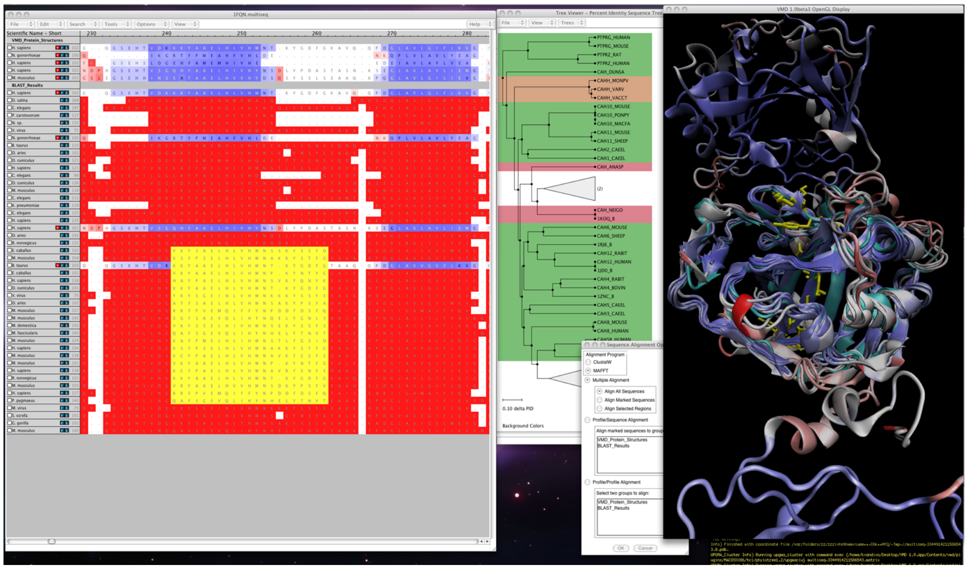
1 Introduction

MultiSeq (shown in Fig. 1) is a unified bioinformatics analysis environment that allows one to organize, display, and analyze both sequence and structure data for proteins and nucleic acids. Special emphasis is placed on analyzing the data within the framework of evolutionary biology. MultiSeq was created to allow biomedical researchers to study the evolutionary changes in sequence and structure of proteins across all three domains of life, from bacteria to humans. The comparative sequence and structure metrics, and analysis tools introduced in the article by O'Donoghue and Luthey-Schulten [1] are part of MultiSeq. In particular, the Luthey-Schulten group has included a structure-based measure of homology Q_H (see Appendix B), that takes into account the effect of insertions and deletions and has been shown to produce accurate structure-based phylogenetic trees. The STAMP structural alignment algorithm, kindly provided by our colleagues Russell and Barton, is included [2]. As a result, Multiple Alignment is an invaluable tool for relating protein structure to its function or misfunction.



Alignment Program
○ ClustalW
● MAFFT
● Multiple Alignment
● Align All Sequences
○ Align Marked Sequences
○ Align Selected Regions
○ Profile/Sequence Alignment
Align marked sequences to group
VMD_Protein_Structures
BLAST_Results
○ Profile/Profile Alignment
Select two groups to align:
VMD_Protein_Structures
BLAST_Results

Figure 1: MultiSeq In VMD

In any publication of scientific results based completely or in part on the use of MultiSeq, please reference:

Elijah Roberts, John Eargle, Dan Wright, and Zaida Luthey-Schulten. MultiSeq: Unifying sequence and structure data for evolutionary analysis. BMC

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

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

Bioinformatics, 2006, 7:382.

1.1 Installation

MultiSeq is part of the standard VMD release. You can download VMD from `http://www.ks.uiuc.edu/Research/VMD/`. Although BLAST is not necessary for the overall function of MultiSeq, it is highly recommended to have BLAST installed locally (i.e. accessible through file browsing on your local computer). See `http://www.scs.illinois.edu/~schulten/multiseq/` for links to tutorials with additional information on BLAST installation.

ClustalW is the default sequence alignment tool and is packaged with MultiSeq. However, MAFFT (available from `http://mafft.cbrc.jp/alignment/software/`) can be used for doing sequence alignment if it is installed on your computer system. (MAFFT version 6.811 has been tested. Newer versions are expected to work as well and should be used if possible.)

Paths to all locally installed software and databases are set via the File — Preferences menu in the MultiSeq window. The Preferences menu has a 'Metadata' tab and a 'Software' tab. The 'Software' tab is where file paths can be provided.

MultiSeq uses a collection of databases that need to be downloaded to your computer system. The first time you run MultiSeq you will be asked to create a folder to store these databases, and the databases will then be downloaded from our servers. When you subsequently run the plugin, it will check to insure that you have the most recent versions of the databases.

2 The MultiSeq Graphical Environment

MultiSeq is accessed as an extension within VMD. To begin MultiSeq, launch VMD and:

1. In the VMD main window, click on the Extensions Menu.

2. In Extensions, select Analysis → MultiSeq.

(alternatively, if you are a fan of command lines, you can type '`multiseq`' into the VMD terminal window)

The main MultiSeq window (see Fig. 2) will appear (note that the first time you run MultiSeq, you will be prompted to download necessary databases before seeing the main window).

3 Using and Managing Data

To begin analyzing proteins in MultiSeq, data from sequence[3] and structure[4] files is required. Import Data (from the File menu within MultiSeq) allows you

[3] FASTA files.

