

Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions

One sentence summary: *Small heat-shock proteins avoid dysfunctional co-assembly using mechanisms that cause minimal disruption to their conserved interfaces*

Georg K.A. Hochberg^{1,2,3}, Dale A. Shepherd^{1,2,4}, Erik G. Marklund^{1,2,5}, Indu Santhanagoplan⁶,
Matteo T. Degiacomi^{1,7}, Arthur Laganowksy^{1,8}, Timothy M. Allison¹, Eman Basha^{2,9},
Michael T. Marty^{1,10}, Martin R. Galpin¹, Weston B. Struwe¹, Andrew J. Baldwin¹,
Elizabeth Vierling² & Justin L.P. Benesch^{1,*}

¹ *Department of Chemistry, Physical & Theoretical Chemistry Laboratory, University of Oxford,
Oxford, OX1 3QZ, U.K.*

⁶ *Department of Biochemistry & Molecular Biology, University of Massachusetts,
Amherst, MA 01003, U.S.A.*

² These authors contributed equally.

* Correspondence to justin.benesch@chem.ox.ac.uk, +44 1865 285420

Present addresses: ³ Department of Ecology and Evolution, University of Chicago, Illinois 60637 U.S. ⁴ Waters Corporation, Stamford Ave., Wilmslow, SK9 4AX, U.K. ⁵ Department of Chemistry – BMC, Uppsala University, Box 576, 75123, Uppsala, Sweden. ⁷ Department of Chemistry, Durham University, South Rd, Durham, DH1 3LE, U.K. ⁸ Center for Infectious and Inflammatory Diseases, Institute of Biosciences and Technology, Texas A&M Health Science Center, Houston, Texas 77030, U.S.; Department of Chemistry, Texas A&M University, College Station, Texas 77842, U.S.; Department of Microbial Pathogenesis & Immunology, College of Medicine, Texas A&M Health Science Center, Bryan, Texas 77807, U.S. ⁹ Botany Department, College of Science, Tanta University, Tanta, Egypt. ¹⁰ Department of Chemistry and Biochemistry University of Arizona, 1306 E. University Blvd, Tucson, AZ 85721, U.S.

Abstract

Oligomeric proteins assemble with remarkable selectivity, even in the presence of closely related proteins, in order to perform their cellular roles. We show that most proteins related by gene duplication of an oligomeric ancestor have evolved to avoid hetero-oligomerization, and that this correlates with their acquisition of distinct functions. We report how co-assembly is avoided by two oligomeric small heat-shock protein paralogs. A hierarchy of assembly, involving intermediates that are populated only fleetingly at equilibrium, ensures selective oligomerisation. Conformational flexibility at non-interfacial regions in the monomers prevents co-assembly, allowing interfaces to remain largely conserved. Homomeric oligomers must overcome the entropic benefit of co-assembly and, accordingly, homomeric paralogs comprise fewer subunits than homomers that have no paralogs.

Main text

Many proteins associate into selective homo- or heteromers in order to function (1). New assemblies are most often created by gene duplication of a pre-existing homomer (2). The resulting oligomeric paralogs initially co-assemble because both have the same sequence (and hence structure and interfaces) as their ancestor (Fig. 1A) (3). This co-assembly can easily become entrenched if evolution of the two resulting duplicates is functionally constrained to maintain the interaction (4, 5), implying that heteromerisation should be the most likely fate of oligomeric paralogs. However, when we interrogated the human, *Arabidopsis*, yeast, and *E. coli* interactomes (Supplementary Materials, Data S1), we found that the majority of oligomeric paralogs in fact do not form heteromers (i.e do not co-assemble) (Fig. 1B), despite overlapping localization and expression profiles (Fig S1A,B). Moreover, we found that when paralogs cannot co-assemble, they share lower sequence identity and fewer common functions than paralogs that can (Fig. 1C,D). This suggests that heteromerisation acts as a constraint on the functional divergence of oligomeric paralogs (6). Relieving this constraint is therefore a key step in the evolutionary trajectories of oligomeric proteins towards evolving novel functions.

To interrogate how this occurs, we examined the selective assembly of two paralogous small heat-shock proteins (sHSPs), molecular chaperones found across the tree of life that are key to the cell's ability to respond to stress (7, 8). A duplication event led to land plants having two classes of cytosolic sHSPs (class-1 and -2, Fig. 1E, S2) that both assemble as dodecamers but cannot form heteromers between classes (9). Both are required for thermo-tolerance *in vivo* (10), and have different mechanisms of action (11, 12). We chose one paralog of each class from *Pisum sativum*, HSP18.1 and HSP17.7 (hereafter WT-1 and WT-2, respectively). Both proteins comprise an N-terminal region, an α -crystallin domain and a C-terminal tail, and both form homo-12-mers (12) using three independent interfaces: the α -crystallin domain mediates the formation of an isologous α \cdot α dimer, these dimers assemble into oligomers through heterologous contacts between the α -

crystallin domain and the C-terminal tails from neighbouring dimers ($\alpha\cdot C$), and interactions between the N-terminal regions (N \cdot N) (Fig. 1F) (13). Their complex, multi-interface architecture makes these proteins an ideal system to investigate how evolution acts to regulate the biophysical properties of oligomers to develop a set of selective interfaces that allows them to diverge functionally.

Small-angle X-ray scattering experiments indicated that both proteins form tetrahedral oligomers (Fig. S3), implying that there are no major differences in quaternary structure that prevent co-assembly. Nonetheless, when we obtained native mass spectra of a mixture of WT-1 and WT-2 after prolonged incubation (Fig. 1G, upper) or initiating re-assembly from their subunits (Fig. S4A), we failed to detect any hetero-12-mers in either case. However, both homo-12-mers underwent continual dissociation and re-association, though WT-1 did so >10 times faster than WT-2 (Fig S4). These facile quaternary dynamics show that heteromers are in principle kinetically accessible and so, despite the similarity in quaternary architectures of WT-1 and WT-2, must be thermodynamically unfavourable.

To identify the sequence-determinants of selective assembly, we aligned class-1 and -2 sHSPs, and noted conserved differences in their C-terminal tails (Fig. S5). We then engineered a chimera with the class-1 N-terminal region and α -crystallin domain linked to the class-2 C-terminal tail (N¹ α 1^C2, see Table S1) and incubated it with WT-2. This small change in sequence produced a series of hetero-12-mers formed between WT-2 and N¹ α 1^C2 (Fig. 1G, lower). These represent a proxy for class-1 and -2 co-assembly, and allowed us to interrogate the functional consequences of heteromerisation. We incubated purified sHSPs with pea leaf lysate under heat-shock conditions to form reversible aggregates (14), mimicking their action *in vivo* (10, 11). WT-2 partitioned significantly faster into the insoluble fraction than WT-1 (Fig. 1H, S6). The rate measured for the heteromers of N¹ α 1^C2 and WT-2, however, was intermediate to WT-1 and WT-2 homomers. The functional differentiation of the two proteins therefore depends on their selective homomerisation, demonstrating the operational necessity of avoiding co-assembly.

The hetero-12-mers formed by swapping C-terminal tails comprised only even numbers of each type of subunit (Fig. 1G, lower), implying that either the $\alpha\alpha$ or the N•N interface must also be selective. To determine which, we engineered an N-terminal chimera, $N2\alpha1^C1$, and incubated it with WT-1. This produced a series of hetero-12-mers comprising odd and even numbers of each subunit (Fig. S7A). While N•N contacts therefore are not thermodynamically selective (and hence the $\alpha\alpha$ interface must be), we noticed that dissociation of $N1\alpha1^C2$ oligomers was as fast as WT-1 (Fig. S7B), whereas dissociation of $N2\alpha1^C1$, was slow (Fig. S7A). This means that the promiscuous N•N contacts, not the thermodynamically selective $\alpha\alpha$ and $\alpha\alpha$ interfaces, control the kinetic stability of the 12-mers.

Our subunit-exchange data indicate that, over the functional temperature range, hetero-12-mers formed via N•N contacts during assembly would decompose into homomers on the timescale of minutes to hours (Fig. S4E). Yet, we had observed no long-lived heteromers in our assembly experiment, even at low temperatures (Fig S4A). To resolve this apparent conflict, we generated constructs of WT-1 and WT-2 lacking the N-terminal region and measured their stoichiometries using native ion mobility mass spectrometry (IM-MS). Both were polydisperse, spanning dimers to 12-mers (Fig. 2A, S8A). Constructs instead lacking the C-terminus only formed monomers and dimers (Fig. 2B, S8B). $\alpha\alpha$ contacts therefore likely form early and ensure rapid self-selective oligomerisation, while N•N contacts subsequently stabilize the 12-meric fraction (Fig. S8C, see Supplemental Text). This hierarchy obviates the need for kinetically stable N•N contacts to be selective, and avoids long-lived heteromers that would compromise the rapid stress response of sHSPs in the cell.

To understand the thermodynamic basis of selectivity at the $\alpha\alpha$ interface, we examined chimeric versions of the N-terminal truncations. $\alpha1^C2$ formed polydisperse oligomers, but $\alpha2^C1$ did not assemble beyond a dimer (Fig. 2C, S8D-F). Selectivity in the $\alpha\alpha$ interface is therefore directional, arising from an unfavourable association between the WT-1 C-terminal tail and WT-2. We quantified this effect directly by excising the core domains of both proteins ($\alpha1$ and $\alpha2$, Table S1) and measuring

their affinity for each other's C-terminal tails. Whereas $\alpha 1$ bound peptides mimicking each tail equally well, $\alpha 2$ had a much lower affinity for a WT-1 than WT-2 peptide ($\Delta\Delta G > 6 \text{ kJmol}^{-1}$, Fig. S9).

We next turned our attention to the $\alpha\alpha$ interface, which is selective (Fig. S10A) despite high sequence conservation (Fig. S5B). Crystal structures revealed $\alpha 1$ and $\alpha 2$ to be extremely alike (Fig. 3A,B, Table S2). The dimer interface is formed in both homodimers by salt bridges centred on the $\beta 8$ - $\beta 9$ loop (L8/9) that are fully conserved between the two proteins; and by reciprocal strand-exchange between $\beta 6$ and $\beta 2$. The latter involves only one obvious class-specific contact: between the π -systems of a histidine on $\beta 6$ and a tryptophan on $\beta 2$ in WT-1 that is absent in WT-2 (Fig. 3C,D). In 2- μs molecular dynamics (MD) simulations, both homodimers and a modelled heterodimer were stable. The interfaces of the heterodimer featured equivalent overall numbers of interacting side-chains, hydrogen bonds, and level of structural flexibility compared to both homodimers (Fig. S10B-E, S11). Remarkably, the α -crystallin domain is therefore selective with only minimal differences in the number or type of contacts at its interface.

To investigate the origin of this selectivity, we performed calorimetric measurements and found that there are differences in the relative contributions from entropy and enthalpy to the favourable free energy of dimerization in $\alpha 1$ and $\alpha 2$ (Fig. S12A-C). This suggests subtle differences in their association mechanisms that may impart selectivity. To quantify which parts of the dimer are responsible for selectivity, we divided the core domain into three segments (Fig. 3E, Table S1): the β -sandwich (S), which includes the L8/9 interface and $\beta 2$ from the $\beta 6\beta 2$ interface; $\beta 6$ (B); and the loop (L) connecting $\beta 6$ to the β -sandwich. We shuffled these segments between $\alpha 1$ and $\alpha 2$ (Fig. 3E) and, for the 36 pairwise combinations of chimeric and wild-type constructs, determined the corresponding free energy of dimerization, $\Delta G_{\alpha,\alpha}$ by performing quantitative IM-MS titration experiments (Fig. S12D-G). From the overall dataset, we identified statistically significant intermolecular interactions between $\beta 6$ and the β -sandwich (B·S), and the loop and the β -sandwich (L·S). Summed (B+L·S, Fig 3F), these interactions contribute $\approx 11 \text{ kJmol}^{-1}$ to the stability of the dimer, except when $^S 2$ encounters $^{B1}L1$,

which unilaterally destabilizes the dimer by ≈ 7 kJmol⁻¹ (Fig. 3F, left). The L·S and B·S components contribute nearly equally to dimer stability (Fig. 3F, middle and right), a surprising observation considering that the loop is not part of the interface.

Because the $\alpha 1$ and $\alpha 2$ dimer structures did not reveal differences that account for our experimental thermodynamic data, we performed steered MD simulations in which we gradually detached $\beta 6$ from $\beta 2$, and estimated the resulting free-energy profile (Fig. 3G). As predicted by our thermodynamic data, we found that the heteromeric $B^{+L1}\cdot S2$ interface was significantly easier to break than the other combinations. We also noticed that in unconstrained simulations of the $\alpha 1$ monomer (performed in triplicate) the β -sandwich remained rigid (Fig. 3H, S13A,C,D), while the loop distorted and formed intra-molecular contacts (Fig S13D). In the $\alpha 2$ monomer, the loop more closely retained its conformation from the dimer (Fig. 3I, S13B-D), but $\beta 2$ detached from the β -sandwich and became highly flexible (Fig S13C,D,E).

Our data imply that the loop in $\alpha 1$, and $\beta 2$ in $\alpha 2$, have a propensity to sample conformations in the monomers that are limited upon formation of a dimer interface (Fig. S13D). In both homodimers only one side of each B+L·S interface is restrained in this way, while in the heterodimer both sides of the $B^{+L1}\cdot S2$ interface are (Fig. 4), making it easier to break apart. Conversely, to dimerize, dynamic regions must undergo a structural transition from their monomeric conformations. In homodimers, only one side of each interface would have to do this, with the other being pre-ordered for dimerization. In a heterodimer, this conformational complementarity would be absent for the $B^{+L1}\cdot S2$ interface, also leading to a slow association rate. These effects would therefore combine to discourage the formation of heterodimers and instead ensure self-selection.

If this mechanism is correct, with the loop making a large contribution to the instability of the heterodimer (Fig. 3E), it should be a major regulator of the monomeric structure. Indeed, the conformations of simulated chimeric monomers lie between the extremes occupied by $\alpha 1$ and $\alpha 2$, and the segment that shifts the structure the most is the loop, not the interfacial segments (Fig. S13F).

Similarly, chimeric dimers incorporating segments that do not change conformations in our simulations (β^1 , β^2 , and β^3 , Fig. 3E) should be more stable than both α^1 and α^2 . This prediction is borne out in their experimental melting temperature being ≈ 5 °C higher (Fig. S13G).

We mined our MD trajectories for specific contacts that were more abundant in one class over the other, and identified 11 and 3 that involved residues that displayed class-specific evolutionary conservation in α^1 and α^2 , respectively. Strikingly, we found that the majority of these are outside of the dimer interface: in α^1 7 out of 11 conserved sites either attach β^2 to the sandwich or promote curling of the loop, while in α^2 one maintains an extended loop conformation (Fig. S14), and another makes β^2 prone to detach in the monomer. Thus non-interfacial regions, and their effects on the structure of dissociated monomers, determine selectivity in the α -crystallin domain of class-1 and -2 sHSPs across land plants. This is consistent with the observation that non-interfacial residues can affect interface stabilities (15).

To homomerize, paralogs must overcome a substantial entropic benefit of co-assembly arising from the number of ways distinguishable subunits can be arranged. This mixing entropy increases with the number of subunits in the oligomer such that the energetic cost of homomerization rises logarithmically (Fig. 4A, Supplementary Text). Combining this contribution with the strength of interactions we quantified experimentally, allowed us to generate a model predicting the stability of all possible combinations of the two sHSPs and their chimeras, dependent only on their stoichiometry and constituent α -C and α - α interfaces (Supplementary Text, Fig. S15). We used this model to calculate the difference in stability between every possible heteromer and the corresponding homomers along the assembly pathway (Fig. 4B). The selective interactions in the α -C and α - α interfaces narrowly overcome the entropic benefit of co-assembly for all stoichiometries (Fig. 4C), resulting in a predicted population of hetero-12-mers at equilibrium that is just below detectable levels (Fig. 4C, right).

Homomers are therefore only marginally more stable than heteromers, even though the paralogs have diverged for >400 million years (16). The number and type of selective interactions we found is the minimum required for a tetrahedron (17), with half of the oligomeric interfaces (N·N and those involving C2) remaining promiscuous. These observations imply that selectivity is difficult to evolve, perhaps because most substitutions that disfavour co-assembly, also disfavour self-assembly (18).

Our model predicts that this would be more problematic for oligomers with more subunits, for which the entropic barrier to self-assembly is higher (Fig. 4A). Using a dataset of oligomeric architectures based on curated crystal structures (17) and combining it with our list of paralogs (Fig. 1B, Data S2), we found that self-selective paralogs comprise fewer subunits than homomers that have no paralogs (Fig. 4D). The data are well explained by the probability that selectivity evolves after duplication being inversely proportional to the mixing entropy (Supplemental Text). Applying this relationship to scale the stoichiometry distribution of oligomers without paralogs renders it indistinguishable from the self-selective set (Fig. 4D). This indicates that this fundamental thermodynamic bias acts as a significant evolutionary constraint across oligomeric proteins. The mechanisms for selectivity we have uncovered for the sHSPs studied here are some, of possibly many, ways in which proteins have evolved to escape co-assembly.

References and notes

1. J. A. Marsh, S. A. Teichmann, Structure, dynamics, assembly, and evolution of protein complexes. *Annu Rev Biochem*, (2014).
2. J. B. Pereira-Leal, E. D. Levy, C. Kamp, S. A. Teichmann, Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol* **8**, R51 (2007).
3. E. Kaltenecker, D. Ober, Paralogous interference affects the dynamics after gene duplication. *Trends Plant Sci* **20**, 814-821 (2015).
4. G. Diss *et al.*, Gene duplication can impart fragility, not robustness, in the yeast protein interaction network. *Science* **355**, 630-634 (2017).
5. G. C. Finnigan, V. Hanson-Smith, T. H. Stevens, J. W. Thornton, Evolution of increased complexity in a molecular machine. *Nature* **481**, 360-364 (2012).
6. C. R. Baker, V. Hanson-Smith, A. D. Johnson, Following gene duplication, paralogous interference constrains transcriptional circuit evolution. *Science* **342**, 104-108 (2013).
7. W. E. Balch, R. I. Morimoto, A. Dillin, J. W. Kelly, Adapting proteostasis for disease intervention. *Science* **319**, 916-919 (2008).
8. K. Richter, M. Haslbeck, J. Buchner, The heat shock response: life on the verge of death. *Mol Cell* **40**, 253-266 (2010).
9. K. W. Helm, G. J. Lee, E. Vierling, Expression and native structure of cytosolic class II small heat-shock proteins. *Plant Physiol* **114**, 1477-1485 (1997).
10. F. McLoughlin *et al.*, Class I and II small heat-shock proteins protect protein translation factors during heat stress. *Plant Physiol*, (2016).

11. M. Kirschner, S. Winkelhaus, J. M. Thierfelder, L. Nover, Transient expression and heat-stress-induced co-aggregation of endogenous and heterologous small heat-stress proteins in tobacco protoplasts. *Plant J* **24**, 397-411 (2000).
12. E. Basha, C. Jones, V. Wysocki, E. Vierling, Mechanistic differences between two conserved classes of small heat shock proteins found in the plant cytosol. *J Biol Chem* **285**, 11489-11497 (2010).
13. G. R. Hilton, H. Lioe, F. Stengel, A. J. Baldwin, J. L. Benesch, Small heat-shock proteins: paramedics of the cell. *Top Curr Chem* **328**, 69-98 (2013).
14. E. W. Wallace *et al.*, Reversible, specific, active aggregates of endogenous proteins assemble upon heat stress. *Cell* **162**, 1286-1298 (2015).
15. T. Perica *et al.*, Evolution of oligomeric state through allosteric pathways that mimic ligand binding. *Science* **346**, 1254346 (2014).
16. E. R. Waters, E. Vierling, The diversification of plant cytosolic small heat shock proteins preceded the divergence of mosses. *Mol Biol Evol* **16**, 127-139 (1999).
17. S. E. Ahnert, J. A. Marsh, H. Hernandez, C. V. Robinson, S. A. Teichmann, Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aaa2245 (2015).
18. C. D. Aakre *et al.*, Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594-606 (2015).
19. A. Chatr-Aryamontri *et al.*, The BioGRID interaction database: 2017 update. *Nucleic Acids Res* **45**, D369-D379 (2017).
20. S. V. Rajagopala *et al.*, The binary protein-protein interaction landscape of Escherichia coli. *Nat Biotechnol* **32**, 285-290 (2014).
21. C. The UniProt, UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169 (2017).

22. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
23. C. The Gene Ontology, Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* **45**, D331-D338 (2017).
24. M. Uhlen *et al.*, Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
25. J. Liu *et al.*, Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell* **24**, 4333-4345 (2012).
26. S. Ghaemmaghami *et al.*, Global analysis of protein expression in yeast. *Nature* **425**, 737-741 (2003).
27. G. Orfanoudaki, A. Economou, Proteome-wide subcellular topologies of E. coli polypeptides database (STEPdb). *Mol Cell Proteomics* **13**, 3674-3687 (2014).
28. B. T. Korber *et al.*, Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. *J Virol* **68**, 7467-7481 (1994).
29. A. Laganowsky *et al.*, Crystal structures of truncated alphaA and alphaB crystallins reveal structural mechanisms of polydispersity important for eye lens function. *Protein Sci* **19**, 1031-1043 (2010).
30. F. D. Kondrat, W. B. Struwe, J. L. Benesch, Native mass spectrometry: towards high-throughput structural proteomics. *Methods Mol Biol* **1261**, 349-371 (2015).
31. F. Sobott, H. Hernandez, M. G. McCammon, M. A. Tito, C. V. Robinson, A tandem mass spectrometer for improved transmission and analysis of large macromolecular assemblies. *Anal Chem* **74**, 1402-1407 (2002).

