BMC Plant Biology

DATABASE Open Access

sHSPdb: a database for the analysis of small Heat Shock Proteins

Emmanuel Jaspard[1,2,3*] and Gilles Hunault[4]

Abstract

Background: small Heat Shock Proteins (sHSP) is a wide proteins family. SHSP are found in all kingdoms and they play critical roles in plant stress tolerance mechanisms (as well as in pathogenic microorganisms and are implicated in human diseases).

Results: sHSPdb (small Heat Shock Proteins database) is an integrated resource containing non-redundant, full-length and curated sequences of sHSP, classified on the basis of amino acids motifs and physico-chemical properties. sHSPdb gathers data about sHSP defined by various databases (Uniprot, PFAM, CDD, InterPro). It provides a browser interface for retrieving information from the whole database and a search interface using various criteria for retrieving a refined subset of entries. Physicochemical properties, amino acid composition and combinations are calculated for each entry. sHSPdb provides automatic statistical analysis of all sHSP properties. Among various possibilities, sHSPdb allows BLAST searches, alignment of selected sequences and submission of sequences.

Conclusions: sHSPdb is a new database containing information about sHSP from all kingdoms. sHSPdb provides a classification of sHSP, as well as tools and data for the analysis of the structure - function relationships of sHSP. Data are mainly related to various physico-chemical properties of the amino acids sequences of sHSP. sHSPdb is accessible at http://forge.info.univ-angers.fr/~gh/Shspdb/index.php.

Keywords: Small heat shock proteins, Amino acids sequence, Physico-chemical properties

Background

Heat Shock Proteins (HSP) are proteins whose expression is increased when cells are exposed to elevated temperatures or other stress. Among HSP, sHSP belong to the superfamily of protein chaperones. They counteract the irreversible aggregation of misfolded proteins. sHSP are ubiquitous proteins found in all living organisms, from bacteria to mammals and especially in plants [1–4]. Moreover, sHSP display a variety of subcellular localization and/or tissue distribution [3, 5].

sHSP contain : (i) a highly variable (in both length and in sequence) N-terminal domain; (ii) a very conserved central domain called the Alpha Crystallin Domain (ACD) involved in dimerization leading to the building block of higher order sHSP structure; (iii) a short C-terminal domain that links monomers inside dimmers [6–8]. A study of all sHSP merged have shown that the average length of ACD is 90 ± 10 residues [9] and this length is much more constant than the N- and C-terminal domains. However, we have classified sHSP into 21 classes and this value of average length of ACD does not apply anymore when considering sHSP class by class. The N-terminal part of ACD is likely not necessary for dimerization or chaperone activity, but seems required for higher order aggregates formation [10].

As members of the chaperone network, sHSP play important roles in cells exposed to heat and other stress, but their functional molecular mechanisms are not fully elucidated. sHSP monomers (12–42 kDa) usually associate into large homo-oligomers of mostly 24 subunits [9, 11]. sHSP quaternary structures have a high degree of plasticity due to changes in their oligomeric state under different cellular conditions or to exchange of their subunits. Loosening of subunit organization leads to more dynamic properties thus enhancing the available chaperone sites for the client

* Correspondence: emmanuel.jaspard@univ-angers.fr
[1] Université d'Angers, UMR 1345 IRHS, SFR 4207 QUASAV, Angers, France
[2] INRA, UMR 1345 IRHS, Beaucouzé, France
Full list of author information is available at the end of the article

proteins. For example, it was shown that the sequence RLFDQ (found at the N-terminal part of ACD of HSP27 - class number 12) contributes to the higher order assembly of their subunits as well as their structural stability [12]. Structural studies of numerous sHSP have brought details in monomer, transition of the oligomeric state [13–15] and sHSP chaperone mechanism [12, 16].

However, such approaches do not allow the comparison of more than a few sHSP at a time. On the contrary, sHSPdb offers access to large-scale analysis of sHSP. Moreover, since sHSP is a very large protein family there is a wide discrepancy in annotation, definition and terminology: sHSPdb provides a classification of sHSP based on selective motifs, physicochemical properties and previous classification [17–19]. sHSPdb has been conceived with multiple functions to search and describe the different sHSP. sHSPdb provides a global vision of sHSP physicochemical properties, amino acid usage, literature as well as statistical analysis on any sHSP dataset built by the user. Such a resource will be invaluable for computational analyses of sHSP structure - function relationships, especially regarding the specific roles of each of the three domains. sHSPdb is useful to gather information about sHSPs from all kingdom. sHSPdb provides the largest set of sHSP sequences, a new classification and a precise description of each class of sHSP. Physico-chemical properties and statistical analysis provided by sHSPdb represent an actual plus-value.

Construction and content

Data sources and characteristics of the dataset

To fill sHSPdb, we used a two-stage process. The first step filled automatically the fields of the tables of the database using PHP and perl scripts, getting the information from text and XML files from the following databases: NCBI/Proteins, NCBI/CDD, NCBI/Taxonomy, EBI/Picr, UNIPROT, Interpro, SANGER/Pfam, AMIGO. Starting with the NCBI accession number, we retrieved the « *GenPept* » file to derive the GI number, the eventual PFAM, CDD and Interpro [8, 20, 21] identifiers and the textual information. Then, using the cross-references of EBI/Picr we obtained the Uniprot accession number and name to complete and double-check the data. The main sources to fill sHSPdb are information contained in « *GenPept* » files and in the corresponding Uniprot files. The primary request to get « *GenPept* » files from the public database NCBI contained a construction of Boolean and keywords. This allowed the retrieving of ≈ 17,300 « *GenPept* » files stored in sHSPdb. Nevertheless, there was a very high redundancy of sequences for many organisms. The second step was a manual check to ensure that all information and links were pertinent and relevant to the sHSP issue. The pertinence of the files annotation and that of the sequences were checked using: (i) the sHSPdb « *Search* » option (see below); (ii) BLAST homology analysis and multiple alignments software; (iii) MOCAR (http://forge.info.univ-angers.fr/~gh/wstat/Mocar/), a program developed specifically for our sHSP classification based on pattern motif recognition in any Fasta sequences dataset; (iv) the « *Analyze* » and the « *Statistical analysis* » tools implemented in sHSPdb; (v) Finally, the motifs determined for each class of sHSP were further used to perform PHI-Blast allowing to get additional sequences.

Sequence conflict (in particular in the case of AGI entries for *Arabidopsis thaliana*) was checked using annotations from the Uniprot files. Identical sequences were removed using a program developed for sHSPdb feeding. This allowed detecting truncated, ambiguous (e.g., annotated « unnamed protein », « hypothetical », « unknown product »..., although some few populated sHSP classes may contain such annotated files) or false sHSP sequences. Sequences containing the undefined amino acid symbol (X) were removed. All files are stored in sHSPdb but only non-redundant, full-length and annotated ones are accessible.

The two holding chaperones (holdases) families, i.e., HSP31 and HSP33, cannot be considered formally as sHSP since they do not harbor the ACD domain, and interact with client proteins through Cys and disulfide bonds formation [22]. However, these redox regulated molecular chaperones have physico-chemical properties close to those of sHSP and also contribute to protein homeostasis under stress conditions. Moreover, they are redox regulated molecular chaperones: they protect both thermally unfolded and oxidatively damaged proteins from irreversible aggregation and play an important role in the bacterial defense system toward oxidative stress. Therefore, the two classes HSP31 and HSP33 were incorporated into sHSPdb.

The main goal of sHSPdb is the analysis of the structure-function relationships of sHSP. This is why only amino acids sequences are implemented. Today the user can nevertheless BLASTX any nucleotide sequence against sHSPdb to check the existence of similar or homologous sHSP.

