bioRxiv preprint doi: https://doi.org/10.1101/081745; this version posted October 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Functional determinants of protein assembly into homomeric complexes

L. Therese Bergendahl* and Joseph A. Marsh

MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK

**Correspondence to therese.bergendahl@igmm.ed.ac.uk*

bioRxiv preprint doi: https://doi.org/10.1101/081745; this version posted October 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Approximately half of proteins with experimentally determined structures interact with other copies of themselves and assemble into homomeric complexes. A number of functional benefits to homomerization have been proposed, and it is often taken for granted that it has been selected for evolutionarily. However, there has been little systematic analysis of the relationship between homomer structure and function. Here, utilizing the large numbers of structures and functional annotations now available, we have investigated how proteins that assemble into different types of homomers are associated with different biological functions. We observe that homomers from different symmetry groups are significantly enriched in distinct functions, and we can often provide simple physical and geometrical explanations for these associations in regards to substrate recognition or physical environment. One of the strongest associations is in dihedral complexes and we suggest that much of this can be related to allosteric regulation. We provide a physical explanation for why allostery is related to dihedral complexes: it allows for efficient propagation of conformational changes across isologous (*i.e.* symmetric) interfaces. Overall we demonstrate clear relationships between protein function and homomer symmetry that have important implications for understanding protein evolution, as well as for predicting protein function and quaternary structure.

bioRxiv preprint doi: https://doi.org/10.1101/081745; this version posted October 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

One of the fundamental challenges in the biological sciences is in understanding the relationship between protein structure and function. This problem is highly relevant not only to the understanding of the evolution of a protein's biological role, but for protein structure and function prediction, protein design, and the prediction of the phenotypic impact of mutations.

Within protein families there is an enormous functional diversity, and sequences, folds and domains can all code for a number of different functionalities depending on their surroundings[1–3]. In the dynamic and crowded environment that is a living cell[4,5], proteins are in constant contact with each other and often carry out their functions as part of larger protein complexes[6–9]. The way that the subunits are organised to form the quaternary structure of a protein complex is a crucial piece of the protein structure-function relationship puzzle, alongside sequences, folds and domains.

Most of the structural information on protein complexes that we have available today is for homomers, i.e. protein complexes that are formed by the assembly of multiple copies of a single type of polypeptide chain. Analysis of published crystal structures shows that roughly 45% of eukaryotic proteins and 60% of prokaryotic proteins can form homomeric complexes[10], and more than three quarters of the published X-ray crystal structures of protein complexes are homomeric[9]. Whilst the high fraction of homomers does reflect biases in protein structure determination, and the fraction of heteromeric complexes (*i.e.* those formed from multiple distinct polypeptide chains) within

bioRxiv preprint doi: https://doi.org/10.1101/081745; this version posted October 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cells is probably higher, it is clear that homomerization is an extremely common biological phenomenon.

Why do proteins assemble into homomeric complexes? A number of benefits have been proposed[6,11–14]. For example, the construction of a large complex from a number of smaller units, rather than one large peptide chain, provides a means for regulation, as the kinetics of homomer assembly is heavily dependent on monomer concentration. Homomerization can also provide both coding efficiency and error control, as a smaller genetic space is used compared to a monomeric protein of a similar size. From a purely physical perspective, the formation of a homomeric complex can potentially lead to higher stability, as a larger proportion of protein surface area is buried, minimizing any energetically unfavourable solvent interactions.

Another notable feature of homomeric structures is that the large majority are symmetric. This is striking, considering the irregularity of individual protein molecules and the lack of any underlying symmetry of the specific amino acids. Symmetry is routinely used as a tool to predict both structure and reactivity of many chemical systems[15] and it has been suggested that, as symmetry contributes to the form of the underlying energy landscape, a symmetric organisation of protein subunits can be associated with increased stability in protein complexes[16,17]. Recent large-scale simulations and directed evolution experiments have indeed also concluded that symmetric assemblies are in fact the most energetically stable and favourable, at least for homodimers, and therefore more likely to overcome the entropic cost of complex formation. Hence there is an overrepresentation of symmetric assemblies that are available for evolution[18–20]. The drive towards formation of

bioRxiv preprint doi: https://doi.org/10.1101/081745; this version posted October 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

symmetric structures is also evident when considering ordered protein complex assembly pathways[21–24]. Protein assembly seems to follow an analogous mechanism to protein folding, as the most common homomer assembly intermediates are in their most energetically favourable, often symmetric, conformations[21].

From the perspective of the specific functions carried out by protein complexes, there are a number of benefits that could drive the evolution of symmetric homomers. One of the more obvious functional aspects is when active sites are located directly at homomeric interfaces. There are a number of examples of shared active sites, such as that of the homotetrameric dihydropicolinate synthase[25]. As the formation of a protein complex leads to a modification of residue interactions, it also alters the conformational space available, leading to a change in the overall protein dynamics. Access to more conformations can lead to new cooperative mechanisms that facilitate allosteric changes at active sites located at different units in the assembly[26]. The textbook example is haemoglobin, where the oxygen affinity at a specific active site on a monomer in the tetramer is heavily dependent on the coordination of oxygen at the analogous active sites in the complex[27].

As is the nature of evolutionary questions, there are many difficulties in trying to determine what aspects contribute to homomeric quaternary structures, and functional advantages are often assumed without any direct evidence. Lynch has in fact suggested that the observed distribution of homomer quaternary structures could be explained by purely stochastic, non-adaptive processes[28,29]. A recent coarse-grain modelling study also shows an example of oligomerization without any obvious functional benefits, emerging purely as

bioRxiv preprint doi: https://doi.org/10.1101/081745; this version posted October 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a side-effect of the thermodynamic stability of the assemblies[30]. Clearly there is an important open question as to how much of protein homomerization can be attributed to evolutionary adaptation.

