

# Evolution towards simplicity in bacterial small heat shock protein system

Piotr Karaś, Klaudia Kochanowicz, Marcin Pitek, Przemysław Domanski, Igor Obuchowski, Barłomiej Tomiczek\*, Krzysztof Liberek\*

Intercollegiate Faculty of Biotechnology UG-MUG, University of Gdansk, Gdańsk, Poland

**Abstract** Evolution can tinker with multi-protein machines and replace them with simpler single-protein systems performing equivalent functions in an equally efficient manner. It is unclear how, on a molecular level, such simplification can arise. With ancestral reconstruction and biochemical analysis, we have traced the evolution of bacterial small heat shock proteins (sHsp), which help to refold proteins from aggregates using either two proteins with different functions (IbpA and IbpB) or a secondarily single sHsp that performs both functions in an equally efficient way. Secondarily single sHsp evolved from IbpA, an ancestor specialized in strong substrate binding. Evolution of an intermolecular binding site drove the alteration of substrate binding properties, as well as the formation of higher-order oligomers. Upon two mutations in the  $\alpha$ -crystallin domain, secondarily single sHsp interacts with aggregated substrates less tightly. Paradoxically, less efficient binding positively influences the ability of sHsp to stimulate substrate refolding, since the dissociation of sHps from aggregates is required to initiate Hsp70-Hsp100-dependent substrate refolding. After the loss of a partner, IbpA took over its role in facilitating the sHsp dissociation from an aggregate by weakening the interaction with the substrate, which became beneficial for the refolding process. We show that the same two amino acids introduced in modern-day systems define whether the IbpA acts as a single sHsp or obligatorily cooperates with an IbpB partner. Our discoveries illuminate how one sequence has evolved to encode functions previously performed by two distinct proteins.

**\*For correspondence:**

barłomiej.tomiczek@biotech.ug.edu.pl (BT);  
krzysztof.liberek@ug.edu.pl (KL)

**Competing interest:** The authors declare that no competing interests exist.

**Funding:** See page 18

**Preprint posted**

30 May 2023

**Sent for Review**

22 June 2023

**Reviewed preprint posted**

21 August 2023

**Reviewed preprint revised**

23 November 2023

**Version of Record published**

08 December 2023

**Reviewing Editor:** Christian R Landry, Université Laval, Canada

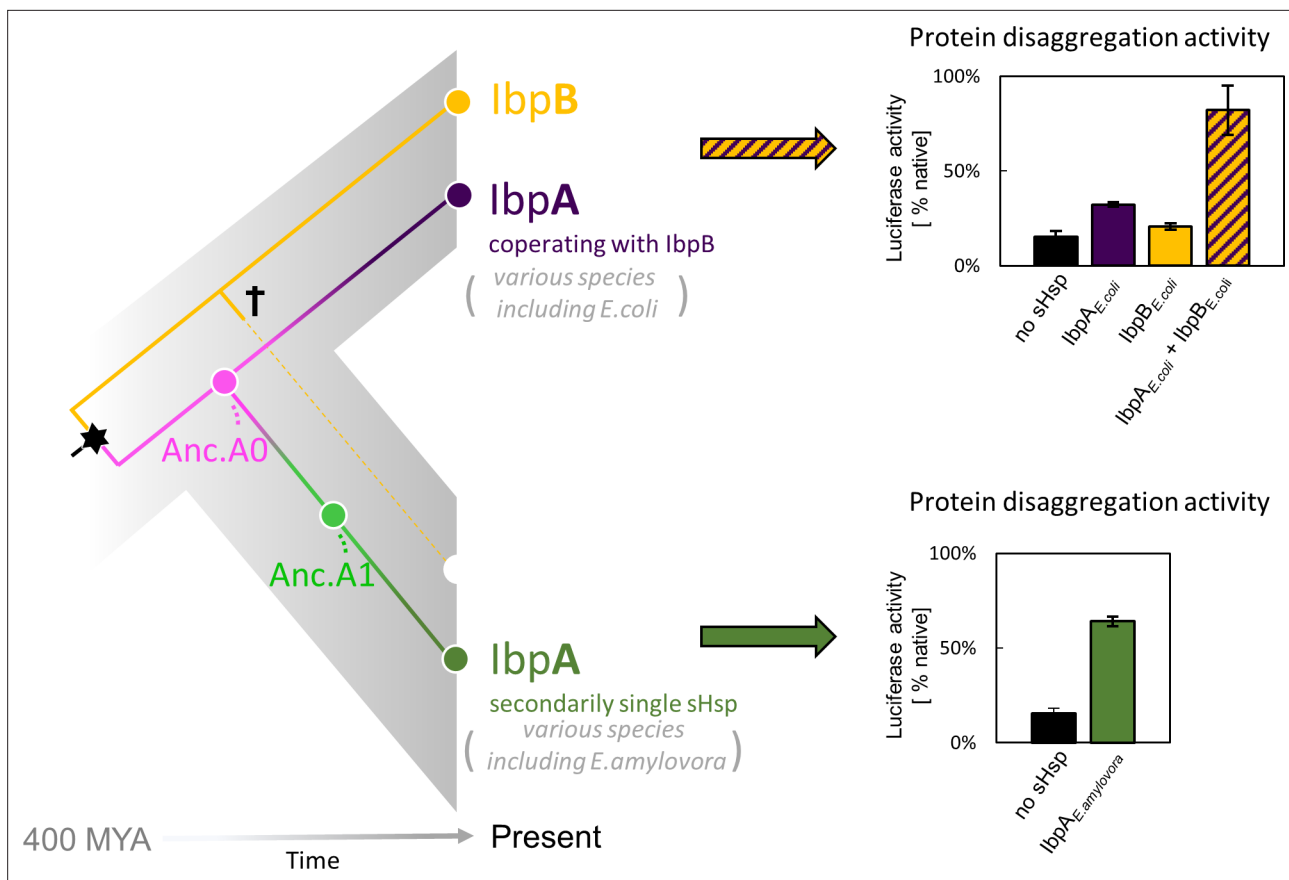
© Copyright Karaś et al. This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

## eLife assessment

This **valuable** study advances our understanding of the evolution of protein complexes and their functions. Through **convincing** experimental and computational methodologies, the authors show that the specialization of protein function following gene duplication can be reversible. The work will be of interest to investigators working in biochemical evolution and those working on heat shock proteins.

## Introduction

Gene birth and loss is a hallmark of protein family evolution, however, molecular determinants and genetic mechanisms enabling that process are not well understood (*Fernández and Gabaldón, 2020; Worth et al., 2009*). Gene loss and differential retention of paralogues reshapes the divergence of organisms both in *Animalia*, *Archaea*, and *Bacteria* (*Fernández and Gabaldón, 2020; Irazzo et al., 2019; Puigbò et al., 2014*). Examples of gene loss often involve adaptive changes in response to changing environmental niches, like loss of genes encoding olfactory receptors in primates (*Demuth and Hahn, 2009*) or differential retention of paralogous genes encoding venom toxins in different rattlesnake lineages (*Dowell et al., 2016*). In prokaryote genomes, gene loss is one of the main evolutionary processes accelerating sequence divergence leading to functional innovations (*Puigbò et al.,*



**Figure 1.** small heat shock protein (sHsp) systems in *Enterobacteriaceae* and *Erwiniaceae*. Left - schematic phylogeny of sHsps in *Enterobacteriales*. Gene duplication resulting in IbpA + IbpB two-protein system is marked with a star, while the loss of *ibpB* gene in *Erwiniaceae* clade is marked with a cross. Anc.A<sub>0</sub> - reconstructed last common ancestor of IbpA from *Erwiniaceae* and *Enterobacteriaceae*, expressed as a part of two-protein system. Anc.A<sub>1</sub> - reconstructed last common ancestor of secondarily single IbpA from *Erwiniaceae*. Right - representative extant sHsps' ability to stimulate luciferase refolding. sHsps were present during the luciferase thermal denaturation step. Refolding of denatured luciferase was performed by the Hsp70-Hsp100 chaperone system. Activity of luciferase was measured after 1 hr refolding at 25 °C and shown as an average of 3 repeats. Error bars represent standard deviation.

The online version of this article includes the following source data for figure 1:

**Source data 1.** Spreadsheet containing raw data for the luciferase refolding graphs shown in **Figure 1**.

**2014**). In complex protein systems, execution of a cellular function can be shared between several proteins. In bacteria, the cost of maintaining additional gene copy is very high and maintaining a low gene count is important for keeping the replication energy costs low (**Kempes et al., 2017; Lever et al., 2015; Lynch and Marinov, 2015**). Still, it remains unclear how a multi-protein system can undergo simplification. Here, we asked what are the molecular events that enabled the gene loss, and how one of the biochemical functions has been taken over by the other protein. We investigated these questions using sHsp system, which underwent simplification within *Enterobacteriales* (which include common bacteria species like *Escherichia coli*, *Salmonella enterica* and *Erwinia amylovora*) as a model (**Figure 1**).

sHsps are a family of ATP-independent molecular chaperones present in all living organisms with various copy numbers (ten representatives in human) (**Haslbeck and Vierling, 2015**). They bind misfolded proteins and sequester them into refolding – prone assemblies, preventing uncontrolled aggregation and helping to maintain proteostasis at stress conditions. sHsp is composed of a highly conserved  $\alpha$ -crystallin domain (ACD), in a form of so-called  $\beta$ -sandwich, flanked by less conserved, unstructured N- and C-terminal regions (**Haslbeck and Vierling, 2015; Haslbeck et al., 2019; Reinle et al., 2022**). Their smallest functional unit is usually a dimer, formed by the interaction between ACDs of two neighboring sHsps. Stable sHsp dimers in turn tend to form variable and dynamic higher-order

oligomers, stabilized by N- and C-terminal region interactions. Particularly the interaction between the IXI motif, highly conserved in sHsps C-termini, and the cleft formed by  $\beta 4$  and  $\beta 8$  strands of ACD is critical for oligomer formation (Haslbeck and Vierling, 2015; Kennaway et al., 2005; Mani et al., 2016; Simmons and Ochoterena, 2000). Oligomers of bacterial sHsps reversibly dissociate into smaller forms when the temperature increases. It is considered their activation mechanism, probably uncovering substrate interaction sites. The mechanism of sHsps' interaction with misfolded substrates is not yet fully understood, but both N-termini and  $\beta 4$ – $\beta 8$  cleft region have been found to play a role in this process (Basha et al., 2006; Fuchs et al., 2009; Jaya et al., 2009; Lee et al., 1997; Reinle et al., 2022).

In most *Enterobacterales* a two-protein sHsps system exists, consisting of lbpA and lbpB proteins (Mogk et al., 2003; Obuchowski et al., 2019). lbpA and lbpB have originated via duplication, form a heterodimer partnership, and are functionally divergent from one another (Obuchowski et al., 2019; Piróg et al., 2021; Figure 1A and B). lbpA is specialized in tight substrate binding (sequestrase activity), while lbpB is required for the dissociation of both sHps from the aggregates, a step necessary to initiate Hsp70-Hsp100 dependent substrate disaggregation and refolding (Obuchowski et al., 2021; Ratajczak et al., 2009). In a subset of *Enterobacterales* (*Erwiniaceae*), as a result of *lbpB* gene loss, the secondarily single-protein sHsp (lbpA) system has emerged (Figure 1). The term 'secondarily single' is used in order to distinguish it from single-protein lbpA from clades in which the duplication did not occur (for example *Vibrionaceae*; Obuchowski et al., 2019). How did lbpA evolve to become independent of its partner? In this study, using ancestral reconstruction, we identify mutations, which allowed the secondarily single lbpA to be fully functional without its partner in substrate sequestration and handover to Hsp70-Hsp100-mediated disaggregation and refolding.

