

Evolution towards simplicity in bacterial small heat shock protein system

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2

November 23, 2023 (this version)

Reviewed preprint version 1

August 21, 2023

Sent for peer review

June 22, 2023

Posted to bioRxiv

May 25, 2023

Piotr Karaś, Klaudia Kochanowicz, Marcin Pitek, Przemysław Domanski, Igor Obuchowski, Bartłomiej Tomiczek ✉, Krzysztof Liberek ✉

Intercollegiate Faculty of Biotechnology UG-MUG, University of Gdansk, Abrahama 58, 80-307 Gdansk, Poland

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

Abstract

Evolution can tinker with multi-protein machines and replace them with simpler single-protein systems performing equivalent functions in 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 formation of higher-order oligomers. Upon two mutations in the α -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 system 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.

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.

Gene birth and loss is a hallmark of protein family evolution, however molecular determinants and genetic mechanisms enabling that process are not well understood (Fernandez & Gabaldon, 2020 [↗](#); Worth *et al.*, 2009 [↗](#)). Gene loss and differential retention of paralogues reshapes the divergence of organisms both in *Animalia*, *Archaea* and *Bacteria* (Fernandez & Gabaldon, 2020 [↗](#); Iranzo *et al.*, 2019 [↗](#); Puigbo *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 & 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 (Puigbo *et al.*, 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 low gene count is important for keeping the replication energy costs low (Kempes *et al.*, 2017 [↗](#); Lever *et al.*, 2015 [↗](#); Lynch & 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 small heat shock protein (sHsp) system, which underwent simplification within *Enterobacterales* (which include common bacteria species like *Escherichia coli*, *Salmonella enterica* and *Erwinia amylovora*) as a model (**Fig. 1** [↗](#)).

sHsps are a family of ATP – independent molecular chaperones present in all living organisms with various copy numbers (ten representatives in human) (Haslbeck & 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 are composed of a highly conserved α -crystallin domain (ACD), in a form of so-called β -sandwich, flanked by less conserved, unstructured N – and C – terminal regions (Haslbeck & 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 IXI motif, highly conserved in sHsps C-termini, and the cleft formed by β 4 and β 8 strands of ACD is critical for oligomer formation (Haslbeck & Vierling, 2015 [↗](#); Kennaway *et al.*, 2005 [↗](#); Mani *et al.*, 2016 [↗](#); Strozec *et al.*, 2012 [↗](#)). 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 β 4 – β 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 IbpA and IbpB proteins (Mogk *et al.*, 2003 [↗](#); Obuchowski *et al.*, 2019 [↗](#)). IbpA and IbpB have originated via duplication, form a heterodimer partnership and are functionally divergent from one another (Obuchowski *et al.*, 2019 [↗](#); Pirog *et al.*, 2021 [↗](#)) (**Fig. 1A,B** [↗](#)). IbpA is specialized in tight substrate binding (sequestrase activity), while IbpB is required for 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* (*Erwinaceae*), as a result of *ibpB* gene loss, the secondarily single-protein sHsp (IbpA) system has emerged (**Fig. 1** [↗](#)). The term “secondarily single” is used in order to distinguish it from single-protein IbpA from clades in which the duplication did not occur (for example *Vibrionaceae*) (Obuchowski *et al.*, 2019 [↗](#)). How did IbpA evolve to become independent of its partner? In this study, using ancestral reconstruction, we identify mutations, which allowed the secondarily single IbpA to be fully functional without its partner in substrate sequestration and handover to Hsp70-Hsp100-mediated disaggregation and refolding.

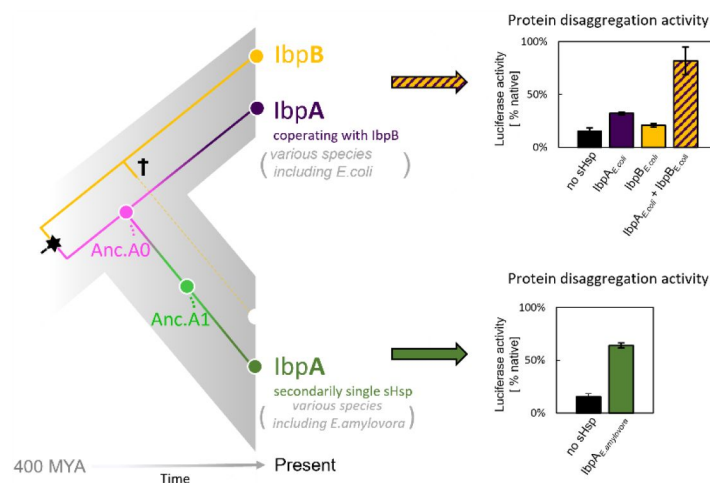


Fig. 1.

sHsp systems in *Enterobacteriaceae* and *Erwinia*

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 *Erwinia* clade is marked with a cross. AncA₀ – reconstructed last common ancestor of IbpA from *Erwinia* and *Enterobacteriaceae*, expressed as a part of two-protein system. AncA₁ – reconstructed last common ancestor of secondarily single IbpA from *Erwinia*. 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 1h refolding at 25 °C and shown as an average of at least three repeats ± standard deviation.

New activity of *Erwiniaceae* IbpA has evolved in parallel to *ibpB* gene loss

To better understand the evolution of sHsps after gene loss, we reconstructed the IbpA ancestors from before and after the loss of its IbpB 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 IbpA sequences from *Enterobacterales* (supplementary file 1), from which we inferred the phylogeny of IbpA using maximum likelihood method (Fig. 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 & Pupko, 2011 [↗](#); Cohen *et al.*, 2008 [↗](#); Pupko *et al.*, 2002 [↗](#); Simmons & Ochoterena, 2000 [↗](#)). Next, we resurrected (i.e., expressed and purified) the last ancestor of IbpA present before (AncA₀) and after (AncA₁) the differential gene loss (Figs. 2 [↗](#), supplementary file 3).

Similarly, to modern-day IbpA proteins both AncA₀ and AncA₁ were fully folded, and reversibly deoligomerized into smaller species under elevated temperature (Fig. 3 – figure supplement 1). Moreover, both ancestral proteins were able to sequester aggregating firefly luciferase in sHsp-substrate assemblies. AncA₀ exhibited sequestrase activity on the level comparable to IbpA from *Escherichia coli* (IbpA_{*E.coli*}). AncA₁ was moderately efficient in this process and IbpA from *Erwinia amylovora* (IbpA_{*E.amyl*}) was the least efficient sequestrase (Fig. 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 (Fig. 3B [↗](#)). Ancestral proteins' interaction with the aggregated substrates was stronger than in the case of extant *E. amylovora* IbpA, but weaker than in the case of extant *E. coli* IbpA (Fig. 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. AncA₁ stimulated luciferase refolding, however its effectiveness was around half of both analyzed extant sHsp systems (single IbpA from *E. amylovora* or IbpA + IbpB system from *E. coli*), similar to extant IbpA from *E. coli* without its IbpB partner. AncA₀, in contrast, inhibited luciferase refolding in comparison to control (no sHsps at substrate aggregation step) (Fig. 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) (Fig. 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 (Fig. 3 – figure supplement 3 A–D). However, the last property (the influence on refolding) required higher Hsp70 system concentration to observe AltAll variants activity (Fig. 3 – figure supplement 3 C,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 A₀ and A₁ nodes.

We performed a molecular evolution analysis to test for positive selection across IbpA phylogeny using both branch models and branch-site models in codeml (Jeffares *et al.*, 2015 [↗](#); Yang, 1998 [↗](#), 2007 [↗](#); Yang & Nielsen, 2002 [↗](#)). The analysis shows a significantly increased ratio of nonsynonymous to synonymous substitutions, after the gene loss, at the branch leading to A₁ with both tests. This result indicates that the new IbpA functionality likely arose due to an episode of positive selection rather than genetic drift (Fig. 4A [↗](#), Supplementary file 4 A,B). The result of the

