RESEARCH ARTICLE SUMMARY

PROTEIN STRUCTURE

Principles of assembly reveal a periodic table of protein complexes

Sebastian E. Ahnert,* Joseph A. Marsh,* Helena Hernández, Carol V. Robinson, Sarah A. Teichmann†

INTRODUCTION: The assembly of proteins into complexes is crucial for most biological processes. The three-dimensional structures of many thousands of homomeric and heteromeric protein complexes have now been determined, and this has had a broad impact on our understanding of biological function and evolution. Despite this, the organizing principles that underlie the great diversity of protein quaternary structures observed in nature remain poorly understood, particularly in comparison with protein folds, which have been extensively classified in terms of their architecture and evolutionary relationships.

RATIONALE: In this work, we sought a comprehensive understanding of the general principles underlying quaternary structure organization. Our approach was to consider protein complexes in terms of their assembly. Many protein complexes assemble spontaneously via ordered pathways in vitro, and these pathways have a strong tendency to be evolutionarily conserved. Furthermore, there are strong similarities between protein complex assembly and evolutionary pathways, with assembly pathways often being reflective of evolutionary histories, and vice versa. This suggests that it may be useful to consider the types of protein complexes that have evolved from the perspective of what assembly pathways are possible.

RESULTS: We first examined the fundamental steps by which protein complexes can assemble, using electrospray mass spectrometry experiments, literature-curated assembly data, and a large-scale analysis of protein complex structures. We found that most assembly steps can be classified into three basic types: dimerization, cyclization, and heteromeric subunit addition. By systematically combining different assembly steps in different ways, we were able to enumerate a large set of possible quaternary structure topologies, or patterns of key interfaces between the proteins within a complex. The vast majority of real protein complex structures lie within these topologies. This enables a natural organization of protein complexes into a "periodic table," because each heteromer can be related to a simpler symmetric homomer topology. Exceptions are mostly the result of quaternary structure assignment errors, or cases where sequence-identical subunits can have different interactions and thus introduce asymmetry. Many of these asymmetric complexes fit the paradigm of a periodic table when their assembly role is considered. Finally, we implemented a model based on the periodic table, which predicts the expected frequencies of each quaternary structure topology, including those not yet observed. Our model correctly predicts quaternary structure topologies of recent crystal and electron microscopy structures that are not included in our original data set.

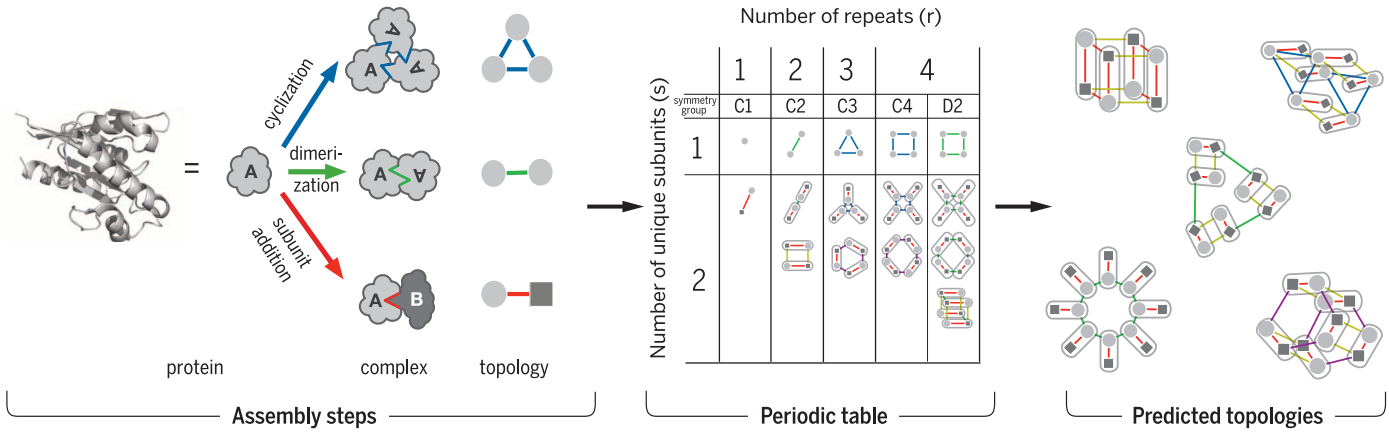
ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.aaa2245

