Protein Flexibility Facilitates Quaternary Structure Assembly and Evolution

Joseph A. Marsh[1¤*], **Sarah A. Teichmann**[1,2]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, [2] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

Abstract

The intrinsic flexibility of proteins allows them to undergo large conformational fluctuations in solution or upon interaction with other molecules. Proteins also commonly assemble into complexes with diverse quaternary structure arrangements. Here we investigate how the flexibility of individual protein chains influences the assembly and evolution of protein complexes. We find that flexibility appears to be particularly conducive to the formation of heterologous (i.e., asymmetric) intersubunit interfaces. This leads to a strong association between subunit flexibility and homomeric complexes with cyclic and asymmetric quaternary structure topologies. Similarly, we also observe that the more nonhomologous subunits that assemble together within a complex, the more flexible those subunits tend to be. Importantly, these findings suggest that subunit flexibility should be closely related to the evolutionary history of a complex. We confirm this by showing that evolutionarily more recent subunits are generally more flexible than evolutionarily older subunits. Finally, we investigate the very different explorations of quaternary structure space that have occurred in different evolutionary lineages. In particular, the increased flexibility of eukaryotic proteins appears to enable the assembly of heteromeric complexes with more unique components.

Citation: Marsh JA, Teichmann SA (2014) Protein Flexibility Facilitates Quaternary Structure Assembly and Evolution. PLoS Biol 12(5): e1001870. doi:10.1371/journal.pbio.1001870

Academic Editor: Gregory A. Petsko, Brandeis University, United States of America

Received November 25, 2013; **Accepted** April 17, 2014; **Published** May 27, 2014

Funding: J.A.M. was supported by a Long-term Fellowship from the Human Frontier Science Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: A_{rel}, relative solvent accessible surface area.

* E-mail: joseph.marsh@igmm.ed.ac.uk

¤ Current address: MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, United Kingdom.

Introduction

The assembly of proteins into protein complexes is ubiquitous within the cell [1–3]. This provides many potential benefits, such as allosteric regulation, co-localization of distinct biological functions, and protection from aggregation or degradation [4–6]. Alternatively, protein oligomerization may in some cases result from random mutations combined with neutral drift [7]. The individual polypeptide constituents of a protein complex—that is, the subunits—can be assembled into a wide variety of symmetric and asymmetric quaternary structure topologies [8–11]. Recent work has demonstrated the biological importance of the assembly process by showing that many protein complexes assemble via ordered pathways that have a strong tendency to be evolutionarily conserved [12,13].

The intrinsic flexibility of proteins is intimately related to their assembly into complexes. For example, flexibility is often crucial for binding—either for facilitating the structural changes that are induced upon binding or for allowing the intrinsic fluctuations within the unbound state that enable a conformational selection binding mechanism [14]. The flexibility of the unbound state also generally correlates with the magnitude of binding-induced conformational changes [15,16]. However, although the role of flexibility in simple binary interactions is becoming quite well understood, there has been little investigation into how subunit flexibility relates to the diversity of observed quaternary structure topologies. How does flexibility facilitate the assembly of multiple proteins into a protein complex? And given that quaternary structures can evolve in a process analogous to assembly [12,13,17], has flexibility been important for this evolution?

The structures of a huge number of protein complexes are now available. Although many structure-based methods are available for characterizing protein flexibility and dynamics, we are primarily interested in the intrinsic flexibility of monomers before they assemble into a complex. Because there are no unbound-state structures available for most individual proteins observed as subunits of protein complexes, it has previously been difficult to characterize their flexibility. Algorithms for predicting intrinsic disorder from protein sequences can provide some useful information, and have revealed a significant tendency for the subunits of large multiprotein complexes to be disordered in isolation [18–20].

We recently introduced a simple method for predicting the intrinsic flexibility of proteins from their structures. This method relies on the fact that the folding of a protein from its unfolded state is driven primarily by the intramolecular burial of surface area [21]. Proteins that bury less surface area within their folds will tend to retain more conformational entropy and be more flexible

Author Summary

Proteins often interact with other proteins and assemble into complexes. Here we show that the flexibility of individual proteins is important for their recruitment to complexes, as it facilitates the formation of asymmetric interfaces between different subunits. The role of flexibility becomes increasingly important as a greater number of distinct proteins are packed together within a single complex: the more distinct subunits, the more flexible those subunits need to be. A consequence of this is that, when a protein complex gains a new subunit during evolution, the newer subunit will tend to be more flexible than the older subunits. This suggests that we may be able to partially reconstruct the evolutionary history of a protein complex by considering the flexibility of its subunits. We also find that the types of protein complexes an organism forms are closely related to the flexibility of its proteins, with eukaryotic species, and particularly animals, using their increased flexibility to assemble complexes involving more distinct components.

[22]. Therefore, a simple proxy for surface-area burial, the relative solvent-accessible surface area (A_{rel}), is highly predictive of various flexibility measures, including those calculated from protein structures and those derived directly from experimental measurements [22]. In fact, the correlation between A_{rel} and independent measures of flexibility is as strong or stronger than the correlation of those different flexibility measures with each other. A_{rel} also shows a strong correspondence with the extent of conformational changes that occur upon complex assembly [16] or disassembly [23].

A_{rel} is a simple ratio describing how much solvent-accessible surface area a protein is exposing compared to what we expect for a typical folded, monomeric, crystallizable protein of the same molecular weight. Roughly speaking, A_{rel} values of 0.8–0.9 are observed for the most compact, rigid proteins, whereas A_{rel} values greater than 1.2 are seen for highly flexible proteins that undergo large conformational changes upon binding [16].

Although A_{rel} involves major simplifications, it is important to emphasize that its use as a measure of flexibility arises from fundamental energetic principles—it is not merely a probe of globularity. In fact, some proteins are highly efficient at burying enough intramolecular surface area to become quite rigid, while retaining fairly extended overall conformations. As discussed previously, by assuming constant energy per unit of surface area buried, A_{rel} can be directly related to the difference in conformational entropy with respect to an idealized folded state [22]. Furthermore, its remarkable agreement with various computational and experimental flexibility measures strongly supports its utility for large-scale analyses.

There is another major benefit for our purposes here: when A_{rel} is calculated for the bound subunits of protein complexes (i.e., by considering the subunits in isolation, ignoring any interfacial contacts), there is a very strong correlation between the A_{rel} values of bound subunits and those same proteins in their unbound states [16]. This is illustrated here in Figure S1A. Crucially, this means that the conformation of a protein subunit in its bound state can be used to predict its flexibility in its unbound, monomeric state.

The highly flexible proteins identified with this method also show some correspondence with intrinsic disorder: protein subunits predicted to be disordered in isolation tend to have substantially higher A_{rel} values [16,24]. Furthermore, although the overall sequence determinants of intrinsic disorder are quite different from A_{rel} [22], there is still a significant correspondence between the A_{rel} values of bound subunits and the fraction of residues predicted to be disordered (Figure S1B). In essence, it appears that A_{rel} is able to capture the entire spectrum of protein flexibility associated with binding, of which intrinsic disorder represents one extreme end [25].

It should be noted that, with an approach like this, it can be difficult to distinguish between scenarios where flexibility itself is required for assembly, as opposed to flexibility being a consequence of the structural requirements of a protein complex. For example, proteins that form larger intersubunit interfaces have less surface area available to bury intramolecularly, and are therefore likely to be more flexible in isolation. Similarly, proteins with more elongated shapes will generally be more flexible, and so it may not be possible to differentiate a conformational necessity for elongated shapes within the complex from a requirement for intrinsic subunit flexibility.

In this study, we have used A_{rel} to quantitatively investigate the relationships between intrinsic subunit flexibility and the structure, assembly, and evolution of protein complexes. We find that subunit flexibility is strongly associated with the formation of heterologous interfaces required for the assembly of asymmetric, cyclic, and heteromeric complexes. This has major implications for understanding the evolution of protein complexes, as it means that subunit flexibility is often reflective of their evolutionary histories. Moreover, this relationship between flexibility and assembly is also manifested in the very different evolutionary explorations of quaternary structure space observed for prokaryotes and eukaryotes.

