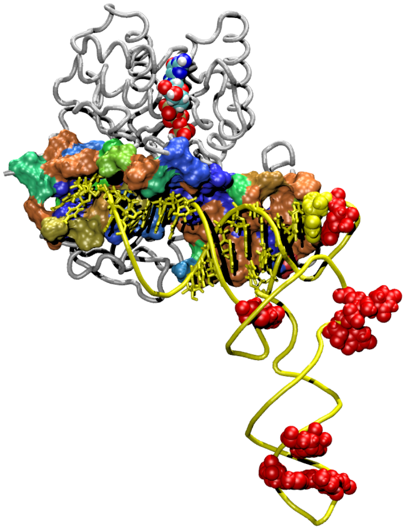
University of Illinois at Urbana-Champaign
Luthey-Schulten Group
NIH Resource for Macromolecular Modeling and Bioinformatics
Computational Biophysics Workshop

Evolution of Translation
EF-Tu: tRNA



VMD Developer: John Stone

MultiSeq Developers

Elijah Roberts
John Eargle
Dan Wright

Tutorial Authors

Ke Chen
John Eargle
Zhaleh Ghaemi
Jonathan Lai
Zan Luthey-Schulten

August 2014

A current version of this tutorial is available at
`http://www.scs.illinois.edu/~schulten/tutorials/ef-tu`

Contents

1 Introduction

1.1 The Elongation Factor Tu

The mechanism of translation, by which proteins are produced in the cell, is one of the oldest and most fundamental cellular processes. Coding regions of DNA are transcribed to mRNA molecules, which will ultimately be translated into protein by the ribosome. Proteins are composed of linear chains of amino acids joined with peptide bonds between adjacent amino and carboxyl groups. The ordering of amino acids is directly translated from the sequence of mRNA nucleotides, ultimately tracing back to the DNA sequence. The translation between *codons*, triplets of nucleotides, and individual amino acids is known as the *genetic code*. This translation is mediated by molecules in the cell known as transfer RNAs, or tRNAs. The tRNAs each bind a specific amino acid and also recognize a particular codon on the mRNA in the ribosome. The genetic code is maintained by the enzymes which 'charge' each tRNA with a specific amino acid, the aminoacyl- tRNA synthetases (AARS). The recognition element on the tRNA that interacts with a codon on the mRNA is known as an 'anticodon', and is also a triplet of nucleotides. There are three binding sites for tRNA inside the ribosome: the aminoacyl (A) site, peptidyl transferase (P) site, and the exit (E) site.

The elongation factor Tu (EF-Tu), the subject of this tutorial, acts as a ferry for the tRNA from the AARS to the ribosome. **Note:** The molecule in bacteria and eukaryotes known as EF-Tu has a different name in the archaea and eucarya: EF1A. These two terms are used extensively in this tutorial; they refer to the same molecule in different domains of life.

With the number of copies in *Escherichia coli* estimated to be 10 times the number of ribosomes, elongation factor Tu is one of the most abundant proteins in the cell. The EF-Tu binds promiscuously to any tRNA that has been aminoacylated. The enzyme protects the ester linkage between the tRNA and is cognate amino acid *en route* to the ribosome. The EF-Tu:tRNA complex then associates with the ribosome, allowing the anticodon region of the tRNA to associate with the codon of the mRNA in the A-site of the ribosome. If the codon-anticodon pairing is correct, the EF-Tu undergoes a conformational change, hydrolyzing a bound GTP (guanosine triphosphate) molecule to GDP, and releasing the tRNA fully into the ribosome. If the codon-anticodon pairing is not correct, the EF-Tu:tRNA dissociates from the ribosome with a high probability. Therefore the EF-Tu acts both as a ferry for tRNA and maintains accuracy in the translation process. More details on the translation process in the ribosome can be found in the specific tutorial on the ribosome: Evolution of Translation: Ribosome.

This tutorial is designed around the evolutionary analysis of EF-Tu:tRNA. Typically such an analysis is a prelude to further energetic and network analyses. Throughout this tutorial you will be learning about the advanced features of MultiSeq, including running BLAST searches on genomes, creating QR representative sets, performing ClustalW sequence and profile alignments, creating maximum likelihood phylogenetic trees with RAxML, and finally making use of the new TCL scripting functionality of MultiSeq. We include with this tutorial a copy of the paper **Dynamics of Recognition between tRNA and Elongation Factor Tu** by Eargle *et al* [2] in the directory 1.EFTu_paper, and we will refer to this paper as we examine the conservation of the EF-Tu in binding the tRNAs. This tutorial should take approximately 1.5 hours to complete.

1.2 Getting Started

1.2.1 Requirements

MultiSeq must be correctly installed and configured before you can begin using it to analyze the EF-Tu. There are a few prerequisites that must be met before this section can be started:

- VMD 1.9.1 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 1.5 GB of free space on your local hard disk.

1.2.2 Copying the tutorial files

This tutorial requires certain files which can be downloaded from the tutorial website.

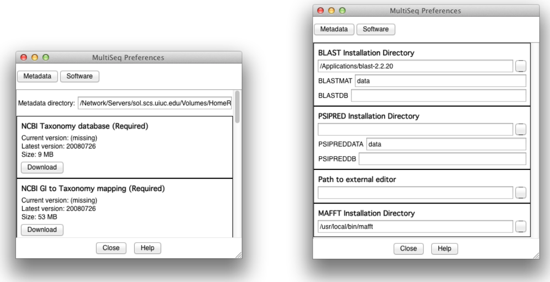
You should copy this entire directory to a location on your local hard disk. The path to the directory must not contain any spaces. For the remainder of this tutorial, this directory on your local drive will be referred to as TUTORIAL_DIR.

1.2.3 Working directory

A directory TUTORIAL_DIR/working_directory has been created in the tutorial directory structure. We will refer to this directory repeatedly throughout the tutorial. You can save all your intermediate files to this directory, and they will all be in one place when you need them later. If you cannot complete a section, the end of each section will inform you what file you should copy to your working_directory in order to continue with the tutorial.

1.2.4 Preferences

The directory and path variables needed to configure Multiseq are defined in the Preference Menu (Figure 1). Please check that all of the variables are pointing to the correct directories on your local machine.



Metadata directory: /Network/Servers/sol.scs.uiuc.edu/Volumes/HomeR

NCBI Taxonomy database (Required)
Current version: (missing)
Latest version: 20080726
Size: 9 MB

NCBI GI to Taxonomy mapping (Required)
Current version: (missing)
Latest version: 20080726
Size: 53 MB

BLAST Installation Directory
/Applications/blast-2.2.20
BLASTMAT data
BLASTDB

PSIPRED Installation Directory
PSIPREDDATA data
PSIPREDDB

Path to external editor

MAFFT Installation Directory
/usr/local/bin/mafft

Figure 1: Multiseq preferences windows: a) Metadata b) Software

1.3 Configuring BLAST for MultiSeq

The instructions for configuring MultiSeq and BLAST are available in the previous tutorial **Evolution of Translation: Class I Aminoacyl-tRNA Synthetases:tRNA complexes**. Note that we are using a legacy version of BLAST instead of BLAST+.

2 Comparative Analysis of EF-Tu

Modern comparative genomics is a relatively new field - it was only born after the appearance of the first complete bacterial genome, *Haemophilus influenzae*, in 1995. Since then bacterial genomes have been sequenced at an exponential rate with a doubling time of about 20 months, and archaeal genomes have been sequenced with a doubling time of about 34 months [7]. With this rapid growth in the number of complete genomes, we are provided with the information of not only which genes are present (in order to identify the "orthologs" between different genomes) or present in multiple copies (in order to identify the "paralogs" within one genome) in any particular genome but also which ones are absent. By exploring the structure of bacterial gene space, researchers have developed the notion of clusters of orthologous genes (COGs) [3,8] and observed that only a small number of genes and universally conserved acrossed all three domains of life. One such gene, EF-Tu, is key to translation.

There are more than 30,000 bacterial and archaeal genomes available on the NCBI site and DOE JGI. The information is available in many different formats, depending on the particular needs of the researcher. We have provided you with a single BLAST database containing select bacterial and archaeal genomes in the directory 1.2.blast_database. You will be using this database to search for the EF-Tu from all of these organisms in order to perform an evolutionary analysis on them. The creation of this database is beyond the scope of this tutorial. For users interested in building BLAST databases from FASTA files, an appendix 5 is provided at the end with detailed instructions.

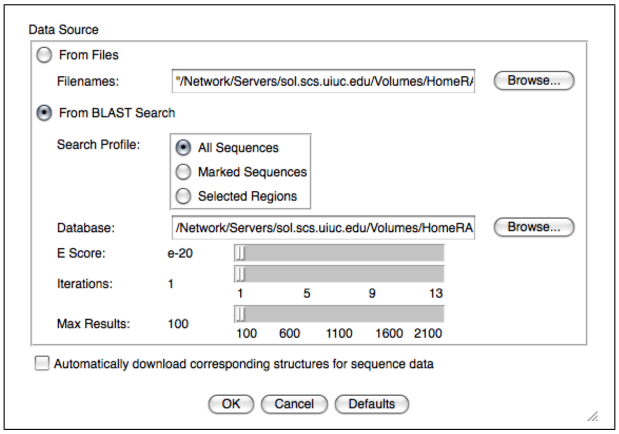
In order to perform a comparative analysis on the EF-Tu between bacteria and archaea, we will first load BLAST results based on a archaeal sequence into MultiSeq and perform a multiple sequence alignment. We will repeat this process for a set of bacterial sequences. Finally, we will perform a profile-profile alignment between the two domains of life to determine the conserved residues between them.

2.1 Finding archaeal EF1A sequences

BLAST. BLAST is an acronym for Basic Local Alignment Search Tool. It was originally developed by Steven Altschul, Eugene Myers, and others at the NIH for rapid searching of biological databases using a protein or nucleotide query sequence [1]. BLAST provides a much faster, but less accurate alignment than those obtained by standard dynamic programming alignment algorithms such as Needleman-Wunsch (global) and Smith-Waterman (local).

1 Open VMD and click Extensions → Analysis → MultiSeq. We are going to load an archaeal sequence EF1A sequence in order to perform a BLAST search over all sequenced genomes of bacteria and archaea. Click on File → Import Data. Make sure the From Files radio button is selected, then click the Browse button.

2 Navigate to the 1.1.blast_sequences directory and select the file `archaea_EF1A.fasta`. This file contains a single sequence of an archaeal EF1A which we will use to perform the BLAST search. Make sure Automatically download corresponding structures for sequence data is unchecked, and then click OK. The sequence will appear in the MultiSeq main window.