Finally, sHSPdb contains ≈ 4800 curated, non-redundant and full-length sHSP sequences. More than 3200 sequences (67 %) are classified into 21 classes on the basis of their corresponding unique sequence motif. Roughly 400 additional sequences are classified on the basis of various criteria, eye inspection and our expertise. The remaining unclassified sHSP sequences have been assigned to a class 99 with ≈ 530 of these sequences being annotated Hsp20. Even if we expect to which class sequences of class 99 belong, they are not yet classified because they do not match exactly the regular expression of any of the defined motifs. Since there is no program available to automatically determine such complex discriminants motifs for any set of sequences, we are currently developing appropriate methods to address this question. In particular, we are using the recent constraint programming approach, based on a two-stage program whose last stage reduces to the classical minimum set covering problem, though, as for now, the complete and « perfect » characterization of all classes is not achieved. Partials results have already been published [23, 24].

Sequences are mainly issued from ≈ 2355 Bacteria, ≈ 2259 Eukaryota (among which ≈ 1050 Viridiplantae) and ≈ 129 Archaea (Table 1).

sHSPdb contains also ≈ 14,240 non-accessible entries (i.e., not retained after our curation). This dataset is important when the database is updated and for further analyses of sHSPdb since it may serve for statistical studies and forbid wrong entries in the curated set.

Organization and main features of sHSPdb

A link to a guided tour is proposed in the homepage to learn how to use the interface of sHSPdb. This guided tour proposes various scenarii of search. The user can select any entries through various parameters and conduct further analyses using the implemented tools. For this purpose, sHSPdb has three main features: (i) the browse mode that allows the user to consult all or part of the database; (ii) the search mode based upon multiple search criteria; (iii) the export mode to retrieve sequences in different formats.

The browse mode

It allows consulting the whole database (Fig. 1). The « *Summary* » option provides the NCBI-GenPept accession number and the Uniprot accession number, the name of the sequence and of the organism, the putative function of the sHSP (if any). The « *Details* » option provides more information (up to 20 fields from the GenPept or the Uniprot files). The accession numbers, the name of the organism, the PFAM, the CDD and the InterPro numbers provide a link to their relevant website. A series of physicochemical properties (length/pI/MW/FoldIndex/GRAVY/mean net charge at pH7/mean hydrophilicity/mean hydrophobicity < H>/flexibility/bulkiness/accessibility of residues/transmembrane tendancy of residues) are given by selecting the « *Physicochemical properties* » option. The « *AA comp* » option displays the amino acid composition of sHSP. Finally, the « *Fasta* » option provides all sequences in Fasta format. The « *Align sequences* » option allows to align selected sequences.

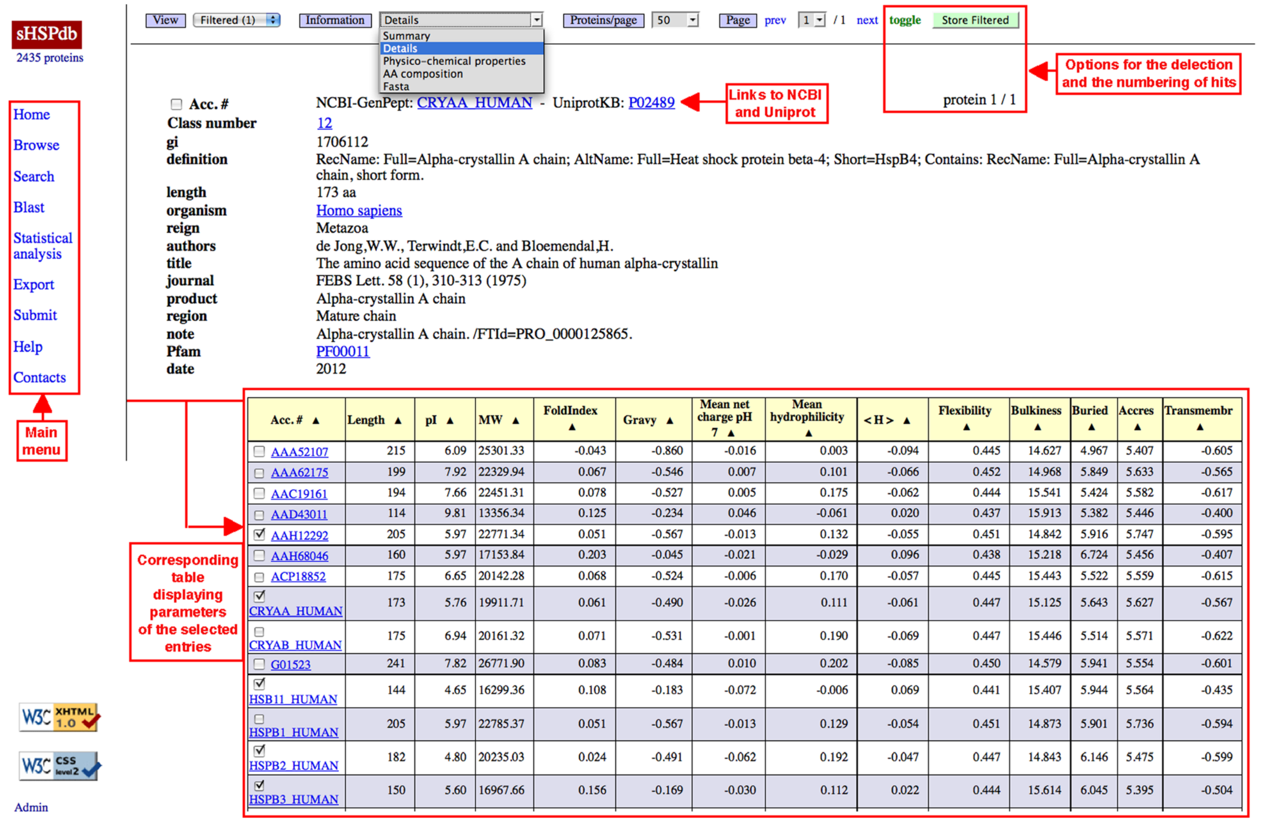
The search mode (multiple search capabilities)

It corresponds to an advanced search with multiple parameters and permits retrieval of very fine subsets of data (example of search: « sHSP of the PF00011 family from *Arabidopsis thaliana*, containing the motif of Class 1, associated to the key word « chromosome », published by an author called Walker in 2012, with a length comprised between 150 and 200 amino acids »). The search can be made by using a sophisticated text expression accepting wildcards and Booleans - this search applies to all fields of the whole database; by accession number (accepting wildcards) - one or multiple accession number(s) separated by a space can be searched at one time; by organism; by PFAM or CDD numbers; by date; by range of length of amino acids sequence. The user can also retrieve sequences by entering amino acids motif either exact or degenerated using regular expressions with sophisticated syntax, allowing to specify complex motifs. The known motifs (Table 2) implemented in sHSPdb result from our analysis.

The user can perform a similarity search by using BLASTP implemented in sHSPdb. This allows obtaining additional information and retrieving the best scoring sequences through the classical BLAST output interface. It is also possible to BLASTX a nucleotide sequence against sHSPdb to check the existence of similar or homologous sHSP in order, for example, to construct a putative sHSP cDNA sequence from EST.