Results and Discussion

Cyclic and Asymmetric Homomers Are Associated with Increased Subunit Flexibility

We first consider simple homomeric complexes, which are comprised of just a single type of self-interacting subunit. To investigate the relationship between flexibility and symmetry, we group the homomers into the following major classes:

(1) Twofold dimeric complexes, represented by the C_2 symmetry group, are characterized by a single twofold axis of rotational symmetry, which results in an isologous (i.e., symmetric or head-to-head) interface between the two subunits. Such isologous interfaces are extremely common, which has been suggested to be due to their inherent energetic favourability [26,27].

(2) Cyclic complexes, represented by the $C_{\mathrm{n}\ (\mathrm{n}>2)}$ symmetry groups, possesses higher order rotational symmetry, with the subunits forming closed rings via heterologous (i.e., asymmetric or head-to-tail) interfaces. Note that although the C_2 complexes do have twofold rotational symmetry, here we will only refer to complexes with at least threefold symmetry as cyclic, due to their distinct interface properties.

(3) Dihedral complexes, represented by the $D_{\mathrm{n}\ (\mathrm{n}>1)}$ symmetry groups, can be thought of as a doubling of the other topologies through the addition of a new twofold rotational axis (e.g., dimerization of C_3 gives D_3). All dihedral complexes therefore have isologous interfaces corresponding to this twofold axis. Dihedral complexes with at least six subunits usually (but not always) have a mixture of both isologous and heterologous interfaces. Dihedral complexes appear to be particularly conducive to facilitating allosteric regulation, as the isologous interfaces associated with the twofold axis provide a simple way to transmit conformational changes between subunits [9].

(4) Asymmetric complexes, represented by the trivial symmetry group C_1, can be formed in various ways but are characterized by

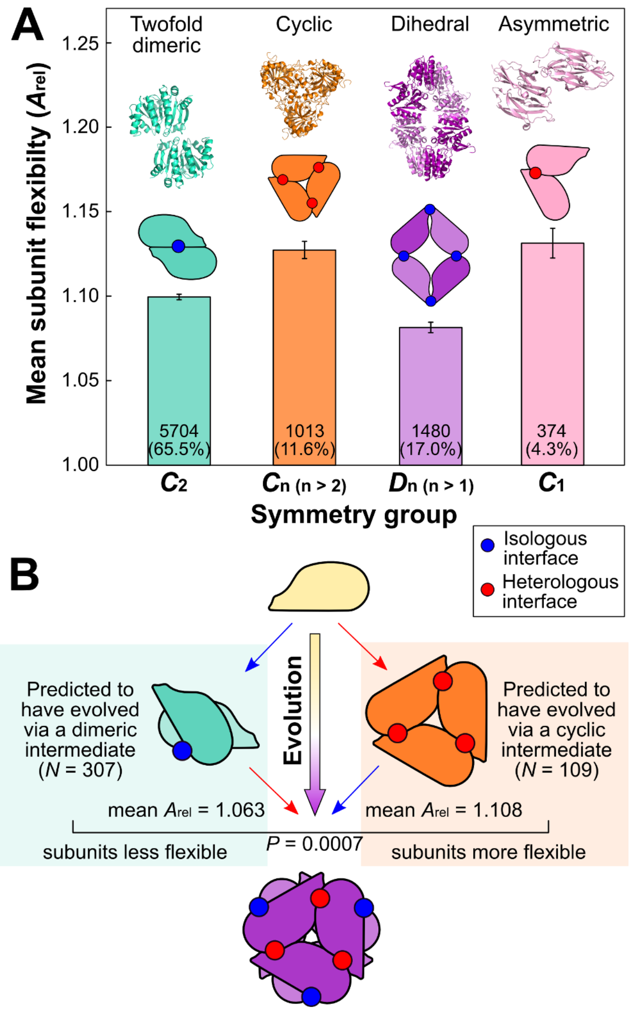

Figure 1. Relating the flexibility of homomeric subunits to quaternary structure topology and evolution. (A) Comparison of subunit flexibility, as measured by A_{rel}, for homomers from different symmetry groups. An example from each symmetry group is shown above. The numbers and percentages of each group within the total set of homomeric complexes are shown on the bars. These groups comprise all complexes in the PDB except the rare cubic (0.9%) and helical (0.6%) symmetry groups. Error bars represent SEM. Boxplots for each group along with the p values between groups are provided in Figure S2A. (B) There are two possible evolutionary pathways for a dihedral hexamer (D_3): via a twofold dimer (C_2) intermediate (left) or via a cyclic (C_3) intermediate (right). When considering all such complexes where two different evolutionary pathways are possible, we observe a strong tendency for those that evolved via a cyclic intermediate to have more flexible subunits. Interestingly, the subunits of complexes with predicted dimeric intermediates are less flexible than those from twofold dimeric complexes ($A_{rel}=1.063$ versus 1.099, $p=5\times10^{-7}$, Wilcoxon rank-sum test) and those from complexes with predicted cyclic intermediates are less flexible (but not significantly so) than those from cyclic complexes ($A_{rel}=1.108$ versus 1.127, $p=0.7$). One potential explanation for this is that lower subunit flexibility might be associated with a greater propensity for evolving higher order quaternary structures via dimeric or cyclic intermediates.
doi:10.1371/journal.pbio.1001870.g001

the existence of different subunits in nonequivalent positions (e.g., the asymmetric dimer shown in Figure 1A in which a heterologous interface involving two distinct surfaces is formed).

In Figure 1A, we compare the mean flexibilities, as measured by A_{rel}, of homomeric subunits from these different groups. Most strikingly, we observe a highly significant tendency for the subunits of cyclic and asymmetric complexes to be more flexible than those forming twofold dimeric or dihedral topologies. Much weaker trends are observed if sequence-based intrinsic disorder predictions are used instead of A_{rel} (Figure S3A). Furthermore, when we group the homomers from different symmetry classes by total number of subunits, we observe very little correspondence with subunit flexibility (Figure S4).

What is the origin of this relationship between flexibility and symmetry? A possible explanation is that both cyclic and asymmetric complexes are associated with heterologous intersubunit interfaces involving two distinct surfaces. When forming an asymmetric, heterologous interface, it is easy to imagine how flexibility could be highly beneficial, as it allows for conformational changes of one surface with respect to the other, thus enabling tight intersubunit packing.

In contrast, twofold dimeric and dihedral homomers form isologous interfaces involving self-complementary surfaces. A basic property of an isologous interface is that any conformational change that occurs on one side of the interface must also occur on the other, in order to preserve interface symmetry. This general requirement for equivalent conformational changes on both halves of an isologous interface is likely to make intrinsic flexibility much less advantageous. Therefore, we hypothesize that a major role of subunit flexibility is to facilitate the conformational changes required for heterologous interfaces.

Increased flexibility and conformational changes upon binding are also known to be associated with larger interfaces [16,28,29]. This concept is especially intuitive when using A_{rel} as a measure of flexibility, as flexible proteins that bury less intramolecular surface area will have more surface available to participate in intermolecular interactions. Thus, one might hypothesize that the increased flexibility associated with asymmetric and cyclic quaternary structures could arise from a requirement for larger interfaces. However, we show in Figure S5 that the symmetry groups associated with increased subunit flexibility do not show a similar association with larger interfaces.

Previously we noted that flexibility shows a significant correspondence with secondary structure: α proteins tend to be more flexible than β proteins [22]. Therefore, in Table S1 we demonstrate that the trends observed here are consistent across different secondary structure classes.

Subunit Flexibility Reflects the Evolutionary Histories of Homomeric Complexes

The diverse quaternary structures observed in nature are not independent of each other: homomers can evolve from one topology to another [7,12,30]. Previously it has been shown that the relative sizes of a homomer's interfaces can be used to predict its evolutionary history, as the largest interface will nearly always have formed first [12,31]. This means there are multiple possible evolutionary pathways when considering certain quaternary structure topologies. For instance, although all cyclic complexes have exclusively heterologous interfaces and all dihedral complexes have some isologous interfaces, dihedral complexes with at least six subunits can simultaneously have both isologous and heterologous interfaces. In cases where the isologous interfaces are the largest in the complex, the complex will be predicted to have evolved via a dimeric intermediate (Figure 1B, left pathway). On the other hand, if a heterologous interface is the largest, the complex will almost certainly have evolved via a cyclic intermediate (Figure 1B, right pathway).

