Online Sources of Sequence Alignments

This is by no means an exhaustive list, but includes several sources of protein multiple sequence alignments for use in Chimera. Why you might want such alignments:

- you have a protein structure and want to show which parts are more or less conserved in its family or superfamily
- you have a protein structure and want to find other structures in the same family or superfamily and compare them
- you are interested in a particular family or superfamily and want to look up a sample of structures and compare them

I would then work with the sequences and structure(s) using Chimera, for example to: evaluate conservation, superimpose structures, evaluate conformational variability, morph between different structures. Be aware that different sources have different definitions of ***family*** and ***superfamily***, but everyone agrees that a family is more closely related, a superfamily more diverse.

The list is organized as follows:

Descriptions here are minimal. I recommend consulting literature references and/or other documentation provided on the sites to better understand their contents and methodology.

Points to consider:

- How diverse a set of sequences is appropriate for your purposes? Must the proteins have identical functions, or are similar structures and related functions sufficient?

- How many related structures are known? It is generally more convenient for sequence-structure work to use sequence alignments of known structures (as in (A)) but that is only possible when several structures have been solved. If only a single structure (perhaps even a model) is available, you will have to use sequences without known structures to map conservation.

- One advantage of using precalculated alignments from a database, or an alignment calculated by a server such as **ConSurf**, is that in your own published work, you can cite the source rather than having to describe in detail and justify how you generated your own alignment.

← (A) Alignments containing proteins of known structure

These databases contain sequence alignments of proteins with experimentally determined 3D structures. Typically the names in the alignment are structure identifiers, which makes it easy to autofetch all the structures with a single step in Chimera (from the sequence alignment window, choose **Structure... Load Structures**). Of course, you can just fetch a subset of the structures individually with the **open**

command or **File... Fetch by ID**.

- **HOMSTRAD**
  http://tardis.nibio.go.jp/homstrad/
  - alignments are for families, generally narrow sets of proteins with high similarities
  - download *.pir or *.ali sequence alignment (either variant of aligned NBRF/PIR format is accepted by Chimera)
  - can also download a coordinate file in which each structure is a different chain; this is not as convenient as separate models in Chimera, so I prefer to fetch structures as described above, then superimpose them using the sequence alignment
  - can autofetch structures from the PDB since the sequences are named with PDB IDs

  Examples:
  - rubredoxin family and alignment file rub.pir
  - periplasmic binding protein -- sugar family and alignment file sugbp.pir

  Tip: **getting rid of extra chains**
  - Often structures include additional chains that are not associated with the sequence alignment and not needed for the intended analyses. These chains may be additional copies of the same protein or different macromolecules. Here is a trick for removing such unassociated chains in Chimera:
    1. With the mouse in the sequence alignment, draw a box that includes at least one associated residue from each structure. That will select the associated residues.
    2. Click into the main graphics window and press the keyboard up arrow key to promote the selection from residues to chains.
    3. Invert the selection to contain unwanted atoms: *Command*: **sel invert** or *Menu*: **Select... Invert (all models)**
    4. Delete the selection: *Command*: **del sel** or *Menu*: **Actions... Atoms/Bonds... delete**

- **PASS2**
  http://caps.ncbs.res.in/pass2/
  - structure-based alignments of SCOP superfamilies filtered at 40% sequence identity (SCOP release 1.75 from Jun 2009)
  - *.ali files (aligned NBRF/PIR format)
  - superimposed structures can also be downloaded
  - PASS2 sequence names can be converted to SCOP IDs (*e.g.* for Chimera fetch by ID) by changing any "-" (hyphen) to "_" (underscore)

  Examples:
  - WD40 repeat-like superfamily and alignment file 50978.ali
  - periplasmic binding protein-like I superfamily and alignment file 53822.ali

- **CATH**
  http://www.cathdb.info/
  - structure-based multiple alignments of close and distant structural clusters are available for some T and H groups in aligned FASTA format, with sequences named by domain ID, which includes PDB ID
  - choose Alignments tab, save FASTA format as plain text
  - "core" alignment has "domains in cluster" number of sequences, "expanded" has many more

  Examples:

    - porin (Topology)
    - crystallins (Homologous superfamily)

- **SISYPHUS**
  http://sisyphus.mrc-cpe.cam.ac.uk/sisyphus/
    - manually curated structure-based alignments for proteins with non-trivial relationships such as permutations
    - you can browse listings within the alignment categories: fragment, homologous, fold
    - aligned FASTA format
    - sequence names start with PDB IDs

← (B) Alignments that do not necessarily contain proteins of known structure

If the corresponding tree in New Hampshire (aka Newick) format is available, it can be loaded after the sequence alignment has been opened.

