A Bioinformatics Approach to the Analysis of the Glycogen Phosphorylase Protein Family

Jieming Shen, Rutgers, The State University of New Jersey
Mentor: Dr. Hugh Nicholas, Biomedical Initiative, Pittsburgh Supercomputing Center
June 16, 2006

Introduction

Alignments of multiple sequences of proteins are used to determine the evolutionary relationships between protein homologues and to provide clues as to the function of a protein family as a whole. The goal of our project is twofold. We hope to conduct a validation study of a recently developed method for multiple sequence alignment called ProbCons. ProbCons is a progressive multiple sequence alignment tool based on the technique of probabilistic consistency scoring for multiple sequence alignment (1). In order to assess the accuracy of this new method, multiple sequence alignments of the glycogen phosphorylase protein family will be conducted using the ProbCons algorithm as well as a combination of more established methods for alignment. The individually obtained outputs from each of the programs we use will be compared to one another to evaluate the accuracy of the ProbCons alignment.

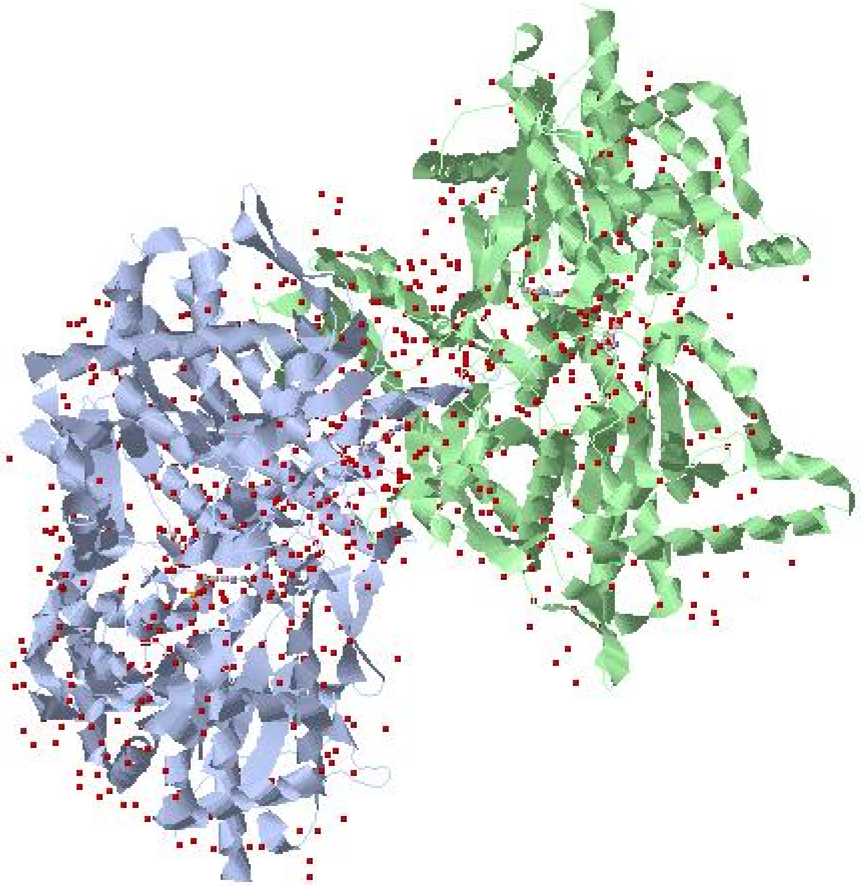
Central to analysis is the idea that two species will acquire differentiating mutations proportional to their divergence time. By incorporating probability data to account for the chance that such a sequence mutation will occur, we can determine the amount of time that has elapsed since two species diverged. Applying this strategy to the analysis of multiple sequences, we can gain insight into the way that related species evolved from their common ancestor and construct a phylogenetic tree for the homologous sequences of a protein family. Multiple sequence alignment is also used to identify the residues that are conserved or variable among the homologues. Features that are most often conserved represent areas essential to protein functionality and are probably responsible for the molecule's biochemical activity. Examination of the conserved regions in a three-dimensional model of the protein also provides insight into the types of reactions involving that specific protein. Another application of the data gathered by multiple sequence alignments involves using the information to build a profile for a protein family from which new or distant members can be identified and classified (5).

Creating multiple sequence alignments can be challenging because the process is both biologically and computationally complex. Since the aligning of sequences represents the first step in sorting through sequence data, it is crucial to ensure that error is minimized by maintaining high accuracy of the alignments. For biologists, the challenge lies in determining the alignment program or combination of alignment programs to use in order to ensure that the most biologically correct alignment is produced for a specific case (3).

In order to assess the usefulness of the ProbCons method, we will independently apply this method and various other alignment algorithms to the glycogen phosphorylase protein family, which will serve as the testing grounds for our evaluation. We chose the enzyme glycogen phosphorylase for study because it is one of the proteins expressed differentially in Drosophila embryos during the development of the ventral furrow. For our secondary goal of the examination of the glycogen phosphorylase protein family, we would like to understand how the expression of different protein features is associated with the overall physiological event as well as how the protein isozymes in are related different species.