We considered those homomers with both isologous and heterologous interfaces that therefore have at least two possible evolutionary pathways. These were split into those predicted to have evolved via either twofold dimeric (C_2) or cyclic ($C_{n\ (n>2)}$) intermediates. Interestingly, complexes with dimeric intermediates are nearly three times as abundant as those with cyclic intermediates, consistent with the finding that isologous interfaces are generally more ancient [31,32], and therefore would be expected to be larger.

We also observe a significant tendency for subunits that assemble via cyclic intermediates to be more flexible than those that assemble via dimeric intermediates (mean $A_{rel} = 1.108$ versus 1.063, $p = 0.0007$, Wilcoxon rank-sum test). In other words, those complexes in which a heterologous interface is the largest will tend to have more flexible subunits, further demonstrating the relationship between subunit flexibility and heterologous interface formation. This also reveals a fascinating connection between subunit flexibility and evolutionary history: just as the evolution of a complex is related to the sizes of its interfaces, it is also reflected in the flexibility of its subunits.

Finally, it is interesting to specifically consider those dihedral complexes predicted to have evolved via dimeric intermediates. If we consider each dimeric precursor together as an individual "subunit," we can calculate an A_{rel} value for the dimer, just as we would for an individual subunit. Given that increased flexibility of individual subunits is associated with assembly into cyclic complexes, we might expect the dimeric precursors of $D_{n\ (n>2)}$ complexes (e.g., trimers or tetramers of dimers) to have higher A_{rel} values than those from D_2 (i.e., dimer of dimers) complexes. However, the A_{rel} values from the two groups of dimeric precursors are nearly identical (1.086 for $D_{n\ (n>2)}$, 1.088 for D_2, $p = 0.5$, Wilcoxon rank-sum test), suggesting that flexibility at the level of dimeric subcomplexes is not as closely related to quaternary structure as is monomer flexibility.

Flexibility Enables Packing of Distinct Heteromeric Subunits

Although homomeric interfaces between identical chains can either be isologous or heterologous, heteromeric interactions between dissimilar subunits are inherently heterologous. Therefore, just as flexibility appears to facilitate the packing of heterologous homomeric interfaces, flexibility might also promote the formation of heterologous interfaces in heteromers.

To address this, we group protein complexes by their total number of nonhomologous subunits and plot the mean subunit flexibility as measured by A_{rel} (Figure 2). In this figure, homomers and homologous heteromers (i.e., heteromers where all the distinct chains are homologous) are represented by a single column (blue), whereas other heteromers can have varying numbers of nonhomologous subunits. There is a very striking association between subunit flexibility and an increasing number of nonhomologous subunits per complex, thus confirming the importance of flexibility in heteromer assembly.

Despite this strong trend, it should be noted that not all subunits of large multiprotein complexes are highly flexible. Although flexibility appears to be important for assembling multiple subunits of different shapes within a single complex, not all subunits need be flexible to achieve this packing. For instance, of those heteromers with four nonhomologous subunits, 13/19 have at least one subunit with A_{rel}<1.1.

Previously, it was noted that protein complexes with more distinct components tend to be enriched in intrinsic disorder [19]. Here, although we observe a slight tendency for predicted disorder to increase in heteromeric complexes (Figure S3C), the trend is much stronger with A_{rel}. This further suggests that a range of protein flexibility, of which intrinsic disorder forms part, is important for assembly.

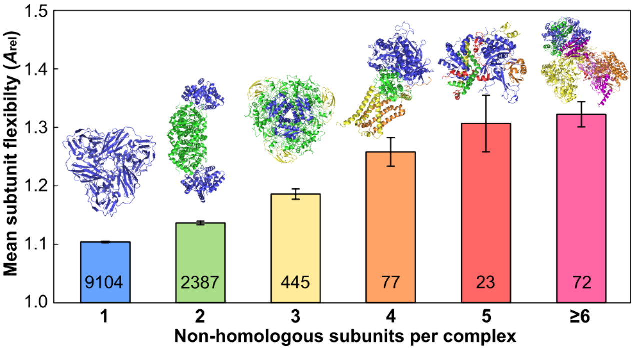

Figure 2. Comparison of subunit flexibility from protein complexes with varying numbers of nonhomologous subunits. Examples of complexes with varying numbers of nonhomologous subunits are shown above. The numbers of unique chains in each group are shown on the bars. Error bars represent SEM. Boxplots for each group are provided in Figure S2B.
doi:10.1371/journal.pbio.1001870.g002

Flexibility Facilitates the Evolution of New Heteromeric Subunits

The above results have major implications for our understanding of quaternary structure evolution. If we consider a simple scenario in which a heteromer evolves in a sequential manner, gaining a new subunit with each step, then the simplest way to account for this would be if the newly added subunits are more flexible than those from the ancestral complex. This is illustrated in Figure 3A. A similar model was anticipated by Hegyi et al., who suggested that the propensity for intrinsic disorder should be greater in evolutionarily more recent subunits due to the increased disorder propensity in complexes with many subunits [19].

Do the evolutionarily more recent subunits of protein complexes have a significant tendency to be more flexible? To test this, we employed a comparative genomic approach in an attempt to partially reconstruct the evolutionary histories of human heteromers. If an ortholog of a human gene encoding a protein subunit is present in the genome of a given species, then we can assume that that protein was present in the last common ancestor with humans. Of course, the presence of orthologs in an ancestral species does not necessarily mean they interacted [33–38]. However, when orthologs of different subunits of the same human complex are present in yeast, the vast majority also form a complex in yeast [39]. Therefore, using the orthologs present in different species taken from the Ensembl Compara [40] and OMA [41] databases, we can say with strong confidence that certain subunits of protein complexes are highly likely to have been present in an ancestral species.

Although we can identify the presence of some subunits in ancestral species with relative simplicity, it is much more difficult to conclusively show that a given subunit was not present, even if no ortholog is detected. For example, the identification of orthologs can be complicated by genome annotation errors or fast evolutionary divergence rates. Moreover, genes can be lost in evolution, so the absence of a gene does not mean that it was not present in an ancestral species. To compensate for these complications, we employed an extremely conservative approach to the identification of subunits that were likely absent in an ancestral species. For each human subunit, we identified the evolutionarily most divergent species in which it might possibly