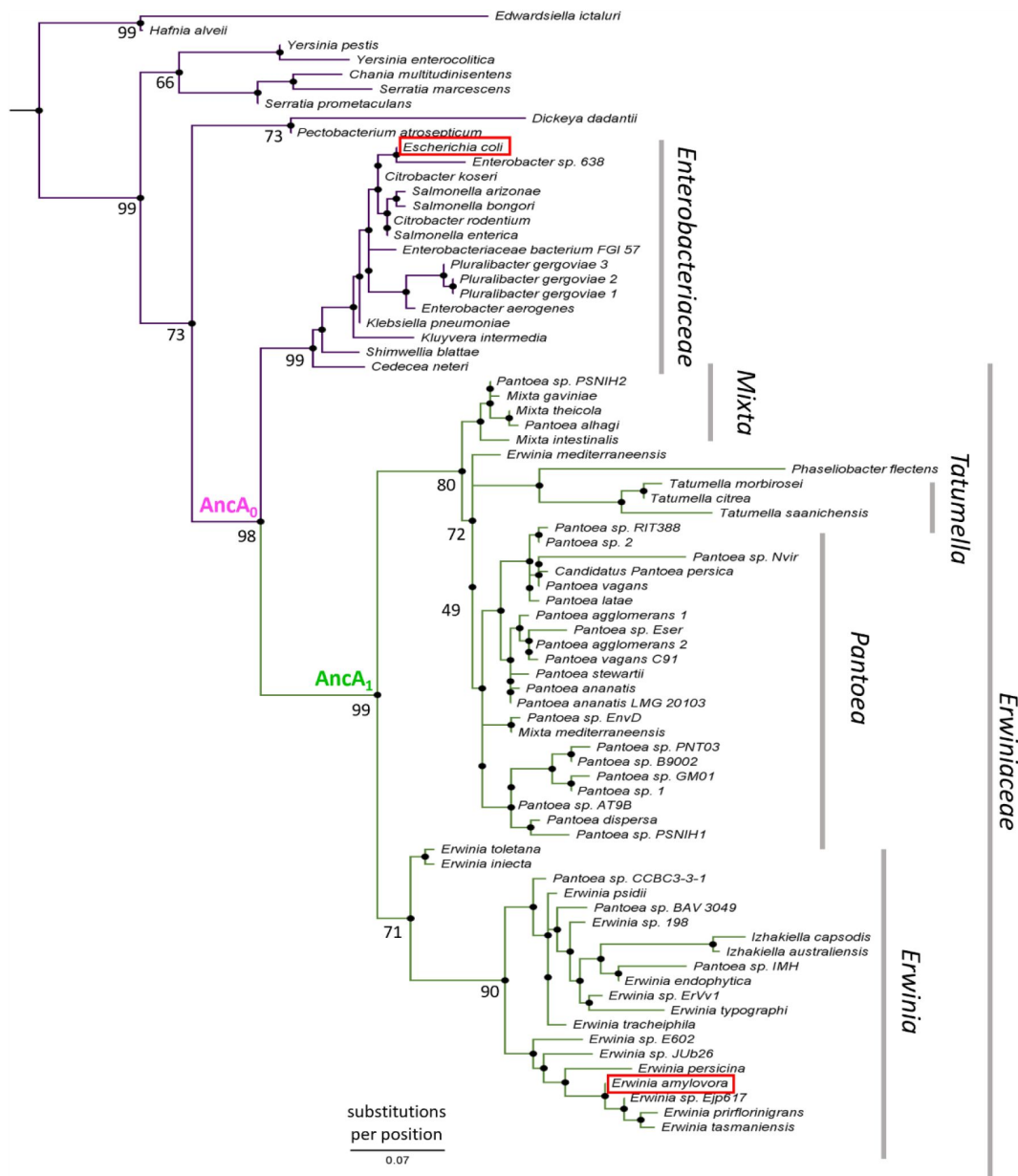


Fig. 2.

IbpA phylogeny in *Enterobacterales*:

Phylogeny was reconstructed from 77 IbpA orthologs from *Enterobacterales* using Maximum Likelihood algorithm with JTT + R3 substitution model. *AncA₀* – node representing the last common ancestor of IbpA from *Erwinia* and *Enterobacteriaceae*. *AncA₁* – node representing the last common ancestor of IbpA from *Erwinia*. Bootstrap support is noted for the major nodes. Extant IbpAs from *E. coli* and *E. amylovora* are marked with a red frame. Scale bar – substitutions per position.

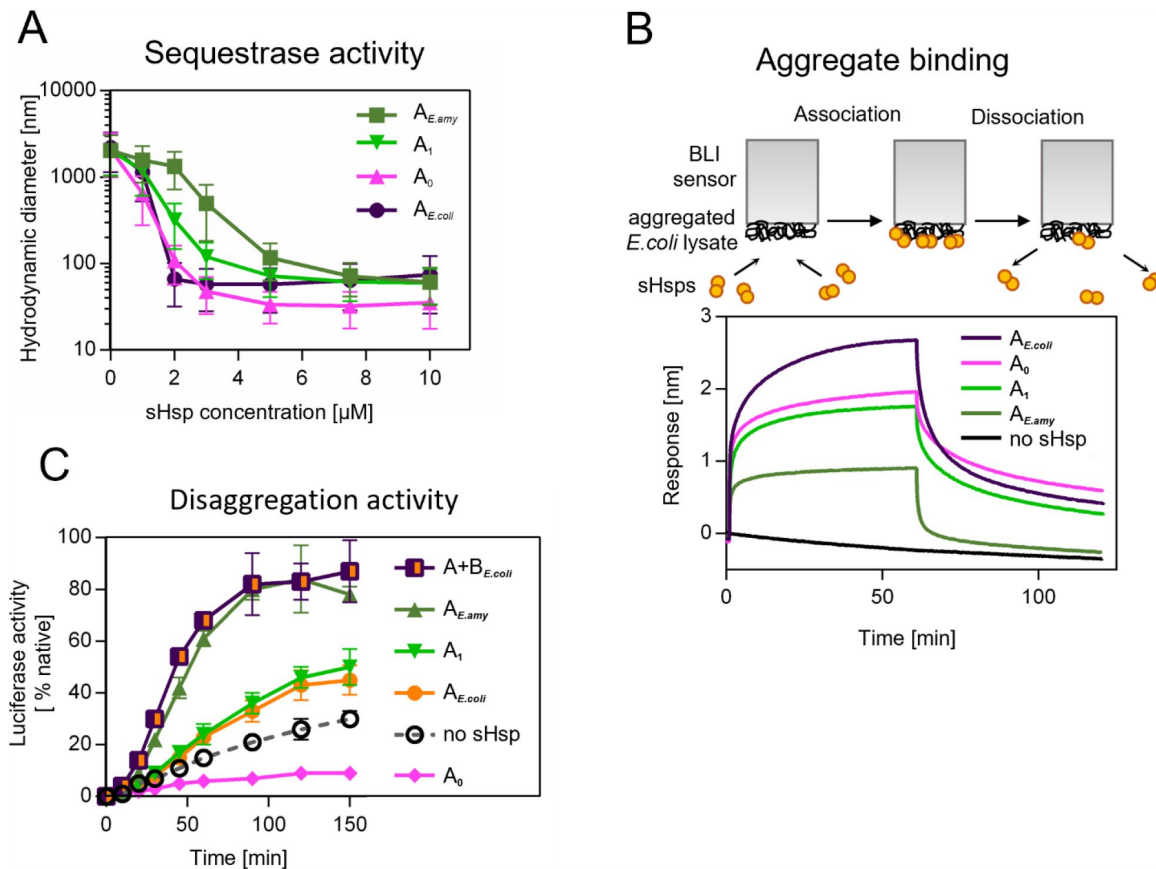


Fig. 3.

Functional changes during the evolution of secondarily single sHsp in *Erwiniaceae*.

(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. Results are shown as average hydrodynamic diameter $\pm$ standard deviation. (B) Binding of extant and ancestral sHsps to heat-aggregated *E. coli* proteins. *E. coli* proteins were heat aggregated and immobilized on a 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 at least three repeats $\pm$ standard deviation.

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 minimum number of mutations that are responsible for a change in functional properties of IbpA.

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 seven out of ten substitutions as probable candidates. Three substitutions were removed from analysis based on the low conservation of these positions in extant proteins (Fig. 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 $AncA_1$ (Fig. 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 (Fig. 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 $AncA_1$ (Fig. 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 (Fig. 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 (Fig. 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 α -crystallin domain properties

Substitutions Q66H and G109D, responsible for gaining single sHsp activity, are located in the α -crystallin domain (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 (Fig. 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 α -crystallin domain (ACD) dimers in complex with C-terminal peptide using AlphaFold2 and *in silico* mutagenesis and subjected them to 0.5 μ 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 overall contact pattern remain similar upon their introduction (Fig. 5 – figure supplement 1) and no major differences in the overall ACD domains structure were observed (Fig. 5 – figure supplement 2). To explore the possibility that identified substitutions affect strength of this interaction, we analyzed binding of purified ACDs of $AncA_0$ and $AncA_0$ Q66H G109D to C-terminal peptide using biolayer interferometry. Titrations of immobilized C-terminal peptide by different ACDs (Figs. 5B [↗](#), 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 μ M to 7.1 μ 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 formation of sHsp oligomers (Fu *et al.*, 2005 [↗](#); Mani *et al.*, 2016 [↗](#); Strozecka *et al.*, 2012 [↗](#)), we used dynamic light scattering to investigate how Q66H and G109D substitutions influence the size of oligomers formed by $AncA_0$ 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 $AncA_0$ deoligomerization (Figure 5 – figure supplement 4 A,B).

