

Original Research Article

# Computational Proteomic Data Mining of Heat Shock Protein- $\beta$ 1 using Mathematica Software

Rashid Saif<sup>1,2\*</sup>, Saeeda Zia<sup>2,3</sup>, Kinza Qazi<sup>4</sup>, Tania Mahmood<sup>1</sup>, Fatima Asif<sup>1</sup>, Aniqaj Ejaz<sup>1</sup> and Talha Tamseel<sup>4</sup>

## Abstract

<sup>1</sup>Institute of Biotechnology, Gulab Devi Educational Complex, Ferozepur Road, Lahore, Pakistan

<sup>2</sup>Decode Genomics, 264-Q, Johar Town, Lahore, Pakistan

<sup>3</sup>Department of Mathematics, National University of Computer and Emerging Sciences, Lahore, Pakistan

<sup>4</sup>Department of Bioinformatics and Computational Biology, Virtual University of Pakistan, Lahore, Pakistan

\*Corresponding Author's E-mail:  
[rashid.saif37@gmail.com](mailto:rashid.saif37@gmail.com)

Omics data mining approach helps us to discover the hidden patterns of today's molecular life. Current study is conducted to compare and analyze 75 homologous *Hspb1* protein sequences from distinct 62 species. Multiple sequence alignment and phylogenetic analysis are performed which revealed that, the fungi, bacteria, plants, human and other mammals appeared in different clades on the basis of orthology and paralogy being used through Mathematica software. Sequence based clustering analysis and physicochemical properties are observed which might serve as an alternative technique to characterize the protein further. Structure and motif analysis are also examined to have insight of conserved domains of this protein, similarly, orthology of human *Hspb1* presented optimal alignment scores which demonstrate important characteristics of this protein. *Hspb1* interacting network and pathway analysis are also analyzed with other proteins through the built-in Mathematica algorithms through gene expression enrichment scores. This tool may be further use to analyze the bigger genomics, transcriptomics and metabolomics dataset to check the reliability of clustering algorithm of this software with other contemporary packages in the field of bioinformatics and computational biology.

**Keywords:** Mathematica, Cluster analysis, Proteomic data mining, Orthology, Network and Pathway analysis, Gene expression enrichment scores

## INTRODUCTION

Heat shock proteins (HSPs) are produced in stressful conditions both during intrinsic and extrinsic stimuli. One of HSPs is heat shock protein beta 1 (*Hspb1*), alternative names include CMT2F, HMN2B, Hsp25/27/28, SRP27, HS.76067, stress-response protein 27 and estrogen-regulated 24 kDa protein whose molecular weight ranges from 12-40 kDa (Asea and Brown, 2008). This diversified protein acts as a protein chaperone, activates proteasomes degrading junk proteins, act as an antioxidant and plays a key role in the inhibition of apoptosis (Cao et al., 1999). It was also detected as a surface membrane protein in some cancer cell types which act as a biomarker and therapeutic target (Vidyasagar et al., 2012). Mathematica software was employed for evolutionary and protein data studies in which 75 *Hspb1* sequences of 62 different species in FASTA format was taken and various functions of this software were considered for multiple sequence alignment, phylogenetic analysis, for exploration of

physicochemical properties, pattern matching, pathway network and analysis etc (Satpathy et al., 2013).

A typical phylogenetic analysis of protein sequence data involves distinct steps i.e. data collection, inference of homology, sequence alignment and alignment trimming (Rokas, 2011). The aim of phylogenetic trees is to depict historical (i.e. evolutionary) relationships. Phylogenetic analysis is sometimes regarded as being an intimidating complex process that requires expertise. In fact, it is a straightforward process that can be learned quickly and applied effectively. Owing to the technological advancement of the sequencing techniques in molecular biology and the ability to collect large amounts of data from disparate organisms, evolutionary studies are still of higher interests (Soltis and Soltis, 2003). Similarly, pathway and network analysis is one of the techniques used in bioinformatics to predict a relation of protein with other proteins present within a cell or within other cells. It is synonymously related to enrichment and gene set

analysis. Through these studies, entire metabolic map of a cell could be redesign by rational and directed metabolic methods (Schilling et al., 1999).

Mathematica software is a computer algebra system or program, used mainly in computing fields of applied mathematics (Buchmann and Shipman, 1991). One perspective to view Mathematica is from the world of data science such as bioinformatics and system biology. The Wolfram bioinformatics solution effortlessly integrates specialized capabilities such as built-in computable genome and protein data and sequence alignment with sophisticated computational aesthetics and award winning design. Mathematica presents the results beautifully, instantly creating top of the line interactive visualizations and publications quality documents.

Our aim is to do firstly, the clustering analysis of species, on the basis of *Hspb1* by Mathematica software and cross checking its results with software. Secondly, to specifically demonstrate the outcomes of Mathematica related to its exclusive protein analysis.

## **MATERIALS AND METHODS**

### **Mathematica software**

Proteomic and cladistics in silico analysis using "Mathematica software" such as clustering tree, physicochemical properties, structure and motif analysis, pathway and network exploration and much more was reconnoitered. Protein analysis has wide range options to explore about it such as its properties, phylogenetic methodologies like clustering tree, dendrogram and other non-hierarchical visualizations e.g. nearest neighbor graph, clustering components etc. Working on it encompasses know-how with its algorithms and functions, downloading <https://www.wolfram.com/mathematica/trial/> and installation procedure which requires activation key. By clicking option of "other ways of activation" and then on "manual activation", the Wolfram Mathematica keygn.exe folder is copied. Clicking the windows button and typing cmd.exe in search bar, paste the copied folder, type math id written in wolfram product activation window. System will provide an activation key and password. This will activate the software.