Here we take advantage of the large amount of data that is now available on both protein structure and function, and carry out a systematic analysis on the functional determinants of protein assembly into symmetric homomers. We demonstrate significant relationships between the structural symmetry of complexes and the functions they perform. We also present results suggesting that complex symmetry is linked to specific allosteric functions. These results have important implications for understanding the evolution of protein complexes, as well as potential for being utilised in both the characterisation and prediction of cellular functions in homomeric complexes[28,29].

Results

Specific protein functions are enriched in distinct protein symmetry groups

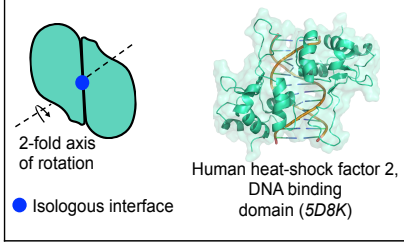
Homomeric protein complexes often assemble into energetically stable structures that are symmetric at the quaternary structure level. This means that group theory can provide a useful metric to characterise and separate most homomeric proteins with respect to a closed or helical symmetry group. Even if a protein does not comply with any symmetry conditions as a whole, it is also worth pointing out that its functional module (for example a DNA-binding- or catalytic site) might be related to the homomeric symmetry groups

bioRxiv preprint doi: https://doi.org/10.1101/081745; this version posted October 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

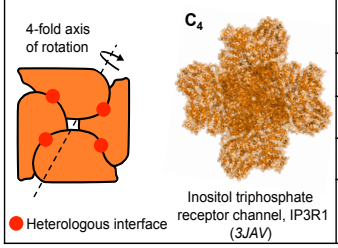
used here. Any function that is specifically associated to a particular symmetry group in this study therefore also has the potential to help characterise for example, asymmetric heteromeric complexes with a symmetric catalytic site and complexes with pseudosymmetry.

The Gene Ontology (GO) project is a large bioinformatics-based initiative with the goal to provide terms for the description of function and characteristics of gene products. The functional annotations of genes in the GO database are based on experiments reported in a number of scientific journals and as such it is not likely to be uniformly based across gene products. However, investigations of over- and under-representation of GO terms has become one of the main uses of the GO to date and enrichment analysis are used routinely on sets of genes from various sources with excellent results[31]. Here we investigate the enrichment of GO terms in a large set of non-redundant non-viral protein structures containing only a single type of polypeptide chain (*i.e.* monomers and homomers). Below we first introduce the closed symmetry groups that are commonly seen in homomeric complexes, followed by the top most significantly enriched functions associated with each group. We determine enrichment of GO terms within the set of annotated structures in order to avoid bias in functionalities that are more highly associated with the PDB than the overall ontology. All of the significantly enriched GO terms within each symmetry group are provided in Supplementary Table 1, and further example protein structures from each symmetry group are provided in Supplementary Figure 3.

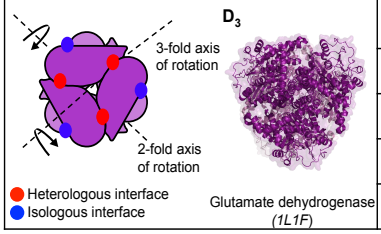
a

Dimeric (C_2) (4830)	GO Annotation	% in symmetry group	% in all homomers	P-value
	*biosynthetic process*	31	23	4×10^{-48}
	cofactor binding	12	7.4	2×10^{-43}
	heterocycle biosynthetic process	16	11	4×10^{-40}
	transcription, DNA-templated	6.3	3.5	1×10^{-30}
	pyridoxal phosphate binding	4.2	2.0	4×10^{-30}

b

Cyclic ($C_{N\ (N>2)}$) (652)	GO Annotation	% in symmetry group	% in all homomers	P-value
	*passive transmembrane transporter activity*	7.5	0.9	2×10^{-32}
	transmembrane transporter activity	12	2.4	1×10^{-31}
	integral component of membrane	11	3.1	6×10^{-22}
	membrane	20	8.9	2×10^{-20}
	monovalent inorganic cation transmembrane transporter activity	4.4	0.5	5×10^{-19}

c

Dihedral (D_N) (1451)	GO Annotation	% in symmetry group	% in all homomers	P-value
	*single-organism metabolic process*	61	38	1×10^{-73}
	oxidation-reduction process	30	16	4×10^{-52}
	carboxylic acid metabolic process	0.2	0.1	3×10^{-26}
	single-organism catabolic process	20	11	3×10^{-25}
	protein tetramerization	2.5	0.4	3×10^{-23}

d

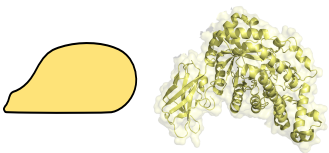
Monomeric (6063)	GO Annotation	% in symmetry group	% in all homomers	P-value
	*hydrolase activity*	28	22	4×10^{-48}
	protein metabolic process	16	11	2×10^{-43}
	hydrolase activity, hydrolyzing O-glycosyl compounds	5.5	3.2	4×10^{-40}
	macromolecule modification	9.4	6.7	1×10^{-30}
	macromolecule metabolic process	29	24	4×10^{-30}