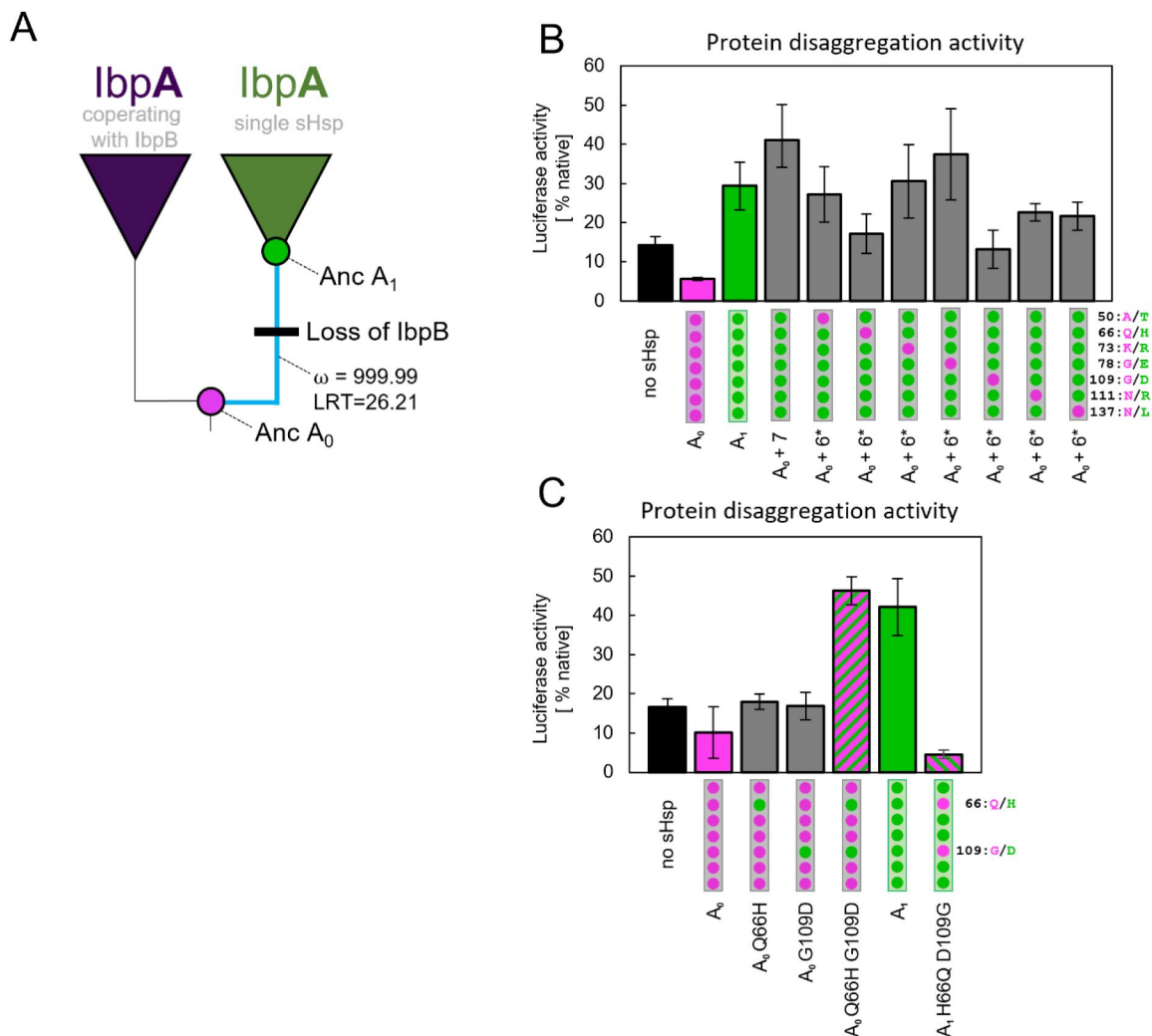


Fig. 4.

Substitutions at positions 66 and 109 that occurred between nodes A_0 and A_1 are crucial for ancestral sHsps to work as a single protein.

Luciferase refolding assay was performed as in [fig. 1](#). Activity of luciferase was measured after 1h refolding at 25 °C and shown as an average of at least three repeats $\pm$ standard deviation. (A) Schematic phylogeny of *Enterobacteriales* IbpA showing increased ratio of nonsynonymous to synonymous substitutions (ω) on the branch between nodes $AncA_0$ and $AncA_1$. Loss of cooperating IbpB is marked on a tree. Value of Likelihood Ratio Test (LRT) is given for the selection model. (B) Identification of substitutions necessary for $AncA_0$ to obtain $AncA_1$ – like activity in luciferase disaggregation; seven candidate mutations were introduced into $AncA_0$ ($AncA_0 + 7$); subsequently, in series of six mutants, each of the candidate positions was reversed to a more ancestral state ($AncA_0 + 6^*$ variants) (C) Effect of substitutions at positions 66 and 109 on the ability of $AncA_0$ and $AncA_1$ to stimulate luciferase refolding.

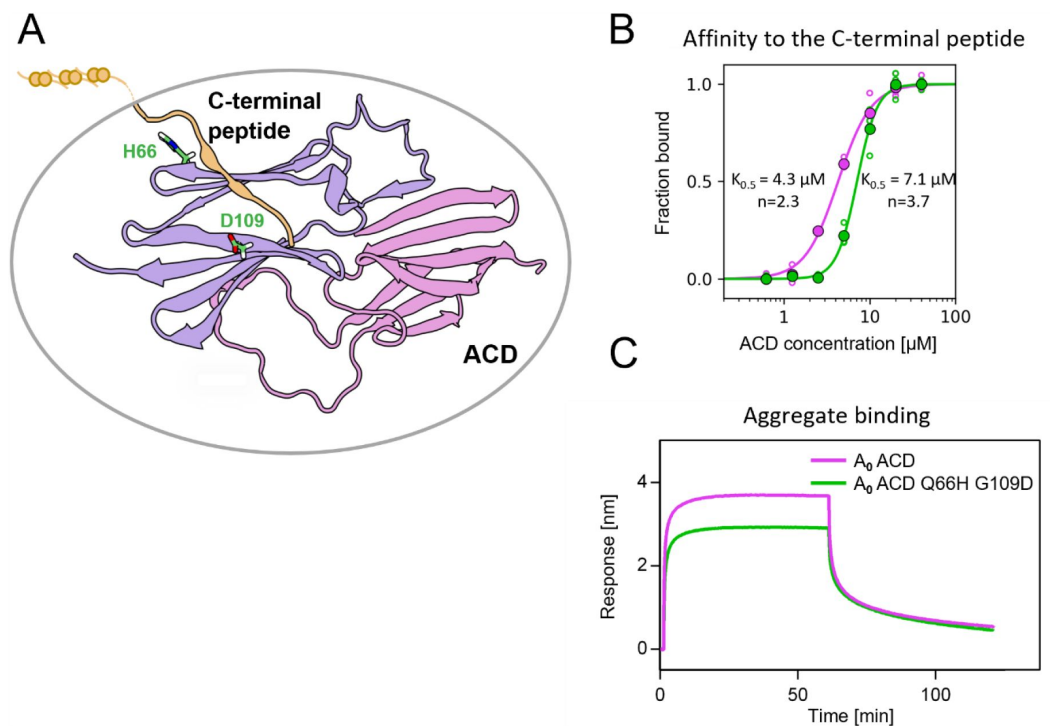


Fig. 5.

Substitutions at positions 66 and 109 decreased the affinity of AncA₀ ACD to C-terminal peptide and aggregated substrate.

(A) Structural model of complex formed by AncA₀ Q66H G109D α-crystallin domain dimer (purple and lilac) and AncA₀ C-terminal peptide (orange). (B) Effect of Q66H G109D substitutions (green) on AncA₀ (purple) ACD's affinity to the C-terminal peptide assayed by 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 cooperative binding model (Hill equation). Values of fitted binding affinities [$K_{0.5}$] (AncA₀ 4.3 ± 0.2 μM, AncA₀ Q66H G109D 7.1 ± 0.2 μM) and Hill coefficients [n] (AncA₀ 2.3 ± 0.17, AncA₀ Q66H G109D 3.7 ± 0.34) are indicated on the plot. (C) Effect of Q66H G109D substitutions on AncA₀ ACD's affinity to aggregated *E. coli* proteins bound to BLI sensor. Analysis was performed as in Fig. 3A.

β 4- β 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 ACD of AncA₀ binds protein aggregates and whether this interaction is influenced by Q66H G109D substitutions. We observed that AncA₀ ACD efficiently binds to either aggregated *E. coli* lysate or aggregated luciferase, and this binding was weakened by analyzed substitutions (Figs. 5C, 5 – figure supplement 5). This suggests that ACD of bacterial sHsps interacts with the substrate, most likely through the β 4- β 8 cleft.

These results allow us to conclude that substitutions Q66H and G109D in AncA₀ substantially increased the sHsp ability to stimulate Hsp70-Hsp100-dependent substrate refolding by weakening the interaction of β 4- β 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 (Fig. 5 – figure supplement 5 B,C).

Identified substitutions define the mode of action of extant sHsps

As more ancestral, AncA₀-like state in positions 66 and 109 is conserved in IbpA of *E. coli* while more modern, AncA₁-like state is conserved in IbpA of *E. amylovora*, we decided to ask whether this difference is sufficient to explain functional differences between the two extant proteins. Therefore, we introduced AncA₁-like substitutions into IbpA_{*E. coli*} and AncA₀-like substitutions into IbpA_{*E. amylovora*}. Resulting IbpA_{*E. coli*} Q66H G109D, in comparison to wild type IbpA_{*E. coli*}, exhibited 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* IbpA. At the same time, IbpA_{*E. amylovora*} H67Q D110G significantly less efficiently stimulated luciferase refolding, while exhibiting increased ability to bind aggregated substrates in comparison to wild type IbpA_{*E. amylovora*} (Fig. 6A-C [↗](#)). Still, both new IbpA variants exhibited properties characteristic for sHsps, namely reversible thermal deoligomerization and sequestrase activity (Fig. 6 - figure supplement 1 A,B).