[4] The ASTRAL database (http://astral.stanford.edu) is a compendium of protein domain structures derived from the PDB database. It divides each protein structure into its domain

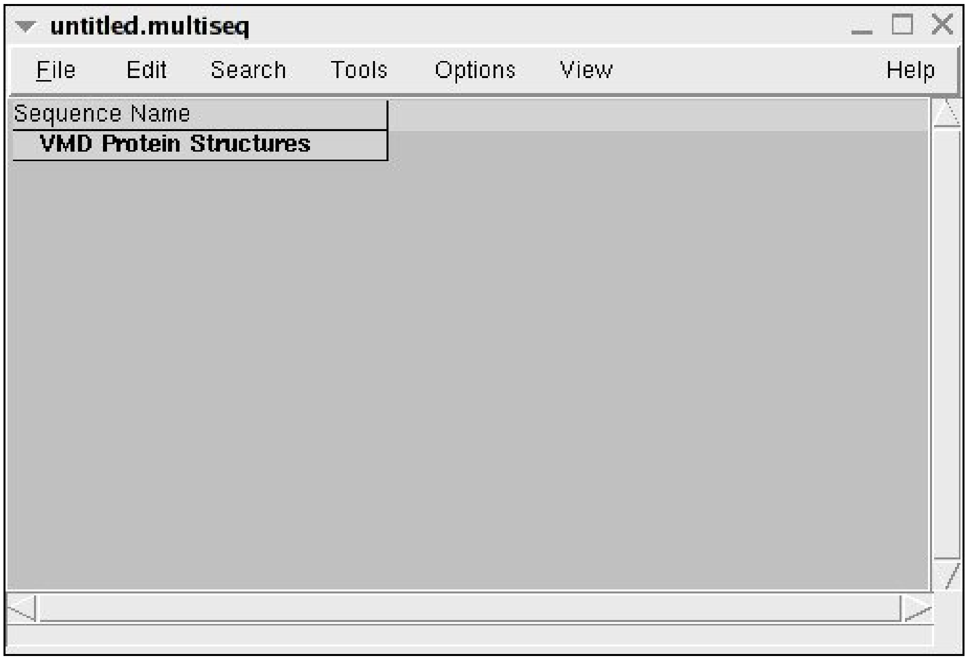


Figure 2: Main MultiSeq Window With No Structures Loaded

to load structure and sequence files, both locally and via a network connection.

Various structure and trajectory files, such as PDB and PSI, can be loaded via the New Molecule function of the VMD Main window, but Import Data allows you to load sequence files as well. Additionally, Import Data has BLAST searching capabilities, if a local copy of BLAST is installed.

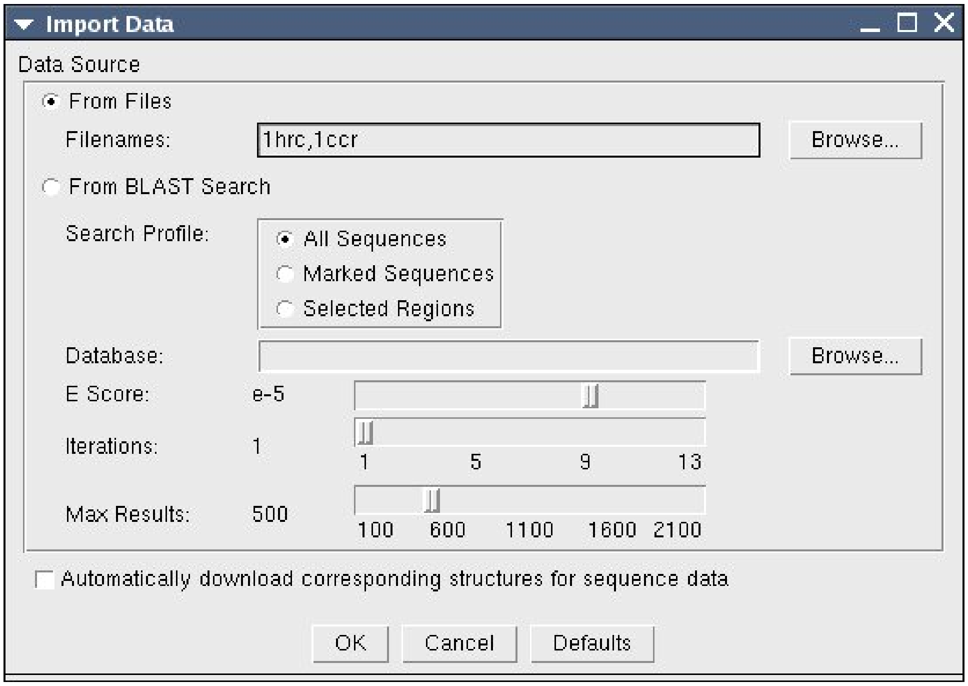
3.1 Importing from files

Structure[5] and Sequence files can be loaded into MultiSeq via Import Data. PDB files are structure files, whereas FASTA is a sequence file format. To load these files:

1. Make sure From Files is selected as a Data Source.
2. In the Filenames: dialogue, either type in the location of the file, or hit the browse button to locate the file. Another option is to simply type in

components. For example, AspRS is divided into three separate PDB files: one containing the catalytic domain, one with the insertion domain, and one for the anticodon binding domain. The names of the files contain the PDB extension, the letter a for ASTRAL, and a number, which corresponds to which domain it is in the original PDB file. The PDB is the single worldwide repository for the processing and distribution of 3-D structure data of large molecules of proteins and nucleic acids.

[5] See VMD Manual for supported formats



Import Data

Data Source

● From Files

Filenames: 1hrc,1ccr

○ From BLAST Search

Search Profile:

● All Sequences

○ Marked Sequences

○ Selected Regions

Database:

E Score: e-5

Iterations: 1

Max Results: 500

□ Automatically download corresponding structures for sequence data

Figure 3: Import Data Window

the PDB or SCOP id. This option requires a network connection for your computer to obtain files from PDB or ASTRAL directly.