Data Source

○ From Files

Filenames: "/Network/Servers/sol.scs.uiuc.edu/Volumes/HomeR/

● From BLAST Search

Search Profile: ● All Sequences ○ Marked Sequences ○ Selected Regions

Database: /Network/Servers/sol.scs.uiuc.edu/Volumes/HomeRA

E Score: e-20

Iterations: 1

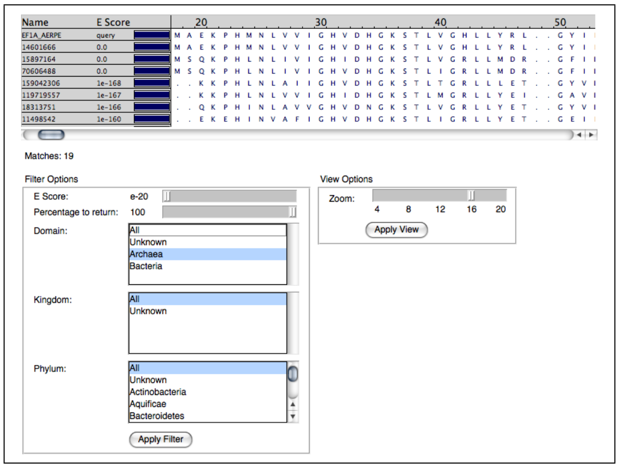
Max Results: 100

☐ Automatically download corresponding structures for sequence data

Figure 2: Import Data Window

3 Select File → Import Data. Select the From BLAST Search radio button, and make sure All Sequences is marked. Now click Browse next to the Database field. Navigate to the 1.2.blast_database directory, and select the file AB_genomes.faa.psq. (Actually, it can be any of the files with extension .pXX). Click OK. Move the E Score slider all the way to the left so that it reads e-20. The E Score, or expectation score, is a measure of the number of different alignments with scores equivalent to or better than the actual alignment score to occur by chance. The smaller the E Score, the lower the probability that this alignment occurred by chance, and the more significant the alignment. Move the Max Results slider until it reads 100. Make sure Automatically download corresponding structures for sequence data is unchecked and then click OK. The BLAST search should take less than a minute. When it completes, the BLAST Search Results will appear on the screen

4 When the BLAST Search Results window appears, take a look at the results it has returned by scrolling down in the top section. Note that the E Scores returned are all smaller than your cutoff of e-20. Many of these results are actually bacterial sequences, as the number of bacterial genomes far outweighs the number of archaeal genomes at the time of this writing. Therefore we must filter these results to only return archaea sequences. In the Domain window, unselect All and select Archaea, then click the Apply Filter button. You will see that the number of results is reduced to 19. Click the Accept button.



Name	E Score	20 … 30 … 40 … 50
EF1A_AERPE	query	M A E K P H M N L V V I G H V D H G K S T L V G H L L Y R L . . G Y I
14601666	0.0	M A E K P H M N L V V I G H V D H G K S T L V G H L L Y R L . . G Y I
15897164	0.0	M S Q K P H L N L I V I G H I D H G K S T L V G R L L M D R . . G F I I
70606488	0.0	M S Q K P H L N L I V I G H V D H G K S T L I G R L L M D R . . G F I I
159042306	1e-168	. . K K P H L N L A I I G H V D H G K S T L T G R L L L E T . . G Y V I
119719557	1e-167	. . K K P H L N L V V I G H I D H G K S T L M G R L L Y E I . . G A V I
18313751	1e-166	. . Q K P H I N L A V V G H V D N G K S T L V G R L L Y E T . . G Y V I
11498542	1e-160	. . E K E H I N V A F I G H V D H G K S T L I G R L L Y E T . . G E I

Matches: 19

Filter Options

E Score: e-20

Percentage to return: 100

Domain: All, Unknown, Archaea, Bacteria

Kingdom: All, Unknown

Phylum: All, Unknown, Actinobacteria, Aquificae, Bacteroidetes

View Options

Zoom:

Figure 3: BLAST Search Results Window

5 Now the filtered results have been added to the MultiSeq main window. You see that they are all categorized under BLAST Results. Right click on the BLAST Results header on the left side of the MultiSeq window, under where it says Sequence Name. When the menu pops up, select Mark Group. Now all of the BLAST results have been marked. Click Options

→ Grouping → Taxonomy. In the window that appears, select the Marked Sequences radio button and choose the classification level as domain. Then click OK. Make sure that all the BLAST results have been recategorized under the Archaea group. Delete the original sequence EF1A_AERPE under Sequences on the left side of the MultiSeq window by clicking on the sequence title to highlight it and then hitting the delete key. Then right click on the group title Sequences and select Delete Group.

6 To further organize the results, click Options → Grouping → Taxonomy, and this time select phylum as the classification level. Click OK. You will see that the 19 sequences have been recategorized under their respective phyla. You should have 7 sequences under the Crenarchaeota phylum, 10 sequences under the Euryarchaeota phylum, and one sequence each under the phyla Korarchaeota & Nanoarchaeota.

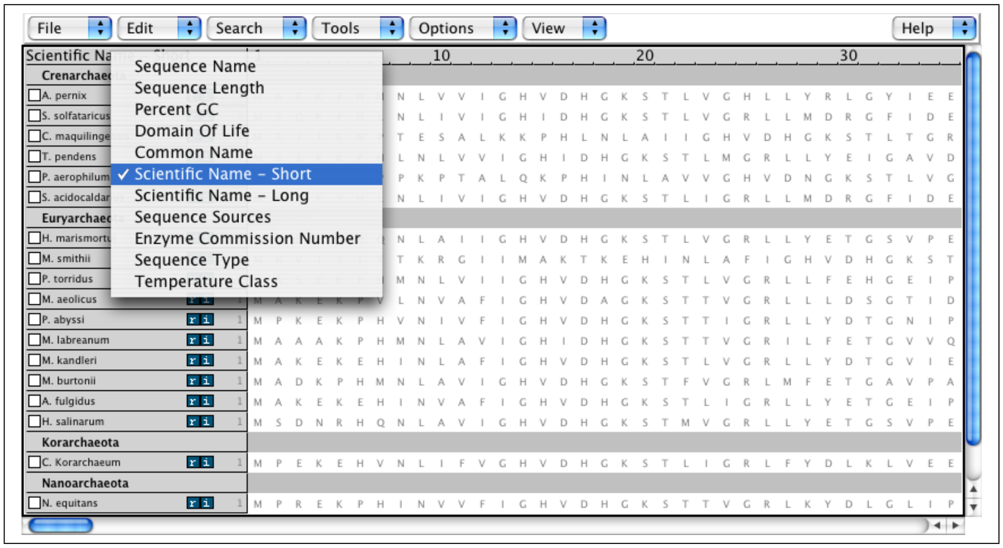


Figure 4: Archaea sequences grouped by phylum

7 Finally, it is usually easier to read the sequences with their scientific names. Click Sequence Name and choose Scientific Name – Short so that the description before each sequence shows the name of organism from which it is derived instead of the gi number. You may also change the description of the sequences to whichever you feel familiar with from the list.

2.2 Aligning archaeal sequences and removing redundancy

1 Now we will perform a multiple sequence alignment of the downloaded archaeal sequences. Click on Tools → Sequence Alignment. Choose the traditional ClustalW alignment program by click on the radio button before it. Make sure the Multiple Alignment radio button is selected and select Align All Sequences. Click OK.

2 Next we will color the alignment to more easily view the similarities and differences between the sequences. Click on View → Coloring → Apply to All. Then click View → Coloring → Sequence Similarity → BLOSUM80. The sequences will now be colored by sequence similarity using the BLOSUM 80 substitution matrix. Take a moment to examine the resulting alignment. The dark blue columns represent residues that are highly similar across organisms. The red columns represent regions that are divergent across all sequences. Do the sequences for the archaeal EF1A appear to be very similar? What do you think this means?

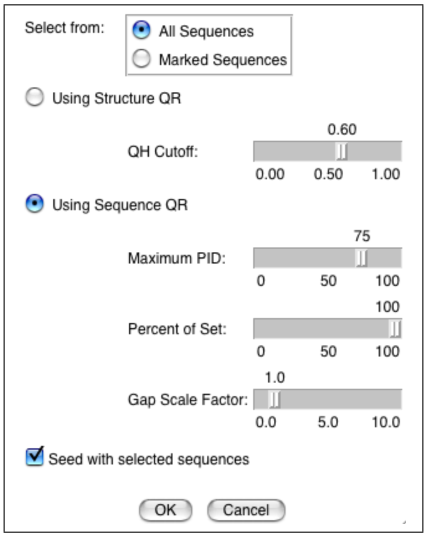
BLOSUM. BLOSUM is an acronym for BLOcks of amino acid SUbstitution Matrix [4]. It is well understood that changes in DNA occur and accumulate over evolutionary time. As a result, the same genes compared across multiple species will not be identical, but will display insertions, deletions, or substitutions of nucleotides. Since DNA codes for protein, similar changes occur in protein sequences. Over long periods of time, substantial changes can accumulate, such that two modern proteins with the same ancestral protein may be quite different. Substitution matrices are a method developed to measure the probability that two proteins are actually related, and are not similar by random chance alone. To create the BLOSUMs, alignments of highly conserved ungapped regions of proteins were analyzed from the BLOCKS database, and the probabilities of particular substitutions were determined. Sets of proteins where more than 40% of proteins were identical became BLOSUM40 matrix, while sets where more than 80% were conserved became BLOSUM80. Therefore more highly conserved sequences should use BLOSUM with higher numbers. As the EF-Tu is highly conserved across entire domains of life, we use the BLOSUM80 substitution matrix in our analysis.

3 Let's check the sequence identity of two of the sequences within the Euryarchaeota phylum. Click on the title of the first sequence in the phylum, hold down the SHIFT key, and click on the second sequence to highlight it. The percent identity across the two sequences will be listed in the bottom left corner of the MultiSeq window. Now check the sequence identity of the sequences within the Euryarchaeota phylum by clicking on the title of the first sequence to highlight it, holding down the SHIFT key and then clicking on the title of the last sequence in the phylum (you may have to scroll down). This will highlight all of the sequences in the phylum. Now minimum, average, and maximum percent identity is listed for all the sequences. What is the average, minimum, and maximum percent identity across all the sequences in the Euryarchaeota phylum? In the Crenarchaeota phylum?

4 Finally, we will create a QR representative set of the archaea sequences in order to reduce species bias and reduce the number of sequences in the alignment. We want to make sure a few of the sequences stay in our alignment and are not removed by the QR factorization algorithm. Holding down the CTRL key (COMMAND key on Mac), click on each of these three sequences to highlight them: *Aeropyrum pernix, Haloarcula marismortui, and Sulfolobus solfataricus.*

QR sets. The purpose of the QR set is to reduce the size of a large set down to its most representative elements, while preserving the phylogenetic tree topology of the homologous group. Structural and sequence based QR was developed in the lab of Zan Luthey-Schulten [9, 11]. The method is based on a multidimensional QR factorization of numerically encoded multiple sequence alignments which removes redundancy from the alignments and orders the sequences by increasing linear dependence.