Above results suggest that tight sHsp binding to aggregates negatively affects 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 requires sHsps to be outcompeted from aggregates (Zwirowski *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 (Fig. 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 IbpA_{*E. amylovora*} or IbpA_{*E. coli*} Q66H G109D on aggregates only weakly inhibits Hsp70 binding (Fig. 6E,H [↗](#)). In contrast, the inhibition is much more pronounced when IbpA_{*E. amylovora*} H67Q D110G or IbpA_{*E. coli*} are present on aggregates (Fig. 6F,G [↗](#)).

All above experiments indicate that two specific amino acids in positions 66 and 109 in ACD of IbpA proteins define the mode of IbpA activity. Glutamine 66 and glycine 109 are characteristic for IbpA proteins which bind tightly to substrates and thus are not easily outcompeted from the aggregates by Hsp70s. Such IbpAs require IbpB 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 IbpB.

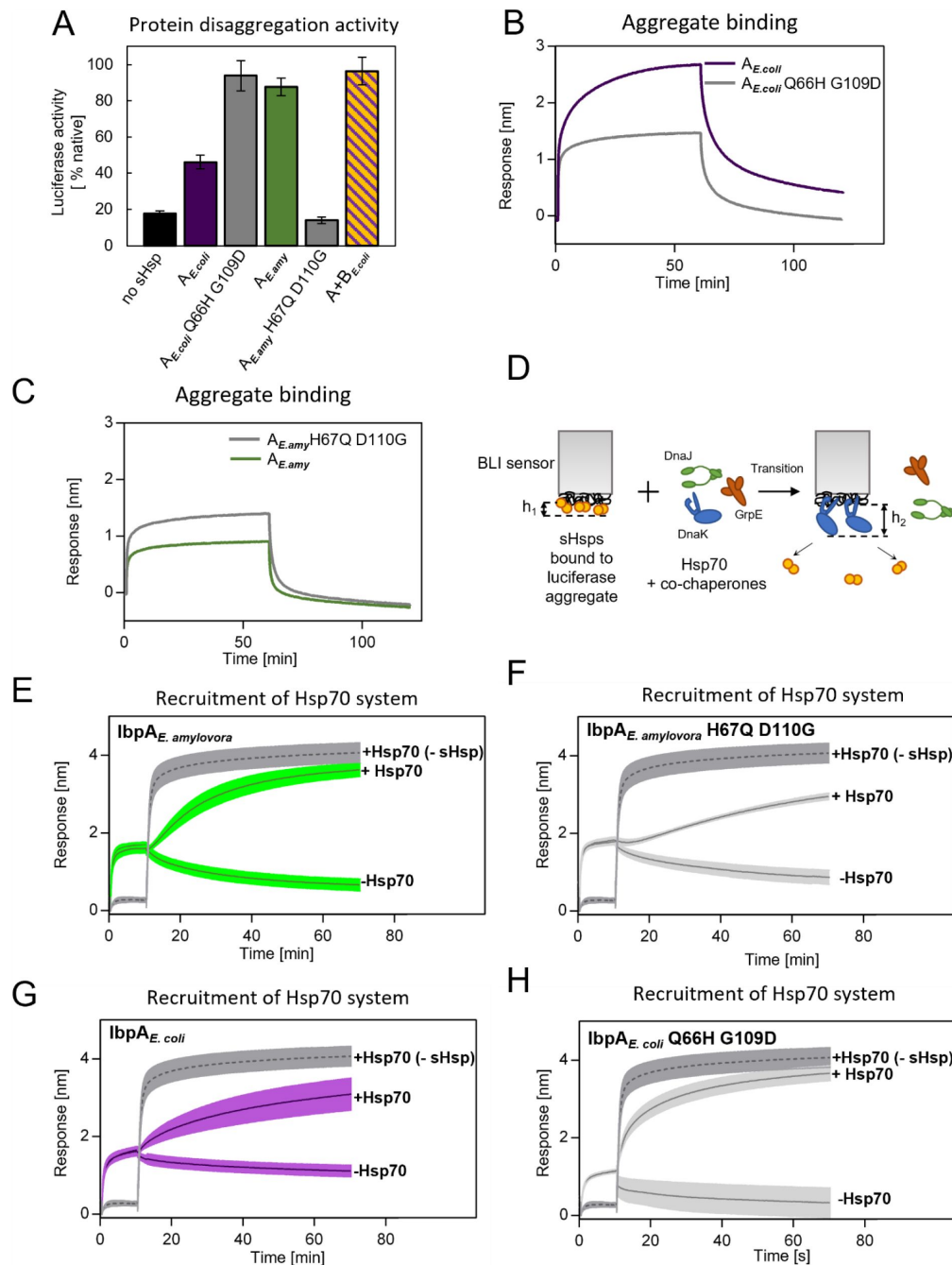


Fig. 6.

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 Fig. 1B. Activity of luciferase was measured after 1h refolding at 25 °C and shown as an average of at least three repeats $\pm$ 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. coli* proteins. Assay was performed as in 3A. (D-H) Effect of substitutions at analyzed positions on inhibition of Hsp70 system binding to aggregates by extant sHsps (D) Experimental scheme. (E-H) Aggregate-bound sHsps differently inhibit Hsp70 binding. BLI sensor with immobilized aggregated luciferase and aggregate bound sHsps was incubated with Hsp70 or buffer (spontaneous dissociation curve). Grey traces present Hsp70 binding to immobilized aggregates in the absence of sHsps. Results are presented as an average of at least three repeats $\pm$ standard deviation.

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 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; Pirog *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 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 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 β 4- β 8 cleft to aggregated substrates as well as to the C-termini of the other sHsps. These interactions might be of particular importance for 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 addition of free C-terminal peptide causes *E. coli* IbpA and IbpB dissociation from the outer shell of sHsp - substrate complex (Zwirowski *et al.*, 2017). It was also shown that, in case of Hsp 16.6 from cyanobacterium *Synechocystis*, substitutions that slightly weakened 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 & 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 & 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 vertical approach in biochemical studies. This study closely follows the evolutionary process, in which mutations in one of the two cooperating proteins tinker it to a point 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 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 (Natarajan *et al.*, 2023 [DOI](#)) the context of *E. coli* IbpA protein did not influence the ability of IbpA_{E.coli} 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

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 [DOI](#)). 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 (Kalyaanamoorthy *et al.*, 2017 [DOI](#); Nguyen *et al.*, 2015 [DOI](#)). 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 [DOI](#)).

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 [DOI](#); Cohen & Pupko, 2011 [DOI](#); Cohen *et al.*, 2008 [DOI](#); Jones *et al.*, 1992 [DOI](#); Pupko *et al.*, 2002 [DOI](#); Simmons & Ochoterena, 2000 [DOI](#)).

Alternative ancestral sequences for AncA₀ and AncA₁ proteins were obtained by substituting most likely amino acid on every uncertain position (defined as a position with more than one amino acid with posterior probability ≥ 0.2) with the amino acid with the second highest posterior probability (Eick *et al.*, 2017 [DOI](#)).

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 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 AncA₀-AncA₁ 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

Protein Purification

Purification of IbpA proteins

pET3a plasmids containing *ancA₀*, *ancA₁*, *ancA₀+7*, *ancA_{0 alt_all}*, *ancA_{1 alt_all}* and *ibpA_{Ea}* genes were ordered from GeneScript. 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, 5mM β-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 β-mercaptoethanol, 6M 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, 50mM NaCl, 10% glycerol, 5mM β-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 β-mercaptoethanol).

Purification of ACD domains

ACDs of IbpA_{Ec}, AncA₀ and AncA₀ Q66H G109D were purified as described previously (Pirog *et al.*, 2021 [DOI](#)) and as a final step dialyzed to buffer G (50 mM Tris pH 7.5, 150mM KCl, 5mM β-mercaptoethanol).

Purification of His₆ – SUMO and His₆ –SUMO-C-terminal peptide of AncA₀ construct

His₆ - SUMO was purified using the Champion™ pET SUMO Expression System. pET28a plasmid containing gene encoding His₆ –SUMO fused with C-terminal peptide of AncA₀ (PEAMKPPRIEIN) was ordered from GeneScript. Proteins were overproduced in *E. coli* BL21(DE3). Cells were then lysed by sonication in Qsonica sonicator (20% amplitude, 2 min process time, 5s pulse-ON time, 10s pulse-OFF time) in lysis buffer L2 (40 mM Tris pH 7.5, 100 mM NaCl, 10 % glycerol, 10 mM imidazole, 2 mM β-mercaptoethanol). Insoluble fractions were separated by centrifugation for 30 min at 70 000 x g and supernatants, containing proteins of interest, were incubated for 1h 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, 2mM β-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, 2mM β-mercaptoethanol) and then dialyzed to buffer C (as above).

DnaK, DnaJ, GrpE, ClpB and IbpA_{Ec} proteins were purified as described previously (Ratajczak *et al.*, 2009 [DOI](#)). IbpB_{Ec} protein was purified as described previously (Pirog *et al.*, 2021 [DOI](#)). His-tagged luciferase used for BLI measurements was purified as described previously (Obuchowski *et al.*, 2019 [DOI](#)).

Purity of purified proteins was assessed with SDS-PAGE electrophoresis with Coomassie Blue staining. Protein concentrations were measured using Bradford reaction, with Bovine Serum Albumin as a standard. In the case of His₆ – SUMO fused with C-terminal peptide of AncA₀, concentration was measured with SDS-PAGE electrophoresis with Coomassie Blue staining coupled with densitometric analysis with Bovine Serum Albumin used as a standard.