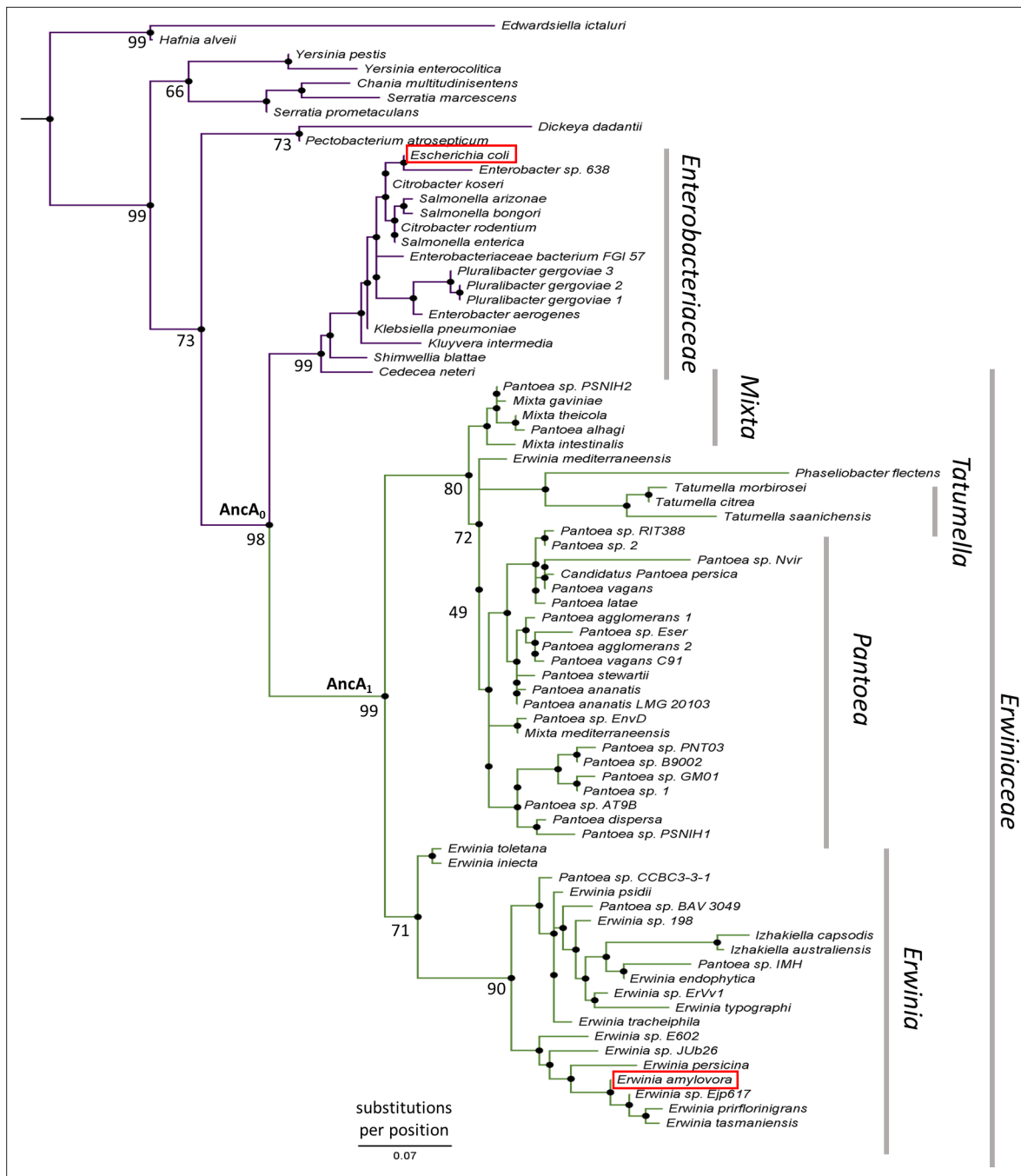
## Results

### New activity of *Erwiniaceae* lbpA has evolved in parallel to lbpB gene loss

To better understand the evolution of sHsps after gene loss, we reconstructed the lbpA ancestors from before and after the loss of its lbpB partner. This technique uses multiple sequence alignments of modern-day proteins from different species to infer amino acid sequences of its common ancestors (Ashkenazy et al., 2012; Pupko et al., 2002) and is widely used to investigate various evolutionary questions (Gaucher et al., 2008; Longo et al., 2020; Thomson et al., 2005; Thornton et al., 2003). We created a multiple sequence alignment of 77 lbpA sequences from *Enterobacterales* (Supplementary file 1), from which we inferred the phylogeny of lbpA using the maximum likelihood method (Figure 2, Supplementary file 2 – phylogenetic tree in Newick format). From that, we inferred ancestral sequences, which have the highest probability of producing the modern-day sequences using the empirical Bayes method (Ashkenazy et al., 2012; Cohen and Pupko, 2011; Cohen et al., 2008; Pupko et al., 2002; Simmons and Ochoterena, 2000). Next, we resurrected (i.e. expressed and purified) the last ancestor of lbpA present before ( $AncA_0$ ) and after ( $AncA_1$ ) the differential gene loss (Figure 2, Supplementary file 3).

Similarly, to modern-day lbpA proteins both  $AncA_0$  and  $AncA_1$  were fully folded, and reversibly deoligomerized into smaller species under elevated temperature (Figure 3—figure supplement 1). Moreover, both ancestral proteins were able to sequester aggregating firefly luciferase in sHsp-substrate assemblies.  $AncA_0$  exhibited sequestrase activity on the level comparable to lbpA from *Escherichia coli* (lbpA<sub>*E. coli*</sub>).  $AncA_1$  was moderately efficient in this process and lbpA from *Erwinia amylovora* (lbpA<sub>*E. amylovora*</sub>) was the least efficient sequestrase (Figure 3A). The differences in sequestrase activity were especially pronounced at lower sHsp concentrations. Next, we tested their ability to bind protein aggregates in real time (Figure 3B). Ancestral proteins' interaction with the aggregated substrates was stronger than in the case of extant *E. amylovora* lbpA, but weaker than in the case of extant *E. coli* lbpA (Figure 3B).

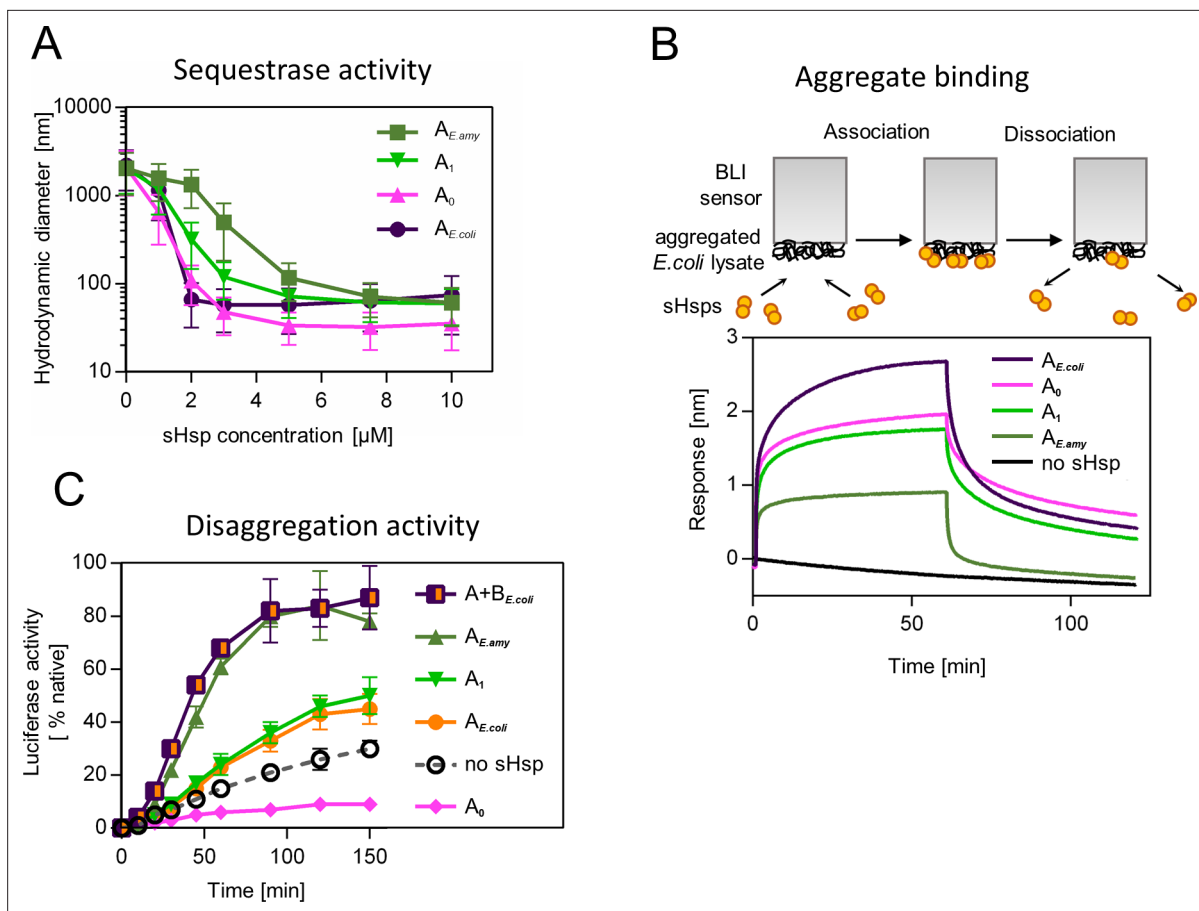
Finally, we asked how the modification of the substrate aggregation process by reconstructed proteins influences subsequent substrate refolding by the Hsp100 and Hsp70 chaperones.  $AncA_1$  stimulated luciferase refolding, however, its effectiveness was around half of both analyzed extant sHsp systems (single lbpA from *E. amylovora* or lbpA + lbpB system from *E. coli*), similar to extant



**Figure 2.** lbpA phylogeny in Enterobacterales. Phylogeny was reconstructed from 77 lbpA orthologs from Enterobacterales using the Maximum Likelihood algorithm with JTT + R3 substitution model. AncA<sub>0</sub> - node representing the last common ancestor of lbpA from Erwiniaceae and Enterobacteriaceae. AncA<sub>1</sub> - node representing the last common ancestor of lbpA from Erwiniaceae. Bootstrap support is noted for the major nodes. Extant lbpAs from *E. coli* and *E. amylovora* are marked with a red frame. Scale bar - substitutions per position.

lbpA from *E. coli* without its lbpB partner. AncA<sub>0</sub>, in contrast, inhibited luciferase refolding in comparison to control (no sHsps at substrate aggregation step) (**Figure 3C**).

To test the robustness of our observations, we repeated the analysis for the alternative ancestors, which have the second highest probability to produce the modern-day sequences (AltAll) (**Figure 3—figure supplement 2, supplementary file 3** – posterior probabilities) (Eick et al., 2017). Both AltAll variants behaved similarly to most likely (ML) variants in reversible deoligomerization, sequestrase activity and stimulation of substrate refolding assays (**Figure 3—figure supplement 3A-D**). However,



**Figure 3.** Functional changes during the evolution of secondarily single small heat shock protein (sHsp) in *Erwinaceae*. **(A)** Sequestrase activity of extant and ancestral sHsps. Luciferase was heat denatured in the presence of different concentrations of sHsps and size of formed sHsps-substrate assemblies was measured by DLS. For every measurement, ten subsequent series of ten 10-s runs were averaged and particle size distribution was calculated by fitting to 70 size bins between 0.4 and 10,000 nm. Results are presented as an average hydrodynamic diameter of measured particles weighted by intensity (Z-average), calculated with Malvern Zetasizer Software 7.13. Error bars represent standard deviation obtained from the distribution. **(B)** Binding of extant and ancestral sHsps to heat-aggregated *E. coli* proteins. *E. coli* proteins were heat aggregated and immobilized on a Biolayer interferometry (BLI) sensor. sHsps were heat-activated before the binding step. **(C)** Extant and ancestral sHsps' ability to stimulate luciferase refolding. Experiment was performed at 25 °C. Luciferase activity at each timepoint was shown as an average of 3 repeats. Error bars represent standard deviation.

The online version of this article includes the following source data and figure supplement(s) for figure 3:

**Source data 1.** Spreadsheet containing raw data for the sequestrase activity graph shown in **Figure 3A**.

**Source data 2.** Spreadsheet containing raw data for the binding and dissociation curve shown in **Figure 3B**.

**Source data 3.** Spreadsheet containing raw data for the luciferase refolding graph shown in **Figure 3C**.

**Figure supplement 1.** Both reconstructed proteins reversibly deoligomerize at heat shock temperature, similarly to extant proteins.

**Figure supplement 1—source data 1.** Spreadsheet containing raw data for DLS curves shown in **Figure 3—figure supplement 1**.

**Figure supplement 2.** Amino acid sequences of most likely (ML) and AltAll variants of reconstructed ancestral proteins.

**Figure supplement 3.** AltAll variants of AncA<sub>0</sub> and AncA<sub>1</sub> have similar properties to most likely (ML) variants.

**Figure supplement 3—source data 1.** Spreadsheet containing raw data for DLS curves shown in **Figure 3—figure supplement 3A**.

**Figure supplement 3—source data 2.** Spreadsheet containing raw data for sequestrase activity measurements shown in **Figure 3—figure supplement 3B**.

**Figure supplement 3—source data 3.** Spreadsheet containing raw data for luciferase refolding graphs shown in **Figure 3—figure supplement 3C**.

**Figure supplement 3—source data 4.** Spreadsheet containing raw data for luciferase refolding graphs shown in **Figure 3—figure supplement 3D**.

the last property (the influence on refolding) required higher Hsp70 system concentration to observe AltAll variants activity (**Figure 3—figure supplement 3C, D**). Together, these data show that reconstructed ML and AltAll ancestors are functional. What is of particular interest, these data clearly point out that the ability of reconstructed sHsps to stimulate Hsp70-Hsp100-dependent substrate refolding arose between  $A_0$  and  $A_1$  nodes.

