Introduction:

Biological applications of computers are vastly increasing as computer technology rapidly improves. One such application is the visualization of the three-dimensional structure of certain proteins. There is a direct correlation between the amino acid structure of proteins and the functionally important structure of these proteins. Through evolution and mutation the amino acid sequences of a number of proteins has changed in many ways, yet the overall function of these proteins has remain unadulterated. However, even through the millions of years and the plethora of mutations certain regions of the amino acid sequences have remained identical throughout a certain protein family. It is these conserved regions which we hope to discover with the anticipation that these regions might expose the critical structural regions of the amino acid sequence of selected proteins.

In the search for a protein group to analyze the Drosophila was of particular interest due to the availability of studies on the organism. One of the key events in the development of a Drosophila embryo is the formation of the ventral furrow. There are vast swings in gene expression levels as well as protein concentrations during the invagination of the Drosophila embryo. One such protein is heat shock protein 23. The concentration of heat shock protein
23 increases over 100 fold in the 15 minutes after gastrulation commences\(^1\). This has sparked interest in the protein and the family to which it belongs.

HSP23 is a member of the alpha-crystallin-related small heat shock protein family. The function of these proteins is vast and varied. One function of interest is that these proteins act as chaperones which are critical in the stress response of numerous cells. They protect against oxidative stresses as well as possessing anti-apoptotic properties. In recent studies, a link has been established between members of this protein family and some neurological disorders\(^2\) demonstrating the important biological role of this protein family.

One key area of information associated with any protein is its three-dimensional configuration. With this information, drugs can be designed to block the active site of any protein or modulate the activity of the protein through allosteric interactions. One method to achieve this goal is to crystallize every individual protein in the family. Unfortunately, this method is both tedious and sometimes extremely challenging. Thus, other more efficient approaches are preferred to investigate the 3D structure of proteins. This approach is one of that takes place entirely \textit{in silico}.

In the \textit{IProclass} database\(^3\), there are 122 proteins from the alpha-crystallin-related small heat shock protein family belonging to organisms within the Metazoan and Fungal classification. The goal of my research project is to compare these protein sequences with different software packages and analyze the similarities and differences between the sequences. In this manner, I will identify regions of the amino acid sequences essential to the function of the
proteins and determine the non-conserved sequences/regions that result in
diversity of the protein family.

**Methods:**

The method for the sequence manipulation is essentially a seven step process. Following the sequence analysis visual manipulation occurs. The first step in the sequence manipulation will be the retrieval of the amino acid sequences from the *IProclass*\(^3\) database. The 122 sequences will then be compiled into a text file prior to analyses. The first of these was a sequence alignment by a program called T-Coffee\(^4\). This program performs an approximate multiple sequence alignment. The next step of processing was through an algorithm entitled MEME\(^5\). MEME is a more specific algorithm that searches throughout the entire sequence and does a pair wise analysis to determine if there are any repetitions throughout all of the sequences.

The combination of the two program outputs will be the input format for Genedoc\(^6\). This program serves as an editing platform that parses through and determines the regions that are highly conserved and, therefore, critical to the viability and functionality of the protein. Also when the conserved regions of the amino acid sequence are isolated the non-conserved regions become evident.

Fourth, I will use programs from the PHYLIP\(^7\) suite. These programs will perform a bootstrap analysis on the amino acid sequences to try and separate the 122 sequences down into even smaller subfamilies. Next, I will analyze the subfamilies with SeqSpace\(^8\) to confirm the subfamilies as well as determine which residues are the most influential to the characteristics of the subfamily.
Finally, a phylogenetic tree will be constructed with the above information to visualize these relationships.

Once the sequences are aligned as closely as possible, a three-dimensional model of the protein can be constructed and color-coded to demonstrate which regions of the protein are conserved between organisms in the family using the visualization program RasMol. These color-coded areas are uniquely important to the function of the protein and can be the main consideration of future experiments and analyses. Also, with knowledge of these regions, models can be extrapolated from models that are already in the existing Protein Data Bank.

**Possible Results and Implications:**

After the first set of algorithms is executed on the sequences, it should be possible to categorize the proteins/sequences into different groups. These groups should provide varied degrees of insight. One possible outcome would be to determine evolutionary pathways between organisms in the same family. Another outcome would be to propose possible functions of the different proteins of the family. This is especially useful in cases where a function has not been determined for a member of the family.

With a functional three-dimensional model of the protein family, a number of hypotheses could be suggested. Testable hypothesis are an invaluable resource to the scientific community. This model would cater itself well to computational biologists for testing in a variety of fields from neuroscience to
oncology. With the above implications, it can be seen that this research is worthy of the time and resources allocated.

References


