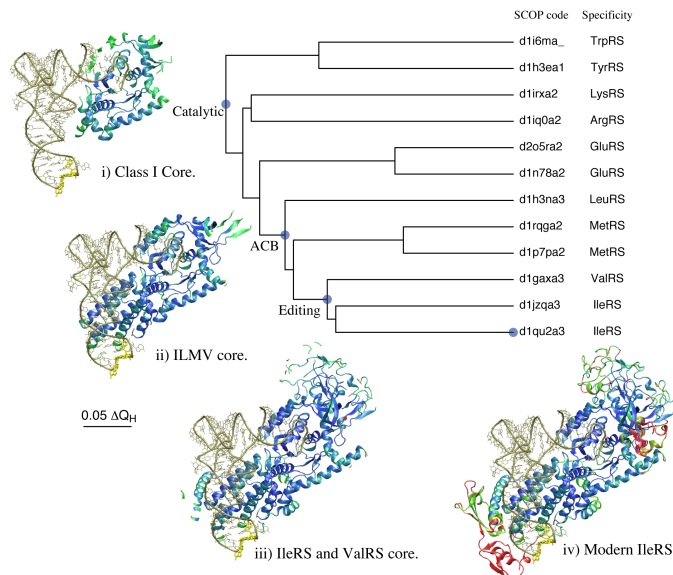


University of Illinois at Urbana-Champaign
Luthey-Schulten Group
NIH Resource for Macromolecular Modeling and Bioinformatics
Computational Biophysics Workshop

Evolution of Translation Class I Aminoacyl-tRNA Synthetase:tRNA complexes



VMD Developer: John Stone

MultiSeq Developers Tutorial Authors

Elijah Roberts
John Eargle
Dan Wright

Li Li
Anurag Sethi
Zan Luthey-Schulten

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

Contents

1	Introduction	4
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 Glutamyl-tRNA Synthetase:tRNA Complex	13
1.4.1	Loading the structure into MultiSeq	13
1.4.2	Selecting and highlighting residues	14
1.4.3	Domain organization of the synthetase	15
1.4.4	Nearest neighbor contacts	15
2	Evolutionary Analysis of aaRS Structures	19
2.1	Loading Molecules	19
2.2	Multiple Structure Alignments	20
2.3	Structural Conservation Measure: Q_{res}	21
2.4	Structure Based Phylogenetic Analysis	24
2.4.1	Limitations of sequence data	24
2.4.2	Structural metrics look further back in time	26
3	Complete Evolutionary Profile of TyrRS	29
3.1	Expanding the genetic code by engineering TyrRS	29
3.2	Comparing archaeal and bacterial TyrRS:tRNA complexes	29
3.3	The structural basis of the altered specificity of the engineered TyrRS	30
3.4	Evolutionary Profile of TyrRS	31
3.4.1	Importing the archaeal sequences	31
3.4.2	Now the other two domains of life	33
3.4.3	Organizing Your Data	34
3.4.4	Aligning to a Structural Profile using ClustalW	35
3.4.5	Curating the sequence alignment	36
3.4.6	Eliminating Redundancy with Sequence QR	37
3.4.7	Phylogenetic Tree of an Evolutionary Profile	38
3.4.8	Insights from the evolutionary profile	39
3.5	Export Data	42
3.6	MultiSeq Sessions	42
4	Evolutionary Analysis of tRNA	43
4.1	tRNA and Modified Bases	43
4.2	Structural Alignment	45
4.3	Alignment Editing	46
4.4	Sequence Alignment	47

<i>CONTENTS</i>	3
4.5 Sequence Tree of tRNA ^{Tyr}	49
5 Acknowledgments	49
6 Appendices	51
6.1 Appendix A: Q	51
6.2 Appendix B: Q_H	52
6.3 Appendix C: Q_{res} Structural Similarity per Residue	54

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 - Archaea, Bacteria, and Eukarya. For example, one can compare the bacterial sequences and structures of a particular biomolecule to its human counterpart in MultiSeq. MultiSeq contains several metrics for the comparison of sequences and structures developed by the Luthey-Schulten group [7, 8, 16] in addition to some of the standard metrics such as percentage identity, sequence similarity, sequence entropy, and RMSD of structures. 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. Q_H is a measure for the structural similarity between pairs of homologs and is based on a metric, Q , developed by Wolynes, Luthey-Schulten, and coworkers [4], to measure the local unfolding of a protein (see Appendix A). In addition to Q , Q_H has also got a gap penalty term that measures how insertions and deletions perturb the aligned core structure of the biomolecule. 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[12], BLAST[1], and ClustalW[17]. 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 of evolution. MultiSeq is an invaluable tool for relating protein structure to function and can be used to generalize the results to homologous molecules in all three domains of life.

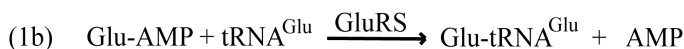
This tutorial showcases the MultiSeq environment and will allow the reader to combine sequence and structure information into evolutionary profiles used on protein:RNA complexes in translation [7, 8, 16, 15]. Evolutionary profiles are compact representative sets that can be used for gene annotation [16], coevolution [11], and energetic analysis [3]. 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), Archaea (A), and Eukarya (E). The essential process of protein synthesis requires twenty sets of synthetases and their corresponding tRNAs for the correct trans-

mission of the genetic information. The aaRSs are responsible for loading the twenty different amino acids (aa) onto their cognate tRNA (tRNA containing the appropriate anticodon). The formation (See Figure 1) of aminoacyl-tRNA (aa-tRNA) occurs via 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[10, 13].

Direct Pathway



Indirect Pathway

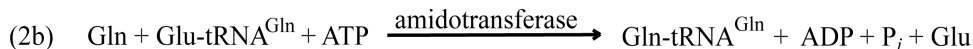


Figure 1: (1) The two steps direct acylation of tRNA by glutamyl-tRNA synthetase. (a) The glutamate is first combined with an ATP molecule to form an “activated” glutamyl-adenylate and then (b) the adenylate reacts with the tRNA to form the “charged” glutamyl-tRNA. (2) The indirect mechanism for charging the tRNA. (a) The tRNA^{Gln} is mischarged with a glutamate which is then (b) converted to a glutamine 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 based on homology of their catalytic domains: class I or class II. Class I aaRSs possess 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 leucyl-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 2). 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.

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 1.8.7 beta 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 500 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_Translation_Class-I/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

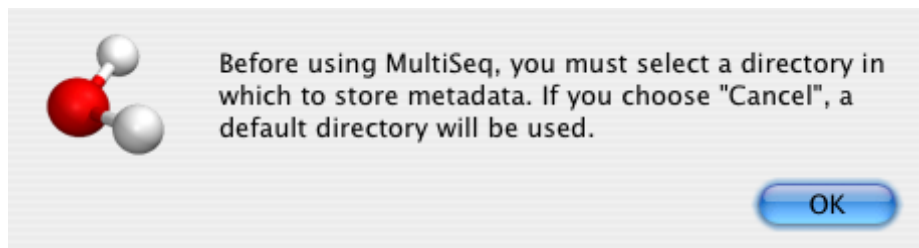
MultiSeq saves user preferences in a file named `.multiseqrc` located in your home directory. The preferences saved include the location of any local databases, previous search options, and others. When you start MultiSeq for the first time, it will ask you if it is ok to create this file and to specify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



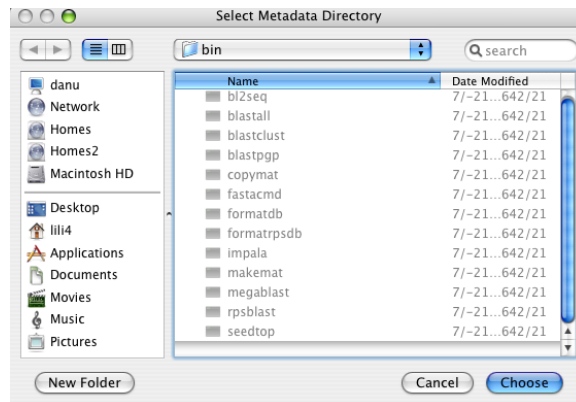
What is metadata? Metadata is a term meaning “data about data”. In MultiSeq the word metadata refers to information about the sequences or structures loaded into the program. MultiSeq knows how to find certain types of sequences or structures in the public metadata databases and can display information 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:

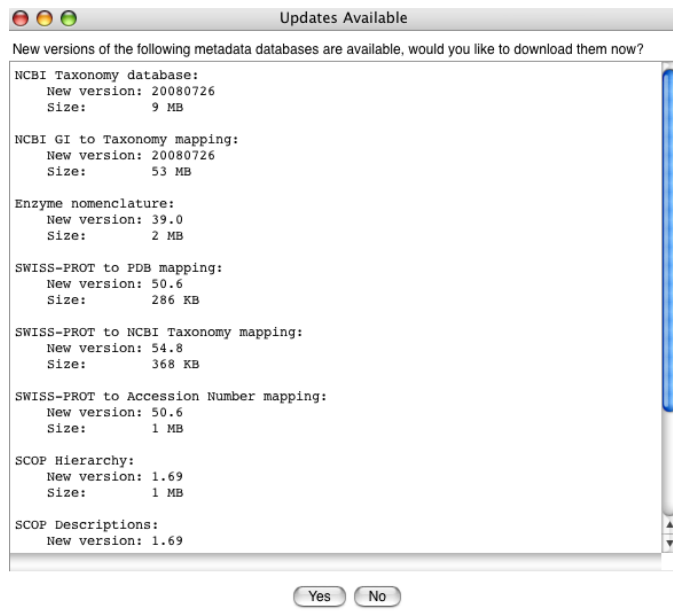
1. Launch VMD.
2. Within the VMD main window, choose the Extensions menu, select Analysis → MultiSeq.
3. MultiSeq will notify you that you must select a directory in which to store metadata databases. Press the OK button.



