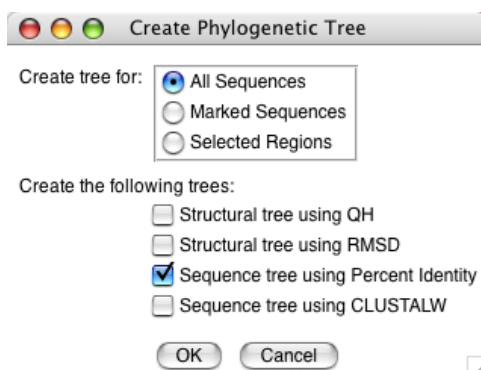
**Core Structure** Now that we have observed the structural conservation patterns, go back to the main MultiSeq window and see where the coloring of the core begins and ends. Using the side-scroll on the bottom of the main MultiSeq window, you can see the core residues number begins at about 120 and ends around 730.

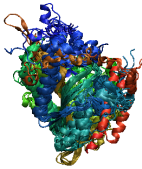
## 2.4 Structure Based Phylogenetic Analysis

### 2.4.1 Limitations of sequence data

In this section we will look at the phylogenetic history of the class II AARS structures. Most common methods of phylogenetic analysis use only information derived from the sequences to build the tree. However, many of the proteins we are looking at diverged before the last universal common ancestral state (LU-CAS), and have evolved independently since then. Consequently, they have a very low level of sequence identity. In fact, many of them have no more sequence relation than would be expected at random (8–10%). This is called the “twilight zone” of sequence identity and makes phylogenetic reconstruction using sequence metrics unreliable for very distantly related proteins. To demonstrate this, construct a sequence based phylogenetic tree of these AARSs by following these steps:


1. In the MultiSeq program window, select the Tools → Phylogenetic Tree menu option.
2. The Create Phylogenetic Tree dialog will appear. Select Sequence tree using Percent Identity as the type of tree to construct and press the OK button.





**Calculating phylogenetic relationships.** The phylogenetic trees in MultiSeq are all distance based trees. This means that they are calculated by using a pairwise metric (e.g. percent identity or  $Q_H$ ) to build a matrix comparing all possible pairs and then transforming this distance matrix into a tree. To do this, MultiSeq uses two treeing methods: UPGMA (Unweighted Pair Group Method with Arithmetic mean) and Neighbor-Joining. Other methods, such as Maximum Likelihood or Maximum Parsimony, may give more accurate results, but are generally much more computationally intensive. MultiSeq does not support computing trees this way, but will allow you to view them after they have been computed. Look up the details of these four tree computation methods on the Internet. Which one would you choose to use?

A phylogenetic tree based on percent sequence identity of the proteins will be calculated and drawn, as shown in Figure 10. Select View → Leaf Color → Domain of Life and then View → Leaf Labels → EC Description to show more information in the tree viewer. Notice that many of the branch points lie below 10% sequence identity (0.05 on the dendrogram). These branch points are unreliable as discussed above.



**How to read a phylogenetic tree.** MultiSeq shows phylogenetic trees as dendrograms. A dendrogram represents the distance between any two nodes of the tree as the total horizontal distance traversed to get from one node to the other. In Figure 10, for example, the distance traversed to get from d1e1oa2 to d11sca2 is 0.8, or twice the distance to their closest common parent node. In this example, that distance represents 20% identity between the two sequence. The distance between any two nodes is shown in the tree status bar when you click on the first node and then Shift click on the second node.

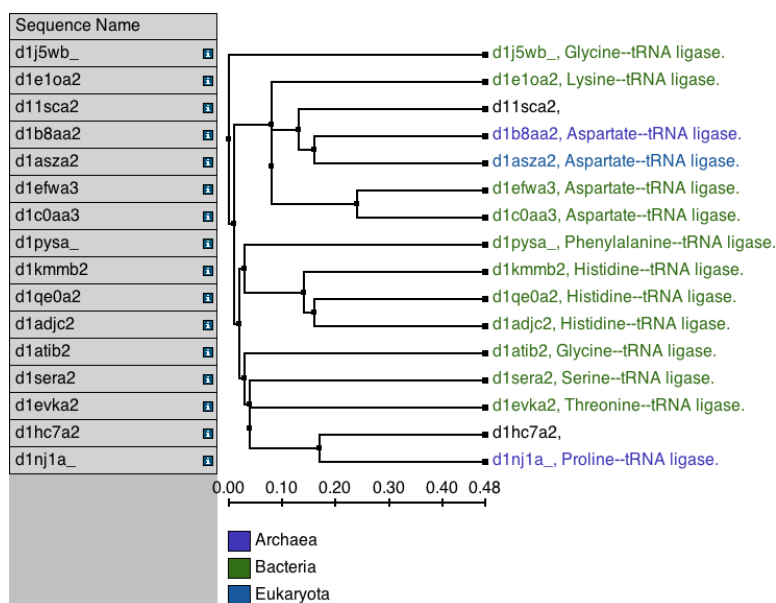


Figure 10: Percent identity sequence phylogenetic tree of 16 diverse AARS structures

#### 2.4.2 Structural metrics look further back in time

In order to reliably compare such distantly related proteins, we need a metric that is based on a property of the protein that is more highly conserved through evolutionary time. As structure has been shown to be more conserved than sequence, a structural metric fits this description. MultiSeq supports using  $Q_H$  and RMSD between aligned proteins to construct structural phylogenetic trees.  $Q_H$  is detailed in the paper titled “Evolutionary profiles derived from the QR factorization of multiple structural alignments gives an economy of information” located in the tutorial distribution at:

[TUTORIAL\\_DIR/papers/odonoghue\\_JMB\\_2005.pdf](TUTORIAL_DIR/papers/odonoghue_JMB_2005.pdf)

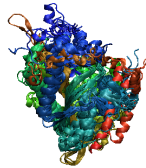
Generate a  $Q_H$  structural phylogenetic tree of the AARSs by performing the following:

1. Select the region of the sequence alignment between elements  $\sim 121$  and  $\sim 751$  where the structural overlap is good. This eliminates any extraneous insertions or deletions that may be the result of unresolved N- or C-terminal regions in the crystal structure and avoids any structural dissimilarities that may be the result of crystal packing artifacts in these regions.

