Assignment 6: MIDTERM: Homology Contest! Exploring Sequence/Structure/Function Relationships (& Related Tools/Databases like SCOP, IMAGE, BLAST, NDB, PDB)

With the rapidly growing information on genomic sequences, *comparative modeling* — structure prediction based on sequence similarity — is becoming increasingly valuable. Indeed, structural and functional genomics, the three-dimensional (3D) structure and functional analysis of genomic products, are rising disciplines in bioinformatics. It has been reported, for example, that a sequence homology of larger than 40% usually implies more than 90% 3D-structure overlap (see below for precise definitions of similarity). Thus, with the growing amount of genomic information, we may eventually be able to predict reliably 3D structures of proteins. Since structural similarity is often preserved more strongly than sequence through evolution, reliable homology-based predictions might provide crucial functional properties of new gene products in the near future.

Through this assignment, you will gain some experience in quantifying and analyzing sequence and 3D structure similarity for proteins. You will also explore sequence and structure databases in search of interesting examples, and learn how to use important computational and database resources. You will have to be resourceful in looking for suitable programs for alignment and structure analysis besides those below; no simple recipes will be given here.

This assignment can be done by teams of two students; choose a partner with complementary skills. You will have to present your results to the class.

The 5 Tasks

Find and demonstrate the following four relationships for proteins:

1. [**EASY**] Two proteins with very *high sequence similarity* (but less than 95%) and very *high structural similarity*. Excluded from consideration are trivial examples, such as involving multiple PDB entries for the same protein.

2. [**EASY**] Two proteins with very *high sequence similarity* (but less than 95%) <u>and</u> *very high structural similarity* <u>but</u> markedly *different biological/functional* properties.

3. [**MODERATE**] Two proteins with *low sequence similarity* but *high structural similarity*. Also comment on the *functional* properties of the pair.

4. [**HARD**] Two proteins with very *high sequence similarity* but very *low structural similarity*. Also comment on the *functional* properties in your example.

For problems 3 and 4 above, the class contest will be won by the students that find the most extreme examples (i.e., the maximal sequence similarity / minimal structural similarity, minimal sequence similarity / maximal structural similarity).

5. [**EASY WARMUP**] Search and identify all the determined structures in the PDB/NDB that contain the nucleic acid sequence TATAAAAG. Discuss these structures and their significance.

For each task, generate color molecular views, report the analyses in detail, and include a description of how you found the example. Also discuss your similarity/dissimilarity criteria (see below), and prepare a class presentation on your results.

Ground Rules

1. Homology, or sequence similarity, will be defined by the percentage of sequence identity.
2. 3D-structure similarity will be defined in two ways:
   (a) the percentage of C^{α} atoms of the proteins that "overlap", i.e., are within 3.5 Å of each other in a rigid-body alignment of the protein;
   (b) the root-mean-square-deviation (RMSD) between C^{α} atoms of the proteins in a rigid-body alignment of the protein. (Recall your experience with RMSD measurements in the previous assignment).

   You should first experiment with overlapping several protein structures to determine what RMSD values and/or percentages of C^{α} overlap indicate random similarity. *Discuss this in your submission.*

Tools of the Trade

1. **Sequence and Structure Databases**. You have already navigated through the structural PDB and NDB databases and various sequence databases. Continue to work with these and the RCSB facilities.
2. **SCOP**. This site for the *Structural Classification of Proteins* (scop.mrc-lmb.cam.ac.uk/scop/ categorizes proteins according to the levels (top-to-bottom) of: class, fold, superfamily, family, domain, and reference PDB structure.
3. **Insight II**. Continue to use Insight II for structure display and analysis.
4. **NCBI Tools like BLAST and Its Cousins**. BLAST is a library of heuristic similarity search programs (Basic Local Alignment Search Tools) that explore relationships involving protein and nucleic-acid sequences and 3D structures. This library contains blastp, blastn, blastx, tblastn, tblastx,

and others, developed at the National Center for Biotechnology Information at the National Library of Medicine of the National Institutes of Health. Get started at their web site www.ncbi.nlm.nih.gov/BLAST/. This page leads to the BLAST suites as well as contains usage information. See, for example, Overview, Manual, BLAST FAQs, References.