### **Hierarchical clustering algorithm**

Cluster analysis is a group of multivariate techniques whose primary function is to group objects (e.g. products or entities) based on the characteristics they possess. It is a means of records based upon attributes that makes them similar. It is based on the union between the two

nearest clusters. Through it, clustering tree is created in this paper to check the trait of Mathematica software algorithm.

### **Edit distance function**

Edit distance (measure of distance between two strings) works by transforming one string into another by deletions, insertions and substitutions in individual character within a string (Gusfield, 1997). In Mathematica software, it is the Levenshtein distance (one of the types of edit distances) being used. Clustering tree uses this function which is one of the built-in functions of Mathematica software.

### **Wolfram alpha**

Wolfram alpha is based on Mathematica software that take input in Wolfram natural language and retrieves data robustly in curated form on Mathematica notebook within couple of seconds without any exceptional efforts, that distinguishes it from other semantic search engines (Hoy, 2010). As a command on Mathematica, it is used for various proteomic and genomic analysis in this research study.

### **Structure and motif analysis**

Molecular structure is a property of protein data that can be determined by molecule plot and serves as an essential entity of Mathematica software while protein motif represents the conserved portion of proteins that has structural and functional significance.

### **Physicochemical characters prediction**

Mathematica gives detailed description of physicochemical properties i.e. its molecular weight, theoretical isoelectric point, extinction coefficient, half-life of the protein, hydrophobicity, aliphatic side chains and GRAVY.

### **System modeler for pathway analysis**

Many different modeling tools of Wolfram are attached to Mathematica software which has several built-in functions related to the field of bioinformatics to analyze gene expressions, pathway analysis, gene annotations and drug designs. These functions are analyzed faster and more accurately by Mathematica software with the help of system modeler. Through it, interaction of *Hspb1* with other genes is also developed in this paper.

```

ClusteringTree[{"BAB17232-HSPB1-Homo sapiens", "NP001247878-HSPB1-Racaca mulatta", "NP001531-HSPB1-Homo sapiens", "ABX80641-HSPB1-Drosophila buzzatii",
"BAF94137-HSPB1-Alligator mississippiensis", "P04792-HSPB1_Homo sapiens", "ABI26632-HSPB1-Carassius auratus", "ABI26639-HSPB1-Carassius auratus",
"BAE93468-HSPB1-Carassius auratus", "AH004458-HSPB1-Rattus sp", "ALC43477-HSPB1-Drosophila busckii", "AAV54182-HSPB1-Sus scrofa",
"AFY13335-HSPB1-Oreochromis niloticus", "NP001266459-HSPB1-Oreochromis niloticus", "AAA28638-HSPB1-Drosophila melanogaster", "AFK93550-HSPB1-Capra hircus",
"NP114176-HSPB1-Rattus norvegicus", "AAA41353-HSPB1-Rattus norvegicus", "P14602-HSPB1-Mus musculus", "AAH64051-HSPB1-Mus musculus",
"NP001008615-HSPB1-Danio rerio", "AAH97148-HSPB1-Danio rerio", "NP001087285-HSPB1-Xenopus laevis", "NP001087283-HSPB1-Xenopus laevis",
"NP001003295-HSPB1-Canis lupus familiaris", "AITS2079-HSPB1-Canis lupus familiaris", "NP001027772-HSPB1-Ciona intestinalis", "AKL81619-HSPB1-Bos grunniens",
"JAC05005-HSPB1-Ceratitidis capitata", "ACD76913-HSPB1-Ceratitidis capitata", "JAB92600-HSPB1-Ceratitidis capitata", "AAB46593-HSPB1-Poeciliopsis lucida",
"013224-HSPB1-Poeciliopsis lucida", "P15991-HSPB1-Cricetulus longicaudatus", "Q3T149-HSPB1-Bos taurus", "NP001020740-HSPB1-Bos taurus",
"Q00649-HSPB1-Gallus gallus", "F58987-HSPB1-Drosophila ficusphila", "Q6R272-HSPB1-Saguinus oedipus", "Q4F877-HSPB1-Oncorhynchus mykiss",
"D6BL40-HSPB1-Oncorhynchus tshawytscha", "D6BL51-HSPB1-Oncorhynchus nerka", "F58961-HSPB1-Drosophila sulfurigaster albostrigata",
"F58997-HSPB1-Drosophila takahashii", "F58959-HSPB1-Drosophila malerkotliana", "F58981-HSPB1-Drosophila ananassae", "A0A0A1X0N2-HSPB1-Bactrocera cucurbitae",
"F58956-HSPB1-Drosophila parabipectinata", "F58993-HSPB1-Drosophila mauritiana", "A0A0688B78-HSPB1-Lucilia cuprina", "F58994-HSPB1-Drosophila parabipectinata",
"F58998-HSPB1-Drosophila tani", "A0A0K8UP97-HSPB1-Bactrocera latifrons", "A0A034NGS6-HSPB1-Bactrocera dorsalis", "JAC40360-HSPB1-Bactrocera dorsalis",
"JAC55533-HSPB1-Bactrocera dorsalis", "JAC52973-HSPB1-Bactrocera dorsalis", "F58982-HSPB1-Drosophila bainaii", "F58990-HSPB1-Drosophila lini",
"F58962-HSPB1-Drosophila repletoides", "F58986-HSPB1-Drosophila elegans", "F58960-HSPB1-Drosophila albomicans", "A0A0C9R0M2-HSPB1-Fopius arisanus",
"G4V8X3-HSPB1-Schistosoma mansoni", "F589A1-HSPB1-Drosophila eugracilis", "G4W8Y3-HSPB1-Kryptolebias marmoratus", "Q2VFP7-HSPB1-Mytilus galloprovincialis",
"F589A0-HSPB1-Drosophila triarvaria", "F58984-HSPB1-Drosophila biarmipes", "F58988-HSPB1-Drosophila jambulina", "F58989-HSPB1-Drosophila kikkawai",
"ADX98507-HSPB1-Larimichthys crocea", "ATU47315-HSPB1-Bubalus bubalis", "NP001072817-HSPB1-Xenopus tropicalis", "P02518-HSPB1-DROSOPHILA MELANOGASTER"}];