Luciferase refolding assay

1.5 μM recombinant firefly luciferase in buffer F (50 mM Tris pH 7.5, 150 mM KCl, 20 mM MgCl_2 , 2.5 mM DTT) was denatured by incubation for 10 min at 44°C alone or in the presence of 10 μM sHsps (3 μM IbpA_{Ec} + 7 μM IbpB_{Ec} in the case of two-protein system from *E. coli*). Denatured luciferase was then incubated at 25 °C with Hsp70 system (1 μM DnaK, 0.4 μM DnaJ and 0.3 μM GrpE), 2 μM ClpB, and ATP regeneration system (5 mM ATP, 0.1 mg/ml creatine kinase and 18 mM creatine phosphate). For experiment presented in Fig.3 – figure supplement 3D, higher concentration of the Hsp70 system was used (2 μM DnaK, 0.8 μM DnaJ and 0.6 μM 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 μl sample volume, scattering angle 173° and wavelength of 633 nm. For every measurement, minimum 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, as previously described (Zwirowski *et al.*, 2017 [\[3\]](#)).

For reversible deoligomerization assay, size of oligomers formed by 10 μM sHsps in buffer F (50 mM Tris pH7.5, 150 mM KCl, 20 mM 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 influence of substitutions on oligomer formation, size of oligomers formed by either 10 μM AncA₀ or 10 μM AncA₀ 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 maximum of a dominant peak of size distribution by volume plotted against temperature $\pm$ standard deviation.

For assembly formation (sequestrase) assay, 1.5 μM 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 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 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 MgCl_2 , 5 mM β -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 1h in 5 μ M sHsp solution in H buffer to measure sHsps association. Sensors were then moved for 1h 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 pH7.5, 150 mM KCl, 20 mM MgCl₂, 2.5 mM DTT) instead of buffer H. Baseline biolayer was measured for 60 s. Sensors were then placed in 5 μ 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 μ M DnaK, 0.28 μ M DnaJ, 0.21 μ M GrpE, 5 mM ATP, 0.1 mg/ml creatine kinase, 18 mM creatine phosphate). Hsp70 system binding and sHsps dissociation were measured for 1h. 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 (50mM Tris pH 7.5, 150 mM KCl, 5% glycerol, 5 mM β -mercaptoethanol) and baseline signal was measured for 60 s. Sensors were then placed in 2.5 μ M His₆-SUMO-C-peptide solution in buffer C and incubated for 15 min. Surplus His₆-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 His₆-SUMO for His₆-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 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 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 & 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 & Kollman, 1992). Integration of equations of motion was performed by leap-frog algorithm (Van Gunsteren & Berendsen, 1988) with a time step of 2 fs. Periodic boundary conditions were applied in all dimensions.

The initial conformation of AncA₀ ACD - C-terminal peptide complex was predicted by ColabFold implementation (Mirdita *et al*, 2022) of AlphaFold-Multimer (Richard *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 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 $1000 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$. 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/> [↗](#)). Contact between interfacial residues was defined as any of the following types of interaction: hydrogen bond, ionic, π -stacking, π -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 “gmxcuster” tool and Jarvis-Patrick clustering method (Jarvis & Patrick, 1973 [↗](#)). RMSD cutoff of 0.2 nm, was used for Jarvis-Patrick algorithm and conformations possessing at least 3 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 to 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) (William *et al*, 1996 [↗](#)) and Blender (<https://www.blender.org/> [↗](#)) were used for structure visualization.

Acknowledgements

This work was supported by a grant of the Polish National Science Centre (OPUS 17 2019/33/B/NZ1/00352). We gratefully acknowledge Poland’s high-performance Infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH, PCSS, CI TASK, WCSS) for providing computer facilities and support. We thank Prof. Max Telford, Prof. Jaroslaw Marszalek and Dr. Agnieszka Kłosowska for helpful discussions.

Additional Files

Figure supplements

Supplementary file 1 Multiple sequence alignment of *Enterobacterales* IbpA orthologs used for phylogenetic analysis and ancestral reconstruction in fasta format

Supplementary file 2 Phylogenetic tree of *Enterobacterales* IbpA protein family in newick format (see **figure 2** [↗](#))

Supplementary file 3 Posterior probability statistics for ancestral sequence reconstruction of Anca₀ and Anca₁ 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 AncA₀ and AncA₁ are marked in bold and italics. Posterior probabilities estimated using FastML program based on 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 AncA₀ and AncA₁ or entire *Erwiniaceae* clade as foreground; NS – Not significant. B) Branch - site model statistics for IbpA orthologs from *Enterobacterales*. Models assumed either branch between nodes AncA₀ and AncA₁ or entire *Erwiniaceae* clade as foreground; NS – Not significant

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 Res* **40**:W580–584 <https://doi.org/10.1093/nar/gks498>
- 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** *J Biol Chem* **281**:39943–39952 <https://doi.org/10.1074/jbc.M607677200>
- Bussi G, Donadio D, Parrinello M (2007) **Canonical sampling through velocity rescaling** *Journal of Chemical Physics* **126** <https://doi.org/10.1063/1.2408420>
- Cohen O, Pupko T (2011) **Inference of gain and loss events from phyletic patterns using stochastic mapping and maximum parsimony--a simulation study** *Genome Biol Evol* **3**:1265–1275 <https://doi.org/10.1093/gbe/evr101>
- Cohen O, Rubinstein ND, Stern A, Gophna U, Pupko T (2008) **A likelihood framework to analyse phyletic patterns** *Philos Trans R Soc Lond B Biol Sci* **363**:3903–3911 <https://doi.org/10.1098/rstb.2008.0177>
- Demuth JP, Hahn MW (2009) **The life and death of gene families** *Bioessays* **31**:29–39 <https://doi.org/10.1002/bies.080085>
- 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** *Curr Biol* **26**:2434–2445 <https://doi.org/10.1016/j.cub.2016.07.038>
- Eick GN, Bridgham JT, Anderson DP, Harms MJ, Thornton JW (2017) **Robustness of Reconstructed Ancestral Protein Functions to Statistical Uncertainty** *Mol Biol Evol* **34**:247–261 <https://doi.org/10.1093/molbev/msw223>
- Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG, Doi 10.1063/1.470117 (1995) **A Smooth Particle Mesh Ewald Method** *Journal of Chemical Physics* **103**:8577–8593
- Fernandez R, Gabaldon T (2020) **Gene gain and loss across the metazoan tree of life** *Nat Ecol Evol* **4**:524–533 <https://doi.org/10.1038/s41559-019-1069-x>
- Fu X, Zhang H, Zhang X, Cao Y, Jiao W, Liu C, Song Y, Abulimiti A, Chang Z (2005) **A dual role for the Nterminal region of Mycobacterium tuberculosis Hsp16.3 in self-oligomerization and binding denaturing substrate proteins** *J Biol Chem* **280**:6337–6348 <https://doi.org/10.1074/jbc.M406319200>
- 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** *Biochem J* **425**:245–255 <https://doi.org/10.1042/BJ20090907>
- Gaucher EA, Govindarajan S, Ganesh OK (2008) **Palaeotemperature trend for Precambrian life inferred from resurrected proteins** *Nature* **451**:704–707 <https://doi.org/10.1038/nature06510>

- 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** *J Biol Chem* **277**:46310–46318 <https://doi.org/10.1074/jbc.M208926200>
- Haslbeck M, Vierling E (2015) **A first line of stress defense: small heat shock proteins and their function in protein homeostasis** *J Mol Biol* **427**:1537–1548 <https://doi.org/10.1016/j.jmb.2015.02.002>
- Haslbeck M, Weinkauff S, Buchner J (2019) **Small heat shock proteins: Simplicity meets complexity** *Journal of Biological Chemistry* **294**:2121–2132 <https://doi.org/10.1074/jbc.REV118.002809>
- Hess B, Bekker H, Berendsen HJC, Fraaije JGEM, Doi 10.1002/(Sici)1096-987x(199709)18:12<1463::Aid-Jcc4>3.3.Co;2-L (1997) **LINCS: A linear constraint solver for molecular simulations** *Journal of Computational Chemistry* **18**:1463–1472
- Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, Grubmüller H, MacKerell AD (2017) **CHARMM36m: an improved force field for folded and intrinsically disordered proteins** *Nature Methods* **14**:71–73 <https://doi.org/10.1038/nmeth.4067>
- 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** *Nat Commun* **10** <https://doi.org/10.1038/s41467-019-13429-2>
- Jarvis RA, Patrick EA (1973) **Clustering Using a Similarity Measure Based on Shared near Neighbors** *Ieee T Comput* :1025–1034 <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** *Proc Natl Acad Sci U S A* **106**:15604–15609 <https://doi.org/10.1073/pnas.0902177106>
- 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 Mol Biol* **1201**:65–90 https://doi.org/10.1007/978-1-4939-1438-8_4
- Jones DT, Taylor WR, Thornton JM (1992) **The rapid generation of mutation data matrices from protein sequences** *Comput Appl Biosci* **8**:275–282 <https://doi.org/10.1093/bioinformatics/8.3.275>
- Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) **Comparison of Simple Potential Functions for Simulating Liquid Water** *Journal of Chemical Physics* **79**:926–935 <https://doi.org/10.1063/1.445869>
- Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermini LS (2017) **ModelFinder: fast model selection for accurate phylogenetic estimates** *Nat Methods* **14**:587–589 <https://doi.org/10.1038/nmeth.4285>
- 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**10
- Kennaway CK, Benesch JL, 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** *J Biol Chem* **280**:33419–33425 <https://doi.org/10.1074/jbc.M504263200>