CONCLUSION: This work explains much of the observed distribution of known protein complexes in quaternary structure space and provides a framework for understanding their evolution. In addition, it can contribute considerably to the prediction and modeling of quaternary structures by specifying which topologies are most likely to be adopted by a complex with a given stoichiometry, potentially providing constraints for multi-subunit docking and hybrid methods. Lastly, it could help in the bioengineering of protein complexes by identifying which topologies are most likely to be stable, and thus which types of essential interfaces need to be engineered. ■

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. E-mail: saraht@ebi.ac.uk
Cite this paper as S.E. Ahnert *et al.*, *Science* 350, aaa2245 (2015). DOI: 10.1126/science.aaa2245

Protein assembly steps lead to a periodic table of protein complexes and can predict likely quaternary structure topologies. Three main assembly steps are possible: cyclization, dimerization, and subunit addition. By combining these in different ways, a large set of possible quaternary structure topologies can be generated. These can be arranged on a periodic table that describes most known complexes and that can predict previously unobserved topologies.



RESEARCH ARTICLE

PROTEIN STRUCTURE

Principles of assembly reveal a periodic table of protein complexes

Sebastian E. Ahnert,[1]* Joseph A. Marsh,[2,3]* Helena Hernández,[4] Carol V. Robinson,[4] Sarah A. Teichmann[1,3,5]†

Structural insights into protein complexes have had a broad impact on our understanding of biological function and evolution. In this work, we sought a comprehensive understanding of the general principles underlying quaternary structure organization in protein complexes. We first examined the fundamental steps by which protein complexes can assemble, using experimental and structure-based characterization of assembly pathways. Most assembly transitions can be classified into three basic types, which can then be used to exhaustively enumerate a large set of possible quaternary structure topologies. These topologies, which include the vast majority of observed protein complex structures, enable a natural organization of protein complexes into a periodic table. On the basis of this table, we can accurately predict the expected frequencies of quaternary structure topologies, including those not yet observed. These results have important implications for quaternary structure prediction, modeling, and engineering.

Evolution has given rise to an enormous variety of protein complexes (*1*–*3*). The organizing principles that underlie this diversity remain poorly understood, particularly in comparison with protein folds, which have been classified extensively in terms of their architecture (*4*–*6*) and evolution (*7*, *8*). However, network models have shown considerable promise in recent years for characterizing and comparing protein complexes. For example, complexes are often represented as networks of associations between proteins, with little consideration for structure or stoichiometry. Alternatively, a graph representation, which we introduced several years ago, can be used to capture the main features of quaternary structure topology (*9*). In this model, the nodes are the polypeptide chains, defined by their amino acid sequence and often referred to as subunits, and the edges are the interfaces between physically interacting chains, weighted according to size.

Many protein complexes assemble spontaneously via ordered pathways in vitro, and we have shown that these assembly pathways have a strong tendency to be evolutionarily conserved (*10*, *11*). Furthermore, there are strong similarities between protein complex assembly and evolutionary pathways, with assembly pathways often being reflective of evolutionary histories, and vice versa (*12*). Thus, quaternary structure evolution essentially can be thought of as an assembly process occurring on an evolutionary time scale. This suggests that it may be useful to consider the types of protein complexes that have evolved from the perspective of assembly pathways.

In this work, we attempted to understand and explain the organization of protein complexes in quaternary structure space, using the principles of assembly. First, by characterizing the assembly pathways of a large number of protein complexes, we found that assembly can be explained generally by three basic steps: dimerization, cyclization, and subunit addition. Combinations of these steps allow us to exhaustively enumerate possible quaternary structure topologies within a given region of quaternary structure space.

To achieve this, we considered each polypeptide chain as a distinct self-assembly building block and considered all the ways in which interfaces can be distributed across the chains that are present in the complex. The large variety of possible topologies generated by this approach were then compared to observed structures. We found that ~92% of known protein complex structures are compatible with this model.

A major benefit of this assembly-centric view of protein complexes is that it enables a natural organization of complexes into a "periodic table," ordered by the number of subunit repeats (r) and the number of subunit types that are unique within a given complex (s). Exceptions are primarily the result of quaternary structure assignment errors or cases where sequence-identical subunits can have different interactions and thus introduce asymmetry. Many of these asymmetric complexes fit the paradigm of a periodic table when their assembly role (rather than their subunit identity) is considered.