We performed a molecular evolution analysis to test for positive selection across lbpA phylogeny using both branch models and branch-site models in codeml (*Jeffares et al., 2015; Yang, 1998; Yang, 2007; Yang and Nielsen, 2002*). The analysis shows a significantly increased ratio of nonsynonymous to synonymous substitutions, after the gene loss, at the branch leading to  $A_1$  with both tests. This result indicates that the new lbpA functionality likely arose due to an episode of positive selection rather than genetic drift (**Figure 4A, Supplementary file 4 A,B**). The result of the branch-site test indicates possible positive selection acting on all sites substituted at the branch leading to  $A_1$  with  $pp > 0.5$ , therefore, we aimed to identify the minimum number of mutations that are responsible for a change in functional properties of lbpA.

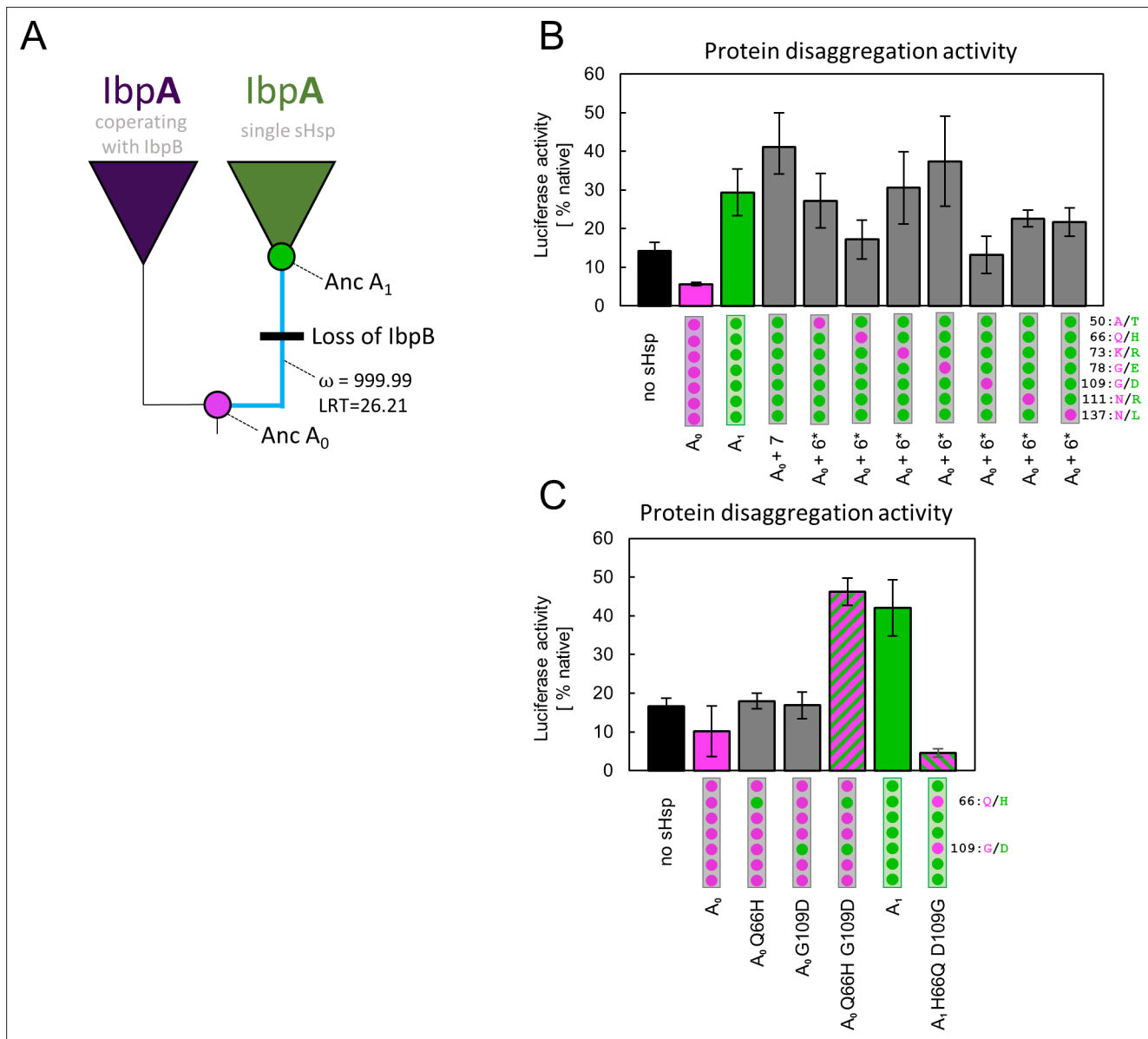
### Identification of residues defining ancestral sHsps activities

In order to identify amino acids responsible for the observed new functionality of  $AncA_1$ , we compared the sequences of two ancestral proteins, selecting 7 out of 10 substitutions as probable candidates. Three substitutions were removed from the analysis based on the low conservation of these positions in extant proteins (**Figure 4—figure supplement 1**). The remaining seven were introduced into  $AncA_0$ . Resulting  $AncA_0 + 7$  protein stimulated Hsp70-Hsp100-dependent luciferase refolding at the level comparable to  $AncA_1$  (**Figure 4B**). To further specify key mutations, we prepared seven additional variants. In each variant, a different position in  $AncA_0 + 7$  was reversed to a more ancestral state. The substantial decrease in luciferase refolding stimulation was observed for positions 66 and 109 (**Figure 4B**). Next, each of these substitutions on its own (Q66H or G109D) was separately introduced into  $AncA_0$ . This was not sufficient to increase  $AncA_0$  ability to stimulate luciferase refolding. However, when both substitutions were introduced simultaneously, the resulting sHsp exhibited activity similar to  $AncA_1$  (**Figure 4C**). What is more, when in  $AncA_1$  these two positions were reversed to  $AncA_0$ -like state, the resulting sHsp lost the ability to stimulate luciferase refolding (**Figure 4C**). All analyzed proteins, namely  $AncA_0 + 7$ ,  $AncA_0$  Q66H G109D and  $AncA_1$  H66Q D109G, possess biochemical properties characteristic for sHsps, exhibiting reversible thermal deoligomerization and sequestrase activity (**Figure 4—figure supplement 2 A,B**). All these results show that substitutions Q66H and G109D are both sufficient and necessary for the increase in activity observed for ancestral sHsps between  $A_0$  and  $A_1$  nodes.

### Identified substitutions influence $\alpha$ -crystallin domain properties

Substitutions Q66H and G109D, responsible for gaining single sHsp activity, are located in the ACD within  $\beta 4$  and  $\beta 8$  strands, which form a cleft responsible for the interaction with unstructured C-terminal peptide of the neighboring sHsp dimer (**Figure 5A**). In order to identify possible structural underpinnings of the single sHsp activity we have predicted structures of  $AncA_0$  and  $AncA_0$  Q66H G109D ACD dimers in complex with C-terminal peptide using AlphaFold2 and in silico mutagenesis and subjected them to 0.5  $\mu$ s equilibrium molecular dynamics (MD) simulations. Analysis of the C-terminal peptide interface contact probabilities in MD trajectories showed that both substituted residues contact the C-terminal peptide, although the overall contact pattern remain similar upon their introduction (**Figure 5—figure supplement 1**) and no major differences in the overall ACD domains structure were observed (**Figure 5—figure supplement 2**). To explore the possibility that identified substitutions affect the strength of this interaction, we analyzed the binding of purified ACDs of  $AncA_0$  and  $AncA_0$  Q66H G109D to C-terminal peptide using biolayer interferometry. Titrations of immobilized C-terminal peptides by different ACDs (**Figure 5B, Figure 5—figure supplement 3**) allowed us to determine the dissociation constants. These two substitutions increased the  $K_{0.5}$  of ACD binding to the C-terminal peptide from 4.3  $\mu$ M to 7.1  $\mu$ M at the same time increasing the Hill coefficient of the interaction from 2.3 to 3.7, indicating a modest decrease in affinity, accompanied by an increase in binding cooperativity.

As this interaction is known to play a crucial role in the formation of sHsp oligomers (*Fu et al., 2005; Mani et al., 2016; Simmons and Ochoterena, 2000*), we used dynamic light scattering to



**Figure 4.** Substitutions at positions 66 and 109 that occurred between nodes A<sub>0</sub> and A<sub>1</sub> are crucial for ancestral small heat shock proteins (sHsps) to work as a single protein. Luciferase refolding assay was performed as in **Figure 1**. Activity of luciferase was measured after 1 hr refolding at 25 °C and shown as an average of 3 repeats (6 in the case of no sHsp and proteins AncA<sub>0</sub> and AncA<sub>1</sub> in panel C). Error bars represent standard deviation. **(A)** Schematic phylogeny of *Enterobacteriales* lbpA showing increased ratio of nonsynonymous to synonymous substitutions ( $\omega$ ) on the branch between nodes AncA<sub>0</sub> and AncA<sub>1</sub>. Loss of cooperating lbpB is marked on a tree. Value of the Likelihood Ratio Test (LRT) is given for the selection model. **(B)** Identification of substitutions necessary for AncA<sub>0</sub> to obtain AncA<sub>1</sub>-like activity in luciferase disaggregation; seven candidate mutations were introduced into AncA<sub>0</sub> (AncA<sub>0</sub> + 7); subsequently, in series of six mutants, each of the candidate positions was reversed to a more ancestral state (AncA<sub>0</sub> + 6\* variants) **(C)** Effect of substitutions at positions 66 and 109 on the ability of AncA<sub>0</sub> and AncA<sub>1</sub> to stimulate luciferase refolding.

The online version of this article includes the following source data and figure supplement(s) for figure 4:

**Source data 1.** Spreadsheet containing raw data for the luciferase refolding graph shown in **Figure 4B**.

**Source data 2.** Spreadsheet containing raw data for the luciferase refolding graph shown in **Figure 4C**.

**Figure supplement 1.** Amino acid sequence differences between AncA<sub>0</sub> and AncA<sub>1</sub>; Differing positions used in further analysis were marked in red, an differing positions omitted from further analysis due to low conservation in extant *Erwiniaceae* were marked in bold.

**Figure supplement 2.** AncA<sub>0</sub> Q66H G109D, AncA<sub>0</sub> + 7 and AncA<sub>1</sub> H66Q D109G exhibit sequestrase activity and reversibly deoligomerize at heat shock temperature.

**Figure supplement 2—source data 1.** Spreadsheet containing raw data for DLS curves shown in **Figure 4—figure supplement 2A**.

Figure 4 continued on next page

Figure 4 continued

**Figure supplement 2—source data 2.** Spreadsheet containing raw data for sequestrase activity measurements shown in **Figure 4—figure supplement 2B**.

investigate how Q66H and G109D substitutions influence the size of oligomers formed by AncA<sub>0</sub> at different temperatures. In agreement with decreased affinity between ACD and the C-terminal peptide, we have shown that these substitutions slightly decrease the oligomer size and facilitate AncA<sub>0</sub> deoligomerization (**Figure 5—figure supplement 4 A,B**).

$\beta$ 4- $\beta$ 8 cleft in certain sHsps, in addition to its interaction with C-terminal peptide, was also shown to participate in interactions with substrates and partner proteins (**Fuchs et al., 2009; Jaya et al., 2009; Lee et al., 1997; Reinle et al., 2022**). Therefore, we tested whether the ACD of AncA<sub>0</sub> binds protein aggregates and whether this interaction is influenced by Q66H G109D substitutions. We observed that AncA<sub>0</sub> ACD efficiently binds to either aggregated *E. coli* lysate or aggregated luciferase, and this binding was weakened by analyzed substitutions (**Figure 5C, Figure 5—figure supplement 5**). This suggests that ACD of bacterial sHsps interacts with the substrate, most likely through the  $\beta$ 4- $\beta$ 8 cleft.

These results allow us to conclude that substitutions Q66H and G109D in AncA<sub>0</sub> substantially increased the sHsp ability to stimulate Hsp70-Hsp100-dependent substrate refolding by weakening the interaction of  $\beta$ 4- $\beta$ 8 cleft with both the C-terminal peptide and the aggregated substrates. Despite its ability to bind aggregated substrates in biolayer interferometry assay, analyzed ACDs do not exhibit sequestrase activity and were unable to positively influence substrate refolding by the Hsp70-Hsp100 system (**Figure 5—figure supplement 5B, C**).