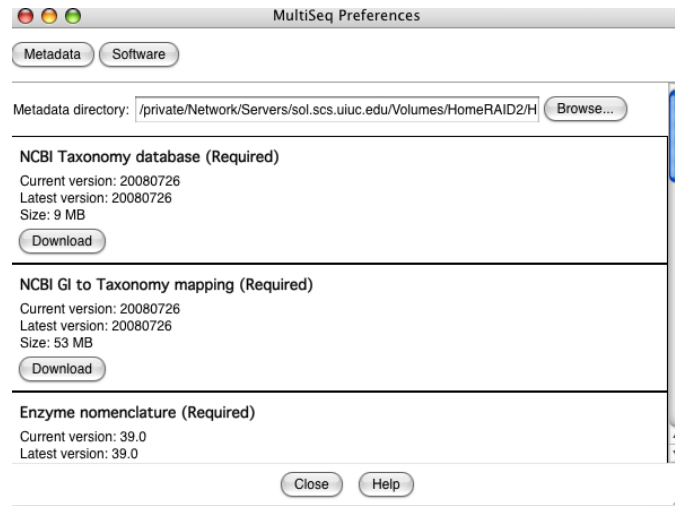
4. You will then be prompted to select the metadata directory. If the directory already contains the metadata databases, MultiSeq will use them. If not, MultiSeq will download them into the directory. If you are following this tutorial from a CD, choose the TUTORIAL_DIR/multiseqdb directory in the dialog 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.



5. 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.



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



7. 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.

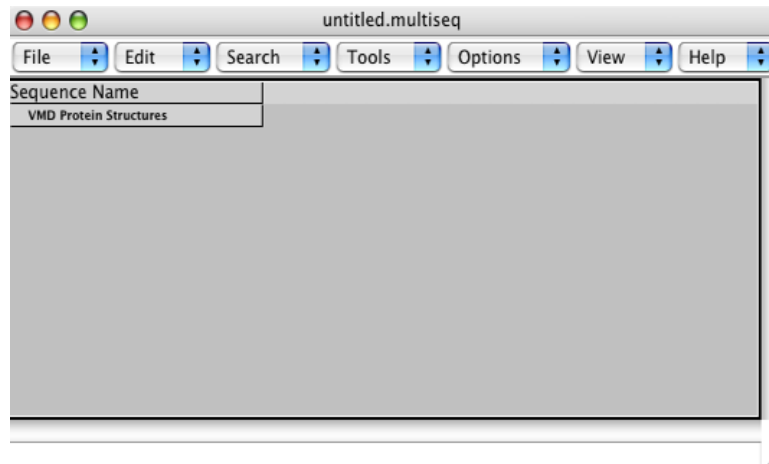


Figure 3: The MultiSeq program window

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:

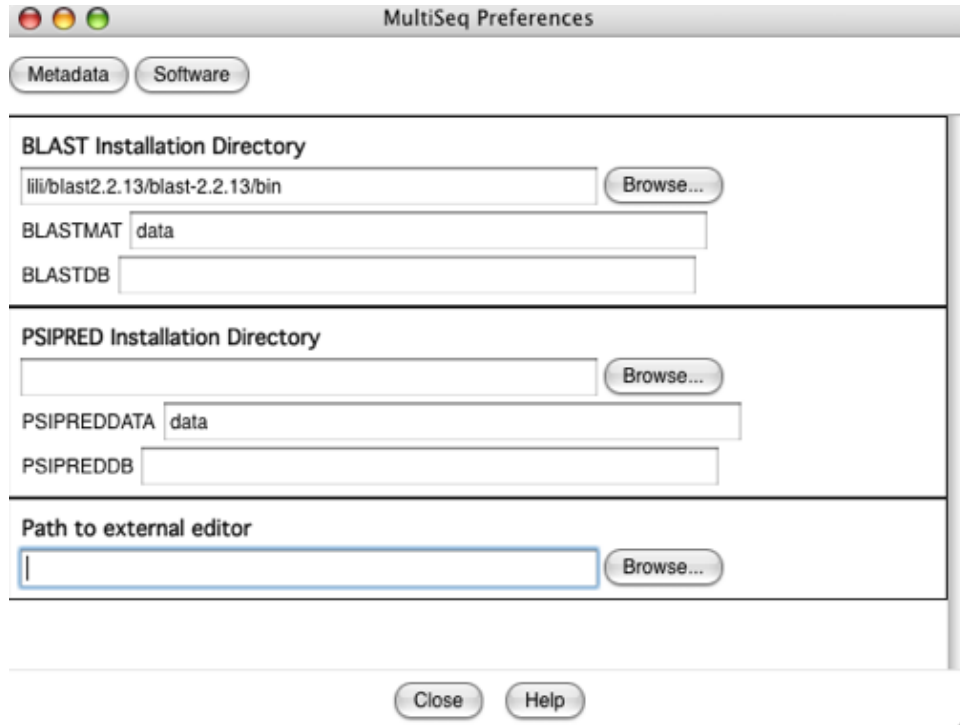
```
/Tutorials/class-I/blast-install/
```

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.



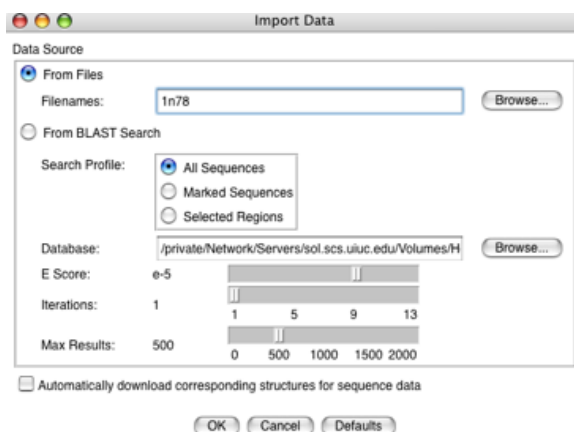
7. Press the Close button to save these changes. MultiSeq is now configured to use your local installation of BLAST.

1.4 The Glutanyl-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 glutamyl-tRNA synthetase (GluRS) as complexed with glutamyl-adenylate analog and tRNA^{Glu} (PDB code: 1n78). 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 “1n78”. 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/1n78.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. “1n78,1asy,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.

You should now have the GluRS:tRNA complex loaded in MultiSeq, as shown in Figure 4. When you load a structure into VMD, MultiSeq represents each chain of the molecule as a separate row showing the one character code for each residue in the columns. In 1n78, the crystallographic unit contains two nearly identical complexes. Therefore, you can see two molecules of protein (the A chain and the B chain) and they are named as **1n78_A** and **1n78_B** in the MultiSeq program window. There are also two molecules of tRNA and named **1n78_C** and **1n78_D**, respectively.

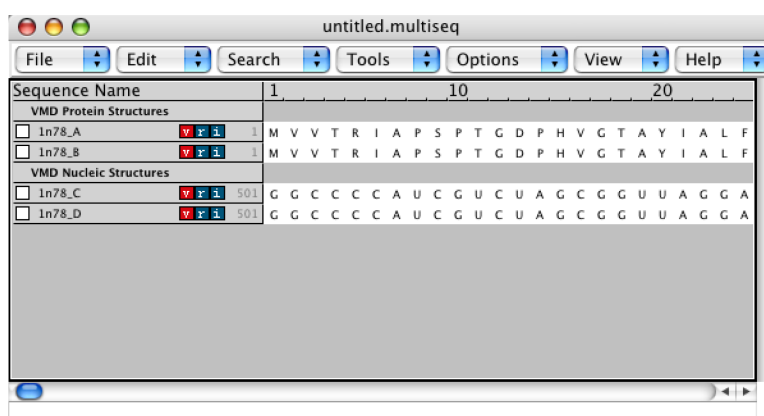


Figure 4: MultiSeq showing the loaded structure 1N78

1.4.2 Selecting and highlighting residues

Click on one of the residues in the sequence named **1n78_A**. The residue should appear highlighted in both the MultiSeq window and the Open GL display. If you can't see it in the Open GL display, try changing the representation used for highlighting the current selection by selecting to the View → Highlight Style → VDW menu option. Notice that MultiSeq also shows the resID of the currently selected residue in the status bar at the bottom of the MultiSeq window. These are the same resID numbers as in the PDB file and can be very useful during 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.

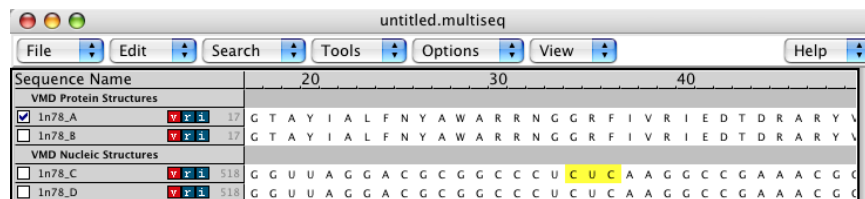
1.4.3 Domain organization of the synthetase

All of the aaRSs are multidomain proteins, but the exact number and fold of each domain is specific to each synthetase. GluRS has a catalytic domain (comprised of residues 1–79 and 181–306), a four helix-junction domain (residues 307–374), an anticodon-binding domain (residues 375–468), and a CP1 insertion (residues 81–186, CP1 referred to connected-peptide 1). Interestingly, the CP1 insertion interrupts the sequence of the catalytic domain. 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 glutamate is comprised of C534, U535, and C536. Select these bases in Multiseq and they will be highlighted. Note how the anticodon-binding domain of the enzyme attaches itself to the anticodon in the tRNA; zoom in on the anticodon. The CUC anticodon decodes GAG codon, which encodes glutamate. 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 those residues in the GluRS that recognize the anticodon. To make this process easier, MultiSeq provides a function that allows you search for residues in contact with a selected region. To do this, first click the checkbox to the left of the name of sequence **1n78_A**. The sequence should appear checked as shown below.



This is called marking a sequence; multiple 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.

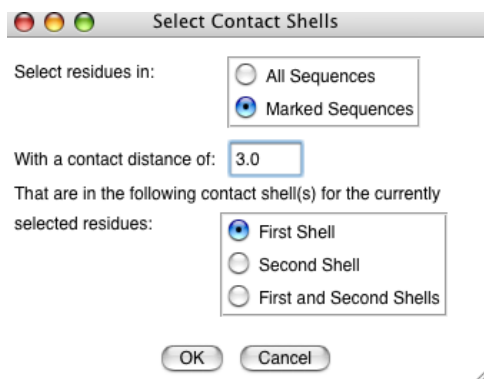


Figure 5: Select Contacts Shells dialog

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. They include R358, R417, R435, L442 and T444. As you may noticed, many of them are positively charged residues, which can stabilize the RNA:protein interaction by electrostatic forces. Among these residues, R358 is particularly interesting: it is responsible for discriminating tRNA^{Glu} and tRNA^{Gln}[14]. Also notice the π - cation interaction between residue R435 of the synthetase and bases U535 of the tRNA. What other types of interactions between the protein and tRNA can you recognize?