Table 1 Main taxonomy of the organisms in sHSPdb. Only groups containing more than 10 sequences are indicated

129 Archaea	2355 Bacteria	2259 Eukaryota	
21 Crenarchaeota	198 Actinobacteria	1180 Opisthokonta	
77 Euryarchaeota	11 Aquificae	1051 Viridiplantae	
31 Thaumarchaeota	95 Bacteroidetes	Opisthokonta	Viridiplantae
	12 Chlamydiae		
	22 Chloroflexi	203 Fungi	12 Funariaceae
	111 Cyanobacteria	977 Metazoa	49 Pinales
	14 Deinococcus		285 Liliopsida
	279 Firmicutes		687 Eudicotyledons
	15 Planctomycetes		
	1501 Proteobacteria		
	39 Spirochaetes		
	14 Thermotogae		



□ Acc. #	NCBI-GenPept: CRYAA_HUMAN - UniprotKB: P02489
Class number	12
gi	1706112
definition	RecName: Full=Alpha-crystallin A chain; AltName: Full=Heat shock protein beta-4; Short=HspB4; Contains: RecName: Full=Alpha-crystallin A chain, short form.
length	173 aa
organism	Homo sapiens
reign	Metazoa
authors	de Jong,W.W., Terwindt,E.C. and Bloemendal,H.
title	The amino acid sequence of the A chain of human alpha-crystallin
journal	FEBS Lett. 58 (1), 310-313 (1975)
product	Alpha-crystallin A chain
region	Mature chain
note	Alpha-crystallin A chain. /FTId=PRO_0000125865.
Pfam	PF00011
date	2012

Acc. # ▲	Length ▲	pI ▲	MW ▲	FoldIndex ▲	Gravy ▲	Mean net charge pH 7 ▲	Mean hydrophilicity ▲	<H> ▲	Flexibility ▲	Bulkiness ▲	Buried ▲	Accres ▲	Transmembr ▲
□ AAA52107	215	6.09	25301.33	-0.043	-0.860	-0.016	0.003	-0.094	0.445	14.627	4.967	5.407	-0.605
□ AAA62175	199	7.92	22329.94	0.067	-0.546	0.007	0.101	-0.066	0.452	14.968	5.849	5.633	-0.565
□ AAC19161	194	7.66	22451.31	0.078	-0.527	0.005	0.175	-0.062	0.444	15.541	5.424	5.582	-0.617
□ AAD43011	114	9.81	13356.34	0.125	-0.234	0.046	-0.061	0.020	0.437	15.913	5.382	5.446	-0.400
☑ AAH12292	205	5.97	22771.34	0.051	-0.567	-0.013	0.132	-0.055	0.451	14.842	5.916	5.747	-0.595
□ AAH68046	160	5.97	17153.84	0.203	-0.045	-0.021	-0.029	0.096	0.438	15.218	6.724	5.456	-0.407
□ ACP18852	175	6.65	20142.28	0.068	-0.524	-0.006	0.170	-0.057	0.445	15.443	5.522	5.559	-0.615
☑ CRYAA_HUMAN	173	5.76	19911.71	0.061	-0.490	-0.026	0.111	-0.061	0.447	15.125	5.643	5.627	-0.567
□ CRYAB_HUMAN	175	6.94	20161.32	0.071	-0.531	-0.001	0.190	-0.069	0.447	15.446	5.514	5.571	-0.622
□ G01523	241	7.82	26771.90	0.083	-0.484	0.010	0.202	-0.085	0.450	14.579	5.941	5.554	-0.601
☑ HSB11_HUMAN	144	4.65	16299.36	0.108	-0.183	-0.072	-0.006	0.069	0.441	15.407	5.944	5.564	-0.435
□ HSPB1_HUMAN	205	5.97	22785.37	0.051	-0.567	-0.013	0.129	-0.054	0.451	14.873	5.901	5.736	-0.594
☑ HSPB2_HUMAN	182	4.80	20235.03	0.024	-0.491	-0.062	0.192	-0.047	0.447	14.843	6.146	5.475	-0.599
☑ HSPB3_HUMAN	150	5.60	16967.66	0.156	-0.169	-0.030	0.112	0.022	0.444	15.614	6.045	5.395	-0.504

Fig. 1 Browse interface for accessing the whole sHSPdb content. The « Search » mode provides an advanced search with multiple parameters and permits retrieval of very fine subsets of data. The user can also retrieve sequences by entering amino acids motif either exact or degenerated using regular expressions with sophisticated syntax. Among other information, the « Summary » option provides links to general databases, the name of the sHSP and its putative function. The « Details » option provides more information. A series of physicochemical properties are given by selecting the « Physicochemical properties » option. The « AA comp » option displays the amino acid composition of sHSP. Finally, the « Fasta » option provides all sequences in FASTA format

Many options are proposed for the output: (i) when entries have been retrieved, the information can be displayed in any of the views of the browse mode and the fields such as accession number, taxonomy, PFAM, CDD and InterPro numbers are linked to their original web site according to the relevant information found at the NCBI; (ii) the user can select or deselect any entry resulting from his search (« *Toggle* » option) and the final selection can be stored. This is true for all displays (browse mode, BLAST output, …); (iii) one original point is the re-ordering (ascending or descending sort) of data displayed in tables, for a better comparison and analysis of data, using any shown information as the sort criterion; (iv) the user can automatically align selected sequences.

The export mode

All sequences in SHSPdb can be exported trough the « *Export* » interface. The « *View* » menu displays the entries in the order of the last actions made by the user (all, filtered, selected or blasted). The « *Format* » menu displays three exporting formats: FASTA, XML and Excel (CSV).

The analyze tool

One original feature of sHSPdb is the possibility to obtain numerous information of any self-built dataset using the left panel. The right panel allows selecting the desired information. For categorical data, results are sorted tables of counts, respectively with respect to alphabetic order and values of counts; for quantitative data, classical descriptive statistics and graphics (such as means, median, histograms…) are provided.

The statistical analysis tool

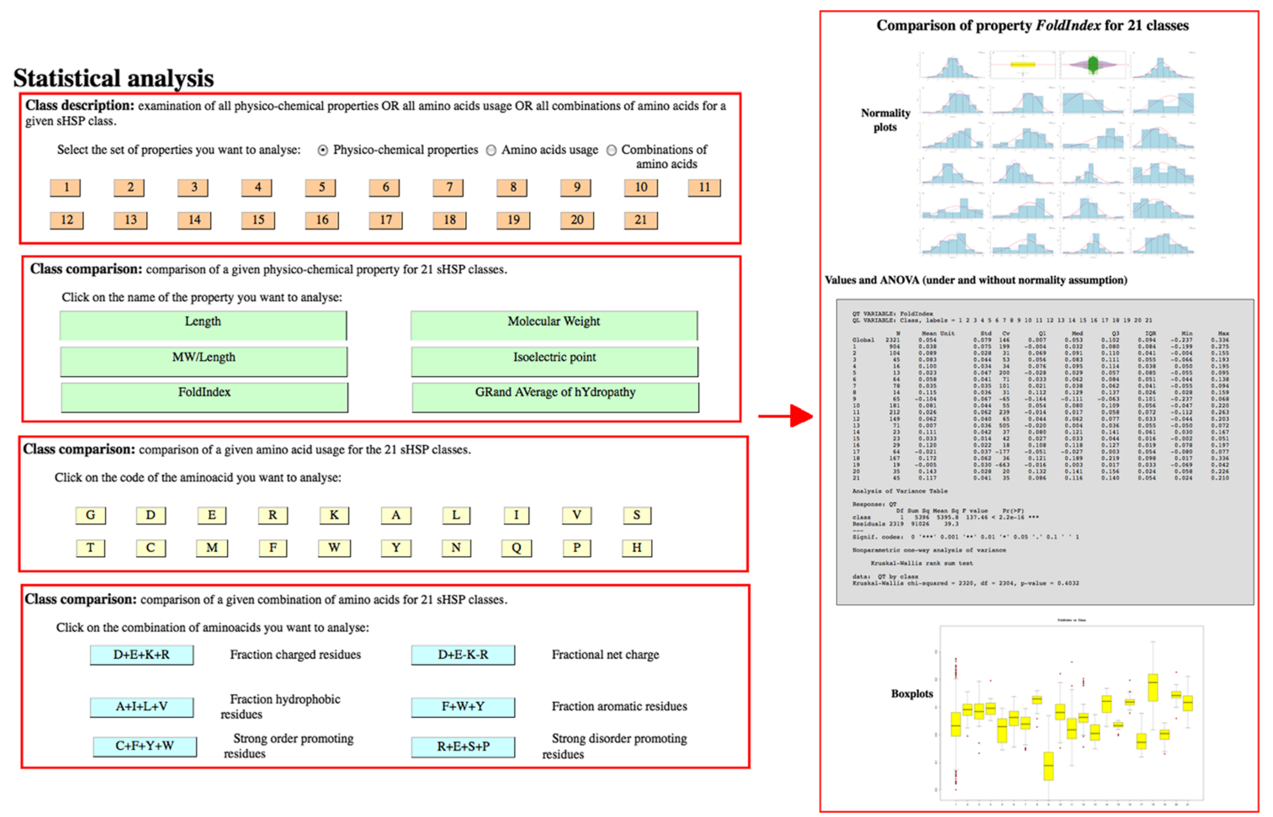
Another original feature of sHSPdb is the possibility to dynamically generate plots of physico-chemical properties or amino acids usage of sHSP using different criteria. Therefore, one can immediately make multiple comparisons of self datasets. The user can access to statistical data through the menu « *Statistical analysis* ». This option displays « *Class description* » (examination