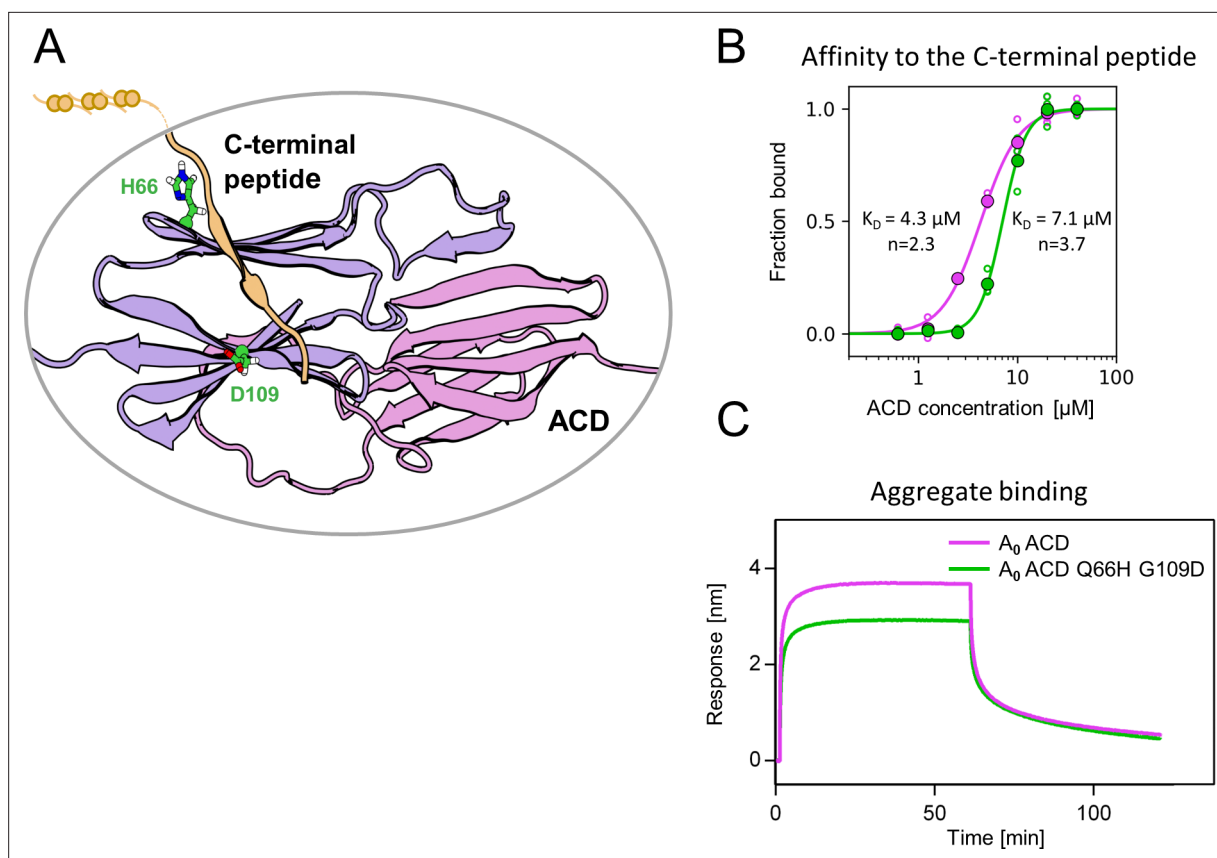
### Identified substitutions define the mode of action of extant sHsps

As more ancestral, AncA<sub>0</sub>-like state in positions 66 and 109 is conserved in lbpA of *E. coli* while more modern, AncA<sub>1</sub>-like state is conserved in lbpA of *E. amylovora*, we decided to ask whether this difference is sufficient to explain functional differences between the two extant proteins. Therefore, we introduced AncA<sub>1</sub>-like substitutions into lbpA<sub>*E. coli*</sub> and AncA<sub>0</sub>-like substitutions into lbpA<sub>*E. amylo*</sub>. Resulting lbpA<sub>*E. coli*</sub> Q66H G109D, in comparison to wild-type lbpA<sub>*E. coli*</sub>, exhibited an increased ability to stimulate Hsp70-Hsp100-dependent luciferase refolding, as well as a decreased ability to bind aggregated substrates becoming more similar to modern *E. amylovora* lbpA. At the same time, lbpA<sub>*E. amylo*</sub> H67Q D110G significantly less efficiently stimulated luciferase refolding, while exhibiting increased ability to bind aggregated substrates in comparison to wild-type lbpA<sub>*E. amylo*</sub> (**Figure 6A–C**). Still, both new lbpA variants exhibited properties characteristic of sHsps, namely reversible thermal deoligomerization and sequestrase activity (**Figure 6—figure supplement 1A, B**).

Above results suggest that tight sHsp binding to aggregates negatively affects the subsequent Hsp70-Hsp100-dependent substrate refolding process. It is initiated by binding of the Hsp70 system (DnaK and cochaperones DnaJ and GrpE) to aggregates that require sHsps to be outcompeted from aggregates (**Żwirowski et al., 2017**). To gain insight into the competition between sHsps and Hsp70 we modified the biolayer interferometry experiments and introduced the sensor with sHsps bound to luciferase aggregates into a buffer containing Hsp70 system (**Figure 6D**). Although biolayer interferometry cannot distinguish between proteins bound to the sensor, we took advantage of the differences in the thickness of the protein layers specific for sHsp or Hsp70 binding and also in the binding kinetics. The analysis of the Hsp70 binding to the aggregates covered with sHsps clearly shows that the presence of lbpA<sub>*E. amylo*</sub> or lbpA<sub>*E. coli*</sub> Q66H G109D on aggregates only weakly inhibits Hsp70 binding (**Figure 6E and H**). In contrast, the inhibition is much more pronounced when lbpA<sub>*E. amylo*</sub> H67Q D110G or lbpA<sub>*E. coli*</sub> are present on aggregates (**Figure 6F and G**).

All the above experiments indicate that two specific amino acids in positions 66 and 109 in ACD of lbpA proteins define the mode of lbpA activity. Glutamine 66 and glycine 109 are characteristic of lbpA proteins which bind tightly to substrates and thus are not easily outcompeted from the aggregates by Hsp70s. Such lbpAs require lbpB partner cooperation to function properly. Substitutions at these positions to histidine (position 66) and aspartic acid (position 109) allowed for the emergence of a single sHsp which binds to aggregating substrate less tightly and can be outcompeted from the aggregates by Hsp70s in the absence of lbpB.





**Figure 5.** Substitutions at positions 66 and 109 decreased the affinity of AncA<sub>0</sub> ACD to C-terminal peptide and aggregated substrate. **(A)** Structural model of complex formed by AncA<sub>0</sub> Q66H G109D α-crystallin domain dimer (purple and lilac) and AncA<sub>0</sub> C-terminal peptide (orange). **(B)** Effect of Q66H G109D substitutions (green) on AncA<sub>0</sub> (purple) ACD's affinity to the C-terminal peptide assayed by Biolayer interferometry (BLI). Biolayer thickness at the end of the association step was used to calculate the fraction of bound peptide. Filled circles represent means of triplicate measurements, individual data points are shown as hollow circles and were fitted to the cooperative binding model (Hill equation). Values of fitted binding affinities [ $K_{0.5}$ ] (AncA<sub>0</sub>  $4.3 \pm 0.2 \mu\text{M}$ , AncA<sub>0</sub> Q66H G109D  $7.1 \pm 0.2 \mu\text{M}$ ) and Hill coefficients [ $n$ ] (AncA<sub>0</sub>  $2.3 \pm 0.17$ , AncA<sub>0</sub> Q66H G109D  $3.7 \pm 0.34$ ) are indicated on the plot. **(C)** Effect of Q66H G109D substitutions on AncA<sub>0</sub> ACD's affinity to aggregated *E. coli* proteins bound to BLI sensor. Analysis was performed as in **Figure 3A**.

The online version of this article includes the following source data and figure supplement(s) for figure 5:

**Source data 1.** PDB file of the structural model of complex formed by AncA<sub>0</sub> Q66H G109D α-crystallin domain dimer and AncA<sub>0</sub> C-terminal peptide shown in **Figure 5A** and **Figure 5—figure supplement 2**.

**Source data 2.** Spreadsheet containing raw values of biolayer thickness at the end of the association step (for binding to the C-terminal peptide and nonspecific binding to His<sub>6</sub>-SUMO), used for fitting the curve shown in **Figure 5B**.

**Source data 3.** Spreadsheets containing raw data for binding and dissociation curves used to obtain biolayer thickness at the end of association curve values shown in **Figure 5—source data 2**.

**Source data 4.** Spreadsheets containing raw data for binding and dissociation curves shown in **Figure 5C**.

**Figure supplement 1.** AncA<sub>0</sub> residues 66 and 109 contact the C-terminal peptide, although overall contact pattern remains similar upon introduction of Q66H G109D substitutions.

**Figure supplement 2.** Comparison of the structural models of AncA<sub>0</sub> (purple) and AncA<sub>0</sub> Q66H G109D (green) α-crystallin domain (ACD) dimers complexed with the C-terminal peptides.

**Figure supplement 2—source data 1.** PDB file of the structural model of complex formed by AncA<sub>0</sub> α-crystallin domain dimer and AncA<sub>0</sub> C-terminal peptide shown in **Figure 5—figure supplement 2**.

**Figure supplement 3.** Substitutions Q66H G109D decrease affinity of AncA<sub>0</sub> ACD to AncA<sub>0</sub> C-terminal peptide.

**Figure supplement 4.** Substitutions Q66H G109D influence AncA<sub>0</sub> oligomerization.

**Figure supplement 4—source data 1.** Spreadsheet containing raw data for DLS curves shown in **Figure 5—figure supplement 4A–C**.

**Figure supplement 4—source data 2.** Spreadsheet containing raw data for graph shown in **Figure 5—figure supplement 4D**.

Figure 5 continued on next page

Figure 5 continued

**Figure supplement 5.** AncA<sub>0</sub> ACD and AncA<sub>0</sub> ACD Q66H G109D can bind to aggregated luciferase, but do not exhibit sequestrase activity or ability to stimulate luciferase refolding.

**Figure supplement 5—source data 1.** Spreadsheet containing raw data for binding and dissociation curves shown in **Figure 5—figure supplement 5A**.

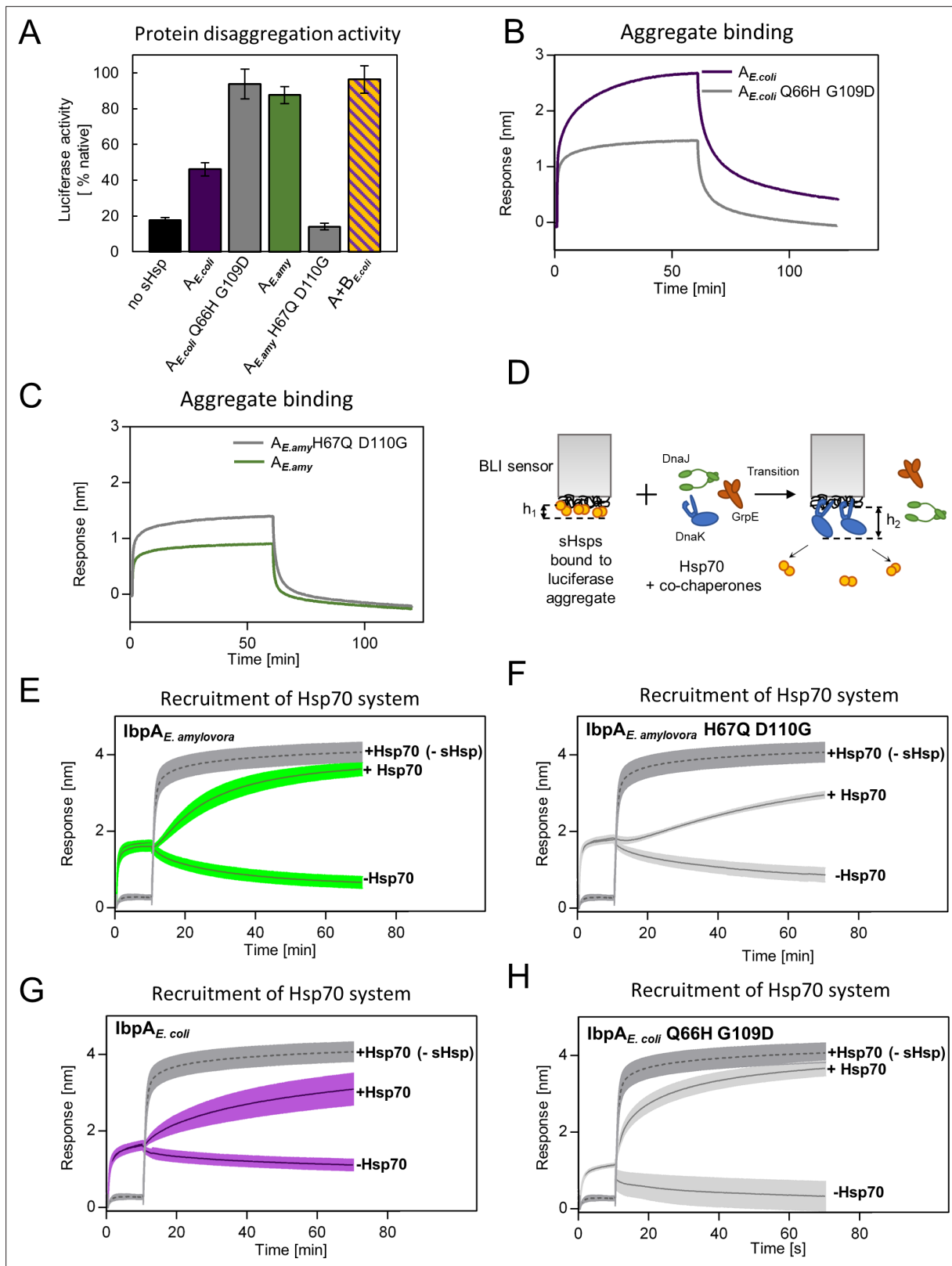
**Figure supplement 5—source data 2.** Spreadsheet containing raw data for sequestrase measurements shown in **Figure 5—figure supplement 5B**.