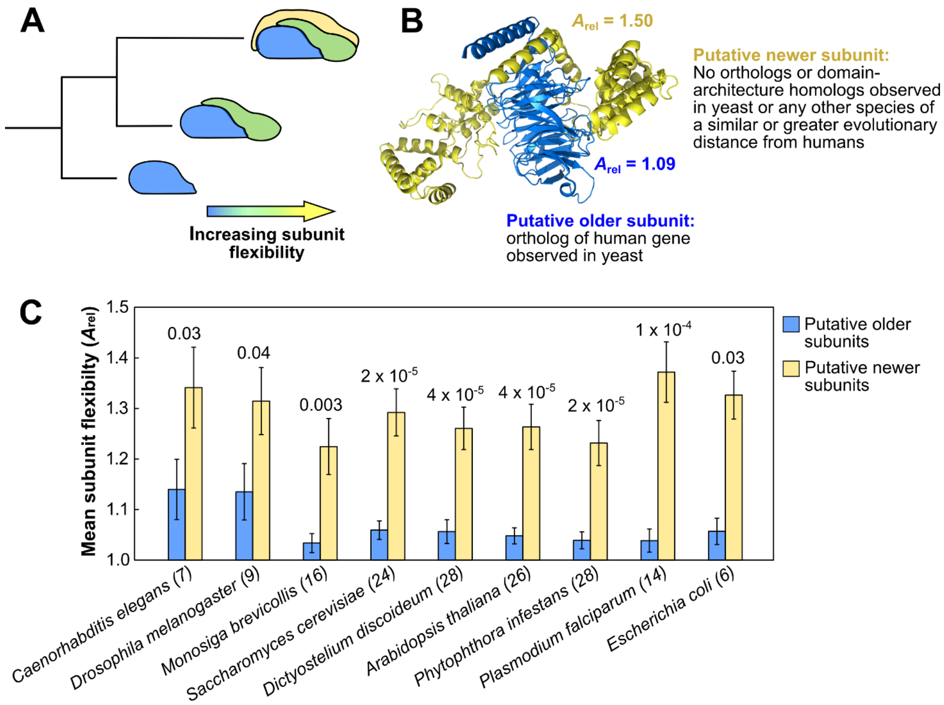

Figure 3. The importance of protein flexibility for the evolution of new heteromeric subunits. (A) Model for the evolution of heteromeric complexes in which new subunits of increasing flexibility are sequentially gained. (B) Example of a protein complex (Gβ5-RGS9, PDB ID: 2PBI) in which different relative ages can be assigned to different subunits. There is an ortholog of Gβ5 (blue) in yeast, whereas no orthologs or domain-architecture homologs of RGS9 (yellow) can be detected in yeast or any other species of a similar or greater evolutionary distance from humans (the most distant ortholog is observed in *Caenorhabditis elegans* and the most distant homolog sharing the same domain architecture is seen in the single-celled eukaryote *Capsaspora owczarzaki*, which is more closely related to humans than yeast). (C) Pairwise comparisons of the flexibility of putative older and putative newer subunits of human (or closely related) protein complexes, with respect to different species. No species more closely related to humans than *C. elegans* and *Drosophila melanogaster* are shown as there are none where >5 complexes with putative older and newer subunits can be identified. The full set of species considered is provided in Table S2. The *p* values calculated with the Wilcoxon signed-rank test are shown for each species, and the numbers of complexes from each species are shown in parentheses. Error bars represent SEM.
doi:10.1371/journal.pbio.1001870.g003

have been present. This was done by considering not just orthologs, but also homologous proteins that share the same domain architectures. These can be of much greater sequence divergence than simple orthologs. Thus, if any ortholog or domain-architecture homolog of a human subunit is present in a given species, we presume that it might possibly (but not necessarily) have formed part of a similar complex in the last common ancestor.

Combining these two approaches, we considered each human (or closely related) protein complex from the perspective of different species of varying evolutionary relatedness to humans. Proteins for which an ortholog could be identified in a given species were considered to be the "putative older subunits." In contrast, proteins for which no ortholog or homolog could be detected in that species, or any other species of similar or greater evolutionary divergence from humans, were considered to be the "putative newer subunits." An example of a complex in which two subunits could be confidently assigned as having different evolutionary ages is shown in Figure 3B.

In Figure 3C, we compare the flexibilities of the putative older and newer subunits for several species (all species are provided in Table S2). In this analysis, only those complexes in which both older and newer subunits could be identified were considered. For nearly all species, there is a very strong tendency for the newer subunits to be more flexible than the older subunits, thus supporting our hypothesis that subunit flexibility reflects the relative evolutionary age of subunits.

We can also combine the observations made for different species into a nonredundant set of 61 complexes where both older and newer subunits can be identified. In this case, the newer subunits are also far more flexible than the older subunits ($A_{\text{rel}} = 1.213$ versus 1.082, $p = 6 \times 10^{-6}$, Wilcoxon signed-rank test). Similarly, in the large majority of complexes (48/61), the newer subunit(s) are more flexible than the older subunit(s) ($p = 8 \times 10^{-6}$, binomial test).

Although many subunits from protein complexes of known structure are truncated forms of full proteins (e.g., individual domains), a strong tendency for newer subunits to be more flexible is still observed when only full-length or nearly full-length proteins are considered ($A_{\text{rel}} = 1.245$ versus 1.115, $p = 0.007$, $N = 19$). It has also been observed that evolutionarily newer proteins are generally shorter than older proteins [42,43]. If shorter proteins tended to be more flexible, this could influence our results. However, we find that even when we consider only those cases where the putative newer subunits are longer than the older subunits, the newer subunits are still more flexible ($A_{\text{rel}} = 1.221$ versus 1.115, $p = 0.007$, $N = 24$).

An additional concern is that some fast-evolving proteins may have diverged beyond detectable homology, yet still share structural and functional similarity and possibly still interact within the same complex. If there existed a tendency for more flexible proteins to evolve at a faster rate, then more flexible proteins might simply appear to be more recent due to their lower conservation. Generally it is thought that, although the more flexible regions of a given protein tend to evolve more quickly than its more rigid regions, there is little correspondence between flexibility and evolutionary conservation at the global protein level [17]. We address this further in Figure S6, showing that there is no clear propensity for evolutionarily newer proteins to be more flexible overall (i.e., when not considered at the individual complex level), although there is a slight tendency for the most highly flexible proteins to be less conserved.

Finally, there is a completely different way by which we can assess the propensity for evolutionarily more recent subunits to be more flexible. As an alternative to the scheme in Figure 3A, we can hypothesize that existing subunits might have evolved to become more flexible in order to accommodate new, more rigid subunits. To address this, we "normalize A_{rel}" for the variation that occurs between homologous proteins that form subunits of different complexes, and for the variation that occurs between evolutionarily unrelated protein families (Figure S7). This analysis shows that very little of the trend in Figure 2 can possibly be explained by increasing flexibility of existing subunits, thus strongly supporting the scenario in Figure 3A.

Evolutionary Exploration of Quaternary Structure Space Is Related to Proteome Flexibility

The observation that subunits gained later in evolution tend to be more flexible raises interesting questions about proteome and interactome evolution. Specifically, it suggests that the average flexibility of proteins in an organism might increase over the course of evolution as new proteins are acquired and the number of protein complex interactions increases. Therefore, it is interesting to first consider how quaternary structure varies in evolution, by comparing the proportion of homomeric and heteromeric complexes in bacteria, archaea, and eukaryotes (Figure 4A). Interestingly, a far greater percentage of eukaryotic complexes in our dataset are heteromeric (29.3%), as compared to bacterial (6.4%) or archaeal (8.7%) complexes ($p<10^{-34}$, Fisher's exact test). This is consistent with the previous observation that heteromers are enriched in vertebrates relative to unicellular organisms [44]. Although gene duplications in eukaryotes are known to have resulted in many homologous heteromers [45], these still comprise only a small fraction of the total heteromers (Figure 4A). These huge differences strongly suggest that heteromeric topologies are much more frequently utilized in eukaryotes than prokaryotes. Moreover, this is compatible with the fact that eukaryotes also generally have larger genomes. The larger number of protein-coding genes therefore provides more different proteins with which to form complexes.

Next, to explore the evolutionary relationship between flexibility and quaternary structure, we grouped complexes by their species of origin and plotted the number of nonhomologous subunits per complex against the mean subunit flexibility (Figure 4B; values for all species provided in Table S3). There is a striking distinction between prokaryotes and eukaryotes: the eukaryotes tend to have more flexible subunits that form complexes with more unique components, whereas bacterial and archaeal complexes have fewer, less flexible subunits. Although there are certainly some biases in the complexes crystallized from different species, the consistency of the division between prokaryotes and eukaryotes suggests that it is reflective of real evolutionary differences.

There are two eukaryotes that cluster with the prokaryotes: the plant *Arabidopsis thaliana* and the protozoan *Plasmodium falciparum*. This is quite interesting given that these two species are the most evolutionary divergent eukaryotes, relative to the more closely related yeast and metazoans [46]. When all 174 other plant complexes (excluding *A. thaliana*) are considered together, they have more nonhomologous subunits per complex (1.172) than observed in any of the prokayotes, but very low subunit flexibility (mean A_{rel} of 1.067). From this limited evidence, it is difficult to tell whether these results reflect genuine evolutionary differences. However, this does hint that some of this divergence may have occurred in the fungi/metazoa lineage.

