Characterization of APE/Ref-1 Protein Family

By Jay Shukla\textsuperscript{1,2}

Mentor: Dr. Hugh B. Nicholas Jr.\textsuperscript{3}

\textsuperscript{1}Bioengineering & Bioinformatics Summer Institute, Center for Computational Biology & Bioinformatics, University of Pittsburgh, Pittsburgh, PA 15261

\textsuperscript{2}Department of Biomedical Engineering, New Jersey Institute of Technology, Newark, NJ 07103

\textsuperscript{3}Biomedical Initiative, Pittsburgh Supercomputing Center, Pittsburgh, PA 15213
Overview

- Summary of Proteins
- Function of APE/Ref-1 Proteins
- Steps behind analysis
  - Multiple Sequence Alignment
  - Phylogenetic & Principal Component Analysis
  - 3D Modeling
Quick Summary of Proteins

- Protein structure is based on an amino acid chain
- Proteins vary in length as well as in sequence of the amino acids
- Function depends directly on sequence, length AND the way the protein folds

\[ \alpha\text{-helix} \]
\[ \beta\text{-sheet} \]
Function of APE/Ref-1 Protein Family

- Involved in repairing damaged DNA
- Involved in transcription regulation of numerous proteins
- Responds to oxidative stress and protects cells from genotoxic and cytotoxic effects of oxidizing agents
- Protein is located in numerous organisms, and exact function/mechanism varies
What are we trying to find/create?

1. Identify the motifs that are characteristic of the protein family
2. Analyze phylogenetic & entropy values to identify distinct subfamilies
3. Identify the residues that are characteristic of each subfamily
Purpose

- Get a better understanding of the protein’s function in various organisms
- Better understand the mechanism behind the protein
- Learn about the residues that are significant to the protein’s function vs. the ones that are not
- Possibly get a better understanding of the protein’s evolutionary history
Multiple Sequence Alignment

[MSA]

- IProClass identified 124 proteins within the APE/Ref-1 family
- Using TCoffee and MEME, a multiple sequence alignment was created
- The programs identified 20 highly conserved motifs within the sequences
- GeneDoc was used to compile the information gathered by TCoffee and MEME
Full Multiple Sequence Alignment

Let’s take a closer look!
PHYLIP & SeqSpace Analysis

- PHYLIP analyzes the MSA and creates a phylogenetic tree
- Clusters sequences based on similar patterns of substitutions
- SeqSpace uses a principal component analysis and complements the PHYLIP
- Four subfamilies were created (AA, 2, 3, 4)
Principal Component Analysis of APE1

Ape1 Protein Co-ordinates  Ape1 Position Co-ordinates

Group 2
Group 3
Group 4

Principal Component 3

Principal Component 2
Phylogenetic Tree Featuring Four Subfamilies

What happened here??
Residue Analysis

GEnt calculates the entropy of entire family
Also calculates entropy distance of each subfamily using the equation:

\[ \text{Group} = \sum_i p_i \times \log_2 \left( \frac{p_i}{q_i} \right) + \sum q_i \times \log_2 \left( \frac{q_i}{p_i} \right) \]

Residues with high scores signify the residues that are unique to the specific subfamily
AA Subfamily Specific Amino Acids Identified by Cross Entropy

Alignment Index Subfamily Amino Acid / Not Subfamily Amino Acid

- 436 A / H
- 9 D / N
- 405 H / P
- 440 T / W
- 293 G / W
- 99 E / K
- 14 W / C
- 442 Y / E
Multiple Sequence Alignment with Characteristic Residues
AA Group

- Contains 15 proteins
- Composed mostly of AAE, APE1 proteins
- Thought to be five characteristic residues within the subfamily
RasMol was used to visually display the protein

Two views were used:
- One shows the motifs that are unique to the APE/Ref-1 Protein Family
- One shows the residues characteristic of the AA subfamily

Now it's possible to see where the residues are located relative to the protein's structure
Protein from AA Group with APE/Ref-1 Motifs Highlighted
Protein from AA Group with Unique Residues Highlighted

9 D/W
14 W/C
99 E/K
442 Y/E
Protein from AA Group with Motifs and Residues Highlighted
The Next Step

- The next step would be to continue to analyze the residues and their locations.
  - Can be done computationally (molecular mechanics & dynamics or quantum mechanics) and in a chemical lab.
- Further refinement of the MSA to clearly determine the location of the five unplaced sequences.
Resources


Resources (cont.)


The Room of Essential Biology. www4.ocn.ne.jp/~bio/biology/protein.htm

Acknowledgements

Mentor: Hugh B. Nicholas Jr. PhD
214 Mental Support: Nick Morsillo, Deanna Nachreiner, Jordan Torok
Rajan Munshi
Everyone at the Pittsburgh Supercomputing Center
Everyone in the BBSI
Supplemental Slides: Stochastic Sampling

- Randomly selects a segment from each sequence but one
- Compares each segment to the left out sequence to score each segment
- Score is then used to randomly select one of the segments
- Process is repeated until left with the most represented pattern
Randomly select a sequence segment from every sequence but **one** to make into a pattern and evaluate the information content of the pattern.

Cyc_Orsya: \( \text{ln}[\text{pkkyi}]\text{PGTKMvfpgkl} \)

Cyc_Tepy: \( \text{hv}\text{PGTKMafaglp}[\text{adkdr}]\text{ad} \)

Cy2_Argtc: \( \text{kk}\text{iPGN}[\text{KMala}]\text{giskpeeldn} \)

Cy2_Rhoru: \( \text{fvleksgdpkAKSKMtfkltkd} \)

C550_Parde: \( \text{pw1}[\text{vkmtd}]\text{dkGATKMtflkm} \)

C550_Nitwi: \( \text{pkakvPGTKMv}[\text{fagik}]\text{kdsel} \)
Refine the Pattern: Picking the next word

Early scores and selection of the next word to add to the pattern

- Sum the scores to create a scale from zero to the total of the scores
- Pick a random number between zero and the total
- The interval on the scale containing the random number marks the word to be included in the pattern for the next cycle

Late scores and selection of the next word to add to the pattern