**Figure supplement 5—source data 3.** Spreadsheet containing raw data for luciferase refolding graphs shown in **Figure 5—figure supplement 5C**.

## Discussion

In this study, we traced, at the molecular level, how the two-protein sHsp system, a part of the cellular protein refolding machinery, underwent simplification in a way that its biochemical functions are performed by a single protein. In most *Enterobacterales* two sHsps (IbpA and IbpB) drive the sequestration of misfolded proteins into the reactivation-prone assemblies (Obuchowski *et al.*, 2019). Together, IbpA and IbpB form a functional heterodimer, in which IbpA specializes in substrate binding, preventing the substrates from creating large aggregates (sequestrase activity), while IbpB promotes sHsps dissociation from the aggregates required for subsequent Hsp70-Hsp100-dependent substrate refolding (Obuchowski *et al.*, 2019; Piróg *et al.*, 2021). We showed that, in parallel to the *ibpB* gene loss in *Erwiniaceae*, new functions of IbpA have emerged, i.e., a lower substrate sequestrase activity, which correlates with the efficient substrate refolding. We have identified two amino acid substitutions (Q66H and G109D) responsible for this new IbpA functionality. Selection analysis shows that these two substitutions were likely driven by positive selective pressure. This indicates that this change possibly had an adaptive character in the ancestral background. It is important to note, however, that models used for the analysis do not account for the variation of synonymous substitution rate and multinucleotide substitution events, which in some cases might lead to false positive results (Lucaci *et al.*, 2023). Because of that, the alternative hypothesis that substitutions Q66H and G109D occurred in the common ancestor of *Erwiniaceae* due to genetic drift, enabling the subsequent loss of *ibpB* gene, cannot be fully discounted. Functional differences observed between modern-day sHsps from *E. coli* and *E. amylovora* are at least partially defined by the presence of specific amino acids in these two positions and can be diminished by their swapping between extant proteins. Their occurrence in the last common ancestor of *Erwiniaceae* IbpA resulted in a decreased affinity of ACD's  $\beta$ 4- $\beta$ 8 cleft to aggregated substrates as well as to the C-termini of the other sHsps. These interactions might be of particular importance for the stabilization of sHsps on a surface of sequestered aggregated substrates leading to the formation of so-called protective shell preventing further uncontrolled aggregation. Apparent role of C-termini and ACD interaction is in agreement with earlier studies, showing that the addition of free C-terminal peptide causes *E. coli* IbpA and IbpB dissociation from the outer shell of sHsp - substrate complex (Żwirowski *et al.*, 2017). It was also shown that, in case of Hsp 16.6 from cyanobacterium *Synechocystis*, substitutions that slightly weakened the interaction between ACD and C-termini lead to increased stimulation of luciferase refolding. However, abolishing this interaction resulted in a non-functional protein, most likely due to the loss of the sequestrase activity (Giese and Vierling, 2002). Destabilization of the protective shell of secondarily single IbpA by weakening these interactions had an effect functionally analogous to a role of IbpB in the two-protein system, facilitating IbpA dissociation from the substrate (Obuchowski *et al.*, 2019). Two identified substitutions also weaken the ACD interaction with aggregated substrates which is an additional factor shifting the sHsps balance towards dissociation, a step necessary to initiate Hsp70-Hsp100 dependent disaggregation and refolding.

Our results show how ACD substitutions can fine-tune sHsp system by exerting pleiotropic effects on ACD-C-terminal peptide and ACD-substrate interactions. These relatively small changes strongly influence the effectiveness of sHsp functioning in complex process of aggregated protein rescue by molecular chaperones. It might be particularly important in the case of a conserved interaction, like the one between C-terminal peptide and ACD, when excessive changes of affinity may be detrimental to the overall protein function (Giese and Vierling, 2002). Our approach enabled us to find functional residues in sHsp system, which would not have been possible by using conventional mutagenesis and highlights the importance of using the vertical approach in biochemical studies. This study closely follows the evolutionary process, in which mutations in one of the two cooperating proteins tinker it



**Figure 6.** Differences at positions 66 and 109 determine functional differences between extant IbpA proteins from *E. coli* and *E. amylovora*. **(A)** Effect of substitutions at position 66 and 109 (and homologous) on the ability of IbpA from *E. amylovora* and *E. coli* to stimulate luciferase refolding. Assay was performed as in **Figure 1B**. Activity of luciferase was measured after 1 h refolding at 25 °C and shown as an average of 3 repeats. Error bars represent standard deviation. **(B, C)** Effect of substitutions at analyzed positions on binding of IbpA from *E. coli* **(B)** and *E. amylovora* **(C)** to heat - aggregated *E.*

Figure 6 continued on next page

Figure 6 continued

*coli* proteins. Assay was performed as in 3 A. (D–H) Effect of substitutions at analyzed positions on inhibition of Hsp70 system binding to aggregates by extant small heat shock proteins (sHsps) (D) Experimental scheme. (E–H) Aggregate-bound sHsps differently inhibit Hsp70 binding. Biolayer interferometry (BLI) sensor with immobilized aggregated luciferase and aggregate bound sHsps was incubated with Hsp70 or buffer (spontaneous dissociation curve). Dark gray traces (dashed line) represent Hsp70 binding to immobilized aggregates in the absence of sHsps. Results are presented as an average of 5 (in the case of IbpA<sub>E.coli</sub> in panel G), 4 (in the case of IbpA<sub>E.coli</sub> Q66H G109d with Hsp70 in panel H), 2 (in the case of IbpA<sub>E.coli</sub> Q66H G109d without Hsp70 in panel H) or 3 repeats (in the case of the remaining curves). Error bars (shades) represent standard deviation.

The online version of this article includes the following source data and figure supplement(s) for figure 6:

**Source data 1.** Spreadsheet containing raw data for the luciferase refolding graph shown in **Figure 6A**.

**Source data 2.** Spreadsheet containing raw data for binding and dissociation curve shown in **Figure 6B**.

**Source data 3.** Spreadsheet containing raw data for binding and dissociation curve shown in **Figure 6C**.

**Source data 4.** Spreadsheet containing raw data for binding and dissociation curves shown in **Figure 6E**.

**Source data 5.** Spreadsheet containing raw data for binding and dissociation curves shown in **Figure 6F**.

**Source data 6.** Spreadsheet containing raw data for binding and dissociation curves shown in **Figure 6G**.

**Source data 7.** Spreadsheet containing raw data for binding and dissociation curves shown in **Figure 6H**.

**Figure supplement 1.** IbpA<sub>E.coli</sub> H66Q G109D and IbpA<sub>E.amyyl</sub> H67Q D110G exhibit sequestrase activity and reversibly deoligomerize at heat shock temperature: (A) Reversible deoligomerization of IbpA<sub>E.coli</sub> H66Q G109D and IbpA<sub>E.amyyl</sub> H67Q D110 at heat shock temperature.

**Figure supplement 1—source data 1.** Spreadsheet containing raw data for DLS curves shown in **Figure 6—figure supplement 1A**.

**Figure supplement 1—source data 2.** Spreadsheet containing raw data for sequestrase activity measurements shown in **Figure 6—figure supplement 1B**.

to a point where it becomes independent of its partner, enabling the simplification of a more complex system by partner loss while maintaining its overall function.

Following, with molecular precision, the genetic events associated with the gene loss allowed us to answer several questions about protein family evolution. The first question concerns how the lost function is incorporated into a remaining partner protein. The results of our experiments indicate that even though the primary function of a partner protein is maintained, it is altered so that it does not interfere with the newly incorporated one. In our case substrate binding properties, as well as the formation of higher-order oligomers were altered in a way to keep the sequestrase function maintained but allow for more efficient stimulation of Hsp70-Hsp100-dependent substrate refolding. The second question concerns the context-dependence of mutations in protein evolution. We successfully transplanted the mutations that appeared in *E. amylovora* into *E. coli* IbpA ortholog, artificially creating an efficient single protein system from a protein that normally needs a partner for efficient substrate refolding. In contrast to other studies (Mogk et al., 2003) the context of *E. coli* IbpA protein did not influence the ability of IbpA<sub>E.coli</sub> Q66H G109D to work without IbpB partner. This brings the question about the accessibility of adaptive solutions. In case of *E. coli* IbpA, there was no need for adaptive changes because the gene loss did not occur in this clade, suggesting that the loss of a protein can push another one towards an adaptation, which leads to finding efficient molecular innovations.

## Materials and methods

### Key resources table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Software, algorithm	Clustal Omega	Clustal Omega	RRID:SCR_001591	European Bioinformatics Institute
Software, algorithm	iq-tree	IQ-TREE		
Software, algorithm	FastML	FastML	RRID:SCR_016092	Tel Aviv University
Software, algorithm	codeml	codeml	RRID:SCR_014932	Part of PAML software package
Software, algorithm	Gromacs 2019.2	GROMACS	RRID:SCR_014565	
Software, algorithm	AlphaFold-Multimer	AlphaFold-Multimer		ColabFold implementation

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Software, algorithm	PyMol	PyMol	RRID:SCR_000305	
Recombinant DNA reagent	pET3a – <i>AncA</i> <sub>0</sub> (plasmid)	This paper		pET3a plasmid carrying <i>AncA</i> <sub>0</sub> gene was ordered from Genscript on the basis of reconstructed amino acid sequence. For the sequence of synthesized gene see <b>Supplementary file 5</b>
Recombinant DNA reagent	pET3a – <i>AncA</i> <sub>1</sub> (plasmid)	This paper		pET3a plasmid carrying <i>AncA</i> <sub>1</sub> gene was ordered from Genscript on the basis of reconstructed amino acid sequence. For the sequence of synthesized gene see <b>Supplementary file 5</b>
Recombinant DNA reagent	pET3a – <i>AncA</i> <sub>0alt_all</sub> (plasmid)	This paper		pET3a plasmid carrying <i>AncA</i> <sub>0alt_all</sub> gene was ordered from Genscript on the basis of reconstructed amino acid sequence. For the sequence of synthesized gene see <b>Supplementary file 5</b>
Recombinant DNA reagent	pET3a – <i>AncA</i> <sub>1alt_all</sub> (plasmid)	This paper		pET3a plasmid carrying <i>AncA</i> <sub>1alt_all</sub> gene was ordered from Genscript on the basis of reconstructed amino acid sequence. For the sequence of synthesized gene see <b>Supplementary file 5</b>
Recombinant DNA reagent	pET3a – <i>AncA</i> <sub>0</sub> + 7 (plasmid)	This paper		pET3a plasmid carrying <i>AncA</i> <sub>0</sub> + 7 gene was ordered from Genscript on the basis of reconstructed amino acid sequence. For the sequence of synthesized gene see <b>Supplementary file 5</b>
Recombinant DNA reagent	pET3a – <i>IbpA</i> <sub>Ea</sub> (plasmid)	This paper		pET3a plasmid carrying gene encoding <i>Erwinia amylovora</i> <i>IbpA</i> protein was ordered from Genscript on the basis of amino acid sequence obtained from PDB database (accession number E5BAR7). For the sequence of synthesized gene see <b>Supplementary file 5</b>
Recombinant DNA reagent	pET28a- <i>AncA</i> <sub>0</sub> ACD (plasmid)	This paper		pET28a plasmid carrying <i>AncA</i> <sub>0</sub> ACD gene was ordered from Genscript on the basis of reconstructed amino acid sequence. For the sequence of synthesized gene see <b>Supplementary file 5</b>
Recombinant DNA reagent	pET28a- <i>AncA</i> <sub>0</sub> Q66H H109D ACD (plasmid)	This paper		pET28a plasmid carrying <i>AncA</i> <sub>0</sub> ACD Q66H G109D gene was ordered from Genscript on the basis of reconstructed amino acid sequence. For the sequence of synthesized gene see <b>Supplementary file 5</b>
Recombinant DNA reagent	Champion pET SUMO Expression System (plasmid)	Thermo Fisher Scientific	Cat. # K30001	
Recombinant DNA reagent	pET28a – <i>His</i> <sub>6</sub> -SUMO-C-peptide (plasmid)	This paper		pET28a plasmid carrying <i>His</i> <sub>6</sub> -SUMO-C-peptide gene was ordered from Genscript on the basis of reconstructed amino acid sequence of <i>AncA</i> <sub>0</sub> C-terminal peptide. For the sequence of synthesized gene see <b>Supplementary file 5</b>
Peptide, recombinant protein	QuantiLum Recombinant Luciferase	Promega	Cat. # E1701	
Peptide, recombinant protein	Creatin Kinase	Roche	Cat. # 10127566001	
Peptide, recombinant protein	DnaK	DOI:10.1016 /j. jmb.2008.12.009		DnaK protein from <i>E. coli</i>
Peptide, recombinant protein	DnaJ	DOI:10.1016 /j. jmb.2008.12.009		DnaJ protein from <i>E. coli</i>
Peptide, recombinant protein	ClpB	DOI:10.1016 /j. jmb.2008.12.009		ClpB protein from <i>E. coli</i>