32. M. F. Bush *et al.*, Collision cross sections of proteins and their complexes: a calibration framework and database for gas-phase structural biology. *Anal Chem* **82**, 9557-9565 (2010).
33. J. T. Hopper, K. Sokratous, N. J. Oldham, Charge state and adduct reduction in electrospray ionization-mass spectrometry using solvent vapor exposure. *Anal Biochem* **421**, 788-790 (2012).
34. G. R. Hilton *et al.*, C-terminal interactions mediate the quaternary dynamics of alphaB-crystallin. *Philos Trans R Soc Lond B Biol Sci* **368**, 20110405 (2013).
35. M. V. Petoukhov *et al.*, New developments in the program package for small-angle scattering data analysis. *J Appl Crystallogr* **45**, 342-350 (2012).
36. A. J. Baldwin *et al.*, The polydispersity of alphaB-crystallin is rationalized by an interconverting polyhedral architecture. *Structure* **19**, 1855-1863 (2011).
37. M. T. Degiacomi, M. Dal Peraro, Macromolecular symmetric assembly prediction using swarm intelligence dynamic modeling. *Structure* **21**, 1097-1106 (2013).
38. D. W. A. Buchan, F. Minneci, T. C. O. Nugent, K. Bryson, D. T. Jones, Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res* **41**, W349-W357 (2013).
39. D. Svergun, C. Barberato, M. H. J. Koch, CRY SOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Crystallogr* **28**, 768-773 (1995).
40. L. Breiman, Random forests. *Mach Learn* **45**, 5-32 (2001).
41. M. T. Marty *et al.*, Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Anal Chem* **87**, 4370-4376 (2015).

42. A. J. Baldwin, H. Lioe, C. V. Robinson, L. E. Kay, J. L. Benesch, alphaB-crystallin polydispersity is a consequence of unbiased quaternary dynamics. *J Mol Biol* **413**, 297-309 (2011).
43. P. D. Adams *et al.*, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221 (2010).
44. M. J. Abraham *et al.*, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19 - 25 (2015).
45. K. Lindorff-Larsen *et al.*, Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950-1958 (2010).
46. H. W. Horn *et al.*, Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J Chem Phys* **120**, 9665-9678 (2004).
47. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J Chem Phys* **126**, (2007).
48. H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola, J. R. Haak, Molecular-Dynamics with Coupling to an External Bath. *J Chem Phys* **81**, 3684-3690 (1984).
49. M. Parrinello, A. Rahman, Polymorphic transitions in single-crystals - a new molecular-dynamics method. *J Appl Phys* **52**, 7182-7190 (1981).
50. B. Hess, P-LINCS: A parallel linear constraint solver for molecular simulation. *J Chem Theory Comput* **4**, 116-122 (2008).
51. B. Hess, H. Bekker, H. J. C. Berendsen, J. G. E. M. Fraaije, LINCS: A linear constraint solver for molecular simulations. *J Comput Chem* **18**, 1463-1472 (1997).

52. J. P. Ryckaert, G. Ciccotti, H. J. C. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* **23**, 327-341 (1977).
53. K. A. Feenstra, B. Hess, H. J. C. Berendsen, Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J Comput Chem* **20**, 786-798 (1999).
54. R. W. Hockney, J. W. Eastwood, *Comput simul using part* (A. Hilger, Bristol England ; Philadelphia, ed. Special student, 1988), pp. xxi, 540 p.
55. U. Essmann *et al.*, A smooth particle mesh Ewald method *J Chem Phys* **103**, 8577-8592 (1995).
56. G. M. Torrie, J. P. Valleau, Non-physical sampling distributions in Monte-Carlo free-energy estimation: Umbrella sampling. *J Comput Phys* **23**, 187-199 (1977).
57. J. S. Hub, B. L. de Groot, D. van der Spoel, g_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J Chem Theory Comput* **6**, 3713-3720 (2010).
58. S. Pronk *et al.*, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845-854 (2013).
59. X. Daura *et al.*, Peptide Folding: When Simulation Meets Experiment. *Angewandte Chemie Int Ed* **38**, 236-240 (1999).
60. A. Gelman, D. B. Rubin, Inference from Iterative Simulation Using Multiple Sequences. *Statist Sci* **7**, 457-472 (1992).
61. V. Gabelica, N. Galic, F. Rosu, C. Houssier, E. De Pauw, Influence of response factors on determining equilibrium association constants of non-covalent complexes by electrospray ionization mass spectrometry. *J Mass Spectrom* **38**, 491-501 (2003).

62. T. N. Starr, J. W. Thornton, Epistasis in protein evolution. *Protein Sci* **25**, 1204-1218 (2016).
63. H. McWilliam *et al.*, Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res* **41**, W597-600 (2013).
64. F. Abascal, R. Zardoya, D. Posada, ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104-2105 (2005).
65. S. Guindon *et al.*, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321 (2010).
66. D. T. Jones, W. R. Taylor, J. M. Thornton, The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275-282 (1992).
67. F. Sobott, J. L. Benesch, E. Vierling, C. V. Robinson, Subunit exchange of multimeric protein complexes. Real-time monitoring of subunit exchange between small heat shock proteins by using electrospray mass spectrometry. *J Biol Chem* **277**, 38921-38929 (2002).
68. H. Ke *et al.*, Genetic investigation of tricarboxylic acid metabolism during the *Plasmodium falciparum* life cycle. *Cell Rep* **11**, 164-174 (2015).
69. E. D. Levy, E. Boeri Erba, C. V. Robinson, S. A. Teichmann, Assembly reflects evolution of protein complexes. *Nature* **453**, 1262-1265 (2008).
70. J. A. Marsh *et al.*, Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell* **153**, 461-470 (2013).
71. H. Akaike, A new look at the statistical model identification. *IEEE Trans Autom Control* **19**, 716-723 (1974).
72. M. Kimura, Evolutionary Rate at the Molecular Level. *Nature* **217**, 624-626 (1968).

73. N. Tokuriki, F. Stricher, J. Schymkowitz, L. Serrano, D. S. Tawfik, The stability effects of protein mutations appear to be universally distributed. *J Mol Biol* **369**, 1318-1332 (2007).
74. M. Lynch, Evolutionary diversification of the multimeric states of proteins. *Proc Natl Acad Sci U S A* **110**, E2821-2828 (2013).
75. R. L. van Montfort, E. Basha, K. L. Friedrich, C. Slingsby, E. Vierling, Crystal structure and assembly of a eukaryotic small heat shock protein. *Nat Struct Biol* **8**, 1025-1030 (2001).
76. K. K. Kim, R. Kim, S. H. Kim, Crystal structure of a small heat-shock protein. *Nature* **394**, 595-599 (1998).
77. W. Kabsch, Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J Appl Crystallogr* **26**, 795-800 (1993).

Acknowledgments

We thank Carol Robinson, Jason Schnell, Philipp Kukura, David Staunton (all University of Oxford) and Brian Metzger (University of Chicago) for helpful discussions. We acknowledge access to B21 and help from Mark Tully and James Douth at the Diamond Synchrotron (JLPB for SM9384-2); and the ARCUS cluster at Advanced Research Computing, Oxford. We thank the following funding sources: Engineering & Physical Sciences Research Council (GKAH for a studentship, JLPB for EP/J01835X/1); Carl Trygger's Foundation (EM); Swiss National Science Foundation (MTD for P2ELP3_155339) Biotechnology & Biological Sciences Research Council (AJB for BB/J014346/1, JLPB for BB/K004247/1 and BB/J018082/1); National Institutes of Health (EV for RO1 GM42761); Massachusetts Life Sciences Center (EV for a New Faculty Research Award); Royal Society (JLPB for a

University Research Fellowship). All data to support the conclusions is available in the manuscript or the Supplementary Materials, and is deposited with DOI 10.5287/bodleian:54jBVeAzw.

Supplementary Materials

Materials and Methods; Supplementary Text; Figs. S1 to S15; Tables S1 to S2; References (19-77);
Data S1 to S2.

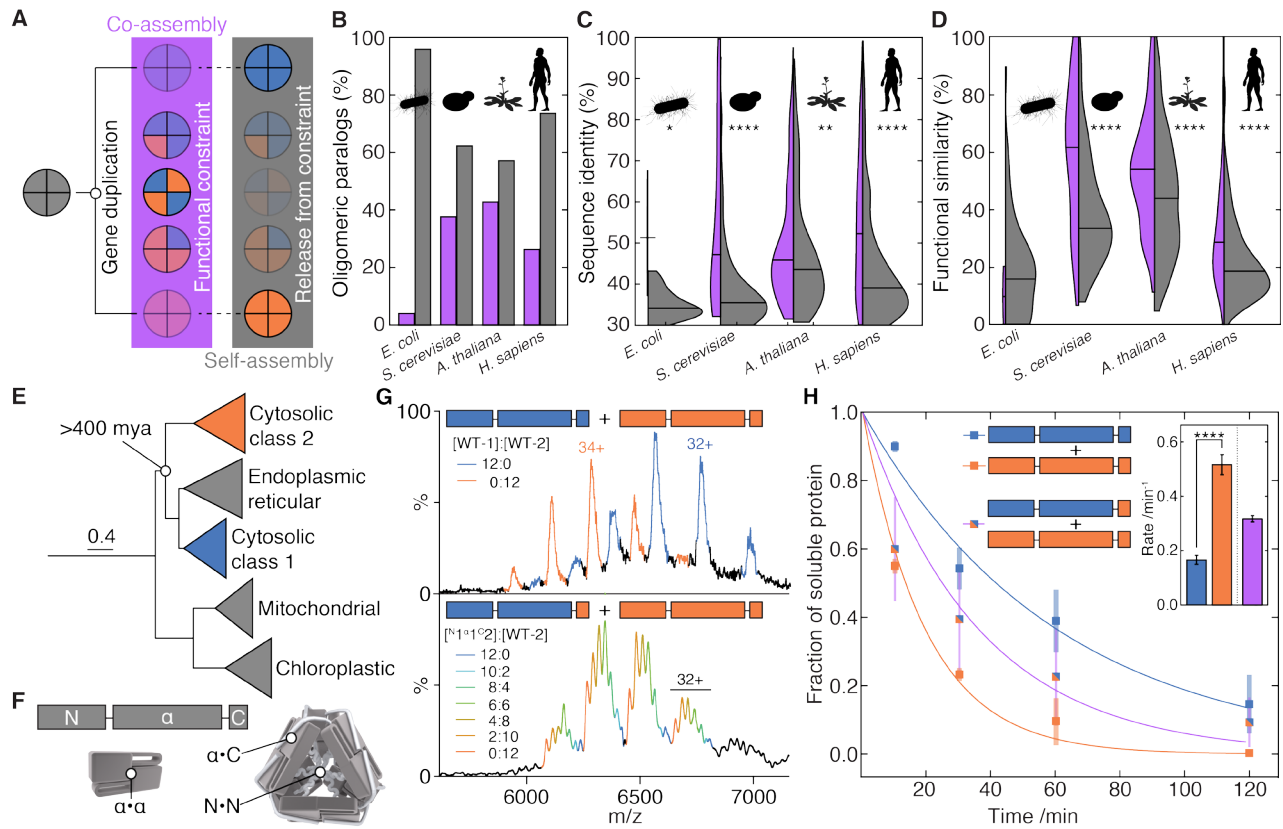


Figure 1. Self-selective assembly allows oligomeric paralogs to evolve distinct functions.

A) After gene duplication, oligomeric paralogs co-assemble into and predominantly populate heteromers, constraining their functions to be compatible with co-assembly. If they subsequently evolve the ability to assemble self-selectively into homomers, their functions are free to diverge.

B) Percentage of pairs of oligomeric paralogs that either co-assemble into heteromers (purple) or only self-assemble into homomers (grey) in *E. coli* (73 pairs in dataset), *Saccharomyces cerevisiae* (215 pairs), *Arabidopsis thaliana* (742 pairs), and *Homo sapiens* (1086 pairs).

C) Pairwise sequence identity is higher between co-assembling paralogs (purple) than between self-assembling paralogs (grey). Horizontal lines denote medians. * $p < 0.05$, ** $p < 0.01$, **** $p < 0.0005$, Mann-Whitney rank sums test.

D) Pairwise functional similarity of co-assembling (purple) and self-assembling (grey) pairs of paralogs as measured by the intersection over the union of their gene ontology annotations. Horizontal lines denote medians. **** $p < 0.0005$, Mann-Whitney rank sums test.

E) Maximum-likelihood phylogeny of select clades of plant sHSPs. Scale bar indicates average number of substitutions per site.

F) Schematic of the three different interfaces used by sHSP to assemble into oligomers.

G) Mass spectrum of WT-1 and WT-2 after prolonged incubation plotted in the mass-to-charge (m/z) dimension. WT-1 (blue) and WT-2 (orange) 12-mers are observed, with varying numbers of charges. No peaks corresponding to heteromers are detected (**upper**). Hetero-12-mers are formed via exchange of dimers if WT-2 is mixed with $N1^{\alpha}1^{\zeta}2$, resulting in additional peaks for each charge state (**lower**). One charge-state is labelled for each 12-mer.

H) When mixed prior to incubation with pea-leaf lysate at 42 °C, WT-1 and WT-2 partition into aggregates at different rates (**** $p < 0.0005$). When WT-2 is incubated with $N1^{\alpha}1^{\zeta}2$, subunits from both proteins partition at the same, intermediate rate (**inset**). Heteromers thus function differentially to segregated WT oligomers. Error bars in the raw data are standard deviations from three independent experiments; error bars in the inset are standard deviations calculated from 1000 bootstrap replicates of the fit.

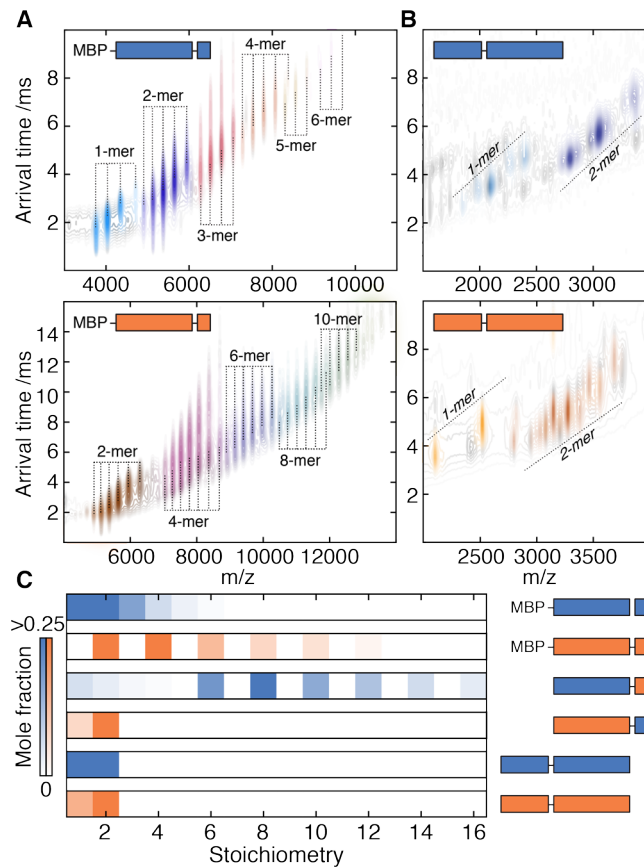


Figure 2. *Oligomeric interfaces form in a hierarchical order.*

A) IM-MS spectra of truncated constructs of WT-1 (**upper**) and WT-2 (**lower**) lacking the N-terminal region. The two dimensions of separation (m/z and arrival time, which depends on collision cross-section) separate charge-state series corresponding to a series of stoichiometries (coloured individually). Both truncated proteins assemble into polydisperse ensembles. MBP – maltose binding protein.

B) IM-MS spectra of truncated constructs of WT-1 (**upper**) and WT-2 (**lower**) lacking the C-terminal tail. Both proteins do not assemble beyond dimers. Truncations on the exposed N-terminus result in several charge-series for monomers and dimers that are separated in the arrival time dimension (see Fig. S8 for detailed assignments).

C) Distribution of stoichiometries populated by truncated constructs extracted from spectra in **A**, **B**, **Fig S6**. The C-terminal tail is required for assembly beyond dimers, whereas the N-terminus is

required for monodisperse 12-mers. The $\alpha^2\text{C}1$ construct (Fig. S8E) does not oligomerize, indicating an unfavourable $\alpha\text{-C}$ interaction.

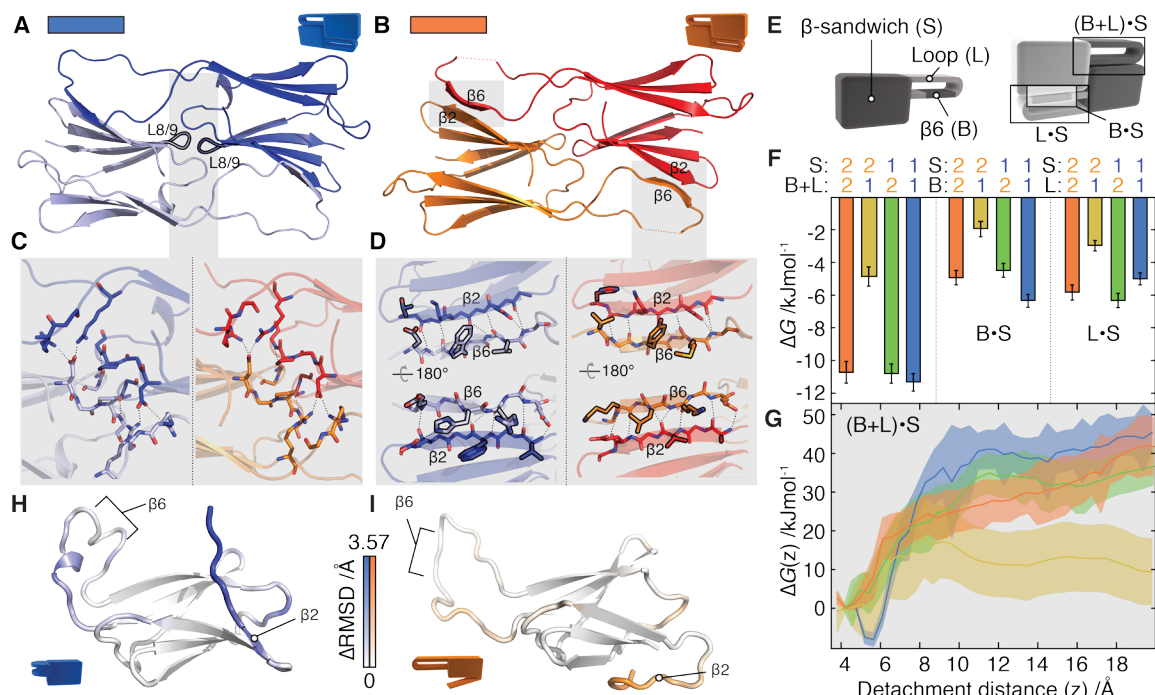


Figure 3. Selectivity in the structurally conserved α -crystallin domain.

A and **B**) $\alpha 1$ and $\alpha 2$ dimers have an identical fold (backbone RMSD = 1.2 \AA) in which two highly similar interfaces (labelled L8/9 and $\beta 6\cdot\beta 2$) connect monomers.

C) The L8/9 interface is centred on the loop between $\beta 8$ and $\beta 9$ (black outline) and is indistinguishable in the two proteins. Inter-chain hydrogen bonds are shown as dashed lines.

D) The two $\beta 6\cdot\beta 2$ interfaces in the dimer are formed by exchange between the $\beta 6$ and $\beta 2$ strands. Side-chains that differ between $\alpha 1$ and $\alpha 2$ at homologous positions are outlined in black. The π -stacking interaction specific to $\alpha 1$ is shown as a dotted red line.

E) Constructs were designed by swapping the β -sandwich, loop, and $\beta 6$ strand (**left**). These were used to assess the strength of the $\beta 6\cdot\beta 2$ interface, and deconvolve the contribution from the loop and $\beta 6$ strand (**right**).

F) Global thermodynamic model of dimerization based on experimentally determined $\Delta G_{\alpha,\alpha}$ values in Fig S12G. The combined loop and $\beta 6$ from $\alpha 1$ interact less favourably with $\beta 2$ from $\alpha 2$ than all other

combinations (**left**). $\alpha 2$ and $\alpha 1$ partition contributions to $\Delta G_{\alpha,\alpha}$ differently (**shaded**). Error bars are standard deviations from 1000 bootstrap replicates of the model fit.

G) In a simulated heterodimer, the free energy barrier is significantly reduced for the $\alpha 2 \cdot \alpha 1$ pair (yellow), but indistinguishable from the homodimers in the case of $\alpha 1 \cdot \alpha 2$ (green) when the $\beta 6 \cdot \beta 2$ interface is disrupted along a reaction coordinate that separates them. Shaded area corresponds to the standard error of the mean.

H,I) Median monomeric conformations determined by principal component analysis coloured according to structural difference. This is calculated at each residue from the C α RMSD between $\alpha 1$ and $\alpha 2$ monomers, minus the RMSD between repeats for each monomer. Positive Δ RMSD values indicate conformational differences between proteins that cannot be explained by the variations intrinsic to each protein, and only those with $p < 0.05$ (after Bonferroni correction, permutation test) are coloured. Differences are apparent in the loop surrounding $\beta 6$ and in $\beta 2$. In $\alpha 1$ the loop curls up, whereas in $\alpha 2$ the $\beta 2$ strand detaches readily from the remainder of the β -sandwich.

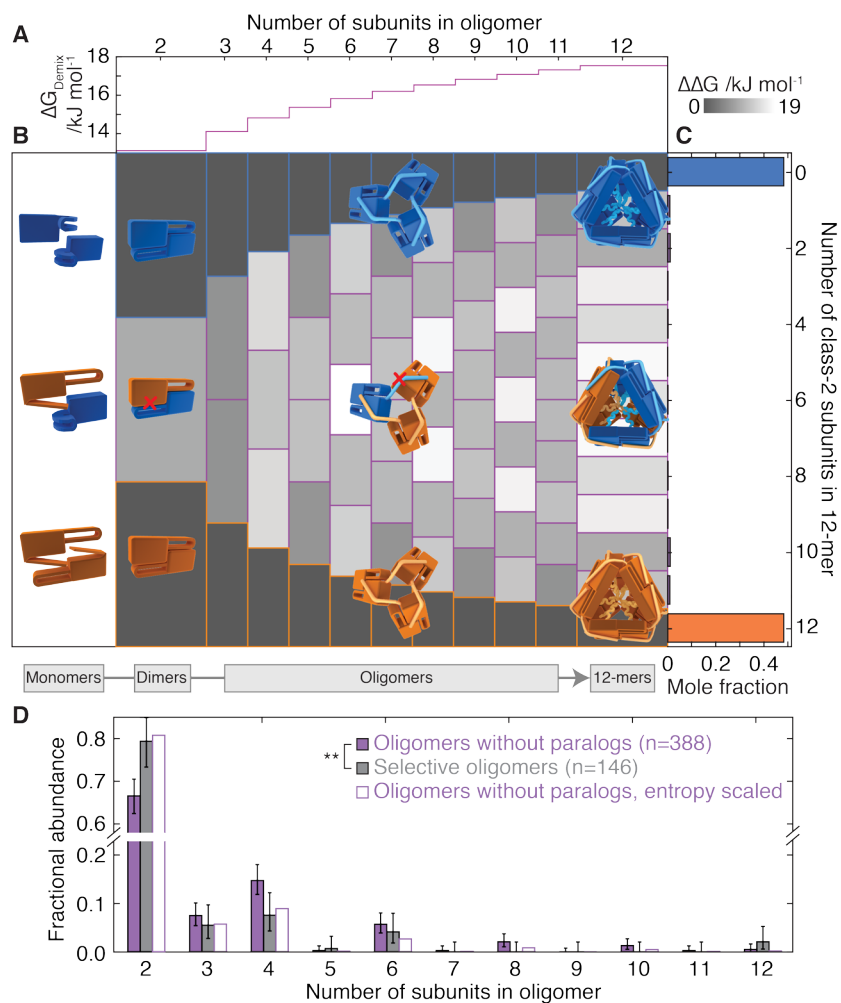


Figure 4. Selective interfaces overcome unfavourable entropy of homomerization.

