

Contents

1	Introduction	3
1.1	Accessing Multiseq	4
1.2	Installation	4
1.2.1	BLAST	5
1.2.2	PSIPRED	5
1.2.3	MAFFT	6
2	Using and Managing Data	8
2.1	Importing from files	9
2.2	Sequences and BLAST searching	9
3	Working in the Environment	11
3.1	Title Display	11
3.2	Grouping	11
3.2.1	Moving Sequences into a group	11
3.3	Visualization Menu	12
3.4	Representation Menu	12
3.5	Info Viewer	12
3.6	Selecting vs. Marking	12
4	File Menu	13
4.1	New Session	13
4.2	Load Session	14
4.3	Save Session	14
4.4	Export Data	14
4.5	Save Screenshot	14
4.6	Preferences	14
4.7	Choose Working Directory	14
4.8	Cleanup Representation	15
5	Edit Menu	15
5.1	Enable Editing	15
5.2	Remove Gaps	15
5.3	Edit In Text Editor	15
6	Search Menu	15
6.1	Find, Find Next, Find Previous	15
6.2	Select Contact Shells	15
6.3	Select Non-Redundant Set	16
6.4	Select Residues	17

7	Tools Menu	18
7.1	Performing Alignments	18
7.1.1	Structure Alignments	18
7.1.2	Sequence Alignments	20
7.2	Phylogenetic Tree	20
7.2.1	Tree Viewer	20
7.3	Plot Data	22
8	Options Menu	23
8.1	Atom Picking	23
8.2	Grouping	23
9	View Menu	23
9.1	Zoom	24
9.2	Coloring	24
9.3	Highlight Style	25
9.4	Highlight Color	26
9.5	Color Scale	26
9.6	Zoom Window	26
10	Appendices	26
10.1	Appendix A: Q	26
10.2	Appendix B: Q_H	28
10.3	Appendix B: Q_{res} Structural Similarity per Residue	30

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. 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 as well as analysis tools introduced in the article

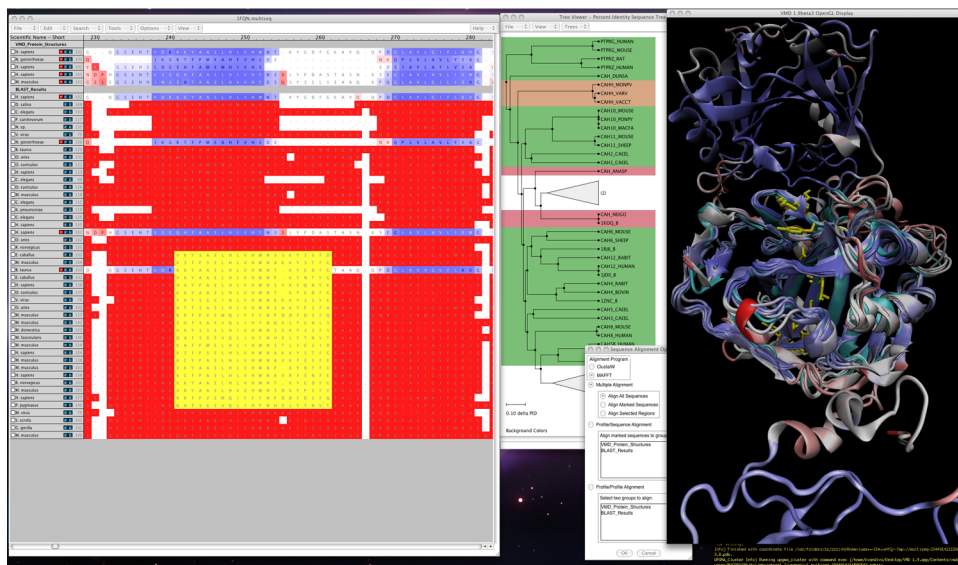


Figure 1: MultiSeq In VMD

by O’Donoghue and Luthey-Schulten¹ are part of MultiSeq. In particular, the Luthey-Schulten group has included a structure-based measure of homology Q_H (see 10.2), which takes the effect of insertions and deletions into account and has been shown to produce accurate structure-based phylogenetic trees. Multiple Alignment is an invaluable tool for relating protein structure to its function or malfunction. Therefore, the STAMP structural alignment algorithm, kindly provided by our colleagues Russell and Barton, is included².

For 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 Bioinformatics, 2006, 7:382.