Sequence Name		110	120	130
<b>VMD Structures</b>				
<input type="checkbox"/> d1nj1a_	48	.	P	H G F M I R K N T L K I
<input type="checkbox"/> d1evka2	268	.	N	D G W T I F R E L E V F
<input type="checkbox"/> d1c0aa3	139	P E M	A	Q R L K T R A K I T S L
<input type="checkbox"/> d1adjc2	14	.	G	K E L R M H Q R I V A T
<input type="checkbox"/> d1efwa3	138	R R M	Q	E N L R L R H R V I K A
<input type="checkbox"/> d1kmbb2	16	.	P	G E T A I W Q R I E G T
<input type="checkbox"/> d1asza2	242	V T N	Q	A I F R I Q A G V C E L
<input type="checkbox"/> d11sca2	125	R R P	F	A V M R I R D E L E R A

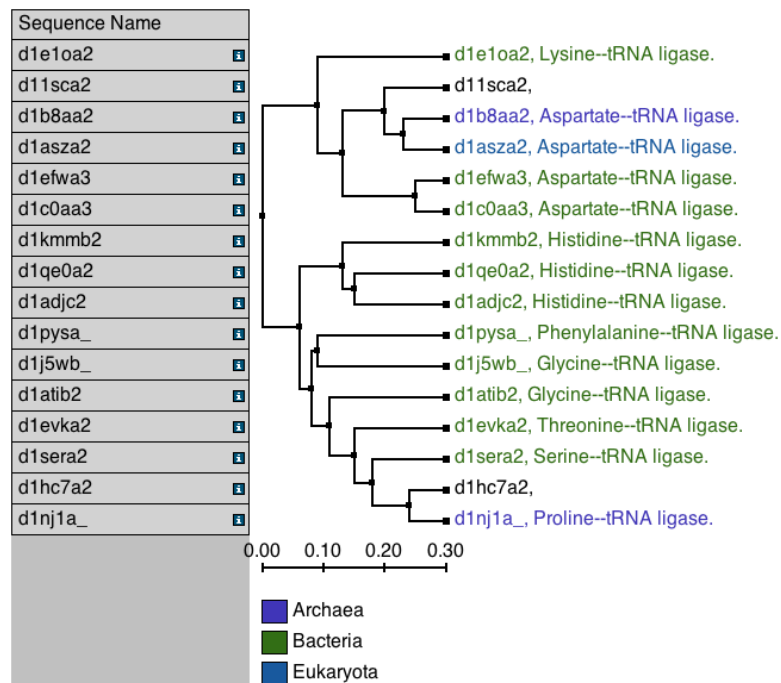
2. Select the Tools → Phylogenetic Tree menu option.
3. In the Create Phylogenetic Tree dialog select the Selected Regions radio button to use only the selected region during calculation of the tree
4. Make sure only the Structural tree using QH checkbox is checked and press the OK button.

MultiSeq calculates and displays the  $Q_H$  tree for the selected structural regions. Compare this tree, shown in Figure 11, to the percent identity sequence tree generated earlier. Notice how the branch points are much more evenly spaced, not bunched together on the left of the tree. This indicates that there is indeed a phylogenetic relationship between the structures that is elucidated when using the structural tree. What phylogenetic relationships can you discern from the tree?

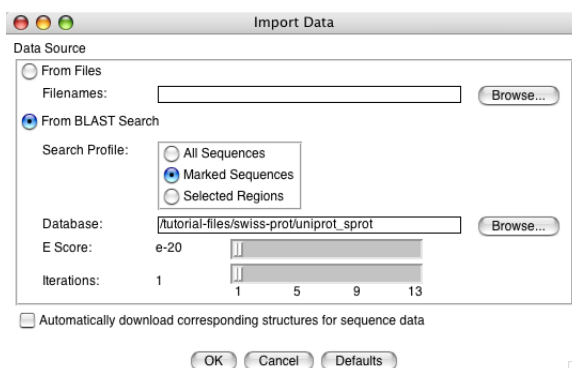


**Evolutionary relationship of the AARSs.** Compare the structural tree you created to the one shown on page 560 of the “On the evolution of structure in aminoacyl-tRNA synthetases” paper. Do you notice any differences between your tree and the one shown in the paper? You can right-click on nodes in the MultiSeq tree viewer and choose Rotate Left or Rotate Right to rearrange the tree. Try to get your tree to look like the one in the paper. Can you do so? If not, what are some reasons that might be the case?



Figure 11:  $Q_H$  structural phylogenetic tree of 16 diverse AARS structures





When the search is complete, a new dialog called BLAST Search Results appears (see Figure 12). As you may have noticed, over a hundred sequences were returned by BLAST. To restrict the results to only the Archaea, do the following:

1. In the Filter Options section and in Domain list, unselect the All list item by clicking on it and the select the Archaea list item.
2. Press Apply Filter button.

The dialog now displays only the 27 sequences from the domain Archaea, a much more manageable number. Press the Accept button at the bottom of the window to bring these sequences into MultiSeq.

### 3.1.2 Now the other domains of life

Now perform the same search for the eukaryal structure by unmarking d1b8aa2, marking d1asza2, and then repeating the above steps. Be sure to select Eukaryota in the Domain list this time. You should find around 12 eukaryal sequences. After that, bring in the bacterial sequences. Unmark d1asza2 and mark both d1c0aa3 and d1efwa3 to perform a BLAST profile search using those two bacterial sequences as the profile. Perform the same steps as you did before; this will bring in 257 sequences. You can immediately tell that the Bacteria are over-represented in the sequence databases. Eliminating this bias is important in obtaining good results and will be discussed later in more detail. Doing so, though, requires more computational time than is reasonable for this tutorial. Instead, simply delete all but the first 20 bacterial sequences from MultiSeq.