Figure 1: The top five most significant positively enriched GO terms within each symmetry group are tabulated with their associated P-value from Fisher's Exact Test. **(a) Symmetric dimers are associated with annotated functions that involve small and/or twofold symmetric binding partners.** Cartoon of a symmetric dimer illustrating the twofold rotation element and isologous interface (blue) with the DNA-binding domain of a heat-shock transcription factor serving as an example. **(b) Higher-order cyclic protein complexes are required for specialist architectures and are enriched in functional terms involving membrane structures.** A C_4 symmetric complex with the fourfold rotation element and *heterologous* interfaces (red) highlighted. These interfaces are formed by a head-to-tail orientation of the protein subunits (orange). An inositol-1,4,5-triphosphate activated trans-membrane Ca^{2+} ion channel illustrates an example. **(c) Protein complexes with dihedral symmetries have a mixture of both isologous and heterologous interfaces and are enriched in metabolic**

processes. Dihedral complexes in group D_N constitutes either *N* symmetric dimers or two symmetric *N*-mers. A D_3 dehydrogenase illustrates an example of a dimer-of-trimers with heterologous (red) interfaces in the head-to-tail trimers that form a dimer with isologous (blue) interfaces. (d) **Monomers preferentially act together with large substrates.** A β-amylase monomer from B. cereus illustrates an example (yellow).

Symmetric dimers

In our analysis we separate the protein complexes with a single axis of rotation into those with a twofold rotation (C_2 point group, i.e. symmetric dimers), and cyclic assemblies with higher-order rotational axes ($C_{n,\ (n>2)}$). The logic behind this distinction is that C_2 dimers contain isologous (*i.e.* symmetric or head-to-head) interfaces, whereas cyclic homomers form ring-like structures via heterologous (*i.e.* asymmetric or head-to-tail) interfaces that allow them to adopt closed ring-like structures (Figure 1a and 1b).

Our dataset has 4830 non-redundant C_2 dimers, making C_2 dimers by far the most commonly observed symmetry group, comprising 66.3% of homomers. This mirrors previous studies, where small complexes with even numbered stoichiometries have been found to be much more common in the Protein Data Bank (PDB) than larger complexes with odd numbers of subunits[6,32]. An early hypothesis by Monod et al. suggested that an isologous interface is more likely to evolve from a random mutation, as each amino acid occurs twice[33]. It has however since been shown that this effect is somewhat counteracted by the correspondingly low mutation rate of these interface residues, and the prevalence of symmetric dimers seen today is also suggested to be a result of their increased stability[18].

The C_2 dimers in our set are generally associated with GO functional terms that involve the binding to small substrates. Apart from the functions that

require binding to small inorganic and organic cofactors, such as pyridoxal phosphate, the most significantly enriched term is “biosynthetic process” ($P = 4.29 \times 10^{-48}$ Fisher’s Exact test). This term is defined as the energy demanding conversion of small molecules to more complex substances. One example of the latter is the heterocycle biosynthetic process, which is also significantly associated with the C_2 symmetric structures in our set ($P = 4.28 \times 10^{-40}$, Fisher’s Exact test). Interestingly however, the dimeric structures are also significantly associated with DNA-templated transcription ($P = 1.31 \times 10^{-30}$, Fisher’s exact test). The DNA macromolecule does clearly not qualify as a “small” substrate. However, many transcription factors, like nuclear receptors, bind to double-stranded DNA at palindromic sequences or inverted repeats[34], which have local twofold symmetry. Thus, a twofold symmetric homomer provides an ideal mechanism to recognize such a site. This is consistent with, for example, the modular view of transcription factors, where the specific DNA-binding regions can only bind to DNA once a transcription factor dimer has been formed[35]. In Figure 1a, the C_2 symmetrical DNA-binding domain of human HSF2 is used as an illustration[36].

Cyclic homomers

Our dataset contains 652 higher-order cyclic ($C_{n(n>2)}$) assemblies, comprising 8.9% of homomers. All of the most significantly enriched functionalities in cyclic complexes presented in Figure 1b are related to membranes ($P = 4.81 \times 10^{-19}$ to 1.98×10^{-32} Fisher’s Exact test). This is unsurprising, as a majority of membrane proteins have been reported previously to have cyclic symmetries[13]. As the cell membrane is effectively two-dimensional, it makes structural sense to find that the cyclic homomers are enriched in this group,

forming a defined channel or pore through the membrane with the symmetry axis in its centre. An axis of rotation perpendicular to the membrane plane would imply that the different subunits would be in the same relative position with respect to the membrane bilayer. This set-up has been suggested to be a simplification of the insertion process of the protein in the membrane[13], although membrane complexes have also been observed to form co-translationally[37]. Whilst the much larger set of C_2 complexes is found to associate with a large variety of functions, processes and structures, higher-order cyclic complexes are significantly specialized towards functions that require some type of directionality and/or the formation of a channel structure. The recently reported structure of the IP_3R1 ion channel[38] is used to illustrate an example in Figure 1b.

Dihedral homomers

If a protein complex has assembled so that it has two orthogonal axes of rotational symmetry, one being twofold, it is dihedral (Figure 1c) and denoted D_n, where *n* is the highest order of rotation. These complexes can be thought of as resulting from the assembly of either two C_n homomers or *n* C_2 dimers. In our data, 1451 complexes belong to this point group, and it is therefore the second most common symmetry group, comprising 19.9% of homomers. The twofold symmetry axis means that dihedral complexes always have some isologous interfaces but, depending on the mode of assembly, heterologous interfaces can also be present. For example, a “dimer of trimmers” structure with D_3 symmetry will have both isologous and heterologous interfaces, as illustrated with the UDP- glucose dehydrogenase (*3TF5*) in Figure 1c.

Apart from the functional term associated with the formation of the protein complex itself (protein tetramerization, $P = 2.89 \times 10^{-23}$), the most significant GO terms are all related to metabolic processes. Strikingly, more than 60% of the dihedral complexes are associated with the GO term for single-organism metabolic process ($P = 1.44 \times 10^{-73}$).