A) Selective homomerization is entropically unfavourable and requires an energetic penalty upon forming heteromeric contacts to suppress heteromerization. Shown is the theoretical magnitude of this penalty per subunit (ΔG_{Demix}) required to populate heteromers at only 2% of all oligomers. It increases logarithmically with the size of the oligomer, making it more challenging for larger oligomers to be selective.

B) Empirically derived stabilities of all possible heteromers along the assembly pathway compared to homomers of the same size ($\Delta\Delta G = \Delta G_{\text{heteromer}} - \Delta G_{\text{homomer}}$). The upper and lower tiles of each column correspond to homomers of WT-1 and WT-2, respectively. Those in between represent heteromers, with increasing numbers of WT-2 subunits (downwards). The $\Delta\Delta G$ values are positive for all heteromers, meaning that energetic penalty to co-assembly we quantified in selective interactions is larger than the positive entropy of heteromerization.

C) The equilibrium population of homo- and hetero-12-mers calculated based on the values in **B** results in mole fractions of hetero-12-mers just below detectable levels. >96% of subunits partition into homomers, compared to only 0.05% based on the binomial distribution of hetero-oligomers that would arise in the absence of selective interfaces.

D) The oligomeric stoichiometries populated by selective oligomeric paralogs (grey fill) are smaller with a particular excess of dimers than for a control set of oligomers that have no paralogs (purple).
** $p < 0.005$, Mann-Whitney rank sums test. Error bars represent 90% Clopper-Person confidence interval, n denotes sample size. Applying a scaling according to ΔG_{Demix} to the control set reproduces closely the observed selective distribution (purple outline, $p = 0.0005$, Akaike information criterion).

Supplementary Materials for

Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions

Georg K.A. Hochberg, Dale A. Shepherd, Erik G. Marklund, Indu Santhanagoplan, Matteo T. Degiacomi, Arthur Laganowsky, Timothy M. Allison, Eman Basha, Michael T. Marty, Martin R. Galpin, Weston B. Struwe, Andrew J. Baldwin, Elizabeth Vierling, Justin L.P. Benesch

correspondence to: justin.benesch@chem.ox.ac.uk

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S15
Tables S1 to S2

Other Supplementary Materials for this manuscript includes the following:

Data S1 to S2

Materials and Methods

Bioinformatic analysis

For *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* we used the BioGRID database as the interactomics dataset (19). We first identified all proteins that have been annotated as capable of self-interaction based on data obtained using any method labeled as “direct” in the dataset. For *E. coli* we used a dataset that includes primary yeast two-hybrid data, as well as data reported in the literature as summarized by the authors (20). In each dataset, we first recorded the BioGRID identifiers of all proteins that could self-assemble (i.e. reported an interaction if used as both bait and target in a yeast two-hybrid experiment) and downloaded their sequences from UniProt (21). Genes whose BioGRID identifiers mapped to more than one UniProt identifier were excluded from further analysis. To find paralogs within the set of all homomers within each organism, we conducted a reciprocal BLAST search (22), using our list of all homomeric proteins as both the query and database within each organism and a BLAST E-value cut-off of 10^{-10} . We only assigned hits as paralogs if the aligned portion of the hit covered at least 60% of the sequence the shorter protein (either the query or the hit). Since this means that, in all cases, both members of any pair of paralogs can form homomers, we took this as evidence that they evolved from ancestor that could do the same. This is a necessary shorthand in the absence of outgroup information (4), and will include some pairs of oligomeric paralogs whose last common ancestor was not oligomeric.

For each homomer in our database, we then asked if the interactomics dataset reports an interaction with each of its paralogs and noted the identity of an interacting pair, ensuring that each pair was only counted once. If any pair of paralogs was found not to interact, we recorded it as a selective pair only if both proteins had been mentioned together in at least one publication (as a proxy for whether the interaction has actually been tested) in the BioGRID database. In the *E.coli* dataset, we only recorded an interaction as selective if their interaction was actually tested in the dataset. In all datasets we recorded a pair as interacting even if it was only found to interact in one of two reciprocal experiments (i.e. protein A and B are found to interact when protein A is used as bait, but not found to interact when protein B is used as bait). This is conservative with respect to calling any pair selective, but probably overestimates the number of interacting pairs due to false positive interactions. This stringent procedure ensures that our estimates of the fraction of selective paralogs are a lower bound.

Finally, we recorded the percent sequence identity over sequence aligned in the blast hit for each pair. For the gene ontology (GO) term analysis, we downloaded the corresponding annotations for each organism from www.geneontology.org (23), and divided the intersection of their GO terms by the

union of their GO terms for each pair of paralogs. We discarded pairs in which one or both paralogs did not have any annotations from this analysis.

To evaluate if pairs of selective paralogs that do not co-assemble are co-expressed, we downloaded tissue-specific baseline gene-expression datasets from the EMBL Expression Atlas for humans (32 different tissues)(24) and *Arabidopsis* (4 different tissues)(25). We counted a pair of paralogs as co-expressed in a tissue if both their expression levels were 10-fold higher than the detection threshold reported in the dataset. Pairs for which expression data were missing for at least one paralog were excluded from the analysis. For the two single-celled organisms, we instead computed whether a pair of selective paralogs is localized to the same cellular compartment. For yeast, we used a dataset based on GFP fusion constructs that contains data for five separate growth conditions (one in rich media, four under various types of environmental stress). Pairs in which the localization profile of at least one protein was annotated as ambiguous or below threshold in the original database were excluded from the analysis (26). For *E. coli*, we used the K12 dataset from STEPdb (27), which contains a single growth condition. All statistical tests were performed in the `scipy.stats` (v 0.17.1) Python package. The data is available in Data S1.

Number of self-selective oligomeric paralogs versus the number of times selectivity evolved

The percentages we show report on the number of paralogs that no longer co-assemble (Fig. 1B), not on the number of times selectivity has evolved. The latter number will be substantially smaller because selectivity can evolve after an initial gene duplication event, resulting in two selective paralogs A and B. If these paralogs then each duplicate again (as is the case in our phylogeny), the resulting progeny A1, A2, B1 and B2 will inherit their ancestors' selectivity: A1 and A2 will co-assemble, but neither will be able to co-assemble with B1 or B2 and vice versa. In this manner, subsequent gene duplications can cause the number of selective pairs of paralogs to greatly exceed the number of times selectivity evolved. Establishing the true frequency with which selectivity evolves after duplication would require a more explicit phylogenetic treatment that is beyond the scope of this work (4). Regardless of this caveat, our data are unambiguous about selective assembly between oligomeric paralogs being very common and thus important to understand in proteomes across life.

Site entropy analysis

Site entropy p-values were computed using the Entropy-Two server on the HIV-sequence database website www.hiv.lanl.gov which employs a procedure similar to as described previously (28). We used an alignment of 39 class-1 and 28 class-2 sHSPs that spans the α -crystallin domain and the C-terminus. Briefly, the method works by first computing site-specific Shannon entropies for each position in a

background alignment (in our case the combined class-1 and class-2 alignment). Shannon entropies are then also computed for an alignment of interest (in our case an alignment containing either just class-1 or just class-2 sequences), and the latter subtracted from the former to obtain an entropy difference. Sites at which there is little variation (and hence conservation) in the alignment of interest will have larger (positive) values for this difference than sites where there is more variation in the alignment of interest. To obtain a p -value, the procedure is repeated 10000 times with sequences being randomly assigned without replacement into either the background or alignment of interest. The p -value is then computed as the fraction of time the random samples have larger entropy differences than the observed values. To compute p -values on a log scale, a pseudo count of 1 was added for all sites to the number of times the random samples have larger entropy differences than the observed values. This guarantees the resulting fraction never equals zero (and hence allows for a log transformation of the ratio).

Protein expression and purification

Isotopically labelled protein was produced by growing cells in M9 media supplemented with ^{13}C -glucose as the only carbon source, and either ^{14}N - or ^{15}N -labelled ammonium chloride, and purified as described below.

Full-length sHSPs

WT-1 and WT-2 were expressed and purified as described previously (12). Chimeric full-length proteins were cloned into pET28b using the NdeI and BamHI sites to produce an untagged gene-product and were purified using the same protocol as for the WT proteins (12), but taking different ammonium sulfate cuts and adjusting the ion exchange buffers as appropriate.

Cys-1 was generated by incorporation of the mutations E81C and V151C, based on the αC interfaces observed in HSP16.9 from *Triticum aestivum* (PDB ID: 3L1G) and purified in similar manner to WT-1, except that buffers were supplemented with 2 mM DTT. Cys-1 assembles into a monodisperse 12-mer with comparable collision-cross section ($87.4 \pm 3.9 \text{ nm}^2$) to WT-1 ($87.5 \pm 4.7 \text{ nm}^2$), demonstrating the quaternary structure to be unperturbed by mutation.

^{15}N - ^{13}C was lysed in 20 mM Tris pH 7.5, 1 mM EDTA, 1 mM NaN_3 . Ammonium sulfate cuts were taken at 0–20%, 20–40%, 40–60%, and 60–95%, and most of the chimeric protein salted out between 60 and 95%. The 60–95% fraction was dialysed against the lysis buffer overnight, filtered and loaded onto a 26/60 cation-exchange column (GE) at 1 ml min^{-1} . The protein was eluted with lysis buffer + 1 M NaCl using first a 0–12% gradient over 80 mL, then a 12–100% gradient over 100 mL. Fractions containing

the chimera were pooled, concentrated and loaded onto a Superdex 200 10/300 size-exclusion column (GE) equilibrated in 200 mM ammonium acetate.

$N1^{\alpha}2^{\beta}1$ was lysed in the same buffer as for $N2^{\alpha}1^{\beta}1$. The 0–60% ammonium sulfate cut was taken, and dialysed against lysis buffer overnight. The dialysed protein solution was centrifuged at 20000 g for 15 minutes to pellet insoluble protein, filtered and loaded onto a 26/60 cation exchange column (GE). All subsequent steps were as for $N2^{\alpha}1^{\beta}1$.

$N1^{\alpha}1^{\beta}2$ was lysed in PBS. Ammonium sulphate cuts were taken in PBS, and the protein partitioned into the 90-95% fraction at 4°C. The protein was dialysed overnight into 50 mM tris, 1 mM EDTA pH 10, which caused the protein to precipitate. Precipitate was re-solubilized in 1M NaCl, 50 mM Tris pH 9, yielding soluble protein of very high purity. The solution was concentrated and loaded onto a Superdex 200 10/300 size-exclusion column (GE) equilibrated in 200 mM ammonium acetate.

α -crystallin domain constructs

All α -crystallin domain constructs were codon-optimised for expression in *E. coli*, synthesised (IDT, Belgium) and cloned into a modified pET28b vector containing a TEV protease cleavable N-terminal his-tag using the BamHI and XhoI restriction sites using an InFusion kit (Clontech, CA). Proteins were expressed and purified as described previously (29). Purified protein was concentrated, and used directly or snap-frozen and stored at -80 °C until use.

N-terminal truncation constructs

All N-terminal truncation constructs were codon-optimised for expression in *E. coli* and synthesised (IDT, Belgium). $\alpha1^{\beta}1$ and $\alpha2^{\beta}2$ were cloned into a modified pET28b vector containing a TEV cleavable, N-terminal his-MBP tag. Additionally, $\alpha1^{\beta}1$ was cloned into the same pET28b vector we used the α -crystallin domain constructs, containing a TEV cleavable N-terminal his-tag. The proteins were lysed as per the α -crystallin domain constructs, and purified in one step on 5-mL HisTrap columns (GE). All three constructs were not soluble in ammonium acetate after tag removal, so the tag was left attached for MS analysis.

$\alpha1^{\beta}2$ and $\alpha2^{\beta}1$ were cloned into a pET28b vector containing a TEV-protease cleavable N-terminal his-tag using the BamHI and XhoI restriction site, and purified as per the α -crystallin domain constructs.

C-terminal truncation constructs

$N2^{\alpha}2$ and $N1^{\alpha}1$ were amplified from full-length constructs by PCR and cloned into a modified pET15 vector containing a TEV cleavable C-terminal GFP-his tag using NdeI and NheI restriction sites. Proteins were otherwise purified as per the α -crystallin domain constructs.

Native (ion mobility) mass spectrometry

Native MS experiments were performed using methods described previously (30), employing Q-ToF2 and Synapt instruments (Waters, Wilmslow, UK), modified for the analysis of large protein ions (31). Protein concentrations were 5-30 μM unless stated otherwise. Parameters for all experiments were: capillary voltage: 1.3–1.7 kV; sample cone: 10–80 V; extractor cone: 10–20V, acceleration into collision cell: 10–80 V (no activation) and 80–200 V (for CID). The collision cell was pressurized with argon at ≈ 35 μbar . All IM-MS data were recorded on a modified first-generation Synapt HDMS instrument (Waters, UK) fitted with either with a travelling-wave IM device, or a radially confining drift tube (32).

For spectra of α -crystallin domain constructs, including peptide-binding experiments, a vial of isopropanol was introduced into the source region for charge-reduction (33). Instrument settings for all experiments involving were as previously described (34), and collisional cross sections were measured as described previously (32).

Small angle X-ray scattering experiments

SAXS data were collected at the B21 bending-magnet instrument at the Diamond Light Source (Harwell, UK). Samples were prepared in 200 mM ammonium acetate to a concentration of 5 mgml^{-1} and 2 successive 2-fold dilutions. Protein and corresponding buffer solutions were exposed to the beam in a 1.6 mm diameter quartz capillary at 15 $^{\circ}\text{C}$. The sample capillary was held in vacuum, and subjected to a cleaning cycle between each measurement. A Pilatus 2M two-dimensional detector was used to collect 180-frame exposures of 1 s from each sample and the corresponding buffer. The detector was placed at 3.9 m from the sample, giving a useful Q-range from 0.012 \AA^{-1} to 0.4 \AA^{-1} . Two-dimensional data reduction consisted of normalisation for beam current and sample transmission, radial sector integration, background buffer subtraction and averaging. Each frame was inspected manually and discarded if signs of radiation damage were apparent. Data scaling, merging and Guinier analysis were performed in PRIMUS (35).

Small angle X-ray scattering learning algorithm

The determination of a protein quaternary architecture on the basis of its SAXS curve can be interpreted as a classification problem. If the data generated by different architectures contain distinguishable features, a trained learning algorithm should be able to correctly classify them. For both WT-1 and WT-2 we therefore generated a randomly oriented but symmetrical ensemble of 1000 tetrahedral, 1000 double-ring and 1000 single-ring assemblies. We constructed these models using a

two-step assembly modeling strategy aimed at minimizing the number of missing atoms within. First, based on methods we have described previously (36, 37), we produced 12-mers by symmetry operations on the appropriate protomer (α -crystallin domain structure, including bound C-terminus modeled by structural alignment). We subsequently modeled in the missing N-terminal regions assuming symmetry matching to the overall structure such that atoms avoid co-penetration. N-terminal structures were obtained by homology using the PSIPRED webserver (38).

For WT-1, the α -crystallin domain protomer comprised residues 49-143, and a model of the N-terminal region 1-48 was obtained using PDB ID: 3HHV as template. For WT-2, the protomer comprised residues 47-140, and PDB ID: 1BUC was used. We simulated SAXS data between 0 and 0.3 \AA^{-1} for all the resulting 6000 models using Crysol (39). We used 800 signals per topology per protein to train a Random Forests learning algorithm (40), with remainder retained for testing the performance of the algorithm. In the testing phase, classification accuracies of 95% and 99% were achieved for WT-1 and -2, respectively.

Chaperone functional partitioning assays

Fresh pea leaves (*Pisum sativum*) were flash frozen in liquid nitrogen, and ground using a mortar and pestle under liquid nitrogen. Frozen cell paste was re-suspended to a concentration of 400 mgmL^{-1} in PBS containing 400 μgmL^{-1} cycloheximide, 20 μM triptolide (both Sigma Aldrich) and one tablet of cOmplete Protease Inhibitors (Roche, UK) per 500 mL of lysate.

Chaperones were incubated at 42 $^{\circ}\text{C}$ in PBS for 30 min prior to being mixed with lysate to allow for subunit exchange. Re-suspended leaf material was spun at 19000 g at 4 $^{\circ}\text{C}$ for 20 min to pellet cell debris, and passed through a 0.22 μM filter. sHSPs were added to a final concentration of 0.2 mgmL^{-1} of each, to a final total concentration of 200 mgmL^{-1} lysate. The mixtures were incubated at 42 $^{\circ}\text{C}$ in a water bath for 2 hr. At particular time-points, 30 μL were withdrawn and the aliquots spun at 10000 g and 4 $^{\circ}\text{C}$ for 30 min to pellet the insoluble fraction. The 15 μL of the supernatant was dissolved in LDS sample buffer immediately. The pellet was washed in 15-30 μL PBS and then re-suspended in a total volume of 30 μL . 15 μL of the re-suspended pellet were dissolved in LDS for gel analysis.

All samples were heated at 95 $^{\circ}\text{C}$ for 10 min prior to gel loading. Gel band intensities were analysed in ImageLab (BioRad). The intensity of each band in the soluble fraction was expressed as a fraction of the intensity of the band at 0 min. Intensities of the insoluble fraction were not analysed due to sample loss when washing the pellets.

The data were fit to an exponential decay to extract pseudo-first-order rate constants. In the mixtures of chaperones, an F-test was performed to determine whether it was justified to fit the intensities of each component with a separate exponential. Degrees of freedom were $n-2$ for the more complex model. An additional degree of freedom was removed from both models to account for the categorical variable that distinguishes intensities from the two different proteins.

Subunit exchange of full-length proteins

Unlabelled and labelled samples in 200 mM ammonium acetate were equilibrated to the desired temperature using a water bath for >20 min before mixing. Labelled and unlabelled samples were then mixed and incubated at the required temperature in a water bath. At the desired time-points, aliquots were taken and snap-frozen in liquid nitrogen. Samples were kept in liquid nitrogen until analysis at room temperature. Spectra for individual time-points were fitted using UniDec (41). Kinetic traces were fitted to a model describing concurrent subunit exchange of monomers and dimers based on a previous implementation (42). Fits were bootstrapped 1000 times to obtain means and standard deviations of all fitted parameters.

X-ray crystallography

$\alpha 2$ was crystallized in 200 mM CaCl_2 , 100 mM sodium cacodylate pH 6.5, 40 % (v/v) polyethylene glycol (PEG) 200 in hanging-drop plates at room temperature. $\alpha 1$ was crystallised in 200 mM Li_2SO_4 , 100 mM Tris-HCl, pH 8.5, 30 % (w/v) PEG 400 in hanging-drop plates at room temperature. Crystals of both $\alpha 2$ and $\alpha 1$ were flash-frozen without addition of cryo-protectant. Diffraction images were processed using XDS. Structures were solved by molecular replacement using the HSP16.9 dimer (PDB: 1GME) for $\alpha 1$, and $\alpha 1$ for $\alpha 2$ as search models. Manual refinement and validation was carried out in Phenix (43). Crystallographic parameters and PDB codes for the structures are given in Table S2.

Unconstrained molecular dynamics simulations

MD simulations were carried out using the GROMACS simulation package (44). A periodic cell corresponding to a rhombic dodecahedron with a shortest distance of 12 Å between the protein and the box edge was made for each wild-type dimer and monomer. The proteins, modelled with the Amber99sb-ildn force field (45), were solvated in Tip4p-ew water (46), with NaCl added to a concentration of 0.154 M. First, all systems were subjected to steepest-descent minimization followed by 100 ps *NVT* simulation at 300 K, with position restraints applied to all heavy atoms in the protein, using a force constant of 1000 $\text{kJmol}^{-1}\text{nm}^{-2}$, and the v-rescale thermostat (47) with a time-constant of

200 fs. A 1-ns *NVT* simulation without position restraints allowed for initial relaxation, followed by 1 ns of *NPT* simulation with Berendsen pressure-coupling (48) to equilibrate the pressure. Dimers and monomers were simulated in the *NPT* ensemble, using a Parrinello-Rahman barostat (49) for 2 μ s each. To improve sampling, both wild-type monomers were simulated in triplicate, starting from three different configurations taken from the last part of the preceding equilibration. All simulations in this study employed the LINCS algorithm (50, 51) to constrain bond lengths, SHAKE (52) to keep water molecules rigid, and virtual interaction sites (53) for hydrogens, allowing for leap-frog integration (54) of the forces in 4-fs time-steps. The particle mesh Ewald method (55) was used for electrostatic interaction outside of a 10-Å cut-off, and shifted Lennard-Jones interactions were used to model van der Waals interactions within 10 Å.

A heterodimer was constructed from the α 1 and α 2 crystal structures by first superimposing a α 2 monomer onto the dimeric α 1 model, then subjecting the resulting dimer to energy minimisation. To relax the dimeric structure, two subsequent simulations with position-restraints applied to the protein were conducted. The first was run for 100 ps, using the heavy atoms from the starting configuration as reference coordinates for the restraints. The second was run for 25 ns, using the backbone atoms from dimeric α 1 as reference coordinates for restraining the inserted α 2 monomer in order to allow for the side chains to relax into the dimeric structure while keeping the backbone in place. *NVT* and *NPT* relaxation, and 2 μ s of *NVT* production simulation were then carried out, following the methods described above. Chimeric constructs comprising all six combinations of the class-1 and -2 sandwich, loop, and β 6 strand – identical to the constructs used in the experiments – were made by swapping parts of the α 1 and α 2 structures. These chimeras were each preprocessed and simulated for 2 μ s, once only, in the same way as the wild-type monomers as described above.