Table 2 sHSP classification in sHSPdb

sHSP class (Usual name)	Localization (plants sHSP)	Examples of representant and/or annotation	Motif allowing to classify sequences[a]	Motif position in sequences	Number of sHSP classified[b]	Main taxonomy
Class 1 (Class I)	Cytosol-nucleus	TaHsp16.9 PsHsp18.1	[AGPSTV][ADEGHMNQTV][ILPSVY][KR][AFGNPQST][IV][ADEHKNQ][IFV][ACGNSTY][ADEGV]?(.{1,6})?$	C-terminal	362	Viridiplantae
Class 2 (Class II)	Cytosol-nucleus	17–17.9 kDa HSP	[ILNV].{0,1}[KNR].{0,1}[PT]P[APQ].{1,2}[AKR].{0,6}[IV].{0,5}$	C-terminal	111	Viridiplantae
Class 3 (Class_IV)	Endoplasmic reticulum		WH.{8,10}[FS][IW]R[QR][FL].{2,2}P.{3,3}[^K]	Central	68	Viridiplantae
Class 4 (Class_V)			[^E]A[RS]AA$	C-terminal	46	Viridiplantae
Class 5 (Class_VI)		21.7–22.3 kDa	D.{2,3}D.{1,2}PLW	N-terminal	64	Viridiplantae
Class 6 (Class M)	Mitochondria	HSP23	[LM].{10,10}[KR].{1,2}[DEKQ][DEKNQ][DEGHPQ].{2,2}[DEN].{3,3}[IV].[^H][DENQ]$	C-terminal	118	Viridiplantae
Class 7 (Class P)	Plastids	HSP26 to HSP26.8 CPsHSP21	[DEFHNQSTY][AGKPRST][KT][IV][FHILMTV][DEG][IRV][DEHKNQS][ILV][EKNQR]G?$	C-terminal	102	Viridiplantae
Class 8 (Class Px)	Peroxysome		[IV].{3,3}[^D][KR]L$	C-terminal	52	Viridiplantae
Class 9 (Class RA)	Ribosome Associated	sHSP15 RNA-binding S4	P[^P]K.{2,2}R[QR].{7,27}$	C-terminal	247	Bacteria
Class 10 (HSP20/IbpA/IbpB)		Hsp20 - IbpA/B - HspA/B - HspB6	P[GKP][FHMY][DN][ILV]	N-terminal	545	Bacteria Fungi
Class 11 (HSP20/HspC)			[FILMV][^P].?[DEQ][FILMPV][ADEKNQS][^S][FILMV][FLVW][ADEGHIMNQS]	N-terminal	679	Bacteria Archaea Fungi
Class 12 (HSP27 /aA-crystallin /aB-crystallin)		HspB1 aA or aB -crystallin HspB4 or 5	[AGHNPQRSTY][DHKR][AFILMPV].[DES][DQ].F[AG]	N-terminal	534	Metazoa Bacteria
Class 13 (beta-9)			[LM][LP][SV].{2,2}L.[ADEN].{11,14}F[KQ]	N-terminal	56	Metazoa
Class 14 (HSP Beta_3)			PVRY[EKQ]	N-terminal	84	Metazoa
Class 15 (HSP Beta_7)			EIKI$	C-terminal	26	Metazoa
Class 16 (HSP Beta_11)		HSP family B	[ADE]TFW	N-terminal	76	Metazoa
Class 17 (HspQ)			K[FY][AFG][IL]G	N-terminal	88	Bacteria
Class 18 (HSP42)		Hsp42p	DEEL.{1,16}$	C-terminal	16	Fungi
Class 19 (HSP30)			[PST][AEIV][HLR][PST][LH]W[PT][AEHT]	N-terminal	24	Metazoa
Class 20 (HSP31)		HchA « Holdases »	TG.{4,4}E.{11,11}G	N-terminal	147	Bacteria Fungi
Class 21 (HSP33)		33 kDa chaperonin « Holdases »	C.[FY]C.{3,3}Y	C-terminal	135	Bacteria

[a] Sequences classified via the motif of the class. Regular expression syntax for amino acid motifs: [XYZ] means X or Y or Z; X? means X is optional; [^XY] means not X nor Y; $ means C-terminal position;.means any amino acid; X{n,m} means n to m times X. [b] Number of sequences retrieved in sHSPdb using the motif indicated

of all physico-chemical properties or all amino acids usage or all combinations of amino acids for a given sHSP class) or « *Class comparison* » (comparison of a given physico-chemical property for the sHSP classes). Plots are automatically generated and are clickable to be resized (Fig. 2). The statistical analysis includes both descriptive (means, medians…) and inferential (parametric or non parametric ANOVA) computations and graphics



Statistical analysis

Class description: examination of all physico-chemical properties OR all amino acids usage OR all combinations of amino acids for a given sHSP class.

Select the set of properties you want to analyse: ● Physico-chemical properties ○ Amino acids usage ○ Combinations of amino acids

Class comparison: comparison of a given physico-chemical property for 21 sHSP classes.

Click on the name of the property you want to analyse:

Class comparison: comparison of a given amino acid usage for the 21 sHSP classes.

Click on the code of the aminoacid you want to analyse:

Class comparison: comparison of a given combination of amino acids for 21 sHSP classes.