The pathway to the glycogen phosphorylase activation is triggered by hormone activity. The enzyme catalyzes the first step of glycogen breakdown, eventually leading to the release of glucose subunits which can be used by living organisms as a source of energy (2). Because of its involvement in the basic metabolic process of glycogenolysis, glycogen phosphorylase is found in a large number of organisms. To varying extents, the functions many of the homologues in the protein family have already been studied, and in many species, the three-dimensional structure for the protein has already been solved. This makes glycogen phosphorylase an ideal object of experimentation on which to test and apply our methods for protein family computational analysis.

Figure 1. A three-dimensional representation of human liver glycogen phosphorylase, which exists as a dimer composed of two identical subunits, is shown below.



Methodology

About 369 sequence entries in total belonging to the glycogen phosphorylase family exist in the iProclass database. Of these, 107 belong to the domain Eukaryotae, 248 to Bacteria, 14 to Archaea, and 0 to Viruses. Two initial alignments of these sequences will be obtained—one using the ProbCons algorithm and one using Tree-based Consistency Objective Function for alignment Evaluation (T-COFFEE). The T-COFFEE algorithm, 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 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 (5). ProbCons also incorporates a progressive multiple sequence alignment but the alignment it generates is based on hidden Markov model-derived posterior probabilities and three-way alignment consistencies (1).

The same glycogen phosphorylase protein sequences will be aligned using yet a third program, Multiple EM for Motif Elicitation (MEME). MEME is an algorithm which sorts through sequences and finds and reports alignments regardless of their placement along the protein. It is concerned with local alignment rather than global alignment, and its results are assumed to be highly accurate. The alignment outputs from the first two algorithms will be compared to those obtained by MEME, and the three alignments will be superimposed to determine the result that is closest to the MEME alignment, and therefore more accurate.

The integration of results obtained from all three alignment methods will be achieved with the aid of the GeneDoc program, which will overlay the different alignments and provide visualization with the highlighting of common motifs in each alignment. This visualization will allow for additional refinement of the data to a single alignment (4).

Further analysis of the alignment data will be obtained from the SeqSpace program, a statistical technique involving a “top down” principle component analysis, which finds columns in the alignment with highly similar patterns of variation and partitions them into groups (4). The result is a “top down” tree based on sequence to sequence similarities and variations.

In addition, the PHYLIP suite software will be used to conduct a bootstrap analysis of the glycogen phosphorylase family. PHYLIP will calculate a phylogenetic tree for the protein family by means of a re-sampling method based on a random number generator. The phylogenetic tree generated from its tabulation of statistics a “bottom up” tree and can be contrasted with the “top down” tree generated by SeqSpace. Ideally and in most instances, the results obtained from these two methods are the same.

The GEnt program will be used to conduct a group entropy analysis to identify unique features in each partitioned group from the previous step and distinguish their group to group differences. Each set of unique features will be superimposed onto an existing three-dimensional structure representative of that protein group using a visualization program such as RasMol or VMD. Areas that are conserved across the protein family or unique to a group will be color-coded, and this color-coded mapping of regions will allow for the formation of hypotheses relating the location and structure of the critical regions in each group to the function and evolution of the glycogen phosphorylase protein family.

Potential Results

Analysis of the sequence alignments obtained from several different algorithms for the glycogen phosphorylase protein family will determine the method or combination of methods that produce the most accurate alignment. The performance of the relatively new ProbCons algorithm in comparison to that of the more established T-COFFEE will be quantified by experimental results. In addition, the conserved and differentiated regions of the glycogen phosphorylase sequence will be identified, and glycogen phosphorylase homologues will be grouped based on this information. The groupings will aid in the determination of residues unique to a specific group as well as assist the tracing of patterns of evolutionary descent for the entire protein family. A three-dimensional visualization of representative proteins with conserved and unique regions highlighted will aid in the formulation of additional hypotheses concerning protein function.

References

1. Do, C.B., Mahabhashyam, M.S.P., Brudno, M., and Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*. 15 (2005):330-340.

2. Johnson, L.N. Glycogen phosphorylase: control by phosphorylation and allosteric effectors. *The FASEB Journal*. 6 (1992): 2274-2282.

3. Lassmann, T., and Sonnhammer, E.L.L. Automatic assessment of alignment quality. *Nucleic Acids Research*. 33 (2005): 7120-7128.

4. Nicholas Jr., H.B. Glutathione S-Transferase Subfamily Differences: Remodeling the Subunit and Domain Interfaces.

5. Notredame, C., Higgins, D.G., and Heringa, J. T-Coffee: A novel method for multiple sequence alignments." *J. Mol. Bio.* 302 (2000): 205-217.