There are a number of reasons why metabolic enzymes might form dihedral complexes. Active sites of enzymes are often found at the interfaces between two or more subunits[39,40]. As a dihedral arrangement brings together a number of subunits, this leads to a potentially larger concentration of active sites. This can be directly beneficial to the kinetics of its enzymatic function. If the substrate concentration is high, the rate of enzyme activity is a function of not only its rate-constant and enzyme concentration, but also the number of active sites present[41]. The bringing together of many subunit also leads to a generally large enzyme complex being formed from relatively small units. A large protein complex is more likely to be able to provide the correct functional groups for catalysis, have the correct complementarity to its substrate, and also be bulky enough to provide a low dielectric environment for its catalytic process if needed[40].

One crucial aspect of enzymatic activity that must be mentioned is that it is often beneficial for the enzyme to be able to undergo conformational rearrangements that not only facilitate the process itself, but also change the affinity of the substrates if needed. Bringing together a large number of subunits and forming active sites at interfaces could facilitate this type of allosteric control mechanisms[42,43]. In order to address this important issue, we

also investigated our dataset with respect to allostery, as discussed later in the manuscript.

Cubic homomers

Introducing higher-order rotational symmetry elements leads to so-called cubic complexes. The cubic symmetry groups all have one threefold rotational axis combined with a non-perpendicular axis that can be twofold, as in tetrahedral (T) complexes, fourfold, for complexes with octahedral (O) symmetry, or fivefold, forming an icosahedral (I) protein complex. These large complexes are assembled from 12, 24 and 60 subunits respectively (Figure S1). These symmetry groups make up only 1.1 % of the homomers in our dataset, including 56 tetrahedral, 22 octahedral and 4 icosahedral complexes (most known icosahedral structures are viral capsids, which were excluded from our analysis). The top results from our enriched functions in the cubic set are all associated with homeostasis and metal ion binding in general. This is due to the fact that, despite our filtering for sequence redundancy, the set contains a number of octahedral ferritins from diverse organisms. This highlights a potential functional advantage of cubic homomers: the formation of large hollow shells ideally suited for storage purposes[6,44,45]. The 24 monomers in ferritin are organised to form a large octahedral capsule capable of storing and transporting iron oxide. The ferritin complex from *Chlorobium tepidum* is shown as an example in Supplementary Figure 1[46].

Helical and asymmetric homomers

In our non-redundant dataset of homomers with functional associations, the vast majority (96.2%) belong to a closed symmetry group. Of the exceptions,

a tiny fraction (15 or 0.2%) have open helical (H) symmetry, where the symmetry elements contain a rotational axis as well as a translational element along the direction of the rotational axis. These are often found in proteins involved in the formation of long fibres, such as microtubules and actin filaments[47]. The rest (260 or 3.6%) are asymmetric and contain no rotational axes, nor helical symmetry elements. Both helical and asymmetric complexes are *non-bijective* according to the nomenclature of the recently published Periodic Table of Protein Complexes, as sequence-identical subunits are required to exist in topologically non-equivalent positions[23]. Interestingly, however, the majority of these non-bijective homomers are the result of quaternary structure assignment errors[23,48], often due to the PDB biological assembly erroneously being defined to be the same as the asymmetric unit. Overall, there were no functional terms significantly enriched in these non-bijective homomers, which lacked closed symmetry (Supplementary Figure 1).

Monomers

Finally we consider the functions related to monomers, i.e. structures made up of single chain peptides that do not assemble into homomers (Figure 1d). In our non-redundant set, 6063 proteins were identified as being monomeric. Thus, in our dataset of protein structures containing only a single type of polypeptide chain, 54.6% are homomeric and 45.4% are monomeric. The five functions that were most significantly positively enriched in this set ($P = 4.09 \times 10^{-30}$ to 4.29×10^{-48}, Fisher's exact test) were functions that involve hydrolase activity, protein metabolism and the more general terms for macromolecular modification and metabolism. Interestingly, the most significantly negatively enriched functions in the monomer set ($P = 5.60 \times 10^{-96}$ to 2.85×10^{-78},

Fisher's exact test) were related to metabolic processes involving small molecules and organic substances.

Looking into the specific hydrolase activity functional term, and calculating the enrichment of monomeric structures within each subclass (according to the enzyme classification scheme by the Enzyme Commission[49]), monomers are negatively associated with hydrolase activity on amide bonds that are not associated to peptides (Figure S2, $P = 4.51 \times 10^{-5}$, Fisher's Exact test), as well as positively enriched for glycosylase activity ($P = 1.58 \times 10^{-4}$). Overall this suggests a preferential binding to large hydrolase substrates such as peptides and oligosaccharides. Macromolecular modifications are processes that are specified by the GO consortium as a modification of a biological macromolecule, such as a polypeptide or polysaccharide, which alters its properties. Including the protein metabolism term, the top monomeric structures are all therefore clearly related to processes, components and functionalities that involve large macromolecular substrates. This makes logical sense from a structural point of view, especially comparing with complexes of higher-order symmetry, as a monomer would be more likely to be able to accommodate large substrates. The only functionality that involved binding to a large molecule that was enriched in the symmetric dimers was DNA-binding, which is related to the dimeric and often symmetric nature of DNA. In contrast, if we look through our full enrichment analysis for RNA binding, we do indeed find it ranked #13 of the positively enriched functions in the monomeric structures ($P = 4.71 \times 10^{-16}$, Fisher's Exact test).

Allostery is related to the types of intersubunit interfaces formed

The term allostery is generally used to describe a dynamic, cooperative process that involves a change at one site of a protein as a result of an event, such as binding of a substrate, at a distant site of the protein. Allostery has often been associated with symmetric complexes and symmetry is integral to the first and most widely used model of allostery, the Monod-Wyman-Changeaux (MWC) model[33]. In this model, allosteric movements are explained as taking place in a concerted manner where the allowed conformations in the protein are restricted to those where the domains that are allosterically coupled remains symmetric with respect to each other. Despite its simplicity, the MWC model has been remarkably successful at explaining experimentally observed allostery in proteins[50]. Although allostery can occur in monomeric proteins[26], it appears that symmetry provides a very simple mechanism for allosteric regulation that has been commonly utilized throughout evolution.