3. Hit the OK button.

If you would like to load multiple files/structures/sequences at once, you can separate each with a comma.

3.2 Sequences and BLAST searching

You can conduct a BLAST search from within MultiSeq if you have the BLAST program installed on your computer. You will need to install BLAST if you haven't already done so, and you will have to configure MultiSeq to know where BLAST is installed (via File — Preferences — Software)

1. Before you open the Import Data window, you have the option of either selecting a set of sequences, or a region within a sequence.

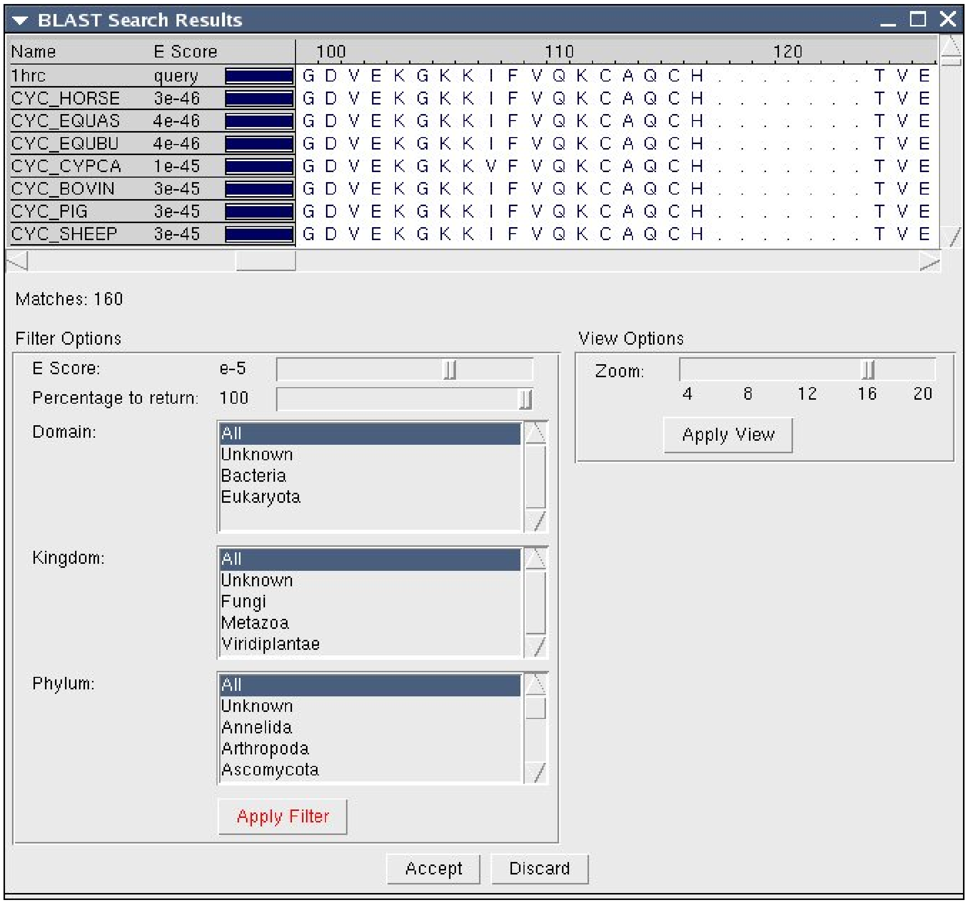
2. Go to File and then Import Data and select From BLAST Search, and either All Sequences, Marked Sequences, or Selected Regions.

3. In the Databases, either type the location of the database, or use the Browse button to locate it. This could be something like a Swiss PROT database or otherwise. Once you give MultiSeq the name of a database, it will remember it for future searches.

4. Select the E Score, Iterations, and Max Results.

5. If you want MultiSeq to automatically download structure information for sequences found via the BLAST search, mark the checkbox for that.

6. Hit the OK button.



Name	E Score	100 … 110 … 120
1hrc	query	G D V E K G K K I F V Q K C A Q C H . . . . T V E
CYC_HORSE	3e-46	G D V E K G K K I F V Q K C A Q C H . . . . T V E
CYC_EQUAS	4e-46	G D V E K G K K I F V Q K C A Q C H . . . . T V E
CYC_EQUBU	4e-46	G D V E K G K K I F V Q K C A Q C H . . . . T V E
CYC_CYPCA	1e-45	G D V E K G K K V F V Q K C A Q C H . . . . T V E
CYC_BOVIN	3e-45	G D V E K G K K I F V Q K C A Q C H . . . . T V E
CYC_PIG	3e-45	G D V E K G K K I F V Q K C A Q C H . . . . T V E
CYC_SHEEP	3e-45	G D V E K G K K I F V Q K C A Q C H . . . . T V E

Matches: 160

Filter Options

E Score: e-5

Percentage to return: 100

Domain: All, Unknown, Bacteria, Eukaryota

Kingdom: All, Unknown, Fungi, Metazoa, Viridiplantae

Phylum: All, Unknown, Annelida, Arthropoda, Ascomycota

View Options

Zoom: 4 8 12 16 20

Figure 4: BLAST Search Results

MultiSeq will then begin a BLAST search. This may take several minutes. When the search is done, a new window called BLAST Search Results will appear. The results do not immediately appear in the main MultiSeq window, because you can apply further filters on the retrieved sequences. The BLAST Search Results window is divided into three main parts: the sequence viewer, Filter Options, and View Options.

The sequence viewer is a read-only display of the sequences that match your BLAST search. The number of matches is listed below the sequence viewer.

You can use the Zoom to change how much of each sequence you see. You can change the zoom level and Apply View and you will see fewer or more sequences in the sequence viewer portion of the window.