- 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** *EMBO J* **16**:659–671 <https://doi.org/10.1093/emboj/16.3.659>
- 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 <https://doi.org/10.1093/femsre/fuv020>
- Longo LM, Despotovic D, Weil-Ktorza O, Walker MJ, Jablonska 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** *Proc Natl Acad Sci U S A* **117**:15731–15739 <https://doi.org/10.1073/pnas.2001989117>
- Lucaci AG, Zehr JD, Enard D, Thornton JW, Kosakovsky Pond SL (2023) **Evolutionary Shortcuts via Multinucleotide Substitutions and Their Impact on Natural Selection Analyses** *Mol Biol Evol* **40**:10
- Lynch M, Marinov GK (2015) **The bioenergetic costs of a gene** *Proceedings of the National Academy of Sciences* **112**:15690–15695 <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** *Sci Rep* **6** <https://doi.org/10.1038/srep24019>
- 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 <https://doi.org/10.1038/s41592-022-01488-1>
- Miyamoto S, Kollman PA, DOI 10.1002/jcc.540130805 (1992) **Settle - an Analytical Version of the Shake and Rattle Algorithm for Rigid Water Models** *Journal of Computational Chemistry* **13**:952–962
- Mogk A, Deuerling E, Vorderwulbecke S, Vierling E, Bukau B (2003) **Small heat shock proteins, ClpB and the DnaK system form a functional triade in reversing protein aggregation** *Mol Microbiol* **50**:585–595
- Natarajan C, Signore AV, Bautista NM, Hoffmann FG, Tame JRH, Fago A, Storz JF (2023) **Evolution and molecular basis of a novel allosteric property of crocodilian hemoglobin** *Curr Biol* **33**:98–108 <https://doi.org/10.1016/j.cub.2022.11.049>
- Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies** *Mol Biol Evol* **32**:268–274 <https://doi.org/10.1093/molbev/msu300>
- Obuchowski I, Karas P, Liberek K (2021) **The Small Ones Matter-sHsps in the Bacterial Chaperone Network** *Front Mol Biosci* **8** <https://doi.org/10.3389/fmolb.2021.666893>
- Obuchowski I, Pirog 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 Genet* **15** <https://doi.org/10.1371/journal.pgen.1008479>
- Parrinello M, Rahman A, Doi 10.1063/1.328693 (1981) **Polymorphic Transitions in Single-Crystals - a New Molecular-Dynamics Method** *Journal of Applied Physics* **52**:7182–7190

- Pirog A, Cantini F, Nierzwicki L, Obuchowski I, Tomiczek B, Czub J, Liberek K (2021) **Two Bacterial Small Heat Shock Proteins, IbpA and IbpB, Form a Functional Heterodimer** *J Mol Biol* **433** <https://doi.org/10.1016/j.jmb.2021.167054>
- Puigbo P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV (2014) **Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes** *BMC Biol* **12** <https://doi.org/10.1186/s12915-014-0066-4>
- 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 <https://doi.org/10.1093/bioinformatics/18.8.1116>
- Ratajczak E, Zietkiewicz S, Liberek K (2009) **Distinct activities of Escherichia coli small heat shock proteins IbpA and IbpB promote efficient protein disaggregation** *J Mol Biol* **386**:178–189 <https://doi.org/10.1016/j.jmb.2008.12.009>
- Reinle K, Mogk A, Bukau B (2022) **The Diverse Functions of Small Heat Shock Proteins in the Proteostasis Network** *J Mol Biol* **434** <https://doi.org/10.1016/j.jmb.2021.167157>
- Richard E, Michael ON, Alexander P, Natasha A, Andrew S, Tim G, Augustin Ž, Russ B, Sam B, Jason Y, et al (2022) **Protein complex prediction with AlphaFold-Multimer** *bioRxiv* <https://doi.org/10.1101/2021.10.04.463034>
- Sievers F et al. (2011) **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega** *Mol Syst Biol* **7** <https://doi.org/10.1038/msb.2011.75>
- Simmons MP, Ochoterena H (2000) **Gaps as Characters in Sequence-Based Phylogenetic Analyses** *Systematic Biology* **49**:369–381
- Strozecka J, Chrusciel E, Gorna E, Szymanska A, Zietkiewicz S, Liberek K (2012) **Importance of N- and Cterminal regions of IbpA, Escherichia coli small heat shock protein, for chaperone function and oligomerization** *J Biol Chem* **287**:2843–2853 <https://doi.org/10.1074/jbc.M111.273847>
- Thomson JM, Gaucher EA, Burgan MF, De Kee DW, Li T, Aris JP, Benner SA (2005) **Resurrecting ancestral alcohol dehydrogenases from yeast** *Nat Genet* **37**:630–635 <https://doi.org/10.1038/ng1553>
- Thornton JW, Need E, Crews D (2003) **Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling** *Science* **301**:1714–1717 <https://doi.org/10.1126/science.1086185>
- Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) **GROMACS: fast, flexible, and free** *J Comput Chem* **26**:1701–1718 <https://doi.org/10.1002/jcc.20291>
- Van Gunsteren WF, Berendsen HJC (1988) **A Leap-Frog Algorithm for Stochastic Dynamics** *Mol Simulat* **1**:173–185 <https://doi.org/10.1080/08927028808080941>
- Virtanen P et al. (2020) **SciPy 1.0: fundamental algorithms for scientific computing in Python** *Nat Methods* **17**:261–272 <https://doi.org/10.1038/s41592-019-0686-2>
- Waskom ML (2021) **seaborn: statistical data visualization** *Journal of Open Source Software* **6** <https://doi.org/10.21105/joss.03021>

- William H, Andrew D, Klaus S (1996) **VMD: Visual molecular dynamics** *Journal of Molecular Graphics* **14**:33–38 [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5)
- Worth CL, Gong S, Blundell TL (2009) **Structural and functional constraints in the evolution of protein families** *Nat Rev Mol Cell Biol* **10**:709–720 <https://doi.org/10.1038/nrm2762>
- Yang Z (1998) **Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution** *Mol Biol Evol* **15**:568–573 <https://doi.org/10.1093/oxfordjournals.molbev.a025957>
- Yang Z (2007) **PAML 4: phylogenetic analysis by maximum likelihood** *Mol Biol Evol* **24**:1586–1591 <https://doi.org/10.1093/molbev/msm088>
- Yang Z, Nielsen R (2002) **Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages** *Mol Biol Evol* **19**:908–917 <https://doi.org/10.1093/oxfordjournals.molbev.a004148>
- Zwirowski S, Klosowska A, Obuchowski I, Nillegoda NB, Pirog A, Zieatkiewicz S, Bukau B, Mogk A, Liberek K (2017) **Hsp70 displaces small heat shock proteins from aggregates to initiate protein refolding** *Embo Journal* **36**:783–796 <https://doi.org/10.15252/embj.201593378>

Article and author information

Piotr Karaś

Intercollegiate Faculty of Biotechnology UG-MUG, University of Gdansk, Abrahama 58, 80-307 Gdansk, Poland
ORCID iD: [0000-0001-5270-7938](https://orcid.org/0000-0001-5270-7938)

Klaudia Kochanowicz

Intercollegiate Faculty of Biotechnology UG-MUG, University of Gdansk, Abrahama 58, 80-307 Gdansk, Poland
ORCID iD: [0000-0002-2041-0301](https://orcid.org/0000-0002-2041-0301)

Marcin Pitek

Intercollegiate Faculty of Biotechnology UG-MUG, University of Gdansk, Abrahama 58, 80-307 Gdansk, Poland
ORCID iD: [0000-0002-1300-4364](https://orcid.org/0000-0002-1300-4364)

Przemyslaw Domanski

Intercollegiate Faculty of Biotechnology UG-MUG, University of Gdansk, Abrahama 58, 80-307 Gdansk, Poland

Igor Obuchowski

Intercollegiate Faculty of Biotechnology UG-MUG, University of Gdansk, Abrahama 58, 80-307 Gdansk, Poland
ORCID iD: [0000-0001-6350-1273](https://orcid.org/0000-0001-6350-1273)

Bartłomiej Tomiczek

Intercollegiate Faculty of Biotechnology UG-MUG, University of Gdansk, Abrahama 58, 80-307 Gdansk, Poland

For correspondence: bartlomiej.tomiczek@biotech.ug.edu.pl

ORCID iD: [0000-0001-9295-663X](https://orcid.org/0000-0001-9295-663X)

Krzysztof Liberek

Intercollegiate Faculty of Biotechnology UG-MUG, University of Gdansk, Abrahama 58, 80-307 Gdansk, Poland

For correspondence: krzysztof.liberek@ug.edu.pl

ORCID iD: [0000-0002-7532-9279](https://orcid.org/0000-0002-7532-9279)

Copyright

© 2023, 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.

Editors

Reviewing Editor

Christian Landry

Université Laval, Canada

Senior Editor

Christian Landry

Université Laval, Canada

Reviewer #1 (Public Review):

The work in this paper is in general done carefully. Reconstructions are done appropriately and the effects of statistical uncertainty are quantified properly. I was glad to see that the tree and alignment are now deposited.