Steered molecular dynamics simulations and umbrella sampling

In order to qualitatively probe the interfaces of the heterodimer in relation to those in the homodimers, we ran a series of umbrella sampling simulations (56) to examine the free energy profile for dissociating the β 6 strand from the neighbouring strand in the β -sheet on the other monomer. To generate starting configurations along the reaction coordinate defined by the centre-of-mass separation of the two strands, we performed 200-ns pulling simulations (one for each homodimer, and one for each side of the heterodimer) where the strands were gradually separated. Specifically, a force was applied between the respective centre of mass of the backbone atoms of the β 6 strand and the hydrogen-bonding residues in the sheet to separate the two parts, using a force constant of 5000 kJ mol⁻¹ and a reference distance that increased by 2 nm over the course of the simulations. Starting structures for 21 umbrella-sampling simulations per interface along the reaction coordinate were

taken from these pulling simulations. Each umbrella simulation was run for 0.5 μs , where the inter-strand distance was restrained to its starting point at the reaction coordinate, using a force constant of 3000 kJ mol^{-1} . While these simulations can reveal differences between the interfaces, we stress that the rupture of only part of the interface, and the limited simulations time for each umbrella simulation, may not allow for exhaustive sampling and thereby compromise the quantitative accuracy of the free-energy profile. Free-energy profiles in the form of potentials of mean force were produced from the umbrella simulation data using the gmx wham tool (57), with bootstrapping for error estimates.

Analysis of MD trajectories and calculation of contact maps

As the dimers remained very similar to their starting structure throughout the 2- μs simulations, and since they had been subjected to prior equilibration, their full trajectories data could be used for analysis. For monomer simulations however, the first 1 μs of each replica was discarded (as this time-period encompassed the transition away from the dimer structure), and the remaining frames pooled together. The RMSD of $C\alpha$ with respect to the starting structure, RMSFs of $C\alpha$, the number of hydrogen bonds, and contact analysis were calculated using GROMACS analysis tools (58). The number, occupancy, and identity of contacts between non-hydrogen atoms were determined for the dimers and the wild-type monomers, using a distance criterion of $<3.5 \text{ \AA}$. This corresponds to the second minimum in the O-O radial distribution function for water, and can be expected to capture most direct or hydrogen-mediated contacts between non-hydrogen atoms in proteins, while excluding most other interactions. From the existence function, we calculated the fraction of time each pair of residues were in contact to generate the contact maps.

Principal component analysis of monomer trajectories

In order to capture the principal structural and dynamical differences between the monomers, we carried out principal component analysis (PCA) on all the combined monomer trajectory data from all wild-type monomer simulations, having first reduced the data to only contain α -carbons in order to make the molecular topologies of the class-1 and -2 data compatible. The trajectories of the wild type monomers, chimeras, and the monomeric units of the dimers, were projected onto the first principal component.

We also used principal component analysis to find representative structures for our triplicate monomer simulations. The three repeats of each monomer were assembled and aligned into a single trajectory. The PCA was then calculated on all atoms of this combined trajectory. To find a representative structure, we calculated the median value for all components and then found the frame

that minimized the Euclidean distance, weighted by the eigenvalues of each component, between its principal components and the median components. We found that both trajectories converged on a single most representative frame if seven or more components were used. We used this method rather than cluster averages, because we noticed that in particular in $\alpha 2$, cluster averages were dominated by its extremely flexible β -2 strand and yielded structures that were quite unrepresentative in their loops. Nevertheless, for comparison we also show the top three cluster averages for both monomers using a previously described algorithm (59), and extracting the cluster centroids. This was done for all dimers and the wild-type monomers. A relatively large cut-off of 2 Å (3 Å for $\alpha 2$) was used to obtain few, but large, clusters. The cluster averages of the dimers were used to compute the per-residue RMSD between each monomer and its corresponding dimeric conformation (see below).

Statistical analysis of monomer conformations

To compare the monomeric conformations of $\alpha 1$ and $\alpha 2$ across our three replicates for each monomer we first computed the Euclidian distance between the α -carbons across frames of different repeats (but not within the same repeat) of the same protein using every fiftieth frame. Averaging these distances for each site gave us the RMSD for this site between replicates for the same monomer ($\text{RMSD}_{\text{Repeat}}$). We next calculated an analogous quantity, comparing the two different monomers across all repeats ($\text{RMSD}_{\text{Protein}}$). Sites in regions that consistently adopt distinguishable conformations between the two proteins across repeats should have larger RMSDs when comparing different monomers than when comparing the different repeats for the same monomer. To find which states have statistically significantly larger RMSDs when comparing monomers we used a permutation test. Our calculated RMSD values were assigned randomly without replacement to either the between-repeat or between-monomer category, and the mean RMSD calculated 10000 times for each site. The achieved significance level was calculated as the fraction of all permutations in which the RMSD difference ($\text{RMSD}_{\text{Protein}} - \text{RMSD}_{\text{Repeat}}$) was greater than the observed difference. The values we plot on the structures in Figure 3G,H are the RMSD values for sites that were statistically significant at $p < 0.05$, after Bonferroni correction for multiple testing.

The standard error of the mean for each site in Fig. S14D was calculated by analogy to the variance of multiple Metropolis coupled Monte Carlo chains (60) by first combining the within and between repeat variance through propagation of errors into the total variance σ_{Tot}^2 :

$$\sigma_{\text{Tot}}^2 = \frac{F-1}{FN} \sum_N \sigma_{\text{Tra}}^2 + \sigma_{\text{Rep}}^2$$

Where F is the number of frames, N is the number of repeats of each protein (3 in our case), σ_{Tra}^2 is the standard variance calculated from the frames of each repeat and σ_{Rep}^2 is the variance of the means of the repeats. The standard error of the mean was then calculated from this variance as usual.

Titration experiments

Peptides were obtained from Biomatik (Canada), and binding experiments were carried out and analysed as described previously (34). The concentration of the domain was kept at 10 μM in each experiment. Peptide was dissolved from powder and used at concentrations of 10, 20, 40, 80 and 160 μM . Protein concentrations of α -crystallin domain constructs were measured using a BCA kit (Pierce) before use in titrations. To obtain $\Delta G_{\alpha,\alpha}$ for homodimerization, we recorded mass spectra at varying concentrations for each domain (Fig S6A), extracted the relative monomer and dimer intensities and fit them to a model describing homodimerization. To extract $\Delta G_{\alpha,\alpha}$ for heterodimerization, we kept the concentration of one domain constant while varying the other (Fig. S13E).

$\Delta G_{\alpha,\alpha}$ values were obtained by using $\Delta G_{\alpha,\alpha} = -RT \ln K_D$, where the dissociation constants, K_D , for homodimers were calculated by fitting the relative amount of dimer to the expression:

$$[D] = \left(4[P_0] + K_D - \sqrt{(8[P_0]K_D + K_D^2)} \right) / 8$$

where $[D]$ is the concentration of dimer and $[P_0]$ is the total protein concentration. The concentration of heterodimer AB formed between any two monomers A and B at given concentrations of both monomers was calculated by providing an initial guess for the concentration of free monomers $[A]$ and $[B]$ of both components and evaluating the system:

$$[A_2] = [A]^2 / K_{D,1} ; [B_2] = [B]^2 / K_{D,2} ; [AB] = 2[A][B] / K_{D,3}$$

The factor of 2 in the equilibrium describing the formation of $[AB]$ corrects for the entropy of mixing that makes a heterodimer comprising two distinguishable monomers intrinsically more favourable than a homodimer that dimerizes in exactly the same way. Specifically, it derives from the subtraction of $RT \ln 2$ from the $\Delta G_{\alpha,\alpha}$ of heterodimers to make them comparable to homodimer values (see below).

The correct concentrations $[A]$ and $[B]$ for a given set of K_D s, and total concentrations of A and B were found by considering:

$$A_{\text{Tot}} = [A] + 2[A_2] + [AB] ; B_{\text{Tot}} = [B] + 2[B_2] + [AB]$$

and minimizing the difference between the calculated and experimental A_{Tot} and B_{Tot} by varying $[A]$ and $[B]$. K_D s for the heterodimers, $K_{D,3}$, were determined by fixing $K_{D,1}$ and $K_{D,2}$ to the values determined

in titrations of homodimers. At each step of the fit, the relative amounts of A, B, A₂, B₂ and AB were found using the procedure outlined above. Two additional parameters were introduced as multipliers of the experimental total concentrations of A and B at each titration point to account for different ionization and transmission efficiencies of the two types of monomers, and the error in their respective concentrations. This correction is made by optimizing the parameters such that the $\Delta G_{\alpha,\alpha}$ s for the two homodimers to match those obtained for the same homodimers in isolation. The ratio of these two parameters was never above 5 or below 0.2, well within the range previously reported in the literature for MS-based ligand-binding experiments (61). Fits were boot-strapped 1000 times to obtain means and standard deviations of all fitted quantities. This was carried out by assembling a synthetic dataset by sampling the original set and including random replacements. The ensuing standard deviation provides a robust uncertainty measure of the fitted parameter. We could not detect any heterodimer formed between α_1 and α_2 in our titration experiments. We could hence only estimate a lower bound for its dissociation constant. We estimated this value to be 400 μ M, based on this value resulting in a predicted abundance of about 1% of the heterodimer in the highest concentrations of our titration series, and assigned a 10% error (40 μ M).

Thermodynamic modelling of dimerization

To determine which interactions govern the energetics of dimerization, we collected all 36 $\Delta G_{\alpha,\alpha}$ values from our titration experiments and used them to evaluate the most significant interfaces that govern the energetics of dimerisation. Because each protein tested comprised a β -sandwich, (S), a β 6 strand, (B) and a loop (L) from either the class-1 or -2 source, the free energy of association can be expressed as the sum of the free energy change associated with specific interfaces formed between the three types of motif (S, B and L). We considered both inter- and intra- molecular terms, such that:

$$\Delta G_{\alpha,\alpha} = \Delta G_{I-II}^{Inter} + \Delta G_I^{Intra} + \Delta G_{II}^{Intra}$$

where the subscripts I and II label arbitrarily the specific monomer under consideration. At its most complex, this framework accommodates 12 possible contributions to the free energy in the set of chimeras, 6 inter- and 6 intramolecular:

$$\begin{aligned} \Delta G_{I-II}^{Inter} &= \Delta G_{S \cdot S}^{Inter} + \Delta G_{B \cdot B}^{Inter} + \Delta G_{L \cdot L}^{Inter} + \Delta G_{S \cdot B}^{Inter} + \Delta G_{L \cdot S}^{Inter} + \Delta G_{B \cdot L}^{Inter} \\ \Delta G_{I,II}^{Intra} &= \Delta G_{S \cdot B}^{Intra} + \Delta G_{L \cdot S}^{Intra} + \Delta G_{B \cdot L}^{Intra} + \Delta G_S^{Intra} + \Delta G_B^{Intra} + \Delta G_L^{Intra} \end{aligned}$$

For the interfaces with two components there will be four possible combinations: both parts could come from the class-1 source, both from class-2, or one part from each. The free energy terms that

depend on a single structural element, $\Delta G_{B,L,S}^{Intra}$, can only be either class-1 or class-2. In this framework therefore, 42 individual free energies are required (9 interfaces with 4 combinations, and 3 interfaces with 2 combinations). For each tested combination of dimers, the specific combination of interfaces can be determined, and the free energy calculated by summing the individual contributions.

To determine the interfaces that are required to explain globally the experimental data, we developed an iterative optimisation procedure. Initially we considered each of the twelve contributions individually, and optimised the four free energies in order to minimise the following chi-squared function, where the sum is over the 36 experimentally measured free energies:

$$\chi^2 = \sum (\Delta G_{Calc} - \Delta G_{Exp})^2$$

The procedure was then repeated with all possible pairs of interfaces, optimising eight individual free energies (2 interfaces, 4 combinations), then all triplets, and quartets of interfaces (12 and 16 individual free energies, respectively). The best fit from each, increasingly complex, model was obtained.

However, increasing the number of free parameters in a model is expected to naturally decrease the χ^2 . To prevent over-fitting, and thereby determine the most suitable model, we performed forward-selection stepwise regression, using F-tests to determine whether, in each case, the reduction in χ^2 on increasing the model complexity was statistically justified. We obtained uncertainty estimates for the fitted parameters by boot-strapping (as above) 1000 times. The correlation between the experimental data and the $\Delta G_{\alpha,\alpha}$ values calculated from our model gave an R^2 of 0.83.

The analysis revealed that the statistically appropriate description of the data includes three specific interfaces, B·S (inter molecular), L·S (inter molecular) and B·S (intra molecular) (Fig. 3E), resulting in twelve individual free energies. The intra-molecular term arises because our dataset involves chimeric proteins, and describes an epistatic interaction between different types of β -strand and β -sandwich within one monomer (62). While it therefore does not aid explaining how dimers of the two classes avoid each other, and can be ignored in our description of the mechanism of discrimination, it represents a significant contribution (2.6±0.5 kJmol⁻¹ for ^{B2}·^{S2}; 4.2±0.3 kJmol⁻¹ for ^{B2}·^{S1}; 4.9±0.3 kJmol⁻¹ for ^{B1}·^{S2}; 3.5±0.3 kJmol⁻¹ for ^{B1}·^{S1}) to the overall free energy. Notably, the $\Delta G_{B,L,S}^{Intra}$ terms were not found to contribute significantly to the observed dimerization free energies, consistent with a model dominated by inter-molecular interactions.

Differential scanning calorimetry

DSC analysis was carried out on a VP-Cap-DSC instrument (Malvern, UK). 400 μL of each protein, at concentrations between 50 and 100 μM (monomer) in 200 mM ammonium acetate, pH 6.9, were injected for each run. Some samples were unstable at 4 $^{\circ}\text{C}$, so the run was performed between 17 and 110 $^{\circ}\text{C}$ at a scan rate of 0.2 $^{\circ}\text{C min}^{-1}$. α^1 (S1^B2^L1) and α^2 (S2^B1^L2) aggregated before any unfolding transition could be observed and were excluded from the analysis. DSC data was analysed in Origin (OriginLab, MA).

Isothermal titration calorimetry

ITC experiments were carried out on an iTC200 MicroCalorimeter (Malvern, UK). α^1 and α^2 were concentrated and dialysed against 200 mM ammonium acetate, pH 6.9 overnight. The final concentrations were 360 μM for α^1 and 1.25 mM for α^2 as determined by a BCA assay (ThermoFisher, MA). The cell was filled with the dialysis buffer and the protein was loaded into the syringe. The injection volume was 2 μL and the temperature of the cell was set to 298 K. ITC data were analysed in Origin (OriginLab, MA).

Phylogenetic identification of class-specific conservation in the α -crystallin domain

87 sHSP sequences from a diverse set of plants were aligned using the T-coffee server (63), and lineage specific insertions removed manually. The best-fit substitution model was determined using the Prottest server (64). The ML tree was calculated in PhyML (65) using the α -crystallin domain and C-terminal parts of the sequence, and the JTT model of protein evolution (66), with empirically derived character frequencies and a four-category gamma distribution to account for among-site rate variation. To find highly conserved residues specific to each clade, we recalculated the α -parameter of the gamma distribution for just class-1, and class-2 sequences, respectively. We then recorded which rate category had the highest posterior probability for each site in either clade (meaning that this site evolves slowly). This allowed us to select sites that had the highest posterior probability for belonging in the lowest rate category in at least one clade, ignoring sites in which the same residue was conserved in both. This method is more reliable than the site-entropy approach used for illustrative purposes in Figure S5, because it is not subject to phylogenetic sampling bias: site-entropies can be made artificially low by only sampling sequences from closely related species (or artificially high by systematically discarding closely related sequences). Using relative evolutionary rates does not suffer from this problem, because it explicitly takes into account the evolutionary relationships between the sequences in the sample and thereby corrects for any sampling bias.

Simulation of mass spectra

To simulate theoretical mass spectra for the hetero-oligomers formed between WT-2 and N1 α 1C2 (Fig. 1G), we first extracted the relative abundances, average charge states, and widths of the charge state distributions of all oligomeric species from the experimental data using UniDec (41). To simulate the spectrum, we followed a previously described approach, employing a Gaussian peak shape, and imposing experimental average charge-states and charge-state distribution widths for each oligomeric species (67). For the abundances we used the output from our thermodynamic model described above.

Structural bioinformatics

To determine the stoichiometry of oligomers that have selective paralogs and oligomers that do not, we examined a recently curated dataset (17). We extracted the stoichiometry corresponding to proteins in our list of homomers in each species if they were also annotated as a homomer in the structural database. This excludes structures in which the protein of interest was co-crystallized with another protein. We only used stoichiometries that were annotated as “correct” or “unknown”, and ignored the protein if an equivalent stoichiometry had already been assigned to one of its paralogs. This is a conservative approximation for how often selectivity evolves for each stoichiometry, because it undercounts cases in which selectivity evolves repeatedly between paralogs of the same stoichiometry.

As the reference set for comparison, we used proteins that had no detectable paralogs in our BLAST search. We did this, rather using oligomers that co-assemble with all their paralogs because the interaction datasets are very sparse, meaning that data for all pairwise interactions in a group of paralogs is rarely available. This means we cannot confidently assert that any particular protein co-assembles with all its paralogs and thus has never been subject to the constraints of selective assembly. Using oligomers without paralogs only partially circumvents this problem, because our BLAST strategy will miss ancient paralogs. It is thus possible that our control set of oligomers contains some selective oligomers. Further, it is conceivable that there are biases in the types of oligomers we recover that arise from our choice of reference set. For example, oligomers with paralogs may for example be systematically older or younger than oligomers without, and this could affect their average oligomeric size. While we are not aware of such a bias having been reported, our structural bioinformatics results should be considered bearing in mind that the approach is non-phylogenetic. It therefore captures broad trends whose exact magnitudes should be regarded with caution. Our data is available in Data S2.

Supplementary Text

Deduction of assembly order of interfaces using truncated sHSPs

To understand the order in which different interfaces form during assembly it would have been ideal to follow their formation kinetically in real time and at high resolution in full-length sHSPs to watch at which point each interface is formed. Particularly for homomers, however, such an experiment is still beyond the capabilities of modern structural biology. As an alternative, we performed a kinetic-arrest experiment using truncated constructs that lack regions of sequence necessary to form particular interfaces. We could then assess how far the assembly process can proceed without particular interfaces. This approach is analogous to deleting the genes corresponding to different enzymes in a metabolic pathway, and then watching which intermediate products accumulate to determine the order in which the enzymes act (68).

Using this approach we made three distinct observations: if we delete the C-terminal region (and hence remove the possibility of the α C interface forming while retaining N·N interactions) the proteins do not oligomerise (Fig. 2B). Conversely, if we delete the N-terminal region (and hence remove the possibility of the N·N interactions while retaining the α C interfaces), then oligomerisation is possible but monodisperse 12-mers do not form (Fig. 2A). Lastly, subunits that possess N-termini preferentially partition into 12-mers, whereas subunits that do not have N-termini do not show this preference (Fig S8C).

Together, these observations imply that N·N interactions cannot form unless α C are also made, and that N·N interactions preferentially stabilize the 12-mer. We thus conclude that N·N interactions most likely do not form in early assembly intermediates that are smaller than 12-mers. In our enzyme pathway analogy, N-terminal contacts would be akin to a late-acting enzyme that cannot perform its function if the one immediately preceding it is yet to conduct its chemical transformation (68).

In addition to these experiments with truncation mutants, we performed kinetic experiments to monitor the timescale of subunit exchange at equilibrium, and hence quantified the rate at which the oligomers dissociate. From these experiments, we observed that the N·N interactions were rate-limiting in oligomer dissociation (Fig. S5,7). Dissociation was sufficiently slow such that, should heteromeric N·N interactions form through random collisions during assembly, they would be kinetically stable on a timescale of tens-of-minutes to hours. Yet, we observed no such long-lived intermediates during assembly. Combined with our truncation experiments, this leads us to conclude that stable N·N contacts are most likely not formed in the early stages of assembly, allowing the two

selective interfaces to completely sort the subunits before N·N form to provide the final thermodynamic and kinetic stabilization of 12-mers.

There may be alternative explanations that can account for our data. It is, for example, conceivable that truncation of the C-terminus somehow results in a conformational change in the N-termini that renders them unable to form their usual interface. Hierarchical assembly pathways similar to the one we describe are, however, thought to be an almost universal feature of protein complexes (69, 70), giving us confidence in our interpretation.

Boltzmann stability of heteromers between class 1 and class 2

Forming heteromers between two distinguishable types of subunit, A and B, is associated with a favourable entropy of mixing that reflects the greater number of ways, $\binom{N}{i}$, in which two types of subunits can be arranged into a heteromer relative to a homomer, where N is the total number of subunits in the oligomer, i is the number of type A in a particular heteromer (and $N-i$ is therefore the number of type B).

Opposing this favourable entropy change upon mixing may be energetic penalties associated with interfaces between A and B subunits that are less favourable than A·A or B·B interfaces. The energetic penalty associated with forming any particular heteromer (e.g. a 12-mer with 6 class 1 and 6 class 2 subunits, A₆B₆) depends on the number of unfavourable contacts made in the heteromer. Within an A₆B₆ 12-mer, there are 924 ways, $\binom{12}{6}$, in which the subunits can be arranged. However these different arrangements are not necessarily energetically equivalent. If we consider the tetrahedron of dimers formed by the sHSPs in this study, we could consider the A₆B₆ 12-mers formed using exclusively homomeric $\alpha\cdot\alpha$ interfaces, with all heteromeric contacts being made between dimers via $\alpha\cdot C$ contacts. Alternatively, the subunits could be arranged such that only heterodimers are formed. These arrangements (and those between these two extremes) are not energetically equivalent because an unfavourable $\alpha\cdot\alpha$ interface is associated with an energetic penalty different from that for an unfavourable $\alpha\cdot C$ interaction. We used a penalty of 11 kJ mol⁻¹ for the difference between a homomeric and heteromeric dimer interface. This derives from our experimental measurements of the $\Delta G_{\alpha,\alpha}$ values for WT-1 (28.2 kJ mol⁻¹), WT2 (30.1 kJ mol⁻¹) and the heterodimer between WT-1 and WT-2 (19.2 kJ mol⁻¹). This value is larger than the calculated stability of the heterodimer based on our thermodynamic model of the dimer interface (calculated as 7 kJ mol⁻¹) because of the imperfect correlation between our experimental values and the model coefficients (R^2 of 0.83). For the penalty of associated with an unfavorable $\alpha\cdot C$ interaction we used 7 kJ mol⁻¹ based on the data in Figure S7C, using the average of the penalties we measured for the first (≈ 6 kJ mol⁻¹) and second (≥ 8 kJ mol⁻¹)

WT-1 peptide binding to a 2 α dimer. To obtain an overall energetic penalty for each heteromer, we enumerated all possible arrangements of subunits within it and noted the number of unfavourable $\alpha\alpha$ and αC interfaces in each.