5 To perform the QR factorization, click Search → Select Non-Redundant Set. In the resulting window, make sure All Sequences is selected. Choose the radio button Using Sequence QR, and adjust the Maximum PID slider to be 75. Note that the Percent of Set slider automatically goes to 100 when you do this; these two settings are mutually exclusive. This will only keep a sequence if it shares less than 75 percent identity with all the other sequences in the set. Check the Seed with selected sequences check box at the bottom of the window. This command will force MultiSeq to keep the marked sequences as part of the QR set. Click OK.



Select from: ● All Sequences ○ Marked Sequences

○ Using Structure QR

QH Cutoff: 0.60

● Using Sequence QR

Maximum PID: 75

Percent of Set: 100

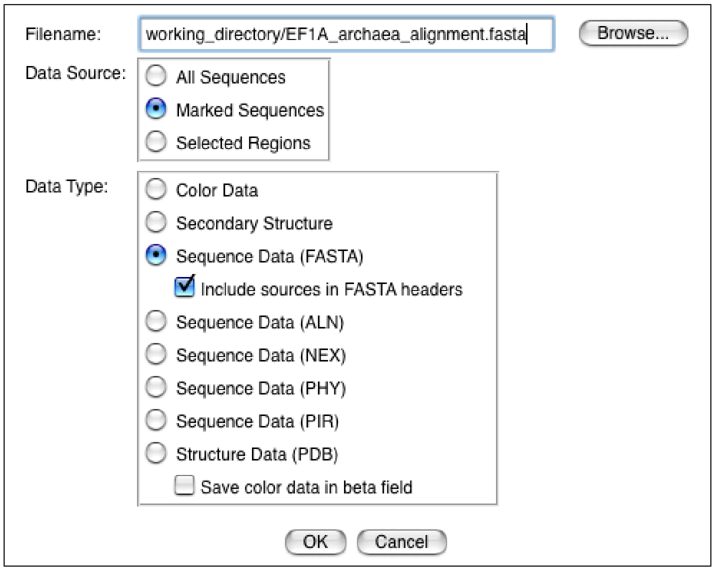
Gap Scale Factor: 1.0

☑ Seed with selected sequences

Figure 5: Select Non-Redundant Set using QR facterization

6 As a result, you should see 17 out of 19 sequences are highlighted. We will create a new group out of these sequences. Choose Options → Grouping → From Selection. Type in a new group title of QR75 and click OK to create the new group.

7 We will now export the archaea alignment information to a file in order to use it later. Right click on the group title QR75 and choose Unmark All. Again right click on the group title QR75 and select Mark Group. Click on File → Export Data (Figure 6). Click on Marked Sequences to select it. Make sure Sequence Data (FASTA) is marked, and Include sources in FASTA headers is checked. Click the Browse button next to the Filename text box, and navigate to your working_directory. Type in the filename `EF1A_archaea_alignment.fasta` and click Save. Click OK again to save the file. If you have been unable to complete this section, you will find the file `EF1A_archaea_alignment.fasta` in the 1.1.blast_sequences directory.



Filename: working_directory/EF1A_archaea_alignment.fasta [Browse...]

Data Source:
○ All Sequences
● Marked Sequences
○ Selected Regions

Data Type:
○ Color Data
○ Secondary Structure
● Sequence Data (FASTA)
☑ Include sources in FASTA headers
○ Sequence Data (ALN)
○ Sequence Data (NEX)
○ Sequence Data (PHY)
○ Sequence Data (PIR)
○ Structure Data (PDB)
□ Save color data in beta field

[OK] [Cancel]

Figure 6: Exporting Selected Regions

2.3 Finding bacteria EF-Tu sequences

1 Click File → New Session, and click Yes when you are asked to confirm.

2 Click on File → Import Data. Make sure the From Files radio button is selected, then click the Browse button.

3 Navigate to the 1.1.blast_sequences directory and select the file `bacteria_EFTu.fasta`. This file contains a sequence of bacteria EF-Tu from *E. coli* which we will use as reference to perform the BLAST search. Make sure Automatically download corresponding structures for sequence data is unchecked, and then click OK. The sequences will appear in the MultiSeq main window.

4 Now we will perform a BLAST search using the database from all the sequenced archaea and bacteria genomes. Select File → Import Data. Select the From BLAST Search radio button and make sure All Sequences is marked. Move the E Score slider until it reads e-20. Move the Max Results slider until it reads 200. Make sure Iterations is set to 1. Uncheck the Automatically download corresponding structures for sequence data option and then click OK. This BLAST search may take several minutes. When it completes, the BLAST Search Results will appear.

> **Reference to EF-Tu Paper.** While you are waiting on the BLAST search, read the Results and Discussion section of the paper **Dynamics of recognition between tRNA and Elongation Factor Tu** by Eargle *et al*; specifically the **Evolutionary analyses of EF-Tu and tRNA** section (pages 1386-1387). This is located on the CD in TUTORIAL_FILES. Pay particular attention to the discussion of conserved residues on the interface of EF-Tu and tRNA, as we will make use of this information when we examine the structure of the interface. Also read the section Local nonbonded interaction energies starting on page 1391 and note Figure 5 in the paper, which shows the interaction energy per residue masked by sequence identity. [2]

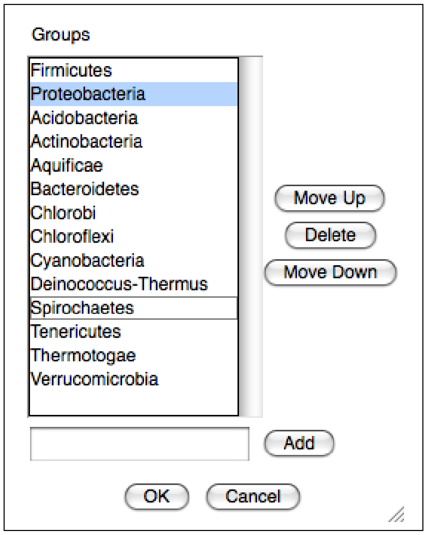
5 The current BLAST search should have returned 200 results. Under Filter Options make sure E Score is set to e-20 and Percentage to return is set to 100. Under Domain unselect All and select Bacteria. Click the Filter button. Now 196 results are left. Click the Accept button.

2.4 Performing ClustalW Multiple Sequence and Profile-Profile Alignments

There are 17 major bacterial phyla, including *Firmicutes*, *Proteobacteria*, *Cyanobacteria*, *Actinobacteria*, and *Aquificae*, among others. EF-Tu is a highly conserved house-keeping gene across *Bacteria* and *Archaea*. However, divergences do occur during evolution, and a complete multiple sequence alignment of all sequences in a domain of life would disregard the similarities within particular subgroups. For example, bacterial sequences within a phylum should be more similar to each other than to those in another phylum. As a result, we will first perform several multiple sequence alignments of EF-Tu sequences within selected bacterial phyla to preserve the similarity, and then use profile-profile alignments to align these multiple sequence alignments to each other. Going through this process yields a more accurate alignment than simply performing a multiple sequence alignment on all sequences returned by the BLAST search, especially when the number of sequences is large.

1 Now that the results have been imported into the MultiSeq window, we will delete the original query sequence as well as the multiseq group associated to it. Now we will sort the BLAST sequences by taxonomy. Click on Options → Grouping → Taxonomy. Classify the sequences by Phylum as you did for the archaea EF1A sequences, making sure to choose the All Sequences radio button.

2 The groups in MultiSeq are by default ordered alphabetically. However, there is an easy way to rearrange the ordering of groups so that it is convenient to find the group of interest in a long list. Click on Options → Grouping → Custom. In the resulting group list (Figure 7), click on the Firmicutes title to highlight it. Now click the Move Up button repeatedly to move the Firmicutes group to the top of the list. Now move the Proteobacteria to the second position in the list. Click OK.



Groups

Firmicutes
Proteobacteria
Acidobacteria
Actinobacteria
Aquificae
Bacteroidetes
Chlorobi
Chloroflexi
Cyanobacteria
Deinococcus-Thermus
Spirochaetes
Tenericutes
Thermotogae
Verrucomicrobia

Figure 7: Rearranging the order of groups

3 We are going to perform several ClustalW alignments in order to create a global multiple sequence alignment between all the bacterial sequences. Sequences within phyla should be more similar, and we would like our alignment to reflect that. Therefore we will first perform several ClustalW alignments on individual phylum. Scroll up the MultiSeq window to the top where you will find the Firmicutes group. Right click on the group title and choose Unmark All. Again right click on the group title and choose Mark Group. Now choose Tools → Sequence Alignment. In the resulting window, make sure ClustalW alignment program, Multiple Alignment are selected, and choose the Align Marked Sequences radio button. Click OK to perform a multiple sequence alignment on the Firmicutes group.

4 Scroll down the window to find the second group Proteobacteria, and repeat the above step.

5 We are not going to perform step 3&4 for all the remaining sequences phylum by phylum. Instead, we will group them into one group called "Additional Bacteria". Scroll down the MultiSeq window until you see the Acidobacteria group. Click on the title of the first sequence in this group to highlight it. Now scroll to the bottom of the MultiSeq window. While holding down the SHIFT key, click on the title of the final sequence in the list to highlight all the sequences in between. Now select Options → Grouping → From Selection. Type in the group name Additional Bacteria and click OK to place all the remaining sequences in a single group. Delete the empty groups that remain by right click on the group name and then choose Delete Group, or you may also use the Options → Grouping → Custom to do the deletion.

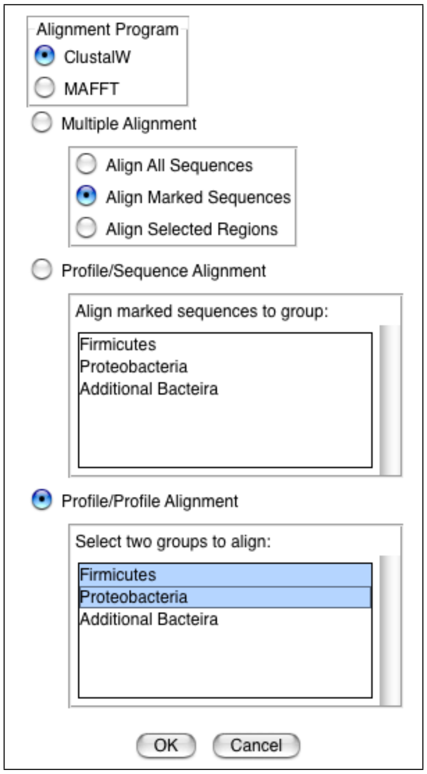
6 Now we will perform a final multiple sequence alignment on the additional group of sequences. Right click on the group title Additional Bacterial and choose Unmark All. Again right click on the group title and choose Mark Group. As you did before, perform a ClustalW multiple sequence alignment on the marked sequences.