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

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

1.1 Accessing Multiseq

Multiseq is part of the standard VMD release. You can download VMD from <http://www.ks.uiuc.edu/Research/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, 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)

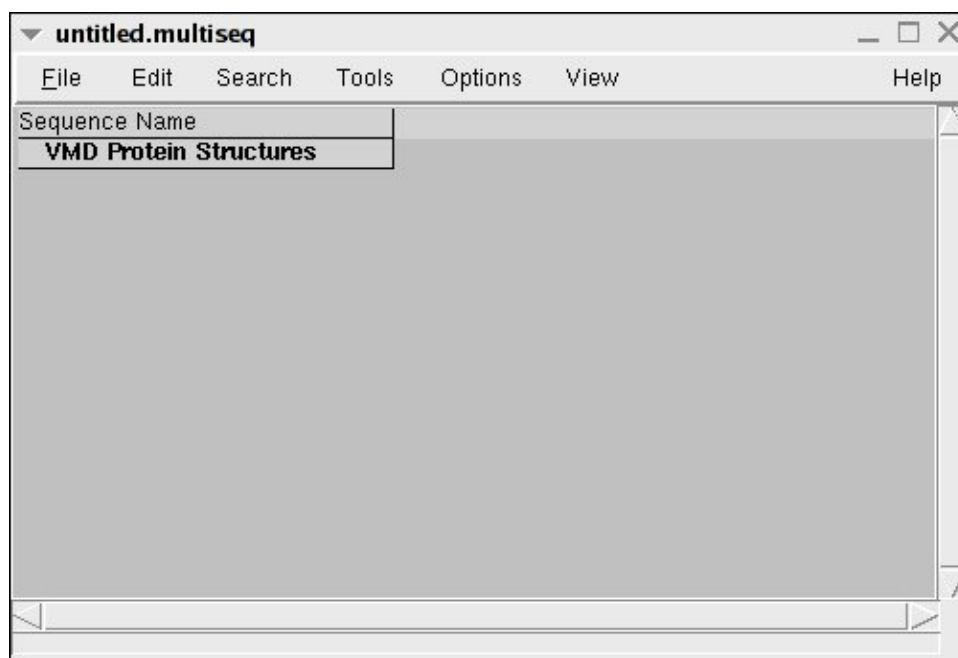


Figure 2: Main Multiseq Window With No Structures Loaded

1.2 Installation

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 as metadata. When you subsequently run the plugin, it will check to insure that you have the most recent versions of the databases and Multiseq will ask to download updates as needed. To manually download database updates or to change the Metadata directory go to File →

Preferences to bring up the preferences dialog. The directory can be changed in the section entitled **Metadata directory:** and each database Multiseq uses is listed underneath with a corresponding Download button for manual downloads.

1.2.1 BLAST

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). However, the newest BLAST release, BLAST+, is not backwards compatible with BLAST and Multiseq only supports BLAST versions that are pre-BLAST+. Therefore, this user guide provides steps to installing the latest legacy version of BLAST.

To install BLAST:

1. Go to <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/2.2.25/>
2. Choose the appropriate architecture and OS for your system to download
3. Create a directory on your local hard disk into which BLAST will be installed.
4. Extract the archive into the directory that you created for BLAST
5. You must set the BLAST installation location in MultiSeq. From the MultiSeq program window, choose File → Preferences to bring up the preferences dialog.
6. Click on the Software button in the upper left portion of the dialog to show the software preferences.
7. Click on the Browse... button in the **BLAST Installation Directory** section and select the directory into which you installed BLAST and make sure to include the bin folder within BLAST in the directory. *Note: on Linux and Mac OS X you may have a directory called blast-2.2.25 underneath your installation directory. If so, pick this directory in the browse dialog.*
8. In the BLASTDB section, you can input a file path to a database that BLAST will use by default. If blank, you must input the directory of the database when using BLAST search from File → import data dialog.

1.2.2 PSIPRED

Multiseq also uses the PSIPRED algorithm to predict secondary structure of proteins. To install PSIPRED:

1. Go to <http://bioinfadmin.cs.ucl.ac.uk/downloads/psipred/>

2. Download the `pripred321.tar.gz` file
3. Create a directory into which PSIPRED will be installed.
4. Extract the archive into the created directory
5. Follow the install instructions in the readme
6. Go into MultiSeq and choose File → Preferences to bring up the preferences dialog
7. Click on the software button in the preferences dialog
8. In the **PSIPRED Installation Directory** section, type the path that you have created for PSIPRED
9. In the PSIPREDDDB section, input the path for a BLAST configured database that PSIPRED will use

Note: PSIPRED also requires PSI-BLAST and Impala software from the NCBI toolkit to be installed and in your PATH to function properly. PSIPRED calls PSI-BLAST by its pre-BLAST+ call blastpgp. If you installed the legacy BLAST mentioned in the previous section, copy blastpgp, impala, and makemat from the bin and place it in the folder PSIPRED calls it from. The default is /usr/local/bin.

Further Note: If you receive an error claiming that PSIPRED is not configured correctly due to not having weight.dat4, you must download a legacy version of PRED26 and install it in the same way you did the newest version.