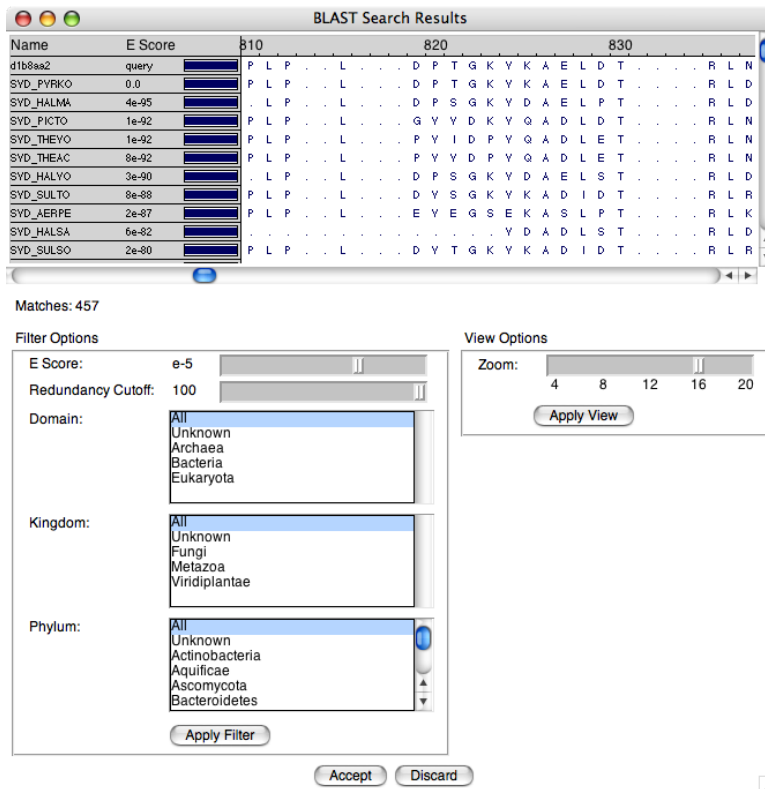


Figure 12: Blast Search Results Dialog

### 3.2 Organizing Your Data

At this point you may be overwhelmed by all of the data in the MultiSeq program window. In order to construct an evolutionary profile and observe sequence signatures specific to a particular domain of life, MultiSeq has various tools that help in the organization of data. One such tool allows you to automatically group sequences and structures by domain of life:

1. Select the **Options** → **Grouping** → **Taxonomy...** menu option. A new dialog called **Group Sequences by Taxonomy** appears.
2. Choose **All Sequences..**
3. Select domain as the level by which to group the data.
4. Press the **OK** button.

The sequences will now be grouped in the MultiSeq program window by domain of life. These groupings are mostly correct, however there is one sequence in

the Eukaryota that is incorrectly grouped. **SYDM\_YEAST** is actually a mitochondrial sequence that should be grouped with the Bacteria in any analyses. MultiSeq doesn't recognize this, since the sequence came from yeast, a Eukaryote. Move the sequence yourself by selecting the sequence and dragging it up the Bacteria group with your mouse.

### 3.3 Finding a Structural Domain in a Sequence

If you examine the size of some of the sequences you brought into MultiSeq, you will notice that they are somewhat longer than the sequences of the structural domains we were using before. That is because the sequences include the entire synthetase, not just the catalytic domain. Evolutionary analyses, on the other hand, should generally be done a domain at a time since domains can evolve at different rates and can even be interchanged among organisms. Here you will learn how to identify which parts of a sequence belong to a domain by mapping a structural domain onto a sequence. We will accomplish this by using ClustalW to align the sequences and then use MultiSeq to keep only the portions of the sequences that align to the catalytic domain in the structures. Since sequence alignments work best on closely related sequences, we will perform the operation on each domain of life separately, once again starting with the Archaea:

1. Make sure you don't have anything marked by right clicking on the **Archaea** group heading and choosing Unmark All.
2. Now mark all of the archaeal sequences by right clicking on the **Archaea** group heading and choosing Mark Group.
3. Bring up the gap removal dialog by choosing Edit → Remove Gaps... from the menu.
4. Select Marked Sequences and All Gaps and press OK.
5. Next, select the Tools → ClustalW Sequence Alignment menu option.
6. In the dialog, select Multiple Alignment and Align Marked Sequences.
7. Press the OK button. ClustalW will take a few seconds to perform the alignment.

You now have a sequence alignment of the archaeal sequences. Try coloring them by sequence similarity:

1. Select View → Coloring → Apply to Marked in the menu.
2. Then select View → Coloring → Sequence Similarity → BLOSSUM 30. It will take a few seconds for the coloring to be computed.

Now we want to delete the portions of the sequence that are not part of the catalytic domain. Normally, MultiSeq prevents you from making changes to the

sequence. Turn on sequence editing by choosing **Edit** → **Enable Editing** → **Full** from the menu. Be careful not to make any inadvertent changes. The catalytic domain is the C-terminal domain in the archaeal protein, so select the portion of the archaeal sequence alignment that corresponds to the N-terminal region, as shown in Figure 13, and press the **Delete** key. You have just deleted everything from the archaeal sequences except the catalytic domain.