The paper identifies which mutations are crucial for the functional differences between the ancestors tested. This is done quite carefully - the authors even show that the same substitutions also work in extant proteins.

These substitutions very slightly lower the affinity and increase the cooperativity of the C-terminal peptide binding to the alpha crystallin domain - a key oligomeric interaction. These relatively minor changes nevertheless apparently affect the subunit exchange behaviour and oligomerization of the sHSP.

Lastly, the authors use likelihood methods to test for signatures of selection. This reviewer is not a fan of these methods, as they are easily misled by common biological processes (see PMID 37395787 for a recent critique). The paper is relatively careful in the interpretation of this test though, and I think the importance of the other findings does not depend on the action of selection along this branch.

Reviewer #2 (Public Review):

This was an interesting study, and I enjoyed seeing different experimental approaches used to compare the properties of the different native proteins, the ancestral reconstructions, and

the other mutants. In the original manuscript, I felt that the authors had over-simplified their explanations, as the differences between the ancestral proteins, and the changes induced by the two mutations, only partially explain the differences between IbpA proteins from the two different species. However, with their revised version, I think the presentation and discussion of their results are much better. Overall, I think this represents a valuable contribution to the field, providing convincing mechanistic evidence as to how these small heat shock proteins have evolved.

Author Response

The following is the authors' response to the original reviews.

Reviewer #1 (Recommendations For The Authors):

My main request is to show the phylogeny in the main text, so the reader knows what nodes are being compared.

Full phylogeny was added to the main text as Fig. 2. Additionally, phylogenetic tree in Newick format is presented as a Supplementary file 2.

I also suggest the authors check their figure legends carefully. At least in figure one, I think there is some mix-up with the letter labelling of the panels.

Our mistake. Figure legend was corrected. In this version of the manuscript Figure 1 was split into Fig. 1 and Fig. 3. Corrected version is presented in the legend to Fig. 3.

And lastly, I urge the authors to deposit the tree, alignment, and reconstructed sequences in a public repository.

Alignment in fasta format and phylogenetic tree in Newick format were added as supplementary files to the publication (supplementary file 1 and supplementary file 2, respectively). Reconstructed sequences (both Most likely and AltAll variants) were shown as a figure supplement (Figure 3 – figure supplement 2). Posterior probabilities for all positions of the reconstructed sequences were added as a supplementary file (supplementary file 3).

Reviewer #2 (Recommendations For The Authors):

-I find the term "secondarily single sHsp" to be a little confusing, especially because it is often used in relation to IbpA/B, but it is just IbpA in another species. I think it would be more clear for the reader to consistently refer to it as Erwiniaceae IbpA vs Escherichia IbpA, or something similar.

In the introduction we clarified (page 4 lines 11-13) that the term “secondarily single” IbpA refers to IbpA that lacks partner protein as a result of *ibpB* gene loss. This is in opposition to “single-protein” IbpA from a clade in which gene duplication leading to creation of two – protein sHsp system did not occur (like Vibrionaceae or Aeromonadaceae) - see Obuchowski et al., 2019.

-Figure 1B. The labels are not defined. What is L? A and B refer to IbpA and IbpB but this should be made more clear to the reader. Why is this panel only referred to in the Introduction and not the Results? Why is there a second panel for E.amy, rather than including it in the same panel, as for other experiments? What are the error bars? (That goes for every error bar in the paper, none are defined).

Labels in Fig.1B were corrected; “L” was used in reference to “luciferase alone” and it has been corrected for consistency to “no sHsp”. The sHsps activity measurements (obtained in the same experiment) were split into two separate panels as a correspondence to the two branches of the simplified tree in Fig. 1. The figure was modified to make it clearer and avoid confusion. Definitions of error bars were added to this and other figures.

-"AncA0 exhibited sequestrase activity on the level comparable to IbpA from Escherichia coli (IbpAE.coli). AncA1 was moderately efficient in this process and IbpA from Erwinia amylovora (IbpAE.amyl) was the least efficient sequestrase (Fig. 1D)." - First, this should be referring to Fig. 1C. Second, the text doesn't quite match the panel. A0 appears to have the strongest sequestrase activity over most concentrations. Can the authors comment on in what concentration range these differences are most meaningful?

Figure legend was corrected. Descriptions of panels C and D were fixed. Now these data are presented in panels A and B of a new Fig. 3. In our opinion differences in sequestration are most meaningful at lower sHsp concentrations (in this case lower than 5 μ M), as with high enough sHsp concentration even less effective sequestrases seem to be able to effectively sequester aggregated proteins. Comment about it was added to the main text (page 5, line 6)

-"Ancestral proteins' interaction with the aggregated substrates was stronger than in the case of extant E. amylovora IbpA, but weaker than in the case of extant E. coli IbpA (Fig. 1C)." - Is this referring to Fig. 1C, or to the unlabelled panel on the bottom right panel of Fig 1 (that is not referred to in the legend)? Can the authors comment on why they think the 2 ancestral proteins are much more similar to each other than they are to either of the native IbpAs?

Due to our mistake descriptions of panels C and D were switched.

Figure 1 was rearranged and split into Figures 1 and 3. Former figure S1 (full phylogeny) was inserted into the main text, as Fig. 2, per request of reviewer #1. Former panel 1D (now 3B) was rearranged, as graph was not apparent to be a part of that panel and looked as if it was unlabeled.

The fact that the two ancestral proteins are more similar to each other than to the extant E. coli and E. amylovora proteins in their interaction with model substrate might be caused by higher sequence identity between the two ancestral proteins than between ancestral and extant proteins (10 amino acid differences between AncA0 and AncA1 compared to 20 differences between AncA1 and IbpA from E. amylovora or 11 differences between AncA0 and IbpA from E. coli). One also has to remember that this property is only one aspect of sHsp activity – proteins AncA0 and AncA1 are much less similar to each other if other activities such as sequestrase activity are considered. Substrate affinity and sequestrase activity are connected to each other, but there isn't a strict correlation, as can be seen in the case of free ACD domains, which strongly bind aggregated substrate while effectively lacking sequestrase activity (fig. 5 A, fig. 5 – figure supplement 4 A,B).

-Figure 1E should have E. coli IbpA and IbpB, by themselves, included for comparison. Strangely, it seems, by comparison to Fig 1B, that the "inhibitory" activity of A0 is not present in the E. coli protein, and the authors should comment on this. Similarly, A1 disaggregation looks like it might not be significantly different than the E. coli protein. Can the authors comment on why disaggregation might be so low in A1 compared to E.amy?

E. coli IbpA alone was added to Fig. 1E (Fig. 3C in the new version) as suggested.

AncA1 indeed exhibits similar activity to extant IbpA from *E. coli*, which, at the conditions of the experiment, does not possess inhibitory effect observed for AncA0. This suggests that:

-There was an additional increase in ability to stimulate luciferase disaggregation between AncA1 and extant IbpA from *E. amylovora*

-There was also an increase of ability to stimulate luciferase refolding between AncA0 and extant *E. coli* IbpA, albeit to a significantly lesser degree than in the *Erwiniaceae* branch.

It is quite likely that after separation of *Erwiniaceae* and *Enterobacteriaceae* sHsp systems, they underwent further optimization through evolution. This might have led to observed higher effectiveness of modern IbpAs from both clades in refolding stimulation in comparison to the reconstructed ancestral proteins.

Despite the above, effects of substitutions on positions 66 and 109 on activities of the extant *E. coli* and *E. amylovora* proteins suggests that the two identified positions still play key role in differentiating extant IbpAs from *Erwiniaceae* and *Enterobacteriaceae*.

Nevertheless, additional mutations that lead to increased ability to stimulate luciferase reactivation must have occurred in both *Erwiniaceae* and *Enterobacteriaceae* branches of the phylogeny during evolution. These substitutions would be a worthwhile subject of further study.

| *-Fig 1D - lysate should be lysate.*

The typo was corrected.

| *-What is the bottom right panel in Fig 1? It doesn't seem to be referred to in the legend.*

This panel was intended to be the part of figure 1D, but it was not clearly visible. This figure was rearranged to make it clearer. Now these data are presented as Fig. 3B.

| *-Sequences are provided for the ancestral proteins, but I don't see them anywhere for the alternative ancestral proteins. How similar are the Anc proteins to the AltAlls? If they are very similar, this may not tell us anything about "robustness".*

Sequences of alternative proteins are added as a figure supplement (Fig. 3 - figure supplement 2). Full sequences of ML and alternative ancestors with posterior probabilities for each reconstructed position are presented in supplementary file 3

The testing of the robustness to statistical uncertainty was intended to test to what extent properties of reconstructed ancestral proteins could be influenced by uncertainty present in a given reconstruction due to probabilistic nature of the process. Relatively high similarity between ML and AltAll sequences would indicate low uncertainty of the reconstruction (most likely due to high conservation during evolution). In such a case similar properties of AltAll and ML proteins would simply indicate that they are robust to the level of uncertainty present in a given reconstruction (which may be low). It would not tell us much about "general" robustness to mutations, but it was not relevant to research questions considered.

| *-If the functional gain by IbpA comes down to only two amino acid substitutions, I'm not convinced this would be meaningfully reflected in any tests of positive selection.*

After considering Reviewer #1's comments about limitations of models used for selection analysis we added acknowledgment in the discussion (page 9, line 9 - 13) that results

indicating positive selection in our dataset should not be considered conclusive (see answer to Reviewer #1's public review below).