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Peptide, recombinant protein	IbpA <sub>Ec</sub>	DOI:10.1016 /j.jmb.2008.12.009		IbpA protein from <i>E. coli</i>
Peptide, recombinant protein	IbpB <sub>Ec</sub>	DOI: 10.1016 /j.jmb.2021.167054		IbpB protein from <i>E. coli</i>
Peptide, recombinant protein	GrpE	DOI: j.jmb.2008.12.009		GrpE protein from <i>E. coli</i>
Peptide, recombinant protein	His-tagged luciferase	DOI: 10.1371/journal.pgen.1008479		
Commercial assay or kit	Luciferase Assay System	Promega	Cat. # E151A, E152A	
Strain, strain background ( <i>Escherichia coli</i> )	<i>E. coli</i> BL21 (DE3)	ther		Laboratory collection

## Reconstruction of IbpA phylogeny

Amino acid sequences of 77 IbpA orthologs from *Enterobacterales* were obtained from NCBI and UniProt databases and aligned using Clustal Omega (Sievers et al., 2011). Alignment was trimmed manually. JTT + R3 was identified as the best-fit model by iq-tree, using Bayesian Information Criterion, and was used in the analysis (Kalyanamoothy et al., 2017; Nguyen et al., 2015). The phylogenetic tree was inferred using iq-tree on the basis of 328 iterations of ML search with 100 rapid bootstraps replicates (Nguyen et al., 2015).

## Reconstruction of ancestral IbpA amino acid sequences

Ancestral sequence reconstruction was performed on the basis of multiple sequence alignment of 77 amino acid sequences of IbpA orthologs from *Erwiniaceae* and *Enterobacteriaceae* as well as a phylogenetic tree of those orthologs (see above). Marginal reconstruction of ancestral sequences was performed with FastML program based on ML algorithm and Bayesian approach using JTT substitution matrix with gamma parameter (Ashkenazy et al., 2012; Cohen and Pupko, 2011; Cohen et al., 2008; Jones et al., 1992; Pupko et al., 2002; Simmons and Ochoterena, 2000).

Alternative ancestral sequences for AncA<sub>0</sub> and AncA<sub>1</sub> proteins were obtained by substituting the most likely amino acid on every uncertain position (defined as a position with more than one amino acid with posterior probability  $\geq 0.2$ ) with the amino acid with the second highest posterior probability (Eick et al., 2017).

## Analysis of natural selection

Analysis of natural selection was performed using codeml. First, Pal2Nal was used to obtain codon alignment based on the multiple sequence alignment of amino acid sequences of IbpA orthologs from *Enterobacterales* (see above) as well as corresponding nucleotide sequences obtained from the NCBI database. Resulting codon alignment was then trimmed manually and used together with the phylogenetic tree obtained earlier (see above) for the selection analysis.

For branch model analysis, models M0 (null hypothesis) and two-ratio (with either AncA<sub>0</sub>-AncA<sub>1</sub> branch or *Erwiniaceae* clade as foreground) were used. For branch-site model analysis, models A null (null hypothesis) and A were used, with foreground branches selected as above. Statistical significance of different models was estimated with Likelihood Ratio Test (LRT) (Jeffares et al., 2015; Yang, 1998; Yang, 2007; Yang and Nielsen, 2002).

## Protein purification

### Purification of IbpA proteins

pET3a plasmids containing AncA<sub>0</sub>, AncA<sub>1</sub>, AncA<sub>0</sub> + 7, AncA<sub>0</sub><sub>alt\_allr</sub>, AncA<sub>1</sub><sub>alt\_allr</sub> and IbpA<sub>Ea</sub> genes were ordered from GeneScript (Supplementary file 5). Point mutations were introduced using site-directed mutagenesis and confirmed by sequencing. Proteins were overproduced in *E. coli* BL21(DE3). Cells were then lysed by sonication in Qsonica sonicator (13% amplitude, 2 min 30 s process time, 15 s

pulse-ON time, 45 s pulse-OFF time) in lysis buffer L1 (50 mM Tris pH 7.5, 50 mM NaCl, 5 mM EDTA, 10% glycerol, 5 mM  $\beta$ -mercaptoethanol). Insoluble fraction containing proteins of interest was separated by centrifugation (75,000 x g, 30 min, 4 °C) and resolubilized in buffer A (40 mM Tris pH 7.5, 50 mM NaCl, 10% glycerol, 5 mM  $\beta$ -mercaptoethanol, 6 M urea) and then centrifuged (75,000 x g, 30 min, 4 °C). Supernatant was loaded on Q-Sepharose chromatography column equilibrated with buffer A and eluted in 50 mM–500 mM NaCl gradient. Fractions containing proteins of interest were then dialyzed to buffer B (40 mM Tris pH 8.5, 50 mM NaCl, 10% glycerol, 5 mM  $\beta$ -mercaptoethanol) and loaded on Q-Sepharose chromatography column equilibrated with buffer B. Flow-through fraction was collected and dialyzed to buffer C (50 mM Tris pH 7.5, 150 mM KCl, 5% (v/v) glycerol, 5 mM  $\beta$ -mercaptoethanol).

### Purification of ACD domains

ACDs of IbpA<sub>Ec</sub>, AncA<sub>0</sub>, and AncA<sub>0</sub> Q66H G109D were purified as described previously (Piróg *et al.*, 2021) and as a final step dialyzed to buffer G (50 mM Tris pH 7.5, 150 mM KCl, 5 mM  $\beta$ -mercaptoethanol).

### Purification of His<sub>6</sub>-SUMO and His<sub>6</sub>-SUMO-C-terminal peptide of AncA<sub>0</sub> construct

His<sub>6</sub>-SUMO was purified using the Champion pET SUMO Expression System. pET28a plasmid containing gene encoding His<sub>6</sub>-SUMO fused with C-terminal peptide of AncA<sub>0</sub> (PEAMKPPRIEIN) was ordered from GeneScript (**Supplementary file 5**). Proteins were overproduced in *E. coli* BL21(DE3). Cells were then lysed by sonication in Qsonica sonicator (20% amplitude, 2 min process time, 5 s pulse-ON time, 10 s pulse-OFF time) in lysis buffer L2 (40 mM Tris pH 7.5, 100 mM NaCl, 10% glycerol, 10 mM imidazole, 2 mM  $\beta$ -mercaptoethanol). Insoluble fractions were separated by centrifugation for 30 min at 70,000 x g and supernatants, containing proteins of interest, were incubated for 1 hr with Ni-NTA resin equilibrated with buffer L2. Resins were then washed with the buffer D (40 mM Tris pH 7.5, 100 mM NaCl, 10% glycerol, 40 mM imidazole, 2 mM  $\beta$ -mercaptoethanol). Proteins of interest were eluted from the columns with the buffer E (40 mM Tris pH 7.5, 100 mM NaCl, 10% glycerol, 400 mM imidazole, 2 mM  $\beta$ -mercaptoethanol) and then dialyzed to buffer C (as above).

DnaK, DnaJ, GrpE, ClpB, and IbpA<sub>Ec</sub> proteins were purified as described previously (Ratajczak *et al.*, 2009). IbpB<sub>Ec</sub> protein was purified as described previously (Piróg *et al.*, 2021). His-tagged luciferase used for BLI measurements was purified as described previously (Obuchowski *et al.*, 2019).

Purity of purified proteins was assessed with SDS-PAGE electrophoresis with Coomassie Blue staining. Protein concentrations were measured using the Bradford reaction, with Bovine Serum Albumin as a standard. In the case of His<sub>6</sub>-SUMO fused with C-terminal peptide of AncA<sub>0</sub>, concentration was measured with SDS-PAGE electrophoresis with Coomassie Blue staining coupled with densitometric analysis with Bovine Serum Albumin used as a standard.

QuantiLum Recombinant Luciferase was purchased from Promega. Creatin Kinase from rabbit muscle was purchased from Roche.

### Luciferase refolding assay

1.5  $\mu$ M recombinant firefly luciferase in buffer F (50 mM Tris pH 7.5, 150 mM KCl, 20 mM MgCl<sub>2</sub>, 2.5 mM DTT) was denatured by incubation for 10 min at 44 °C alone or in the presence of 10  $\mu$ M sHsps (3  $\mu$ M IbpA<sub>Ec</sub> + 7  $\mu$ M IbpB<sub>Ec</sub> in the case of two-protein system from *E. coli*). Denatured luciferase was then incubated at 25 °C with Hsp70 system (1  $\mu$ M DnaK, 0.4  $\mu$ M DnaJ, and 0.3  $\mu$ M GrpE), 2  $\mu$ M ClpB, and ATP regeneration system (5 mM ATP, 0.1 mg/ml creatine kinase and 18 mM creatine phosphate). For experiment presented in **Figure 3—figure supplement 3D**, higher concentration of the Hsp70 system was used (2  $\mu$ M DnaK, 0.8  $\mu$ M DnaJ, and 0.6  $\mu$ M GrpE). At different timepoints luciferase activity was measured with GLOMAX 20/20 luminometer, using the Luciferase Assay System from Promega. Results are presented as averages of at least three independent repeats  $\pm$  standard deviation.

### DLS measurements

Dynamic Light Scattering measurements were performed using Malvern Instruments ZetaSizer Nano S instrument, at 40  $\mu$ l sample volume, scattering angle of 173° and wavelength of 633 nm. For every measurement, a minimum of 10 subsequent series of ten 10 s runs were averaged and particle

size distribution was calculated by fitting to 70 size bins between 0.4 and 10,000 nm, as previously described (Żwirowski et al., 2017).

For reversible deoligomerization assay, size of oligomers formed by 10  $\mu\text{M}$  sHsps in buffer F (50 mM Tris pH7.5, 150 mM KCl, 20 mM  $\text{MgCl}_2$ , 2.5 mM DTT) were measured by DLS. First measurement was performed at 25 °C and then the sample was heated to 44 °C, cooled to 25 °C, heated to 44 °C and cooled to 25 °C, with measurements performed after each change in temperature. Results are presented as size distribution by volume.

For measuring the influence of substitutions on oligomer formation, the size of oligomers formed by either 10  $\mu\text{M}$  AncA<sub>0</sub> or 10  $\mu\text{M}$  AncA<sub>0</sub> Q66H G109D in buffer F (as above) was measured by DLS at 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, and 45°C. Results were presented as size distribution by intensity (for temperatures 25 °C, 35°C, and 45°C) or as an average hydrodynamic diameter corresponding to the maximum of a dominant peak of size distribution by volume plotted against temperature  $\pm$  standard deviation.

For assembly formation (sequestrase) assay, 1.5  $\mu\text{M}$  firefly luciferase in buffer F was denatured alone or in the presence of different sHsp concentrations by incubation at 44 °C for 10 min. Size of obtained luciferase aggregates was then measured by DLS at 25 °C. Results are presented as an average hydrodynamic diameter of measured particles weighted by intensity (Z-average)  $\pm$  standard deviation.