In the Filter Options you can tweak the parameters to reduce or expand the number of sequence matches. Once you have changed a parameter you can hit Apply Filter and see which sequences match.

Once you have a collection of sequences that you want to import, you can hit the Accept button at the bottom and they will be added to the MultiSeq window.

4 Working in the Environment

MultiSeq provides a unique working environment for the analysis of proteins.

4.1 Title Display

By default, for each sequence loaded into Multiseq, you will be shown the "sequence name" as the title for each row in the main window. Sometimes this is not as useful as, for instance, the scientific name of the sequence might be. Multiseq allows you to change the displayed title for each sequence by by left clicking on the header of the titles and choosing a different option. This can be seen in Figure 5. If you choose an option where a sequence does not have a value, Multiseq will show you the <Sequence Name> in angle brackets.

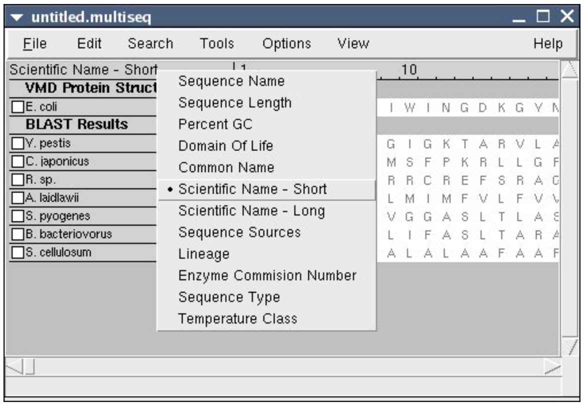


Figure 5: Choosing Data To Display As Sequence Title

4.2 Grouping

While working with the Sequence Viewer in MultiSeq, you may notice certain patterns or trends. As a result you would like to put certain sequences closer to others to analyze such motifs. MultiSeq allows such grouping based on taxonomy or you can customize the groupings. Right clicking on a group name (such as VMD Protein Structures will bring up a context menu where you can manage groups.

4.3 Info Viewer

Whenever you load a sequence or structure into MultiSeq an '`i`' box will appear next to the protein's ID. If you click on this box, a new window will appear called the Info Viewer (See Fig. 6). Within this window information regarding the species the protein is from will appear. If you have PSIPred installed and configured, you can predict the secondary structure at the bottom of the Info window.

4.4 Selecting vs. Marking

As you browse the menus of MultiSeq you will notice options for Selected Sequences or Marked Sequences. "Selecting Sequences" is when you highlight a portion of the sequence(s) in the sequence viewer using the mouse. This can be either the entire sequence or a portion. However "Marking Sequences" allows you to more easily select an entire sequence by simply checking the box next to the protein ID.

5 Edit Menu

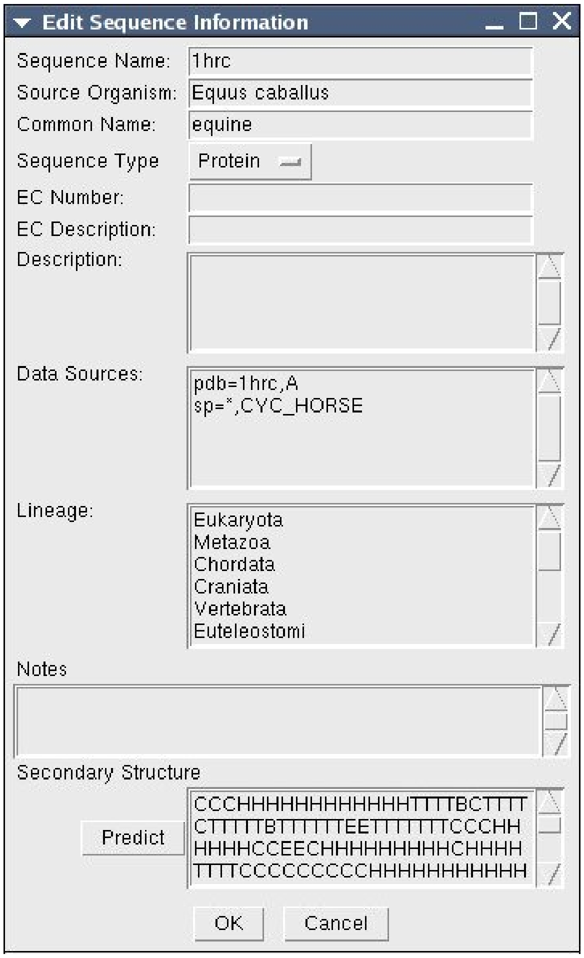
Along with the copy/cut/paste options that you expect to see in an edit menu, this menu also provides a power sequence editor.

If you want to edit a sequence, Enable Editing. If you are just wanting to align sequences, you can probably choose to just enable gap editing. Once you have enabled editing, you can then use the mouse to choose a residue (or residues). Hit the space bar to insert a gap, or, if you have enabled full editing, you can insert a residue by typing the desired character.

If you want to truly edit the sequences manually, you can choose to Edit In Text Editor. VMD's text editor will be loaded, and you can change the sequence data. Dashes are gaps and the sequence characters can be changed as you see fit.

6 Search Menu

Find, Find Next, Find Previous Via Search, you can find and highlight residues in sequences. When you use Find, all of the residues will be



Sequence Name: 1hrc

Source Organism: Equus caballus

Common Name: equine

Sequence Type: Protein

EC Number:

EC Description:

Description:

Data Sources:
pdb=1hrc,A
sp=*,CYC_HORSE

Lineage:
Eukaryota
Metazoa
Chordata
Craniata
Vertebrata
Euteleostomi

Notes

Secondary Structure

```
CCCHHHHHHHHHHHHTTTTBCTTTT
CTTTTTBTTTTTTEETTTTTTTCCCHH
HHHHCCEECHHHHHHHHHCHHHH
TTTTCCCCCCCCCHHHHHHHHHHH
```

Figure 6: Edit Sequence Information

highlighted, and you can then cycle through them by using Find Next and Find Previous.