7 Next, we will perform a profile-profile alignment between the aligned groups. Choose Tools → Sequence Alignment. Now, choose the Profile/Profile Alignment radio button and click on Firmicutes and Proteobacteria to highlight them (Figure 8). Click OK to perform a profile alignment on these two groups.

8 Now select all the sequences in the Firmicutes and Proteobacteria using the SHIFT key as you did before, and combine them into a single group by choosing Options → Grouping → From Selection and giving them the title Profile1. Now perform a profile-profile alignment between the Profile1 group and the Additional Bacteria group by choosing Tools → Sequence Alignment as you did before.

9 Finally, we will perform a QR factorization on these aligned bacterial sequences to reduce the size of the group to a smaller size. While MultiSeq is more than capable of handling many hundreds or even thousands of sequences, for the purposes of this tutorial we do not need more than a few sequences. As before, we want to seed the QR set with a few bacteria that we wish to keep in the alignment. Holding down the CTRL key (COMMAND on Mac), select the organism *Buchnera aphidocola* under profile 1 and the first sequence in the species *Thermus thermophlius.* To perform the QR factorization, click Search → Select Non-Redundant Set. In the resulting window (Figure 5), make sure All Sequences is selected. Choose the radio button Using Sequence QR, and adjust the Maximum PID slider to be 75. Note that the Percent of Set slider automatically goes to

100 when you do this; these two settings are mutually exclusive. This will only keep a sequence if it shares less than 75 percent identity with all the other sequences in the set. Check the Seed with selected sequences check box at the bottom of the window. This command will force MultiSeq to keep the selected sequences as part of the QR set. Click OK. Now several sequences are highlighted. We will create a new group out of these sequences. Choose Options → Grouping → From Selection. Type in a new group title of QR75 and click OK to create the new group.



Alignment Program
● ClustalW
○ MAFFT

○ Multiple Alignment
○ Align All Sequences
● Align Marked Sequences
○ Align Selected Regions

○ Profile/Sequence Alignment
Align marked sequences to group:
Firmicutes
Proteobacteria
Additional Bacteira

● Profile/Profile Alignment
Select two groups to align:
Firmicutes
Proteobacteria
Additional Bacteira

Figure 8: Profile-Profile alignment

10 Save the multiple sequence alignment of the QR set by clicking File → Export Data (Figure 6). Make sure Selected Regions is selected, as well as Sequence Data (FASTA) and Include sources in FASTA headers. Click the Browse button and navigate to the working_directory. Type in the filename `EFTu_bacteria_alignment.fasta` and click Save. Click OK again to save the file. If you have been unable to complete this section, you will find the file `EFTu_bacteria_alignment.fasta` in the 1.1.blast_sequences directory.

2.5 Creating Multiple Sequence with MAFFT

In the previous steps of the tutorial, we have used ClustalW to align the sequences. While ClustalW is perfectly fine for tens of sequences, it tends to be too slow to align hundreds of sequences. For more than a hundred sequences, we recommend that one use MAFFT.

MAFFT. is a multiple sequence alignment program for amino acid or nucleotide sequences.It takes standard fasta format input files and the type of input sequences (amino acid or nucleotide) is automatically recognized. It offers a range of multiple alignment methods, L-INS-i (accurate; for alignment of $<\sim$200 sequences), FFT-NS-2 (fast; for alignment of $<\sim$10,000 sequences), etc. One of the main advantages of MAFFT is that its scoring system is designed to allow large gaps. Thus MAFFT is suitable for LSU rRNA and SSU rRNA alignments that sometimes have variable loop regions. For more information, please refer to http://mafft.cbrc.jp/alignment/software/ [5, 6].

1 To enable the MAFFT function, please download the MAFFT program from the website in the above information box, and install it on your machine. Then configure MAFFT the same way as you did for BLAST. To make sure that MAFFT is installed correctly, pleases open a terminal and type “mafft”. If the terminal prompt returns “Input file? (fasta format)” then MAFFT has been installed correctly. Afterwards, please check the MAFFT directory path variable under the Preference menu correctly points to the MAFFT executable.

2 In order to compare results from both ClustalW and MAFFT alignments, we will go through the steps learnt in section 2.3 again to get the 196 bacteria EF-Tu sequences.

3 Choose Tools → Sequence Alignment. Select the alignment program MAFFT instead of ClustalW, then do Mutiple Alignment for all sequences. Save the result in fasta format.

Instead of going through all the steps in section 2.4, MAFFT completes the alignment in one single step within 30 seconds. MAFFT obviously wins over when the speed of the two alignment programs is compared. Then you can open two VMD/Multiseq sessions to compare the alignments you obtained from MAFFT and ClustalW. Use any coloring, phylogenetic reconstruction and scripting techniques you have learnt in the tutorial. Judge the results based on biological knowledge of EF-Tu.

2.6 Conservation of EF-Tu among the Bacteria

EF-Tu participates in one of the oldest and most fundamental processes in the cell: the mechanism of translation, by which proteins are made from messenger

RNA. Therefore we might expect that portions of the molecule have certain properties which must have been preserved over evolutionary time. The most recent common ancestor of bacteria and archaea, from which both domains evolved, existed over three billion years ago. In this section, we will use the coloring features of MultiSeq to examine the evolutionary conservation between the aligned sequences of bacteria and archaea, and examine how this conservation correlates with the structure and function of the molecule.

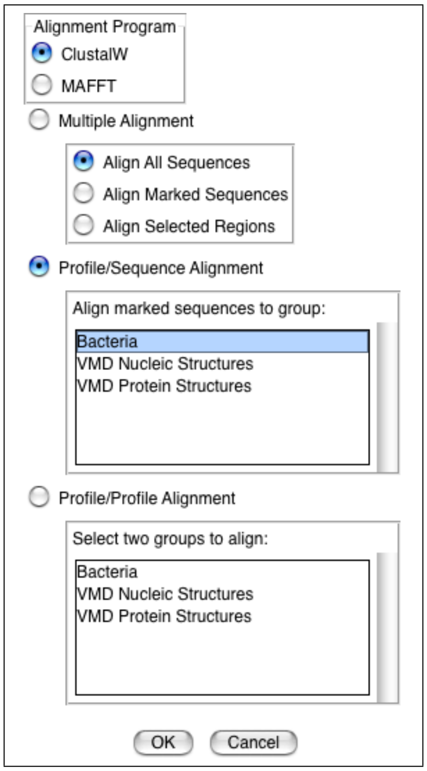
1 We must first load the bacterial sequences that we created in the previous section. Let's create a new session of MultiSeq by clicking on File → New Session and clicking Yes at the confirmation window. This will erase all the current data in MultiSeq. Now go to File → Import Data. Make sure From Files is selected, and click the Browse button. Navigate to the working_directory and open the `EFTu_bacteria_alignment.fasta` file that we created earlier. Click OK to import the aligned sequences. Now right click on the group title Sequences and click on Rename Group. Rename the group to be Bacteria and click OK.

2 Go to File → Import Data and click on the Browse button. Choose PDB Files in the Enable drop down menu in the File Chooser windows. The structure is located on the CD in the directory 1.4.eftu_structures. Load the structure `1B23_bact_tRNA.pdb` into MultiSeq by highlighting it and clicking Open. Then click OK. Go back to the MultiSeq window and note that two sequences have been added at the bottom of the sequence list, under groups VMD Protein Structures and VMD Nucleic Structures.

3 In order to color the structure by sequence identity, we must align the protein sequence to the existing bacterial alignment. This will map the alignment conservation information onto the structure, and allow us to color the structure by sequence conservation. Right click on the group title VMD Protein Structures and select Unmark All. Now mark the structure `1B23_bact_tRNA.pdb` in the group VMD Protein Structures. Click Tools → Sequence Alignment. Select Profile/Sequence Alignment and in the box labeled Align marked sequences to group select the Bacteria group. Click OK (Figure 9).

4 Right click on the Bacteria group and choose Mark Group. This will select all the sequences in the Bacteria group but not the single sequence in the VMD Nucleic Structures group, which we want to exclude.

5 Now we will color the alignment by sequence identity, and examine the conservation in the corresponding structures. Click on View → Coloring → Apply to Marked, and then View → Coloring → Sequence Identity. The sequences are now colored by identity in MultiSeq, and the EF-Tu structure is colored in the same manner in the VMD Display. Dark blue highlights residues which are conserved, and dark red highlights residues which are highly divergent. Light blue and light red indicate regions of intermediate conservation (Figure 10).



Alignment Program

● ClustalW

○ MAFFT

○ Multiple Alignment

● Align All Sequences

○ Align Marked Sequences

○ Align Selected Regions

● Profile/Sequence Alignment

Align marked sequences to group:

Bacteria

VMD Nucleic Structures

VMD Protein Structures

○ Profile/Profile Alignment

Select two groups to align:

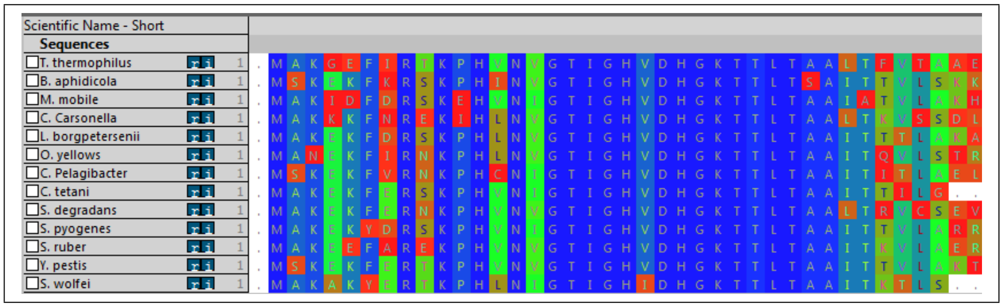
Bacteria

VMD Nucleic Structures

VMD Protein Structures

Figure 9: Aligning *T. aquaticus* sequence to Bacteria profile

6 Click on the `1B23_bact_tRNA.pdb` label under the VMD Protein Structures group and drag it down to the top of the Bacteria group. This will add that sequence to this group, and allow the conservation information to be mapped onto the structure.



```
Scientific Name - Short
Sequences
□T. thermophilus    1 . M A K G E F I R T K P H V N V G T I G H V D H G K T T L T A A L T F V T A A E
□B. aphidicola      1 . M S K E K F K R S K P H I N V G T I G H V D H G K T T L T S A I T T V L S K K
□M. mobile          1 . M A K I D F D R S K E H V N I G T I G H V D H G K T T L T A A I A T V L A K H
□C. Carsonella      1 . M A K K K F N R E K I H L N V G T I G H V D H G K T T L T A A L T K V S S D L
□L. borgpetersenii  1 . M A K E K F D R S K P H L N V G T I G H V D H G K T T L T A A I T T T L A K A
□O. yellows         1 . M A N E K F I R N K P H L N V G T I G H V D H G K T T L T A A I T Q V L S T R
□C. Pelagibacter    1 . M S K E K F V R N K P H C N I G T I G H V D H G K T T L T A A I T I T L A E L
□C. tetani          1 . M A K E K F E R S K P H V N I G T I G H V D H G K T T L T A A I T T I L G . .
□S. degradans       1 . M A K E K F E R N K P H V N V G T I G H V D H G K T T L T A A L T R V C S E V
□S. pyogenes        1 . M A K E K Y D R S K P H V N I G T I G H V D H G K T T L T A A I T T V L A R R
□S. ruber           1 . M A K E E F A R E K P H V N V G T I G H V D H G K T T L T A A I T K V L A E R
□Y. pestis          1 . M S K E K F E R T K P H V N V G T I G H V D H G K T T L T A A I T T V L A K T
□S. wolfei          1 . M A K A K Y E R T K P H L N I G T I G H I D H G K T T L T A A I T K T L S . .
```