BLAST, one of the most popular tools among molecular biology researchers, has evolved rapidly since its inauguration in 1990. BLAST searches a database in two stages, finding small sequence lengths that match the target exactly and then attempting to extend the length of the match from this subset of sequences in the database. Not only are the alignment algorithms improving continuously (e.g., allowing alignments of DNA or protein sequences with insertions or deletions in Gapped BLAST; forming families of aligned sequences and quick profiles of them in Position-Specific Iterated (PSI)-BLAST; or incorporating biological-function hypotheses into sequence queries to restrict the analysis to subset of protein sequences as in Pattern-Hit Initiated (PHI)-BLAST), but performance has been greatly accelerated. Algorithmic features include dynamic programming tools, hidden Markov models, and various optimization strategies.

To align two protein or nucleotide sequences, go to the link of BLAST 2 sequences (www.ncbi.nlm.nih.gov/gorf/bl2.html) and set up the computation according to the instructions. Take care to choose the options of the computation with care, and explore different options. The server will send the results to the web browser being used.

Some available programs are:

- blastp: compares an amino acid query sequence against a protein sequence database.
- blastn: compares a nucleotide query sequence against a nucleotide sequence database.
- blastx: compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
- tblastn: compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).
- tblastx: compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

See www.ncbi.nlm.nih.gov/BLAST/newblast.html#introduction for further information.

Other similarity programs are available (such as MEME *and* MAST *from SDSC); use anything appropriate for the task.*

5. **An Image Library**. The *Image Library of Biological Macromolecules* organized by the Institute for Molecular Biotechnology in Jena, Germany (www.imb-jena.de/IMAGE.html) offers a colorful library of biomolecular images corresponding to structures available in databases like the NDB and PDB. Besides detailed colorful illustrations of the structure in a variety

of styles, relevant structural information and publication links are available. Basic tutorials on structural biology are under preparation at this site.

HINTS For the Assignment

1. Scan the literature for related papers on comparative or homology modeling but do not repeat known examples. You **CAN** be original with some work.
2. Large changes in 3D structure despite high sequence similarity can result from the following situations:
   - mutations in critical regions of the proteins such as active sites
   - mutations in ligand binding sites (as in immunoglobulins)
   - mutations in regions that connect two secondary-structural elements (as in helix-loop-helix motifs)
   - structure determination of the same system at different environmental conditions (e.g., different solvent, different crystal packing forms for mutant proteins)
   - proteins containing the same subunits but a different number of subunits, with a structure/fold/topology that depends critically on that number.

   Search PDB and SCOP for examples in this spirit.
3. Look for groups of proteins in the same family, or for proteins sharing the same fold in the SCOP site. The structural classification information should generate ideas.
4. General structure alignment via Insight is not very sophisticated and may be entirely unsuitable for sequences of disparate lengths and for structures with two similar subdomains adopting a different relative orientation. Search for suitable programs for these cases (e.g., from the RCSB, www.rcsb.org and from SDSC) and also write/use your own programs to perform certain analyses, such as structure similarity measurements upon alignment (e.g., criterion 2a under **Ground Rules**).

Background Reading

- D. Baker and A. Sali, "Protein Structure Prediction and Structural Genomics" *Science* **294**, 93–96 (2001). [**From Coursepack**].
- B. Honig and A. Nicholls, "Classical Electrostatics in Biology and Chemistry", *Science* **268**, 1144–1149 (1995) [**From Coursepack**].
- D. Case, "NMR Refinement", in P. von Ragué Schleyer (Editor-in Chief), N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, III, editors, *Encyclopedia of Computational Chemistry*, volume 3, pages 1866–1876. John Wiley & Sons, West Sussex, UK, 1998.