Finally, by combining the periodic table with our enumeration, we introduced a model to predict the expected frequencies of different quaternary structure topologies. Not only does this model effectively replicate the relative frequencies of known protein complex structures, it also predicts the new topologies that are most likely to be observed in the future.

A survey of transitions in the assembly pathways of protein complexes

To understand the principles that underlie quaternary structure organization, it is useful to begin by considering the different ways in which protein complexes can assemble. We therefore first sought to determine the assembly and disassembly ["(dis)assembly"] pathways for as many protein complexes as possible. Previously, we have used electrospray mass spectrometry to characterize the (dis)assembly of eight homomers (*10*) and eight heteromers (*11*, *13*). Whereas the homomers followed simple pathways, more diversity was observed for the heteromeric complexes. For this reason, in this study, we experimentally characterized the (dis)assembly pathways of nine additional heteromers with widely varying quaternary structures (Fig. 1). In all of these cases, well-defined intermediate subcomplexes could be identified under at least one set of experimental conditions. All eight homomers and 15 of the 17 heteromers characterized by electrospray mass spectrometry to date have stoichiometries under native conditions that are consistent with the published biological units in the Protein Data Bank (PDB).

We also searched the literature for protein complexes of known structure for which experimental (dis)assembly data are available, as we have done previously (*10*, *11*). Often, these are cases where at least two different oligomeric states have been observed under equilibrium conditions. In total, we identified 11 homomers and 13 heteromers for which some (dis)assembly information is available in the literature.

We obtained further information on protein assembly by considering the large number of protein complexes of known structure. We searched for pairs of protein complexes where the quaternary structure of one complex could be described as a subset of the other. Such pairs include, for example, a homodimer and a homotetramer with highly similar or identical sequences, suggesting that the tetramer assembles via a dimeric intermediate. Also included are homomer-heteromer pairs, where the heteromer has acquired a subunit with respect to the homomers. In total, this approach identified 154 homomers and 263 heteromers with putative structure-based assembly information.

We recognize that the structure-based pathways do not represent direct characterization of assembly. Instead, they indicate that two or more different quaternary structure states have been observed, and we assume that assembly transitions can occur between them. Even for biophysically characterized assembly pathways, we do not always have evidence that they are physiologically relevant. However, the fact that the biophysical and structure-based pathways have a strong tendency to reflect evolutionary history (*10*) and to

[1]Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge CB3 0HE, UK. [2]Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. [3]European Molecular Biology Laboratory–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [4]Physical and Theoretical Chemistry Laboratory, Department of Chemistry, University of Oxford, South Parks Road, Oxford OX1 3QZ, UK. [5]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
*These authors contributed equally to this work. †Corresponding author. E-mail: saraht@ebi.ac.uk