Figure 10: Sequence Identity across a bacterial profile

7 Examine the structure file in the VMD display. Note that the EF-Tu has been colored by sequence identity. Let's change the drawing method of the EF-Tu chain P to Surf. Now examine the interface between the tRNA and the EF-Tu. Does the interface look more conserved than the rest of the protein? Why do you think it is important that this interface is conserved?

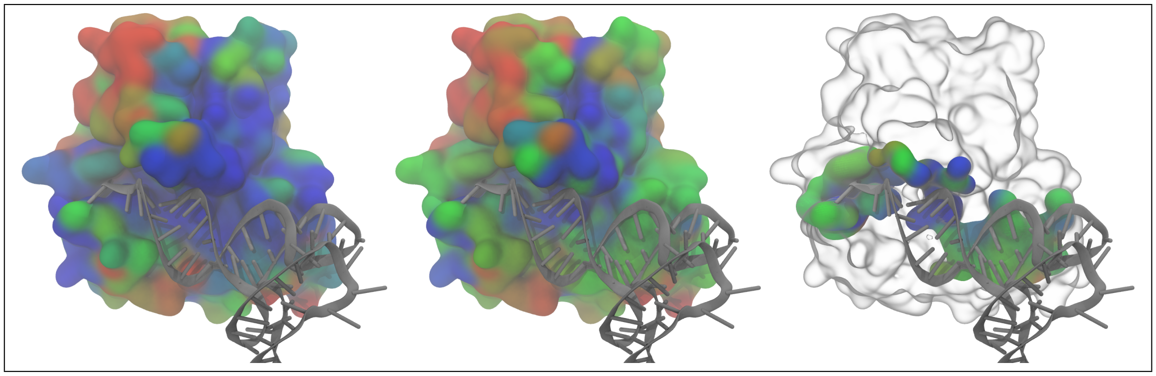


Figure 11: Sequence Identity across A) Bacterial profile and B) Bacterial and Archaea profile mapped onto EF-Tu structure. Sequence Identity of C) Bacterial and Archaea profile mapped onto residues at the interface between EF-Tu and the tRNA.

2.7 Finding conserved residues across the bacterial and archaeal domains

1 We must first load the archaeal sequences that we created earlier. Import the EF1A_archaea_alignment.fasta file from the working_directory. Now right click on the group title Sequences and click on Rename Group. Rename the group to be Archaea and click OK.

2 Now we are ready to perform the profile-profile alignment between the bacterial and archaeal sequences. As you did in the previous section, use ClustalW to perform a profile-profile alignment between the Archaea and Bacteria groups.

3 Let's mark all the sequences with the exception of the tRNA sequence that is still present in MultiSeq window. Do this by right clicking on a group title and selecting Mark All. Now unmark the single sequence in the VMD Nucleic Structures group.

4 Save the complete multiple sequence alignment by clicking File → Export Data. Make sure Marked Sequences is selected, as well as Sequence Data (FASTA) and Include sources in FASTA headers. Click the Browse button and navigate to the working_directory. Type in the filename EFTu_all_alignment.fasta and click OK. Click OK again to save the file. If you have been unable to complete this section, you will find the file EFTu_all_alignment.fasta in the 1.1.blast_sequences directory.

5 Now we will color the global alignment by sequence identity, and examine the conservation in the structure. Click on View → Coloring → Apply to Marked, and then View → Coloring → Sequence Identity. The sequences are now colored by identity in MultiSeq. Dark blue highlights sequences which are conserved across both bacteria and archaea, while dark red highlights sequences which have little or no conservation either within or across both domains of life. Light blue and light red indicate regions of intermediate conservation.

6 To more easily visualize the colored alignment, we will use the Zoom Window feature of MultiSeq. Click on View → Zoom Window to open the Zoom Window (Figure 12). Enlarge the small window that appears by stretching it to the right. You can now easily see the entire alignment across both groups. Note that within the bacteria most of the protein is highly conserved, and the same is true for within the archaeal group. There are a few sections where both archaea and bacteria share the same sequence identity. Note the black box in the Zoom Window. This shows the portion of the Zoom Window currently displayed in the main MultiSeq window. If you click anywhere in the Zoom Window you will move the black box and can visualize the close-up of the alignment at that point. Click on the places where both bacteria and archaea share the same residues. What do you think is significant about these regions of the protein? When you are done, close the Zoom window.

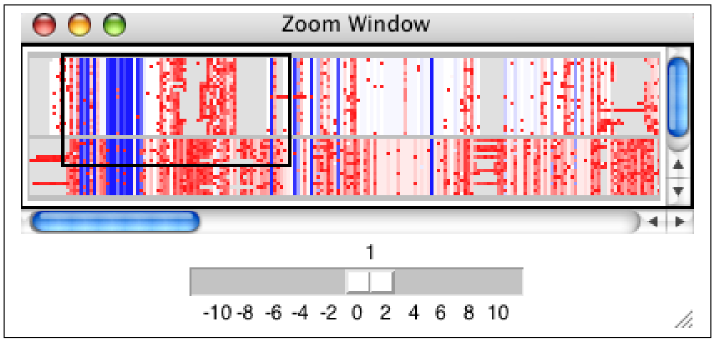


Figure 12: Zoom window

7 Take a look at the VMD Display. The coloration on the EF-Tu structure has changed to be sequence identity across bacteria and archaea, rather than simply within the bacteria (Figure 11B).

8 Now click on View → Coloring → Apply by Group, and color by sequence identity as you did before. Switch back and forth between coloring by group and coloring by marked. Is the interface between the EF-Tu and the tRNA more or less conserved when the archaeal sequences are taken into account?

9 Finally, examine Figure 13, extracted from the EF-Tu paper [2]. The graph shows interaction energies between all EF-Tu residues and the tRNA, as averaged over 16 nanoseconds of a molecular dynamics simulation with NAMD and filtered for sequence identity. In VMD, change the Drawing Method for chain P to NewCartoon, and create a new representation. Display the residues labeled in the graphic below using the VDW drawing method, and Selected Atoms as `chain P and resid 56 87 90 227 300 376 389 390`. Set the Coloring Method to be Trajectory → User → User. Where are these highly conserved residues located in your structure? What property do most of these residues share?

2.8 EF-Tu Interface with the Ribosome

Proteins are dynamic molecules and may exist in multiple conformations. Within the carefully controlled environment of the cell, there are usually only a few stable conformations for a particular protein, each one associated with the presence or absence of a covalently-bound or non-bonded ligand or protein. Two of the stable conformations of the elongation factor include the GTP-bound

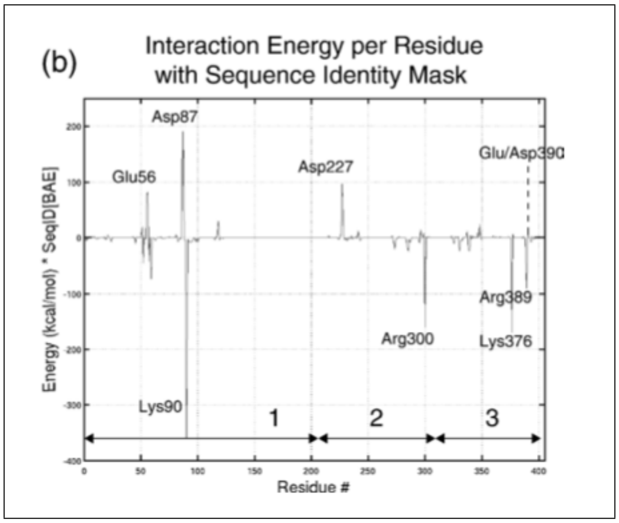


Figure 13: Interaction energies for residues on the EF-Tu:tRNA interface

and GDP-bound conformations. GTP (guanosine triphosphate) is a nucleotide which is bound to the elongation factor in a ternary complex with the ‘charged’ tRNA. Upon correct codon recognition of the tRNA with the mRNA in the ribosome, the elongation factor hydrolyzes the phosphoanhydride linkage between the gamma and beta phosphate groups of the GTP, releasing inorganic phosphate and transforming the GTP into GDP, or guanosine diphosphate. This induces a conformational change in the elongation factor, releasing the tRNA to the ribosome. In this section of the tutorial, we will examine the interface between the EF-Tu and the ribosome using the same sequence identity methods as the previous section.

1 Before loading the structures of EF-Tu bound to the ribosome, we must delete the current structures out of VMD to make sure we do not get confused with multiple EF-Tu structures. In the VMD main window, highlight `1B23_bact_tRNA.pdb` and delete it.

2 We will now load the structures of EF-Tu:tRNA bound to a ribosome. Open the structure files `3FIH.pdb` and `3FIK.pdb` from the `1.5.eftu_cryoEM_ribosome` directory. These represent the small and large subunits of the *E. coli* ribosome, respectively, to which the EF-Tu and tRNA have been fitted. One might also compare these files to the PDB files 2Y18 and 2Y19—which are crystal structures of the EF-Tu:tRNA:ribosome complex. It may take a few minutes for the structures to open, as they are large structures and it takes time for them to be loaded into MultiSeq.