Sequence Name		100	110	120
<b>Archaea</b>				
<input checked="" type="checkbox"/> d1b8aa2	104	-	-	P L P L D P T G K V K
<input checked="" type="checkbox"/> SYD_PYRKO	86	L G F E I L P E K I V V L N . R A E T	P L P L D P T G K V K	
<input checked="" type="checkbox"/> SYD_HALMA	83	T G V E V T P E S V D V I S . E A D P	E L P L D P S G K V D	
<input checked="" type="checkbox"/> SYD_PICTO	79	S G I E I I A D S V E I L N . A A E A	P L P L G V V D K V Q	
<input checked="" type="checkbox"/> SYD_THEVO	82	S G L E V S G E S V E V L N . R S E R	P L P L P V I D P V Q	
<input checked="" type="checkbox"/> SYD_THEAC	82	A G I E I S G T S I S I V N . E A E A	P L P L P V V D P V Q	
<input checked="" type="checkbox"/> SYD_HALVO	83	T G V E V T P E S L D V I A . E A E A	Q L P L D P S G K V D	
<input checked="" type="checkbox"/> SYD_SULTO	83	R G I E L H A E E I T L L S . K A K A	P L P L D V S G K V K	
<input checked="" type="checkbox"/> SYD_AERPE	88	E G V E V K V E R L E V L S . T P V E	P L P L E V E G S E K	
<input checked="" type="checkbox"/> SYD_HALSA	85	G G V E L A P T E L T V V S . E A T D	V P S I E I S K D V D	
<input checked="" type="checkbox"/> SYD_SULSO	83	N G V E V H A K D I E I L S . Y A R S	P L P L D V T G K V K	
<input checked="" type="checkbox"/> SYD_PYRAE	88	S G V E I F P S E I W I L N . K A K .	P L P I D I W S E T .	
<input checked="" type="checkbox"/> SYN_PYRAB	86	G G A E V H V E K L K V I Q A V S E F	P I P E N P . . . . E Q	
<input checked="" type="checkbox"/> SYN_PYRKO	85	T G A E V Q G E K L Q I I Q N V D F F	P I T K D . . . . .	
<input checked="" type="checkbox"/> SYN_PYRHO	86	G G A E V H V E K L E V I Q A V S E F	P I P E N P . . . . E Q	
<input checked="" type="checkbox"/> SYN_THEAC	84	S G Y E I A V D S F R V Y Q K N D V F	P I T K D . . . . .	
<input checked="" type="checkbox"/> SYN_PYRFU	86	G G A E V R V E K L E V I Q A V S E F	P I P E N P . . . . E Q	
<input checked="" type="checkbox"/> SYN_THEVO	84	T G Y E I S I H K F T V Y Q K N D V F	P I T K D . . . . .	
<b>Bacteria</b>				
<input type="checkbox"/> d1c0aa3	181	R V H K G K F Y A L P Q S P Q L F K Q L L M M S G F D R Y Y		
<input type="checkbox"/> d1efwa3	182	F L V P Y R H E P G L F Y A L P Q S P Q L F K Q M L M V A G		

Figure 13: Selecting the N-terminal domain for deletion

Perform the same series of operations on the Bacteria and Eukaryota groups. The bacteria pose an interesting problem since they have a third domain inserted into the middle of the catalytic domain, as discussed in the introduction. Make sure to delete this domain as well.

### 3.4 Aligning to a Structural Profile using ClustalW

Finally we have a set of sequences and structures of the catalytic domain of the aspartyl-tRNA synthetase loaded. In order to analyze the group as a whole, however, the entire set must be aligned. While sequence alignment methods generally work well for closely related proteins, this set is too diverse to yield a good sequence alignment. What we will do instead is use the structural alignment, which is more accurate for distant proteins, to guide the sequence alignment. The following steps will walk you through that process:

1. Unmark any sequences that are currently marked.
2. Right click on the **Archaea** group heading and choose **Insert Group...**
3. Enter **Structures** as the name of the new group and press **OK**.
4. Move the 4 structures into this new group and mark them.

5. Use Stamp to align the marked structures using the Tools → Stamp Structural Alignment menu option.
6. Unmark the structures.
7. Mark all of the sequences in the **Archaea**, **Bacteria**, and **Eukaryota** groups.
8. Remove all gaps from the marked sequences using the Edit → Remove Gaps... menu option.
9. Bring up the ClustalW dialog by choosing Tools → ClustalW Sequence Alignment from the menu.
10. In the dialog, select Profile Alignment and tell Clustal to align marked sequences to the Structures group.
11. Align the sequences to the structural profile by pressing the OK button. ClustalW will take a minute to perform the alignment.

You now have a complete structural based alignment of the aspartyl-tRNA synthetase catalytic domain. Try coloring it according to sequence identity by choosing View → Coloring → Apply to All and then View → Coloring → Sequence Identity (shown in Figure 14). Play around with the other coloring metrics. Do you understand what they all do? Also try coloring by groups independently. What additional insight do you think you can gain by doing so?