```

**Figure 1.** Mathematica input syntax for “Clustering Tree” showing accession number, protein and species names.

## Open PHACTS

The Open PHACTS Discovery Platform contains all the data sources (ChEBI, ChEMBL, Sure ChEMBL, ChemSpider, ConceptWiki, DisGeNET, DrugBank, Gene Ontology, neXtProt, UniProt and WikiPathways). which already use to combined and linked together to see the relationships between compounds, targets, pathways, diseases and tissues. The platform has been used to answer complex questions in drug discovery and results have been executed on Mathematica platform by making account on Open PHACTS and then entering account key and ID to connect with the notebook. After it, different biochemical and molecular analysis can be channeled.

## Pathway lab and neural network function

Pathway Lab Research Edition is a Mathematica application package designed to analyze, and document biochemical pathways and to visualize associated data related to their focus disease which allows scientists to make decisions regarding the cause of disorder and potential drug targets. These pathway models can integrate kinetic, gene expression, and functional data. Neural network is one of the most appropriate built-in functions of pathway lab for drug discovery, apropos with this article by using related packages, platforms and functions, pathway analysis of *Hspb1* with its related gene is prophesied on the basis of their gene expression enrichment score. With the help of CDF player (computable document format) in Mathematica, it is easy

to visualize the cell signaling relation with in gene or relational pathways with other genes.

## Sequence alignment tool

NCBI is a huge database that houses a series of databases relevant to biotechnology. For sequence alignment one of its popular resource “BLAST (Basic local alignment search tool) 2 Sequence” tool is used <https://blast.ncbi.nlm.nih.gov/Blast.cgi> (Tatusova and Madden, 1999). Non redundant protein sequences database along with Protein-protein BLAST (blastp) algorithm was chosen to BLAST 75 Hspb1 sequences of 62 different species for cross validation of clustering tree produced through Mathematica software.