Molecular Dynamics Flexible Fitting. MDFF was developed in the lab of Klaus Schulten at UIUC to allow static X-ray crystallography or NMR structures to be dynamically fitted to low resolution cryo-EM data, while maintaining the overall stereochemical quality of the structure using the MD force field. [13]

3 In MultiSeq, you will see that many sequences have been added. Most of these sequences correspond to ribosomal proteins. We are interested in the sequence `3FIH_Z`, which is the cryo-EM fitted EF-Tu structure. Make sure you do not accidentally select `3FIK_Z`, which is also in the list. Mark the sequence `3FIH_Z` in MultiSeq.

4 Now it's your turn! As you did before, align this marked sequence to the bacterial multiple sequence alignment. Then add this sequence to the Bacteria group by dragging it down with the mouse. Finally, mark both the Bacteria and Archaea groups and color them by Sequence Identity, making sure to choose View → Coloring → Apply to Marked first.

5 In VMD, open the Representations menu. Make sure that `3FIH.pdb` is showing in the drop-down box, and change the drawing method for Chain Z to Surf. This draws the EF-Tu in the VMD Display window as a solvent-accessible surface, which renders it easy to see next to the ribosome.

6 In the same Representations menu select the structure `3FIK.pdb` from the Selected Molecule drop-down box,. This is the large subunit of the ribosome. Click the Create Rep button, and in the Selected Atoms text box, type `nucleic and resid 2650 to 2670` and hit Enter. Under Coloring Method, choose ResID and in the Drawing Method menu, choose VDW. These rRNA nucleotides are now highlighted in the VMD display. Note how these nucleotides interact with the elongation factor. Is the region of the elongation factor highly conserved at this interaction site? The region of the ribosome you have just highlighted is called the *Sarcin-Ricin loop*. Sarcin and ricin are cytotoxins which each cleave a single bond in this highlighted loop, rendering the ribosome unable to properly interact with the elongation factor and shutting down translation.

3 Computing a Maximum Likelihood Phylogenetic Tree with RAxML

A phylogenetic tree is a way to illustrate the evolutionary relationships between a set of organisms. Leaves of the tree represent organisms, which are connected through internal nodes representing putative ancestors that no longer exist. Two leaves which are connected by a recent ancestor are evolutionarily more closely related than organisms which share a more distant common ancestor. Such a 'tree of life' would have as its root the last universal common ancestor (LUCA)

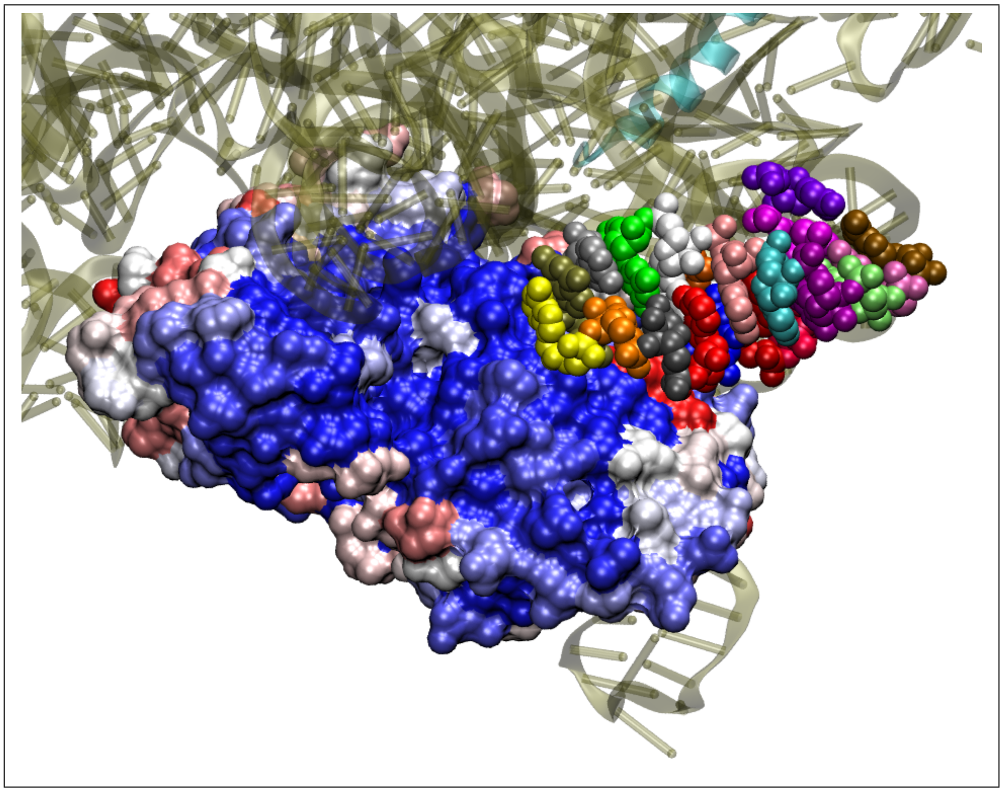


Figure 14: Portion of the ribosomal Sarcin-Ricin loop interacting with the EF-Tu

of all life on Earth, prior to the divergence of the bacterial and eucarya/archaea branches. MultiSeq has the capability to create several types of phylogenetic trees based on sequence and structural similarity (Tools → Phylogenetic Tree), as well as reading in and manipulating phylogenetic trees created externally.

Creating Maximum Likelihood phylogenetic trees is very computationally intensive and may take 15-20 minutes, depending on the speed of your computer. If you would like to go ahead and create the tree, you will find detailed instructions on creating such a tree in Appendix B. Simply complete Appendix B and then continue with the section below. If you wish to skip creation of the tree, we have provided a pre-created tree for you in the directory 2.1.phylogenetic_tree with the filename `RaxML_bipartitions.AB_eftu.boot.tre`. You will also need the file `EFTu_all_alignment.fasta` from the working_directory or 1.1.blast_sequences depending on if you created it or not. If you do not plan on completing Appendix B, copy these files to your working_directory.

3.1 Load the Phylogenetic Tree into MultiSeq

1 We must load the sequences that we created in the first two sections. Let's create a new session of the MultiSeq window. Now go to File → Import Data. Make sure From Files is selected, and click the Browse button. Navigate to the working_directory and open the `EFTu_all_alignment.fasta` file that we created earlier. Click OK to import the aligned sequences.

2 We will load the phylogenetic tree file into MultiSeq to visualize the final tree. Click on Tools → Phylogenetic Tree. Select the All Sequences radio button, and check the From File checkbox. Click the Browse button and navigate to the saved file `RaxML_bipartitions.AB_eftu.boot.tre` in your working_directory. Click OK.

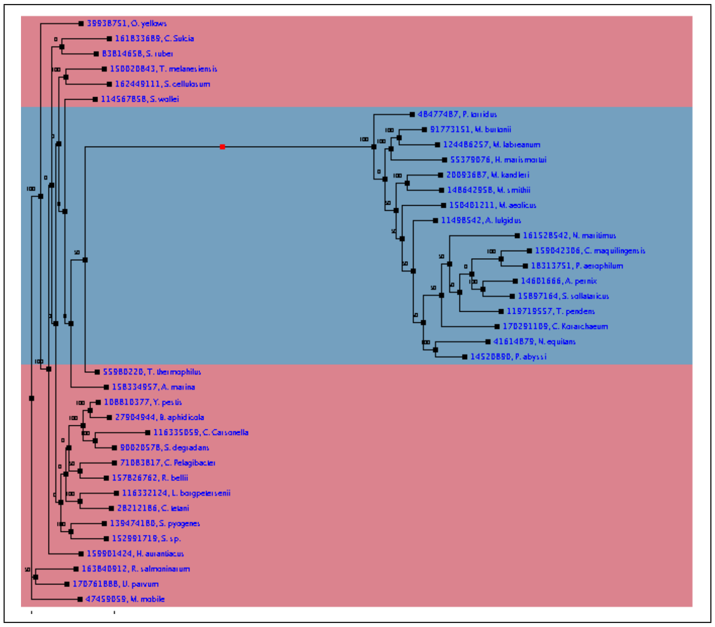


Figure 15: Rerooting the Phylogenetic Tree (Red Mark)

3.2 Reroot and Manipulate the Phylogenetic Tree

1 The phylogenetic tree generated by RaxML for the aligned sequences in MultiSeq will appear in a new window. Note that the initial labels for the tree leaves are the GI numbers for each sequence. Click on View → Background Color → Taxonomy → Domain of Life to color the background by domain of life. Note that the archaea have all been grouped together, but the bacteria are split into two groups. We can 'reroot' the tree to display all the appropriate organisms together. Click anywhere in the long line separating the bacteria from the archaea to create a red mark. Now select View → Reroot tree at selected point to reroot the tree.

2 Now label each abbreviated species name using the View → Leaf Text → Taxonomy menu. Color each leaf by phylum using the View → Leaf Color menu. Do the phyla seem to be grouped together within each domain of life? Remember that we created a very simple maximum likelihood phylogenetic tree using very few replicates, therefore a certain amount of error is expected. To more easily view the grouping by phylum, select View → Collapse By → Taxonomy → Phylum. This will collapse the tree by phylum. All sequences grouped by Phyla are now displayed as triangles. The length of the triangle corresponds to the evolutionary distance between the two most distant sequences in the collapsed set. The bottom point of the triangle represents the shortest branch in the set, while the upper point of the triangle represents the longest branch in the set. Click on View → Expand All to display all the leaves again. Finally, we can rearrange the tree in many different ways. Right click on the root of the tree (the black mark that connects the two domains of life) and select Rotate Up to rotate the bacteria domain above the archaea.