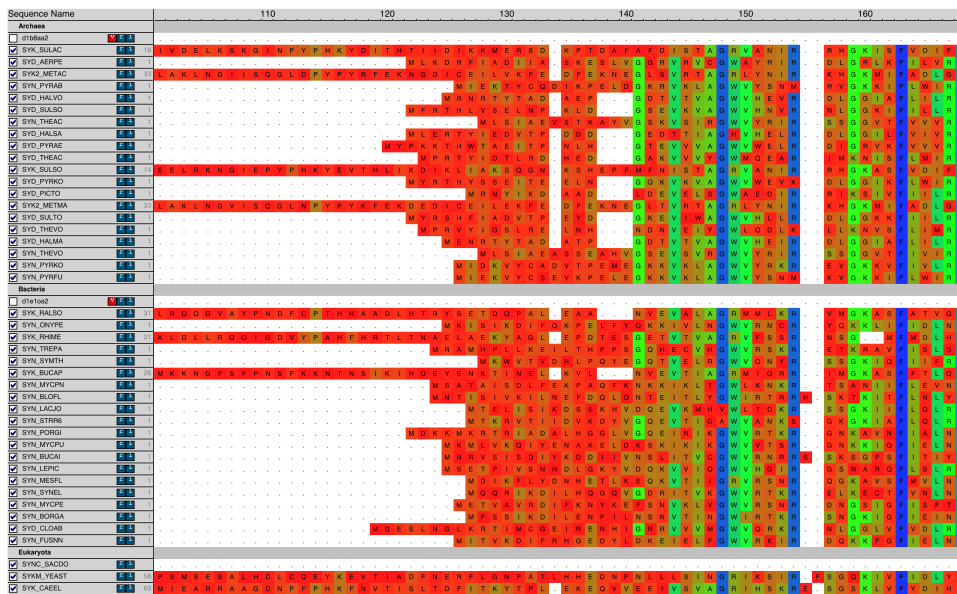


Figure 14: MultiSeq showing all sequences colored by sequence identity

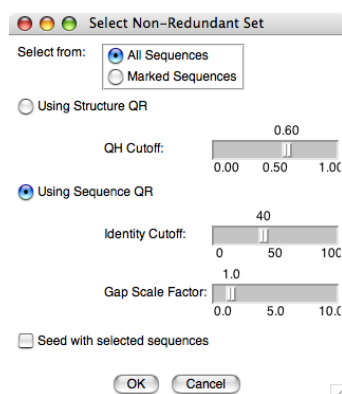


Figure 15: Select Non-Redundant Set dialog

### 3.5 Eliminating Redundancy with Sequence QR

While we now have a structural based alignment of the aspartyl catalytic domains, it is not yet an evolutionarily balanced profile. First, the databases from which we obtained our sequences were biased and, second, the bacteria and archaea generally have more sequence diversity than the eukarya. We need a way to remove any redundancy from our sequences in a systematic and balanced manner. MultiSeq provides the Sequence QR tool (see the accompanying paper) which does just that. Given a set of sequences, it will tell you which ones comprise the most linearly independent set of sequences. Try it by following these steps:

1. Make sure all of the sequences but none of the structures are marked.
2. Choose Search → Select Non-Redundant Set from the menu.
3. Select the Marked Sequences radio button.
4. Mark the Using Sequence QR radio button.
5. Set the Identity Cutoff to be 75.
6. Press the OK button.

A non-redundant set of sequences will be selected for you. You can easily make this into a new group by choosing the Options → Grouping → Create From Selection... menu option. Enter NR Set as the group name. Compare the sequences it picked to the ones it didn't choose. Do you notice any patterns? When you are done, delete everything from MultiSeq except the structures and the non-redundant sequences.



### 3.6 Phylogenetic Tree of an Evolutionary Profile

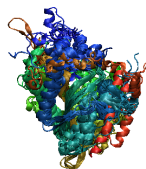
The phylogenetic tree function draws an unrooted dendrogram using sequence identity as the metric. To begin using this function:

1. Go to Tools → Phylogenetic Tree.
2. A window entitled Create Phylogenetic tree will appear
3. Select Create Tree for → All Sequences within the window and check Sequence tree using Percent Identity
4. Press the OK button.

Another window will appear with the dendrogram. Within the new window select the following:

- View → Leaf Color → Domain of Life
- Turn on View → Leaf Labels Name, Species Name, and EC Description.

The tree should appear as shown in Figure 16.



**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 Felsenstein.<sup>a</sup>

<sup>a</sup>J. Felsenstein *Inferring Phylogenies*. Sinauer Associates, Inc.: 2004.

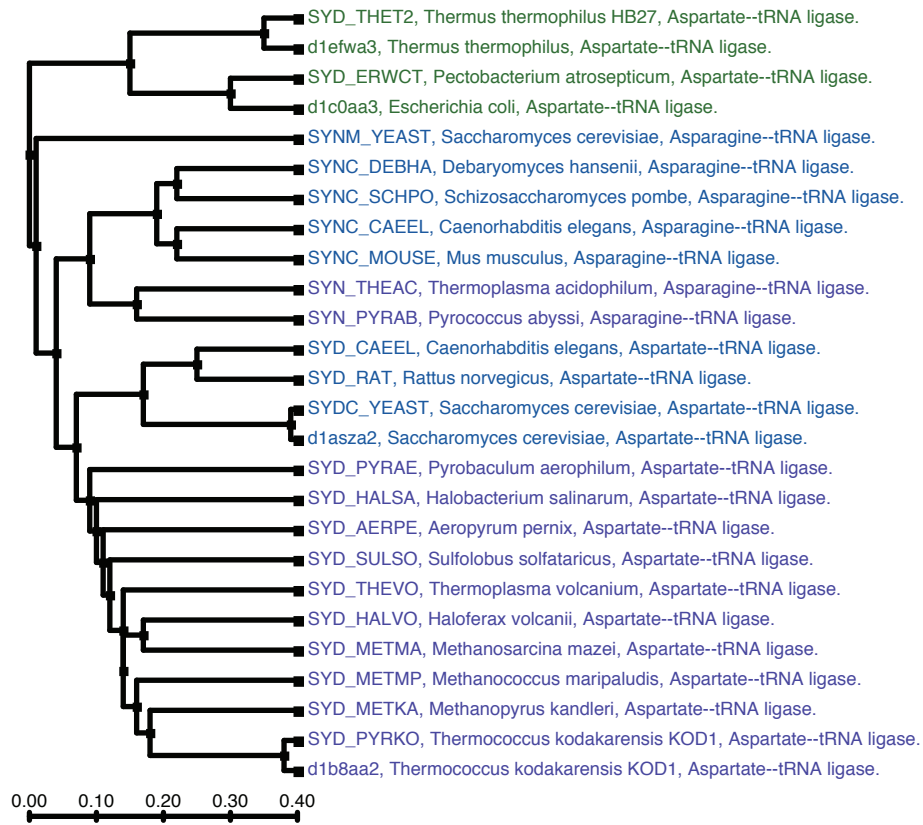


Figure 16: Phylogenetic tree based on sequence identity

### 3.7 Export Data

Data from MultiSeq sessions can be saved in various formats, such that it can be used in other bioinformatics software applications and suites. To save data from a MultiSeq session,

1. Select File→Export Data
2. A new window will appear entitled Export Data
3. Click on the radio button next to the format you want to save to.
4. Hit the OK button.