The eukaryotic species have a much greater spread in nonhomologous subunits per complex. *Bos taurus*, in particular, has more than any other species. A possible explanation for this is that many of these large multiprotein complexes are likely to have been natively purified from bovine tissues. Thus, the complexes tend to contain more of the biologically relevant subunits present *in vivo*, whereas complexes from other organisms are more likely to have been recombinantly produced. Interestingly, we note that *Saccharomyces cerevesiae* also has a relatively large number of nonhomologous subunits per complex, as does *Escherichia coli* when compared to other prokaryotes. These organisms are often used for protein production and so their complexes may also be more likely to have been natively purified. These results highlight the interesting (albeit probably unsurprising) point that protein complexes *in vivo* are likely to have a much greater tendency to contain more distinct subunits than has generally been observed crystallographically.

Figure 4B suggests that the increase in protein flexibility observed in eukaryotes could possibly be explained by the fact that their protein complexes have more distinct components. Therefore, we next compared the flexibility of subunits from bacteria, archaea, and eukaryotes, while controlling for the number of nonhomologous subunits (Figure 4C). Interestingly, the subunits of eukaryotic complexes still tend to be more flexible than those from bacteria. The archaeal subunits are generally intermediate in flexibility to bacteria and eukaryotes, although there are far fewer archaeal complexes in the dataset. Thus, although increased flexibility in eukaryotes is important for facilitating heteromer assembly, much of the increase in eukaryotic proteome flexibility is clearly independent of the physical requirement for packing multiple subunits within individual complexes. Similar relationships between flexibility and nonhomologous subunits are observed for individual species (Figure S8), which suggests that these results are not influenced by any strong species-level bias.

As a complement to this structure-based analysis of flexibility using A_{rel}, we also looked at the relationship between predicted intrinsic disorder and protein–protein interactions. Previous observations have shown a strong tendency for proteins with more interaction partners to possess a greater fraction of intrinsically disordered residues [47–49]. This could be considered somewhat analogous to our observation of increased flexibility in complexes with multiple distinct subunits. In Figure S9, we show that this trend is observed for the bacterial, archaeal, and eukaryotic species with the most experimentally identified protein–protein interactions. These nonstructural results are consistent with our structural analysis, emphasizing the importance of flexibility and disorder for facilitating protein interactions across evolution. They also highlight an increased level of intrinsic disorder in eukaryotes that appears to be independent of the number of interactions made.

Conclusions

In this study, we have demonstrated a close association between intrinsic subunit flexibility and the assembly of protein complexes. The origin of this is simple: because flexibility is largely controlled by how little surface area a protein buries intramolecularly [22], then the more flexible the protein, the more surface area that will be available to participate in intermolecular interactions. This is

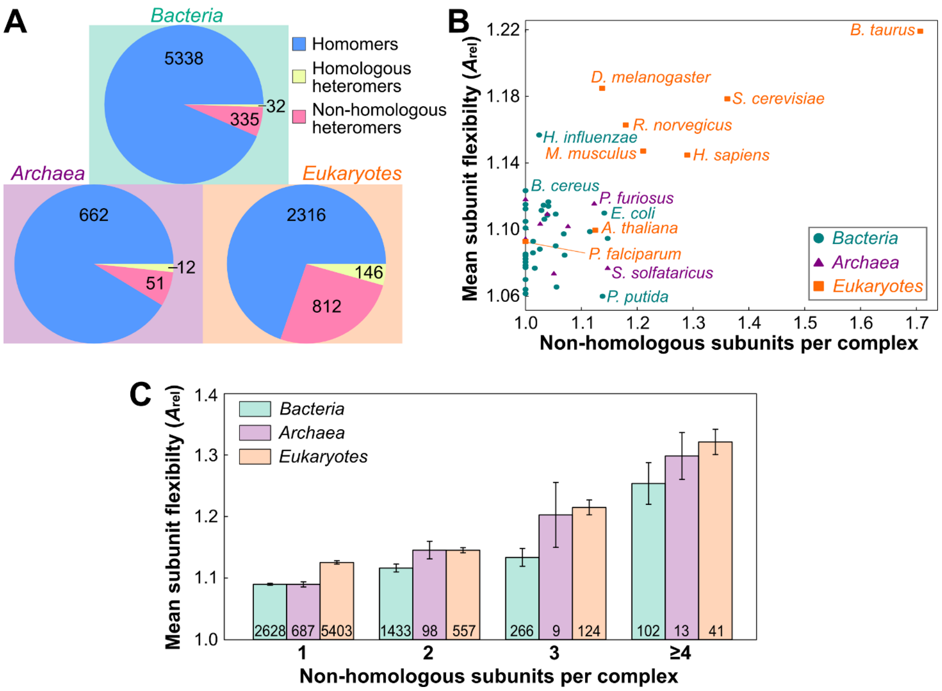

Figure 4. The relationship between evolution, quaternary structure topology, and protein flexibility. (A) Comparison of the numbers of homomers, homologous heteromers (i.e., heteromers where all distinct chains are homologs), and nonhomologous heteromers from bacteria, archaea, and eukaryotes. (B) Comparison of the mean subunit flexibility and number of nonhomologous subunits per complex for the 50 species with the most complexes in our dataset. Values for all species are provided in Table S3. (C) Comparison of subunit flexibility for protein complexes with varying numbers of nonhomologous subunits from bacteria, archaea, and eukaryotes. Error bars represent SEM. A similar species-level analysis is provided in Figure S8.
doi:10.1371/journal.pbio.1001870.g004

why increased flexibility, disorder, and conformational changes upon binding are associated with larger interfaces [16,28,29,50]. The evidence presented here suggests that flexibility is particularly conducive to the formation of heterologous interfaces, in which two distinct surfaces interact with each other. Therefore, flexibility appears to facilitate the assembly of asymmetric, cyclic, and heteromeric complexes.

This work also extends our understanding of protein evolution, as it shows how the evolutionary history of a protein complex can be directly related to the flexibility of its subunits. This suggests that flexibility could potentially be quite useful in the reconstruction of protein complex evolutionary histories. To some extent, our results suggest that the eukaryotic increase in flexibility may have been driven by the evolution of protein complexes with more components. In addition, it is possible that some of the increased flexibility in eukaryotic subunits may be reflective of a greater propensity to form multiple nonconcurrent interactions, as has been seen for intrinsic disorder [49,51,52]. However, the increase in flexibility might also be related to selection for function other than protein complex assembly, increased tolerance due to compartmentalization and chaperones, or simply genetic drift [53].

This new knowledge of the relationship between quaternary structure topology and flexibility could aid the prediction of protein complex topologies from limited information. For example, if some knowledge of intrinsic flexibility is available (based upon sequence, structure, or experiments), this could be used to help assess the relative likelihoods of different quaternary structure arrangements. Similarly, just as flexibility appears to facilitate quaternary structure evolution, it might also prove important for engineering multiprotein assemblies, if the principles of flexibility and interactions can be harnessed to enable the packing of heterologous interfaces.

In the present study, we have interpreted our results as showing that intrinsic flexibility facilitates the assembly and evolution of quaternary structure. However, it is possible that, rather than flexibility being required for assembly, it can to an extent be thought of as arising from the physical requirements of the bound state. That is, the packing of multiple, different-shaped subunits within a single complex may necessitate flexibility. Any protein that could form sufficient intersubunit interactions might be inherently flexible in its unbound state due to a lack of intramolecular contacts. A related issue has recently been discussed by Janin and Sternberg, who suggested that many intrinsically disordered proteins are simply "proteins waiting for a partner" [54]. They propose that actual disorder should be rare *in vivo*, as these proteins will usually be protected by chaperones prior to assembly. Ultimately, more studies will be required to quantify the extent of *in vivo* flexibility and disorder, and to further disentangle the functional importance of unbound-state properties from the conformational requirements of bound subunits.

Methods

Protein Structure Dataset

Biological units of protein crystal structures (<5 Å resolution) were taken from the Protein Data Bank on 2012-08-08, considering chains ≥40 residues. We filtered out backbone-only models and structures containing nucleic acids or >10% nonwater heteroatoms. Heteromers formed by subunit cleavage were also removed by identifying nonidentical chains from the same complex having the same *db_id* assignment. Additionally, protein complexes annotated as having quaternary structure assignment errors [55] were excluded. Symmetry groups were taken directly from the PDB. The number of nonhomologous subunits in a complex was defined on the basis of chains with distinct SUPERFAMILY "family" domain assignments [56]. Complexes in which no subunits had domain assignments were not considered in the "number of nonhomologous subunits" analyses.