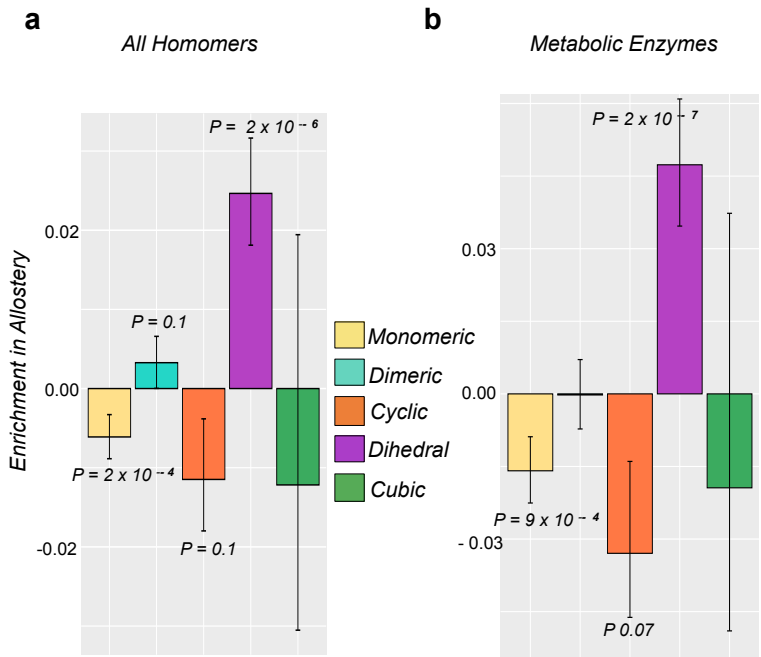


Figure 2: Enrichment of allostery in homomers from different symmetry groups. (a) Enrichment by complexes with closed symmetry in the subset of allosteric proteins**. (b)** The same enrichment profile is seen when controlling for metabolic enzymes. The enrichment is calculated as the difference between the fraction of complexes in the specific set compared to the whole set of homomers. P-values are calculated with Fisher's exact test and error bars represent 68% melded binomial confidence intervals.

Many proteins has an allosteric mechanism that regulates their function, and we decided to carry out our functional investigations with our set split up into those that are putatively allosteric and those that are not. Although there are no GO annotations for allostery, we used the protein complexes included in the Allosteric Database[51] as well as a keyword search to identify further putative allosteric proteins (see Methods). While this approach is not perfect, it does allow us to define a set of proteins that should be highly enriched in allostery. We then assessed whether allostery is associated with different symmetry groups, as presented in Figure 2a. Interestingly, we find that dihedral homomers are by far the most strongly enriched in allostery, with a slight enrichment also observed for the C_2 dimers. Importantly this overall enrichment profile of allostery is still evident when we control for metabolic

enzymes, as illustrated in Figure 2b. This suggests that the trend is not simply due to the fact that metabolic enzymes tend to be allosterically regulated.

In principle, any symmetric arrangement of subunits should be compatible with allostery, so there is a question as to why it is enriched in dihedral, and to a lesser extent dimeric, complexes. Interestingly, these two symmetry groups are both associated with isologous protein interfaces, whereas cyclic complexes have only heterologous interfaces. Therefore, to further investigate whether there is a relationship between isologous interfaces and allostery, we separated the dihedral homomers into tetramers (D_2 symmetry), which have exclusively isologous interfaces, and those with six or more subunits ($D_{n(n>2)}$ symmetry), which can possess a mixture of both isologous and heterologous interfaces, as illustrated in Figure 3a. Analysing at the enrichment in allostery in these two sets of dihedral homomers shows a significant enrichment in allostery for both groups compared to the rest of the homomers, but the enrichment is much stronger for the D_2 set (Figure 3b). This suggests that isologous interfaces drive the enrichment of allostery in dihedral complexes.

We can also look specifically at the dihedral complexes with six or more subunits ($D_{n(n>2)}$ symmetry). Since they can have a mixture of isologous and heterologous interfaces, we wondered whether allosteric complexes would tend to contain more isologous interfaces than non-allosteric complexes. For every $D_{n(n>2)}$ homomer, we calculated the fraction of the total amount of intersubunit interface formed that is isologous and plotted the distributions for allosteric and non-allosteric complexes in Figure 3c. Interestingly, we do observe a significant tendency for allosteric complexes to have a greater

fraction of isologous interfaces (P = 0.045, Wilcoxon rank-sum test), further supporting the idea that isologous interfaces are conducive to allostery.