### 3.8 MultiSeq Sessions

MultiSeq sessions can be saved, closed, and later reloaded into VMD and MultiSeq. This is done by,

- Selecting File→Save Session to save a session.
- Selecting File→New Session to close the current session and start a new MultiSeq session.
- Selecting File→Load Session to load a previously saved session.

MultiSeq sessions are saved into a script with a .multiseq extension. An associated directory is also created. It is within this directory, that various files that contain the alignment data are stored. To save all of the work you have done, go ahead and save the session. You have now completed Part 2 of this tutorial. Close this session of MultiSeq and take a refreshment break! The next part of the tutorial will require a new session of VMD and MultiSeq.

## 4 Evolutionary Analysis of tRNA

### 4.1 tRNA and Modified Bases

As we showed in the introduction, the AARSs charge their cognate tRNA with the amino acid that will subsequently be incorporated on the ribosome into the growing protein chain. In general, the tRNA is made up of 76 ribonucleotides and possess a stable tertiary L-shaped structure under proper pH and ionic conditions. Unlike mRNA and rRNA, tRNA can have as much as 10-15% modified bases. RNA has around 100 known modified bases. Some important modified bases are dihydrouridine (D), pseudouridine (P), and ribosyl thymine (T).

1. Start VMD and MultiSeq.
2. Load 1ASZ-tRNA\_SCer\_D.E.pdb (tRNA:Asp in MultiSeq using Import Data).
3. Notice that there are characters in the alignment that are not A, C, G, or U.

Look at the tRNA structure in the OpenGL window. RNA is transcribed in the 5' to 3' direction so the first nucleotide (U) is at the 5' end of the tRNA molecule. In tRNA, basepaired regions are referred to as “stems”, unbasepaired regions are “loops”, and the structure produced by a stem capped by a loop is called an “arm”.

Since tRNAs have such similar structure, there is a common numbering convention for the nucleotides. When there are insertions or deletions in the molecule, the numbering is not changed. This allows for features of the tRNA to maintain the same numbering across different molecules. The anticodon, for example, is always present at bases 34, 35, and 36.

1. Open the VMD Sequence Viewer from the VMD Main window through Extensions→Analysis→Sequence Viewer.
2. Click the 1-letter code button.

Immediately, you can see the 3-letter codes of several modified bases such as pseudouridine (PSU) and dihydrouridine (H2U), because they do not have 1-letter codes in this viewer. In this pdb file, the tRNA numbering starts with 601, but the second two digits maintain the standard tRNA numbering.

3. Scroll down to base 646. You'll notice that there is no 647. There has been a deletion in the sequence of this tRNA with respect to the standard numbering.
4. Close the Sequence Viewer.

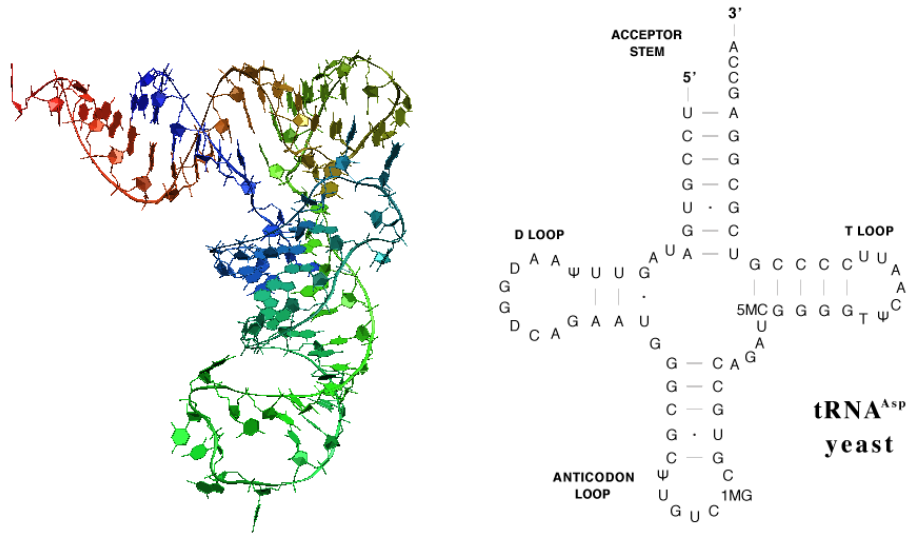
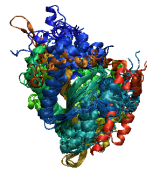


Figure 17: 3D and Cloverleaf view of tRNA

5. Return to the main MultiSeq window and color the molecule through View→Coloring→Element Index.
6. Change the highlighting style through View→Highlight Style→Bonds.
7. Highlight the first seven residues in the alignment window.
8. Next highlight the last eleven residues of the sequence.



**The tRNA cloverleaf.** The stem of the cloverleaf is called the acceptor stem. The first seven bases and the last eleven bases comprise the acceptor stem. The last three bases are referred to as the CCA end and are a common feature of tRNAs. The fourth from the last base tends to be similar across tRNA amino-acid specificity. This base is called the discriminator base. When a tRNA is charged with its cognate amino acid, the amino acid is loaded onto the 3' sugar. This end binds to the catalytic domain of the corresponding AARS.

9. Next highlight bases 11 to 25.

This is the first leaf of the cloverleaf structure. It is called the D-arm because dihydrouridine bases are commonly found in the loop.

10. Highlight bases 26 to 44.

The second leaf, opposite the acceptor stem is the anticodon arm, and the three anticodon bases are located in the middle of the anticodon loop. The anticodon bases are responsible for codon recognition on the mRNA when the charged tRNA is loaded onto the ribosome. In this sequence, the anticodon is GTC. Highlight bases 34 to 36 to reveal the anticodon.