Solvent-accessible surface areas and interface sizes were calculated with AREAIMOL. A_{rel} values were calculated according to $A_{\mathrm{rel}} = A_{\mathrm{s}}/4.44M^{0.77}$, where A_{s} is the solvent-accessible surface area and M is the molecular mass, as in [22]. The A_{rel} values of the dimeric precursors of dihedral complexes were calculated in the same way, except the total solvent-accessible surface area of each dimer was calculated, and the masses of the two subunits were summed. Complexes with two possible assembly pathways were identified as those symmetric homomers with at least six subunits having both heterologous and isologous interfaces >800 Å^2. Homomeric interfaces were identified as being isologous if the correlation between the residue-specific buried surface area for each subunit in an interacting pair was >0.7.

Secondary structure was calculated for each protein chain with STRIDE [57], and the following secondary structure groups were used in Table S1: α proteins (>20% α-helical residues), β proteins (>20% β-strand residues), and αβ proteins (>20% α-helical residues and >20% β-strand residues). Intrinsic disorder was predicted from protein sequences with IUPRED [58], using the "long" setting and threshold of 0.5 for identifying disordered residues.

Protein complexes in which all unique chains share >50% sequence identity were clustered. In addition, to avoid highly similar complexes that vary only slightly in their subunit composition, heteromeric complexes sharing at least four unique chains were clustered. From each cluster, only the complex with the most amino acid residues (ignoring subunit repeats) was selected for the nonredundant set used in this study (8,700 homomers and 1,552 heteromers). However, we note that this sequence-redundancy filtering is not perfect, as proteins can share sequence identity significantly lower than 50%, yet still be quite similar structurally. Therefore, we also created a stricter nonredundant set of protein complexes that are nonhomologous at the structural level by only considering only complexes with unique SUPERFAMILY domain assignments (2,208 homomers and 1,046 heteromers). The main structural analyses from Figures 1A and 2 were repeated with this strict dataset, and the results are essentially the same (Table S1). All complexes used in this study and relevant subunit properties are included in Tables S4 and S5.

Evolutionary Analysis

To map human genes against protein structures, a *blastp* search against all human proteins in Ensembl was performed for each protein chain. All chains with >70% sequence identity to a human protein were considered. Orthologs of these proteins were then identified in a variety of different species with Ensembl Compara [40] and OMA [41] (all species are listed in Table S6). For some species, both databases were used, whereas some species were only available in one or the other. If an ortholog of a human gene that maps to a protein complex subunit was present in a given species, we presumed that that subunit was present in the last common ancestor with humans, and is therefore a "putative older subunit" with respect to that species. The analysis considering full-length and nearly full-length proteins only included chains where at least 75% of the residues from the full-length protein were observed in the crystal structure.

To identify the "putative newer subunits" that were likely not present in an ancestral species, we also considered homologs at the level of domain architecture. This allows us to identify more divergent proteins that might have possibly been playing a similar subunit role in an ancestral complex. Importantly, we do not use this information to say that an ancestral subunit was present, but instead to say that an ancestral subunit might possibly have been present. Using SUPERFAMILY genome-scale domain assignments [59], we asked for each human subunit whether any protein in a given organism has the same set of domains (ignoring N- to C-terminal order) as the full-length human protein. If so, this subunit was excluded as a "putative newer subunit" with respect to that species. Human proteins with no SUPERFAMILY domain assignments were not considered as either newer or older subunits. Finally, in addition to checking that any ortholog or homologs are not present in a given species, we also checked that they were not present in any species of a similar or greater evolutionary distance from humans. This helps to avoid bias from gene loss and genome annotation errors. The ranked evolutionary distance from humans for each species used for this analysis is provided in Table S6.

To generate nonredundant sets of protein complexes having both putative older subunits and putative newer subunits, we only considered a single complex mapping to a given pair of old and new human genes. Similar filtering was performed when the sets of different species were combined. All the sets of putative older and newer subunits are provided in Table S6. Overall, although they include different species, the Ensembl Compara and OMA databases gave very similar results. Table S2 also includes the results for different species calculated with either one or the other databases.

Supporting Information

Figure S1 A_{rel} values of bound subunits from protein complexes are predictive of intrinsic flexibility in the unbound state. (A) Comparison between A_{rel} values of monomeric proteins, A_{rel}(free), and those same proteins (>98% sequence identity, <2% length difference) bound as subunits within homomeric or heteromeric complexes, A_{rel}(bound). In total, 288 homomer and 387 heteromer pairs were identified from the nonredundant dataset used in this study (provided in Table S5). The very strong correlations demonstrate that the A_{rel} of the bound state is highly predictive of the A_{rel}, and thus the intrinsic flexibility, of the free state. The mean difference between A_{rel}(bound) and A_{rel}(free) is 0.9% (mean absolute difference of 2.6%) for homomers and 0.7% (mean absolute difference of 3.0%) for heteromers, suggesting that there is a very slight tendency for A_{rel}(bound) to overestimate A_{rel}(free). These values are consistent with a recent study showing that the accessible surface area of interface residues in the bound state are on average 3.3% higher than in the unbound state [60]. The outliers here are mostly from domain-swapped homomers, where the swapped bound state will have a substantially higher A_{rel} value, but the free state is stabilized by the same intermolecular interactions being formed intramolecularly. Given the overall high correlations and the rarity of outliers observed here, and the fact that domain swapping is only observed in ~5% of protein families [61], the effect of domain swapping on our analyses should be minimal. (B) Fraction of predicted intrinsically disordered residues for bound subunits for which no corresponding monomer structure exists, grouped by A_{rel} value. Error bars represent SEM. The overall correlation (r) between A_{rel} and intrinsic disorder is 0.313 ($N = 9{,}527$). For those subunits for which a corresponding monomer structure does exist (sequence identity >50%), the correlation is much lower ($r = 0.137$, $N = 2{,}695$).
(TIFF)

Figure S2 Boxplot representations of A_{rel} distributions for subunits from different groups of protein complexes. Boxplots are generated in R using standard settings. The *y*-axes are plotted logarithmically. Nonoverlapping notches can be used as a rough indicator of statistically significant differences between two groups. (A) Subunits of homomers from different symmetry groups, as in Figure 1A. The *p* values for the differences between groups are shown calculated with the Wilcoxon rank-sum test. (B) Subunits from heteromers with different numbers of nonhomologous subunits, as in Figure 2.
(TIFF)

Figure S3 Intrinsic disorder is also related to quaternary structure topology, but less so than A_{rel} as a measure of intrinsic flexibility. Comparison of the percentage of residues predicted to be intrinsically disordered for subunits from (A–B) homomeric complexes from different symmetry groups (compare to Figure 1A) and (C–D) complexes with different numbers of nonhomologous subunits (compare to Figure 2A). (A) and (C) show means with SEM and (B) and (D) show boxplots, as in Figure S2. The trends for homomers in (A) and (C) mirror the results using A_{rel}, but are not as strong (compare to *p* values in Figure S2A).
(TIFF)

Figure S4 Subunit flexibility is largely independent of the number of subunits in a homomeric complex. Comparison of subunit flexibility, as measured by A_{rel}, to the number of subunits in homomers from different symmetry groups. The overall correlations (*r*) between A_{rel} and number of subunits are 0.115 for cyclics ($p = 0.0002$), 0.056 for dihedrals ($p = 0.03$), and 0.092 for asymmetrics ($p = 0.07$). Thus, there appears to be a very slight but significant tendency for larger homomers to have more flexible subunits. Error bars represent SEM.
(TIFF)

Figure S5 Interface size is related to symmetry but does not explain the observed flexibility trends. Comparison of interface sizes for homomeric subunits in different symmetry groups: (A) mean interface area per subunit; (B) mean relative interface area per subunit (i.e., what fraction of the surface forms interface). Error bars represent SEM. The trends here show essentially no correspondence with the flexibility results in Figure 1A, demonstrating that the association between flexibility and symmetry is not simply due to a requirement to form larger interfaces.
(TIFF)