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 bends into the active site of the GluRS. Select the residue of position 469 in chain A. This “mysterious” residue is the analog of glutamyl-adenylate. The formation of the glutamyl-adenylate comes from one glutamate molecule and ATP; this adenylated species is “activated” and then transferred 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 out of the active site; the glutamyl-adenylate is trapped between the catalytic domain and the tRNA.

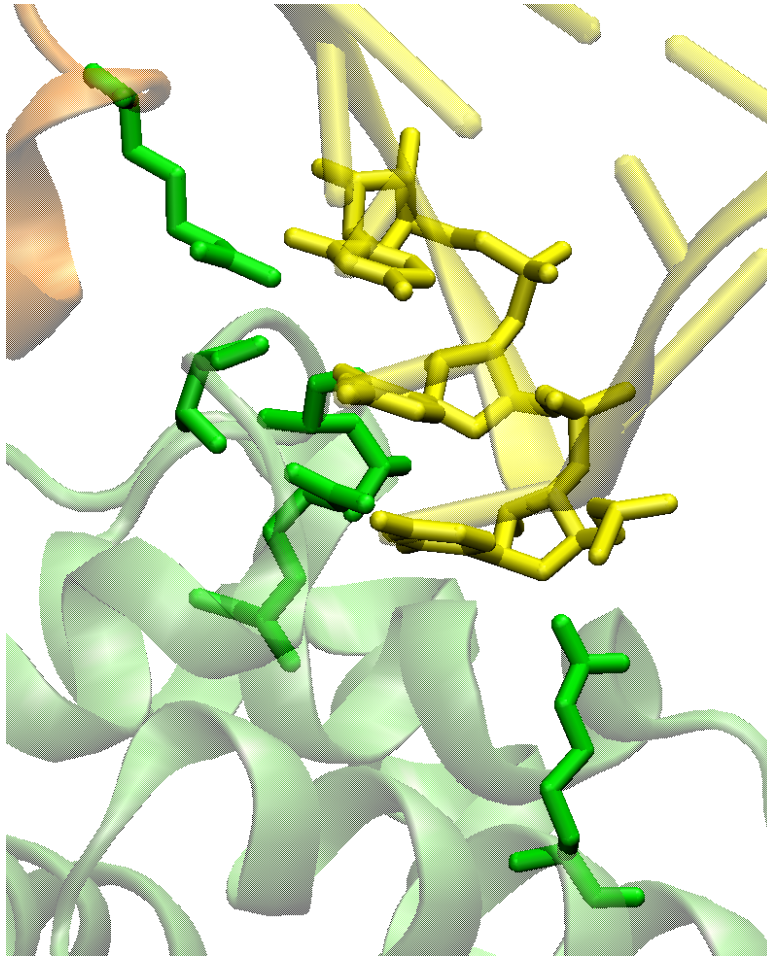
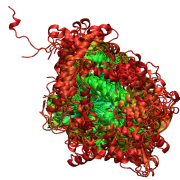
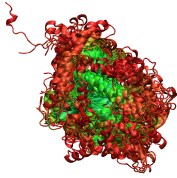


Figure 6: Residues of GluRS (green) within 3.0 Å of the anticodon (yellow)



The chemistry of aaRSs Explore the active site of the GluRS-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?



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 31 class I aaRS structures, representing 11 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 protein family. A sequence based phylogenetic analysis can be used to study recent phylogenetic events. However, sequence alignments are less reliable as the sequence identity reduces below 30% (twilight zone). On the other hand, a structural phylogenetic tree allows examination of more distant evolutionary events such as when specificity was being acquired. 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

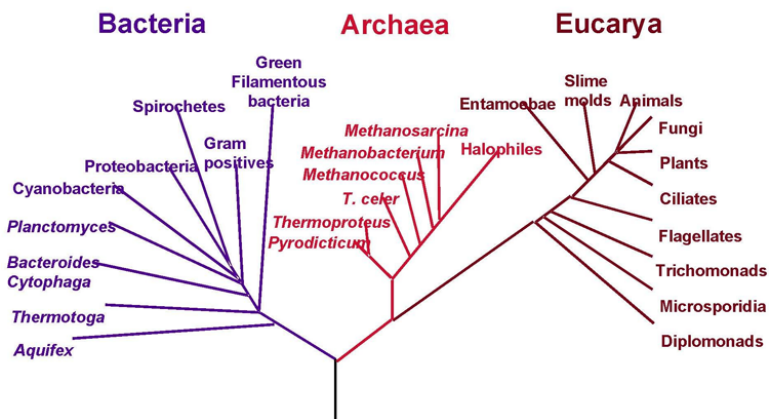


Figure 7: Universal tree of life

2.1 Loading Molecules

To further explore aaRSs, we will now examine the catalytic domain of 31 Class I aaRS 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-1-synthetases directory.

3. There are 31 PDB files you want to load from the directory. You may need to change the filename filter to allow for selection of PDB files¹. 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.
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.

The molecules will appear in the OpenGL display window. We will now walk through the steps for aligning these molecules.



What is the ASTRAL database? The ASTRAL database (<http://astral.berkeley.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, GluRS is divided into two separate PDB files: one containing the catalytic domain, and one for the anticodon binding domain. The names of the files contain the PDB code, the chain name, and number, which corresponds to the structural domain. For example, the anticodon binding domain for one of the GluRS-tRNA complex is: d1j09a1.

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.²

¹Note 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.

²The STAMP manual is available at <http://www.compbio.dundee.ac.uk/manuals/stamp.4.2/>

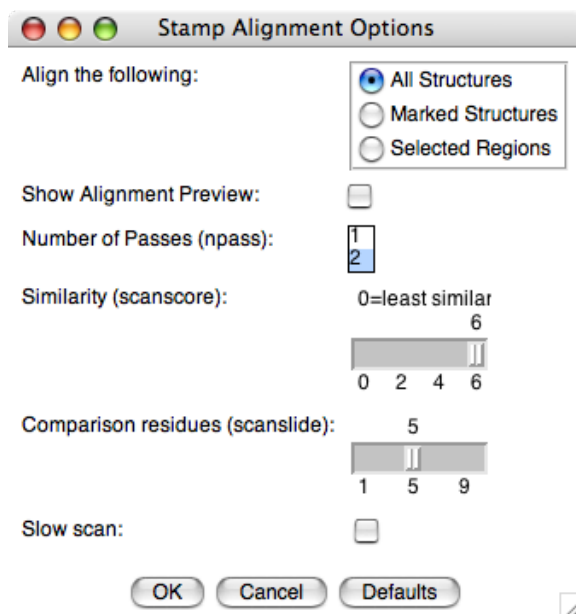



Figure 8: Stamp Alignment Options Window



How molecules are aligned in a multiple structural alignment

MultiSeq uses the program STAMP to align protein molecules. The STAMP algorithm minimizes the C_{α} 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 substructures, 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

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. In this tutorial we use the RGB (red-green-blue) color scale instead of the default RWB (red-white-blue) so that only gaps appear white in the alignment editor.

To change the color scale:

1. From the VMD Main window select Graphics → Colors... to bring up the Color Controls window.
2. In the Color Controls window select the Color Scale tab.
3. Choose RGB from the Method pick list.
4. Close the Color Controls window.



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 Coloring → Apply to all is checked and select Coloring → Qres.

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

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; while blue indicates identical structures ($Q_{res} = 1$) and red for unaligned parts ($Q_{res} = 0$), which often correspond to insertions that are unique to one specificity. For homologous proteins, $Q_{res} \approx 0.7$, hence they are colored bluish green.

You can also view secondary structure information derived from the crystal structures. In a structural alignment, α -helices and β -strands from a given protein should align with similar elements in the other proteins.

To view the secondary structures for the sequences in your alignment:

1. In the MultiSeq window select the sequences by clicking the name of the topmost sequence and then shift-clicking the name of the bottom sequence. All sequences should appear highlighted.

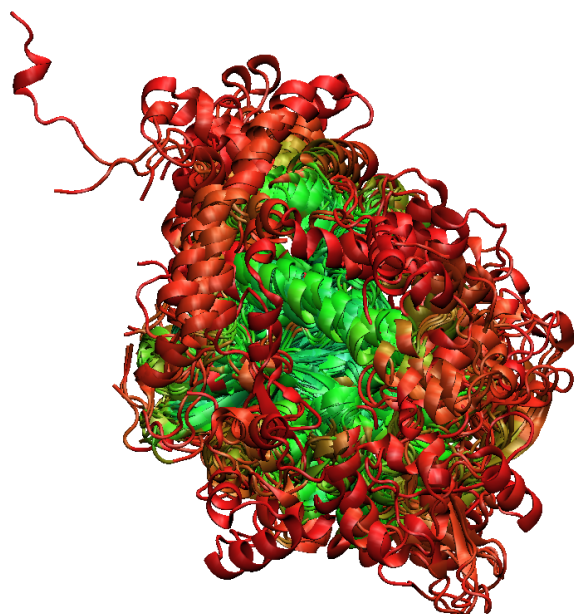
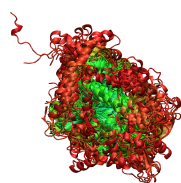


Figure 9: The catalytic domain colored by Q_{res} .

2. Click the `r` box to the right of one of the sequence names and choose Secondary Structure from the popup menu.

You should now be able to see picture representations of α -helices (wavy ribbons), β -strands (fat arrows), and coils (thin lines). Scroll through the alignment and look at how the secondary structure elements align. To view the sequences again follow the above instructions but choose Sequence from the popup menu.



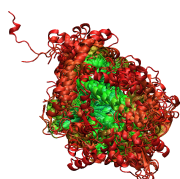
Core Structure Now that we have observed the structural conservation patterns, go back to the main MultiSeq window and see where the coloring of the core begins and ends. Since all the class I aaRS share a homologous core, you would expect the core residues should have a high Q_{res} value and have a green color in the alignment. Using the side-scroll on the bottom of the main MultiSeq window, you can see the core residues begins at about position 160 and ends around 1150 in the alignment (*notice that the position number in the alignment is not always the same as the residue number in each sequence, since the alignment contains gaps*). However, not all sequences in this region are core residues, many of them are insertions, which are characterized by a low Q_{res} value and thus appear to be red in the alignment. For example, there is a long insertion between position 830 to 880 for entry d1wkba3, which corresponds to *Pyrococcus horikoshii* LeuRS.