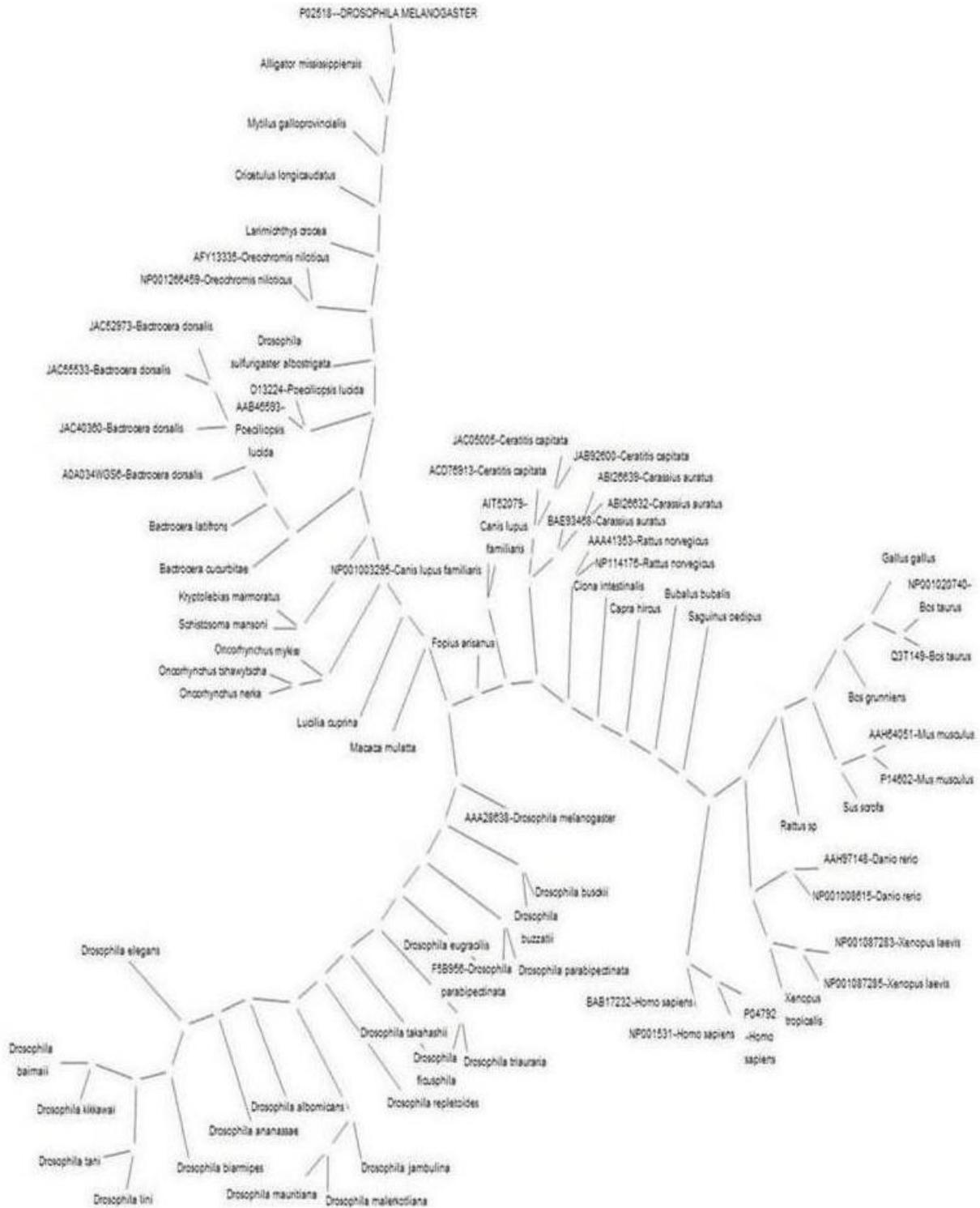
## RESULTS

### Cluster Analysis

#### Clustering tree construction

Evolutionary relationship was observed via clustering tree through Mathematica software. This dendrogram (weighted tree) was constructed through hierarchical approach using built-in edit distance function by joining sub-clusters to nodes and edges, which are labels here with species names (Figure 1).

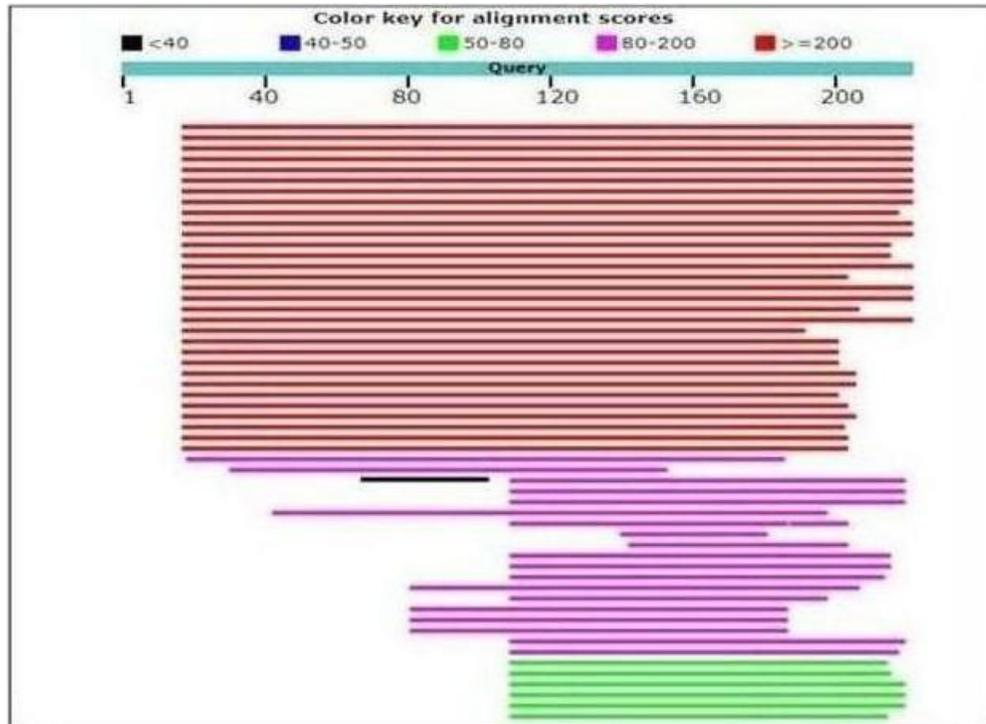
The output of clustering tree showed an unrooted tree without any outgroup, which have diverged clades and



**Figure 2.** Unrooted weighted clustering tree of 75 sequences of 62 different species constructed by hierarchical data visualization through Wolfram Mathematica software, which represents the relatedness among species by grouping closely related species in same clades.

signified its association with the organisms of its own class or the other (Figure 2)

For instance, beginning from lowermost clade, where *Drosophila* species reside and symbolized that all insecta



**Figure 3.** Graphical overview of the Ref sequence mRNA BLAST hits for query sequence. Identity % are shown which revealed that first eight species are closely related and their more than 200 nt are aligned. ABI26632*Carassius auratus* shown in black whose less than 40 nt displays high degree of dissimilarity with Hspb1 of *Homo sapiens*.

are ancestrally interconnected and somewhat far from *Homo sapiens*. In such trees, ancestral lineage is not known due to the absence of central node. All fruit flies (*Bactrocera* species) appeared to be in close proximity with each other. Similarly, clade which has *Homo sapiens*, presented that, besides mammals, Amphibia (*Xenopus laevis*) and one of the species of fish the zebra fish (*Danio rerio*) are also homologous for this particular protein. As the branch length increases and process of speciation occurs, this divides the tree into branches and forms clades and leaves and indicates the distance from the reference species.