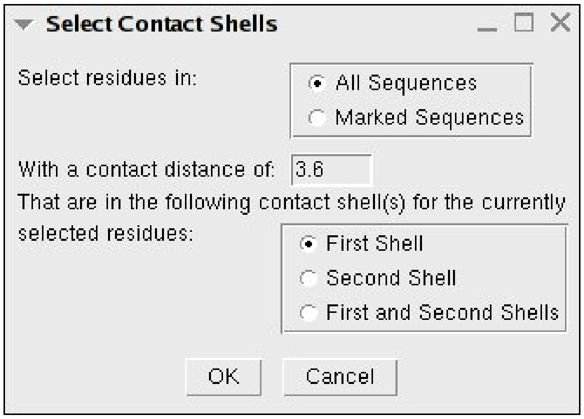
Select Contact Shells See Figure 7.

Select residues in: Lets you choose whether to look through all sequences, or just the ones you have marked.

With a contact distance of: defaults to 3.6.

That are in the following contact shell(s) for the currently selected residues Choose from First, Second, or First and Second shells

Select Non-Redundant Set You can use structure QR or sequence QR to select a non-redundant set (See Fig. 8).



Select Contact Shells

Select residues in:
● All Sequences
○ Marked Sequences

With a contact distance of: 3.6

That are in the following contact shell(s) for the currently selected residues:
● First Shell
○ Second Shell
○ First and Second Shells

Figure 7: Select Contact Shell Window

Select from: Lets you choose whether to look through all sequences, or just the ones you have marked.

Using Structure QR QH Cutoff: Can vary from 0 to 1.

Using Sequence QR Identity Cutoff:

Gap Scale Factor:

Seed with selected sequences If you have selected certain sequences, you can seed the algorithm with these sequences to select a non-redundant set based on them.

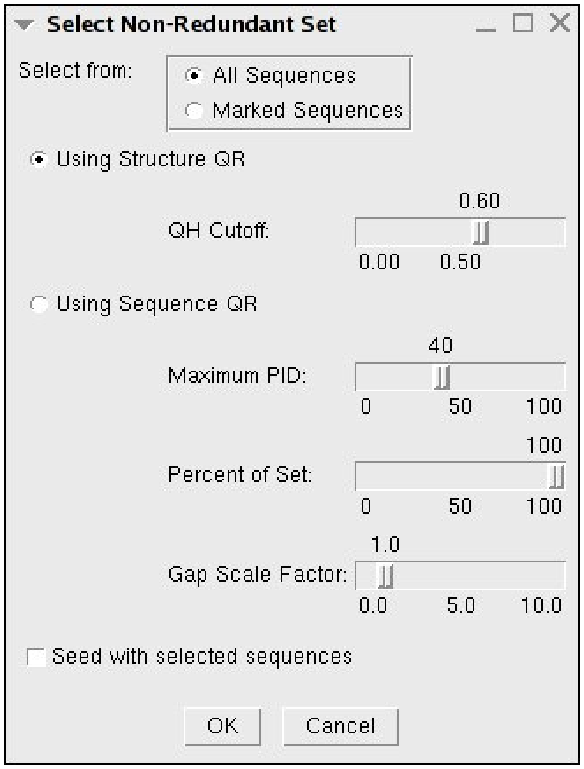
Select Residues The Residue Selection feature (See Fig. 9)lets you analyze conservation, using different measures, and highlight residues in the Sequence Display and Structure Display simultaneously. Residue Selection allows you to examine the conservation on a per residue basis.

There are two options: either Where Sequence Identity is or Where Qres is. Where Sequence Identity is is a sequence identity measure, whereas Where Qres is is a structure measure.

Select residues in: You can choose all sequences or just the marked ones.

Where Sequence Identity is: If this option is selected you can select ‘less than or equal to’ or ‘greater than or equal to’ option, then a number between 0-99.

Where Qres is: If this option is selected you can select ‘less than or equal to’ or ‘greater than or equal to’ option, then a number between zero and one.



Select Non-Redundant Set

Select from: ● All Sequences ○ Marked Sequences

● Using Structure QR

QH Cutoff: 0.60

○ Using Sequence QR

Maximum PID: 40

Percent of Set: 100

Gap Scale Factor: 1.0

□ Seed with selected sequences

Figure 8: Select Non-Redundant Set Window

7 Tools Menu

7.1 Performing Alignments

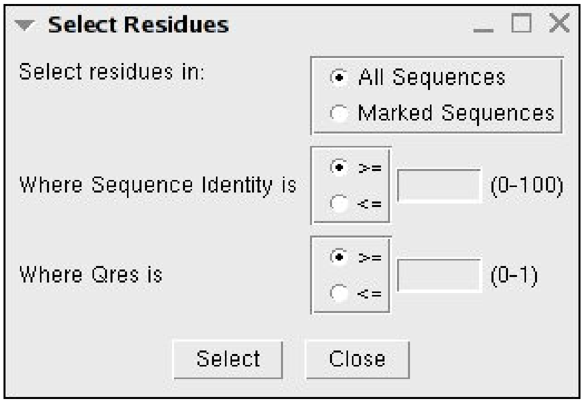
MultiSeq can do both structural and sequence alignments. These options are available via the Tools menu in MultiSeq.