2.4 Structure Based Phylogenetic Analysis

2.4.1 Limitations of sequence data

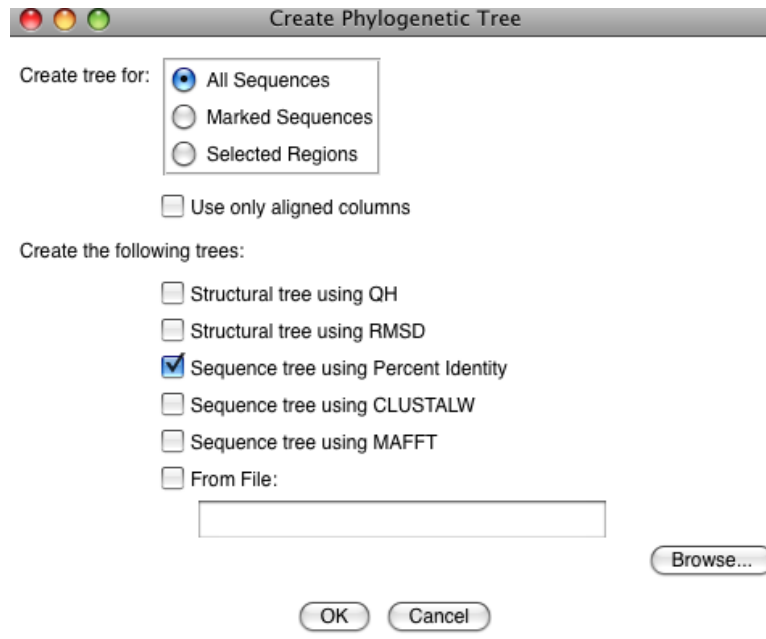
In this section we will look at the phylogenetic history of the class I aaRS structures. Most common methods of phylogenetic analysis use only information derived from the sequences to build the tree. However, the following two reasons restricted the application of these methods to highly divergent sequence data. First, a set of highly divergent sequences may not generate a reliable alignment, this is true for the case of class I aaRS, in which we have to apply the structural alignments, as we did in previous part. Second, many of the proteins we are looking at diverged before the last universal common ancestral state (LUCAS), 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 “midnight zone” of sequence identity and makes phylogenetic reconstruction using sequence metrics unreliable for very distantly related proteins. To demonstrate the second point, 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 → Taxonomy->Domain of Life and then View → Leaf Text → Enzyme->Name to show more information in the tree viewer.



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 d2ts1... to d1jila... is 0.38, or twice the distance to their closest common parent node. In this example, that distance represents 62% 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.

It is important to examine the phylogenetic tree we have built. At first glance, you may notice that many aaRSs with same specificity are grouped together, which is quite reasonable. The IleRS, LeuRS, ValRS and MetRS are grouped close to each other. Similarly, the GluRS and GlnRS, the TyrRS and TrpRS form two individual groups. This observation is consistent with the detailed classification of class I aaRSs[5]. Yet, a closer look brings more questions. For example, the ValRS groups within two IleRSs, which should form a monophyletic group by themselves. Also, you should 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. To resolve these problems, we are going to build a structure based phylogenetic tree.

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

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

1. Select the Tools → Phylogenetic Tree menu option.
2. In the Create Phylogenetic Tree dialog select the All Sequences radio button.
3. Make sure only the Structural tree using Q_H checkbox is checked and press the OK button.

MultiSeq calculates and displays the Q_H tree for the selected structural regions. Comparing this tree (shown in Figure 11) to the sequence tree generated earlier, the structure based tree retains most of the correct features within a given specificity. The structure based tree also makes some improvements on the phylogenetic relationship. For example, in the structure based tree, two IleRSs are grouped within a monophyletic group. You may also notice how the branch points are much more evenly spaced, not bunched together on the left of the tree. This indicates that the phylogenetic history is recorded in the structures, and it is elucidated when using the structural tree. However, the evolutionary relationship between TrpRS and TyrRS is still not well resolved. To overcome that problem, we need to compare the full length TrpRS and TyrRS³.

Our current tree is slightly different from the one we showed in the MMBR paper (page 561), since we are using different structure sets. As you can see, the old dataset contains fewer structures. Our current one is more balanced, less redundant, also has more representatives from the specificities/domains of life that were not resolved earlier.

³That is how we solve the problem in the MMBR paper.

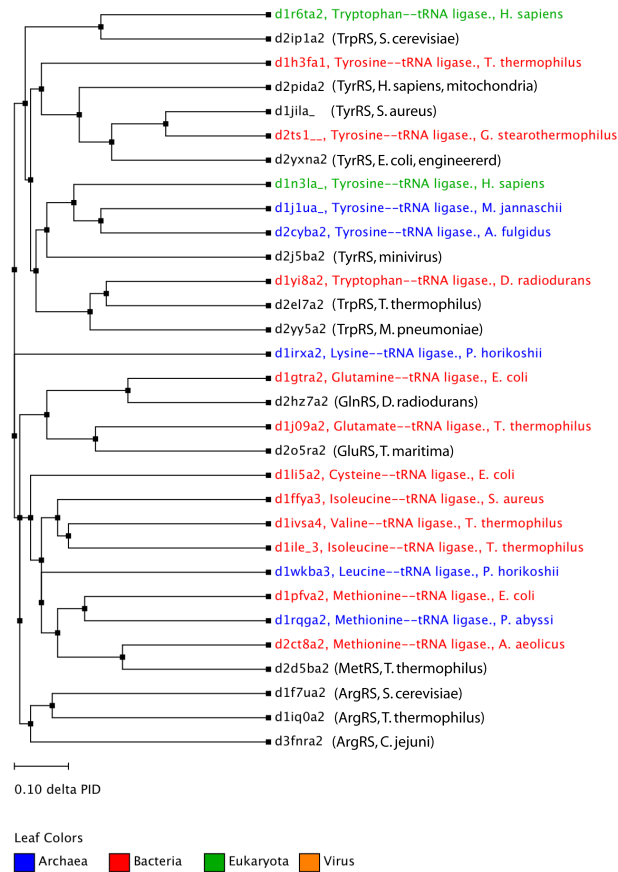


Figure 10: Percent identity sequence phylogenetic tree of 31 diverse aaRS structures. Note here that some aaRS entries do not contain information about their specificities and species names. We add the information manually (shown in parenthesis).

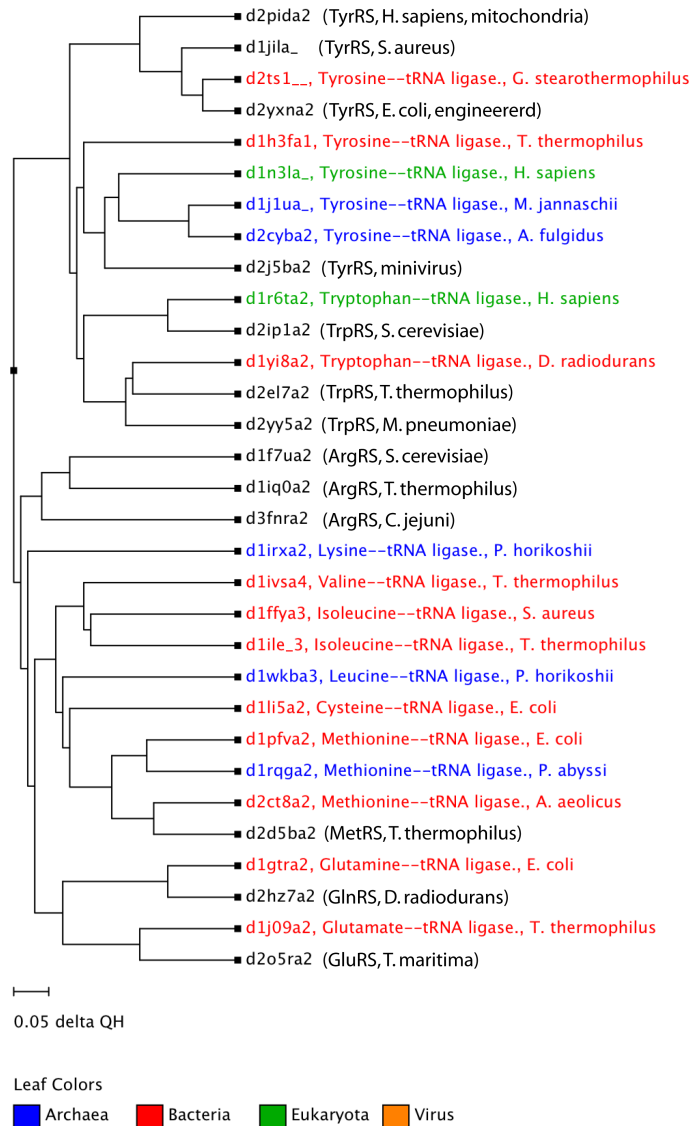


Figure 11: Q_H structural phylogenetic tree of 31 diverse class I aaRS structures. For those ASTRAL structures not present in the metadata, we provided the specificity and the species name of each aaRS manually (shown in parenthesis).

3 Complete Evolutionary Profile of TyrRS

3.1 Expanding the genetic code by engineering TyrRS

So far we have investigated only structures of the catalytic domain of the class I tRNA synthetases. Further analysis of the catalytic domain requires looking at sequences as well. For this tutorial we will be concentrating on one specificity from the class I aaRSs, the tyrosyl-tRNA synthetases.

Methanococcus jannaschii TyrRS is the first synthetase that has been engineered to incorporate an unnatural amino acid into protein in *E. coli* [18]. The overall strategy is to engineer an aaRS that can specifically aminoacylate a tRNA with an unnatural amino acid, and this tRNA (the suppressor tRNA) can deliver the amino acid to a specified position (an amber stop codon) on any gene. Schultz and his colleagues chose TyrRS for this purpose due to several reasons. First of all, archaeal and bacterial TyrRS recognize different parts on tRNA. In particular, the archaeal TyrRS recognizes the C1-G72 base pair, the discriminator base A73, and the anticodon loop; while the bacterial TyrRS relies on the G1-C72 base pair, A73, the anticodon loop as well as the long variable arm. This makes the bacterial TyrRS unable to charge archaeal tRNA^{Tyr} and vice versa. Secondly, TyrRS can recognize the suppressor tRNA, largely due to the similarity between the tyrosine codons (UAU and UAC) and the amber stop codon (UAG). Finally, TyrRS will not hydrolyze the charged unnatural amino acid.