Other fish species like *Larimichthys crocea*, *Oreochromis niloticus*, *Poeciliopsis lucidia*, *Orcorhynchus sp.* and *Kryptolebias marmoratus*) are related to reptile and some insects such as *Alligator mississippiensis* and *Lucilia cuprina*, *Schistosoma mansoni* respectively. But reptile and these insects are far away from *Homo sapiens* due to more leaves divergence and increased branch length while the same fish's showed more degree of similarity with *Homo sapiens*.

This phylogenetic tree construction and evaluation by Mathematica software is an outburst in the field of molecular biology particularly for researchers in this field of quantitative life sciences and bioinformatics. In current study, further analyses are also done on phylogenetic to explore other protein properties and some more

exceptional features for cross validation of this clustering analysis.

### Sequence alignment of Hspb1 for cross validation

Persistently sequence alignment of raw data of 75 sequences through NCBI ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)) integrated with "BLAST2Sequence" tool (Protein BLAST) was done for cross validation of clustering tree results. Query sequence given to BLAST with subject sequences where BAB17232-HSP27-*Homo sapiens* sequence as out group.

Alignment color key is an attractive feature of "BLAST2Sequence" which also shows quality score of the alignment. This aided in exact measurement of relatedness among species. Pair wise alignment was done and then graphical overview was displayed (Figure 3).

Species like *Alligator mississippiensis*, *Mytilus galloprovincialis* and P02518-*Drosopila melanogaster*, which were shown to be in close proximity by clustering tree, similar results were not achieved by BLAST. This difference in identity percentage showing relatedness with *Homo sapiens* suggested that these species are not correlated (Table 1).

**Table 1.** Description of alignment scores showing results of first three species, contradicted with results of clustering tree, next 8 species showing almost 100% identity and last one showing maximum dissimilarity with Hspb1 of *Homo sapiens*.

Description	Max Score	Total Score	Query Cover	E Value	Identity %	Accession
Q2VFP7-HSP27- <i>Mytilus galloprovincialis</i>	84.7	84.7	18%	5e-25	95%	Query_178777
P02518-HSP27- <i>Drosophila melanogaster</i>	84.3	84.3	47%	5e-23	45%	Query_178785
BAF94137-Hsp27- <i>Alligator mississippiensis</i>	262	262	78%	8e-93	67%	Query_178716
NP001247878-HSP27- <i>Macaca mulatta</i>	418	418	92%	2e-154	100%	Query_178712
P04792-HSP27_ <i>Homo sapiens</i>	416	416	92%	2e-153	99%	Query_178714
NP001531-HSP27- <i>Homo sapiens</i>	416	416	92%	3e-153	99%	Query_178713
Q6R2J2-hsp27- <i>Saguinus oedipus</i>	368	368	92%	2e-134	88%	Query_178749
AAV54182-Hsp27- <i>Sus scrofa</i>	363	363	92%	2e-132	88%	Query_178722
P14602-HSP27- <i>Mus musculus</i>	357	357	92%	6e-130	83%	Query_178729
AIU47315-HSP27- <i>Bubalus bubalis</i>	354	354	92%	6e-129	86%	Query_178783
Q3T149-HSP27- <i>Bos taurus</i>	354	354	92%	6e-129	86%	Query_178745
ABI26632-HSP27- <i>Carassius auratus</i>	35.4	35.4	16%	2e-06	42%	Query_178717

### Molecule plot and conserved domain analysis

Molecule plot represented actual structure of protein while domain represented the functional unit of protein. Both are elucidated by Mathematica software simultaneously. Syntax format and their output in next line is shown below as

```
In=["HSPB1", "Domain"] In=["HSPB1", "MoleculePlot"]
Out=[AphaCrystallinHsps] Out=Missing[NotAvailable]
```

Hspb1 is not a known entity for molecule plot in Mathematica software. In general, it initialized the output by getting the required information from Wolfram data server which further had protein data source information from distinct databases such as UniProt, NCBI and PDB. In output, missing is a symbolic object with no default evaluation rules.

Determining protein domain is an integrated property in Mathematica software which analyzes it as a conserved domain (functional part of protein that is common in all of the species). It showed that "Alpha Crystallin Hsps" (a domain of about 90 residues) was conserved in all Hspb1 proteins of these 62 different species (Augusteyn, 2004).