1.2.3 MAFFT

ClustalW is the default sequence alignment tool and is packaged with MultiSeq. However, MAFFT can be used for doing sequence alignment if it is installed on your computer system.

To install MAFFT:

1. Go to <http://mafft.cbrc.jp/alignment/software/>
2. Choose the appropriate OS and download the zip file
3. Unzip and follow the installer instructions
4. Using the MAFFT webpage, determine the installation path(i.e. on Mac the path is /usr/local/bin/)
5. From the MultiSeq program window, choose File → Preferences to bring up the preferences dialog.
6. Click on the Software button in the upper left portion of the dialog to show the software preferences.

7. In the **MAFFT Installation Directory** dialog, copy down the installation path of MAFFT

If you use MAFFT for sequence alignment, note 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 chosen as the desired alignment program.
- 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.

To see an example of what the software preferences dialog should look like, see Fig. 3.

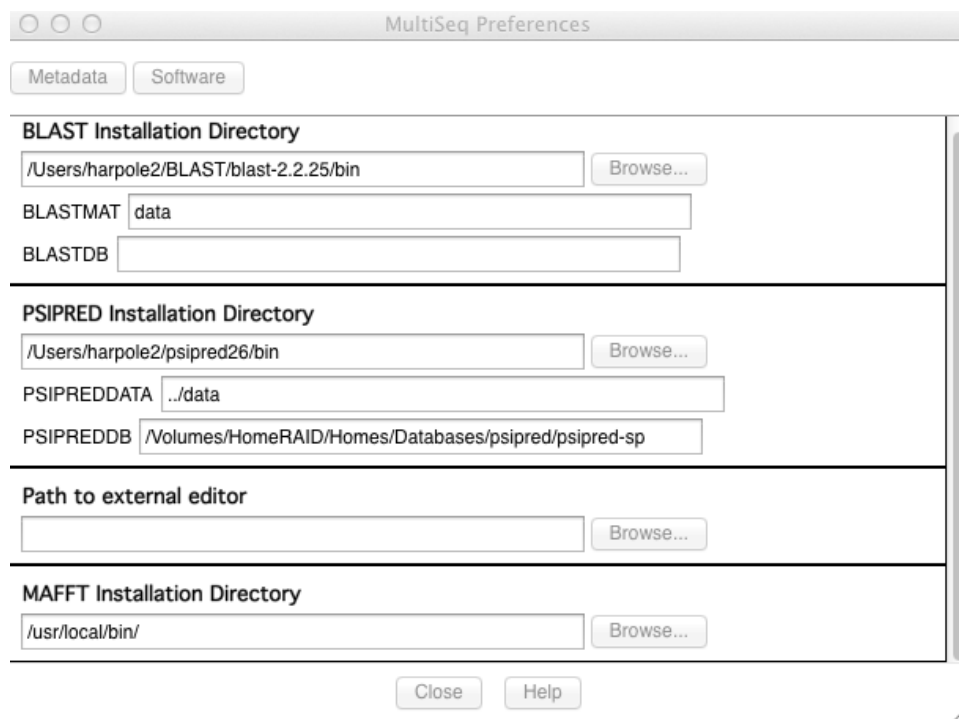


Figure 3: The preferences dialog should look like this when all programs are installed

2 Using and Managing Data

To begin analyzing proteins in MultiSeq, data from sequence³ and structure⁴ files is required. Data can be imported both locally and via a network connection. To import data, go to File → Import data

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.