For oligomers smaller than a 12-mer, we have to assume an oligomeric geometry. For simplicity, we assumed smaller geometries to be topologically approachable as rings of dimers with odd-numbered oligomers containing an unpaired $\alpha\alpha$ interface. In these rings, the αC interactions run along the top and the bottom of the ring, forming two pseudo-vertices. Compared to a tetrahedron, where there are four vertices, this allows for subunits to be arranged with a smaller number of incompatible αC interactions in a heteromer than is possible in a tetrahedron. The ring geometry will therefore lead to a conservative estimate for how unfavourable heteromers are compared to homomers for oligomers with fewer than 12 subunits. In addition, experimental evidence for this architecture comes from our structure of $\alpha 1$, which we observe to pack as a ring of three dimers in the crystal lattice.

To calculate the relative populations of homo- and heteromers formed at equilibrium for an equimolar mixture of class-1 and class-2 subunits we used the Boltzmann distribution:

$$x_i = \frac{\sum_{k=1}^{\binom{N}{i}} e^{-\Delta\Delta G_k^i/RT}}{\sum_{j=0}^N \sum_{k=1}^{\binom{N}{j}} e^{-\Delta\Delta G_k^j/RT}}$$

where x_i is the mole fraction of N -mers containing i WT-2 subunits. k enumerates all $\binom{N}{i}$ possible subunits arrangements in an N -mer containing i WT-2 subunits. $\Delta\Delta G_k^i$ is the energy difference compared to a homomer of subunit arrangement k in an oligomer containing i WT-2 subunits determined as described above. For homomers $\binom{N}{i} = 1$ and $\Delta\Delta G_k^i = 0$, so the numerator of the expression equals 1. In the numerator, $\sum_{j=0}^N$ sums over all possible hetero-oligomers of size N that contain j WT-2 subunits. As the difference between non-selective and selective interfaces tends to zero, $\Delta\Delta G_k \rightarrow 0$, the distribution reduces, in accordance with the expected binomial distribution, to:

$$x_i = \frac{\binom{N}{i}}{\sum_{j=0}^N \binom{N}{j}} = \frac{1}{2^N} \binom{N}{i}$$

corresponding to a distribution that is determined solely by the mixing entropy.

The relative macroscopic stability $\Delta\Delta G_i^{mac}$ of heteromeric N -mers containing i WT-2 subunits compared to homomers was calculated as:

$$\Delta\Delta G_i^{mac} = -RT \ln \sum_{k=1}^{\binom{N}{i}} e^{-\Delta\Delta G_k^i/RT}$$

which reduces to the mixing entropy $-RT\ln\binom{N}{i}$ as $\Delta\Delta G_k \rightarrow 0$.

It is important to note that this simple formulation is only valid for equimolar mixtures of subunits, though it could be extended to allow comparison of the energies of oligomers with a different total number of subunits.

Energetic cost of homomerization

In order for two oligomeric paralogs to be self-selective, their subunits need to make more favorable interactions in homomers than they do in heteromers. The change in free energy associated with replacing WT-1 subunits in a WT-1 homomer with WT-2 subunits, ΔG_{Demix} (which we express as $G_{Homo} - G_{Hetero}$), has to be more positive than the change in entropy associated with the replacement. We approximated the true value of ΔG_{Demix} by calculating the value that results in a molar ratio $\frac{x_1}{x_0}$, where x_0 is the mole fraction of homo-N-mers, and x_1 is the mole fraction of heteromers with 1 WT-2 subunit and 11 WT-1 subunits. Smaller values of $\frac{x_1}{x_0}$ thus reflect a higher degree of self-selectivity.

Using the expression for x_i from above:

$$\frac{x_1}{x_0} = \frac{\sum_{k=1}^{\binom{N}{1}} e^{-\Delta\Delta G_k^i/RT}}{\sum_{k=1}^{\binom{N}{0}} e^{0/RT}} = \sum_{k=1}^{\binom{N}{1}} e^{-\Delta\Delta G_k^i/RT}$$

We next note that for oligomers with closed symmetries all subunit arrangements are energetically equivalent when $i = 1$ (all $\Delta\Delta G_k^1 = \Delta G_{Demix}$) to write:

$$\frac{x_1}{x_0} = N e^{-\Delta G_{Demix}/RT}$$

Rearranging for ΔG_{Demix} :

$$\Delta G_{Demix} = -RT \ln\left(\frac{x_1}{N x_0}\right)$$

The value of ΔG_{Demix} represents the total energetic penalty required per subunit and can be distributed across several heteromeric interfaces.

This expression is somewhat inaccurate for odd-numbered oligomers in which at least one subunit does not have all its interfaces satisfied. In that case not all arrangements are energetically equivalent when $i = 1$ and this expression would underestimate the energy required to suppress heteromerization. For such oligomers (like in the ring-like geometry we used above), it would be necessary to treat the topology of the oligomers explicitly using the approach detailed in the previous

section to determine whether particular ΔG_{Demix} values are sufficient to suppress heteromerization. The expression also assumes that all subunits are associated into oligomers and overestimates the ΔG_{Demix} significantly in cases where the subunit concentration is low enough for monomers to be populated significantly. In our core domains, for example, a ΔG_{Demix} of about 11 kJ mol⁻¹ is sufficient for no detectable heterodimers to form between 1^α and 2^α subunits at low micromolar concentrations, when our formula puts the required value at about 14 kJ mol⁻¹.

For Fig. 4A we set $\frac{x_1}{x_0}$ to 0.02, corresponding to oligomers with eleven WT-1 and one WT-2 subunit, and oligomers with eleven WT-2 and one WT-1 subunit each being populated at 1% of the total.

Generalization to unequal mixtures

To generalize our statistical thermodynamics model to unequal concentrations of the two types of subunit, we began by designating n_k^i the number of oligomers in the ensemble that have i WT-2 subunits, where the index k denotes each of the $\binom{N}{i}$ different arrangements of i WT-2 subunits in an oligomer of size N . The total number of oligomers N_O in the ensemble is then given by:

$$N_O = \sum_{i=0}^N \sum_{k=1}^{\binom{N}{i}} n_k^i$$

The number of ways, W , of achieving a configuration specified by the set $\{n_k^i\}$, is given by:

$$\ln W = N_O \ln N_O - \sum_{i,k} n_k^i \ln n_k^i$$

We introduced three constraints, imposing the conservation of energy, and of the total number of each of WT-1 and WT-2 subunits:

$$E = \sum_{i,k} E_k^i n_k^i \quad n_1 = \sum_{i,k} i n_k^i \quad n_2 = \sum_{i,k} (N - i) n_k^i$$

where E is the total energy of the system, n_1 and n_2 are the total number of WT-1 and WT-2 subunits, respectively, and N is the size of the oligomer.

Then, to find the values of n_k^i that satisfy the expression for $\ln W$ subject to our three constraints, we introduced three Lagrange undetermined multipliers α_1 , α_2 and β to yield:

$$F(\{n_k^i\}, \alpha_1, \alpha_2, \beta) = \ln W + \alpha_1 \left(\sum_{i,k} i n_k^i - n_1 \right) + \alpha_2 \left(\sum_{i,k} (N - i) n_k^i - n_2 \right) - \beta \left(\sum_{i,k} E_k^i n_k^i - E \right)$$

We then maximized F with respect to $n_k^i, \alpha_1, \alpha_2$ and β by differentiating with respect to all four variables and setting the derivative to zero. For α_1, α_2 , and β , this yielded the original constraints, and for n_k^i :

$$\frac{\partial F}{\partial n_k^i} = -\ln\left(\frac{n_k^i}{N_0}\right) + \alpha_1 i + \alpha_2 (N - i) - \beta E_k^i = 0$$

$$\frac{n_k^i}{N_0} = e^{\alpha_1 i + \alpha_2 (N - i) - \beta E_k^i}$$

Letting $a = e^{\alpha_1}, b = e^{\alpha_2}$, we arrived at:

$$\frac{n_k^i}{N_0} = a^i b^{(N-1)} e^{-\beta E_k^i}$$

which is the fractional population of an oligomer of size N with i WT-2 wild-type subunits and subunit arrangement k .

To obtain the fractional population x_i of all oligomers with i WT-2 wild-type subunits we summed over all arrangements.

$$x_i = a^i b^{(N-1)} \sum_{k=1}^{\binom{N}{i}} e^{-\beta E_k^i}$$

To evaluate this expression, we substituted E_k^i with $\Delta\Delta G_k^i$, and need to assign values to a, b , and β . β , as per standard statistical thermodynamics, is equal to $1/k_B T$, while a and b are found by substitution into the expression for n_1 and n_2 , yielding:

$$\sum_{i=0}^N i a^i b^{(N-i)} \sum_{k=1}^{\binom{N}{i}} e^{-\beta E_k^i} = Np$$

$$\sum_{i=0}^N (N - i) a^i b^{(N-i)} \sum_{k=1}^{\binom{N}{i}} e^{-\beta E_k^i} = N(1 - p)$$

where p is the molar fraction of WT-1 subunits.

The values of a and b can then be found numerically for any given p . In the limit where difference between non-selective and selective interfaces tends to zero, $E_k = \Delta\Delta G_k \rightarrow 0$, $a = p$ and $b = 1 - p$, and the distribution reduces to a regular binomial distribution (see previous section).

Statistical analysis of the distribution of stoichiometries

As a crude approximation of the evolutionary process, we assumed that our control distribution is representative of the true distribution of oligomeric sizes in different protein families. Were there no bias in the kinds of oligomers that become selective after gene duplication, our distribution of selective oligomers would be a sample of the underlying control distribution. We can compute the likelihood of the observed frequencies of selective oligomeric sizes by taking the observed frequencies from the control distribution, and computing the likelihood L of the selective dataset using the binomial distribution:

$$L_{selective} = \prod_i \binom{N_s}{k_{s,i}} (k_{c,i}/N_c)^{k_{s,i}} (1 - k_{c,i}/N_c)^{N_s - k_{s,i}}$$

where N_s is the total number of oligomers in the selective dataset, N_c is the total number of oligomers in the control dataset, $k_{s,i}$ is the number of selective oligomers of stoichiometry i and $k_{c,i}$ is the number of oligomers of stoichiometry i in the control dataset.

Our control distribution is a poor fit to the selective distribution (Fig. 4D). Specifically, there are fewer larger oligomers and an excess of dimers, relative to our control distribution, suggesting a bias in the evolutionary process that makes it easier for smaller oligomers to become selective. To test if this bias could stem from the higher energetic barrier to homomerization larger oligomers have to overcome, we considered a model in which the probabilities of observing an oligomer of any particular size in the control distribution are modified according to:

$$p_i = \frac{k_{c,i}}{N_c \log i}$$

where $\log i$ in this expression is proportional to the energetic barrier to homomerization for an oligomer with i subunits (see above). The resulting proportions are then normalized to sum to 1. To assess the fit of this model we computed the new likelihood

$$L_{model} = \prod_i \binom{N_s}{k_{s,i}} (p_i)^{k_{s,i}} (1 - p_i)^{N_s - k_{s,i}}$$

To determine if the introduction of the additional parameter in our model is justified by the improvement in the fit to the data, we computed Akaike information criterion scores (AIC) (71) for both likelihoods, using the number of oligomeric sizes in the control set as the number of parameters, and adding an additional parameter for our log-corrected model.

We then computed the p -value as:

$$p = e^{\frac{AIC_{selective} - AIC_{model}}{2}}$$

Biologically, this model implies that the higher energetic barrier to homomerization leads to a lower average fixation rate for selectivity after gene duplication in the case of larger oligomers. If the evolution of selectivity after gene duplication usually precedes any significant functional diversification, selectivity most likely evolves largely by random genetic drift. Its fixation rate in that case would be determined by the neutral mutation rate at which alleles conferring selectivity are introduced into the population (72). Our model then implies that this neutral mutation rate depends on the energetic barrier that oligomers must overcome to become selective: larger oligomers either need more substitutions to overcome the barrier than smaller oligomers, or they need mutations that confer a large amount of selectivity but are rarer than mutations that have a less drastic effect.

In essence, our model then approximates the mutational distribution of $\Delta\Delta G_k$ s (see our thermodynamic model above). For the effect of mutations on stability this distribution is universally Gaussian (73). Our $(\log i)^{-1}$ correction may thus stem from being of similar shape to the positive side of a similar distribution for $\Delta\Delta G_k$ s of homomerization. We do not consider the particular form we fitted to necessarily be biologically significant, but instead use it to note that a reciprocal relationship can explain, statistically, the difference between the control and the selective dataset. Determining the shape of this distribution more accurately would require a much larger sample set than is currently available.

We also caution that our fit is largely determined by the data for the smallest four oligomeric sizes because of sparse sampling for larger oligomers. There might be other processes that explain the discrepancy between the distributions. For instance, dimeric protein families, in which selectivity evolves often, may have particular functional properties that make the evolution of selectivity especially likely. Or, our sampling may be inadequate to yield reasonable long-term averages for the evolutionary process. Our data may thus be dominated by processes idiosyncratic to the protein families in our limited sample: while mutation rates can vary significantly between different kinds of proteins, our interpretation assumes an average neutral mutation rate that is reasonably approximated by the difference between our distributions. With these caveats in mind, we view our structural bioinformatics results as a hypothesis that warrants more detailed investigation, with more detailed evolutionary models (74), and improved and phylogenetically informed sampling.

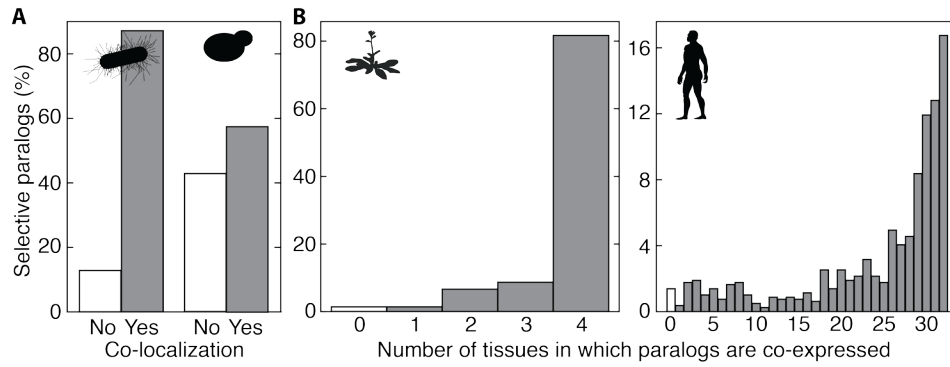


Fig. S1. Paralogs that do not co-assemble are still co-localized and coexpressed. **A)** Fraction of pairs of selective paralogs that localize to the same subcellular compartment in *E. coli* (**left**) and yeast (**right**). **B)** Fraction of pairs of selective paralogs that are co-expressed in the same tissue in *Arabidopsis* (**left**) and humans (**right**). In both cases most pairs of selective paralogs are co-expressed in several tissues.



Fig. S2. Maximum likelihood phylogeny of 87 plant sHSPs. UniProt accession codes are indicated in brackets, and the scale bar indicates the average number of substitutions per site. CI or CII after the species name denotes cytosolic class-1 and -2 sHSPs, respectively. Arabic numbers distinguish between paralogs within the same class. Class-1 and -2 sHSPs were created by a gene duplication event >400 million years ago.

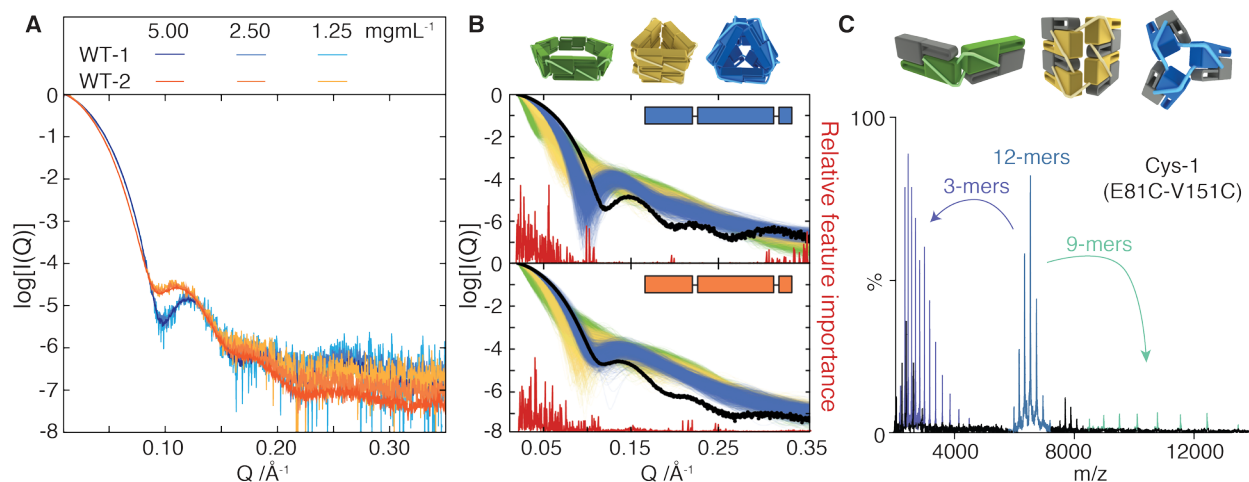


Fig. S3. WT-1 and WT-2 have the same quaternary structures. **A)** Overlay of experimental SAXS curves for WT-1 and WT-2 at three different concentrations. The high degree of overlap at low momentum transfer, Q , suggests they adopt the same quaternary structure. **B)** Both proteins are hexamers of dimers, with all possible arrangements into symmetric 12-mers being variations of tetrahedra, single- or double-rings (13). Experimental SAXS curves (black) of WT-2 (**upper**) and WT-1 (**lower**) overlaid on curves calculated for theoretical models of tetrahedral (blue), double ring (yellow) and circle (green) oligomers. Red lines correspond to the relative importance of features along the theoretical curves used by a learning algorithm (see Supplemental Experimental Procedures for details), trained to distinguish between candidate architectures. Both proteins are classified as tetrahedrons by this algorithm with >95% confidence. **C)** A close ortholog of WT-1 has been crystallized as a double ring (75). To validate a tetrahedral geometry for WT-1, we noted that the number of subunits connected by a closed network of α C contacts is unique to each of the three architectures (upper). We engineered a double mutant, Cys-1, to form a covalent disulphide bond within the α C interface. When the oxidised 12-mers are subjected to gas-phase activation they dissociate into covalently linked trimers (purple) and 9-mers (green). This is only consistent with a tetrahedral architecture. The equivalent construct for WT-2 resulted in highly heterogeneous cross-linked oligomers that were both larger and smaller than the native 12-mers, indicating that in WT-2 the geometry around the engineered cysteine is not unique to 12mers.

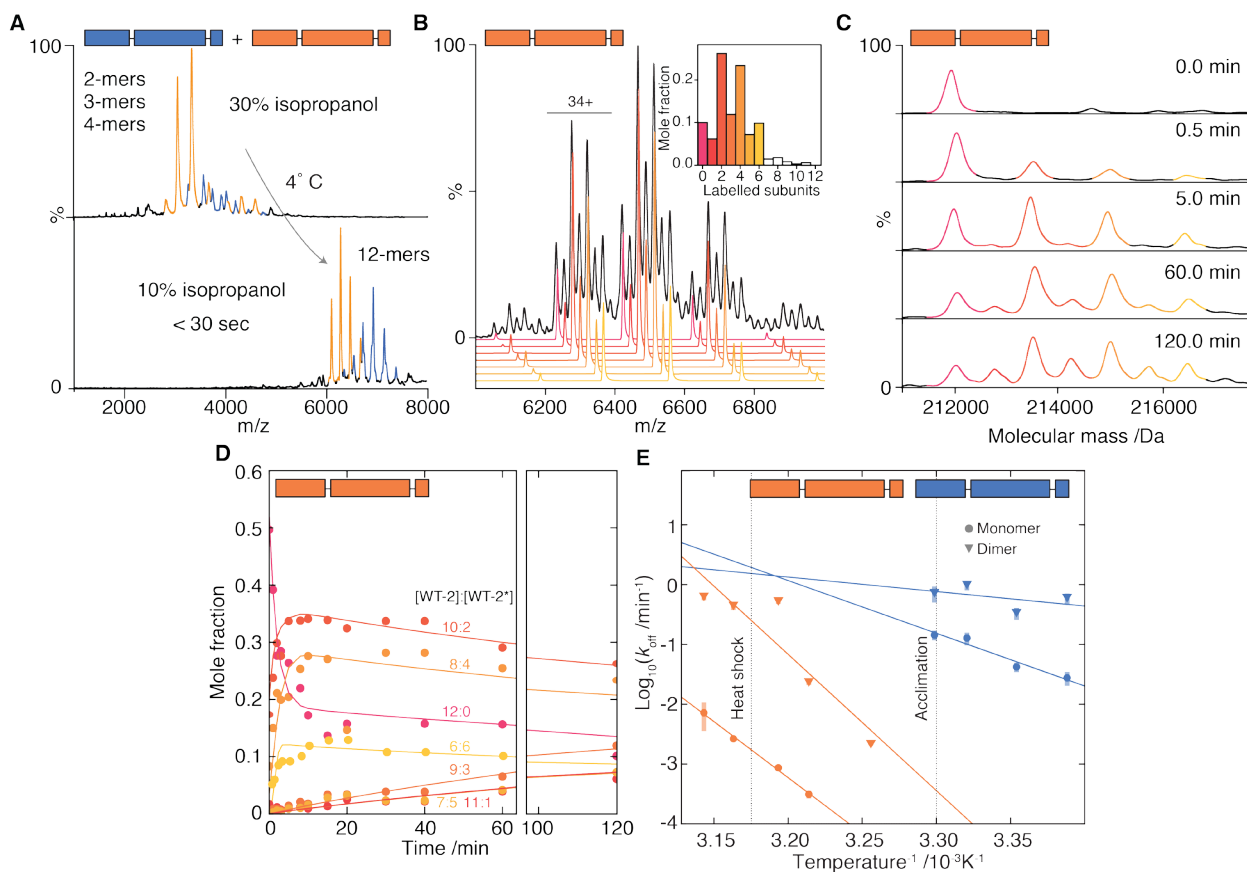


Fig. S4. WT-1 and WT-2 do not hetero-oligomerize even though both continuously recycle their oligomers via subunit exchange. **A)** Disassembly of a mixture of WT-1 and WT-2 in 30% isopropanol results in sub-oligomeric species (**upper**) that reassemble very quickly upon dilution without producing heteromers (**lower**). **B)** Mass spectrum of WT-2 and its ^{13}C -labelled equivalent incubated as a 3:1 ratio at 40°C for 120 min, deconvolved to show the contributions of all exchanging species. 12-mers comprising varying numbers of each subunit are observed, revealing the subunit exchange behaviour of the sHSP. For clarity, only the seven most abundant stoichiometries are shown. Bar chart shows the mole fraction as a function of the number of labelled subunits in the 12-mer (**inset**). **C)** Time-course of mass spectra obtained for the incubation mixture in **(B)**, focusing on the 34+ charge state. 12-mers comprising even numbers of each subunit are observed at earlier times than those comprising odd numbers of each. **D)** Kinetic fit of the time course in **(C)** based on concurrent monomer and dimer exchange. Subunit exchange occurs sequentially. For clarity, only the seven most abundant stoichiometries are shown. **E)** Arrhenius plots for monomer (triangles) and dimer (circles) dissociation rate constants in WT-2 (orange) and WT-1 (blue). The off-rate constants are found by performing the experiments and fits described in **(B-D)** for both proteins at different temperatures. Even though WT-1 exchanges monomers and dimers faster than WT-2, both recycle their oligomers sufficiently fast at equilibrium that, if they are thermodynamically stable, hetero-oligomers should

appear in a mixture within a few minutes at heat-shock or a few hours at room-temperature. Error bars are standard deviations calculated from 1000 bootstrap replicates.