The names of stop codons Stop codons were historically given many different names as they each corresponded to a distinct class of mutants. Amber mutations were the first set of nonsense mutations to be discovered, within bacteriophage T4. It is named after the graduate student, Harris Berstein, who first isolated these mutants (Berstein means "amber" in German). The ochre and opal mutants were isolated later, and their names were given to color names to match the amber mutants. It turned out later that the amber, orche, and opal mutants corresponds to the mutations to the stop codon "UAG", "UAA" and "UGA", respectively.

3.2 Comparing archaeal and bacterial TyrRS:tRNA complexes

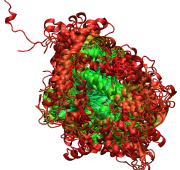
First, we start a new session of Multiseq and load the structure of the *M. jannaschii* TyrRS (PDB code 1J1U)[6] by importing the file `TUTORIAL_DIR/1j1u_dimer.pdb` into Multiseq. Note here although TyrRS forms a homodimer, the original PDB file contains only one molecule of TyrRS and one molecule of tRNA. For the purpose of the tutorial, we have built the dimer complex based on crystallography symmetry (P31 2 1 for 1J1U) by Swiss-PDB viewer⁴. Try to color the

⁴<http://us.expasy.org/spdbv/text/download.htm>

molecules by chain. To do that:

1. Open the Graphical Representations panel in VMD.
2. Change the color in the Color ID list for each chain so that they have different colors.

You can clearly see that a tRNA molecule spans the two subunits of the homodimer: the acceptor stem of the tRNA molecule interacts with one subunit; while the anticodon loop is recognized by the other. You might also notice that the crystal structure is missing some important parts of the TyrRS:tRNA complex, for example, the CCA end in the tRNA and the KMSK loop in the TyrRS. This is probably because that these regions are flexible in the absence of ATP.



How *M. jannaschii* TyrRS specifically recognizes the C1-G72 base pair. Try to use the method we described in the introduction to find out the residues that are responsible for the specificity. *Hint: you may first select the C1-G72 base pair in Multiseq and find residues in its contact shell.*

Next, we will examine the structure of *Thermus thermophilus* TyrRS (PDB code 1H3E)[19], which represents the bacterial type TyrRS. As above, we generated the homodimer of the TyrRS:tRNA complex based on crystallography symmetry. Here, you can load the file `TUTORIAL_DIR/1h3e_dimer.pdb` into Multiseq. You will notice that the *T. thermophilus* structure is significantly different from the archaeal structure: a long arm of tRNA (the variable arm) protrudes outward, and extensive contacts are formed between this part and protein. The archaea tRNA^{Tyr}, which does not have the long variable arm, will not bind stably to the bacterial TyrRS. You can also try to find how the G1-C72 base pair is specifically recognized by *T. thermophilus* TyrRS.

3.3 The structural basis of the altered specificity of the engineered TyrRS

In Schultz's paper, he and his colleagues reported that by introducing four point mutations (Y32Q, D158A, E107T, L162P), they can convert the TyrRS to specifically aminoacylate *O*-methyl-tyrosine. Subsequently, the crystal structure of the engineered TyrRS was solved[20]. In this section, we try to understand the structural basis of the altered specificity. To do this:

1. Delete all the old structures in the MultiSeq program and load three new structures: 1u7d (*apo* wild-type *M. jannaschii* TyrRS), 1u7x (*apo* engineered *M. jannaschii* TyrRS), and 1h3f (*T. thermophilus* TyrRS bound with a tyrosine analog).⁵

⁵As you may notice, the protein structure of 1h3f seems to contain two separate proteins. This is an artifact due to some missing residues in the crystal structure.

2. Delete the B chain for 1u7d and 1u7x, as well as the A chain for 1h3f.
3. Align these three structures by STAMP structural alignment. Color the structure by Qres. You should see that the majority of the catalytic site is blue, indicating a good alignment among these structures.
4. Select the last residue in 1h3f that is shown as a **X**. This is tyrosinol, an analog of tyrosine.
5. Select the four mutated residues: Y32, D158, E107 and L162 in 1u7d and their counterparts in 1u7x.

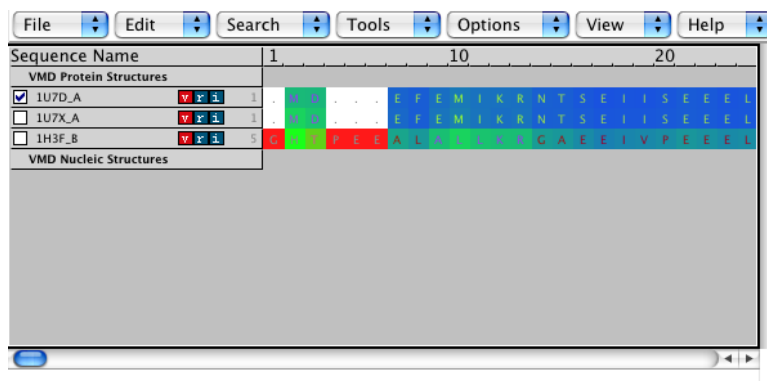
You can see that Y32Q and D158A enlarge the amino acid binding pocket directly, although the other two mutations are not close to the binding pocket and their mechanisms are not clear.

3.4 Evolutionary Profile of TyrRS

3.4.1 Importing the archaeal sequences

In this section, we will closely examine the difference between the archaeal and bacterial TyrRS as well as their evolutionary relationship by generating an evolutionary profile. To do this in MultiSeq, we will perform a BLAST search of a TyrRS structure from each domain of life against the Swiss-Prot database one at a time, starting with the Archaea. Doing the search separately within one domain of life will allow us to be more sensitive in finding only TyrRS sequences. To run the search:

1. In the MultiSeq program window select the 1U7D_A sequence as our source by marking it.

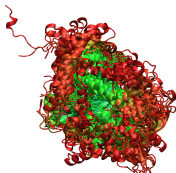


2. Click the File → Import Data menu option.
3. The Import Data dialog will appear. Select the From BLAST Search radio button.

4. In the Search Profile section select the Marked Sequences radio button.
5. Next to the Database field, press the Browse... button and select the file TUTORIAL_DIR/swiss-prot/uniprot_sprot to search over the Swiss-Prot database.
6. Set the E Score to be e-20 and the number of Iterations to be 1⁶.

The screenshot shows the 'Import Data' dialog box. The 'Data Source' section has 'From BLAST Search' selected. Under 'Search Profile', 'Marked Sequences' is selected. The 'Database' field contains '/tutorial_files/swiss-prot/uniprot_sprot'. The 'E Score' is set to 'e-20', 'Iterations' is '1', and 'Max Results' is '500'. There are 'Browse...' buttons for the Database field and a checkbox for 'Automatically download corresponding structures for sequence data'. At the bottom are 'OK', 'Cancel', and 'Defaults' buttons.

7. Now click the OK button. The search may take a minute or two.



E value The Expect value (E value) represents the probability that a certain match or a better one would be expected to occur purely by chance in a search of the entire database. Thus, the lower the E value, the greater the similarity between the input sequence and the match. For a more comprehensive description, you may read the following website: <http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>

⁶We usually set the E value threshold as e-3 or even higher. Here, the extremely low E value is used to screen out TrpRS sequences.

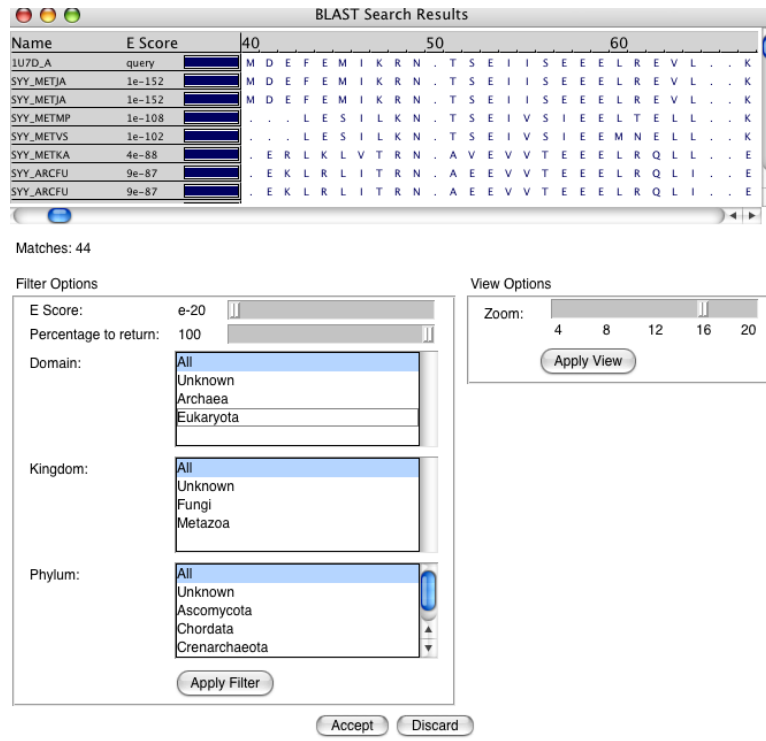


Figure 12: Blast Search Results Dialog

When the search is complete, a new dialog called BLAST Search Results appears (see Figure 12). As you may have noticed, 44 sequences were found 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 38 sequences from the domain Archaea. Press the Accept button at the bottom of the window to bring these sequences into MultiSeq.

3.4.2 Now the other two domains of life

Since there are no eukaryal cytoplasmic TyrRS structures available, we will still use the archaeal one as the seed of BLAST search. This time, select Eukaryota from the Domain list. You will find five eukaryal TyrRS. Bring all of them into Multiseq.

Now perform the same search for the bacterial structure by unmarking d1U7D_A, marking d1h3f_B, and then repeating the above steps with E Score set

to e-5 . Be sure to select **Bacteria** in the **Domain** list this time. This will bring in 270 sequences. You can immediately tell that the **Bacteria** are over-represented in the sequence databases⁷, and if you examine these sequences more carefully, you will notice that many of them are highly similar. Eliminating this bias and redundancy is important in obtaining good evolutionary profile and will be discussed later in more detail. Here, we first use the binary QR to screen out most of the redundant sequences⁸. To do this:

1. In the **Filter Options** section and in **Percentage to return** line, scroll the index to 20. This will give you 54 sequences.
2. Press **Apply Filter** button.
3. Press **Accept** button.

After you obtain the bacterial TyrRS sequences, you should make sure that all their names start with **SY**. Sometimes you may retrieve TrpRS sequences, which start with **SYW**. These sequences should be excluded in the following analysis.

3.4.3 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.