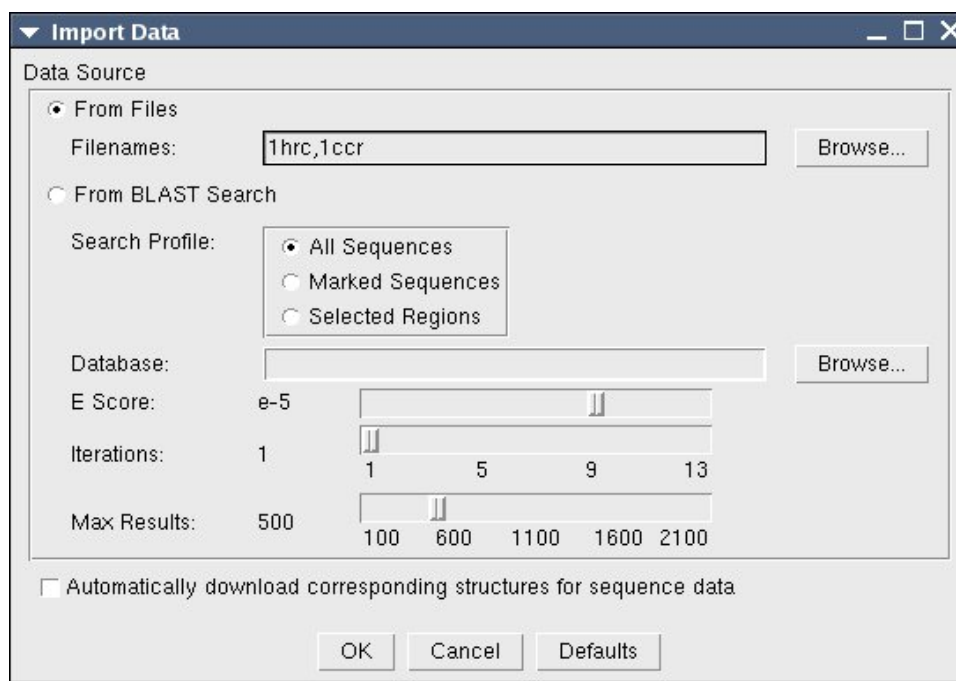


Figure 4: Import Data Window

³FASTA files.

⁴The ASTRAL database (<http://astral.stanford.edu>) is a compendium of protein domain structures derived from the PDB database. It divides each protein structure into its domain components. For example, AspRS is divided into three separate PDB files: one containing the catalytic domain, one with the insertion domain, and one for the anticodon binding domain. The names of the files contain the PDB extension, the letter a for ASTRAL, and a number, which corresponds to which domain it is in the original PDB file. 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.

2.1 Importing from files

Structure⁵ 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** radio 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 the PDB or SCOP id. This option requires a network connection for your computer to obtain files from PDB or ASTRAL directly.
3. Click the **OK** button.

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

2.2 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 and configure BLAST if you haven't already done so (see 1.2.1).

1. Before you open the **Import data** window, you have the option of either selecting a set of sequences by ctrl/shift clicking sequence titles, or a region within a sequence by clicking within the sequences themselves.
2. Go to **File** and then **Import data** and select **From BLAST Search**, and either **All Sequences**, **Marked Sequences**, or **Selected Regions** (see 3.6).
3. In the **Databases** dialog, 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. Click the **OK** button.

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**

⁵See VMD Manual for supported formats

BLAST Search Results

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

Apply View

Accept Discard

Figure 5: BLAST Search Results

Results window is divided into three main parts: the sequence viewer, Filter Options, and View Options (see Fig. 5).

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.

3 Working in the Environment

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

3.1 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. MultiSeq allows you to change the displayed title for each sequence by left clicking on the titles header and choosing a different option. This can be seen in Figure 6. If you

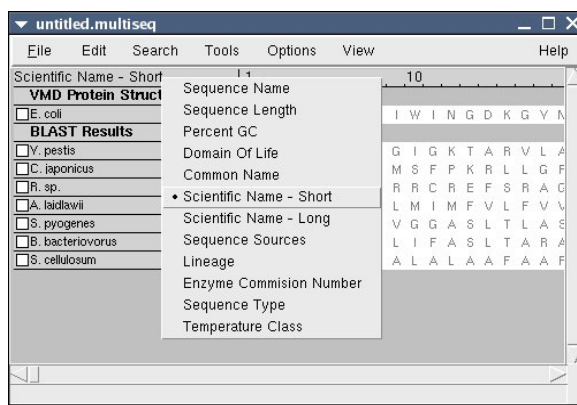


Figure 6: Choosing Data To Display As Sequence Title

choose an option where a sequence does not have a value, MultiSeq will show you the <Sequence Name> in angle brackets.

3.2 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. Right clicking on a group name (such as VMD Protein Structures) will bring up a context menu where you can manage and create new groups.

3.2.1 Moving Sequences into a group

When you want to move a sequence or set of sequences into a different or newly created group, you begin by highlighting the entire sequence by left clicking on the title of that sequence (multiple sequences can be selected using shift-click or ctrl/command-click just like in a file explorer). You then left click on any highlighted sequence and drag your mouse until a newly created black bar is directly below the group you want to place the sequences in. (Note: this technique also works for moving sequences within a group)

3.3 Visualization Menu

Whenever you load a sequence or structure into MultiSeq an ‘v’ box will appear next to the molecule’s ID. If you left-click this box, you can change the representation in VMD OpenGL Display in multiple ways.

Show Molecule sets the molecule to be shown in OpenGL

Hide Molecule hides the molecules in OpenGL

Show Chain sets the active chain representation to be shown in OpenGL

Hide Chain hides the active chain representation in OpenGL

Change Representation changes the OpenGL representation to one of the following: Bonds, VDW, CPK, Lines, Licorice, Trace, New Ribbon, New Cartoon

3.4 Representation Menu

Whenever you load a sequence or structure into MultiSeq an ‘r’ box will appear next to the molecule’s ID. If you left-click this box, you can change the representation in Multiseq in multiple ways

Duplicate simply duplicates the molecule of interest

Sequence displays the Nucleotide/AA sequence

Bar displays the sequence as a length bar with no Nucleotide/AA sequence

Secondary Alignment shows the secondary structure in the Multiseq window

3.5 Info Viewer

Whenever you load a sequence or structure into MultiSeq an ‘i’ box will appear next to the molecule’s ID. If you click on this box, a new window will appear called the Info Viewer (See Fig. 7). Within this window information regarding the taxonomy, common name, sequence type, EC number, and data source of the molecule will appear. If you have PSIPred installed and configured, you can predict the secondary structure at the bottom of the Info window.