7.1.1 Structure Alignments

MultiSeq uses the program STAMP to structurally align protein molecules. The STAMP algorithm minimizes the C_α distance between aligned residues of each molecule by applying globally optimal rigid-body rotations and translations. Also, note that you can perform alignments on molecules that are structurally similar. If you try to align proteins that have no common structures, STAMP will have no means to align them. If you would like further information about how the alignment occurs, please refer to the STAMP manual.

Align the following: Choose which structures you wish to align

Number of passes (npass): Whether one or two fits are to be performed.



Select Residues

Select residues in: ● All Sequences ○ Marked Sequences

Where Sequence Identity is ● >= ○ <= (0-100)

Where Qres is ● >= ○ <= (0-1)

Select Close

Figure 9: Select Residues Window

etc.

The idea is that the initial fit can be used with a conformation biased set of parameters to improve the initial fit prior to fitting using distance and conformation parameters. Default NPASS = 1

Similarity (scanscore): Specifies how the Sc value (STAMP algorithm) is to be calculated. This depends on the particular application. As a general rule of thumb, use SCANSCORE=6 for large database scans, when you are scanning with a small domain, and wishing to find all examples of this domain - even within large structures. Use SCANSCORE=1 when you wish to obtain a set of transformations for a set of domains which you know are similar (and have defined fairly precisely as domains rather than the larger structure that they may be a part of). Default SCANSCORE = 6

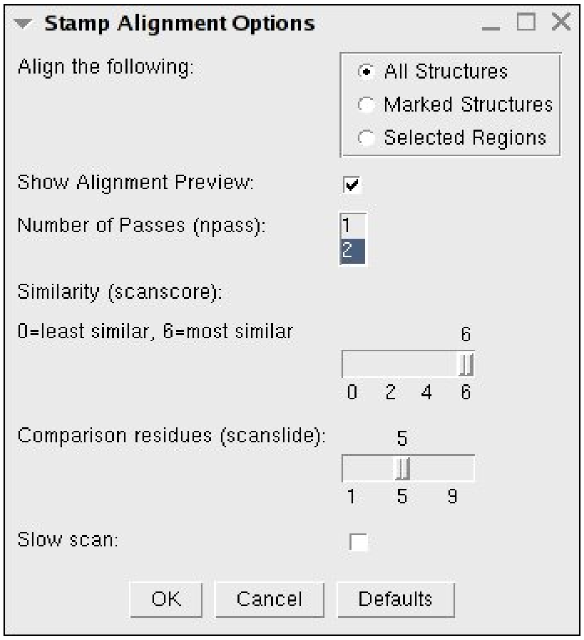
Comparison residues (scanslide): This is the number of residues that a query sequence is 'slid' along a database sequence to derive each initial superimposition. Initially, the N-terminus of the query is aligned to the 1st residue of the databse, once this fit has been performed and refined, and tested for good structural similarity, the N-terminus is aligned with the 1+th position, and the process repeated until the end of the database sequence has been reached. Default SCANSLIDE = 5

Slow scan: If set to TRUE, then the SLOW method of getting the initial fits for scanning will be used (See chapter 1). Default SLOWSCAN = FALSE

Defaults: resets the STAMP parameters to their original values

7.1.2 Sequence Alignments

Sequence alignment in MultiSeq can be done via ClustalW or MAFFT (if you have MAFFT locally installed) (See Fig. 11).



Stamp Alignment Options

Align the following: ● All Structures ○ Marked Structures ○ Selected Regions

Show Alignment Preview: ☑

Number of Passes (npass): 1 2

Similarity (scanscore):
0=least similar, 6=most similar 6

Comparison residues (scanslide): 5

Slow scan: □

Figure 10: STAMP Structural Alignment Window

Once you have decided which program to use (you won't be able to select MAFFT if you haven't configured the path to MAFFT on your local computer via File — Preferences — Software), you can choose from Multiple Alignment, Profile/Sequence Alignment, or Profile/Profile Alignment. Once you have chosen the desired type of alignment, you can set the proper option.

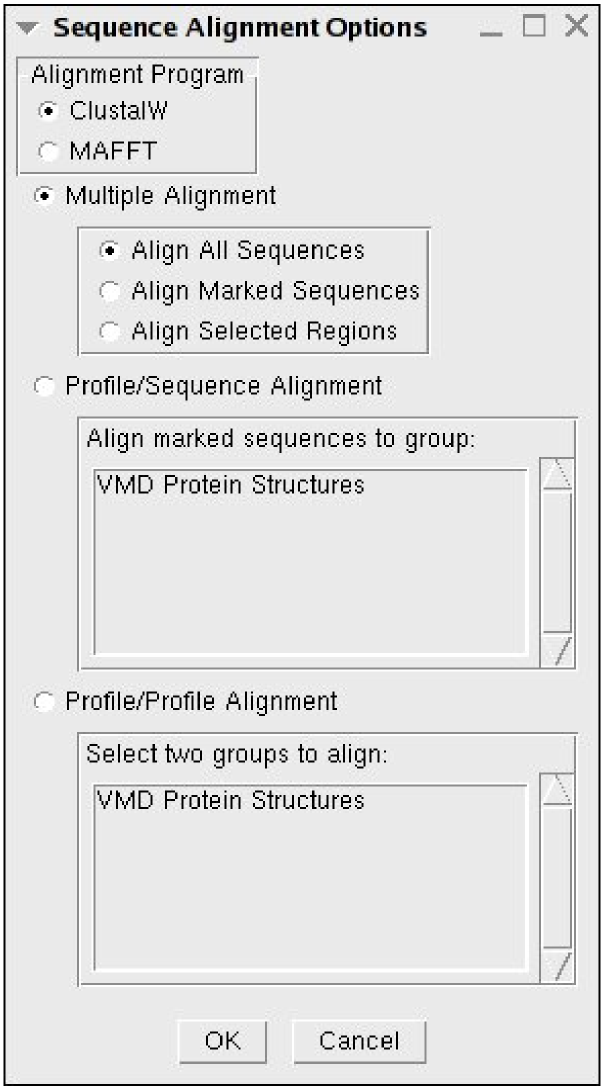
Multiple Alignment Choose which sequences or regions you wish to align.