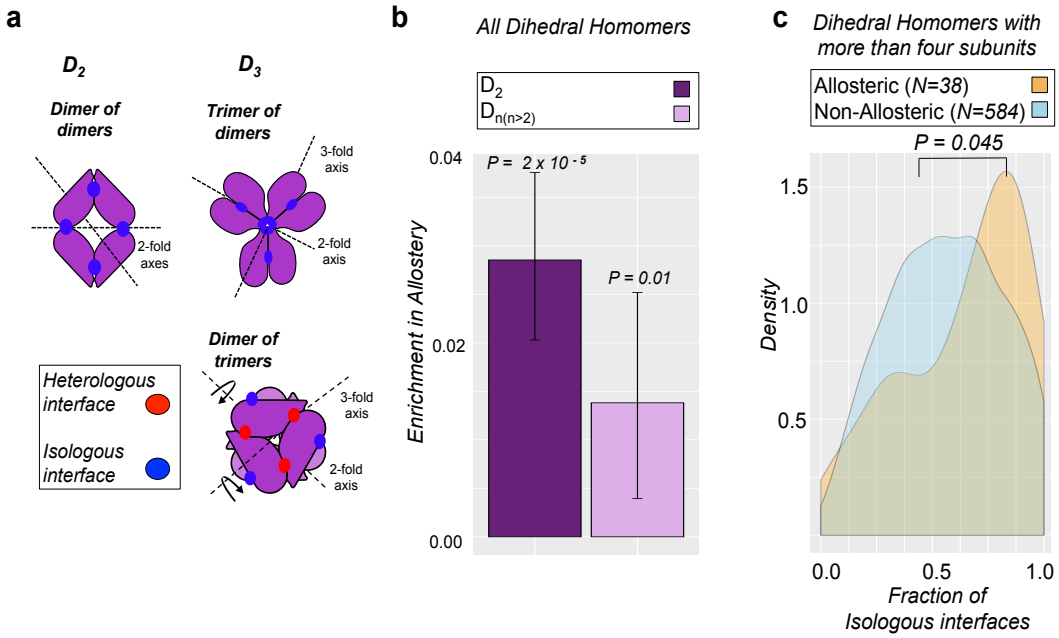


Figure 3: Allostery is associated with the present of isologous intersubunit interfaces. (a) Cartoon representation of the intersubunit interactons D_2 and D_3 complexes. D_2 complexes have two perpendicular twofold symmetry axes and therefore have only isologous interfaces, whereas D_3 complexes can have either all isologous interfaces or a mixture of isologous and heterologous interfaces. **(b)** Comparison of the enrichment in allostery between dihedral tetramers (D_2) vs. dihedral homomers with six or more subunits ($D_{n(n>2)}$). P-values are calculated with Fisher's exact test and error bars represent 68% melded binomial confidence intervals. **(c)** Density plot illustrating the distribution of isologous interfaces in the allosteric (orange) and non-allosteric (light blue) sets of dihedral complexes with more than four subunits. The P-value is from the two-sample Wilcoxon test.

Why would allostery be associated with isologous interfaces? We hypothesize that this is due to the symmetric nature of isologous, or "head-to-head" interfaces, in which two identical protein surfaces interact with each other. If a conformational perturbation occurs on one side of an isologous interface, specific to one subunit, then the symmetry of the interface will be broken, but can easily be restored by an identical change occurring on the other side of the interface. Thus, an isologous interface provides a simple physical mechanism to propagate conformational changes from one subunit to another. In contrast, for a heterologous or "head-to-tail" interface, in which two

different surfaces interact, it is more difficult to induce an identical conformational change from one subunit to another.

The apparent association between allostery and dihedral complexes was previously noted by Goodsell and Olson[6]. They describe two mechanisms by which allostery often occurs: rotation of two rings with respect to each other and pincherlike motions, that are both compatible with dihedral but not cyclic symmetry. In contrast, they suggest that allostery is difficult in cyclic complexes as it must propagate one subunit at a time around the ring. Our results here expand upon this, showing that allostery is facilitated by the nature of isologous interfaces, and that even when we consider only dihedral complexes, those with more isologous interface are more likely to be allosteric.

Discussion

By using the large amount of high quality protein structure data that are now available in combination with functional annotation data, we were able to gain essential insight into the functional specializations of homomeric protein complexes. We have shown that protein subunits tend to organize themselves into homomeric assemblies with clear functional preference depending on their overall symmetry, which can be mostly explained by simple geometrical arguments. These functional benefits are likely to have contributed towards positive evolutionary selection for protein homomerization. However, it is important to emphasize that these represent overall trends. While our results clearly support the idea that homomer quaternary structure can be functionally

beneficial, they do not exclude the likelihood that differences in quaternary structure may often be non-adaptive[28,29].

Finally, this work also shows that the formation of isologous interfaces between subunits of the protein complex can provide specific functional advantages. Essentially, these types of interfaces provide a structural mechanism for transmitting cooperative conformational changes across multiple subunits of the complex, which can contribute to allosteric mechanisms. These results provide an added degree of information that can have future practical advances when it comes to the prediction of both function and malfunction of protein complexes.

Methods

Structural data

Starting with the full set of protein crystal structures in the PDB on 2015-03-19, we selected all of those containing only a single type of polypeptide chain containing at least 30 residues, *i.e.* all of the monomers and homomers. Viral structures and structures with known quaternary structure misassignments[48] were excluded. In order to compile a non-redundant set of structures the chains were clustered into groups according to their sequence similarities and selecting one representative of each cluster. Complexes were filtered for redundancy at the level of 50% sequence identity, which resulted in a set of 13353 non-redundant structures of either monomers or homomeric complexes with functional assignments. In order to evaluate similarities in domain structure, rather than sequence, clustering was also done according to SCOP family assignment. This assured that any functional enrichment was more

likely to be related to complex symmetry than overrepresentation by functional domain. The SCOP domain analysis resulted in analogous functionalities as the sequence analysis, and is presented in Supplementary Table 5 and discussed in the Supplementary Discussion. Symmetry assignments were taken directly from the PDB. Interface symmetry was calculated from the dihedral homomer structures as was done previously[52]. All complexes are provided in Supplementary Table 2.