⁷Here we have used only a subset of the Swiss-Prot database. In reality, you will obtain even more sequences.

⁸Strictly speaking, the QR algorithm should be performed after the sequence alignment of all the available sequences. However, an alignment and the complete Sequence QR factorization of about 300 sequences will take a very long time. So, in order to finish the sequence alignment in a reasonable amount of time, we applied the binary QR method. In the binary QR method, all amino acids are encoded in a single dimension unlike Sequence QR which has a dimension for every amino acid. The second dimension in binary QR encodes the gap positions in the alignment. Binary QR calculates the most representative set of protein sequences based on the pattern of gaps in the BLAST alignment.

3.4.4 Aligning to a Structural Profile using ClustalW

Finally we have a set of sequences and structures of the catalytic domain of the tyrosyl-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. Here, we will build a more reliable structural alignment based on six TyrRS structures. The following steps will walk you through that process:

1. Delete all the loaded TyrRS structures.
2. Load all six PDB files from TUTORIAL_DIR/tyrRS/. They are 1vbma (*Escherichia coli* TyrRS), 1u7da (*Methanocaldococcus jannaschii* TyrRS), 2dlca (*Saccharomyces cerevisiae* TyrRS), 2cyba (*Archaeoglobus fulgidus* TyrRS), 2cyaa (*Aeropyrum pernix* TyrRS) and 1h3fa (*Thermus thermophilus* TyrRS), respectively.
3. These structures should appear in a new group called **VMD Protein Structures**, rename this group as **Structures** by right-click the group name and select **Rename Group...**, enter **Structures** as the name of the new group and press OK.
4. Mark all six structures.
5. Use STAMP to align the marked structures using the **Tools → Stamp Structural Alignment** menu option.
6. Check the quality of the structural alignment by coloring the residues by Q_{res} . You should notice that these structures can be aligned very well.
7. Unmark the structures and mark all of the sequences in the **Archaea**, **Bacteria** and **Eukaryota**. Now, we will align all the sequences to the structural alignment. **We would like to emphasize here that in research, it is better to align sequences from each domain of life separately to generate sequence profiles first and then align these sequence profiles using the structural alignment as a guide.**
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/Sequence Alignment** and tell ClustalW 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 structure based alignment of the tyrosyl-tRNA synthetase. Try coloring it according to sequence identity by choosing View → Coloring → Apply to All and then View → Coloring → Sequence Identity (shown in Figure 13). 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 13: MultiSeq showing all sequences colored by sequence identity

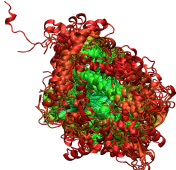
3.4.5 Curating the sequence alignment

After you obtained the sequence alignment, it is important to check it manually. Even the best alignment algorithm will make errors, especially if the sequences are divergent. A basic principle for alignment curating is that those functionally important motifs should be aligned (here, the HIGH and KMSK motifs). As you noticed in Figure 13, the HIGH motif is aligned well⁹. However, for some

⁹Note there is some variations among the HIGH motif. For example, you can see HLGH, HVGH in the alignment.

of the archaeal TyrRS, the KMSK motif (around position 310) are not aligned. To overcome this problem, you have to edit the sequence alignment manually by inserting some gaps:

1. Choose Edit → Enable editing → Gaps Only from the menu.
2. Align KMSK region by adding and deleting gaps. For example, for the entry **SYY_AERPE**, you need to delete five gaps right before the sequence **EIDDDVLAEVKMSKS** by pressing the “delete” button, and add five gaps after it by simply pressing space button. By doing that, we can align KMSK motif and still maintain the rest part of the alignment. Try to align the rest of KMSK motifs. The final alignment should look similar as Figure 14
3. Color the sequences by Sequence Identity again. The KMSK region should appear blue or green.



Conservation of important residues responsible for tyrosine recognition As you would imagine, those residues that are essential for discriminating tyrosine against other amino acids in the TyrRS active site should also be conserved in evolution. Check if Y32 and D158 are conserved, hence perfectly aligned in your alignment. Does that make sense to you? You may also notice that there are some other well aligned parts. Try to mark them out in the structure and think about why they are conserved.

Now we are ready to make a phylogenetic tree.

3.4.6 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.

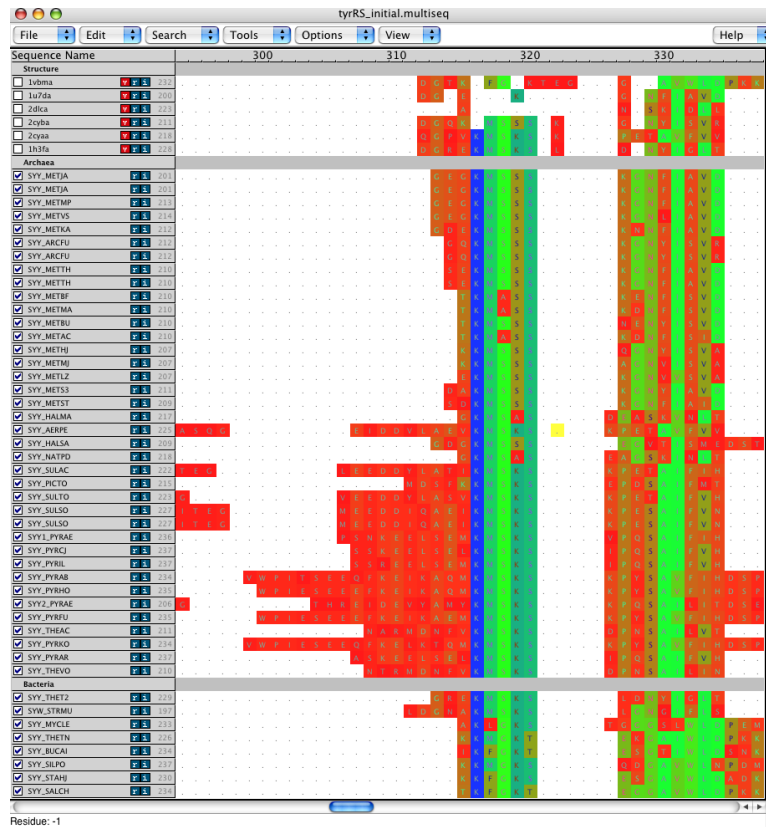


Figure 14: Curating the sequence alignment around KMSK motif

5. Set the Maximum PID (maximum percent identity) to be 50.
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** → **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 non-redundant sequences.

3.4.7 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

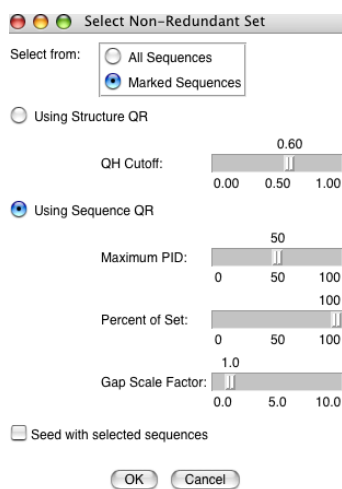


Figure 15: Select Non-Redundant Set dialog

3. Select Create Tree for → All Sequences within the window and check Sequence tree using Percent Identity
4. Press the OK button.

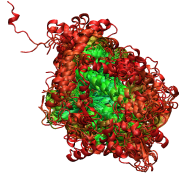
Another window will appear with the dendrogram. Within the new window select the following:

- View → Leaf Color → Taxonomy->Domain of Life
- Turn on View → Leaf Text → Name, Enzyme->Name, and Taxonomy->Species.

The tree should appear as shown in Figure 16.

3.4.8 Insights from the evolutionary profile

Now it is time to think about our results. You can see directly from the tree that the TyrRSs from each domain of life are grouped together, and the eukaryal TyrRSs are more close to the archaeal ones. These suggest that the evolution of TyrRS conforms to the canonical pattern, i.e., there is no horizontal gene transfer between domains of life. This evolutionary profile has also some practical usage. For example, if you are going to expand the genetic code in an eukaryotic system, which one are you going to use, a bacterial TyrRS or an archaeal one? Is it still wise to engineer the *M. jannashii* TyrRS? What do you think of Schultz and his colleagues' choice[2]?



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.^a

^aJ. Felsenstein *Inferring Phylogenies*. Sinauer Associates, Inc.: 2004.

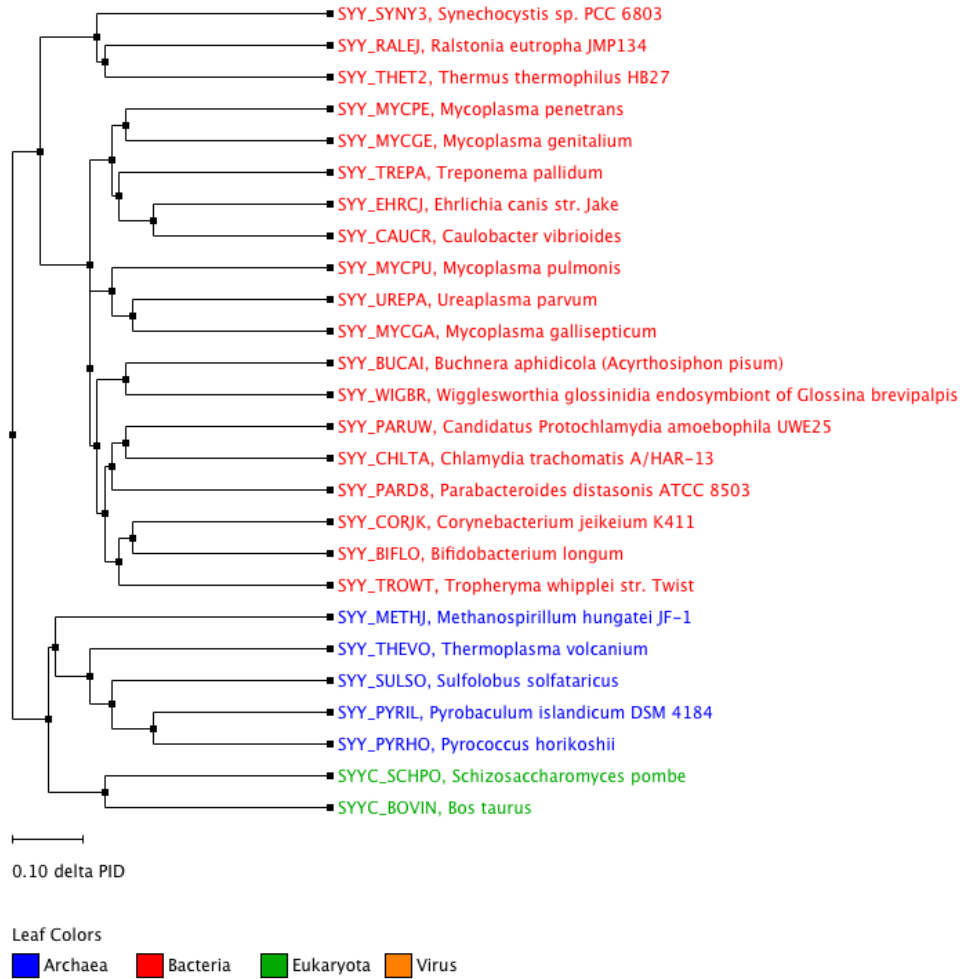


Figure 16: Phylogenetic tree based on sequence identity

3.5 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.