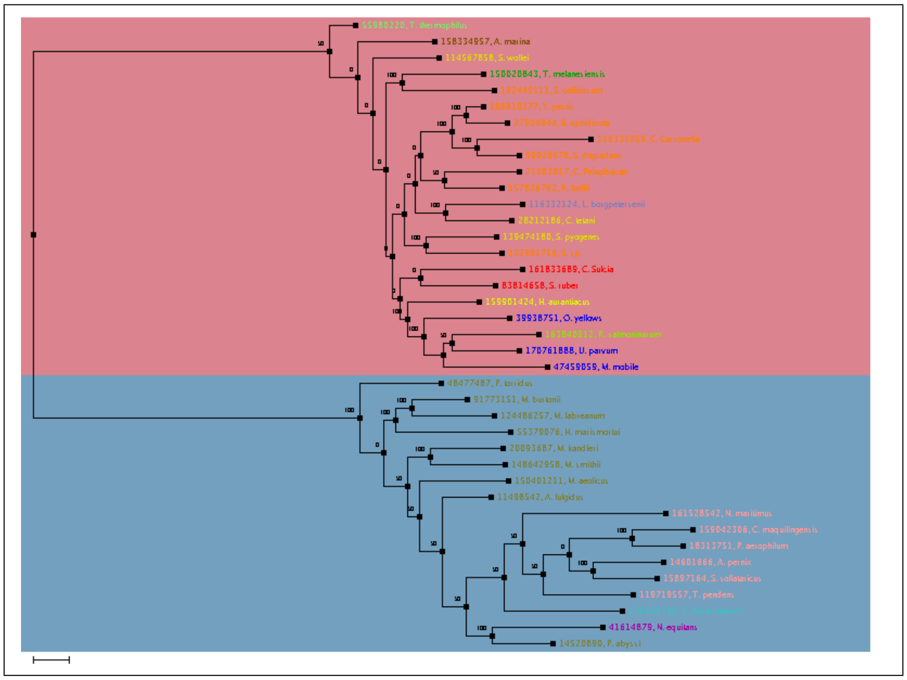


Figure 16: Rerooted Phylogenetic Tree with Leaves colored by Phylum

3 We will save the tree in a format needed for the final section of this tutorial. Click on File → Save in the Tree Viewer window. In the following window, select the Newick file format and make sure both Include internal labels and Include branch lengths are checked. Click the Browse button and navigate to the working_directory. Type in the filename `AB_eftu_rerooted.tre` and click Save. Click OK to save the tree. If you have been unable to complete this section, you will find this file in the directory 2.1.phylogenetic_tree on the CD.

4 MultiSeq TCL Scripting: Genomic Context

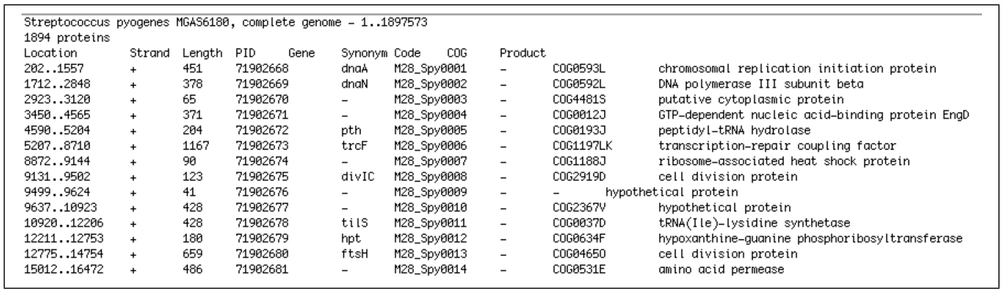
While the content of genes, in terms of their relative mutations, insertions, and deletions is commonly used to compare them, the location of those genes on the genome can also be informative. The genome is not a static structure; duplicates of genes can be introduced during replication, and genetic elements called transposons that move around in the genome can move genes to new locations. As a result, the genomic context of a gene - the position in the genome in which it is located, and the corresponding genes around it, may change. Organisms more closely related to one another and sharing a similar environment are more likely to have similar genomic contexts for a particular gene. In this final portion of the tutorial, we will make use of the genomes of bacteria and archaea to calculate the genomic context of EF-Tu.

1 We have provided a script for you to create the genomic context using the original bacterial and archaeal genomes in the 3.1.AB_genomes directory, the FASTA alignment of the bacterial and archaeal alignment set you created in the first section, and the rerooted phylogenetic tree you created in the previous step. The script file is located in the directory 3.genomic_context, and is called `genomeContext.tcl`. Please copy this file to working_directory.

PTT File Format. The genomic context script makes use of the .ptt file generated for each genome by the BLAST program formatdb. An example of the .ptt file format shown in Figure 17. As you can see, the .ptt file lists all annotated genes in the genome in order of their incidence from either the 5' end of a linear chromosome or the origin of replication if the chromosome is circular. Using the sequential gene information in the .ptt file, the genomic context script can create a graphical representation of the context around a particular gene.

2 In VMD, click on Extensions → TK Console. Navigate to the working_directory using the cd command. To load the script into TCL, use the command `source genomeContext.tcl`.

3 Type the following command into TCL to execute the genomic context script (type this all on one line):



```
Streptococcus pyogenes MGAS6180, complete genome - 1..1897573
1894 proteins
Location        Strand  Length  PID      Gene     Synonym Code    COG       Product
202..1557       +       451     71902668          dnaA    M28_Spy0001       -         COG0593L        chromosomal replication initiation protein
1712..2848      +       378     71902669          dnaN    M28_Spy0002       -         COG0592L        DNA polymerase III subunit beta
2923..3120      +       65      71902670          -       M28_Spy0003       -         COG4481S        putative cytoplasmic protein
3450..4565      +       371     71902671          -       M28_Spy0004       -         COG0012J        GTP-dependent nucleic acid-binding protein EngD
4590..5204      +       204     71902672          pth     M28_Spy0005       -         COG0193J        peptidyl-tRNA hydrolase
5207..8710      +       1167    71902673          trcF    M28_Spy0006       -         COG1197LK       transcription-repair coupling factor
8872..9144      +       90      71902674          -       M28_Spy0007       -         COG1188J        ribosome-associated heat shock protein
9131..9502      +       123     71902675          divIC   M28_Spy0008       -         COG2919D        cell division protein
9499..9624      +       41      71902676          -       M28_Spy0009       -         -       hypothetical protein
9637..10923     +       428     71902677          -       M28_Spy0010       -         COG2367V        hypothetical protein
10920..12206    +       428     71902678          tilS    M28_Spy0011       -         COG0037D        tRNA(Ile)-lysidine synthetase
12211..12753    +       180     71902679          hpt     M28_Spy0012       -         COG0634F        hypoxanthine-guanine phosphoribosyltransferase
12775..14754    +       659     71902680          ftsH    M28_Spy0013       -         COG0465O        cell division protein
15012..16472    +       486     71902681          -       M28_Spy0014       -         COG0531E        amino acid permease
```

Figure 17: PTT File Format

`draw_genome_context_of_phylogeny genomeContext.ps EFTu_all_alignment.fasta AB_eftu_rerooted.tre 8000 0.1 ../3.1.AB_genomes 1`

4 The genomic context will be created in the file `genomeContext.ps` in the working_directory. If you have been unable to complete this step, this file can be found on the CD in the directory 3.genomic_context. You will need a PostScript viewer such as Adobe Acrobat Reader to view the file. Open this file.

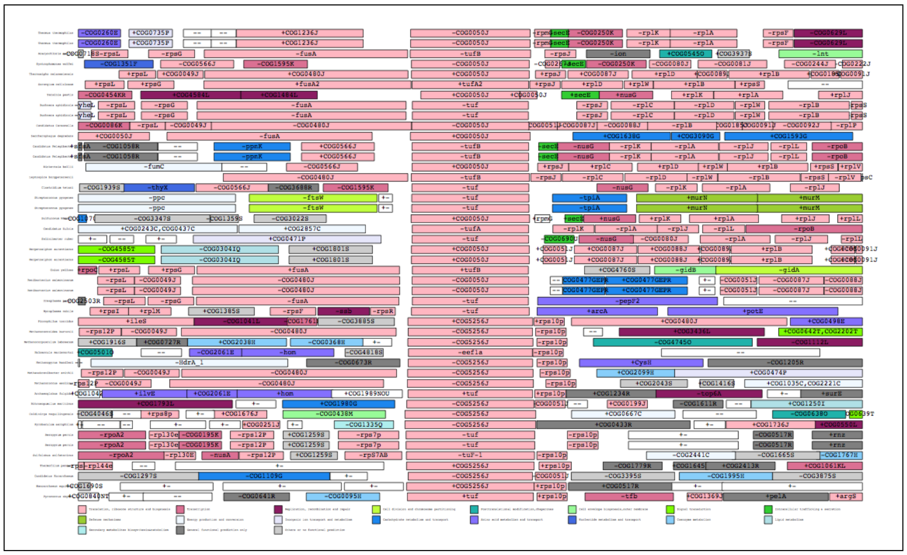


Figure 18: Genomic Context

5 You should still have open the phylogenetic tree in MultiSeq. If not, please reload the phylogenetic tree file `AB_eftu_rerooted.tre` as you did in the previous section. The ordering of organisms in the phylogenetic tree will match the ordering in the genomic context. If you followed the directions in the last step of the previous section, the bacteria should all be listed first, and the archaea second. The first bacteria listed, then, is *T. thermophilus*. The first archaea listed is *P. torridus*

.

6 The center pink boxes which are all aligned represent the EF-Tu gene (for bacteria) or the EF-1A gene (for archaea). The boxes on either side of the center for each organisms represent the genes on either side of the EF-Tu gene in the genome for that organism. You can see by the color of each box and the legend at the bottom of the file. Pink boxes correspond to translation genes, while blue boxes correspond to carbohydrate metabolism, and green boxes represent cellular trafficking.

7 Note first that nearly all of the archaea EF-1A genes are followed by a small pink box with the label `rps10p` or `COG0051J`. This corresponds to the ribosomal protein S10 in both cases, which is a protein attached to the small subunit of the ribosome. In contrast, few of the bacteria have this gene following EF-Tu (rpsJ). Also note that most of the bacteria have translation genes immediately following the EF-Tu gene (pink boxes), which are not present in almost all of the archaea. Clearly, there is no absolute boundary between the two domains of life in terms of the order of genes on the genome around the EF-Tu gene, but there are significant differences. A greater analysis over larger numbers of species can identify more consistent differences between the archaea and bacteria, as well as identify those organisms with rare gene or operon duplication events.

8 If you are interested in translating directly from the COG numbers, which are listed in the genomic context you just created, to actual protein names, the translation file whog is located in the directory 3.genomic_context. This file was downloaded from the FTP site:

ftp://ftp.ncbi.nih.gov/pub/COG/COG/

and contains all the COG (Clusters of Orthologous Groups) numbers along with the title of the corresponding protein. If you open this file in a text editor and search for COG5256 you will find that it refers to Translation elongation factor EF-1alpha, and COG0050 refers to GTPases - translation elongation factors. You can use this file to translate from the COG numbers into the actual protein names.

9 Compare the genomic context you have just calculated with the consensus genomic context below (Figure 19) calculated for a the universal ribosomal proteins in a ribosomal evolution paper [10]. The solid black squares indicate the gene in the genome is found in this position in more than 50%

of all sequenced genomes. Light grey squares indicate this gene is present in only 15% of genomes. As a result, you can see that the ordering of most archaeal ribosomal operons differs from the corresponding bacterial ribosomal operons. While the ordering of genes in the ribosomal operon is highly conserved across the bacterial domain of life, the ordering across both bacterial and archaeal domains of life is more divergent.