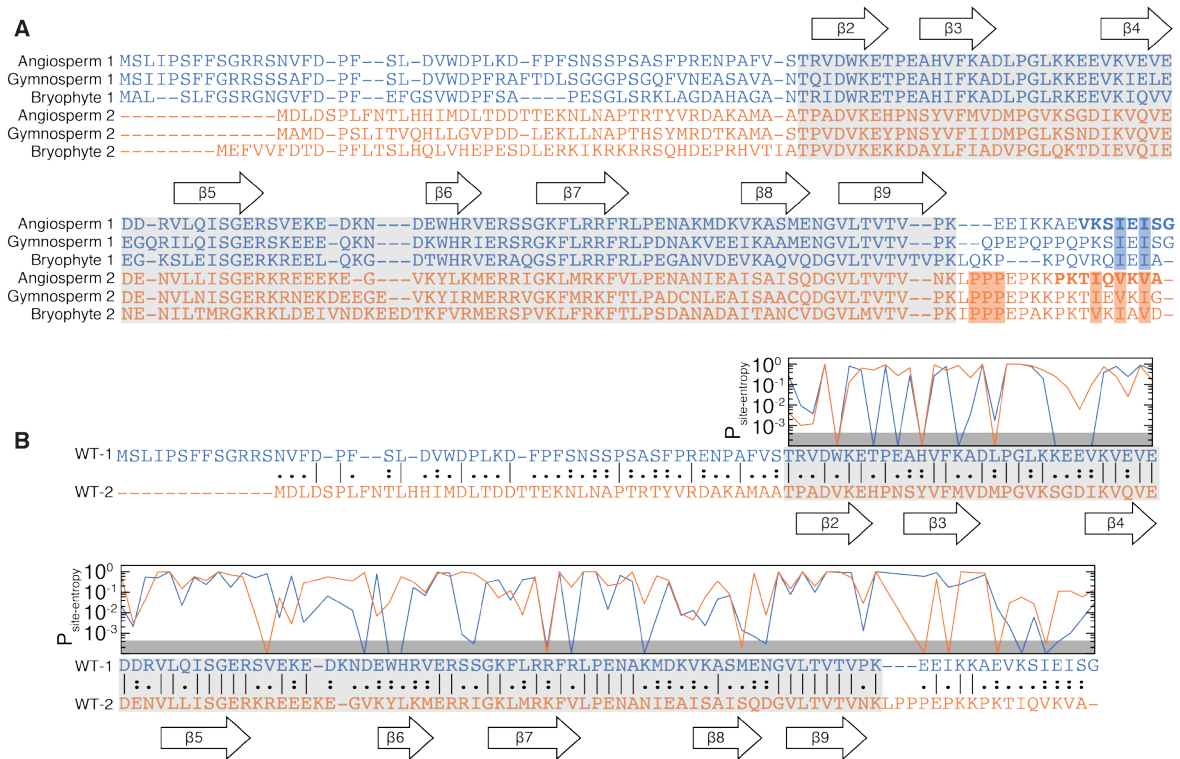


Fig. S5. Sequence alignments of class-1 and class-2 sHSPs. **A)** Sequence alignment of representative class-1 (blue) and class-2 (orange) sHSPs from diverse land plant clades. Both angiosperm sequences are from *Pisum sativum*, bryophyte sequences are both from *Physcomitrella patens*, and gymnosperm sequences are from *Pseudotsuga menziesii* (class 1) and *Picea glauca* (class 2). The α -crystallin domain is shaded grey, and flanked by an N-region and C-terminal tail. Secondary structure elements within the α -crystallin domain are indicated. Class-1 sHSPs incorporate an IXI motif (blue shading) in their C-terminal tails, whereas class-2 sHSP have a longer I/VXI/VXI/V equivalent, as well as a tri-proline repeat (orange shading). **B)** Sequence alignment comparing only WT-1 and WT-2, highlighting the differences between their sequences (| denotes identical, : biochemically similar, and . dissimilar amino acids). The plot above the alignment shows the p -value of a statistical analysis (see Supplementary Experimental Procedures) based on site entropies (derived from an alignment of the class-1 and class-2 sequences in the phylogeny shown in Fig. S2) that identifies sites specifically conserved in either class-1 (blue) or class-2 (orange) sHSPs. Low p -values highlight the differences between the two protein by indicating that a specific amino acid state is conserved only in one class at that site. The shaded area corresponds to $p < 4.3 \times 10^{-4}$ (corresponding to $p < 0.05$ after Bonferroni correction for the 114 sites that were tested). The p -values are only plotted for the α -crystallin domain and C-terminal tail because of poor alignment quality in the N-terminal region.

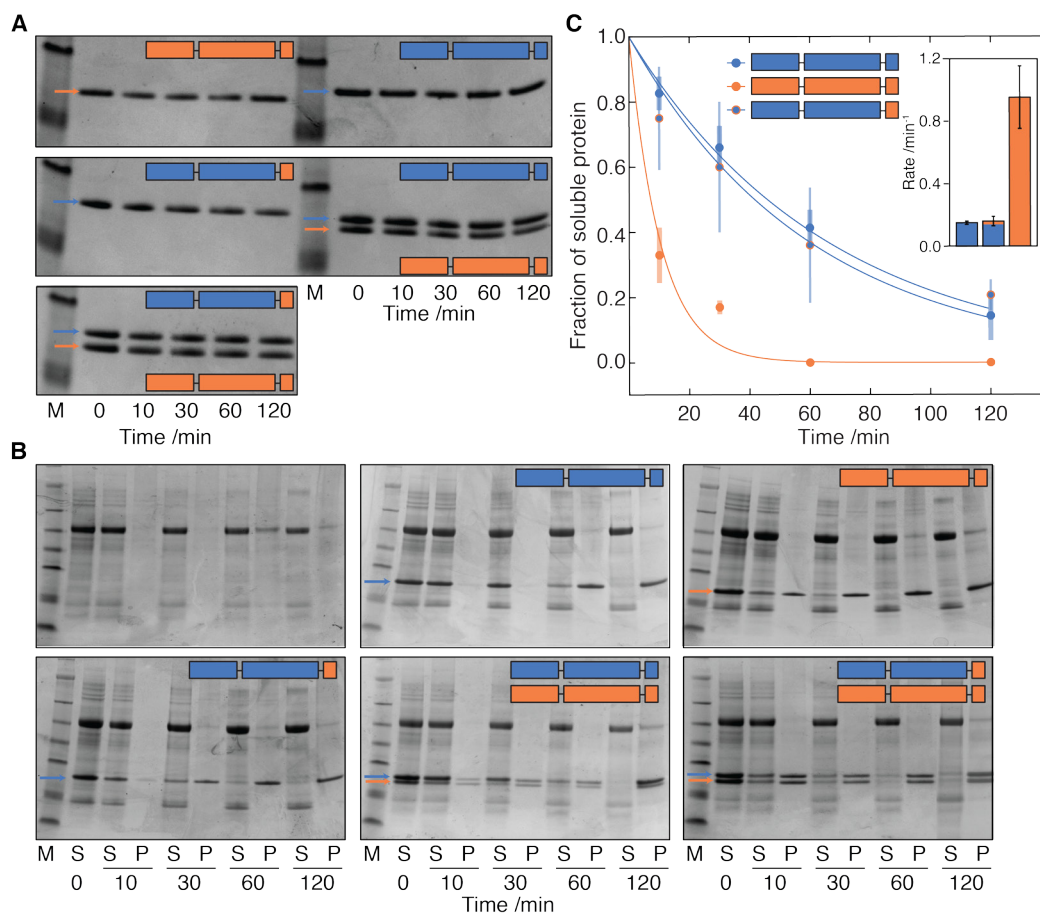


Fig. S6. Hetero-oligomers are functionally impaired compared to homomers. **A,B)** SDS-page based functional assay monitoring the partitioning of chaperone from the soluble fraction (S) into the pellet (P) when incubated without (**A**) or with (**B**) pea leaf lysate. The arrows indicate the locations of the sHSP band(s). In cases where two chaperones are mixed, the lower band always corresponds to WT-2. **A)** Control experiments to test the effect of heating on the proteins in the absence of lysate demonstrate that all three remain soluble for the duration of the experiment. Shown is the protein content of only the soluble fraction, since the pellet was negligible in all cases. M indicates the molecular mass markers at 17, 28 kDa. **B)** Representative denaturing gels showing the partitioning of chaperone between the soluble fraction (S) and pellet (P) when incubated with pea leaf lysate. M indicates the molecular mass markers at 6, 14, 17, 28, 38, 49, 62, 98 kDa. The top left gel shows the partitioning of lysate in the absence of chaperone. Partitioning curves corresponding to experiments involving only one chaperone are shown in **C)**. Experiments corresponding to the two mixtures of chaperones are shown in Fig. 1H. **C)** Partitioning curves and extracted pseudo first-order rate constants (inset) for WT-1, WT-2, and N1 α 1C2. WT-1 partitions at a slower rate than WT-2, and the chimera behaves indistinguishably from WT-1 ($n=3$, \pm standard deviation). Errors in the inset are standard deviations from 1000 bootstrap replicates of the fit.

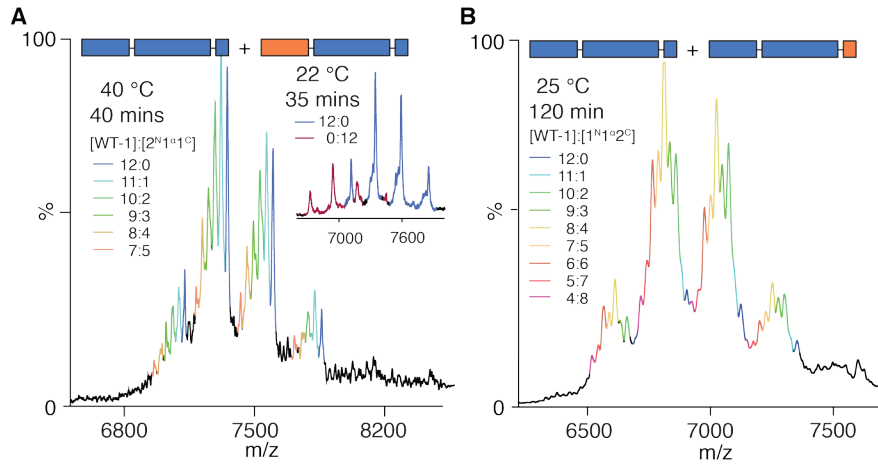


Fig. S7. N-terminal contacts are not selective but determine the kinetic stability of oligomers at equilibrium **A)** A mixture of WT-1 and N²α1C1 results in heteromers populating both odd and even subunits, demonstrating that N·N contacts are not selective. The mixture displays no detectable exchange at 22 °C, and has not reached equilibrium after 40 min at high temperature. This makes its exchange kinetics similar to that of WT-2, whereas WT-1 reaches equilibrium within minutes at this temperature. Breaking of N-terminal contacts is therefore rate-limiting for subunit dissociation during equilibrium subunit exchange. **B)** A mixture between WT-1 and N¹α1C2 exchanges subunits readily at room temperature. Exchange of WT-2 is essentially undetectable at this temperature. The swapped C-terminal tail has therefore not converted the faster exchange behaviour of WT-1 into that of the slower WT-2.

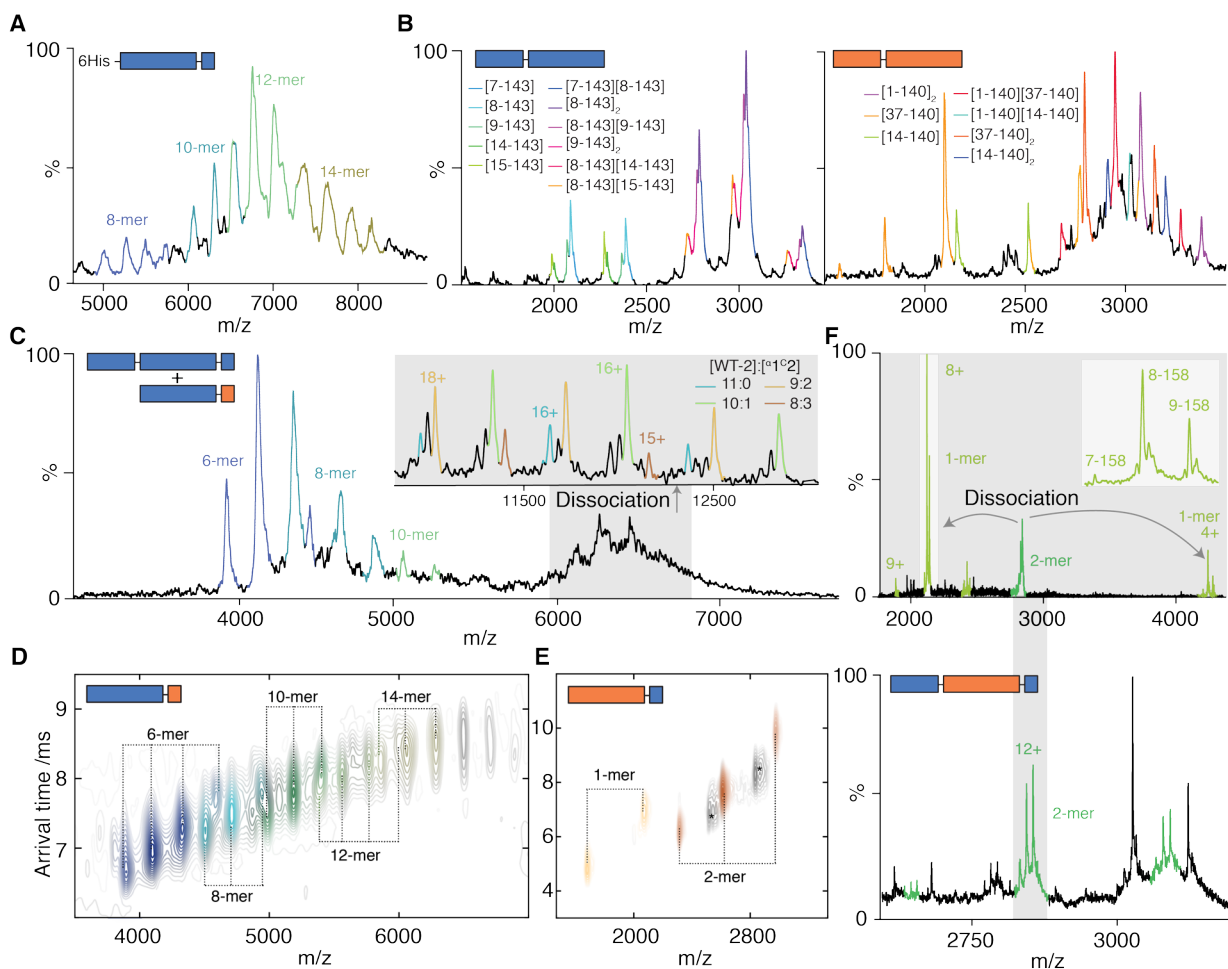


Fig. S8. IM-MS spectra of truncation mutants. **A)** $\alpha^1\text{C1}$, fused to an MBP tag, forms relatively small oligomers, including odd-numbered stoichiometries (shown in Figure 3A). Larger, even-numbered oligomers of $\alpha^2\text{C1}$ are observed if a his-tag is used instead of MBP. This construct is, however, only marginally soluble in ammonium acetate. **B)** $\text{N}^1\alpha^1$ (left) and $\text{N}^2\alpha^2$ (right) do not assemble beyond dimers, leading to exposure of the N-termini. Both proteins suffer from truncations as a result, leading to an ensemble of dimers that incorporate different monomers. Numbers indicate residues remaining after truncation. **C)** Mixture of truncation chimera $\alpha^1\text{C2}$ with WT-1, demonstrating that subunits without N-terminal regions can exchange onto WT-1 12-mers that contain N-terminal regions, but not *vice versa*. This implies that N•N contacts are only made in 12-mers. The region where WT-1 12-mers would normally appear (boxed) shows a broadened of signal, indicating a mixture of subunits of different lengths in 12-mers. The additional stoichiometries populated only by $\alpha^1\text{C2}$ show no sign of broadening, indicating that no WT-1 subunits have incorporated into any stoichiometry other than 12-mers. 11-mers result from tandem-MS of ions in the shaded region, allowing us to identify stoichiometries containing up to 3 truncated subunits (**inset**). Hence, the parent 12-mers contained as

many as 4 truncated subunits. $\alpha 1^C 2$ was chosen for this experiment because it is the only N-terminal truncation that is soluble without a tag that might affect the subunit exchange behaviour. **D)** $\alpha 1^C 2$ assembles into 6-14-mers in even-numbered steps, and therefore has a compatible α -C combination. **E)** $\alpha 2^C 1$ does not assemble beyond a dimer, indicating that its α -C combination is incompatible (starred peaks denote an impurity). **F)** $N 1 \alpha 2^C 1$ has an incompatible α -C combination (**E**) and does not oligomerize despite having an N-terminus (**lower**). Black peaks are impurities. Tandem-MS of the shaded region confirms the dimeric stoichiometry (**upper**), and reveals monomers to be truncated on the N-terminus (**inset**). Numbers indicate residues remaining after truncation.

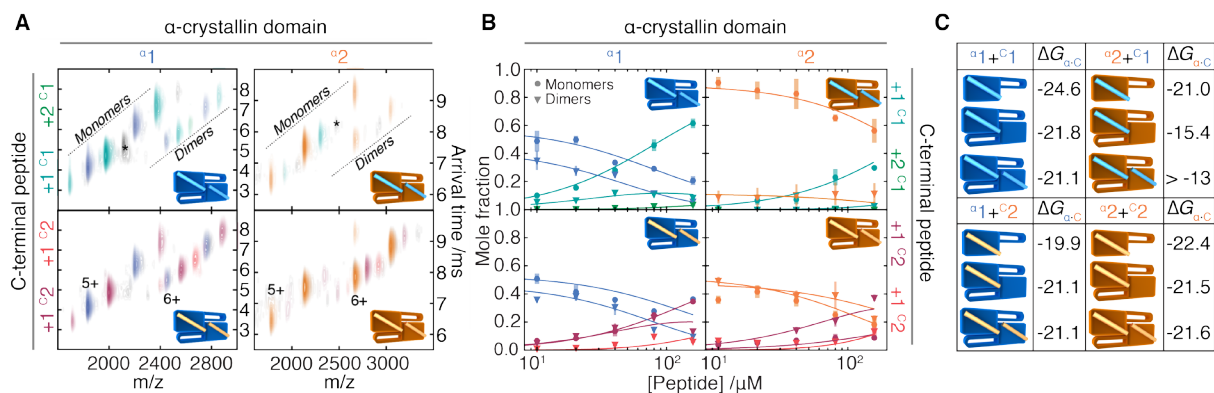


Fig. S9. Thermodynamic basis of selectivity in the α C interface. **A)** IM-MS spectra of α^1 and α^2 incubated with c_1 and c_2 . Sequential binding of peptides to dimers, as well as single peptides binding to monomers can be resolved (* denote impurities). **B)** Isotherms for α^1 and α^2 , monomers and dimers, binding c_1 and c_2 obtained from titration experiments and IM-MS spectra as in **(A)**. Error bars are standard deviations from measurements in triplicate. **C)** Free energies of association, $\Delta G_{\alpha,C}$, obtained from **(B)**. α^2 dimers bind c_1 with much reduced affinity compared to c_2 , indicating selectivity for self-interaction. Measurement errors determined by bootstrapping as <15%. See Table S1 for information about the constructs.