3.6 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 aaRS part 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).

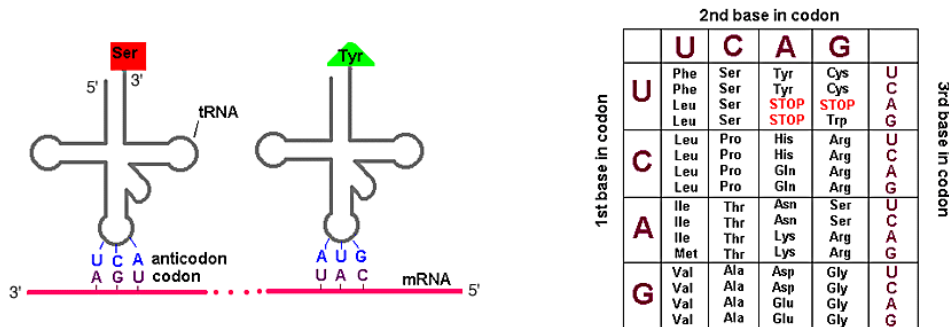


Figure 17: tRNA and genetic code

1. Start VMD and MultiSeq.
2. Change the color scheme to RGB.
3. Load 1ASZ-tRNA_ScEr_D.E.pdb (tRNA^{Asp}) into MultiSeq using Import Data. The file is located here:
/Tutorials/class-I/tutorial-files/trna/
4. 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.

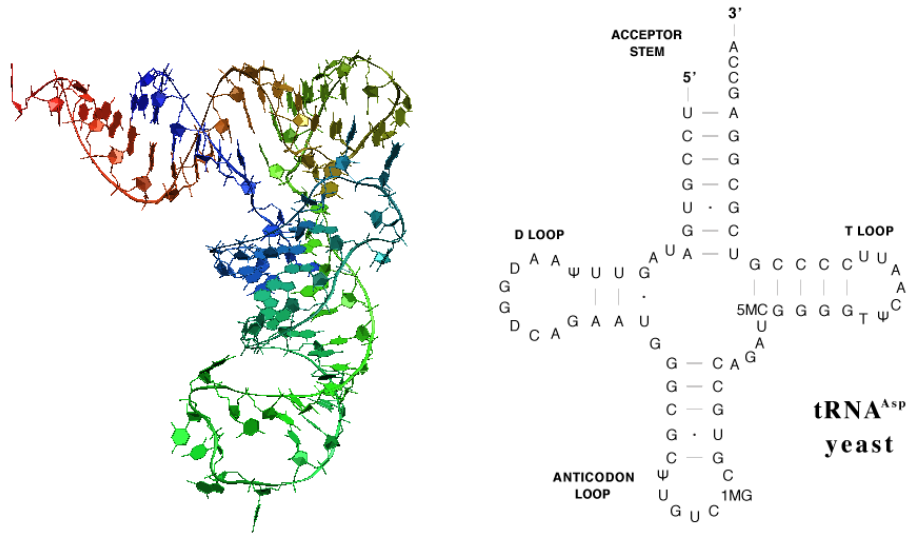
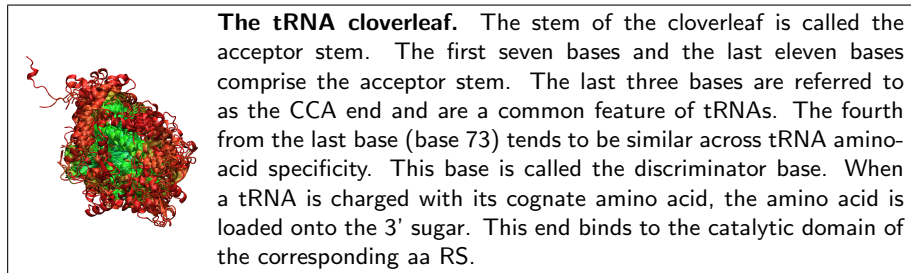


Figure 18: 3D and Cloverleaf view of tRNA

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.
5. Return to the main VMD window and open the Graphics→Representations window. Color the molecule through by Index.
6. Change the highlighting style through View→Highlight Style→Bonds.
7. Highlight the first seven residues in the alignment window.

- Next highlight the last eleven residues of the sequence.



- Next highlight columns 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.

- Highlight columns 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 GUC. Highlight columns 34 to 36 to reveal the anticodon.

- Highlight columns 48 to 64.

The last leaf of the cloverleaf structure is the T arm, so-called because it contains the T Ψ C sequence motif at the 5' end of the T loop. Highlight columns 53 to 55 to see the T Ψ 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 beside the sequence name) for examples of this.

1. Structurally align the tRNAs using Tools→Stamp Structural Alignment with default values. You can set default values by pressing Defaults button.
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 AspRS. 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.

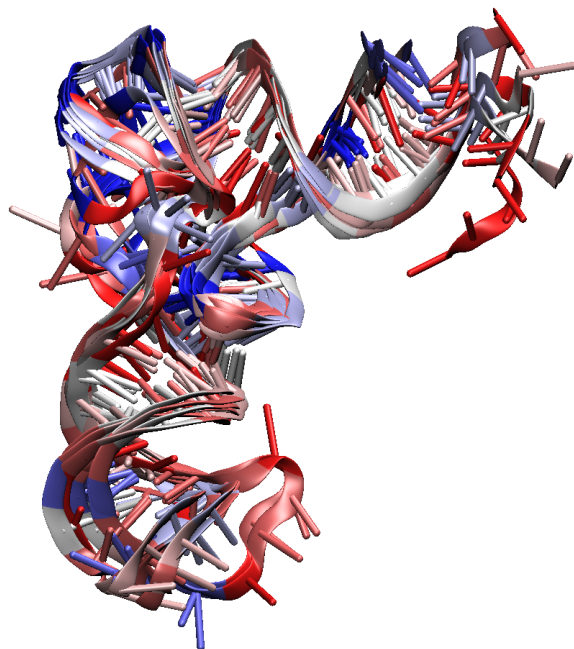


Figure 19: Seven tRNAs aligned with STAMP. Each nucleotide is colored by sequence identity in the alignment.

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

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 (1B23-tRNA_EColi_C_B.pdb) comes from a complex with EF-Tu and has already been charged with a cysteine.

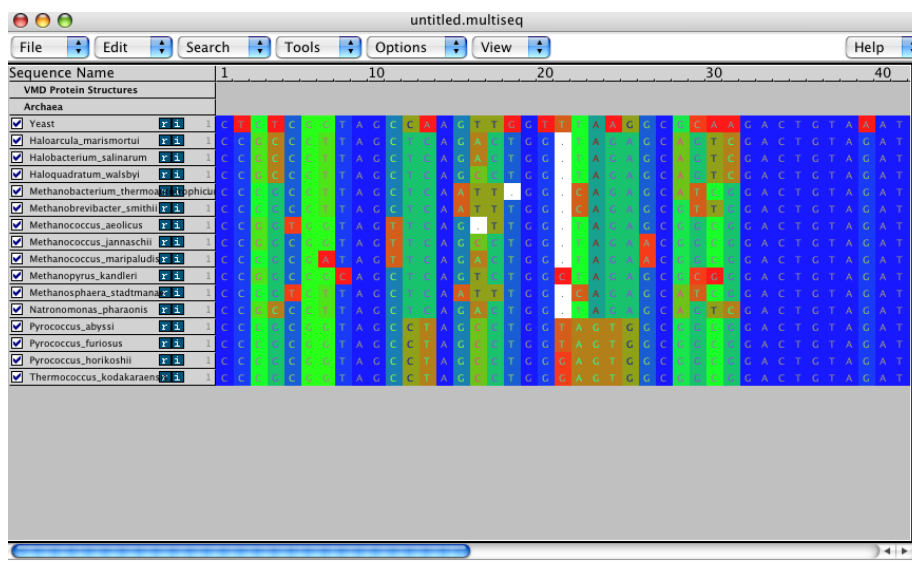
4.4 Sequence Alignment

Bring in tRNA^{Tyr} sequences through **File**→**Import Data** (Gtrnadb.fasta). These data are from genomic tRNA sequences from the Genomic tRNA Database (<http://lowelab.ucsc.edu/GtRNAdb/>)¹⁰. They are genes sequenced from DNA and so have no information about modified base.

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 exception is the yeast tRNA^{Tyr} 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.)
4. Color the alignment by sequence identity through **View**→**Coloring**→**Sequence Identity**.
5. 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?
6. Rename the group with the secondary structure information “Secondary Structure” to keep it separate from the gene sequences.
7. Bring in prealigned tRNA^{Tyr} sequences through **File**→**Import Data** (Bayreuth_tyr.fasta).

¹⁰We only use a subset of total archaea tRNA^{Tyr}.



8. 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>)¹¹. This alignment was made using the standard tRNA numbering which takes basepairing information into account.
9. 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.

1. First, you will remove the gaps from the Archaea alignment.
2. Mark only the sequences in the Archaea group.
3. Now remove the gaps through Edit→Remove Gaps... (Marked Sequences, All gaps).
4. 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.

¹¹This is only a subset of all the tRNA^{Tyr} from the Bayreuth database. We also modified species name manually so that they can be recognized by Multiseq.