### Importing characteristics of Mathematica software

The alternative method to get structure information is to

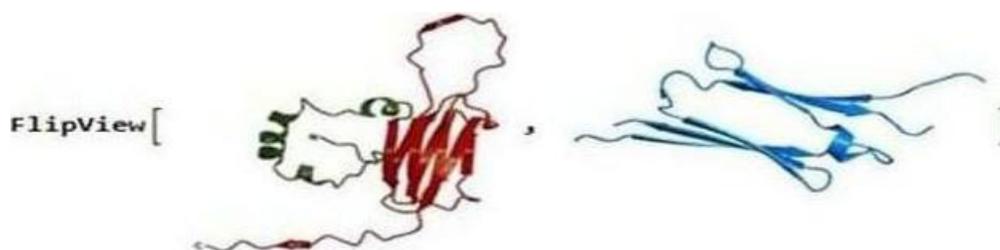
use another Mathematica command "Import". This was done by clicking on the insert option on menu bar, then add file path or file directly already saved in the computer. This can also be done by writing or pasting URL directly rather than to save file. In this case, file path was added for both molecule plot and domain structure (Figure 4).

### Orthology of Hspb1 of *Homo sapiens*

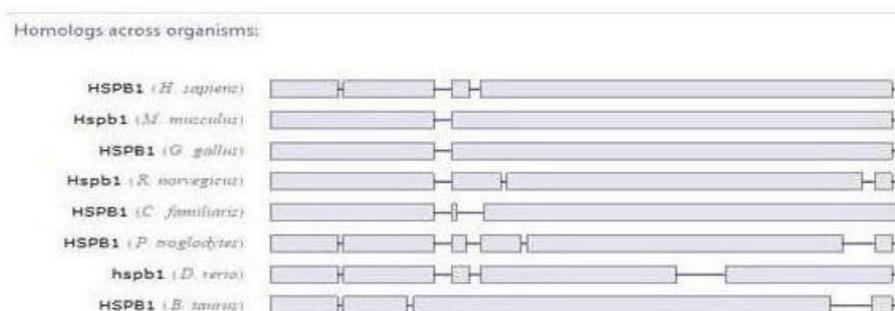
An alternative analysis based on the homology of Hspb1 of *Homo sapiens* across all known organisms (foregoing analysis are based on given data of 62 different species only) was demonstrated using command WolframAlpha [HSPB1]. It retrieves data from Wolfram alpha knowledgebase and understands free-form natural language which is powered by Wolfram language. It revealed that across all organisms, 7 species are closely related and are parallel to the given data used in the preceding analysis except a species of *Pan troglodytes* (Figure 5).

### Physiochemical properties

Properties information in Mathematica software is speci-



**Figure 4.** Red and green color represents the domain which is conserved in all species and blue color represents molecule plot concurrently imported on Mathematica notebook.



**Figure 5.** Showing tremendously related species with *Homo sapiens* Hspb1 analysis across all organisms

**Table 2.** Physiochemical properties analysis of Hspb1 in *Homo sapiens* through Mathematica software.

Properties	<i>Homo sapiens</i>
Domain Positions	{{87,168}}
Sequence Length	205 amino acids
Molecular Weight	22651 Daltons
Biological Processes	{Anti Apoptosis, Cell Motility, Regulation Of Translational Initiation, Response to Unfolded Protein}
Molecular Functions	{Identical Protein Binding}
Domain IDs	29382
Cellular Components	Cell Surface, Cytoplasm, Cytoskeleton, Nucleus
Gene ID	3315
Interacting Genes	AKT1, CRYAA/AB/BB2/GC, DAXX, EIF 461, ILK, PPA1, TGFb1t1, TP53, MAPKAPK5, MAPKAPK2, MAGED1, POP7, MED31, HSPB8, Clorf103, IGSF21
Locus Strings	7q11.23
Chromosome	Chromosome 7
Gene Ontology IDs	0005634, 0005737, 0005856, 0006446, 0006916, 0006928, 0006986, 0009986, 0042802

fied with *Homo sapiens* only. The types of physiochemical parameters include domain position, sequence length, molecular weight, biological processes, molecular functions etc. which were calculated (Saif et al., 2016). Inputs are given as `[[entity, property]]` in tabular form using "Grid" command to explore these characteristics of protein through Mathematica.

### Properties specifications

Gene ID can be used to retrieve this protein on different databases just by writing its ID similar is the case with Domain IDs. Cellular component is another momentous property described that protein is located in or is a subcomponent of a particular cellular component.



```

MIN = {0.64, 0.62, 0.62, 0.77, 0.76, 0.92, 0.55, 0.84, 0.85, 0.69, 0.54, 0.51, 0.73, 0.37, 0.30, 0.43, 0.51, 0.90, 0.61, 0.72, 0.43, 0.67, 0.40, 0.67, 0.68, 0.74,
0.54, 0.57, 1.28, 0.61, 0.33, 0.80, 0.81, 0.42};
MAX = {0.75, 0.71, 0.84, 0.88, 0.88, 1.02, 0.67, 1.10, 1.07, 1.11, 0.89, 0.78, 0.87, 0.78, 0.37, 0.99, 0.81, 1.09, 0.83, 0.91, 0.78, 0.81, 0.91, 0.74, 1.08, 1.07,
0.68, 0.66, 1.42, 0.82, 0.67, 0.98, 0.88, 0.54};

NormDef = {0.67, 0.68, 0.62, 0.88, 0.79, 0.92, 0.60, 0.99, 0.85, 0.69, 0.54, 0.78, 0.75, 0.37, 0.36, 0.47, 0.51, 0.95, 0.61, 0.81, 0.52, 0.72, 0.56, 0.70, 0.77,
0.76, 0.63, 0.58, 1.39, 0.66, 0.51, 0.98, 0.88, 0.50};

CirrboDef = {0.73, 0.68, 0.75, 0.85, 0.79, 0.97, 0.58, 0.95, 0.96, 0.92, 0.81, 0.63, 0.79, 0.67, 0.33, 0.93, 0.78, 1.01, 0.72, 0.84, 0.75, 0.74, 0.62, 0.72, 0.96,
0.86, 0.65, 0.59, 1.35, 0.70, 0.49, 0.91, 0.84, 0.42};

names = {"ATK1", "CRYAA", "CRYAB", "CRYGC", "DAUX", "EIF4G1", "HSP88", "HSP81", "ILK", "PPA1", "TGFBI1", "TP53", "MAPKAPK5", "MAPKAPK3", "MAGED1", "POP7",
"RED31", "Clorf103", "DKFZP56400523", "IGSF21", "PND1", "PPL", "S100P", "SCRIB", "SH3YL1", "STEAP4", "TROP", "TROP", "TRAP", "TRK", "TUBG1", "VIM", "ZBTB33", "ZCCH14", "TNF"};