-The full MSA should be provided as supplemental material.

The full MSA in fasta format is presented in the supplementary file 1.

-For the aggregate binding panels in Figs 3 and 4, it would be helpful to show the native and ancestral proteins for comparison. I know this is a bit redundant, as they're present in Fig 1, but I find it hard to judge the scale of change. This is especially important because A0 and A1 are very similar in Fig 1, so I want to see what kind of difference the 2 mutations make.

Data presented in Fig. 3C (Fig. 5C in the new version) refer to the binding of α -crystallin domains (A0ACD and A0ACD Q66H G109D) and not full length sHsps to E. coli proteins aggregated on a BLI sensor. Our intention was to show the influence of the two crucial substitutions (Q66H G109D) on the properties of A0 ancestral α -crystallin domain.

Figure 4 (Fig. 6 in the new version) represent the effects of the substitutions on the identified positions 66 and 109 on the properties of extant IbpA orthologs from E. coli and E. amylovora, showing that these two positions play a key role in differentiating properties of those extant proteins. Changes in binding to aggregated substrate caused by those substitutions, as shown in Figure 6 B,C (new version), are indeed larger than observed between AncaA0 and AncaA1, as shown in Fig. 3B (new version).

One has to remember, however, that the experiment shown in Fig.3 (new version) shows the effects of all 10 amino acid changes between the nodes A0 and A1 and not only the two analyzed substitutions, as was the case in experiment shown in Fig. 6 B,C (new version). Moreover, due to relatively large number of differences between ancestral and extant sequences (11 differences between AncaA0 and E. coli IbpA, 20 differences between AncaA1 and E. amylovora IbpA), substitutions in the two experiments are introduced into different sequence context.

Because of the above, we believe that direct comparison of the results obtained for ancestral proteins with the results obtained for substitutions introduced into extant proteins would not meaningfully contribute to answering the question of the role of analyzed substitution in the context of extant proteins, while decreasing clarity of presented information.

-Some of the luciferase plots show a time course, but others just show a single %. What is the time point used for the single % plots?

Information was added to appropriate figure legends that for experiments showing a single timepoint the luciferase activity was measured after 1h of refolding.

Reviewer #3 (Recommendations For The Authors):

- 1. In the Introduction, it would be beneficial to explore additional instances where this evolutionary simplification process has been observed in nature. Investigating the prevalence of this phenomenon and identifying other multi-protein systems that have undergone simplification could enhance the understanding of its significance and implications.*

The section of the introduction concerning gene loss and differential paralog retention was expanded with additional examples of gene loss that is considered adaptive (page 3 lines 1 - 12).

1. I am intrigued by the reasons why certain organisms continue to maintain a two-protein system despite the viability of a single-protein system. This aspect is particularly relevant for bacteria, considering the fitness cost associated with maintaining extra gene copies. Do you have any hypotheses or theories that may shed light on this intriguing observation?

Refolding of proteins from aggregates requires the functional cooperation of sHsps and chaperones from Hsp70 system and Hsp100 disaggregase. In two protein sHsps system one sHsp (IbpA) is specialized in substrate binding, while the second one (IbpB) possesses low substrate binding potential and enhances sHsps dissociation from substrates (Obuchowski et al, 2019). Thus, the presence of IbpB reduces the amount of chaperones from Hsp70 system required to outcompete sHsps from aggregated substrates to initiate refolding process. The cost associated with maintaining extra sHsp gene copy (ibpB) in bacteria might be compensated by lower requirement for Hsp70 chaperones for efficient and fast protein refolding following stress conditions.

In this study we have demonstrated how such a system could have been simplified to a single – protein system capable of efficient substrate sequestration as well as stimulation of reactivation. This indeed leads to the question why such single – protein system isn't more prevalent in Enterobacterales.

One possibility may be that there are very specific requirements for efficient reactivation by a single – protein sHsp system. We have shown that new, more efficient IbpA functionality observed in Erwiniaceae required at least two separate mutations. It is possible, that such combinations of two substitutions simply did not occur in Enterobacteriaceae clade, in which IbpA still required partner protein for efficient reactivation stimulation.

One must also remember that experiments performed in this study were performed in vitro in a specific set of conditions, which most likely does not represent whole spectrum of challenges faced by different bacteria. It is possible that two – protein system has some other additional adaptive effects, counterbalancing the additional cost of gene maintenance. It was for example recently shown (Miwa & Taguchi, PNAS, 120 (32) e2304841120) that bacterial sHsps play an important role in regulation of stress response. Two – protein system could potentially allow for more complex regulation.

1. Incorporating X-ray crystallization as an additional technique in the methodology would offer detailed molecular insights into the effects of Q66H and G109D substitutions on ACD-C-terminal peptide and ACD-substrate interactions. The inclusion of such data would further strengthen the results section and provide robust support for your findings. Since the x-ray data might be difficult to collect, the authors might think to get alphafold model or some rosetta score for the model to discuss the finding further.

In response to reviewer comment we added the comparison of the structural models of Anca0 and Anca0 Q66H G109D ACD dimers complexed with the C-terminal peptides, representing middle structures of largest clusters obtained from equilibrium molecular dynamics simulation trajectories based on the AlphaFold2 prediction and in silico mutagenesis (Fig. 5 – figure supplement 2). Model comparison as well as C-terminal peptide – ACD contact analysis did not reveal any major changes in mode of peptide binding or α -crystallin domain conformation, although we do acknowledge that simulation timescale limits the conformational sampling.

Reviewer #1 (Public Review):

The work in this paper is in general done carefully. Reconstructions are done appropriately and the effects of statistical uncertainty are quantified properly. My only slight complaint is that I couldn't find statistics about posterior probabilities anywhere and that the sequences and trees do not seem to be deposited.

Posterior probabilities for all positions of reconstructed proteins were added as a supplementary file 3. MSA of all sequences used for ancestral reconstruction as well as phylogenetic tree in Newick format were added as supplementary files 1 and 2, respectively.

I would also have preferred to have the actual phylogeny in the main text. This is a crucial piece of data that the reader needs to see to understand what exactly is being reconstructed.

Full phylogeny was added to the main text as Fig. 2.

The paper identifies which mutations are crucial for the functional differences between the ancestors tested. This is done quite carefully - the authors even show that the same substitutions also work in extant proteins. My only slight concern was the authors' explanation of what these substitutions do. They show that these substitutions lower the affinity of the C-terminal peptide to the alpha-crystallin domain - a key oligomeric interaction. But the difference is very small - from 4.5 to 7 μ M. That seems so small that I find it a bit implausible that this effect alone explains the differences in hydrodynamic radius shown in Figure S8. From my visual inspection, it seems that there is also a noticeable change in the cooperativity of the binding interaction. The binding model the authors use is a fairly simple logarithmic curve that doesn't appear to consider the number of binding sites or potential cooperativity. I think this would have been nice to see here.

The binding model we used is equivalent to the Hill equation as it accounts for the variable slope of sigmoid function by inclusion of input scaling factor k , which is equivalent to the hill coefficient. Simple one site binding model and two site binding model were also considered but provided worse fits to the data than model including binding cooperativity. Not providing values of fitted parameter k was our mistake, and it was corrected (Fig. 5. with a legend). Additionally, output scaling parameter L is not necessary as fraction bound takes values from 0 to 1, therefore we have fitted the curves again without this parameter. The new values of fitted parameters are very similar to the previous ones. To make text more accessible to the reader, we have used a conventional form of Hill equation. Indeed, AncA0 Q66H G109D ACD displays higher binding cooperativity than more ancestral AncA0 ACD (hill coefficient 2.3 for AncA0 vs 3.7 for AncA0 Q66H G109D). Fitted values of Hill coefficients are higher than one can expect for 2-site ACD dimer, which is probably caused by an experimental setup of BLI, where C-terminal peptide is immobilized on the sensor and ACD is present in solution as bivalent analyte leading to emergence of avidity effects. Both cooperativity and avidity are reflected in the value of Hill coefficient, however as ligand density on the sensor is the same in all experiments only change in ACD binding cooperativity can account for observed difference in the value of Hill coefficients. Difference in the C-terminal peptide binding cooperativity may influence the process of sHsp oligomerization and assembly formation despite similar binding affinity, especially if avidity of multiple binding sites within oligomer is considered.

In addition, we changed the legend to Figure S8 (now called Fig. 5 – figure supplement 4A) to clarify the fact that the differences in average hydrodynamic radius are in fact ferly small. To highlight the observation that there are two populations of particles in AncA0 and AncA0

Q66H G109D measured at 25, 35 and 45 °C with different hydrodynamic diameters, we used % of intensity in DLS measurement. It allows us to show the change in the hydrodynamic diameter distribution that is relatively small. We recognize it was not properly explained in the article and added a clarification in figure description.

Lastly, the authors use likelihood methods to test for signatures of selection. This reviewer is not a fan of these methods, as they are easily misled by common biological processes (see PMID 37395787 for a recent critique). Perhaps these pitfalls could simply be acknowledged, as I don't think the selection analysis is very important to the impact of the work.

We thank the reviewer for pointing to the recent research about limitations of methods used in our work in selection analysis. As per recommendation we added acknowledgment of limitations of methods used to discussion (page 9, line 9 - 13), modifying wording of our conclusions to deemphasize significance of selection analysis results.