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 of tRNA^{Tyr}

Now we start to build a phylogenetic tree of tRNA^{Tyr}. To do that,

1. Delete the yeast tRNA in the Archaea group.
2. Create a non-redundant set of tRNA^{Tyr} with Maximum PID set to 80. Move these sequences into a new group named “NR set”.
3. Select all the sequences except the last three columns (the CCA end)¹².
4. Create a sequence-based phylogenetic tree through Tools→Phylogenetic Tree. Using Selected Regions and Sequence tree using Percent Identity.
5. In the Tree Viewer window, choose View → Leaf Color → Taxonomy->Domain of Life

You will notice the bacterial tRNA^{Tyr} form a monophyletic group while the archaeal and eukaryal tRNA^{Tyr} are mixing together. This is not totally unexpected, since a phylogenetic tree of tRNA is based on an alignment composed of only 76 nucleotides, which contains much less information than in a typical protein or the ribosomal RNA alignment. The clear division between the bacterial tRNA^{Tyr} against the tRNA^{Tyr} from two other domains of life is largely attributed to an insertion (the variable arm) that is unique to bacterial tRNA^{Tyr}. Try to find this insertion in the sequence alignment. Does that remind you something we have mentioned in the previous part?

You have now completed this tutorial. We hope you find it interesting and have learned something from these bioinformatic analysis. In this tutorial, we focus on the molecules responsible for the aminoacylation reaction, which is the first step in protein synthesis. The aminoacyl-tRNA is then transported to ribosome for protein synthesis, which is mediated by elongation factor Tu. We will cover these topics in the next two tutorials.

5 Acknowledgments

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

¹²Here we left out the CCA end because it is not necessarily present in the DNA sequence of tRNA gene. Some organisms will add the CCA sequence to the tRNA after transcription.

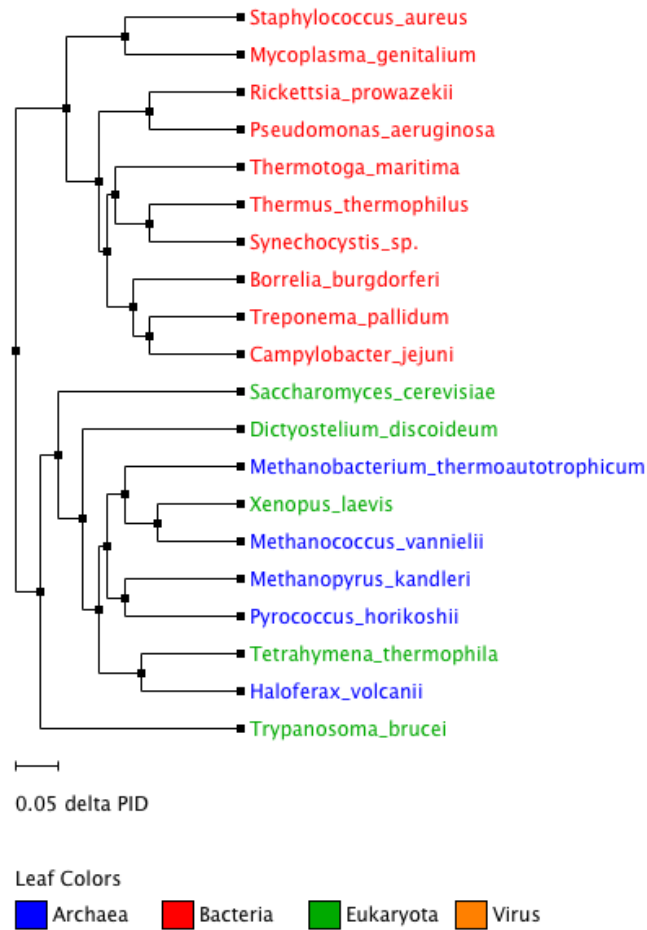


Figure 20: A phylogenetic tree of tRNA^{Tyr} based on sequence percentage identity

6 Appendices

6.1 Appendix A: Q

Q is a structure-based metric that was developed by Wolynes, Luthey-Schulten, and coworkers to study protein folding. It computes the fraction of similar contact distances between any conformation of a protein and its native structure (typically its X-ray or NMR structure). 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^α (or P) atoms.

r_{ij}^{nat} is the C^α - C^α (or P - P) distance between residues i and j in the native state of a protein (or RNA).

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

N is the number of residues of the protein (or RNA) being considered.

In MultiSeq, Q has been generalized to measure the fraction of similar contact distances between all the aligned residues in two homologous proteins. This term computes the fraction of $C_\alpha - C_\alpha$ (or $P - P$) pair distances that are the same or similar between two aligned structures.

6.2 Appendix B: Q_H

The following text is in the article “On the evolution of structure in aminoacyl-tRNA synthetases.” [7, 9].

Homology Measure

In addition to RMSD, we employ a structural homology measure based on Q defined by differences in pairwise residue distances r_{ij} which was developed by Wolynes, Luthey-Schulten, and coworkers in the field of protein folding [4]. Our adaptation of Q is referred to as Q_H (where H stands for homologs), and the measure is designed to include the perturbations due to gaps on the aligned region of the protein: $Q_H = \aleph(q_{aln} + q_{gap})$, 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:

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

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

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

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

This term computes the 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 the protein “a”, and $r_{i'j'}$ is the $C_\alpha - C_\alpha$ distance between residues i' and j' in the protein “b”. This term is restricted to aligned positions, e.g. where i is aligned to i' and j is aligned to j' , and the summation is over all unique, non-nearest neighbor residue pairs.

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 . In constructing the q_{gap} term, we hypothesized that the more the gap residues deviated from the nearest gap edge, the lower the value of structural similarity between the two proteins. In protein “a”, therefore, the

contact distance, $r_{g_a j}$, between a residue j and the gap residue g_a , is compared with the contact distances, $r_{g'_a j'}$ and $r_{g''_a j'}$, between residue j' of protein “b”, which is aligned to residue j , and the gap edges, represented by residues g'_a and g''_a in protein “b”. The “max” function takes whichever gap edge, g'_a or g''_a , that produces a larger contribution to Q_H . The outer summation is over all inserted residues in protein “a”, g_a , while the inner summation is over all non-nearest neighbor aligned residues. The definition is analogous for insertions in protein “b”.

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

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

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

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_{c_gaps} 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.

6.3 Appendix C: Q_{res} Structural Similarity per Residue

Here we define another metric, called Q_{res} , that is derived from Q which is used to measure the structural conservation of the environment of each residue in the alignment. Q_{res} is a measure of the similarity of the C_α - C_α distances between a particular residue and all other aligned residues, excluding nearest neighbors, in a set of aligned 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_α - C_α distance between residues i and j in protein n and $r_{i'j'}^{(m)}$ is the C_α - C_α 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.

References

- [1] S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, Sep 1997.
- [2] J. W. Chin, T. A. Cropp, J. C. Anderson, M. Mukherji, Z. Zhang, and P. G. Schultz. An expanded eukaryotic genetic code. *Science*, 301:964–967, Aug 2003.
- [3] J. Eargle, A. Black, A. Sethi, L. Trabuco, and Z. A. Luthey-Schulten. Dynamics of Recognition between tRNA and Elongation Factor Tu. *J. Mol. Biol.*, 377(5):1382–1405, 2008.
- [4] M. P. Eastwood, 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.
- [5] M. Ibba and D. Soll. Aminoacyl-tRNA synthesis. *Annu. Rev. Biochem.*, 69:617–650, 2000.
- [6] T. Kobayashi, O. Nureki, R. Ishitani, A. Yaremchuk, M. Tukalo, S. Cusack, K. Sakamoto, and S. Yokoyama. Structural basis for orthogonal tRNA specificities of tyrosyl-tRNA synthetases for genetic code expansion. *Nat. Struct. Biol.*, 10:425–432, 2003.
- [7] P. O’Donoghue and Z. Luthey-Schulten. On the evolution of structure in the aminocyl-tRNA synthetases. *Microbiol. Mol. Bio. Rev.*, 67:550–573, 2003.
- [8] P. O’Donoghue and Z. Luthey-Schulten. Evolutionary profiles derived from the QR factorization of multiple structural alignments gives an economy of information. *J. Mol. Biol.*, 346:875–894, Feb 2005.
- [9] P. O’Donoghue and Z. Luthey-Schulten. Evolutionary profiles derived from the qr factorization of multiple structural alignments gives an economy of information. *J. Mol. Biol.*, 346:875–894, 2005.
- [10] P. O’Donoghue, A. Sethi, C. R. Woese, and Z. A. Luthey-Schulten. The evolutionary history of Cys-tRNACys formation. *Proc. Natl. Acad. Sci. U.S.A.*, 102:19003–19008, Dec 2005.
- [11] E. Roberts, A. Sethi, J. Montoya, C. R. Woese, and Z. Luthey-Schulten. Molecular signatures of ribosomal evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 105:13953–13958, Sep 2008.
- [12] R. B. Russell and G. J. Barton. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, 14:309–323, Oct 1992.

- [13] A. Sauerwald, W. Zhu, T. A. Major, H. Roy, S. Palioura, D. Jahn, W. B. Whitman, J. R. Yates, M. Ibba, and D. Sll. RNA-dependent cysteine biosynthesis in archaea. *Science*, 307:1969–1972, Mar 2005.
- [14] S. Sekine, O. Nureki, A. Shimada, D. G. Vassylyev, and S. Yokoyama. Structural basis for anticodon recognition by discriminating glutamyl-tRNA synthetase. *Nat. Struct. Biol.*, 8:203–206, Mar 2001.
- [15] A. Sethi, J. Eargle, A. A. Black, and Z. Luthey-Schulten. Dynamical networks in tRNA:protein complexes. *Proc. Natl. Acad. Sci. U.S.A.*, 106:6620–6625, Apr 2009.
- [16] A. Sethi, P. O’Donoghue, and Z. Luthey-Schulten. Evolutionary profiles from the qr factorization of multiple sequence alignments. *Proc. Natl. Acad. Sci. USA*, 102:4045–4050, 2005.
- [17] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, Nov 1994.
- [18] L. Wang, A. Brock, B. Herberich, and P. G. Schultz. Expanding the Genetic Code of Escherichia coli. *Science*, 292(5516):498–500, 2001.
- [19] A. Yaremchuk, I. Kriklivyi, M. Tukalo, and S. Cusack. Class I tyrosyl-tRNA synthetase has a class II mode of cognate tRNA recognition. *EMBO J.*, 21:3829–3840, 2002.
- [20] Y. Zhang, L. Wang, P. G. Schultz, and I. A. Wilson. Crystal structures of apo wild-type *M. jannaschii* tyrosyl-tRNA synthetase (TyrRS) and an engineered TyrRS specific for O-methyl-L-tyrosine. *Protein Sci.*, 14:1340–1349, May 2005.