11. Highlight bases 48 to 64.

The last leaf of the cloverleaf structure is the T-arm, so-called because it contains the T $\Psi$ C sequence motif at the 5' end of the T-loop. Highlight bases 53 to 55 to see the T $\Psi$ C motif.

## 4.2 Structural Alignment

Load up the other six structures from pdb files. The names of the files include PDB code, organism, amino-acid specificity, and domain of life. The format is *pdbcode-tRNA\_species\_specificity\_domain*. Two of the tRNAs are tRNA-Asp, two are tRNA-Cys, and three are tRNA-Phe. Species information is provided because the PDB code is associated with the protein, and there are cases where the AARS and tRNA in a crystal structure have been taken from different organisms. Look at the taxonomy information for 1B23 and 1TTT (click on the *i* button) for examples of this.

1. Structurally align the tRNAs using Tools→Stamp Structural Alignment with default values.
2. Color the alignment by View→Coloring→Sequence Identity.
3. Scroll across the alignment and notice the two largest gapped regions.

One is at the anticodon loop around column 40 and the other concerns the CCA end at the right side of the alignment. The two tRNA-Asp structures were both bound to AARS molecules in the crystal. Their anticodon loops unwind and flip out for recognition by the Asp-AARS. The CCA end is poorly aligned because CCA and the discriminator base are single-stranded RNA and can experience a lot of motion. These issues cause problems with the structural alignment.

To fix these misalignments, you will use the alignment editing features of the Multiple Alignment plugin.

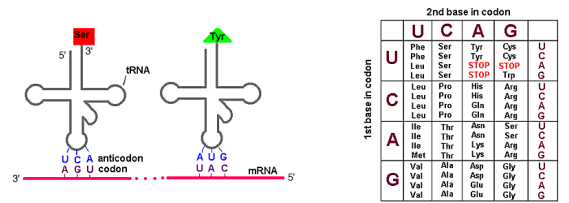


Figure 18: 7 tRNAs aligned with STAMP

### 4.3 Alignment Editing

1. Turn on gap editing through Edit→Enable Editing→Gaps Only. This activates gap editing mode allowing you to add or delete gaps in the alignment.
2. Remove the five-space gaps at the anticodon loop by selecting the base at the right edge of the gap and pressing the BACKSPACE key on your keyboard five times.
3. Alternately, you can highlight the five-space region and press BACKSPACE to delete the whole region at once.
4. Now scroll to the CCA end. Line up the discriminator base and the CCA ends of the sequences.

You may notice that one of the CCA ends is actually CCX. This tRNA (1TTT-tRNA\_Ser\_F\_E.pdb) comes from a complex with EF-Tu and has already been charged with a cysteine.

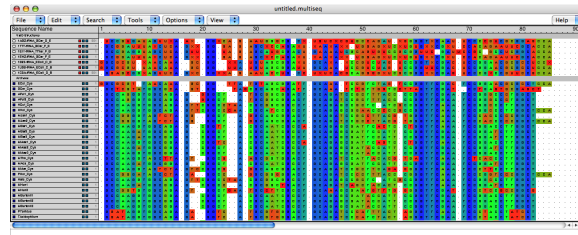
### 4.4 Sequence Alignment

Bring in tRNA-Cys sequences through File→Import Data (Jgi-Gtrnadb.fasta). These data are genomic tRNA sequences from the Joint Genome Institute (<http://img.jgi.doe.gov/cgi-bin/pub/main.cgi>) and the Genomic tRNA Database (<http://lowelab.ucsc.edu/GtRNadb/>). They are genes sequenced from DNA and so have no modified base information.

Also, since the sequences have not been transcribed, they contain thymine bases (T) instead of uracils (U). Almost all of the sequences are from Archaea. The exceptions are E. coli and yeast that appear at the top of the set.

1. Rename the group “Archaea” by right-clicking on the group divider (marked “Sequences”), and choosing Rename Group....
2. Right-click the group divider again and mark all of the sequences in the Archaea group.
3. Align the RNA sequences using ClustalW through Tools → ClustalW Sequence Alignment (Marked Sequences, etc.)

- Color the alignment by sequence identity through View→Coloring→Sequence Identity.



- Secondary structure information has already been generated for these sequences. Load it in through File→Import Data (SecondaryArchaea.fasta). Acceptor stem basepairing is represented with A, D-Stem with D, anticodon stem with C, T-Stem with T, the anticodon with N, and the TΨC motif with P. Are all of these regions aligned well?
- Rename the group with the secondary structure information “Secondary Structure” to keep it separate from the gene sequences.
- Bring in prealigned tRNA-Cys sequences and their secondary structure data through File→Import Data (Bayreuth.fasta, Secondary.fasta).
- Move the secondary structure line into the bottom of the “Secondary Structure” group. These sequences come from the Bayreuth tRNA Compilation (<http://www.staff.uni-bayreuth.de/~btc914/search/index.html>). This alignment was made using the standard tRNA numbering which takes basepairing information into account.
- Compare this alignment against its secondary structure. Is it more consistent with the basepairing information than the previous alignment? Are there any problems with this prealigned data? The misaligned sequences will not significantly affect the tree.

Next we will improve the alignment of our initial sequences by aligning them against the Bayreuth alignment.

- First, you will remove the gaps from the Archaea alignment.
- Mark only the sequences in the Archaea group.
- Now remove the gaps through Edit→Remove Gaps... (Marked Sequences, All gaps).
- To align the Archaea sequences against the Bayreuth alignment click through Tools→ClustalW Sequence Alignment (Profile Alignment, Align marked sequences to: “Sequences”).