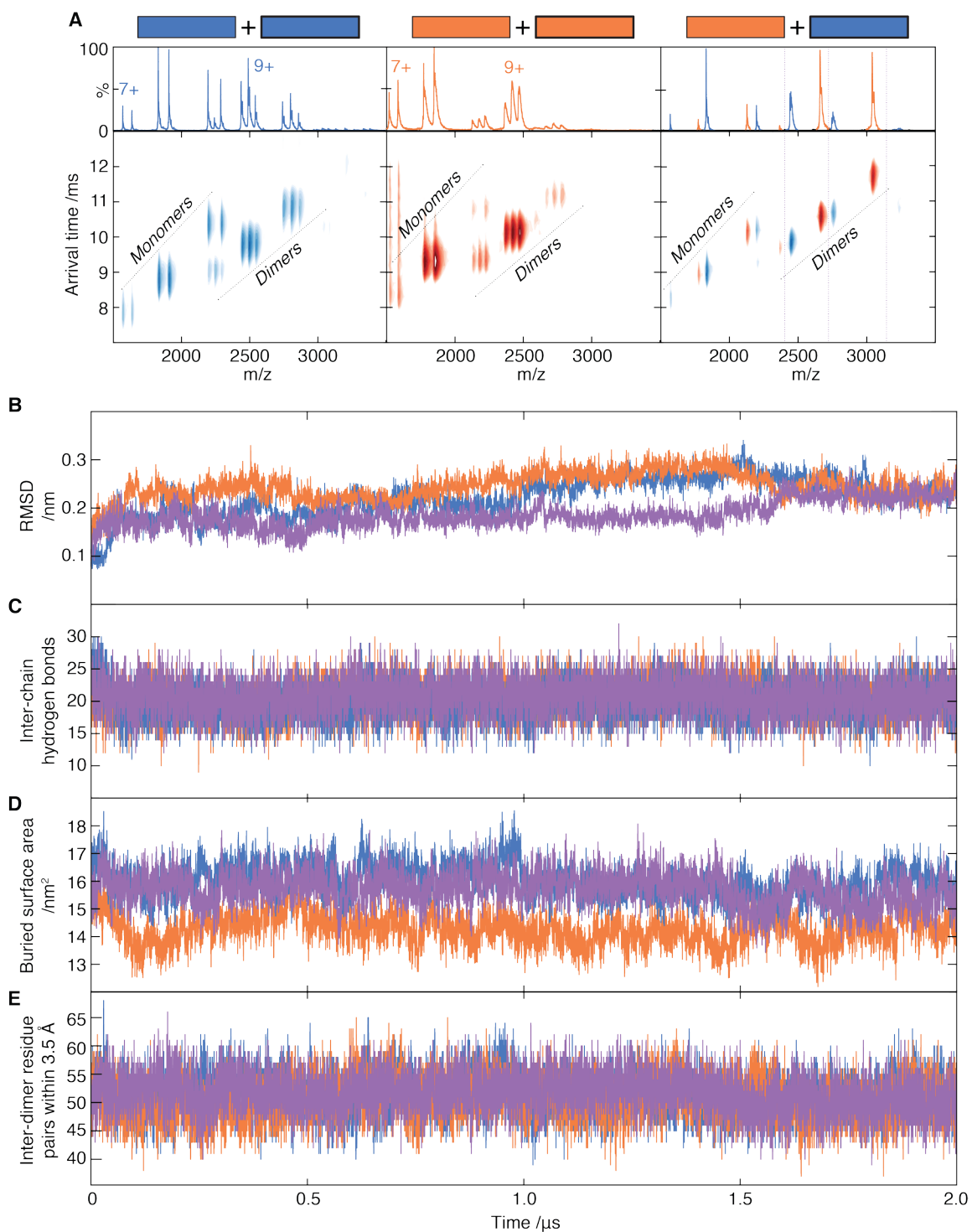


Fig. S10. Heterodimers do not form, despite a seemingly compatible dimer interface. **A)** Mass spectra of $\alpha 1$ or $\alpha 2$ incubated individually (**left** and **middle**) with their ^{13}C -labelled equivalent (thicker outline in schematic), and a mixture of $\alpha 1$ and $\alpha 2$ (**right**). Both $\alpha 1$ and $\alpha 2$ form monomers and dimers at low μM concentrations. For each case, three peaks are observed for each dimer charge state, the middle one corresponding to dimers comprising a heavy and a light subunit as a result of subunit exchange.

Exchange is extremely rapid for homodimers (<1 min, the dead time of our experiment). When $\alpha 1$ and $\alpha 2$ are mixed, only two charge state series are detected for the dimers, corresponding to both homodimers (**right**, orange and blue peaks underlined as dimers). Formation of heterodimers would result in a third peak for each dimer charge state in between the red and blue peaks (purple dotted line). $\alpha 1$ and $\alpha 2$ therefore do not form detectable levels of heterodimers. **B-E)** The $\alpha 1$ and $\alpha 2$ homodimers behave indistinguishably from a $\alpha 1 \alpha 2$ heterodimer in MD simulations. **B)** Root mean square deviation of the $\alpha 1$ (blue), $\alpha 2$ (orange), and the heterodimer (purple) as a function of simulation time. **C)** Number of inter-chain hydrogen bonds as a function of simulation time. The three dimers make an indistinguishable number of hydrogen bonds across the dimer interface. **D)** Buried surface area of the three dimers as a function of simulation time. $\alpha 2$ buries less surface area than $\alpha 1$, but the heterodimer buries as much as $\alpha 1$. **E)** Number of residue pairs that make contact (defined as a <3.5 Å distance between non-hydrogen atoms) between monomers, as a function of simulation time.

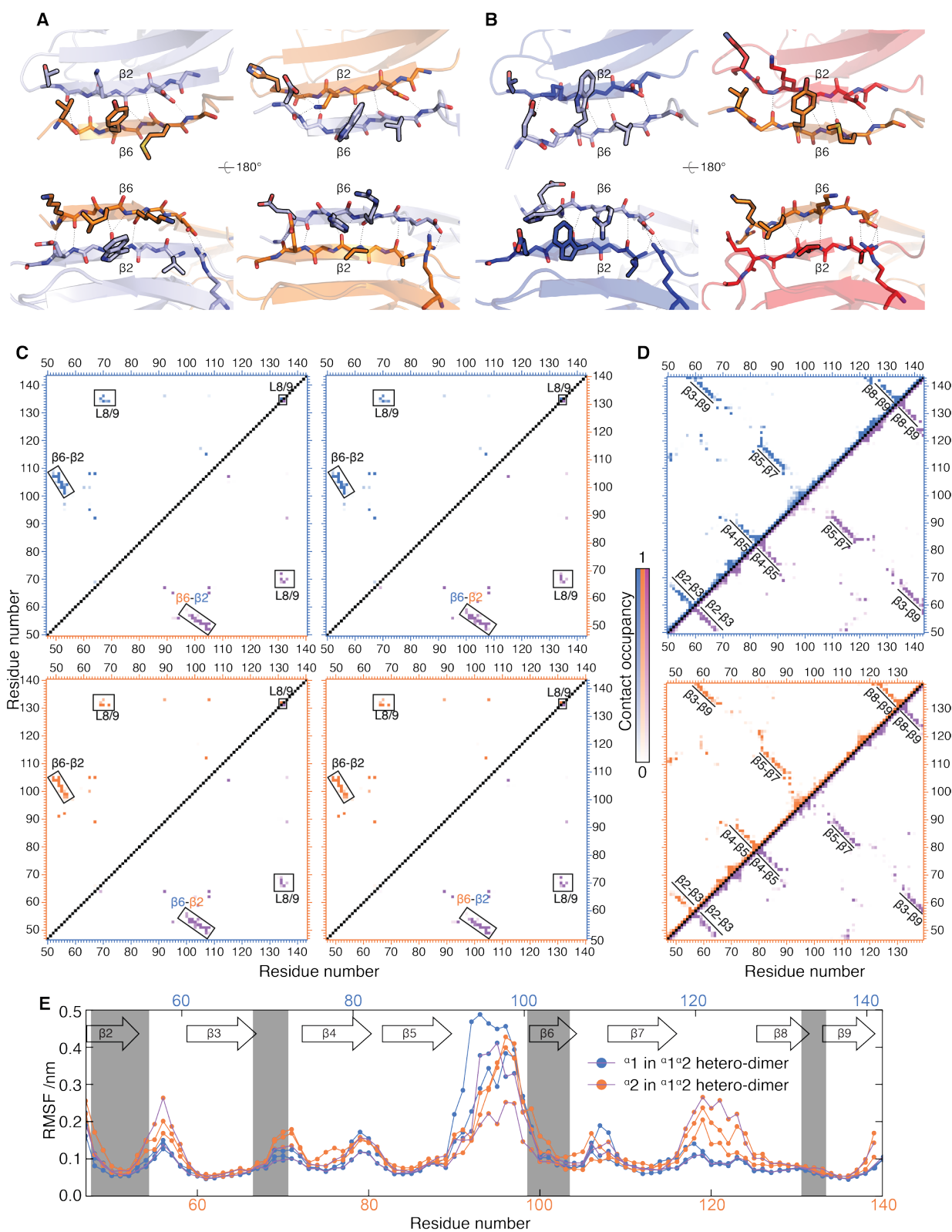


Fig. S11. A simulated heterodimer shows no obvious interfacial incompatibilities and makes very similar inter- and intra-molecular contacts. **A)** Inter-molecular contacts made in the $\beta 6 \cdot \beta 2$ interface in the cluster-average structure of a 2 μ s MD simulation of a heterodimer between $\alpha 1$ (blue) and $\alpha 2$

(orange), shown for the $\alpha 1$ $\beta 2$ interacting with $\beta 6$ on $\alpha 2$ (**left**) and *vice versa* (**right**). **B**) The same region shown as in A, shown for the cluster average structures from 2 μ s MD simulations of the two homodimers. The contacts at the interface rearrange somewhat relative to the crystal structure (Fig. 3D), with the π -stacking interaction not made in the cluster average structure of $\alpha 1$ (though it is found in some frames of the simulation). **C**) Inter-molecular contact maps across the dimer interface of $\alpha 1$ (blue) and $\alpha 2$ (orange) and the heterodimer (purple). The upper left half always shows the contacts made in a homo-dimer, whereas the lower right half always shows contacts made in the heterodimer. **D**) Intra-molecular contact maps comparing the contacts made within the monomers of a heterodimer (lower right half) to those in made by the same protein in a homodimer. The contact maps reveal no obvious differences in the intra- and inter-molecular interactions made by the proteins in their homodimeric, and putative hetero-dimeric, states. **E**) Average root mean square fluctuation plotted per site for all three dimers, with each chain plotted separately. There are no significant differences at the interfaces (shaded).

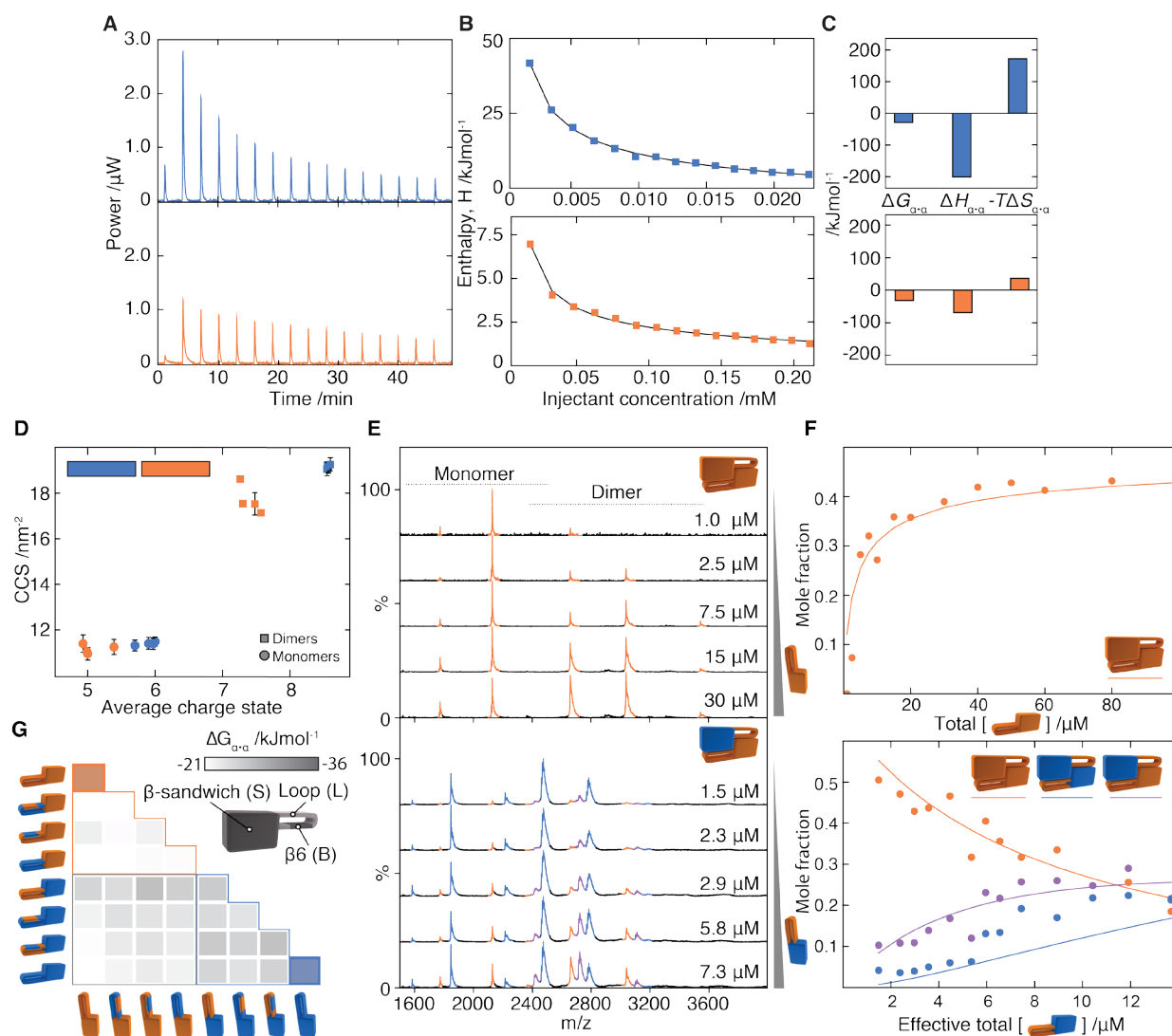


Fig. S12. Measuring the determinants of selectivity in the α -crystallin domain. Raw (A) and integrated (B) isothermal calorimetry traces for dissociation of α_1 and α_2 . C). In both proteins association is exothermic, but the relative contributions of entropy and enthalpy differ, suggesting differences in their association mechanisms. D) The average charge states and collision cross-sections for monomers and dimers of all domain constructs, coloured according to whether they comprise either a class 1 (blue) or class 2 (orange) β -sandwich. The average charge states are low in all cases, indicating folded conformations, and the collision cross-sections in all cases do not change significantly in the chimeras, confirming that all chimeric constructs fold into a shape comparable to that of the WT constructs. Notably, class-1 based constructs have slightly higher charge states, indicative of slightly more facile ionization relative to class-2. E,F) Representative titration experiments to determine $\Delta G_{\alpha,\alpha}$ values for homo- and heterodimers. In both cases, relative intensities are fitted globally to a dimerization model. We first measured for each construct its homodimer $\Delta G_{\alpha,\alpha}$ by obtaining spectra at successive dilutions

(upper). For heterodimers, the concentration of one construct was held constant, while the other was titrated **(lower)**. **G)** Complete set of $\Delta G_{\alpha,\alpha}$ values for each pairwise combination of chimeric and wild-type domains.

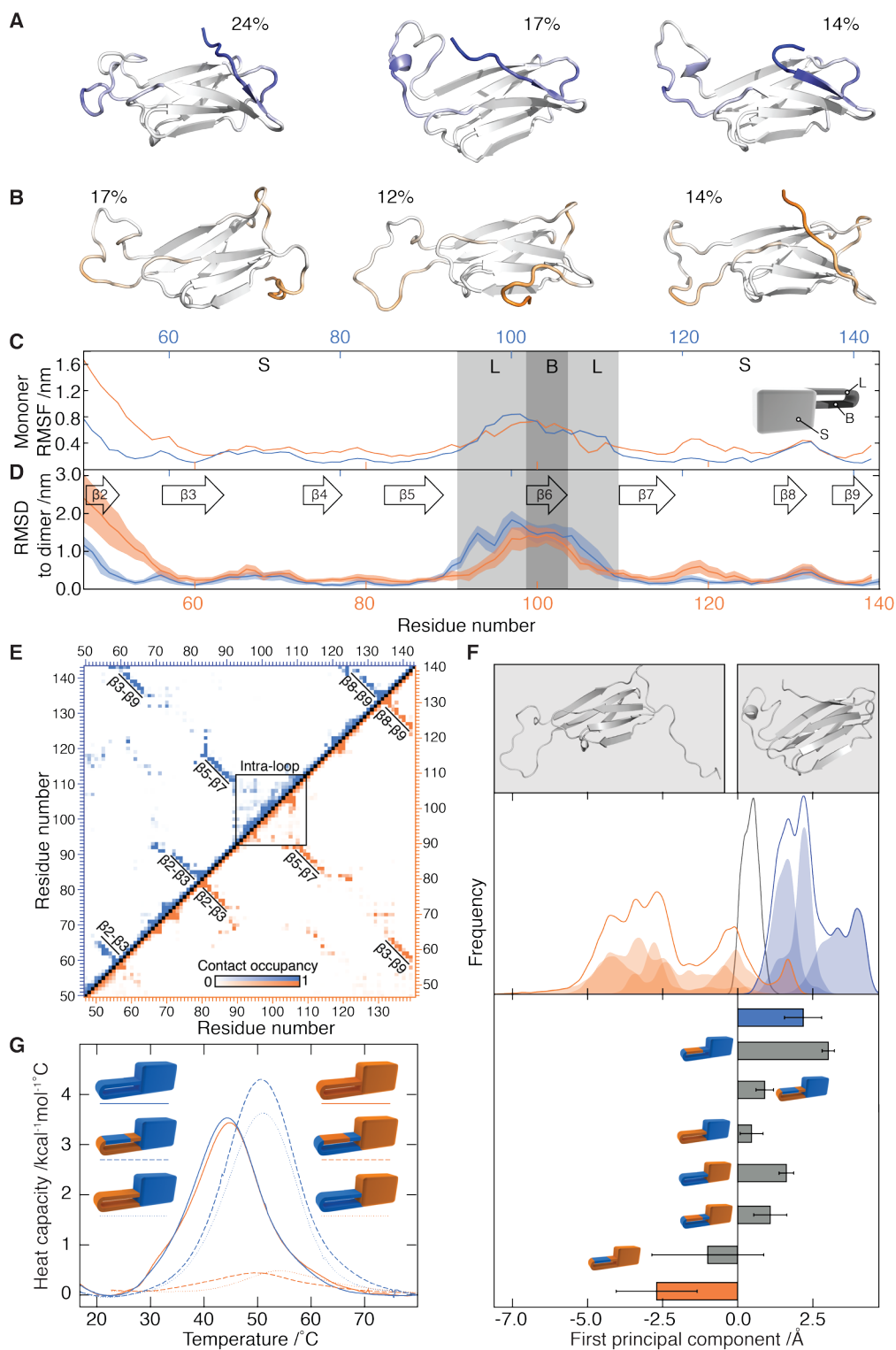


Fig. S13. Different parts of $\alpha 1$ and $\alpha 2$ have a propensity to deform towards their monomeric conformations in MD simulations. **A,B**) Top three cluster average structures of the $\alpha 1$ (**A**) and $\alpha 2$ (**B**) monomer simulations. The relative sizes of the clusters they represent are indicated above each structure. Coloring as in Figure 3H,I. **C**) Root-mean-square fluctuation shown for both proteins. The N-

terminus (containing $\beta 2$) detaches and becomes very mobile in $\alpha 2$ (orange) but not $\alpha 1$ (blue). $\alpha 1$ has a slightly more flexible loop than $\alpha 2$ (light gray shading). **D)** The per residue RMSD of monomeric $\alpha 1$ and $\alpha 2$ when compared to their respective dimeric conformations. $\alpha 1$ is less similar to its dimeric conformation than $\alpha 2$ in the loop (light gray shading). $\alpha 2$ is less similar to its dimeric conformation than $\alpha 1$ around its detaching $\beta 2$. Shaded area represents SEM from three independent replicates (see Supplementary Methods for details). **E)** Contact maps for the monomeric structures of $\alpha 1$ (blue) and $\alpha 2$ (orange) combining all three replicates for each protein. $\alpha 1$ makes more contacts between residues in the loop and $\beta 6$ (square) than $\alpha 2$. $\beta 2$ detaches from $\beta 3$ in $\alpha 2$ but not in $\alpha 1$. **F)** Principal component analysis comparing the conformation of $\alpha 1$ and $\alpha 2$ monomers in the MD trajectories. The three replicates of the wild-type simulations are overlaid as in blue ($\alpha 1$) and orange ($\alpha 2$). The shapes of the combined distributions are shown above as solid blue and orange lines on an arbitrary scale. The largest component captures differences in the loop and $\beta 2$ conformation (**upper**). Projected onto this structure, dimers (black line) fall exactly in between the areas occupied by the two different monomers. Structures shown represent the lowest (**left**) and highest (**right**) value of the component, corresponding to the left-most and right-most point on the x-axis. Chimeric monomers fall between the extremes defined by the wild type domains, indicating that all three parts have distinguishable monomeric conformations between $\alpha 1$ and $\alpha 2$ (**lower**). The $\alpha 1\beta 2$ chimera combines the parts that adopt conformations most similar to those in the dimer in our simulations of the wild-type monomers (**D**) and consistently it occupies the same region along the first component as the dimer structures. Error bars are the median absolute deviations calculated over the trajectory. **G)** Differential scanning calorimetry analysis shows that the melting temperature of $\alpha 1$ chimeras is shifted upwards, indicating stabilized dimers. $\alpha 2$ chimeras (orange lines) that abolish dimerization have poorly defined melting temperatures and diminished $\Delta H_{\text{unfolding}}$ (area under the curve), indicating a loss of structure in the monomer.