Profile/Sequence Alignment This requires certain sequences to be marked, and they will then be aligned to the group that you specify.

Profile/Profile Alignment To align one entire group with another entire group, select this option.

If you choose MAFFT, be aware of the following:

- MultiSeq has been tested with MAFFT version 6.811. It should work with any version of MAFFT reasonably close to that.
- MultiSeq uses the default -auto option for MAFFT.
- Profile-profile and sequence-profile alignment will be done with MAFFT if it is chose as the desired alignment program.



Sequence Alignment Options

Alignment Program

● ClustalW

○ MAFFT

● Multiple Alignment

● Align All Sequences

○ Align Marked Sequences

○ Align Selected Regions

○ Profile/Sequence Alignment

Align marked sequences to group:

VMD Protein Structures

○ Profile/Profile Alignment

Select two groups to align:

VMD Protein Structures

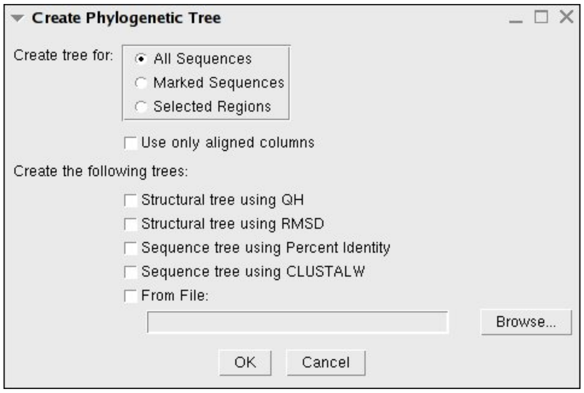
Figure 11: Sequence Alignment Menu Window

- When configuring the path to MAFFT, you need to give the path to the 'bin' directory on a unix-type system. On Windows, give the path that contains the 'mafft.bat' file.

7.2 Phylogenetic Tree

The Phylogenetic Tree feature helps in determining the structure and sequence-based relationships between the aligned domains of proteins.

To do this, it uses a modification of Q that accounts for both gapped and aligned regions. This new metric, Q_H, creates a structure-based phylogeny that is congruent to the sequence-based phylogenies. You can create a Phylogenetic Tree from the Tools menu in MultiSeq (See Fig. 12). Once you choose the sequences or regions you wish to create a tree for, you can choose which trees you want to create. The tree viewer can also create a tree from a data file that you provide (if you have created tree data from an external program, for instance).



Create tree for:
● All Sequences
○ Marked Sequences
○ Selected Regions

□ Use only aligned columns

Create the following trees:
□ Structural tree using QH
□ Structural tree using RMSD
□ Sequence tree using Percent Identity
□ Sequence tree using CLUSTALW
□ From File:

Figure 12: Create Phylogenetic Tree Window

Once you have chosen which trees to create, the Tree Viewer will be shown in simple black and white. But, you can easily use color and Tree View commands to make the data more useful (see Fig. 13).

The Tree Viewer window is very powerful. In the main window, you can right click on any small black box (in front of an individual sequence, or at any joint in the tree) and remove the element/subtree or look at its properties. Additionally, if you have selected a subtree, you can change the shape of the tree. You can collapse/expand a subtree, as shown in Fig. 13 as well.

Menu options include:

File Trees can be loaded and saved in common formats. Additionally, postscript renderings can be created for use in publications.

View If a distance matrix has been created from the data, you can view it. You can also modify the way the tree looks. You can zoom in and out,

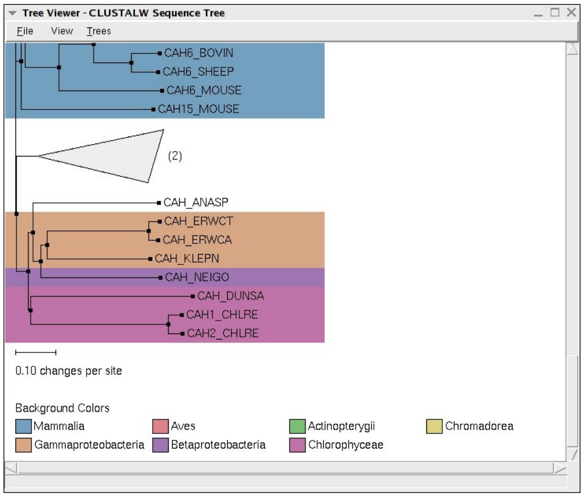


Figure 13: Phylogenetic Tree Viewer - CLUSTALW Sequence Tree

change the scale (which pushes tree leaves left or right for viewability). Orientation will move the labels from the left side of the tree to the right, and you can even choose whether or not you wish the tree to display the labels and nodes.

The Leaf Text option lets you choose the labels that you wish to have displayed, and you can color the labels as well as the tree backgrounds by a variety of different metrics.

You can easily collapse large parts of the tree by choosing a criteria, and, if you have selected a point in the tree, you can make that point the new root node of the tree.

Trees If you have chosen to create multiple trees, you can use this menu to rotate through the trees, or you can jump to one directly.

7.3 Plot Data

Plot Data create graphs of internal MultiSeq data. You can Plot Data from the Tools menu in MultiSeq. Once you choose the sequences or regions you wish to plot, you can choose the data (such as Qres, RMSD, etc) for each residue that you want to display. You can also plot custom data. The data graph will then be displayed (see Fig. 14). If you wish, you can create a postscript file for publication.