5. Look at the secondary structure information to check that gaps were not added to the Bayreuth alignment.

Now you have a full alignment of both sets of data. The Bayreuth alignment had many gap columns that have now been introduced into your initial sequence set. To make the alignment easier to view, remove the gap-only columns by clicking Edit→Remove Gaps...(All Sequences, Redundant Gaps).

#### 4.5 Sequence Tree

You may not expect a phylogenetic tree of tRNA to be very reliable. There is much less information in the 76 nucleotide characters than in a typical protein sequence or the ribosomal RNA.

1. Mark the “Archaea” group and the “Sequences” group.
2. Create a sequence-based phylogenetic tree through Tools→Phylogenetic Tree (Marked Sequences, Sequence tree using Percent Identity).
3. You may need to increase the width of the tree to make the groupings more apparent. Do this through View→Scale Up.

Three main clusters should appear: Archaea, Bacteria, and Eukarya. Despite the small amount of sequence data in a tRNA alignment, there is an obvious high-level pattern. Does it correspond to the ribosomal RNA three-domain tree of life? Why or why not?

1. Create two new groups in the MultiSeq program window by right-clicking an existing group divider and choosing “Insert Group...”.
2. Name these groups “Bacteria” and “Eukarya”. It is useful to view the data with similar sequences grouped together. Since the available tRNA databases do not yet make it simple to retrieve taxonomic information for their sequences, you will need to do this grouping by hand. Fortunately, with a phylogenetic tree, this is relatively easy to do. Locate the cluster with mostly bacterial sequences. It should include E.COLI and BACILLUS.SUBTILIS.
3. Now select all of the branches in this subtree by clicking the upper-most bacterium in the list at the left of the tree window, holding SHIFT, and clicking the bacterium name at the bottom of the group. This selection will also appear in the alignment window.
4. Return to the alignment window then click and drag one of the highlighted sequences into the “Bacteria” group. All of the highlighted sequences will be moved.
5. Repeat this process to move the eukaryal sequences into the “Eukarya” group. This cluster should include SCer.Cys and SCHIZOSACCHA.POM.

6. Select the remaining sequences in the tree and move them to the “Archaea” group. *PLASMODIUM\_FALCIP.* is a bacterium, but it is located outside the other groupings in this tree. Move it into the “Bacteria” group. Can you see any other organisms that seem to be placed in a strange location on the tree? Why might this happen?
7. Now apply sequence identity coloring to each group separately. Click View→Coloring→Apply to Group then View→Coloring→Sequence Identity.

Which columns are conserved across all of life? What secondary structures are conserved? Are there columns that are conserved in some domains and not in others? Are there columns that are highly conserved within each domain where the character varies between domains?

## 5 Appendices

### 5.1 Appendix A: $Q$

The following equation is from the article “Evaluating protein structure-prediction schemes using energy landscape theory” by Eastwood, M.P., C. Hardin, Z. Luthey-Schulten, and P.G. Wolynes in IBM J. Res. Dev. 45: 475-497. 2001.

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

$r_{ij}$  is the distance between a pair of  $C^\alpha$  atoms.

$r_{ij}^{nat}$  is the  $C^\alpha$ - $C^\alpha$  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.

## 5.2 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:

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

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

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

The first term,  $q_{aln}$ , computes the unnormalized fraction of  $C^\alpha$ - $C^\alpha$  pair distances that are the same or similar between two aligned structures.  $r_{ij}$  is the spatial  $C^\alpha$ - $C^\alpha$  distance between residues  $i$  and  $j$  in protein a, and  $r_{i'j'}$  is the  $C^\alpha$ - $C^\alpha$  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  $\sigma_{ij}^2$  terms are computed as:

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

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

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.  $\sigma_{ij}^2$  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  $Q_{aln} = \aleph_{qaln}$ , which is identical to the Q-measure described into the  $Q$  measure described before.

### 5.3 Appendix C: $Q_{res}$ Structural Similarity per Residue

Here we define another metric, called  $Q_{res}$ , that is derived from  $Q$ .  $Q_{res}$  computes the similarity of the  $C_\alpha$ - $C_\alpha$  distances between the residue of interest and all other residues in the protein, excluding nearest neighbors, to the corresponding distances in a given set of proteins. The result is a value between 0 and 1 that describes the similarity of the structural environment of a residue in a particular protein to the environment of that same residue in all other proteins in the set. Lower scores represent low similarity and higher scores high similarity. If the set of proteins represents an evolutionarily balanced set, then structural similarity corresponds to structural conservation. Formally,  $Q_{res}$  is defined as follows:

$$Q_{res}^{(i,n)} = \aleph \sum_{(m \neq n)}^{proteins} \sum_{(j \neq i-1, i, i+1)}^{residues} \exp \left[ -\frac{(r_{ij}^{(n)} - r_{i'j'}^{(m)})^2}{2\sigma_{ij}^2} \right] \quad (1)$$

where  $Q_{res}^{(i,n)}$  is the structural similarity of the  $i^{th}$  residue in the  $n^{th}$  protein,  $r_{ij}^{(n)}$  is the  $C_\alpha$ - $C_\alpha$  distance between residues  $i$  and  $j$  in protein  $n$  and  $r_{i'j'}^{(m)}$  is the  $C_\alpha$ - $C_\alpha$  distance between the residues in protein  $m$  that correspond to residues  $i$  and  $j$  in protein  $n$ . The variance is related to the sequence separation between residues  $i$  and  $j$ ,

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

and the normalization is given by

$$\aleph = \frac{1}{(N_{seq} - 1)(N_{res} - k)} \quad (3)$$

where  $N_{seq}$  is the number of proteins in the set,  $N_{res}$  is the number of residues in protein  $n$ , and  $k$  is 2 when residue  $i$  is the N- or C-terminus otherwise 3.

In order to know which residues correspond to each other across the set of proteins,  $Q_{res}$  requires a multiple sequence alignment (MSA) of the proteins' sequences. Typically the MSA is generated using a structural alignment program.