Comparison and Analysis of Heat Shock Proteins in Organisms of the Kingdom Viridiplantae

Emily Germain1,2, Hugh Nicholas Jr.3

1 Bioengineering & Bioinformatics Summer Institute, Department of Computational Biology, University of Pittsburgh, (2021)
2 Departments of Biomedical Engineering and Biology, Rensselaer Polytechnic Institute, Troy, NY 12180
3 Biomedical Initiative, Pittsburgh Supercomputing Center, Pittsburgh, PA 15213

I. Create a Multiple Sequence Alignment

Multiple sequence alignment created using the T-Coffee global pairwise alignment method, shown with MEME patterns identified. The alignment displays regions that are highly conserved and therefore likely to be critical to the protein structure or function. The regions with higher degrees of variation are those that are less essential to the protein and are tolerant of mutations.

Model small heat shock protein from Methanococcoides janaschii

II. Construct a Phylogenetic Tree and Divide Sequences into Subfamilies

Consensus phylogenetic tree constructed using RNAlign program package with 1000 iterations

Residues highly conserved over family: red – 100% conserved, dark blue – 50% conserved

Sequence Space clusters defining similarity vectors from origin and resulting subfamilies

III. Calculate Group Entropy for Each Subfamily

Group Entropy plot from GEnt for the HSP 17 subfamily


Abstract/Methods

Heat shock proteins are found in every living cell and have a broad range of functions. They are part of the cell’s response to stresses such as extremes of temperature, deprivation of oxygen or glucose, or exposure to toxins. They normally make up about two percent of a cell’s soluble protein, but in a stressed cell they can account for twenty percent. HSPs help proteins denatured by stress to refold back into their proper shape, and most can also chaperone the folding of newly made proteins. They transport other proteins between compartments within the cell and possibly function in the immune response. The full range of functions of HSPs is unknown.

651 sequences for heat shock proteins of organisms of the Kingdom Viridiplantae were extracted from the data available in the Pfam database. These sequences were aligned using two different algorithms: T-Coffee and MEME. T-Coffee created a global multiple sequence alignment, which attempted to line up all the similar regions of families and compare across the entire set. MEME was run using the Zero or One Per Sequence method to identify twenty motifs. The patterns identified are position independent conserved sequence elements that aid in judging the accuracy of the T-Coffee results. These patterns were used to manually refine the results found during the global alignment.

Using RNAlign, a bootstrapping analysis was performed to separate the proteins with similar biochemical activities into distinct subfamilies and quantify how closely related members of a subfamily are. A SVDSpace analysis was calculated which columns of the alignment have the most and least similar sequence variations to confirm groupings. A phylogenetic tree was constructed to visualize these relationships. A cross-entropy analysis was calculated with GEnt to identify which residues are unique to a particular subset of the family and contribute to its specific properties.

After the highly conserved residues were identified, three dimensional graphical models were constructed using VMD and a general representative for the protein family. Important features and residues that define the subfamily are highlighted and the models will be used to form hypotheses about which areas make up the active site or bind to other molecules, which parts are critical to maintaining a functional structure, and how the protein performs its functions.

Conclusions

The evolutionary relationships shown in these heat shock proteins suggest that different variations result from gene duplication. In general, heat shock proteins are more closely related to others in species similar to the one in which they are found, rather than to others of comparable molecular weights in more distantly related species.

Heat shock proteins are highly conserved over the whole family and it is very specific residue alterations that give particular subfamilies their individual properties.

The data collected in this study can be further analyzed by comparing the highly conserved residues found in each group. This can be matched up with data regarding the specific functions of each heat shock protein to generate hypotheses regarding how these specific residues contribute to functional specificity and biochemical properties.

Acknowledgements

Pittsburgh Supercomputing Center
National Institutes of Health
National Science Foundation