8 Options Menu

Atom Picking Normally disabled, but can be turned on.

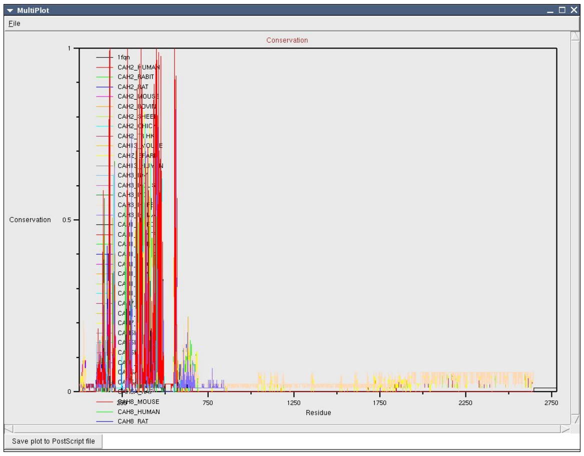


Figure 14: Plot Data - Sequence Conservation

Grouping MultiSeq can automatically create groups and show the sequences in the MultiSeq window sorted accordingly. Just choose the grouping that you want to use.

9 View Menu

The View menu provides several useful options for coding and looking at large amounts of data.

Zoom To change the amount of data seen in the MultiSeq window, you can zoom in and out. As you zoom farther out (choosing a percentage that is smaller) MultiSeq will makes the sequence letters smaller and smaller until you will only see the background colors and then, not even that. If you need to see the entire sequence, the Zoom Window, discussed below, might be more useful.

Coloring You can choose to color the sequences by a wide range of attributes. First, you can choose *what* you want to color by choosing Apply to All, Group, or Marked. Then, you can choose the coloring method that you wish to apply.

For Qres, traditionally, Q has meant "the fraction of similar native contacts" between the aligned residues in two proteins[6], or in two different conformational states of the same protein. When Q = 1, it indicates that the structures are identical. When Q has a low score (*0.1*), it means the

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

structures do not align well, or, in other words, only a small fraction of the C-alpha atoms superimpose. You will discover that homologs typically have Q≥0.4. Q per residue is the contribution from each residue to the overall average Q score. For more information see Appendix A.

For Sequence Identity, the aligned domains are colored by how much of the sequence is conserved. The Sequence Identity coloring method colors each amino acid according to the degree of conservation within the alignment: blue means highly conserved, wheras red means very low or no conservation.

Highlight Style Highlight Style is an option for the OpenGL diplay. The style refers to drawing method in VMD[7]. This option allows a user to highlight residues of a structure in the sequence display and see the areas simultaneously highlighted in the OpenGL display.

Highlight Color Highlight Color is another option for the OpenGL diplay. Alongside Highlight Style, Highlight Color is the color or coloring method[8] used in the OpenGL display when highlighting residues in the Sequence Display. The default Highlight Color is yellow.

Color Scale Once you have chosen a coloring, you might wonder what the specific colors mean. The Color Scale option will show you the scale of colors according to value.

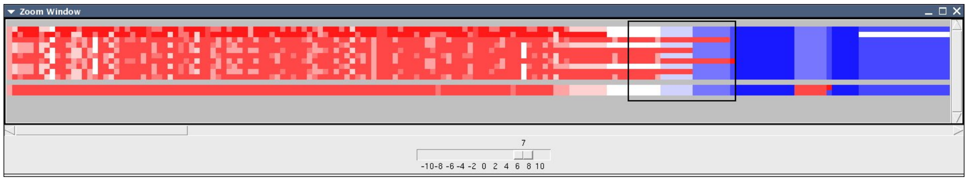


Figure 15: Zoom Window

Zoom Window (See Fig. 15) If you need to see the entire collection of sequences and quickly move from area to area, the Zoom Window will be useful to you. It shows the entire sequence palette. You can choose the zoom factor using the sliding bar at the bottom of the window, and the black box shows you the area of the sequences that are currently visible in the MultiSeq window. To see other areas, just click the mouse and the black box will be moved to the mouse pointer location.

Note: When you have the Zoom Window open, the MultiSeq window will redraw more slowly. If this is a problem for you, just close the Zoom Window and reopen as needed.

[7] For more information about drawing methods, please refer to the VMD manual.

[8] For more information, please refer to the VMD manual

10 Working With Sessions

The Load and Save Session options from the File menu provide a way to save and load all of the files, alignments, and visual representations currently in use within MultiSeq in a convenient package.

10.1 Save Session

You can save a session of MultiSeq, with all of the files, alignments, and visual representations, by simply going to the File Menu and selecting Save Session. You will be prompted to save the session, and will have the opportunity to create a unique name for the session here. Hit the OK button. A file will be generated with a `.multiseq` extention along with a directory filled with various files necessary to load the saved session into MultiSeq. Please note that both the generated file and directory have to be in the same directory location in order to load up the session in the future properly.

10.2 Load Session

Unlike Import Data (also in the File menu), Load Session opens up a previous session of MultiSeq with all of the sequence and structure files aligned, and using previous coloring and drawing methods. To load a previously saved MultiSeq Session, simply select the File menu and Load Session. A file broswer will appear allowing you to select a file with the extension `.multiseq` and make sure it has a corresponding directory of the same name.

11 Other Ways To Export Data

11.1 Save to PostScript

From the File menu, if you choose Save Screenshot, you will be able to save a postscript version of the MultiSeq window.

12 Appendices

12.1 Appendix A: Q

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

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

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

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

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

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

12.2 Appendix B: Q_H

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

Homology Measure

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

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

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

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

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

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

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

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

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