3.6 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.

Edit Sequence Information

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

Predict

```
CCCHHHHHHHHHHHHTTTTBCTTTT
CTTTTBTTTTTEETTTTTTCCCHH
HHHCCEECHHHHHHHHHCHHHH
TTTTCCCCCCCCCHHHHHHHHHH
```

OK Cancel

Figure 7: Edit Sequence Information

4 File Menu

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.

4.1 New Session

This section fully clears all sequences in multiseq along with those linked to the VMD OpenGL Display. Go to File → New Session and a warning will pop up as there is no undo button to recover your work.

4.2 Load Session

Unlike **Import Data** (also in the **File**), **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** → **Load Session**. A file browser will appear allowing you to select a file with the extension `.multiseq` and make sure it has a corresponding directory of the same name.

4.3 Save Session

You can save a session of MultiSeq, with all of the files, alignments, and visual representations, by simply going to the **File** → **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` extension 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.

4.4 Export Data

This function allows you to export specific data from your Multiseq session.

Filename: is the name and path of the file you want to save

Data Source: allows you to decide what sequences to select

Data Type: this option allows you to save specific information in specific formats such as FASTA, ALN, NEX, PHY, PIR, and PDB.

4.5 Save Screenshot

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

4.6 Preferences

See 1.2

4.7 Choose Working Directory

This function allows the user to change the directory that temporary data is written to. It is important to make sure that the user has write permission in the directory that is chosen.

4.8 Cleanup Representation

This function allows you to refresh the multiseq window viewer if you believe that the GUI is not properly displaying your information. This may occur when loading many sequences and doing large alignments or generally dealing with large amounts of data in multiseq.

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.

5.1 Enable Editing

This function allows you to edit sequences manually. 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.

5.2 Remove Gaps

This feature will remove all gaps in an alignment.

5.3 Edit In Text Editor

Allows you to truly edit the sequences manually. 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

6.1 Find, Find Next, Find Previous

In the Search menu, you can find and highlight select residues or residue sequences. When you use Find, all of the residues or sequences will be highlighted, and you can then cycle through them by using Find Next and Find Previous. To input a sequence of residues, type each residue in the Residues to find box with no spaces.

6.2 Select Contact Shells

See Figure 8.

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

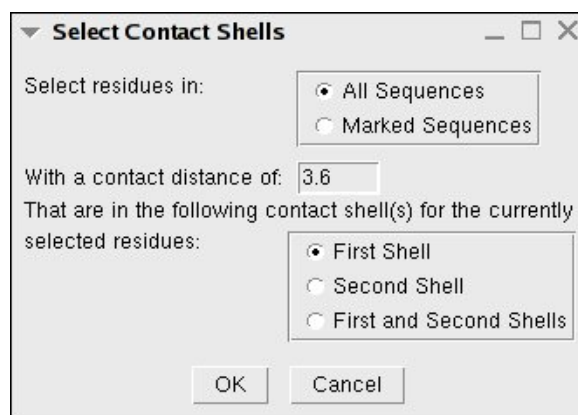


Figure 8: Select Contact Shell Window

With a contact distance of: defaults to 3.6 angstroms.

That are in the following contact shell(s) for the currently selected residues:

Choose from First, Second, or First and Second shells where first shells are defined as the the contacts within the contact distance of the selected sequences and the second shell is the shell within the contact distance of every residue in the first shell.

6.3 Select Non-Redundant Set

You can use structure QR or sequence QR to select a non-redundant set (See Fig. 9).⁶

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

Using Structure QR :

QH Cutoff: Scale from 0 to 1.(see 10.2)

Using Sequence QR :

PID: Is the Maximum Percent Identity required to return a sequence from the search

Percent of Set: Percentage of the set that can be returned such that the least reduntent are choosen first(note: PID and Percent of Set are mutually exculsive and only one can vary in a given search).

⁶For more information about QR, see: Anurag Sethi, Patrick O'Donoghue, and Zaida Luthey-Schulten. "Evolutionary profiles from the QR factorization of multiple sequence alignments" PNAS, 102(11):4045-4055. March, 2005.