```

Figure 8. Pseudo code of *Hspb1* enrichment score.

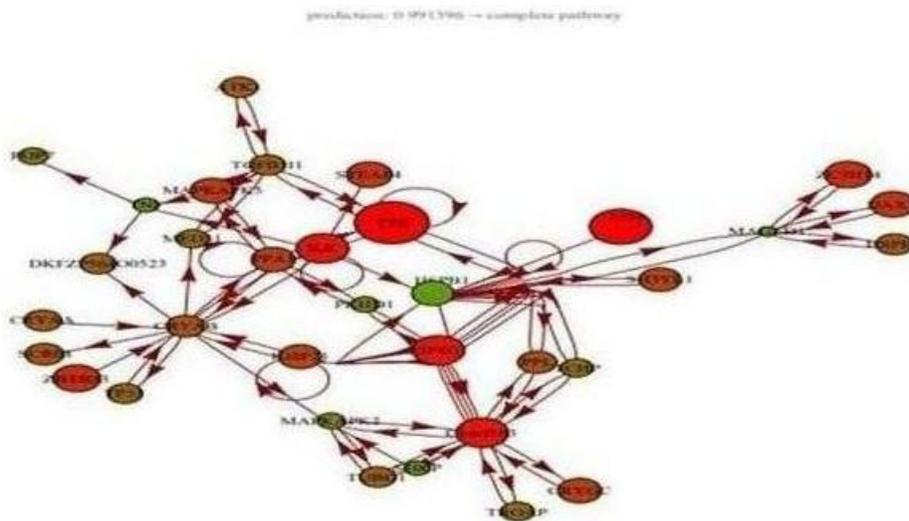


Figure 9. Gene interaction pathway with *Hspb1* based on least enrichment scores.

## Pathway Analysis

Open PHACTS Api was connected to Mathematica notebook and then giving command to provide pathway organisms and its compound, the Open PHACTS picked up required information from its various connected data sources and executed its complete information (Figure 6).

Different genes interconnected with *Hspb1* are obtained from different online packets and subjected to sensitivity analysis, method based on an artificial neural network to find transition phases from homeostasis to altered conditions. Complete pathway with homeostatic conditions (normal functioning of these genes) is shown (Figure 7).

After the complete pathway analysis, down-regulation

and up-regulation of these genes expressions were also observed by enrichment scores analysis to demonstrate and to study the alterations in interconnected genes. Part of the code for whole neural network architecture shown in (Figure 8).

Neural network can compute parameter variations by manipulating control bars as enrichment score can govern the fate of the pathway. Least enrichment score pathway are obtained using CDF player (Figure 9).

## DISCUSSIONS

Traits of Mathematica software in phylogenetic and proteomic data analysis are explored. For this purpose

Wolfram Mathematica version 11 was installed by getting activation key. For required analysis, raw amino acid sequence data of 62 species having 75 sequences in "FASTA" format was obtained from renowned databases of NCBI and Ensembl genome browser. These FASTA sequence are utilized as an input data in aforementioned analysis.

Phylogenetic tree are computed using Mathematica software with heuristic approach on 62 different species resulted in an unrooted, bipartite, connected, simple, loop free, undirected, weighted tree without out-group (Penny et al., 1992). Generated cluster tree are labeled, graph layout are settled in radial visuals, edited and relocated. Cluster tree showed three main clades: first one mainly showed drosophila species, other consist of mammalian species and last one showed leaves of class *Actinopterygii*, insecta, reptiles and mollusk. This analysis was analogous with the parallel analysis done on well-known database NCBI using its "BLAST2Sequence" tool for cross checking by comparing sequences directly. BAB17232-Hspb1-*Homo sapiens* are taken as query reference sequence. Sequence in FASTA format are pasted in respective boxes and automated appliance of algorithm generated superlative results in the form of color key, description (hit table) and alignment scores. These results cross validate that hierarchical algorithm of Mathematica software could be reliable for phylogenetic analysis and dendrogram construction. In sequence alignment, query and identity percent precisely relates the homology of species by inferring E-value which decreases exponentially as the score of matches increases.