- **PFAM** (also tree files)
  http://pfam.sanger.ac.uk/
  http://pfam.janelia.org/
    - for Chimera purposes, I use "seed alignments" (many fewer sequences than the full PFAM alignments)
    - don't get their MSF format (it's apparently wrong), but Selex, Stockholm, FASTA, and the tree format are all fine
    - for some families, corresponding structures are listed; sometimes structures are not listed even though they exist
    - can link out to PFAM from structure pages at the RCSB PDB

  Example:
    - Peripla_BP_1 family seed alignment file PF00532_seed.slx and tree file PF00532_seed.nhx

  Tip: **manual sequence-structure association**
    - Often your structure will not be similar enough to any sequence in the alignment to associate automatically. Some things to try to associate the structure:
        - Use **Structure... Associations** in **Multalign Viewer** to compare all the sequences to the structure and associate it with the best match even though it does not meet the automatic association criteria. The resulting association must be examined, since it may not be good enough to be usable. For example, the forced association of structure **2gbp** with PF00532_seed.slx only associates the first few residues of the structure with the last few residues of a sequence and is obviously wrong.
        - Use **Edit... Add Sequence** in **Multalign Viewer** and add the sequence from the structure to the alignment. This may require several cycles of unsuccessfully adding the sequence, using **Edit... Delete Sequences/Gaps** to remove it, and adding it back again with different alignment parameters. For example, adding the sequence of **2gbp** to PF00532_seed.slx does not work with the default parameters, but is reasonably successful using the **BLOSUM-30** matrix.

- **PANDIT** (again seed alignments from PFAM, but different method used to build trees)
  http://www.ebi.ac.uk/goldman-srv/pandit/
  (Nov 2008: updates stopped, but database may still be useful)

  - see PFAM (above) for family identifier codes
  - no convenient file download; I had to cut-n-paste text from the screen to a file to save the alignment (apparently Selex format is shown) and tree (saving from the link includes a bunch of HTML tags)
  - this may be a version issue, but I couldn't mix-n-match the tree and alignment from PFAM and PANDIT; they were only consistent both from PFAM or both from PANDIT

- **SFLD** (**S**tructure-**F**unction **L**inkage **D**atabase) (family, subgroup, and superfamily alignments, some Chimera "active site" sessions)
  http://sfld.rbvi.ucsf.edu/
  - deep and detailed coverage but of relatively few enzyme superfamilies
  - lists some but not all structures
  - can set up browser (except Safari) to open structures, sequence alignments, sessions directly in Chimera
  - can get an alignment that includes your query sequence if it matches HMMs in the database

← (C) Server-generated multiple alignment from a single input

- **ConSurf** (also tree files)
  http://consurf.tau.ac.il/
  - you can get precalculated results for PDB entries or submit a structure to the server (note server calculations may take a while)
  - the server has many options; if you just use a single structure as input, it will find homologous sequences and create a sequence alignment and tree
  - results files available for download include the sequence alignment (Clustal *.aln format) and TheTree.txt (tree in Newick format)
  - can set up browser (except Safari) to show results directly in Chimera

- the **Blast Protein** tool in Chimera (under **Tools... Sequence** in the menu) uses a web service hosted by the UCSF RBVI. As with any BLAST search (see (D)), the results depend on the search criteria and the database, and may be unbalanced and/or contain significant redundancy.
  - the input can be:
    - a chain from a structure open in Chimera
    - a sequence pasted in as text
    - a sequence from an alignment open in Chimera's **Multalign Viewer**
  - several parameters can be adjusted, including number of iterations (multiple iterations = PSI-BLAST) and database (**pdb** or **nr**)
  - the output is a list of hits, from which all or a user-chosen subset can be retrieved as:
    - a ***pseudo-multiple*** sequence alignment, automatically shown in Chimera's **Multalign Viewer**. A pseudo-multiple alignment from BLAST is not a true multiple alignment, but a consolidation of the pairwise alignments of individual hits to the query, as specified by the BLAST alignment view option "flat query-anchored with letters for identities."
    - structures for hits from **pdb**, automatically superimposed according to the pseudo-multiple sequence alignment

← (D) DIY: Find sequences individually, use alignment server

Issues to consider are how diverse the set of sequences should be, alignment quality, and balance, *i.e.* an alignment could oversample some areas of the intended "sequence space" and undersample others. Imbalance can be reduced by filtering out sequences at some level of sequence identity, and in Chimera, using sequence-weighting options to calculate conservation.

I used the DIY approach to make the alignments in the Chimera "hormone-receptor complex" demo (under **Tools... Demos** in the menu) because I wanted to include sequences for the hormone and receptor from the same six species. The sequences were similar enough to align easily, so I didn't have to worry about tweaking parameters to improve the results.

Look up sequences (I usually save or text-edit the sequences into a single FASTA file):

- UniProt (text search)
  http://www.uniprot.org/
- Entrez Protein (text search)
  http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein
- NCBI BLAST, PSI-BLAST *etc*. (input one sequence to find other potentially related sequences)
  http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins&PROGRAM=blastp

Use a server to align them (order is merely alphabetical):

- ClustalW2
  http://www.ebi.ac.uk/Tools/clustalw2/
- FSA
  http://orangutan.math.berkeley.edu/fsa/
- Kalign
  http://www.ebi.ac.uk/kalign/
- MAFFT
  http://www.ebi.ac.uk/mafft/
  http://mafft.cbrc.jp/alignment/server/
- MUSCLE
  http://www.ebi.ac.uk/muscle/
- PROMALS3D (also takes structures)
  http://prodata.swmed.edu/promals3d
- T-Coffee
  http://www.ebi.ac.uk/t-coffee/
  http://www.tcoffee.org/