### Biolayer interferometry (BLI) measurements

sHsps interactions with aggregated luciferase or aggregated *E. coli* lysate were measured using the Octet K2 system. Anchoring layer of his-tagged luciferase was attached to Octet NTA Biosensors by 5 min incubation in 0.6 mg/ml his-tagged luciferase in denaturing conditions in buffer UF (50 mM Tris pH 7.5, 4.5 M urea, 150 mM KCl, 20 mM  $\text{MgCl}_2$ , 2.5 mM DTT) at 25 °C with 350 rpm shaking. Sensors were then incubated for 5 min as above in buffer H (50 mM Tris pH 7.5, 150 mM KCl, 20 mM  $\text{MgCl}_2$ , 5 mM  $\beta$ -mercaptoethanol) to remove urea and unbound luciferase. The protein aggregate was then formed on the sensor by incubation for 10 min in 0.5 mg/ml His-tagged luciferase or 0.2 mg/ml *E. coli* lysate in buffer H at 44 °C (in case of the luciferase) or 55 °C (in case of the lysate). Sensors were then again incubated in buffer H for 5 min at 25 °C with 350 rpm shaking to remove excess protein. Sensors with attached aggregate were placed in the Octet system in H buffer for 60 s baseline measurement and then placed for 1 hr in 5  $\mu\text{M}$  sHsp solution in H buffer to measure sHsps association. Sensors were then moved for 1 hr to buffer H to measure protein dissociation. Measurements were performed with 1000 rpm shaking at 44 °C (in case of full-length sHsps) or at 25 °C (in case of ACDs). Full-length proteins were preincubated at 44 °C for 10 min before measurement.

sHsps displacement by Hsp70 system was measured using ForteBio BLItz. Sensors with attached aggregates were prepared as described above, with buffer F (50 mM Tris pH 7.5, 150 mM KCl, 20 mM  $\text{MgCl}_2$ , 2.5 mM DTT) instead of buffer H. Baseline biolayer was measured for 60 s. Sensors were then placed in 5  $\mu\text{M}$  sHsp solution in buffer F, previously preincubated for 10 min at 44 °C. sHsps association was measured for 10 min. Sensors were then moved to either buffer F or Hsp70 system in buffer F (0.7  $\mu\text{M}$  DnaK, 0.28  $\mu\text{M}$  DnaJ, 0.21  $\mu\text{M}$  GrpE, 5 mM ATP, 0.1 mg/ml creatine kinase, 18 mM creatine phosphate). Hsp70 system binding and sHsps dissociation were measured for 1 hr. Measurements were performed at room temperature with 2000 rpm shaking.

ACD interactions with C-terminal peptide were measured using Octet K2 system. Octet NTA Biosensors were placed in buffer C (50 mM Tris pH 7.5, 150 mM KCl, 5% glycerol, 5 mM  $\beta$ -mercaptoethanol) and baseline signal was measured for 60 s. Sensors were then placed in 2.5  $\mu\text{M}$  His<sub>6</sub>-SUMO-C-peptide solution in buffer C and incubated for 15 min. Surplus His<sub>6</sub>-SUMO-C-peptide was then removed by incubation in G buffer for 15 min. Sensors were then moved to ACD solution and association was measured for 48 min. After that, ACD dissociation was measured in buffer G for 10 min. Measurements were performed with 1000 rpm shaking at 25 °C. Measurements were performed for different ACD concentrations in triplicates. Biolayer thickness at the end of association stage was corrected for the nonspecific binding using control substituting His<sub>6</sub>-SUMO for His<sub>6</sub>-SUMO-C-peptide and converted to fraction bound by min-max scaling between 0 and 1 using minimal and maximal triplicate averages of biolayer thickness. To determine dissociation constant fraction bound as a function of ACD concentration was fitted to the Hill equation using SciPy implementation (Virtanen et al., 2020) of dogbox algorithm:



$$f_b = \frac{[X]^n}{K_{0.5}^n + [X]^n}$$

Where [X] is the total ACD concentration,  $K_{0.5}$ -ACD concentration required to reach half-maximum binding at equilibrium,  $n$ -Hill coefficient,  $f_b$ -fraction bound. Standard deviations of fitted  $K_{0.5}$  and  $n$  values were derived from the diagonal of the optimized parameters covariance matrix.

## Molecular dynamics (MD)

All MD simulations were performed using Gromacs 2019.2 (Van Der Spoel et al., 2005) and CHARMM36-jul2021 as a force field (Huang et al., 2017). Simulations were performed in the isothermal-isobaric (NPT) ensemble, where the temperature was kept at 310 K using v-rescale thermostat (Bussi et al., 2007) with a time constant of 0.1 ps and the pressure was held at 1 bar using Parrinello-Rahman barostat (Parrinello and Rahman, 1981). Lennard-Jones potential with a cut-off of 1.0 nm was used to describe Van der Waals interactions. Computation of long-range electrostatic interactions was performed using the particle mesh Ewald (PME) method (Essmann et al., 1995) with a Fourier grid spacing of 0.12 nm and a real space cutoff of 1.0 nm. Bonds between hydrogen and protein heavy atoms were constrained by P-LINCS (Hess et al., 1997), and water molecules geometry was constrained by SETTLE (Miyamoto and Kollman, 1992). Integration of equations of motion was performed by leap-frog algorithm (Van Gunsteren and Berendsen, 1988) with a time step of 2 fs. Periodic boundary conditions were applied in all dimensions.

The initial conformation of AncA<sub>0</sub> ACD-C-terminal peptide complex was predicted by ColabFold implementation (Mirdita et al., 2022) of AlphaFold-Multimer (Evans et al., 2022) Substitutions Q66H and G109D were introduced to the complex using PyMol Mutagenesis Wizard (<http://www.pymol.org>). Complexes were then placed in rhombic dodecahedral boxes measuring 10.4 nm in all dimensions and solvated by CHARMM-modified TIP3P water model (Jorgensen et al., 1983). Concentrations of sodium and chloride ions were adjusted to 0.15 M and net zero charge of the system. Each system was subjected to a 3-step energy minimization protocol, where during the first step protein conformation was constrained, during the second step constraint was reduced to protein backbone only and during the third step positions of all protein heavy atoms were restrained with a force constant of 1000 kJ·mol<sup>-1</sup>·nm<sup>-1</sup>. Minimized systems were equilibrated for 10 ns while positions of protein backbone atoms were kept constrained, equilibration was then continued without constraints for further 500 ns. The first 100 ns of equilibration was discarded, and the rest was used for ACD-C-terminal peptide contact determination using GetContacts tool (<https://getcontacts.github.io/>; Fonseca and Ma, 2018). Contact between interfacial residues was defined as any of the following types of interaction: hydrogen bond, ionic,  $\pi$ -stacking,  $\pi$ -cation, or van der Waals between purely hydrophobic residues. The default GetContacts interaction criteria were used for all interaction types except for hydrogen bond detection, where a more stringent 30° cutoff for hydrogen-donor-acceptor angle was used. Contact heatmaps were prepared using seaborn (Waskom, 2021).

The representative conformations of ACD-C-terminal peptide complexes were chosen by clustering the last 400 ns of equilibrium MD trajectories (frames spaced every 0.5 ns) using 'gmx cluster' tool and Jarvis-Patrick clustering method (Jarvis and Patrick, 1973). RMSD cutoff of 0.2 nm, was used for the Jarvis-Patrick algorithm and conformations possessing at least three neighbors in common out of 15 closest conformations were assigned to the same cluster. RMSD calculation was based on coordinates of heavy backbone atoms of the C-terminal peptide interacting ACD monomer without dimerization loop (residues 40–74 and 95–126) and stably interacting region of the C-terminal peptide (residues 132–137). In the case of both simulated complexes the biggest cluster contained more than 90% of simulation frames (92.3% for AncA0 ACD complex and 93.3% for AncA0 Q66H G109D complex) and its middle frame was chosen as representative conformation. Visual Molecular Dynamics (VMD) (Humphrey et al., 1996) and Blender (<https://www.blender.org/>) were used for structure visualization. **Supplementary file 1** **Supplementary file 3**.

## Acknowledgements

This work was supported by a grant from the Polish National Science Centre (OPUS 17 2019/33/B/NZ1/00352). The work of B.T. and P.D. was supported by a National Science Center grant (OPUS 21 2021/41/B/NZ8/02835)(to B.T.). We gratefully acknowledge Poland's high-performance computing

infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2023/016546. We thank Prof. Max Telford, Prof. Jaroslaw Marszalek, and Dr. Agnieszka Kłosowska for their helpful discussions.

## Additional information

### Funding

Funder	Grant reference number	Author
Narodowe Centrum Nauki	OPUS 17 2019/33/B/NZ1/00352	Piotr Karaś Klaudia Kochanowicz Marcin Pitek Igor Obuchowski Krzysztof Liberek
Narodowe Centrum Nauki	OPUS 21 2021/41/B/NZ8/02835	Przemyslaw Domanski Barlomiej Tomiczek
Infrastruktura PL-Grid	PLG/2023/016546	Barlomiej Tomiczek

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

### Author contributions

Piotr Karaś, Conceptualization, Data curation, Formal analysis, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing; Klaudia Kochanowicz, Investigation, Methodology; Marcin Pitek, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing; Przemyslaw Domanski, Investigation, Visualization; Igor Obuchowski, Methodology, Writing - review and editing; Barlomiej Tomiczek, Conceptualization, Formal analysis, Supervision, Validation, Visualization, Writing - original draft, Project administration, Writing - review and editing; Krzysztof Liberek, Conceptualization, Supervision, Funding acquisition, Validation, Writing - original draft, Project administration, Writing - review and editing

### Author ORCIDs

Piotr Karaś  <http://orcid.org/0000-0001-5270-7938>

Marcin Pitek  <http://orcid.org/0000-0002-1300-4364>

Barlomiej Tomiczek  <http://orcid.org/0000-0001-9295-663X>

Krzysztof Liberek  <http://orcid.org/0000-0002-7532-9279>

### Peer review material

Reviewer #1 (Public Review): <https://doi.org/10.7554/eLife.89813.3.sa1>

Reviewer #2 (Public Review): <https://doi.org/10.7554/eLife.89813.3.sa2>

Author Response <https://doi.org/10.7554/eLife.89813.3.sa3>

## Additional files

### Supplementary files

- Supplementary file 1. Multiple sequence alignment of *Enterobacteriales* IbpA orthologs used for phylogenetic analysis and ancestral reconstruction in fasta format.
- Supplementary file 2. Phylogenetic tree of *Enterobacteriales* IbpA protein family in newick format (see **Figure 2**).
- Supplementary file 3. Posterior probability statistics for ancestral sequence reconstruction of AncA<sub>0</sub> and AncA<sub>1</sub> nodes. For each position amino acids reconstructed with posterior probability higher than 0.2 are shown. Single-letter symbols of reconstructed amino acids are followed by posterior probability of reconstruction (in brackets). Positions at which most likely amino acid differ between AncA<sub>0</sub> and AncA<sub>1</sub> are marked in bold and italics. Posterior probabilities were estimated using the FastML program based on the Maximum Likelihood and Empirical Bayes method.
- Supplementary file 4. Statistics for selection analysis. (A) Branch model statistics for IbpA orthologs

from *Enterobacterales*. Models assumed either branch between nodes AncA<sub>0</sub> and AncA<sub>1</sub> or entire *Erwiniaceae* clade as foreground; NS – Not significant. (B) Branch-site model statistics for lbpA orthologs from *Enterobacterales*. Models assumed either branch between nodes AncA<sub>0</sub> and AncA<sub>1</sub> or entire *Erwiniaceae* clade as foreground; NS – Not significant.

- Supplementary file 5. Nucleotide sequences of de novo synthesized genes ordered from Genscript.
- MDAR checklist

### Data availability

All data generated or analysed during this study are included in the manuscript and supporting files. Source data files have been provided for Figs 1, 3, 4, 5, 6 and Figure supplements. Multiple sequence alignment of *Enterobacterales* lbpA orthologs used for phylogenetic analysis and ancestral reconstruction in fasta format is attached as **Supplementary file 1**. Posterior probability statistics for ancestral sequence reconstruction of AncA<sub>0</sub> and AncA<sub>1</sub> nodes are attached as **Supplementary file 3**. Raw data generated in this study are attached as source data files.

