Research Proposal

Title: Multiple Sequence Alignment used to investigate the co-evolving positions in OxyR Protein family.

Name: Minjal Pancholi
Howard University
Washington, DC.
June 19, 2009

Research Mentor: Dr. Hugh Nicholas, Biomedical Initiative, Pittsburgh Supercomputing Center

Introduction:

Sequence alignment of DNA, RNA, and protein can be achieved by multiple sequence alignment techniques. Through speciation event, different species can acquire genes that code for the protein which share a similar ancestral origin. By comparing and analyzing the homologs of these species, we can get a great deal of information about the target protein sequences. Highly conserved proteins are often required for basic cellular function, stability or reproduction. Conservation of protein sequences is indicated by the presence of identical amino acid residues at analogous parts of proteins. These conserved residues can be analyzed through the softwares used for the multiple sequence alignment. Moreover, the time it takes for the species to diverge can be calculated through probability theories and hence, the probability for mutation to occur in certain protein sequence can be calculated.

Coevolving positions are more likely to change protein function when mutated than are positions showing little coevolution (Gloor, 2005). Coevolving positions fall into two general categories. One set is composed of positions that coevolve with only one or two other positions. These positions often display direct amino acid side-chain interactions with their coevolving partner. The other set comprises positions that coevolve with many others and are frequently located in regions critical for protein function, such as active sites and surfaces involved in intermolecular interactions and recognition (Gloor, 2005). Thus, it may be as important to understand the coevolving positions as the conserved positions.

Some of the potential instruments that we might use for the multiple sequence alignment are ClustalW, T-Coffee, and ProbCons programs. The ultimate goal is to find a comparatively high quality alignment of the target protein, by using different programs, and by incorporating additional experimental or computational information during an editing step. The conserved and coevolving residues can then be highlighted in the three dimensional structure, provided an appropriate structure is available.

Thorough analysis of the protein sequence can, then, provide the information related to its divergence in its function through evolutionary time and hence, an important source of information for biomedical/ biochemical studies of the target protein.

Materials and Methods:

The tetrameric OxyR protein is a member of the LysR family of transcription activators and exists in two forms, reduced and oxidized; the latter is the only form able to activate transcription (Scandalios, 2002). Further studies led to the identification of a number of OxyR-activated genes. Protein sequences for OxyR are obtained from the iProClass protein classification database at http://pir.georgetown.edu/iproclass/. ClustalW, T-Coffee, and ProbCons can then be used to gain the initial alignment of the protein sequences. ProbCons is a progressive multiple alignment tool based on the technique of probabilistic consistency scoring for multiple sequence alignment.

T-COFFEE (Tree-based Consistency Objective Function for alignment Evaluation) algorithm is considered a leading multiple alignment system, since its introduction in 2000, generates multiple alignments using a library of alignment information which it has generated from both local and global pair-wise alignments. T-COFFEE also incorporates a progressive strategy optimization method which considers alignments between all sequence pairs, whether or not they have already been aligned, in each step of the alignment process.

ClustalW is the quickest and one of the most popular methods, using a hierarchical method of alignment, or progressive algorithms. This program finds and aligns sequences that are most similar, followed by the next most similar to build a tree (Baxevanis et. al. 2001). T-Coffee and ProbCons take longer to run but generally provide superior alignments.

MEME is another algorithm based on sequence alignment which will be used for additional alignments of the protein sequences. MEME (Multiple EM for Motif Elicitation) sorts through sequences and finds / reports similar or conserved subsequences regardless of their placement along the protein sequence. It serves the local alignment rather than global alignment, and provides the most accurate results among all the other programs. The conserved patterns or motifs identified by the MEME program are used to edit the alignment and improve it.

The results from the different sequence alignments will then be compared with each other and viewed side-by-side in a program called GeneDoc. The best fit or the aligned sequences are selected the way that the sequences of protein remain intact.

Phylogenetic analyses of protein families are used to define the evolutionary relationships between homologous proteins (Palidwar, 2006). The results obtained from the different algorithms can be compared with the statistical data of the phylogenetic tree to establish the evolutionary relationship between different species and to identify groups of paralogous sequences; sequences that have a gene duplication event in their common evolutionary history.

Expected Results:

If an appropriate three dimensional structure is available information from the final edited high quality alignment and the phylogenetic identification of paralogous groups of the protein OxyR can be mapped to this structure. Their biochemical function and structural information can be understood and compared with that of the other paralogous groups of sequences. The conserved regions, the likely important differences among different groups of paralogous sequences, and the coevolving positions within each orthologous set of sequences can be highlighted on the three dimensional protein structure. Homologous sequences share a common ancestor. Orthologous sequences tend to have similar physiological role while paralogous sequences may perform different physiological role (Nicholas, et.al. 2002). Through multiple sequence alignment and subsequent phylogenetic and statistical analysis we can identify orthologous and paralogous sequences of the protein and their evolutionary constraints. The application of information theory to identify coevolutionary relationships among positions in proteins will become an increasingly powerful and important bioinformatic approach (Gloor, 2005). Information about coevolving position can be as important as the conserved positions to establish the relationship with the structure and structural divergence of protein.

References:

Gloor, Gregory, Et. al. “Mutual information in Protein Multiple Sequence Alignment Reveals Two classes of Coevolvig positions.” *Biochemistry* 2005. 44, 7156-7165.

Notredame, Cedric. Et.al. “M-Coffee: combining multiple sequence alignment methods with T-Coffee.” Nucleic Acid Research 2006. Vol. 34, 1692-1699.

Nicholas, Hugh B., Ropelewski, Alexander J., Deerfield, David W. “Strategies for Multiple Sequence Alignment.” *BioTechniques*. 32 (2002): 572-591.

Baxevanis, Andreas D. and Ouellette, B. F. Francis. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. New York: *Wiley-Interscience*, 2001.

Do, Chuong B., Mahabhashyam, Mahathi S.P., Brudno, Michael, and Batzoglou, Serafim. “ProbCons: Probabilistic consistency-based multiple sequence alignment.” *Genome Research*. 15 (2005): 330-340.

Palidwor, Gareth. Et. al.“Taxonomic clolouring of phylogenetic trees of protein sequences.” Bioinformatics. 2006; 7: 79.

Scandalios, John. Et.al. “Oxidative Stress Responses – what have genome-scale studies taught Us.” *Genome Biology* 2002.