Click on the combination of aminoacids you want to analyse:

D+E+K+R
Fraction charged residues
D+E-K-R
Fractional net charge
A+I+L+V
Fraction hydrophobic residues
F+W+Y
Fraction aromatic residues
C+F+Y+W
Strong order promoting residues
R+E+S+P
Strong disorder promoting residues

Comparison of property *FoldIndex* for 21 classes

Values and ANOVA (under and without normality assumption)

```
QT VARIABLE: FoldIndex
QL VARIABLE: Class, labels = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

              N      Mean Unit        Std   Cv        Q1       Med        Q3       IQR       Min       Max
Global     2321     0.054            0.079  146     0.007     0.053     0.102     0.094    -0.237     0.336
1           904     0.038            0.075  199    -0.004     0.032     0.080     0.084    -0.199     0.275
2           104     0.089            0.028   31     0.069     0.091     0.110     0.041    -0.004     0.155
3            45     0.083            0.044   53     0.056     0.083     0.111     0.055    -0.066     0.193
4            16     0.100            0.034   34     0.076     0.095     0.114     0.038     0.050     0.195
5            13     0.023            0.047  200    -0.028     0.029     0.057     0.085    -0.055     0.095
6            64     0.058            0.041   71     0.033     0.062     0.084     0.051    -0.044     0.138
7            78     0.035            0.035  101     0.021     0.038     0.062     0.041    -0.055     0.094
8            14     0.115            0.036   31     0.112     0.129     0.137     0.026     0.028     0.159
9            65    -0.104            0.067  -65    -0.164    -0.111    -0.063     0.101    -0.237     0.068
10          181     0.081            0.044   55     0.054     0.080     0.109     0.056    -0.047     0.220
11          212     0.026            0.062  239    -0.014     0.017     0.058     0.072    -0.112     0.263
12          149     0.062            0.040   65     0.044     0.062     0.077     0.033    -0.044     0.203
13           71     0.007            0.036  505    -0.020     0.004     0.036     0.055    -0.050     0.072
14           23     0.111            0.042   37     0.080     0.121     0.141     0.061     0.030     0.167
15           23     0.033            0.014   42     0.027     0.033     0.044     0.016    -0.002     0.051
16           29     0.120            0.022   18     0.108     0.118     0.127     0.019     0.078     0.197
17           64    -0.021            0.037 -177    -0.051    -0.027     0.003     0.054    -0.080     0.077
18          167     0.172            0.062   36     0.121     0.189     0.219     0.098     0.017     0.336
19           19    -0.005            0.030 -663    -0.016     0.003     0.017     0.033    -0.069     0.042
20           35     0.143            0.028   20     0.132     0.141     0.156     0.024     0.058     0.226
21           45     0.117            0.041   35     0.086     0.116     0.140     0.054     0.024     0.210

Analysis of Variance Table

Response: QT
            Df Sum Sq Mean Sq F value    Pr(>F)
class        1   5396  5395.8  137.46 < 2.2e-16 ***
Residuals 2319  91026    39.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Nonparametric one-way analysis of variance

     Kruskal-Wallis rank sum test

data:  QT by class
Kruskal-Wallis chi-squared = 2320, df = 2304, p-value = 0.4032
```

Fig. 2 The statistical analysis tool. One original feature of sHSPdb is the possibility to dynamically generate plots of physico-chemical properties or amino acids usage of sHSP using different criteria. Therefore, one can immediately make multiple comparisons of self datasets. The user can access to statistical data through the menu « *Statistical analysis* ». This option displays « *Class description* » (examination of all physico-chemical properties or all amino acids usage or all combinations of amino acids for a given sHSP class) or « *Class comparison* » (comparison of a given physico-chemical property for the sHSP classes). Plots are automatically generated and are clickable to be resized

(histograms with density and normal approximation, boxplots, beanplots when accurate…).

The submission form

The user can submit sequence(s) through a submission form. After verification of their relevance, they will be integrated in sHSPdb.

Database implementation

sHSPdb is a fast, interactive, platform independent Web-based database with a user-friendly interface written in PHP. Some options need Javascript to be activated. Information is provided to the user from a MySQL relational database. The statistical computations are performed using the R software called by a PHP script.

Utility and discussion

sHSPdb versus other sHSP datasets

Poulain et al. [9] reported about a dataset including 3787 sHSP sequences. However, this dataset contains some redundancy: 774 sequences from this dataset were aligned and 70 sequences are identical, corresponding to a redundancy of 9 %. It can be thus estimated that the entire dataset correspond roughly to 3400 different sHSP, i.e., less than the number of sHSP contained in sHSPdb. Moreover, this dataset is limited to a flat PDF file with accession numbers and thus does not provide any details or functionalities. Our selection is very stringent since we excluded most of all files annotated « putative », « probable », « unknown », « hypothetical », « unnamed protein product », « synthetic construct », … in order to increase the quality of annotation even if less sequences were retained. We did not consider fragments, i.e., files annotated « partial ». Sequences with such annotation but additional indication and/or classifiable on the basis of a motif were retained.

To our knowledge, there is only one other database dedicated to HSP (HSPIR: http://pdslab.biochem.iisc.ernet.in/hspir/shsp.php) which gather accession numbers and FASTA files for HSP protein sequences, and provides general information about the structure and function of the different families of HSPs [25]. However, HSPIR contains only ≈ 1300 sequences of sHSP, and offers access to only basic search and BLAST tools. sHSPdb appears thus as a unique comprehensive and highly implemented database for the computational and statistical analysis of sHSP.

Classification of sHSP

sHSP constitute a very large protein sub-divised family of proteins. However, there is a wide discrepancy in annotation, definition and terminology. sHSPdb provides a unified and homogeneous classification of sHSP. One of the most important parameter for the sHSP classification is the consensus motif discriminating each class from another (Table 2). The proposed classification in classes is based on the expertise brought by sHSPdb, literature data and previous classifications [1, 2, 9, 17, 19, 26–30].

The quite simple C-terminal motif [ADHKNQ]G.[AILV] matches ≈ 89 % of the sequences and a more sophisticated central motif [APS][EGM].{4,5}[ADENQSK] matches ≈ 99 % of the sequences. Such motifs therefore describe almost all sHSP and are therefore not pertinent for discrimination. Since most of sHSP classes contain the ACD (with exception of sHSP classes 9, 16, 17 and 21), we searched selective motifs either in the N-terminal or the C-terminal domains, if any (Table 3).

This motifs search underlined the difficulties with annotation. For example, the motif [AFLSTVWY]P[AGPS][FHY][DN][ILV] matches ≈ 540 sequences whose annotations are mainly « heat shock protein HSP20, « HSP20 family protein » or « Hsp20/alpha crystallin family protein ». Therefore, the high selectivity of our motifs helps to ameliorate global annotation such as the generic « HSP20 ».

Previous classifications were first based on the intracellular localization of the sHSP: cytoplasmic-nuclear compartment (classes CI, CII and CIII), plastids (class P), endoplasmic reticulum (class ER), mitochondria (class M), peroxisomal (class Po or Px). These seven plant sHSP classes are to date the best characterized of the plant sHSP (they are coded by 14 genes in *Arabidopsis thaliana*). The complete analysis of *Arabidopsis thaliana* genome led to four additional cytoplasmic-nuclear sHSP and a second unique family of sHSP targeted to the mitochondria. Thus, higher plants may contain at least 12 conserved sHSP subfamilies. Organelle-targeted sHSP are so far specific of plants, with some exceptions (mitochondria-targeted sHSP