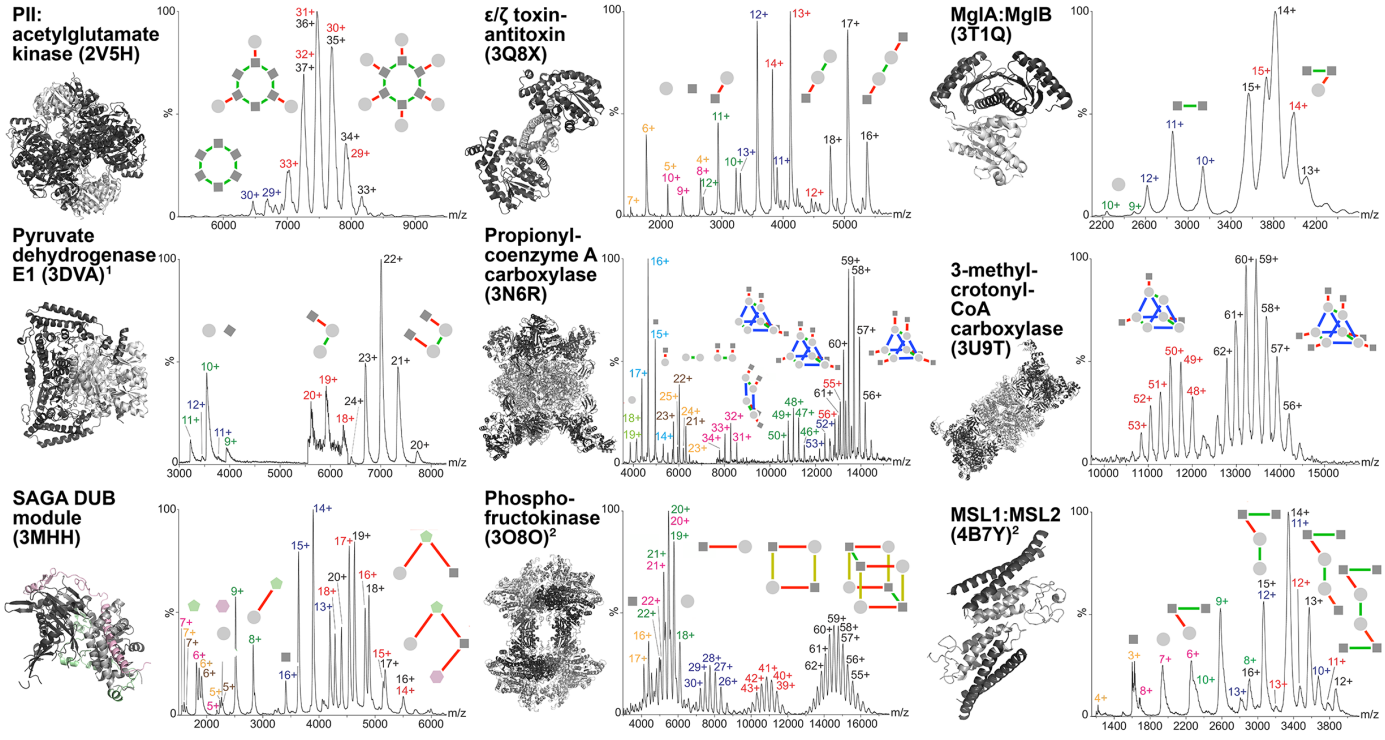


Fig. 1. Mass spectrometry characterization of heteromer (dis)assembly pathways. For each characterized complex, the known three-dimensional structure is shown with a representative mass spectrum, accompanied by graph representations of the full complex and subcomplexes. In all cases, the full complex is represented by the rightmost graph. A full list of subcomplexes is provided in table S1. The structures of 3DVA, 3O8O, and 4B7Y shown here differ from those in the PDB: 3DVA is missing the γ subunit, because it was not present in our sample, and the 4:4 model of 3O8O and the 4:2 model of 4B7Y were built from the unit cell to match the mass spectrometry data. Colors in the graph representations indicate homomeric isologous (green), homomeric heterologous (blue), and heteromeric heterologous (red) interfaces; shapes indicate different subunit types.

be evolutionarily conserved (*11*) does suggest that they have a functional relevance.

Given this large set of assembly data, we next asked what quaternary structure transitions (assembly steps) tend to be observed. For homomeric complexes, we classified all possible transitions into three types (Fig. 2A, left). First, there is dimerization, where a doubling of the complex occurs and a twofold axis of rotational symmetry is formed (e.g., monomer-to-dimer or dimer-to-tetramer). Second, there is cyclization, which involves the assembly of a ring-like quaternary structure with higher-order rotational symmetry (e.g., monomer-to-trimer or monomer-to-tetramer). Third, there is fractional transition, an inherently asymmetric step in which the quaternary structure changes by a non-integer ratio (e.g., dimer-to-timer or trimer-to-tetramer).

For each homomer with assembly data, we identified all the assembly steps that could account for the transitions between the free monomers, the observed subcomplexes, and the full complex (see Methods). The distributions of these three different assembly steps are shown in Fig. 2B. All three data sets show a similar trend, with dimerization being the most common step, cyclization being the next most common, and fractional transitions being rare. This is consistent with previous observations of the favorable assembly and evolutionary transitions between homomers with different symmetries (*10*).

In heteromers, there are two further assembly steps that are possible, in addition to the three steps observed for homomers. These are illustrated in Fig. 2A (right): subunit addition, in which a new subunit is acquired (e.g., monomer-to-heterodimer); and nonstoichiometric transition, in which the types of subunits within the heteromer remain the same, but their relative ratios change (e.g., assembly from 1:1 to 2:1 stoichiometry).

The distributions of all five possible assembly steps for heteromers are shown in Fig. 2C. The same trend is observed among the three homomeric steps, with dimerization being the most common and few fractional transitions. However, across all five possible steps, the most common observed step for heteromers from all three data sets is heteromeric subunit addition.