Figure S6 The observation that evolutionarily more recent subunits are more flexible does not arise from a general tendency for increased flexibility in newer proteins. Although we observed a strong trend for the evolutionarily more recent subunits of protein complexes to be more flexible, it is possible that this could to some extent reflect a general tendency for evolutionarily more recent proteins to be more flexible. This could also arise if more flexible proteins tend to evolve at a faster rate, thus making them less likely to be detected as orthologs. We have addressed this in two ways: (A) comparison of A_{rel} values for human (or closely related) subunits whose most ancient orthologs are of varying evolutionary ages. Error bars represent SEM. There is no clear tendency for newer subunits to be more flexible (although subunits conserved in bacteria do appear to be less flexible), suggesting that our results cannot be explained by a general tendency for newer proteins to be more flexible. Full species names and the different evolutionary groups are provided in Table S6. (B) Comparison of sequence identities for subunits of varying flexibility. Here we grouped subunits by A_{rel} and plotted the mean sequence identities of Ensembl Compara orthologs from different species. This shows that, for the most part, sequence conservation is fairly constant with respect to A_{rel}, although there is some tendency for the most flexible human subunits to be less conserved, particularly when compared to yeast.
(TIFF)

Figure S7 The correspondence between subunit flexibility and the number of nonhomologous subunits per complex is not due to existing subunits evolving to become more flexible. The correspondence between subunit flexibility and the number of nonhomologous subunits per complex could possibly be explained if the existing (i.e., older) subunits of a complex can evolve to become more flexible as new, more rigid subunits are added. To test this, we grouped subunits by their SUPERFAMILY domain architecture. We considered only those groups where evolutionarily related proteins participate in different complexes that have different numbers of nonhomologous subunits. We then plot the relationship between A_{rel} and the number of nonhomologous subunits in three ways (values provided in Table S7): (A) The blue bars are essentially equivalent to Figure 2, although only those subunits that are also considered in (B) and (C) are included here. (B) The pink bars represent the "interfamily normalized" A_{rel} values, in which all variation should be due to evolutionary changes within a domain family. Here, the A_{rel} value for each subunit has been divided by the mean A_{rel} value for all subunits with the same domain architecture. The values are then all scaled by the mean A_{rel} of all subunits in the dataset. If there is a tendency for evolutionarily related proteins to be more flexible when they are part of complexes with more nonhomologous subunits, then we would expect these values to show an increasing trend. However, there is only a very slight trend, which does not explain the variation shown in (A). (C) The yellow bars represent the "intrafamily normalized" A_{rel} values, in which all variation should be due to differences between different types of domains. In these, the A_{rel} value of each subunit has been replaced with the mean A_{rel} value for all subunits with the same domain architecture. Thus we can see that nearly all of the trend in (A) can be explained by differences between evolutionarily unrelated proteins, strongly suggesting that the scheme in Figure 3A is correct and that existing subunits do not generally evolve to become more flexible in order to accommodate new subunits.
(TIFF)

Figure S8 The association between flexibility and the number of nonhomologous subunits per complex is preserved across different species. This plot is essentially the same as Figures 2 and 4C, except it considered separately the nine species with the most heteromers in our nonredundant dataset. A clear trend is observed for nearly all species. Only *M. musculus* and *T. maritima* appear to deviate, although this is likely due to the limited size of the dataset, including the fact that no complexes with >3 nonhomologous subunits are present for these species.
(TIFF)

Figure S9 Increasing intrinsic disorder is associated with a greater number of interaction partners across different species. Comparison of the percentage of residues predicted to be intrinsically disordered for proteins grouped by their number of experimentally identified interaction partners. Experimental protein–protein interactions were taken from STRING v9.0 [62], using only interactions with an experimental evidence confidence score >0.3. Varying the threshold from 0.15 to 0.7 preserved the same general trends. The bacterial, archaeal, and two eukaryotic species with the most interactions are shown here. Error bars represent SEM.
(TIFF)

Table S1 Controlling for structural factors when comparing the flexibilities of subunits from different groups of protein complexes.

This table provides the raw values for the main results in Figures 1A and 2. It also provides the values for these analyses broken down by secondary structure group, and using only the strict structurally nonredundant set of protein complexes, filtered at the domain level.
(XLSX)

Table S2 Pairwise flexibility comparison between putative older and putative newer subunits of protein complexes with respect to all species used in this analysis. These values are the same as used in Figure 3C, except that all species are shown here. We also include the results when only Ensembl Compara or only OMA are used as a source of orthologs.
(XLSX)

Table S3 Comparison of the mean subunit flexibility and number of nonhomologous subunits per complex from different species. These are the same values used in Figure 4B, except that all 263 species with at least five nonredundant complexes in our dataset are shown here.
(XLSX)

Table S4 Homomeric and heteromeric protein complexes used in this study.
(XLSX)

Table S5 Properties of protein complex subunits.
(XLSX)

Table S6 Putative older and newer subunits identified from each species, along with the combined set of nonredundant complexes that have both older and newer subunits. Also included here are the results of the analyses including only full-length or nearly full-length PDB chains, and only complexes in which the newer subunits are longer than the older subunits. The highest sequence identity between a human gene and its ortholog in Ensembl Compara is provided for each older subunit.
(XLSX)

Table S7 A_{rel}, interfamily normalized A_{rel}, and intrafamily normalized A_{rel} values for subunits from different domain families. These are the values used for the analysis in Figure S7.
(XLSX)

Acknowledgments

We thank David Gfeller, Tina Perica, and Xiuwei Zhang for helpful comments on the manuscript.

Author Contributions

The author(s) have made the following declarations about their contributions: Conceived and designed the experiments: JAM SAT. Performed the experiments: JAM. Analyzed the data: JAM. Wrote the paper: JAM SAT.