Table 3 Some length characteristics of sHSP and « alpha-crystalin-unrelated sHSPs »

sHSP class	Sequences length range (min - max)	Motif position[a]	Mean limit values of domains length		
			N-terminal	ACD[b]	C-terminal
Class 1	130–282	132–273	1–50	51–141	142–157
Class 2	129–175	131–151	1–49	50–139	140–156
Class 3	165–328	90–255	1–72	73–164	165–198
Class 4	119–211	115–207	1–42	43–123	124–148
Class 5	172–254	41–118	1–86	87–182	183–194
Class 6	101–279	185–212	1–120	121–202	203–218
Class 7	157–266	215–229	1–129	130–222	223–235
Class 8	115–183	109–177	1–28	29–127	128–144
Class 9	100–175	98–127	*NA*[c]	*NA*	*NA*
Class 10	120–498	8–200	1–40	41–137	138–160
Class 11	111–356	6–109	1–51	52–139	140–156
Class 12	108–374	14–214	1–59	60–151	152–176
Class 13	146–277	26–152	1–53	54–142	143–167
Class 14	141–187	13–54	1–63	64–145	146–152
Class 15	108–262	105–259	1–79	80–160	161–178
Class 16	114–230	34–80	*NA*	*NA*	*NA*
Class 17	102–131	5–6	*NA*	*NA*	*NA*
Class 18	332–453	317–440	1–256	257–349	350–384
Class 19	191–220	6–26	1–81	82–165	166–211
Class 20	133–298	54–163	1–42	43–133	134–156
Class 21	161–344	136–320	*NA*	*NA*	*NA*

[a] Range of starting position of the motif among all sequences of the class. [b] Alpha Crystallin Domain. [c] Not applicable since these HSP do not contain ACD and are defined as « alpha-crystallin-unrelated sHSPs »

from *Drosophila melanogaster* and from *Toxoplasma gondii*) [31, 32]. However, it remains difficult to ascertain the sub-cellular localization of proteins, especially on large numbers of proteins because predictors are not reliable enough. Finally, the taxonomy of organism among each class is homogeneous indicating that our classification of sHSP is accurate (Table 2).

Such classification together with the statistical data automatically generated is thus useful. For example, sHSPdb help for analyzing the respective role of each of the three domains characterizing sHSP in oligomerization and substrate binding on a very large scale. A conserved motif, [ILV].[ILV], located in the C-terminal domain is critical for oligomer formation through contacts with a hydrophobic patch in ACD [9, 13, 33]. The ACD domain is an emblematic signature of sHSP in the literature [19, 28] and it is indeed found in 17 of the 21 classes in sHSPdb. Table 3 and Fig. 3 show the comparison of the mean value of the size of the N-terminal domain, ACD and the C-terminal domain.

Nevertheless, some heat shock proteins lacking an ACD sequence should be in our opinion considered as sHSP and defined as « alpha-crystallin unrelated sHSP ». Therefore, we have implemented 4 classes of sHSP without ACD. This classes of sHSP devoid of the ACD domain fall well within the molecular weight expected for sHSP. Their implementation provides additional information and allows more comparisons and analyses of sHSP.

Short example of use of sHSPdb: analysis of mitochondrial sHSP

A phylogenetic tree of plants mitochondrial sHSP (class 6) has been built (Fig. 4). Five sequences correspond to mitochondrial sHSP from Arabidopsis thaliana (AAM63747, CAA67022, HS235_ARATH, HS23M_ARATH and HS26M_ARATH, all from class 6). The differences between AAM63747 and HS235_ARATH are [S/A]31 and [D/H]69 and the difference between CAA67022 and HS23M_ARATH is [D/H]143. The sequences of ACD and C-terminal domain are very similar (Fig. 5). The sequences of the N-terminal domain of HS235_ARATH and HS23M_ARATH are different from that of HS26M_ARATH and are more intrinsically disordered (negative FoldIndex for HS26M_ARATH). This may have a possible incidence on the dimer formation [34, 35].

Conclusions

sHSPdb harbors a comprehensive dataset available for sHSP, together with tools designed for their online analysis. To our knowledge, there is no equivalent database for sHSP. sHSP are classified into classes on the basis of various parameters, especially on the basis of amino acids motifs that discriminate the classes. sHSPdb thus

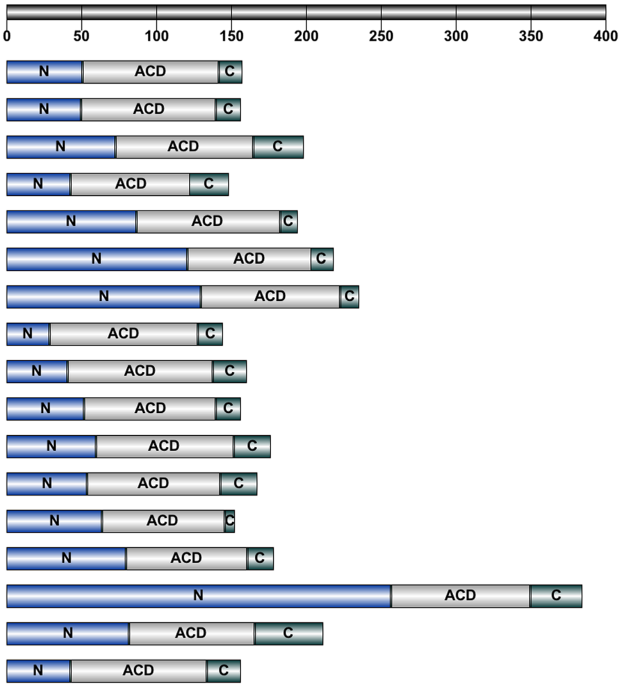


Fig. 3 Comparison of the mean value of the size of the N-terminal domain, ACD and the C-terminal domain. Only HSP classes containing an ACD are represented (see Table 3). The upper scale indicates the size of the sequences in amino acids. ER: endoplasmic reticulum