Within the heteromers, there is a difference between the transitions observed in the mass spectrometry data and those recorded in the other data sets. Specifically, nonstoichiometric transitions are much more common in mass spectrometry data, as evident from the considerable number of subcomplex intermediates with uneven stoichiometry (different numbers of each subunit type) shown in Fig. 1. This can be attributed to two factors: the sensitivity of the mass spectrometry measurements to low-populated assembly intermediates, and the way in which the mass spectrometry experiments are performed—namely, over a range of destabilizing solution conditions designed to progressively disrupt the quaternary structure of the complex. We know that such nonstoichiometric transitions must occur in many cases where they are not observed. For example, consider the transition from an AA homodimer to a BAAB heterotetramer, where there is no interaction between the two B subunits. In this case, an AAB assembly intermediate should form, given that it is highly improbable that two separate B subunits would bind simultaneously. However, this asymmetric subcomplex is unlikely to be observed under non-destabilizing conditions and without highly sensitive mass spectrometry measurements.

Enumeration of the topological space of protein complexes

Next, we explored quaternary structure space by combining different assembly steps to determine which protein complex topologies are possible. Given that the protein complex assembly pathways described above are dominated by dimerization, cyclization, and subunit addition, we focused on these three steps.

An important consideration is interface symmetry. Dimerization results in a twofold axis of rotational symmetry, and therefore the interface formed by dimerization will be isologous (symmetric or head-to-head) and will involve two