Conserved domain and protein structures revealed that molecule plot speak a ribbon like structure of protein while domain exists in the secondary structure (3-dimensional structure by way of coiling and folding) of proteins. So, we deduced that alpha-crystalline (Hsps) domain which act as molecular chaperones, preventing undesired protein-protein interactions and assisting in refolding of denatured proteins is conserved in all species. To demonstrate, molecule plot has been imported along with the secondary structure from external sources on Mathematica notebook simultaneously. Flip view was chosen to view their structures side by side. Now, Hspb1 protein of *Homo sapiens* homology across all organisms was analyzed in natural language and gaining concise but inclusive output both as gene and protein data within no time with Mathematica notebook using command Wolfram alpha (another Wolfram flagship based on Mathematica). For current study, our interest was in result presenting homology across all species in phylogenetic context which demonstrated that *H. sapiens*, *Mus musculus*, *Gallus gallus*, *Rattus norvegicus*, *Canis lupus familiaris*, *Pan troglodytes*, *Danio rerio*, *Bos taurus* are closely related to each other. Among them *Mus musculus* and *Gallus gallus* have the same indels.

Nevertheless, it also provide information about SNPs, splicing structure, amino acid frequency graph, gene/protein ontology characteristics.

Subsequently, exploring the physiochemical properties of Hspb1 protein and the gene which is present on chromosome 7 having the locus string 7q11.23. From the structural point of view Hspb1 is characterized by low molecular weight ranging from 15-40 kDa (Asea and Brown, 2008). Sequence length shows the total number of 205 amino acid in this protein. It also interferes with other gene products which includes (*AKT1*, *CRYAA/AB/BB2/GC*, *DAXX*, *EIF 461*, *ILK*, *PPA1*, *TGFb1t1*, *TP53*, *MAPKAPK5*, *MAPKAPK2*, *MAGED1*, *POP7*, *MED31*, *HSPB8*, *Clorf103*, *IGSF21*).

These gene interactions are studied in detail using "system modeler" based on Mathematica software. List of interacting genes were taken from different packets and 3-Dimensional pathway are obtained using neural network (built-in function of Mathematica) and visualized by CDF player which demonstrated about other related genes that would have an impact if *Hspb1* is disturbed (either upregulated or downregulated). Gene expression enrichment score analysis was also done simultaneously, this approach may assist in understanding of association with disease phenotypes. This in silico research approach to study the theoretical aspects of biomolecules may pave the path to future studies of pharmacogenomics, to study bioinformatics aspects of protein structure, its functions and interactions with other molecules and pathways.

## CONCLUSION

Clustering method and pathway analysis used in this work through Mathematica software demonstrates the ways to ascertain the similarity without aligning the sequences and tremendous view of gene interaction promotes understanding of phenotypic landscapes in more complex organisms without unusual efforts.

## REFERENCES

- Asea AA, Brown IR (2008). Heat shock proteins and the brain: implications for neurodegenerative diseases and neuroprotection. Springer Science & Business Media.
- Augusteyn RC (2004).  $\alpha$ -crystallin: a review of its structure and function. *Clinical and Experimental Optometry* 87: 356-366.
- Buchmann S, Shipman C (1991). Foraging distances flown by honey bee colonies: analyses using Mathematica software. *American Bee Journal* 131: 771.
- Cao Y, Ohwatari N, Matsumoto T, Kosaka M, Ohtsuru A, Yamashita S (1999). TGF- $\beta$ 1 mediates 70-kDa heat shock protein induction due to ultraviolet irradiation in human skin fibroblasts. *Pflügers Archiv* 438: 239-244.
- Gusfield D (1997). Algorithms on strings, trees and sequences: computer science and computational biology. Cambridge university press.

- Hoy MB (2010). Wolfpfram| Alpha: a brief introduction. *Medical reference services quarterly* 29: 67-74.
- Loewe L, Swain P, Gutenkunst R (2013). *Mathematical Tools for Evolutionary Systems Biology*.
- Penny D, Hendy MD, Steel MA (1992). Progress with methods for constructing evolutionary trees. *Trends in Ecology & Evolution* 7: 73-79.
- Rokas A (2011). Phylogenetic analysis of protein sequence data using the Randomized Axelerated Maximum Likelihood (RAXML) Program. *Current protocols in molecular biology* 19.11. 11-19.11. 14.
- Saif R, Khan E, Azhar A, Choudhary S, Hussain T, Babar ME, Awan AR, Tayyab M, Zia S, Wasim M (2016). Insight of Tp53 Mutations and their effect on Protein in Different Feline and Canine Neoplasms. *Advancements in Life Sciences* 3: 42-50.
- Satpathy R, Behera R, Padhi SK, Guru RK (2013). Computational phylogenetic study and data mining approach to laccase enzyme sequences. *Journal of Phylogenetics & Evolutionary Biology*.
- Schilling CH, Schuster S, Palsson BO, Heinrich R (1999). Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnology progress* 15: 296-303.
- Soltis DE, Soltis PS (2003). The role of phylogenetics in comparative genetics. *Plant Physiology* 132: 1790-1800.
- Tatusova TA, Madden TL (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS microbiology letters* 174: 247-250.
- Vidyasagar A, Wilson NA, Djamali A(2012). Heat shock protein 27 (HSP27): biomarker of disease and therapeutic target. *Fibrogenesis & tissue repair* 5:7.