References

1. Robinson CV, Sali A, Baumeister W (2007) The molecular sociology of the cell. Nature 450: 973-982. doi:10.1038/nature06523.
2. Tarassov K, Messier V, Landry CR, Radinovic S, Serna Molina MM, et al. (2008) An in vivo map of the yeast protein interactome. Science 320: 1465–1470. doi:10.1126/science.1153878.
3. Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, et al. (2012) A census of human soluble protein complexes. Cell 150: 1068–1081. doi:10.1016/j.cell.2012.08.011.
4. Monod J, Wyman J, Changeux J-P (1965) On the nature of allosteric transitions: a plausible model. J Mol Biol 12: 88–118.
5. Perica T, Marsh JA, Sousa FL, Natan E, Colwell LJ, et al. (2012) The emergence of protein complexes: quaternary structure, dynamics and allostery. Biochem Soc Trans 40: 475–491.
6. Bershtein S, Mu W, Wu W, Shakhnovich EI (2012) Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. Proc Natl Acad Sci USA 109: 4857–4862. doi:10.1073/pnas.1118157109.
7. Lynch M (2013) Evolutionary diversification of the multimeric states of proteins. Proc Natl Acad Sci USA 110: E2821–E2828. doi:10.1073/pnas.1310980110.
8. Blundell TL, Srinivasan N (1996) Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. Proc Natl Acad Sci U S A 93: 14243–14248.
9. Goodsell DS, Olson AJ (2000) Structural symmetry and protein function. Annu Rev Biophys Biomol Struct 29: 105–153. doi:10.1146/annurev.biophys.29.1.105.
10. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA (2006) 3D complex: a structural classification of protein complexes. PLoS Comput Biol 2: e155. doi:10.1371/journal.pcbi.0020155.
11. Janin J, Bahadur RP, Chakrabarti P (2008) Protein-protein interaction and quaternary structure. Q Rev Biophys 41: 133–180. doi:10.1017/S0033583508004708.
12. Levy ED, Erba EB, Robinson CV, Teichmann SA (2008) Assembly reflects evolution of protein complexes. Nature 453: 1262–1265. doi:10.1038/nature06942.
13. Marsh JA, Hernández H, Hall Z, Ahnert S, Perica T, et al. (2013) Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell 153: 461–470.
14. Boehr DD, Nussinov R, Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol 5: 789–796. doi:10.1038/nchembio.232.
15. Dobbins SE, Lesk VI, Sternberg MJE (2008) Insights into protein flexibility: the relationship between normal modes and conformational change upon protein-protein docking. Proc Natl Acad Sci USA 105: 10390–10395. doi:10.1073/pnas.0802496105.
16. Marsh JA, Teichmann SA (2011) Relative solvent accessible surface area predicts protein conformational changes upon binding. Structure 19: 859–867. doi:10.1016/j.str.2011.03.010.
17. Marsh JA, Teichmann SA (2014) Parallel dynamics and evolution: protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. BioEssays 36: 209–218.
18. Namba K (2001) Roles of partly unfolded conformations in macromolecular self-assembly. Genes Cells 6: 1–12.
19. Hegyi H, Schad E, Tompa P (2007) Structural disorder promotes assembly of protein complexes. BMC Struct Biol 7: 65. doi:10.1186/1472-6807-7-65.
20. Tóth-Petróczy Á, Oldfield CJ, Simon I, Takagi Y, Dunker AK, et al. (2008) Malleable machines in transcription regulation: the mediator complex. PLoS Comput Biol 4: e1000243. doi:10.1371/journal.pcbi.1000243.
21. Chothia C (1975) Structural invariants in protein folding. Nature 254: 304–308.
22. Marsh JA (2013) Buried and accessible surface area control intrinsic protein flexibility. J Mol Biol 425: 3250–3263.
23. Hall Z, Hernández H, Marsh JA, Teichmann SA, Robinson CV (2013) The role of salt bridges, charge density, and subunit flexibility in determining disassembly routes of protein complexes. Structure 21: 1325–1337. doi:10.1016/j.str.2013.06.004.
24. Gunasekaran K, Tsai C-J, Nussinov R (2004) Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. J Mol Biol 341: 1327–1341. doi:10.1016/j.jmb.2004.07.002.
25. Marsh JA, Teichmann SA, Forman-Kay JD (2012) Probing the diverse landscape of protein flexibility and binding. Curr Opin Struct Biol 22: 643–650. doi:10.1016/j.sbi.2012.08.008.
26. Lukatsky DB, Shakhnovich BE, Mintseris J, Shakhnovich EI (2007) Structural similarity enhances interaction propensity of proteins. J Mol Biol 365: 1596–1606. doi:10.1016/j.jmb.2006.11.020.
27. André I, Strauss CEM, Kaplan DB, Bradley P, Baker D (2008) Emergence of symmetry in homooligomeric biological assemblies. Proc Natl Acad Sci USA 105: 16148–16152. doi:10.1073/pnas.0807576105.
28. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. J Mol Biol 285: 2177–2198.
29. Gunasekaran K, Tsai C-J, Kumar S, Zanuy D, Nussinov R (2003) Extended disordered proteins: targeting function with less scaffold. Trends Biochem Sci 28: 81–85.
30. Perica T, Chothia C, Teichmann SA (2012) Evolution of oligomeric state through geometric coupling of protein interfaces. Proc Natl Acad Sci USA 109: 8127–8132. doi:10.1073/pnas.1120028109.
31. Dayhoff JE, Shoemaker BA, Bryant SH, Panchenko AR (2010) Evolution of protein binding modes in homooligomers. J Mol Biol 395: 860. doi:10.1016/j.jmb.2009.10.052.
32. Kim WK, Henschel A, Winter C, Schroeder M (2006) The many faces of protein-protein interactions: a compendium of interface geometry. PLoS Comput Biol 2: e124. doi:10.1371/journal.pcbi.0020124.
33. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, et al. (2001) Identification of potential interaction networks using sequence-based searches for conserved

protein-protein interactions or "interologs." Genome Res 11: 2120–2126. doi:10.1101/gr.205301.

34. Aloy P, Ceulemans H, Stark A, Russell RB (2003) The relationship between sequence and interaction divergence in proteins. J Mol Biol 332: 989–998.
35. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, et al. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res 14: 1107–1118. doi:10.1101/gr.1774904.
36. Beltrao P, Serrano L (2007) Specificity and evolvability in eukaryotic protein interaction networks. PLoS Comput Biol 3: e25. doi:10.1371/journal.pcbi.0030025.
37. Lewis ACF, Jones NS, Porter MA, Deane CM (2012) What evidence is there for the homology of protein-protein interactions? PLoS Comput Biol 8: e1002645. doi:10.1371/journal.pcbi.1002645.
38. Andreani J, Faure G, Guerois R (2012) Versatility and invariance in the evolution of homologous heteromeric interfaces. PLoS Comput Biol 8: e1002677. doi:10.1371/journal.pcbi.1002677.
39. Van Dam TJP, Snel B (2008) Protein complex evolution does not involve extensive network rewiring. PLoS Comput Biol 4: e1000132. doi:10.1371/journal.pcbi.1000132.
40. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, et al. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19: 327–335. doi:10.1101/gr.073585.107.
41. Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C (2011) OMA 2011: orthology inference among 1000 complete genomes. Nucl Acids Res 39: D289–D294. doi:10.1093/nar/gkq1238.
42. Albà MM, Castresana J (2005) Inverse relationship between evolutionary rate and age of mammalian genes. Mol Biol Evol 22: 598–606. doi:10.1093/molbev/msi045.
43. Domazet-Loso T, Tautz D (2003) An evolutionary analysis of orphan genes in Drosophila. Genome Res 13: 2213–2219. doi:10.1101/gr.1311003.
44. Lynch M (2012) The evolution of multimeric protein assemblages. Mol Biol Evol 29: 1353–1366. doi:10.1093/molbev/msr300.
45. Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA (2007) Evolution of protein complexes by duplication of homomeric interactions. Genome Biol 8: R51. doi:10.1186/gb-2007-8-4-r51.
46. Wainright PO, Hinkle G, Sogin ML, Stickel SK (1993) Monophyletic origins of the metazoa: an evolutionary link with fungi. Science 260: 340–342.
47. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J 272: 5129–5148. doi:10.1111/j.1742-4658.2005.04948.x.
48. Patil A, Nakamura H (2006) Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. FEBS Lett 580: 2041–2045. doi:10.1016/j.febslet.2006.03.003.
49. Ekman D, Light S, Björklund ÅK, Elofsson A (2006) What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? Genome Biology 7: R45. doi:10.1186/gb-2006-7-6-r45.
50. Miller S, Lesk AM, Janin J, Chothia C (1987) The accessible surface area and stability of oligomeric proteins. Nature 328: 834–836. doi:10.1038/328834a0.
51. Kim PM, Sboner A, Xia Y, Gerstein M (2008) The role of disorder in interaction networks: a structural analysis. Mol Syst Biol 4: 179–179. doi:10.1038/msb.2008.16.
52. Hsu W-L, Oldfield CJ, Xue B, Meng J, Huang F, et al. (2013) Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. Protein Sci 22: 258–273. doi:10.1002/pro.2207.
53. Fernández A, Lynch M (2011) Non-adaptive origins of interactome complexity. Nature 474: 502–505. doi:10.1038/nature09992.
54. Janin J, Sternberg MJE (2013) Protein flexibility, not disorder, is intrinsic to molecular recognition. F1000 Biol Rep 5: 2. doi:10.3410/B5-2.
55. Levy ED (2007) PiQSi: protein quaternary structure investigation. Structure 15: 1364–1367. doi:10.1016/j.str.2007.09.019.
56. Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313: 903–919. doi:10.1006/jmbi.2001.5080.
57. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23: 566–579. doi:10.1002/prot.340230412.
58. Dosztányi Z, Csizmók V, Tompa P, Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 347: 827–839. doi:10.1016/j.jmb.2005.01.071.
59. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, et al. (2009) SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res 37: D380–D386. doi:10.1093/nar/gkn762.
60. Chakravarty D, Guharoy M, Robert CH, Chakrabarti P, Janin J (2013) Reassessing buried surface areas in protein-protein complexes. Protein Sci 22: 1453–1457. doi:10.1002/pro.2330.
61. Huang Y, Cao H, Liu Z (2012) Three-dimensional domain swapping in the protein structure space. Proteins 80: 1610–1619. doi:10.1002/prot.24055.
62. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39: D561–D568. doi:10.1093/nar/gkq973.