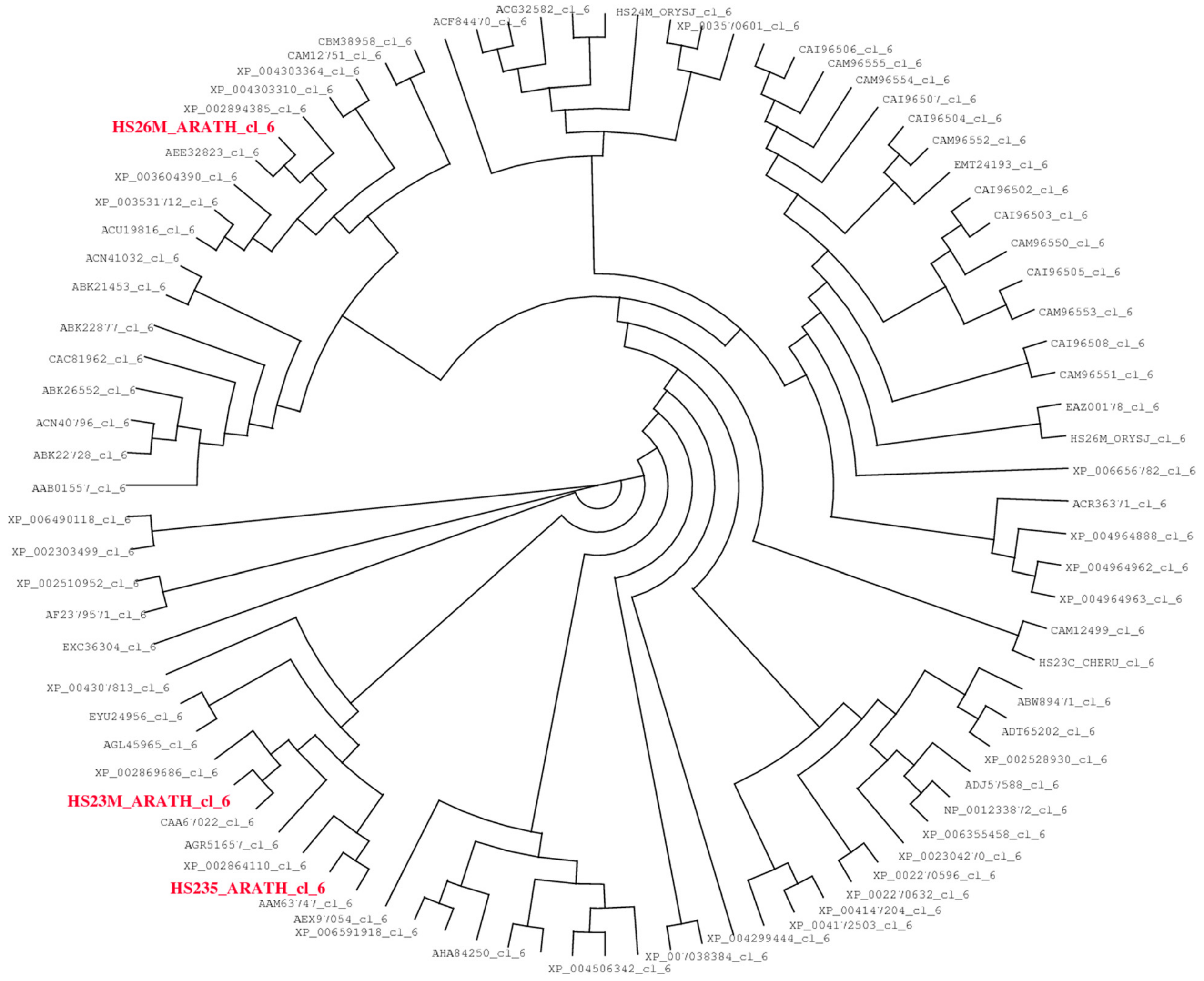


Fig. 4 Phylogenetic tree of plant mitochondrial sHSP. 91 sequences of mitochondrial sHSP from 47 plant organisms were analyzed. The characteristics of the sequences are: a length from 187 to 248 amino acids; an isoelectric point from 4.620 to 10.15; a net charge at pH 7 from – 0.041 to +0.031; a FoldIndex from – 0.044 to 0.138; a hydropathy (GRAVY) from – 0.908 to – 0.274. The alignment was made using Clustal-W and the figure was drawn using Dendroscope [36]

constitute an efficient tool: (i) for the compilation and the organization of growing data concerning sHSP; (ii) for the classification of the various sub-families of sHSP; (iii) for the design of experiments to elucidate the function of this important proteins; (iv) to help the analysis of the sHSP structure-function relationships.

Future developments and perspectives: (i) sHSP physico-chemical properties and sHSP amino acids usage are statistically analyzed for all sHSP classes. We will thus be able to compare the three domains (i.e., the N-terminal, the ACD and the C-terminal), thus bringing additional information to those already determined

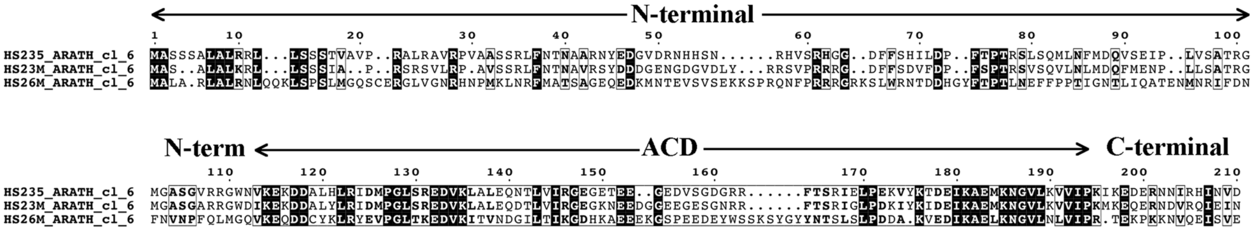

N-terminal
1 10 20 30 40 50 60 70 80 90 100
HS235_ARATH_cl_6 MASSSALALRRL...LSSSTVAVP..RALRAVRPVAASSRLFNTNAARNYEDGVDRNHHSN......RHVSRHGG..DFFSHILDP..FTPTRSLSQMLNFMDQVSEIP..LVSATRG
HS23M_ARATH_cl_6 MAS..ALALKRL...LSSSIA..P..RSRSVLRP.AVSSRLFNTNAVRSYDDDGENGDGVDLY...RRSVPRRRG..DFFSDVFDP..FSPTRSVSQVLNLMDQFMENP..LLSATRG
HS26M_ARATH_cl_6 MALA.RLALRNLQQKLSPSLMGQSCERGLVGNRHNPMKLNRFMATSAGEQEDKMNTEVSVSEKKSPRQNFPRRRGRKSLWRNTDDHGYFTPTLNEFFPPTIGNTLIQATENMNRIFDN
N-term
ACD
C-terminal
110 120 130 140 150 160 170 180 190 200 210
HS235_ARATH_cl_6 MGASGVRRGWNVKEKDDALHLRIDMPGLSREDVKLALEQNTLVIRGEGETEE..GEDVSGDGRR......FTSRIELPEKVYKTDEIKAEMKNGVLKVVIPKIKEDERNNIRHINVD
HS23M_ARATH_cl_6 MGASGARRGWDIKEKDDALYLRIDMPGLSREDVKLALEQDTLVIRGEGKNEEDGGEEGESGNRR......FTSRIGLPDKIYKIDEIKAEMKNGVLKVVIPKMKEQERNDVRQIEIN
HS26M_ARATH_cl_6 FNVNPFQLMGQVKEQDDCYKLRYEVPGLTKEDVKITVNDGILTIKGDHKAEEEKGSPEEDEYWSSKSYGYYNTSLSLPDDA.KVEDIKAELKNGVLNLVIPR.TEKPKKNVQEISVE

Fig. 5 Alignment of three mitochondrial sHSP from *Arabidopsis thaliana*. The alignment was made using Multalin [37] and the figure was drawn using ESPript [38]

by structural methodologies. (ii) We are currently developing software for the analysis of sequence submitted by the users in order to predict if it belongs to any of the sHSP classes. (iii) Since deciphering the molecular functions of sHSP is a major issue, we will provide lexical tools (dictionaries by alphabetic order or occurrence or synonyms…) for a better semantic analysis of the words that describe the known elements of the function of sHSP. (iv) As previously noted, retained proteins that are not fully classified are under study with the help of some predicting values and of a constraint programming software under development.

Availability and requirements

sHSPdb is a free database and visualization tool open to all users with no login requirements and can be accessed at the following URL: http://forge.info.univ-angers.fr/~gh/Shspdb/index.php. The web tool is functional on all modern web browsing environments including Mozilla Firefox, Safari and Google Chrome.

Abbreviations

ACD, alpha crystallin domain; ER, endoplasmic reticulum; NA, not applicable; sHSP, small heat shock proteins; sHSPdb, small heat shock proteins database

Funding

This research was supported by a grant ANR Blanc MITOZEN (ANR-12-BSV8-0021-01) and by a grant from University of Angers (ACS13LES).

Availability of data and materials

The datasets supporting the conclusions of this article are available in the sHSPdb repository freely accessible at http://forge.info.univ-angers.fr/~gh/Shspdb/index.php.

Authors' contributions

EJ served as project advisor. GH implemented the MySQL database, PHP and Perl scripts, and the interface design. EJ collected data, checked data integrity and entered them into the database. GH and EJ wrote the manuscript, checked the accuracy of the database and web interface, read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Université d'Angers, UMR 1345 IRHS, SFR 4207 QUASAV, Angers, France. [2]INRA, UMR 1345 IRHS, Beaucouzé, France. [3]Agrocampus-Ouest, UMR 1345 IRHS, Angers, France. [4]Université d'Angers, Laboratoire d'Hémodynamique, Interaction Fibrose et Invasivité tumorale hépatique, UPRES 3859, IFR 132, F-49045 Angers, France.