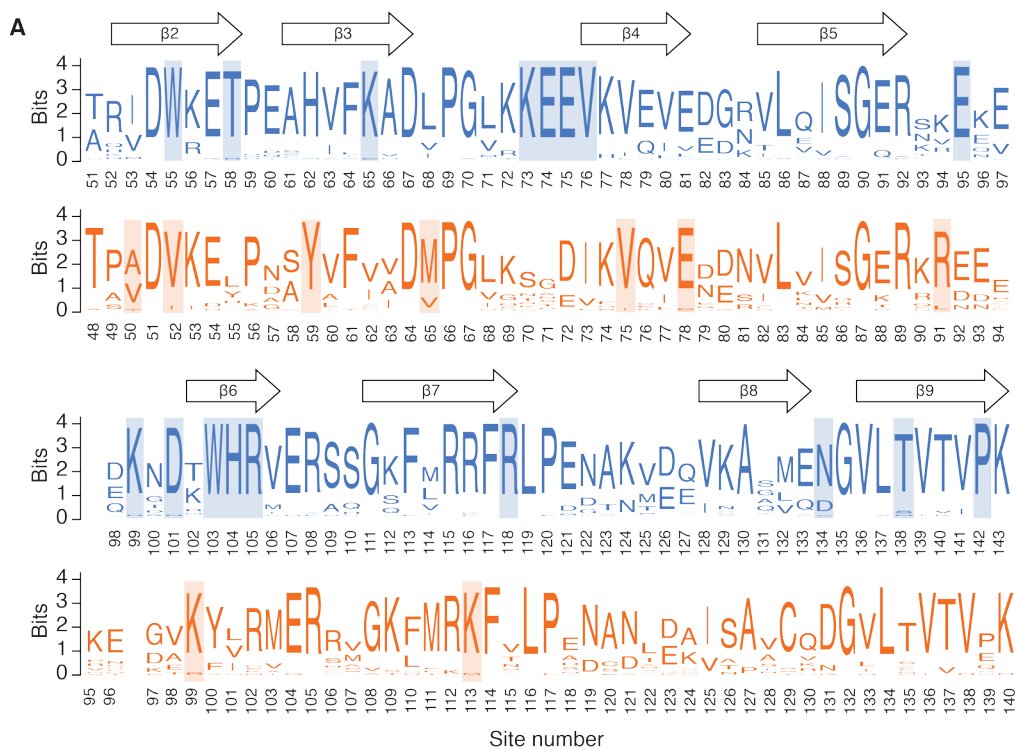


Fig. S14. Class-specific conservation in the α -crystallin domain. **A)** Sequence logos for the α -crystallin domain from WT-1 (upper) and WT-2 (lower). Residues conserved specifically in each class are shaded in blue and orange. Sites were classified as conserved if they fall within the slowest evolving

25% of all sites within each class based on the JTT model of sequence evolution with a four category gamma distribution and empirically derived amino acid frequencies and the tree topology and branch lengths shown in Figure S2 (see Supplementary Methods). Sites at which both classes had the same consensus residue were excluded. If the site was conserved in only one class, but shared the same consensus residue with the other, it was only classified as specifically conserved if, in the non-conserved clade, the site evolved at a rate falling within the 50% fastest evolving sites. We identified 17 and 9 sites that were specifically conserved in class-1 and -2 domains, respectively. **B) Left:** Difference contact maps for $\alpha 1$ and $\alpha 2$ based on the maps shown in Fig. S13E. Sites that have higher occupancy in $\alpha 1$ or $\alpha 2$ are shaded blue and orange, respectively. Interactions that show differences and involve a specifically conserved site have a black outline. **Right:** Representative structures from our MD simulations illustrating the contacts highlighted in the contact map. Conserved residues are shown in blue and orange. W55 packs tightly into the β -sandwich in $\alpha 1$ and prevents it from detaching. $\alpha 2$ has a conserved valine in this position that cannot achieve the same tight packing. K65, R118 and T138 form hydrogen bonds across the $\alpha 1$ β -sandwich, while K73 packs against it without making a hydrogen bond. E95 and D101 make backbone contacts within one of the short helices in $\alpha 1$'s loop. In $\alpha 2$, R91 makes a number of stabilizing hydrogen bonds and Van der Waals interactions across the loop, preventing it from adopting the curled up conformation seen in $\alpha 1$.

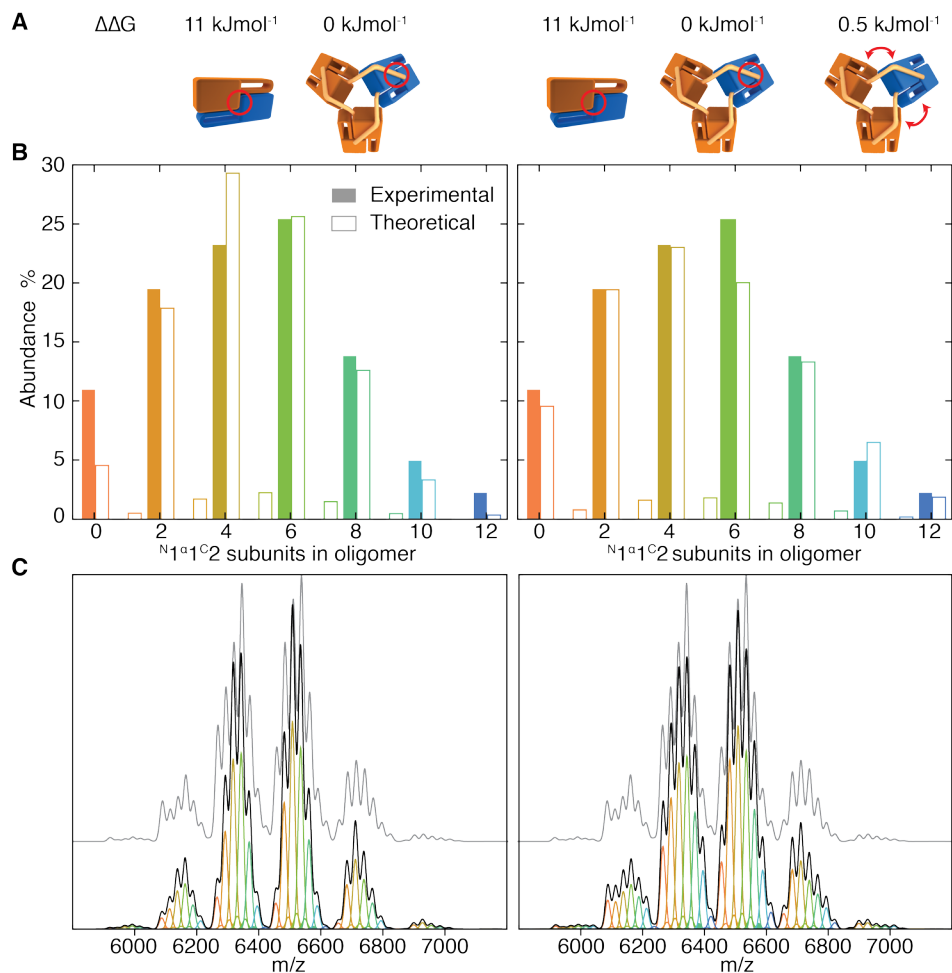


Fig. S15. Empirical measurements of interface stabilities combined with a quantitative model of oligomerization can predict the distribution of heteromers between $1^N1^{\alpha}2^C$ and WT-2, mixed at a molar ratio of 1:2.5. **A)** We developed a quantitative model that describes the formation of heteromers according to their interface stabilities (see Supplementary Experimental Procedures) and tested it against our empirical data from Fig 1G using two different sets of parameters. In our simpler model (**left**), the $\alpha\cdot\alpha$ interface has 11 kJmol⁻¹ selectivity, corresponding to the lower bound we determined for the selectivity of the core domain and there is no selectivity in the $\alpha\cdot C$ interface. For the more complex model (**right**), a small additional selectivity term, optimized at 0.5 kJmol⁻¹, is added for interactions between, for instance, different α -crystallin domains at the vertices of the tetrahedron (76). **B)** Empirical and theoretical distributions of heteromers in the $1^N1^{\alpha}2^C$ and WT-2 mixture. Though both models recapitulate the overall shape of the distribution, the more complex model produces a better fit to the empirical data. Both models predict a low population of oligomers with odd-numbers of either subunit type that we did not observe in our experimental data. **C)** Simulated spectra of the theoretical distributions shown above. The individual peak series corresponding to each heteromer

are shown colored according to **B**, and the sum of all peaks is shown in black. Shown in grey is the back-calculated spectrum of the empirical distribution in **(A)**. Lowly populated 12-mer with an odd number of each subunit type in our theoretical distributions are not detectable as separate peaks in the combined theoretical spectra at the resolution of the MS experiment (and are thus not in conflict with the data in Fig. 1G). Our quantitative model can therefore accurately predict the distribution of a complex mixture of hetero-oligomers based on our empirical measurements of interface stabilities.

Table S1. Sequence details of WT-1 and WT-2 used to construct truncated and chimeric proteins. Numbers refer to amino acid residues included in each region. Letters in brackets refer to nomenclature used in the text for each region. For example, WT-1, which comprises only class-1 amino acids, is referred to as $N1\alpha1C1$. The chimera $N1\alpha1C2$, would differ from WT-1 only in the C-terminal tail.

	Class-1 (HSP 18.1)	Class-2 (HSP 17.7)
Wild type (WT)	1-158	1-157
N-terminal region (N)	1-50	1-47
α -crystallin domain (α)	51-143	48-140
β -sandwich (S)	51-92 + 111-143	48-89 + 108-140
Loop (L)	93-101 + 107-110	90-98 + 104-107
β 6 (β)	102-106	99-103
C-terminal tail (C)	144-158	141-157
Peptide	151-158	149-157

Table S2. Statistics of X-ray data collection and refinement for $\alpha 1$ and $\alpha 2$. The sequences of the crystallization constructs are the same as those listed for $\alpha 1$ and $\alpha 2$ in table S1.

	$\alpha 2$	$\alpha 1$
PDB code	5DS1	5DS2
Crystal parameters		
Space group	C121	P3 ₂ 21
Cell dimensions		
a, b, c /Å	48.86, 82.69 88.90	89.06, 89.06, 142.90
α, β, γ °	90, 92.108, 90	90, 90, 120
Molecules in A.U. ^a	3	6
Data collection		
Synchrotron	In-house ^b	Diamond
Beamline		(104)
Wavelength /Å	1.54	0.843
Resolution /Å	35.02–2.63	77.13–1.85
Reflections	36842 / 10426	566725 / 56563
observed / unique		
Completeness %	98.5 (93.7) ^c	99.9 (99.9)
I/ σ I	7.34 (1.95)	35.12 (1.48)
CC(1/2)	99.1 (88.3)	100.0 (66.9)
Refinement		
Resolution /Å	35.02–2.65	77.13–1.85
R_{work} % ^e	21.5	20.3
R_{free} % ^f	26.4	24.5
Number of non-H atoms		
Protein	2047	4532
Non-protein	1	425
Average B-factor	67.8	52.6
Root mean square deviation		
Bond length /Å	0.015	0.017
Bond angle °	1.85	1.746

a. A.U. = Asymmetric Unit

b. Data were collected at 100 K using a Rigaku Superbright rotating anode diffractometer equipped with a Saturn 944x detector

c. Values in parentheses correspond to the highest resolution shell

d. $R_{mrgd-F} = (\sum |AI(h,P) - AI(h,Q)|) / (0.5 \sum AI(h,P) + AI(h,Q))$
 where $AI = (\sqrt{I}$ if $I \geq 0$ or $-\sqrt{I}$ if $I < 0$) as described previously (77)

e. $R_{work} = \sum |F_o - F_c| / \sum F_o$

f. R_{free} calculated using 5% of the data

Data S1 (separate file). Microsoft Excel file containing all the data in Fig. 1B-D and Fig. S1. Each organism is organized into separate worksheets: one containing all selective pairs of paralogs, and another containing all interacting pairs. For each pair we have listed their UniProt codes, their percentage sequence identity over the aligned region, their percent GO-term overlap, their GO terms, and their expression/localization profiles (see above for detailed descriptions of the source databases). For selective pairs we have also provided the Pubmed codes of papers in which both proteins are mentioned without an interaction being reported. For interacting paralogs we have listed the Pubmed codes for papers that report their interaction.

Data S2 (separate file). Microsoft Excel file containing the data in Fig. 4D. The file contains the UniProt codes, PDB codes and stoichiometries for selective oligomers and oligomers without paralogs (see above for a description of the database used for this analysis).

All data underpinning the entire manuscript is freely available for download, from the following URL:
<https://doi.org/10.5287/bodleian:54jBVeAzw>

References

1. J. A. Marsh, S. A. Teichmann, Structure, dynamics, assembly, and evolution of protein complexes. *Annu Rev Biochem*, (2014).
2. J. B. Pereira-Leal, E. D. Levy, C. Kamp, S. A. Teichmann, Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol* **8**, R51 (2007).
3. E. Kaltenecker, D. Ober, Paralogous interference affects the dynamics after gene duplication. *Trends Plant Sci* **20**, 814-821 (2015).
4. G. Diss *et al.*, Gene duplication can impart fragility, not robustness, in the yeast protein interaction network. *Science* **355**, 630-634 (2017).
5. G. C. Finnigan, V. Hanson-Smith, T. H. Stevens, J. W. Thornton, Evolution of increased complexity in a molecular machine. *Nature* **481**, 360-364 (2012).
6. C. R. Baker, V. Hanson-Smith, A. D. Johnson, Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science* **342**, 104-108 (2013).
7. W. E. Balch, R. I. Morimoto, A. Dillin, J. W. Kelly, Adapting proteostasis for disease intervention. *Science* **319**, 916-919 (2008).
8. K. Richter, M. Haslbeck, J. Buchner, The heat shock response: life on the verge of death. *Mol Cell* **40**, 253-266 (2010).
9. K. W. Helm, G. J. Lee, E. Vierling, Expression and native structure of cytosolic class II small heat-shock proteins. *Plant Physiol* **114**, 1477-1485 (1997).
10. F. McLoughlin *et al.*, Class I and II small heat-shock proteins protect protein translation factors during heat stress. *Plant Physiol*, (2016).
11. M. Kirschner, S. Winkelhaus, J. M. Thierfelder, L. Nover, Transient expression and heat-stress-induced co-aggregation of endogenous and heterologous small heat-stress proteins in tobacco protoplasts. *Plant J* **24**, 397-411 (2000).
12. E. Basha, C. Jones, V. Wysocki, E. Vierling, Mechanistic differences between two conserved classes of small heat shock proteins found in the plant cytosol. *J Biol Chem* **285**, 11489-11497 (2010).
13. G. R. Hilton, H. Lioe, F. Stengel, A. J. Baldwin, J. L. Benesch, Small heat-shock proteins: paramedics of the cell. *Top Curr Chem* **328**, 69-98 (2013).
14. E. W. Wallace *et al.*, Reversible, specific, active aggregates of endogenous proteins assemble upon heat stress. *Cell* **162**, 1286-1298 (2015).
15. T. Perica *et al.*, Evolution of oligomeric state through allosteric pathways that mimic ligand binding. *Science* **346**, 1254-1256 (2014).
16. E. R. Waters, E. Vierling, The diversification of plant cytosolic small heat shock proteins preceded the divergence of mosses. *Mol Biol Evol* **16**, 127-139 (1999).
17. S. E. Ahnert, J. A. Marsh, H. Hernandez, C. V. Robinson, S. A. Teichmann, Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aaa2245 (2015).

18. C. D. Aakre *et al.*, Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594-606 (2015).
19. A. Chatr-Aryamontri *et al.*, The BioGRID interaction database: 2017 update. *Nucleic Acids Res* **45**, D369-D379 (2017).
20. S. V. Rajagopala *et al.*, The binary protein-protein interaction landscape of Escherichia coli. *Nat Biotechnol* **32**, 285-290 (2014).
21. C. The UniProt, UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169 (2017).
22. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
23. C. The Gene Ontology, Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* **45**, D331-D338 (2017).
24. M. Uhlen *et al.*, Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
25. J. Liu *et al.*, Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell* **24**, 4333-4345 (2012).
26. S. Ghaemmaghami *et al.*, Global analysis of protein expression in yeast. *Nature* **425**, 737-741 (2003).
27. G. Orfanoudaki, A. Economou, Proteome-wide subcellular topologies of E. coli polypeptides database (STEPdb). *Mol Cell Proteomics* **13**, 3674-3687 (2014).
28. B. T. Korber *et al.*, Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. *J Virol* **68**, 7467-7481 (1994).
29. A. Laganowsky *et al.*, Crystal structures of truncated alphaA and alphaB crystallins reveal structural mechanisms of polydispersity important for eye lens function. *Protein Sci* **19**, 1031-1043 (2010).
30. F. D. Kondrat, W. B. Struwe, J. L. Benesch, Native mass spectrometry: towards high-throughput structural proteomics. *Methods Mol Biol* **1261**, 349-371 (2015).
31. F. Sobott, H. Hernandez, M. G. McCammon, M. A. Tito, C. V. Robinson, A tandem mass spectrometer for improved transmission and analysis of large macromolecular assemblies. *Anal Chem* **74**, 1402-1407 (2002).
32. M. F. Bush *et al.*, Collision cross sections of proteins and their complexes: a calibration framework and database for gas-phase structural biology. *Anal Chem* **82**, 9557-9565 (2010).
33. J. T. Hopper, K. Sokratous, N. J. Oldham, Charge state and adduct reduction in electrospray ionization-mass spectrometry using solvent vapor exposure. *Anal Biochem* **421**, 788-790 (2012).
34. G. R. Hilton *et al.*, C-terminal interactions mediate the quaternary dynamics of alphaB-crystallin. *Philos Trans R Soc Lond B Biol Sci* **368**, 20110405 (2013).

35. M. V. Petoukhov *et al.*, New developments in the program package for small-angle scattering data analysis. *J Appl Crystallogr* **45**, 342-350 (2012).
36. A. J. Baldwin *et al.*, The polydispersity of alphaB-crystallin is rationalized by an interconverting polyhedral architecture. *Structure* **19**, 1855-1863 (2011).
37. M. T. Degiacomi, M. Dal Peraro, Macromolecular symmetric assembly prediction using swarm intelligence dynamic modeling. *Structure* **21**, 1097-1106 (2013).
38. D. W. A. Buchan, F. Minneci, T. C. O. Nugent, K. Bryson, D. T. Jones, Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res* **41**, W349-W357 (2013).
39. D. Svergun, C. Barberato, M. H. J. Koch, CRY SOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Crystallogr* **28**, 768-773 (1995).
40. L. Breiman, Random forests. *Mach Learn* **45**, 5-32 (2001).
41. M. T. Marty *et al.*, Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Anal Chem* **87**, 4370-4376 (2015).
42. A. J. Baldwin, H. Lioe, C. V. Robinson, L. E. Kay, J. L. Benesch, alphaB-crystallin polydispersity is a consequence of unbiased quaternary dynamics. *J Mol Biol* **413**, 297-309 (2011).
43. P. D. Adams *et al.*, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221 (2010).
44. M. J. Abraham *et al.*, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19 - 25 (2015).
45. K. Lindorff-Larsen *et al.*, Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950-1958 (2010).
46. H. W. Horn *et al.*, Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J Chem Phys* **120**, 9665-9678 (2004).
47. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J Chem Phys* **126**, (2007).
48. H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola, J. R. Haak, Molecular-Dynamics with Coupling to an External Bath. *J Chem Phys* **81**, 3684-3690 (1984).
49. M. Parrinello, A. Rahman, Polymorphic transitions in single-crystals - a new molecular-dynamics method. *J Appl Phys* **52**, 7182-7190 (1981).
50. B. Hess, P-LINCS: A parallel linear constraint solver for molecular simulation. *J Chem Theory Comput* **4**, 116-122 (2008).
51. B. Hess, H. Bekker, H. J. C. Berendsen, J. G. E. M. Fraaije, LINCS: A linear constraint solver for molecular simulations. *J Comput Chem* **18**, 1463-1472 (1997).
52. J. P. Ryckaert, G. Ciccotti, H. J. C. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* **23**, 327-341 (1977).

53. K. A. Feenstra, B. Hess, H. J. C. Berendsen, Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J Comput Chem* **20**, 786-798 (1999).
54. R. W. Hockney, J. W. Eastwood, *Comput simul part* (A. Hilger, Bristol England ; Philadelphia, ed. Special student, 1988), pp. xxi, 540 p.
55. U. Essmann *et al.*, A smooth particle mesh Ewald method *J Chem Phys* **103**, 8577-8592 (1995).
56. G. M. Torrie, J. P. Valleau, Non-physical sampling distributions in Monte-Carlo free-energy estimation: Umbrella sampling. *J Comput Phys* **23**, 187-199 (1977).
57. J. S. Hub, B. L. de Groot, D. van der Spoel, g_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J Chem Theory Comput* **6**, 3713-3720 (2010).
58. S. Pronk *et al.*, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845-854 (2013).
59. X. Daura *et al.*, Peptide Folding: When Simulation Meets Experiment. *Angewandte Chemie Int Ed* **38**, 236-240 (1999).
60. A. Gelman, D. B. Rubin, Inference from Iterative Simulation Using Multiple Sequences. *Statist Sci* **7**, 457-472 (1992).
61. V. Gabelica, N. Galic, F. Rosu, C. Houssier, E. De Pauw, Influence of response factors on determining equilibrium association constants of non-covalent complexes by electrospray ionization mass spectrometry. *J Mass Spectrom* **38**, 491-501 (2003).
62. T. N. Starr, J. W. Thornton, Epistasis in protein evolution. *Protein Sci* **25**, 1204-1218 (2016).
63. H. McWilliam *et al.*, Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res* **41**, W597-600 (2013).
64. F. Abascal, R. Zardoya, D. Posada, ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104-2105 (2005).
65. S. Guindon *et al.*, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321 (2010).
66. D. T. Jones, W. R. Taylor, J. M. Thornton, The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275-282 (1992).
67. F. Sobott, J. L. Benesch, E. Vierling, C. V. Robinson, Subunit exchange of multimeric protein complexes. Real-time monitoring of subunit exchange between small heat shock proteins by using electrospray mass spectrometry. *J Biol Chem* **277**, 38921-38929 (2002).
68. H. Ke *et al.*, Genetic investigation of tricarboxylic acid metabolism during the *Plasmodium falciparum* life cycle. *Cell Rep* **11**, 164-174 (2015).
69. E. D. Levy, E. Boeri Erba, C. V. Robinson, S. A. Teichmann, Assembly reflects evolution of protein complexes. *Nature* **453**, 1262-1265 (2008).

70. J. A. Marsh *et al.*, Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell* **153**, 461-470 (2013).
71. H. Akaike, A new look at the statistical model identification. *IEEE Trans Autom Control* **19**, 716-723 (1974).
72. M. Kimura, Evolutionary Rate at the Molecular Level. *Nature* **217**, 624-626 (1968).
73. N. Tokuriki, F. Stricher, J. Schymkowitz, L. Serrano, D. S. Tawfik, The stability effects of protein mutations appear to be universally distributed. *J Mol Biol* **369**, 1318-1332 (2007).
74. M. Lynch, Evolutionary diversification of the multimeric states of proteins. *Proc Natl Acad Sci U S A* **110**, E2821-2828 (2013).
75. R. L. van Montfort, E. Basha, K. L. Friedrich, C. Slingsby, E. Vierling, Crystal structure and assembly of a eukaryotic small heat shock protein. *Nat Struct Biol* **8**, 1025-1030 (2001).
76. K. K. Kim, R. Kim, S. H. Kim, Crystal structure of a small heat-shock protein. *Nature* **394**, 595-599 (1998).
77. W. Kabsch, Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J Appl Crystallogr* **26**, 795-800 (1993).