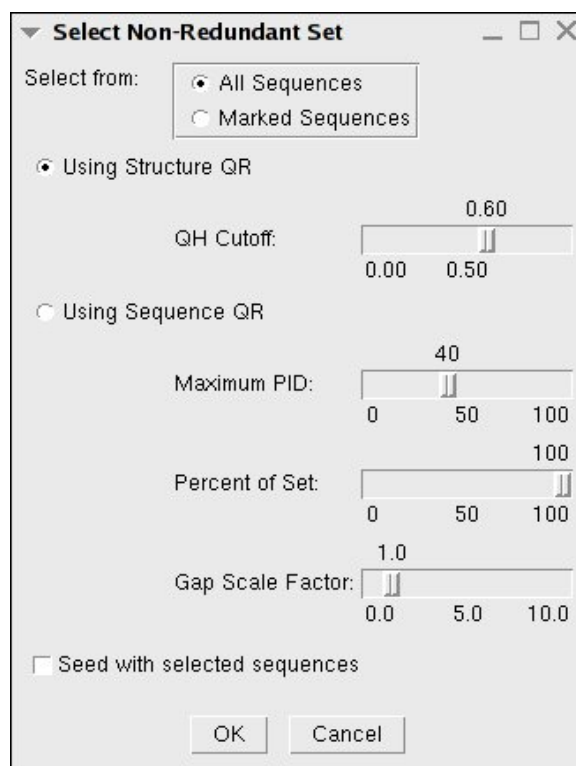


Figure 9: Select Non-Redundant Set Window

Gap Scale Factor: Scale from 0-10 where 0 gives no importance to gaps and 10 weights gaps as being very important for selecting the non-redundant set. DEFAULT = 1

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.

6.4 Select Residues

The Residue Selection feature (See Fig. 10) lets you analyze conservation using different measures, and highlight residues in the Sequence Display and VMD OpenGL 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 Q_{res} is. Where Sequence Identity is is a sequence identity measure, whereas Where Q_{res} is is a structure measure (For more on Q_{res} see 10.3).

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

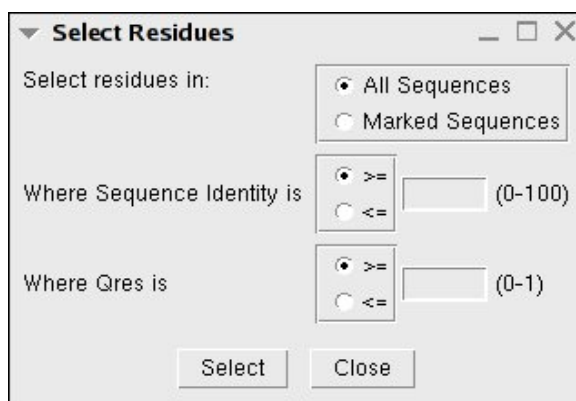


Figure 10: Select Residues Window

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 Q_{res} 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.

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 (<http://www.compbio.dundee.ac.uk/manuals/stamp.4.2/>).

Align the following: Choose which structures you wish to align

Number of passes (npass): Whether one or two fits are to be performed. 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 = 2

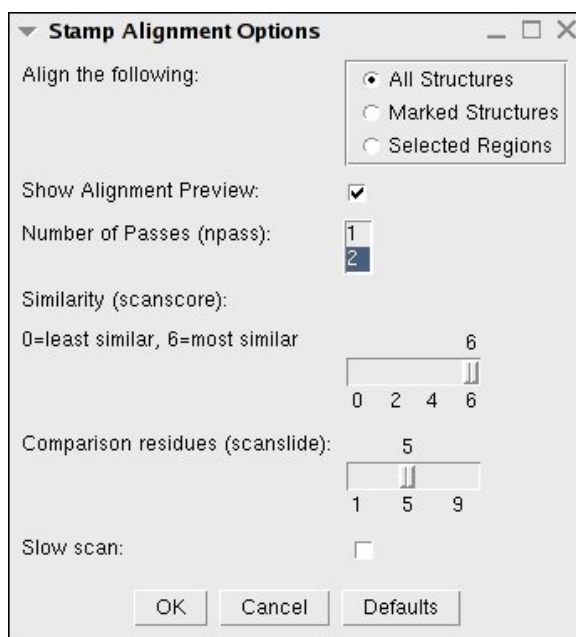


Figure 11: STAMP Structural Alignment Window

Similarity (scanscore): Specifies how the S_c value (STAMP algorithm) is to be calculated. This depends on the particular application. As a general rule of thumb, use $SCANSORE=6$ 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 $SCANSORE=1$ 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 $SCANSORE = 6$

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 database, 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 this box is checked, then the SLOW method of getting the initial fits for scanning will be used (see the manual for more information). 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[For installation information see 1.2.3]) (See Fig. 12).

Once you have decided which program to use, 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 relative to the group that you specify.