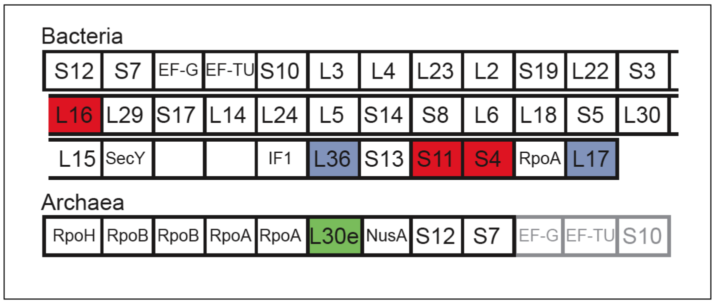


Figure 19: Consensus Genomic Context [10]

5 Appendix A

5.1 Building a BLAST Database

It is possible to build a BLAST database from any file in FASTA format, using the tool formatdb which is included with the standalone BLAST distribution available from the NCBI. As you have already configured BLAST for use with MultiSeq, you will have already installed the formatdb program. For future reference, you can download any genome in the NCBI database by visiting:

ftp://ftp.ncbi.nih.gov/genomes/

Not all sequence information is contained in genome files; many genes are sequenced individually and submitted to the NCBI. You can find a listing of the sequence databases here:

ftp://ftp.ncbi.nih.gov/blast/db/FASTA/

1 To build the BLAST database using the formatdb program, open up a command prompt. On Windows, click Start → `Run` and type cmd in the command window, and click OK. In Linux or Mac OSX, open a terminal window. Navigate to the directory on the CD called 1.2.blast_database using the 'cd' command. Copy the file `all_bacterial_genomes.faa` to your working_directory, and then navigate to your working_directory.

2 Type the following, replacing PATH with the path to the formatdb program on your computer and sequences.fasta with the name of the FASTA file you wish to convert to BLAST format:

Windows: `PATH\formatdb -i sequences.fasta -p T -o T`
Linux/OSX: `PATH/formatdb -i sequences.fasta -p T -o T`

The -i option gives the input file to the program. The -p T option tells formatdb that these are protein sequences, and the -o T option tells formatdb to parse the FASTA headers and index them for sequence retrieval.

3 Once the program completes, you will find seven additional files in this directory with the extensions .phr, .pin, .pnd, .pni, .psd, .psi, .psq. The FASTA sequences have been categorized and indexed by the program to allow for rapid searches by the BLAST alignment algorithm.

6 Appendix B

RAxML was developed by Alexandros Stamakakis for computing ML trees on large datasets [12]. Here we will make use of RAxML to compute an ML tree for the set of sequences we have created. We have precompiled binaries of RAxML for Windows, Mac Intel, Mac PowerPC, Linux x86, and Linux AMD. We have also provided the source code and makefile. RAxML documentation and source code can be downloaded from the following website, although this is not necessary for you at this time:

http://icwww.epfl.ch/ stamatak/index-Dateien/Page443.htm

6.1 Saving QR subset of alignments in PHYLIP and FASTA format

1 We must again load the sequences that we created in the first two sections. Let's open a new MultiSeq window from the VMD Plugins menu. Now go to File → Import Data. Make sure From Files is selected, and click the Browse button. Navigate to the working_directory and open the `EFTu_all_alignment.fasta` file that we created earlier. Click OK to import the aligned sequences.

2 Now we need to export these marked sequences into a format which can be used by RAxML. Select File → Export Data. Select the Marked Sequences radio button, and choose as Data Type the Sequence Data (PHY) option. Now click the Browse button next to the Filename text box and navigate to the working_directory. Type in the filename `QR10_eftu.phy`. Click OK. This will export the sequence data in PHYLIP format, which is the input format for RAxML. If you have been unable to complete the previous sections, you will find this file in the 4.phylogenetic_tree directory on the CD.

6.2 Calculating Maximum Likelihood Trees with RAxML

The simplest method of running RAxML is to copy the binary file from the appropriate directory on the CD to the working_directory where you have saved the exported sequences in PHYLIP format. Please look in the 2.RaxML directory on the CD to locate the binary that is appropriate for your system. You can also compile RAxML for your system by copying the source directory to your computer and running 'make', although this option requires that you have a C compiler such as gcc configured for your system.

1 Based on the installation instructions above, copy the appropriate RAxML binary to your working_directory where you save your PHYLIP file.

2 Open a command prompt on your system. if you are running Windows, click Start → Run and type in 'cmd' into the command box. Click OK. This will open the Windows DOS prompt. If you are running Linux or MacOSX, open a terminal window. Navigate to the working_directory where you have saved the PHYLIP export file.

3 To create the Maximum Likelihood tree, type the following at the command prompt:

```
Windows: raxmlHPC.exe -f d -m PROTMIXWAG -s AB_eftu.phy -n eftu -# 1
Linux/Mac: ./raxmlHPC -f d -m PROTMIXWAG -s AB_eftu.phy -n eftu -# 1
```

The -f option refers to the type of algorithm/function you want RAxML to execute. In this case we are specifying the default (d) hill-climbing algorithm. The -m option refers to the substitution model to be used. The -s option indicates the sequence data file, which must be in PHYLIP format. The -n option gives the output file prefix. Finally, the -# option gives the number of alternative runs on distinct starting trees.

4 The program should take about 5 minutes to create one replicate of the maximum likelihood phylogenetic tree over our reduced set of sequences. As it completes each tree, it assigns a likelihood score to each. At the end of all calculations, it prints which tree is the most likely to the console. This information is also found at the end of the file `RAxML_into.eftu`. Typically RAxML chooses the best tree out of thousands of possible replicates; we are simply restricting the number of computed trees to one for this tutorial. The completed maximum likelihood phylogenetic tree is found in the file `RAxML_result.eftu`. If you have been unable to complete this section, this file can be found in the directory 2.1.phylogenetic_tree on the CD.

6.3 Calculating Bootstrapping Values with RAxML

Bootstrapping is a method to assign confidence scores in your phylogenetic tree. Bootstrapping involves generating many phylogenetic trees using random resamplings of the original sequences. Each phylogenetic tree is then compared with the other trees, and the structural similarities are determined. If a particular subtree persists across many phylogenetic trees despite many sequence permutations, it is assigned a high bootstrapping score. For example, if 50 bootstrapping trees are computed, and a particular subtree is found in 40 of those trees, that subtree will be assigned a bootstrapping value of 80, indicating 80% of trees contained that subtree. Therefore, the highest bootstrapping value that can be assigned is 100, and the lowest is zero. Typically, a bootstrapping value above 70 is required for confidence that a particular subtree accurately represents the phylogeny of the underlying organisms.

1 The next step is to compute the bootstrapping values for the best likelihood tree. You need to provide RAxML with the original alignment in PHYLIP format and the best likelihood tree from the previous step. Usually, many thousands of bootstrapping inferences are calculated for the best likelihood tree, but we will only calculate 2 for this tutorial. To calculate the bootstrapping values, type the following at the command prompt. (The calculation will take about 5 minutes. All commands below should be entered on the same line):

Windows:
```
raxmlHPC.exe -f d -b 12345 -m PROTMIXWAG -s AB_eftu.phy
-n AB_eftu.boot -t RAxML_result.eftu -# 2
```

Linux/Mac:
```
./raxmlHPC -f d -b 12345 -m PROTMIXWAG -s AB_eftu.phy
-n AB_eftu.boot -t RAxML_result.eftu -# 2
```

The -f, -m, -s, and -n options are the same as explained in section 6.2. The -b option gives RAxML a bootstrapping random number seed. The -t option refers to the starting tree in Newick format. Finally, the -# option gives the number of bootstrap analyses RAxML will execute.

The resulting bootstrapping file will be created with the filename `RAxML_bootstrap.AB_eftu.boot`. If you have been unable to complete this section, you will find this file in the directory 2.1.phylogenetic_tree.

6.4 Mapping Bootstrapping Values onto the Highest Scoring Maximum Likelihood Tree

Now that we have generated bootstrapping values for the highest scoring maximum likelihood tree, the final step is to map these values onto the tree so they will be displayed when we load the tree into MultiSeq. This step only takes a few seconds to complete.

1 Finally, we must map the best bootstrapping values onto the best likelihood tree. We do this using the following command:

Windows:
```
raxmlHPC.exe -f b -m PROTMIXWAG -s AB_eftu.phy -n AB_eftu.boot.tre
-z RAxML_bootstrap.AB_eftu.boot -t RAxM_result.eftu
```

Linux/Mac:
```
./raxmlHPC -f b -m PROTMIXWAG -s AB_eftu.phy -n AB_eftu.boot.tre
-z RAxML_bootstrap.AB_eftu.boot -t RAxM_result.eftu
```

The -f b option indicates that RAxML should draw the bipartitions using the bootstrapped trees onto the best scoring maximum likelihood tree. The -m, -s, and -n options are the same as explained in section 6.2. The -z option tells RAxML the filename for the bootstrap result file. Finally, the -t option refers to the starting tree in Newick format.

2 The final phylogenetic tree will be located in the file `RaxML_bipartitions.AB_eftu.boot.tre`. If you have been unable to complete this section, you will find this file in the directory 2.1.phylogenetic_tree.

7 Acknowledgments

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

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, 1990.

[2] J Eargle, A.A. Black, A. Sethi, L.G. Trabuco, and Z. Luthey-Schulten. Dynamics of recognition between trna and elongation factor tu. *J. Mol. Biol.*, 377:1382–1405, 2008.

[3] Michael Y. Galperin, Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin. Expanded microbial genome coverage and improved protein family annotation in the cog database. *Nucleic Acids Research*, 43(D1):D261–D269, 2015.

[4] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *PNAS*, 89:10915, 1992.

[5] Kazutaka Katoh, Kazuharu Misawa, Keiichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. 30(14):3059–3066, 2002.

[6] Masaru Katoh. Comparative genomics on wnt3-wnt9b gene cluster. *International journal of molecular medicine*, 15(4):743–747, 2005.

[7] E.V. Koonin and Y.I. Wolf. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, 36(21):6688–6719, 2008.

[8] Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin. Archaeal clusters of orthologous genes (arcogs): An update and application for analysis of shared features between thermococcales, methanococcales, and methanobacteriales. *Life*, 5(1):818, 2015.

[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(3):875–894, 2005.

[10] E. Roberts, A. Sethi, J. Montoya, C.R. Woese, and Z. Luthey-Schulten. Molecular signatures of ribosomal evolution. *PNAS*, 105(37):13953–13958, 2008.

[11] A. Sethi, P. O'Donoghue, and Z. Luthey-Schulten. Evolutionary profiles from the qr factorization of multiple sequence alignments. *PNAS*, 102(11):4045–4050, 2005.

[12] A. Stamatakis. Raxml-vi-hpc: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[13] L.G. Trabuco, E. Villa, K. Mitra, J Frank, and K Schulten. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure*, 16:673–683, 2008.