Received: 26 February 2016 Accepted: 24 May 2016
Published online: 13 June 2016

References

1. Siddique M, Gernhard S, von Koskull-Döring P, Vierling E, Scharf KD. The plant sHSP superfamily: five new members in Arabidopsis thaliana with unexpected properties. Cell Stress Chaperones. 2008;13:183–97.
2. Sarkar NK, Kim YK, Grover A. Rice sHsp genes: genomic organization and expression profiling under stress and development. BMC Genomics. 2009;10:393.
3. Kampinga HH, Hageman J, Vos MJ, Kubota H, Tanguay RM, Bruford EA, Cheetham ME, Chen B, Hightower LE. Guidelines for the nomenclature of the human heat shock proteins. Cell Stress Chaperones. 2009;14:105–11.
4. Reddy PS, Kishor PBK, Seiler C, Kuhlmann M, Eschen-Lippold L, Lee J, Reddy MK, Sreenivasulu N. Unraveling regulation of the small heat shock proteins by the heat Shock factor HvHsfB2c in barley: its implications in drought stress response and seed development. PLoS One. 2014;9:e89125.
5. Waters ER, Aevermann BD, Sanders-Reed Z. Comparative analysis of the small heat shock proteins in three angiosperm genomes identifies new subfamilies and reveals diverse evolutionary patterns. Cell Stress Chaperones. 2008;13:127–42.
6. Kriehuber T, Rattei T, Weinmaier T, Bepperling A, Haslbeck M, Buchner J. Independent evolution of the core domain and its flanking sequences in small heat shock proteins. FASEB J. 2010;24:3633–42.
7. Stamler R, Kappé G, Boelens W, Slingsby C. Wrapping the alpha-crystallin domain fold in a chaperone assembly. J Mol Biol. 2005;353:68–79.
8. Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Res. 2006;34:D247–251.
9. Poulain P, Gelly JC, Flatters D. Detection and architecture of small heat shock protein monomers. PLoS One. 2010;5:e9990.
10. Feil IK, Malfois M, Hendle J, van Der Zandt H, Svergun DI. A novel quaternary structure of the dimeric alpha-crystallin domain with chaperone-like activity. J Biol Chem. 2001;276:12024–9.
11. van Montfort RL, Basha E, Friedrich KL, Slingsby C, Vierling E. Crystal structure and assembly of a eukaryotic small heat shock protein. Nat Struct Biol. 2001;8:1025–30.
12. Pasta SY, Raman B, Ramakrishna T, Rao CM. Role of the conserved SRLFDQFFG region of α-crystallin, a small heat shock protein. Effect on oligomeric size, subunit exchange, and chaperone-like activity. J Biol Chem. 2003;278:51159–66.
13. Saji H, Iizuka R, Yoshida T, Abe T, Kidokoro S, Ishii N, Yohda M. Role of the IXI/V motif in oligomer assembly and function of StHsp14.0., a small heat shock protein from the acidothermophilic archaeon., Sulfolobus tokodaii strain 7. Proteins. 2008;71:771–82.
14. Bagnéris C, Bateman OA, Naylor CE, Cronin N, Boelens WC, Keep NH, Slingsby C. Crystal structures of α-crystallin domain dimers of αB-crystallin and Hsp20. J Mol Biol. 2009;392:1242–52.
15. Healy EF. A model for heterooligomer formation in the heat shock response of *Escherichia coli*. Biochem Biophys Res Commun. 2012;420:639–43.
16. Jaya N, Garcia V, Vierling E. Substrate binding site flexibility of the small heat shock protein molecular chaperones. Proc Natl Acad Sci U S A. 2009;106:15604–9.
17. Ma C, Haslbeck M, Babujee L, Jahn O, Reumann S. Identification and characterization of a stress-inducible and a constitutive small heat-shock protein targeted to the matrix of plant peroxisomes. Plant Physiol. 2006;141:47–60.
18. de Jong WW, Leunissen JA, Voorter CE. Evolution of the alpha-crystallin/small heat-shock protein family. Mol Biol Evol. 1993;10:103–26.
19. Basha E, O'Neill H, Vierling E. Small heat shock proteins and α-crystallins: dynamic proteins with flexible functions. Trends Biochem Sci. 2012;37:106–17.
20. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 2007;35:D237–240.
21. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009;37:D224–228.
22. Baneyx F. Molecular chaperones. 2008. http://faculty.washington.edu/baneyx/Chaperones/Chaperones.html#Hsp33.
23. Lesaint D, Mehta D, O'Sullivan B. Soft pattern discovery in pre-classified protein families through constraint optimization. 9th Workshop on Constraint-based methods for Bioinformatics (WCB) at CP-2013, Uppsala: 2013.
24. Lesaint D, Mehta D, O'Sullivan B, Vigneron V. A decomposition approach for discovering discriminative motifs in a sequence database. 21st European Conference on Artificial Intelligence (ECAI), IOS Press, Prague: 2014.

25. Ratheesh KR, Nagarajan NS, Arunraj SP, Sinha D, Veedin Rajan VB, Esthaki VK, D'Silva P. HSPIR: a manually annotated heat shock protein information resource. Bioinformatics. 2012;28:2853–5.
26. Scharf KD, Siddique M, Vierling E. The expanding family of Arabidopsis thaliana small heat stress proteins and a new family of proteins containing α-crystallin domains ACD proteins. Cell Stress Chaperones. 2001;6:225–37.
27. Tao P, Guo WL, Li BY, Wang WH, Yue ZC, Lei JL, Zhong XM. Genome-wide identification, classification, and expression analysis of sHSP genes in Chinese cabbage (Brassica rapa ssp pekinensis). Genet MolRes 2015;14: 11975-93.
28. Bondino HG, Valle EM, Ten Have A. Evolution and functional diversification of the small heat shock protein/α-crystallin family in higher plants. Planta. 2012;235:1299–313.
29. Waters ER. The evolution, function, structure, and expression of the plant sHSPs. J Exp Bot. 2013;64:391–403.
30. Zhang J, Liu B, Li J, Zhang L, Wang Y, Zheng H, Lu M, Chen J. Hsf and Hsp gene families in Populus: genome-wide identification, organization and correlated expression during development and in stress responses. BMC Genomics. 2015;16:181.
31. Morrow G, Inaguma Y, Kato K, Tanguay RM. The small heat shock protein Hsp22 of Drosophila melanogaster is a mitochondrial protein displaying oligomeric organization. J Biol Chem. 2000;275:31204–10.
32. de Miguel N, Echeverria PC, Angel SO. Differential subcellular localization of members of the toxoplasma gondii small heat shock protein family. Eukaryot Cell. 2005;4:1990–7.
33. Sun Y, MacRae TH. Small heat shock proteins: molecular structure and chaperone function. Cell Mol Life Sci. 2005;62:2460–76.
34. Chen J, Feige MJ, Franzmann TM, Bepperling A, Buchner J. Regions outside the alpha-crystallin domain of the small heat shock protein Hsp26 are required for its dimerization. J Mol Biol. 2010;398:122–31.
35. Sudnitsyna MV, Mymrikov EV, Seit-Nebi AS, Gusev NB. The role of intrinsically disordered regions in the structure and functioning of small heat shock proteins. Curr Protein Pept Sci. 2012;13:76–85.
36. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. Dendroscope: an interactive viewer for large phylogenetic trees. BMC Bioinf. 2007;22:460.
37. Corpet F. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. 1988;16:10881–90.
38. Gouet P, Robert X, Courcelle E. ESPript/ENDscript: extracting and rendering sequence and 3D information from atomic structures of proteins. Nucleic Acids Res. 2003;31:3320–3.