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

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, by default it uses a modification of Q that accounts for both gapped and aligned regions. This new metric, Q_H (see 10.2), 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. 13). Once you choose the sequences or regions you wish to create a tree for, you can choose which trees you want to create with the default being Q_H . Other options are RMSD (The root mean square deviation measures the distances in angstroms between the C^α atoms of 2 aligned residues), Percent Identity, or Sequence alignment (using either CLUSTALW or MAFFT). 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).

Once you have chosen which tree type 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. 14).

7.2.1 Tree Viewer

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 by clicking on the outgoing line from the subtree, you can change the shape of the tree by rerooting it at selected point. You can collapse/expand a subtree, as shown in Fig. 14 as well

Menu options include:

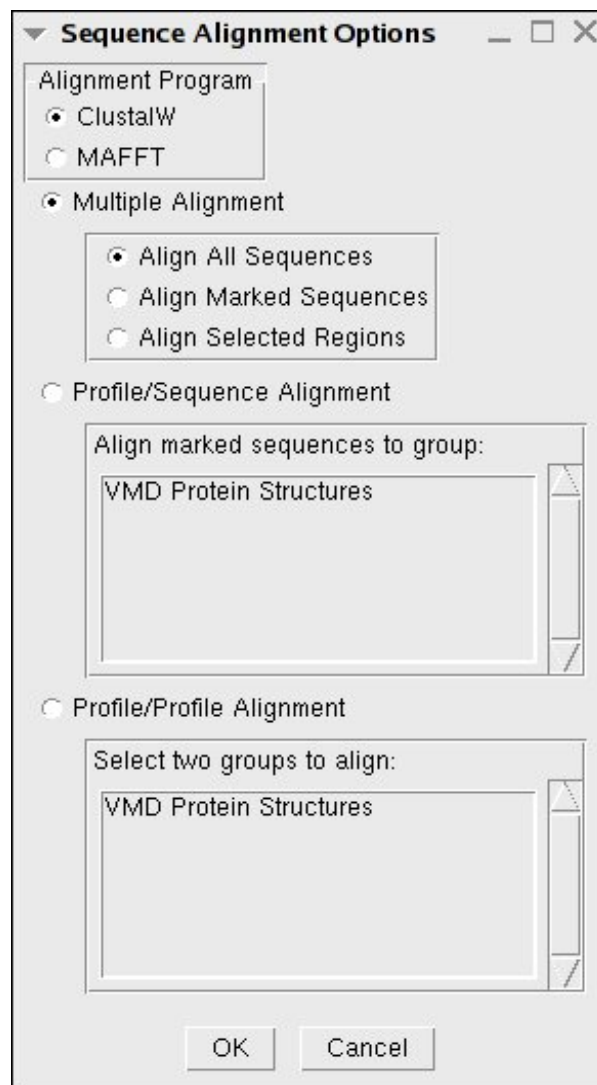


Figure 12: Sequence Alignment Menu Window

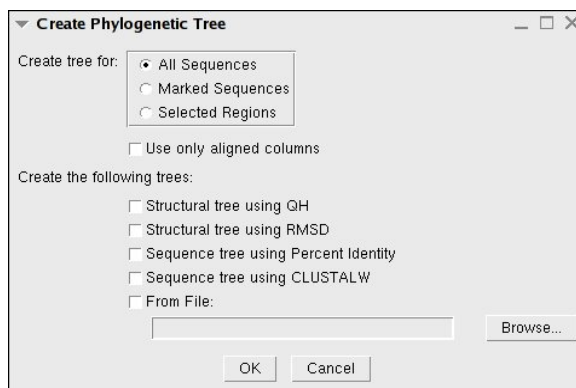


Figure 13: Create Phylogenetic Tree Window

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, change the scale (which pushes tree leaves left or right for viewability). Reverse 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 using **Leaf Color** as well as the tree backgrounds using **Background Color** by a variety of different metrics.

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

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. Also, you can always select the Q_H structure tree no matter what tree type you made to start with.

7.3 Plot Data

Plot Data creates 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 Q_{res} , **RMSD**, **Sequence Conservation**, or **Sequence Identity**) for each residue that you want to display. You can also plot custom data. The data graph will then be displayed (see Fig. 15). If you wish, you can create a postscript file for publication.

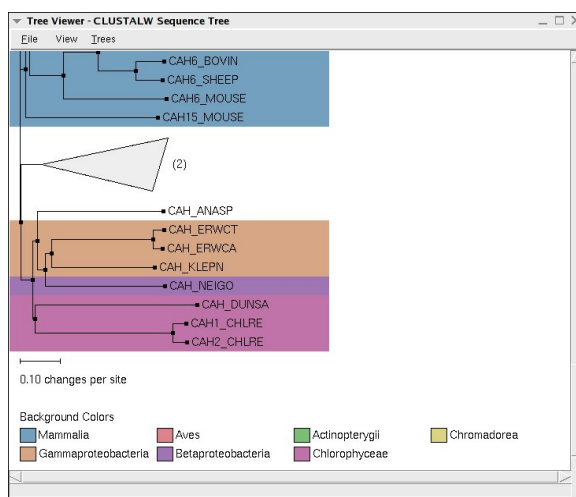


Figure 14: Phylogenetic Tree Viewer - CLUSTALW Sequence Tree

8 Options Menu

8.1 Atom Picking

Normally disabled, but can be turned on. This item allows you to select an atom in VMD OpenGL Display and in the VMD command line, molecular information such as the molecule id, index, rename, chain location, and coordinates appears.