## References

- Ashkenazy H**, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research* **40**:W580–W584. DOI: <https://doi.org/10.1093/nar/gks498>, PMID: 22661579
- Basha E**, Friedrich KL, Vierling E. 2006. The N-terminal arm of small heat shock proteins is important for both chaperone activity and substrate specificity. *The Journal of Biological Chemistry* **281**:39943–39952. DOI: <https://doi.org/10.1074/jbc.M607677200>, PMID: 17090542
- Bussi G**, Donadio D, Parrinello M. 2007. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**:014101. DOI: <https://doi.org/10.1063/1.2408420>, PMID: 17212484
- Cohen O**, Rubinstein ND, Stern A, Gophna U, Pupko T. 2008. A likelihood framework to analyse phyletic patterns. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **363**:3903–3911. DOI: <https://doi.org/10.1098/rstb.2008.0177>, PMID: 18852099
- Cohen O**, Pupko T. 2011. Inference of gain and loss events from phyletic patterns using stochastic mapping and maximum parsimony—a simulation study. *Genome Biology and Evolution* **3**:1265–1275. DOI: <https://doi.org/10.1093/gbe/evr101>, PMID: 21971516
- Demuth JP**, Hahn MW. 2009. The life and death of gene families. *BioEssays* **31**:29–39. DOI: <https://doi.org/10.1002/bies.080085>, PMID: 19153999
- Dowell NL**, Giorgianni MW, Kassner VA, Selegue JE, Sanchez EE, Carroll SB. 2016. The deep origin and recent loss of venom toxin genes in rattlesnakes. *Current Biology* **26**:2434–2445. DOI: <https://doi.org/10.1016/j.cub.2016.07.038>, PMID: 27641771
- Eick GN**, Bridgham JT, Anderson DP, Harms MJ, Thornton JW. 2017. Robustness of reconstructed ancestral protein functions to statistical uncertainty. *Molecular Biology and Evolution* **34**:247–261. DOI: <https://doi.org/10.1093/molbev/msw223>, PMID: 27795231
- Essmann U**, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. 1995. A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **103**:8577–8593. DOI: <https://doi.org/10.1063/1.470117>
- Evans R**, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Židek A, Bates R, Blackwell S, Yim J, Ronneberger O, Bodenstern S, Zielinski M, Bridgland A, Potapenko A, Cowie A, Tunyasuvunakool K, Jain R, Clancy E, Kohli P, et al. 2022. Protein Complex Prediction with AlphaFold-Multimer. *bioRxiv*. DOI: <https://doi.org/10.1101/2021.10.04.463034>
- Fernández R**, Gabaldón T. 2020. Gene gain and loss across the metazoan tree of life. *Nature Ecology & Evolution* **4**:524–533. DOI: <https://doi.org/10.1038/s41559-019-1069-x>, PMID: 31988444
- Fonseca R**, Ma A. 2018. Getcontacts. GitHub. <https://getcontacts.github.io/>
- Fu X**, Zhang H, Zhang X, Cao Y, Jiao W, Liu C, Song Y, Abulimiti A, Chang Z. 2005. A dual role for the N-terminal region of *Mycobacterium tuberculosis* Hsp16.3 in self-oligomerization and binding denaturing substrate proteins. *The Journal of Biological Chemistry* **280**:6337–6348. DOI: <https://doi.org/10.1074/jbc.M406319200>, PMID: 15545279
- Fuchs M**, Poirier DJ, Seguin SJ, Lambert H, Carra S, Charette SJ, Landry J. 2009. Identification of the key structural motifs involved in HspB8/HspB6-Bag3 interaction. *The Biochemical Journal* **425**:245–255. DOI: <https://doi.org/10.1042/BJ20090907>, PMID: 19845507
- Gaucher EA**, Govindarajan S, Ganesh OK. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451**:704–707. DOI: <https://doi.org/10.1038/nature06510>, PMID: 18256669
- Giese KC**, Vierling E. 2002. Changes in oligomerization are essential for the chaperone activity of a small heat shock protein in vivo and in vitro. *The Journal of Biological Chemistry* **277**:46310–46318. DOI: <https://doi.org/10.1074/jbc.M208926200>, PMID: 12297515
- Haslbeck M**, Vierling E. 2015. A first line of stress defense: small heat shock proteins and their function in protein homeostasis. *Journal of Molecular Biology* **427**:1537–1548. DOI: <https://doi.org/10.1016/j.jmb.2015.02.002>, PMID: 25681016

- Haslbeck M**, Weinkauff S, Buchner J. 2019. Small heat shock proteins: Simplicity meets complexity. *The Journal of Biological Chemistry* **294**:2121–2132. DOI: <https://doi.org/10.1074/jbc.REV118.002809>, PMID: 30385502
- Hess B**, Bekker H, Berendsen HJC, Fraaije J. 1997. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**:1463–1472. DOI: [https://doi.org/10.1002/\(SICI\)1096-987X\(199709\)18:12<1463::AID-JCC4>3.0.CO;2-H](https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H)
- Huang J**, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, Grubmüller H, MacKerell AD Jr. 2017. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* **14**:71–73. DOI: <https://doi.org/10.1038/nmeth.4067>, PMID: 27819658
- Humphrey W**, Dalke A, Schulten K. 1996. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **14**:33–38. DOI: [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5)
- Iranzo J**, Wolf YI, Koonin EV, Sela I. 2019. Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. *Nature Communications* **10**:5376. DOI: <https://doi.org/10.1038/s41467-019-13429-2>, PMID: 31772262
- Jarvis RA**, Patrick EA. 1973. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers* **C-22**:1025–1034. DOI: <https://doi.org/10.1109/T-C.1973.223640>
- Jaya N**, Garcia V, Vierling E. 2009. Substrate binding site flexibility of the small heat shock protein molecular chaperones. *PNAS* **106**:15604–15609. DOI: <https://doi.org/10.1073/pnas.0902177106>, PMID: 19717454
- Jeffares DC**, Tomiczek B, Sojo V, dos Reis M. 2015. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in A genome. *Methods in Molecular Biology* **1201**:65–90. DOI: [https://doi.org/10.1007/978-1-4939-1438-8\\_4](https://doi.org/10.1007/978-1-4939-1438-8_4), PMID: 25388108
- Jones DT**, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8**:275–282. DOI: <https://doi.org/10.1093/bioinformatics/8.3.275>, PMID: 1633570
- Jorgensen WL**, Chandrasekhar J, Madura JD, Impey RW, Klein ML. 1983. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**:926–935. DOI: <https://doi.org/10.1063/1.445869>
- Kalyaanamoorthy S**, Minh BQ, Wong TKF, von Haeseler A, Jermin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**:587–589. DOI: <https://doi.org/10.1038/nmeth.4285>, PMID: 28481363
- Kempes CP**, van Bodegom PM, Wolpert D, Libby E, Amend J, Hoehler T. 2017. Drivers of bacterial maintenance and minimal energy requirements. *Frontiers in Microbiology* **8**:31. DOI: <https://doi.org/10.3389/fmicb.2017.00031>
- Kennaway CK**, Benesch JLP, Gohlke U, Wang L, Robinson CV, Orlova EV, Saibil HR, Keep NH. 2005. Dodecameric structure of the small heat shock protein Acr1 from *Mycobacterium tuberculosis*. *The Journal of Biological Chemistry* **280**:33419–33425. DOI: <https://doi.org/10.1074/jbc.M504263200>, PMID: 16046399
- Lee GJ**, Roseman AM, Saibil HR, Vierling E. 1997. A small heat shock protein stably binds heat-denatured model substrates and can maintain A substrate in A folding-competent state. *The EMBO Journal* **16**:659–671. DOI: <https://doi.org/10.1093/emboj/16.3.659>, PMID: 9034347
- Lever MA**, Rogers KL, Lloyd KG, Overmann J, Schink B, Thauer RK, Hoehler TM, Jørgensen BB. 2015. Life under extreme energy limitation: a synthesis of laboratory- and field-based investigations. *FEMS Microbiology Reviews* **39**:688–728. DOI: <https://doi.org/10.1093/femsre/fuv020>, PMID: 25994609
- Longo LM**, Despotović D, Weil-Ktorza O, Walker MJ, Jabłońska J, Fridmann-Sirkis Y, Varani G, Metanis N, Tawfik DS. 2020. Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. *PNAS* **117**:15731–15739. DOI: <https://doi.org/10.1073/pnas.2001989117>, PMID: 32561643
- Lucaci AG**, Zehr JD, Enard D, Thornton JW, Kosakovsky Pond SL. 2023. Evolutionary shortcuts via multinucleotide substitutions and their impact on natural selection analyses. *Molecular Biology and Evolution* **40**:msad150. DOI: <https://doi.org/10.1093/molbev/msad150>, PMID: 37395787
- Lynch M**, Marinov GK. 2015. The bioenergetic costs of a gene. *PNAS* **112**:15690–15695. DOI: <https://doi.org/10.1073/pnas.1514974112>
- Mani N**, Bhandari S, Moreno R, Hu L, Prasad BVV, Suguna K. 2016. Multiple oligomeric structures of a bacterial small heat shock protein. *Scientific Reports* **6**:24019. DOI: <https://doi.org/10.1038/srep24019>, PMID: 27053150
- Mirdita M**, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. *Nature Methods* **19**:679–682. DOI: <https://doi.org/10.1038/s41592-022-01488-1>, PMID: 35637307
- Miyamoto S**, Kollman PA. 1992. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry* **13**:952–962. DOI: <https://doi.org/10.1002/jcc.540130805>
- Mogk A**, Deuerling E, Vorderwülbecke S, Vierling E, Bukau B. 2003. Small heat shock proteins, ClpB and the DnaK system form a functional triade in reversing protein aggregation. *Molecular Microbiology* **50**:585–595. DOI: <https://doi.org/10.1046/j.1365-2958.2003.03710.x>, PMID: 14617181
- Nguyen L-T**, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**:268–274. DOI: <https://doi.org/10.1093/molbev/msu300>, PMID: 25371430
- Obuchowski I**, Piróg A, Stolarska M, Tomiczek B, Liberek K. 2019. Duplicate divergence of two bacterial small heat shock proteins reduces the demand for hsp70 in refolding of substrates. *PLOS Genetics* **15**:e1008479. DOI: <https://doi.org/10.1371/journal.pgen.1008479>, PMID: 31652260

- Obuchowski I**, Karaś P, Liberek K. 2021. The small ones matter—sHsps in the bacterial chaperone network. *Frontiers in Molecular Biosciences* **8**:666893. DOI: <https://doi.org/10.3389/fmolb.2021.666893>, PMID: 34055885
- Parrinello M**, Rahman A. 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **52**:7182–7190. DOI: <https://doi.org/10.1063/1.328693>
- Piróg A**, Cantini F, Nierzwicki Ł, Obuchowski I, Tomiczek B, Czub J, Liberek K. 2021. Two bacterial small heat shock proteins, ibpa and ibpb, form a functional heterodimer. *Journal of Molecular Biology* **433**:167054. DOI: <https://doi.org/10.1016/j.jmb.2021.167054>, PMID: 34022209
- Puigbò P**, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biology* **12**:66. DOI: <https://doi.org/10.1186/s12915-014-0066-4>, PMID: 25141959
- Pupko T**, Pe'er I, Hasegawa M, Graur D, Friedman N. 2002. A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. *Bioinformatics* **18**:1116–1123. DOI: <https://doi.org/10.1093/bioinformatics/18.8.1116>, PMID: 12176835
- Ratajczak E**, Zietkiewicz S, Liberek K. 2009. Distinct activities of *Escherichia coli* small heat shock proteins IbpA and IbpB promote efficient protein disaggregation. *Journal of Molecular Biology* **386**:178–189. DOI: <https://doi.org/10.1016/j.jmb.2008.12.009>, PMID: 19101567
- Reinle K**, Mogk A, Bukau B. 2022. The diverse functions of small heat shock proteins in the proteostasis network. *Journal of Molecular Biology* **434**:167157. DOI: <https://doi.org/10.1016/j.jmb.2021.167157>, PMID: 34271010
- Sievers F**, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**:539. DOI: <https://doi.org/10.1038/msb.2011.75>, PMID: 21988835
- Simmons MP**, Ochoterena H. 2000. Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology* **49**:369–381. DOI: <https://doi.org/10.1074/jbc.M111.273847>, PMID: 12118412
- Thomson JM**, Gaucher EA, Burgan MF, De Kee DW, Li T, Aris JP, Benner SA. 2005. Resurrecting ancestral alcohol dehydrogenases from yeast. *Nature Genetics* **37**:630–635. DOI: <https://doi.org/10.1038/ng1553>, PMID: 15864308
- Thornton JW**, Need E, Crews D. 2003. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* **301**:1714–1717. DOI: <https://doi.org/10.1126/science.1086185>, PMID: 14500980
- Van Der Spoel D**, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. 2005. GROMACS: fast, flexible, and free. *Journal of Computational Chemistry* **26**:1701–1718. DOI: <https://doi.org/10.1002/jcc.20291>, PMID: 16211538
- Van Gunsteren WF**, Berendsen HJC. 1988. A leap-frog algorithm for stochastic dynamics. *Molecular Simulation* **1**:173–185. DOI: <https://doi.org/10.1080/08927028808080941>
- Virtanen P**, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, et al. 2020. Author Correction: SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**:261–272. DOI: <https://doi.org/10.1038/s41592-020-0772-5>, PMID: 32094914
- Waskom ML**. 2021. seaborn: statistical data visualization. *Journal of Open Source Software* **6**:3021. DOI: <https://doi.org/10.21105/joss.03021>
- Worth CL**, Gong S, Blundell TL. 2009. Structural and functional constraints in the evolution of protein families. *Nature Reviews. Molecular Cell Biology* **10**:709–720. DOI: <https://doi.org/10.1038/nrm2762>, PMID: 19756040
- Yang Z**. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15**:568–573. DOI: <https://doi.org/10.1093/oxfordjournals.molbev.a025957>, PMID: 9580986
- Yang Z**, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* **19**:908–917. DOI: <https://doi.org/10.1093/oxfordjournals.molbev.a004148>
- Yang Z**. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**:1586–1591. DOI: <https://doi.org/10.1093/molbev/msm088>, PMID: 17483113
- Żwirowski S**, Kłosowska A, Obuchowski I, Nillegoda NB, Piróg A, Zietkiewicz S, Bukau B, Mogk A, Liberek K. 2017. Hsp70 displaces small heat shock proteins from aggregates to initiate protein refolding. *The EMBO Journal* **36**:783–796. DOI: <https://doi.org/10.15252/embj.201593378>, PMID: 28219929