HST.508 **Problem Set 3** **Due Dec 6**

1.Mammalian Serine Protease Family Portrait.
Consider three important mammalian proteases: chymotrypsin (4CHA), trypsin (2PTN), and elastase (3EST). These enzymes have the same biochemical function and different specificities. An attached PDB file contains their structures superimposed. It can be viewed in RasMol (http://www.openrasmol.org/) or any other protein visualization program.

a. Using the sequence for each protein identify its orthologs from various organisms. Build a multiple sequence alignment of the orthologs and paralogs. I.e.

```
chym_human    ABCDEFGH
chym_chimp    ABCDEFGG
chym_mouse    ABCDEEFG
tryp_human    ABCDEFG
tryp_mouse    ABCDEFG
elast_human   ……….
elast_rat     ……..
```

You can identify orthologous enzymes simply by running blast and selecting those that have the same functional annotation. You can also use databases of orthologous proteins such as INPARANOID or KOG. Use ClustalW (http://www.ebi.ac.uk/clustalw/), Tcoffee or any other appropriate program to build multiple sequence alignments.

b. Study the following residues that lie in the proteins' active site: HIS57, ASP102, SER195. Do these residues overlap in the structure? Are they conserved within each family of orthologs? Across the whole family? Can you find other residues that show similar pattern of conservation?

You can use information (entropy) as a measure of conservation:

$I = -\sum_{a=1..20} f(a)\log f(a)$ where f(a) is a frequency of amino acid ***a*** at a given position of the alignment. You can compute information for every position of the multiple sequence alignment. You can do it separately for each orthologous group, or for all proteins taken together (depending on the question you want to address).

c. Study proteins' specificity pocket. In this family of enzymes, specific recognition of the substrate is performed by a specificity pocket. The specificity pocket is formed by the following residues:

Chymotrypsin	Ser 189, Gly 216, Gly 226
Trypsin	Asp 189, Gly 216, Gly 226
Elastase	Ser 189, Val 216, Thr 226

Find them on the structure. Do they overlap in space? How conserved are they within and between the orthologous groups? Rationalize the conservation pattern of the specificity pocket residues. Do you see a shift in selective pressure (i.e. a residue conserved in one group is not conserved in the other) or a different pattern?

d. By visual analysis of the structure identify a few more residues in the specificity pocket and study their patterns of conservation. Rationalize your observations.

2.Optimal strategy for structural genomics.

The goal of the structural genomics project is to solve for the structures of all proteins. A lot of effort can be saved by solving structures of some selected proteins, while modeling structures of others by homology. The quality of a homology model, however, strongly depends on the similarity between the sequence of the modeled protein (the target) and the sequence of the protein with a known structure (the parent). High sequence similarity is required for a good model.

To minimize efforts and improve the quality of the models one may want to select the set of proteins whose structure needs to be solved, such that more proteins in the family are similar to the selected ones. Your goal is to explore the feasibility and efficiency of this approach.

a. Select a sufficiently large and diverse family in Pfam. Download the alignment and calculate pairwise similarity between every pair of sequences in the family. To compute similarity you can use a BLOSUM matrix or simply sequence identity between the proteins. Develop an algorithm to find the minimal set of proteins whose structures need to be solved, such that other proteins in the family can be accurately modeled. Assume that accurate modeling requires a sequence similarity to be above S_{cutoff}. Define coverage as the number of proteins that have similarity $S > S_{cutoff}$ with at least one protein in the selected set. Try some of the following approaches:
    i. Fixing S_{cutoff} find a minimal set of N_{min} proteins that covers 100% of the family.
    ii. Fixing S_{cutoff} and N, find a set that maximizes the coverage.

   Explore the role of S_{cutoff} and coverage. How many structures have to be solved to provide 80% coverage at a reasonable S_{cutoff}=70% in sequence identify? How much efforts in solving structures can be saved if we learn to model structures at S_{cutoff}=50% or 30% of sequence identity? How much efforts can be solved if we aim at modeling structures for 80% of the proteins in the family? 50% ? Plot of the coverage as a function of N and S_{cutoff}.

b. A naïve strategy is to pick proteins for structure solution at random. How many more structures N_{naive} need to be solved to achieve the same coverage (at a fixed S_{cutoff})?

3.Divergent evolution.

The SCOP database groups proteins into Folds, Super-families and Families.

Folds contain proteins that have the same structure; Super-families contain proteins that have the same structure while being distantly related by sequence; Families contain proteins that are similar in both sequence and structure. While proteins that belong to the same Super-families but different Families have no apparent sequence similarity, they may still be result of a divergent evolution. While having a common ancestor, these proteins have diverged in sequence. They diverged so significantly that their homology cannot be established by pairwise alignment. To establish divergent evolution one may try to find intermediate sequences that have apparent sequence homologous to two distant proteins.

a. Your goal is to find examples of divergent evolution using intermediate sequences. Pick two proteins from SCOP database (scop.berkeley.edu) such that they belong to the same Fold but different Super-families, or same Super-family and different Families (see examples below). Make sure BLAST cannot detect sequence similarity. Look for intermediate sequences (not necessarily of known structure). You may need more than one intermediate: e.g. A-X-Y-B, where A and B are your proteins and X and Y are two intermediates. Be careful not to be tricked by similarity in other domains of the same protein. Focus on individual domains of proteins, rather than on complete proteins (SCOP is built using domains). Report both successful chains of intermediates and failed attempts.

   Sequences of individual SCOP domains can be retrieved from http://astral.berkeley.edu/getseqs/. Two sequences can be aligned at http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi
   To find an intermediate you can run each sequence against a non-redundant database (nr database) of proteins at NCBI Blast website http://www.ncbi.nlm.nih.gov/blast/ and then compare lists of homologous sequences. You may also find ASTRAL database useful for this problem http://astral.berkeley.edu/scopseq-1.69.html

You can try the following cases
1. Immunoglobulin-like beta-sandwich fold: 1fnf and 3cd4
2. TIM beta/alpha-barrel fold: 5tim and 2ebn
3. OB-fold, Nucleic acid-binding proteins: 1mjc and 1cuk