8.2 Grouping

MultiSeq can automatically create groups and show the sequences in the MultiSeq window sorted accordingly.

From Selection Groups all marked or highlighted sequences into one group and prompts you to name the new group

Molecule Type Groups sequences based on macromolecule (Protein, Nucleic Acid)

Taxonomy Groups either all or marked sequences based on any classification desired (i.e. domain, kingdom, species)

Custom Allows the user to move groups up and down relative to each other as well as adding or deleting groups

9 View Menu

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

Mutual Information takes 2 or more align groups and colors them based on mutually shared identities

Q_{res} aka Q per residue colors the sequence based on the contribution from each residue to the overall average Q score. For more information see 10.3.

Residue Type Colors amino acid residues by whether they are acidic or basic.

RMSD takes aligned sequences and colors based on the distanced between two C^α of aligned residues

Sequence Conservation colors residues based on their conservation with all other sequences in Multiseq

Sequence Entropy colors columns of sequences based on the conservation of sequence

Sequence Identity method colors each amino acid according to the degree of conservation within the alignment: blue means highly conserved, whereas red means very low or no conservation in the default RWB color setting.

Sequence Similarity colors aligned sequences based on BLOSUM or a custom substitution matrix

Signatures colors sequences based on sequence signatures. These options include: Minimum fraction conserved to be a group signature:, Maximum fraction of signatures allowed in other groups:, Maximum fraction of gaps allowed in other groups:, Maximum distance of signature from a conserved block:, Minimum length of conserved block:

Custom is a way to import custom tcl scripts with specific coloring commands

Import allows you to import color dat files that had been previously exported using Export Data

Refresh Color is used if you believe that your newest color change was not implemented

9.3 Highlight Style

Highlight Style is an option for the OpenGL display. The style refers to drawing method in VMD⁸. 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.

⁸For more information about drawing methods, please refer to the VMD manual.

9.4 Highlight Color

Highlight Color is another option for the OpenGL display. Alongside Highlight Style, Highlight Color is the color or coloring method used in the OpenGL display when highlighting residues in the Sequence Display⁹. The default Highlight Color is yellow.

9.5 Color Scale

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

9.6 Zoom Window

(See Fig. 16) 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.



Figure 16: Zoom Window

10 Appendices

10.1 Appendix A: Q

In MultiSeq, Q has been generalized to measure the fraction of similar contact distances between all the aligned residues in two homologous proteins or RNA molecules. This term computes the fraction of C^α - C^α (or P - P) pair distances that are the same or similar between two aligned structures. “Evaluating protein structure-prediction schemes using energy landscape theory” by Eastwood,

⁹For more information about coloring methods, please refer to the VMD manual.

M.P., C. Hardin, Z. Luthey-Schulten, and P.G. Wolynes in IBM J . Res. Dev. 45: 475-497. 2001.

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

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

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

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

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

10.2 Appendix B: Q_H

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

Homology Measure

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

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

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

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

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

The first term, q_{aln} , computes the unnormalized fraction of C^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:

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

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

where N_{aln} is the number of aligned residues. N_{gr} is the number of residues appearing in gaps, and n_{gaps} is sum of the number of insertions in protein “a”, the number of insertions in protein “b” and the number of simultaneous insertions (referred to as bulges or c-gaps). n_{cgaps} is the number of c-gaps. Gap-to-gap contacts and intra-gap contacts do not enter into the computation, and terminal gaps are also ignored. 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 $Q_{aln} = \aleph_{qaln}$, which is identical to the Q-measure described into the Q measure described before.

10.3 Appendix B: Q_{res} Structural Similarity per Residue

Here we define another metric, called Q_{res} , that is derived from Q which is used to measure the structural conservation of the environment of each residue in the alignment. Q_{res} is a measure of the similarity of the C^α - C^α distances between a particular residue and all other aligned residues, excluding nearest neighbors, in a set of aligned proteins. The result is a value between 0 and 1 that describes the similarity of the structural environment of a residue in a particular protein to the environment of that same residue in all other proteins in the set. Lower scores represent low similarity and higher scores high similarity. If the set of proteins represents an evolutionarily balanced set, then structural similarity corresponds to structural conservation. Formally, Q_{res} is defined as follows:

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

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

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

and the normalization is given by

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

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

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