Functional data

Functional assignments were taken from the Gene Ontology Annotation (GOA) data set from the UniProt-GOA database (2014-09-27 release), which provides GO functional annotations to PDB chains[31]. In this dataset, a GO term is associated to a chain that maps with at least 90% identity to a UniProt knowledgebase entry. Calculating GO enrichment only within the structure annotated GO term set avoids bias against type of functional terms that are enriched in the PDB. This set was then extended by us to include all associated GO terms (terms that include the *is_a* and *part_of* association in the ontology) from the core ontology database, as curated by the Gene Ontology Consortium. Each GO term was then analysed with respect to the non-viral structures in our set, and the relative ratio of enrichment of protein symmetry groups within each functional term of our set (which were those functions associated with structures in the PDB, rather than the full ontology) was evaluated and tested for significance using Fisher's exact test. Because there are many very similar GO terms, we filtered the GO terms for redundancy, removing terms that share an association with >50% of the same proteins from our structural set. All GO terms, including those we classified as

redundant, are provided in Supplementary Table 1. In order to avoid any bias in our enrichment set caused by there being a larger percentage of monomeric structures, the above analysis was also carried out without monomeric proteins. The general trend in enriched functional terms as discussed in this analysis still holds in the set without monomers, as presented in Supplementary Table 6. Confidence intervals, at 68%, for the enrichments were calculated using a melded binomial procedure, known to provide a good match for P-values of a Fisher's exact test[53]. A Bonferroni correction was used to counteract multiple comparison errors, which resulted in a corrected significance level of 2.1×10^{-6}.

Identification of allosteric proteins

To identify putatively allosteric proteins, we searched PubMed for the terms "allostery", "allosteric" and "allosterism", and then identified PDB structures associated with these abstracts through the NCBI Structure site (http://www.ncbi.nlm.nih.gov/structure). We also accessed all the proteins associated with allostery in the Allosteric Database[51]. The combined set of allosteric proteins was filtered for redundancy and is provided in Supplementary Table 3.

Acknowlegements

We thank György Abrusán and Jonathan Wells for helpful comments on the manuscript. J.M. is supported by a Medical Research Council Career Development Award (MR/M02122X/1).

Additional Information

Author contribution

T.B wrote the raw manuscript text and prepared figures. Both J.M and T.B reviewed and altered the manuscript during its preparation.

Competing financial interests

The authors declare no competing financial interests

References

1. Todd, A. E., Orengo, C. A. & Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective1. *J. Mol. Biol.* **307,** 1113–1143 (2001).
2. Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C. & Teichmann, S. A. Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.* **14,** 208–216 (2004).
3. Tompa, P., Davey, N. E., Gibson, T. J. & Babu, M. M. A Million Peptide Motifs for the Molecular Biologist. *Mol. Cell* **55,** 161–169 (2014).
4. Ellis, R. J. Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Curr. Opin. Struct. Biol.* **11,** 114–119 (2001).
5. Foffi, G., Pastore, A., Piazza, F. & Temussi, P. A. Macromolecular crowding: chemistry and physics meet biology (Ascona, Switzerland, 10–14 June 2012). *Phys. Biol.* **10,** 040301 (2013).
6. Goodsell, D. S. & Olson, A. J. Structural Symmetry and Protein Function. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 105–153 (2000).

7. Janin, J., Bahadur, R. P. & Chakrabarti, P. Protein-protein interaction and quaternary structure. *Q. Rev. Biophys.* **41,** 133–180 (2008).
8. Levy, E. D. & Teichmann, S. Structural, evolutionary, and assembly principles of protein oligomerization. *Prog. Mol. Biol. Transl. Sci.* **117,** 25–51 (2013).
9. Marsh, J. A. & Teichmann, S. A. Structure, Dynamics, Assembly, and Evolution of Protein Complexes. *Annu. Rev. Biochem.* **84,** 551–575 (2015).
10. Marsh, J. A., Rees, H A, Ahnert, S E & Teichmann, S A. Structural and evolutionary versatility in protein complexes with uneven stoichiometry. *Nat. Commun.* **6,** 6394 (2015).
11. Marianayagam, N. J., Sunde, M. & Matthews, J. M. The power of two: protein dimerization in biology. *Trends Biochem. Sci.* **29,** 618–625 (2004).
12. Perica, T. *et al.* The emergence of protein complexes: quaternary structure, dynamics and allostery. *Biochem Soc Trans* **40,** 475–491 (2012).
13. Forrest, L. R. Structural Symmetry in Membrane Proteins. *Annu. Rev. Biophys.* **44,** 311–337 (2015).
14. Ali, M. H. & Imperiali, B. Protein oligomerization: How and why. *Bioorg. Med. Chem.* **13,** 5013–5020 (2005).
15. Gross, D. J. The role of symmetry in fundamental physics. *Proc. Natl. Acad. Sci.* **93,** 14256–14259 (1996).
16. Wolynes, P. G. Symmetry and the energy landscapes of biomolecules. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 14249–14255 (1996).
17. Blundell, T. L. & Srinivasan, N. Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proc. Natl. Acad. Sci.* **93,** 14243–14248 (1996).

