**Characterization of APE/Ref-1 protein family**

Jay Shukla – Biomedical Engineering – The New Jersey Institute of Technology

Mentor: Dr. Hugh Nicholas – The Pittsburgh Supercomputing Center

**Introduction**

The foundations of biochemistry and molecular biology include the concept that the sequence of a protein directly determines its three-dimensional [3D] structure. In turn, the 3D structure determines the potential biochemical activities it could perform (i.e. protein sequence determines protein function). The structure also determines what other molecules the protein is capable of identifying and interacting with. These concepts can be used to obtain further insights into the biochemical function of a specific protein family. Many different organisms have proteins that carry out the same or similar biochemical functions. However, through natural mutations - what can be thought of as nature’s site directed mutagenesis - it has also been discovered that these proteins do not have the same amino acid sequences. Over time, as species diverged, the sequence of the proteins within them changed. But the functions the proteins carried out did not change, or may have altered slightly. This record of variation in sequences that carry out the same function can be studied to discover which sequence elements are essential to the basic structure and function of the protein, and what differences have created what variations in the basic function.

I will use available software applications to compare these protein sequences, which will allow me to identify the regions where the proteins vary. This information indicates which regions within the protein sequence - and which specific amino acid residues - can tolerate variation without changing the proteins function. This information also shows which
amino acid residues are critical to the basic function, and which ones created variations on the basic function when they changed. From this, further analysis of the protein can be done via 3D modeling. The models will help determine how the varying amino acid residues affect protein function. The variation can alter the protein in at least two ways: (i) structurally, by altering the folding process and exposing different areas, and (ii) functionally, by directly playing a role in the biochemistry that is catalyzed by the protein.

The family of proteins that we have decided to analyze is the human apurinic endonuclease/redox-factor 1 (APE/Ref-1), along with all protein sequences related to the family. This protein is located on the chromosome 14 of humans, and is encoded by a 3 kb gene. It is involved in DNA repair (functioning as 5’AP-endonuclease) and in transcriptional regulation of genes (stimulating AP1 and other transcription factors) (Gerhard, 2000).

Methods

First, we will consult the online protein database IProClass (Wu et al, 2004) and collect the protein sequences of the APE/Ref-1 family. We will then create a multiple sequence alignment using two different programs. All of the protein sequences will be aligned using T-Coffee (Notredame, 2000) and through MEME (Bailey et al, 1994). Both of these programs use different algorithms to identify patterns within the sequences, as well as the differences. T-Coffee creates a global multiple sequence alignment, i.e. it creates a detailed juxtaposition of the letters representing different amino acids over all of the sequences in the data. This alignment reveals conserved regions where variation is rare because variation in these regions impairs the protein function and the viability of the organism in which it is found. The alignment also shows less conserved regions where
variation is well tolerated by the protein. From this, we will be able to point out the patterns that exist within all of the proteins of the APE/Ref-1 family. The MEME program directly identifies conserved regions within the proteins without first generating a global alignment. Thus, this determination of what regions are conserved is independent of the alignment and can be used to refine and access the quality of the alignment generated by T-Coffee. We will also incorporate the biochemical and other computational information to help create an accurate reflection of the protein’s evolution.

Next, we will use a phylogenetic bootstrap analysis to identify distinct subfamilies within the family using programs from the PHYLIP (Felsenstein, 2004) suite of phylogenetic analysis programs. These will be defined as proteins that carry out similar but not identical biochemical activities. As a confirmation of these subfamilies, we will use a principle component analysis [PCA] carried out by SeqSpace (Casari et al, 1995). Through these processes, we will have fairly distinct subfamilies. Cross-entropy analysis will then be used to identify the features that distinguish each subfamily. From this, we will also learn the parts of the protein that are critical in differentiation of the protein’s function.

The final step in our analysis will be a 3D model analysis. Using Modeller (Renom, 2000), we will create models of a several of the protein subfamilies. Using these models, we will analyze the structural differences between the subfamilies. At this point, we will be able to determine whether the differentiation of function is caused through a change in structure or through a direct role in the biochemical activity (Perozich et al, 1999).
Possible Results and Interpretation/Implications

After the first two steps, we expect to see a distinct separation between large numbers of the proteins, thereby allowing us to create definite subfamilies. The cross-entropy analysis, supplemented by PCA, will generate testable hypotheses about the amino acid residues in the different subfamilies of proteins that confer different biochemical specificities upon them. The 3D models will then allow us to develop a better understanding of the changes in the biochemical function and how the changes in the sequences contribute to the new role. Finally, we will have a better overall understanding of the APE/Ref-1 protein family and the way in which the proteins carry out their biochemical and physiological roles within various organisms. We will also gain a deeper understanding about how proteins evolve to accommodate new environments in which their host organisms find themselves.
References


Perozich J., Nicholas H. et al. 1999. Relationships within the aldehyde dehydrogenase extended family. Protein Science. 8: 137-146