18. André, I., Strauss, C. E. M., Kaplan, D. B., Bradley, P. & Baker, D. Emergence of symmetry in homooligomeric biological assemblies. *Proc. Natl. Acad. Sci.* **105,** 16148–16152 (2008).
19. Plaxco, K. W. & Gross, M. Protein Complexes: The Evolution of Symmetry. *Curr. Biol.* **19,** R25–R26 (2009).
20. Fraser, N. J. *et al.* Evolution of Protein Quaternary Structure in Response to Selective Pressure for Increased Thermostability. *J. Mol. Biol.* **428,** 2359–2371 (2016).
21. Levy, E. D., Erba, E. B., Robinson, C. V. & Teichmann, S. A. Assembly reflects evolution of protein complexes. *Nature* **453,** 1262–1265 (2008).
22. Marsh, J. A. *et al.* Protein Complexes Are under Evolutionary Selection to Assemble via Ordered Pathways. *Cell* **153,** 461–470 (2013).
23. Ahnert, S. E., Marsh, J. A., Hernández, H., Robinson, C. V. & Teichmann, S. A. Principles of assembly reveal a periodic table of protein complexes. *Science* **350,** aaa2245 (2015).
24. Wells, J. N., Bergendahl, L. T. & Marsh, J. A. Operon Gene Order Is Optimized for Ordered Protein Complex Assembly. *Cell Rep.* **14,** 679–685 (2016).
25. Dobson, R. C. J., Valegård, K. & Gerrard, J. A. The Crystal Structure of Three Site-directed Mutants of Escherichia coli Dihydrodipicolinate Synthase: Further Evidence for a Catalytic Triad. *J. Mol. Biol.* **338,** 329–339 (2004).
26. Hilser, V. J., Wrabl, J. O. & Motlagh, H. N. Structural and Energetic Basis of Allostery. *Annu. Rev. Biophys.* **41,** 585–609 (2012).
27. PERUTZ, M. F. Proteins and nucleic acids-structure and function. x+212 pp. (1962).

28. Lynch, M. The evolution of multimeric protein assemblages. *Mol. Biol. Evol.* **29,** 1353–1366 (2012).
29. Lynch, M. Evolutionary diversification of the multimeric states of proteins. *Proc. Natl. Acad. Sci.* **110,** E2821–E2828 (2013).
30. Jacobs, W. M., Knowles, T. P. J. & Frenkel, D. Oligomers of Heat-Shock Proteins: Structures That Don't Imply Function. *PLOS Comput Biol* **12,** e1004756 (2016).
31. Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. 3D Complex: A Structural Classification of Protein Complexes. *PLoS Comput Biol* **2,** e155 (2006).
32. Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12,** 88–118 (1965).
33. Aranda, A. & Pascual, A. Nuclear Hormone Receptors and Gene Expression. *Physiol. Rev.* **81,** 1269–1304 (2001).
34. Latchman, D. S. Transcription factors: An overview. *Int. J. Biochem. Cell Biol.* **29,** 1305–1312 (1997).
35. Jaeger, A. M., Pemble Iv, C. W., Sistonen, L. & Thiele, D. J. Structures of HSF2 reveal mechanisms for differential regulation of human heat-shock factors. *Nat. Struct. Mol. Biol.* **23,** 147–154 (2016).
36. Wells, J. N., Bergendahl, L. T. & Marsh, J. A. Co-translational assembly of protein complexes. *Biochem. Soc. Trans.* **43,** 1221–1226 (2015).
37. Fan, G. *et al.* Gating machinery of InsP3R channels revealed by electron cryomicroscopy. *Nature* **527,** 336–341 (2015).
38. Wente, S. R. & Schachman, H. K. Shared active sites in oligomeric enzymes: model studies with defective mutants of aspartate transcarbamoylase

produced by site-directed mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **84,** 31–35 (1987).

39. Metzler, D. E. & Metzler, C. M. *Biochemistry: The Chemical Reactions of Living Cells*. (Academic Press, 2001).
40. Hollaway, M. R., Johnson, A. W., Lappert, M. F. & Wallis, O. C. The Number of Functional Active Sites per Molecule of the Adenosylcobalamin-Dependent Enzyme, Ethanolamine Ammonia-lyase, as Determined by a Kinetic Method. *Eur. J. Biochem.* **111,** 177–188 (1980).
41. Stadtman, E. R. in *Advances in Enzymology and Related Areas of Molecular Biology* (ed. Nord, F. F.) 41–154 (John Wiley & Sons, Inc., 1966).
42. Fairman, J. W. *et al.* Structural basis for allosteric regulation of human ribonucleotide reductase by nucleotide-induced oligomerization. *Nat. Struct. Mol. Biol.* **18,** 316–322 (2011).
43. Harrison, P. M., Hoy, T. G., Macara, I. G. & Hoare, R. J. Ferritin iron uptake and release. Structure–function relationships. *Biochem. J.* **143,** 445–451 (1974).
44. Rucker, P., Torti, F. M. & Torti, S. V. Recombinant ferritin: modulation of subunit stoichiometry in bacterial expression systems. *Protein Eng.* **10,** 967–973 (1997).
45. Arenas-Salinas, M. *et al.* The crystal structure of ferritin from Chlorobium tepidum reveals a new conformation of the 4-fold channel for this protein family. *Biochimie* **106,** 39–47 (2014).
46. Carvalho, K. *et al.* Actin polymerization or myosin contraction: two ways to build up cortical tension for symmetry breaking. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368,** 20130005 (2013).

47. Levy, E. D. PiQSi: protein quaternary structure investigation. *Struct. Lond. Engl. 1993* **15,** 1364–1367 (2007).

33. SIB Swiss Institute of Bioinformatics, *Enzyme Nomenclature Database*. Repository (2016) Available at: http://enzyme.expasy.org (Accessed: 16th Oct 2015)

49. Changeux, J.-P. Allostery and the Monod-Wyman-Changeux Model After 50 Years. *Annu. Rev. Biophys.* **41,** 103–133 (2012).

50. Shen, Q. *et al.* ASD v3.0: unraveling allosteric regulation with structural mechanisms and biological networks. *Nucleic Acids Res.* **44,** D527–D535 (2016).

51. Marsh, J. A. & Teichmann, S. A. Protein flexibility facilitates quaternary structure assembly and evolution. *PLOS Biol* **12,** e1001870 (2014).

52. Huntley, R. P. *et al.* The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Res.* **43,** D1057–1063 (2015).

53. Fay, M. P., Proschan, M. A. & Brittain, E. Combining one-sample confidence procedures for inference in the two-sample case. *Biometrics* **71,** 146–156 (2015).