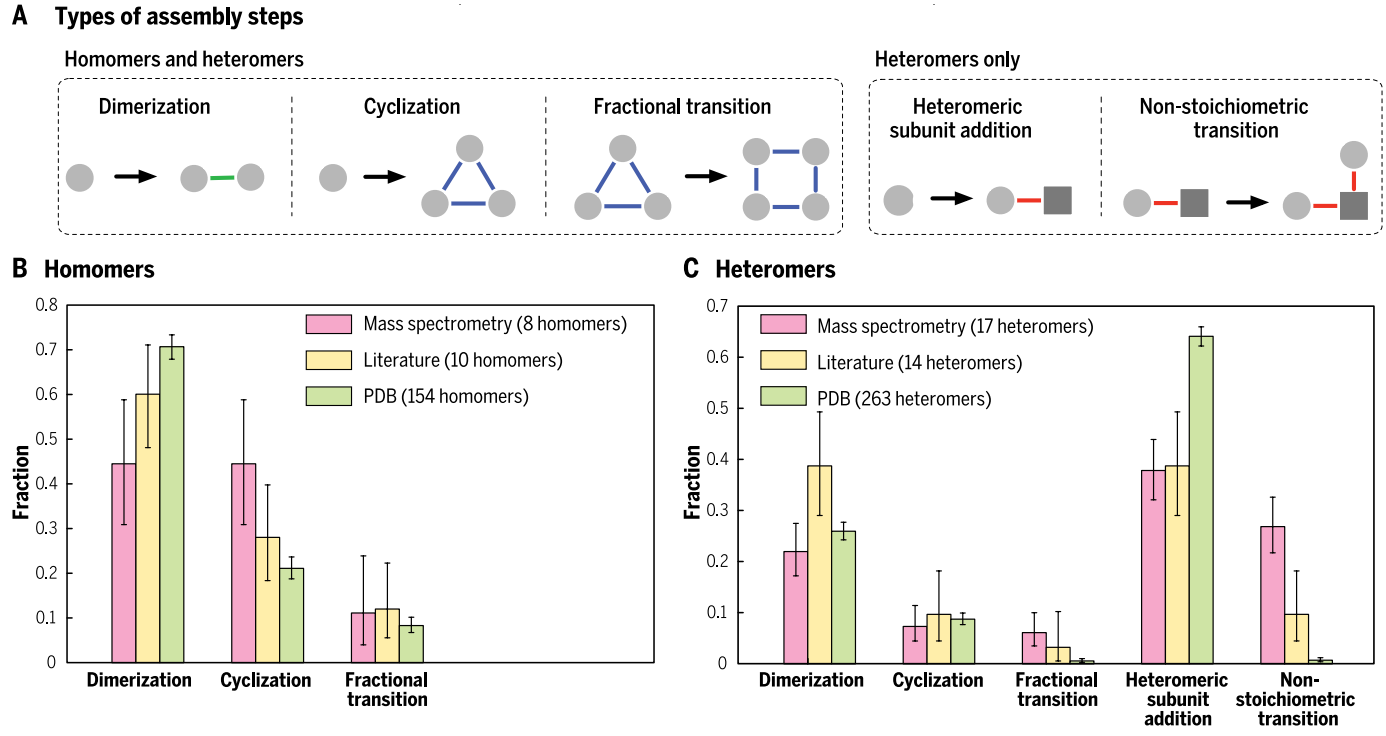


Fig. 2. Types of assembly steps observed in homomeric and heteromeric complexes. (**A**) The five possible types of assembly steps. (**B** and **C**) Distribution of observed assembly steps for homomers and heteromers from mass spectrometry experiments, from assembly pathways identified in the literature, and from complexes with varying quaternary structures in the PDB. Error bars represent 68% Clopper-Pearson confidence intervals.

identical surfaces on subunits of the same type (*14*). In contrast, cyclization results in higher-order rotational symmetry and is associated with interfaces that are heterologous (asymmetric or head-to-tail) and that involve two different surfaces on the same type of subunit. In addition, there are heteromeric interfaces, formed between two distinct polypeptide chains and hence by definition also heterologous.

Proteins are inherently asymmetric at the level of individual polypeptide chains, so we can assume that the same interface surface cannot appear on the same protein twice, or on two structurally different proteins. Together with this fundamental assumption, the three transitions (dimerization, cyclization, and subunit addition) all lead to symmetric protein complexes with even subunit stoichiometry. This is because subunit addition can be viewed as the formation of a larger multiprotein subunit, or protomer, which means that we can extend the homomeric definitions of dimerization and cyclization to homomers formed of these multiprotein subunits, leading to equal multiples of each type of protein (Fig. 3).

Every homomeric complex (of single-protein or multiprotein subunits) can have at most two isologous or heterologous interfaces, because each new homomeric interface imposes a new axis of rotational symmetry. In other words, symmetry constrains the number of homomeric interface types to a maximum of two. One or two interfaces of two possible types give us five scenarios: (i) one isologous, (ii) one heterologous, (iii) two isologous, (iv) two heterologous, and (v) one isologous and one heterologous.

To elucidate all possible heteromeric topologies that can arise under these constraints, we started by enumerating all tree-like topologies of s subunits, in which each subunit type occurs exactly once (Fig. 3). We used trees rather than all possible graphs, because we wanted to distinguish between essential and nonessential ("circumstantial") interfaces in the complex (see next section and Methods for details). For each of the five scenarios described above, we then considered all topologically distinct ways (that is, distinct under symmetry operations on the tree) in which the interfaces can be distributed across the set of subunits and pairs of subunits on the tree. The final step was to construct the topologies of the complexes from these distributions of interfaces across the tree. Some of these are isomorphic (taking into account interface types and subunit identities), which reduces the overall number of topologies.

An important difference between our idealized model and real protein complexes is that real complexes can have more interfaces. However, we can directly relate real protein topologies to the above idealized forms if we consider some of the weaker intersubunit contacts to be circumstantial. In other words, stronger, essential interfaces exist that bind the complex together by themselves. We can distinguish between the essential and circumstantial interfaces by successively cutting away, in increasing order of size, as many interfaces as possible without giving rise to disconnected components of the complex (see Methods for details), thereby producing the simplest possible graph representation of a quaternary structure topology. This contrasts with the previous approach of 3D Complex (*9*), in which all intersubunit interfaces are considered. Thus, this representation effectively sits above the more detailed classification of 3D Complex: A single simplified topology used here can correspond to multiple 3D Complex topologies.

Most real protein complexes are compatible with our model: 92.5% of homomers and 91.7% of heteromers have topologies identified in our exhaustive enumeration (Fig. 4). In these complexes, structurally identical proteins inhabit the same topological environment, meaning the same local environment in terms of the interfaces that they form with other subunits in a complex. We therefore define as bijective those complexes that have a one-to-one correspondence between their polypeptide sequence and their topological environment.

In contrast, all of the real protein complexes not compatible with our enumeration are nonbijective, meaning that sequence-identical subunits exist in nonequivalent topological environments (Fig. 4). The difference between bijective and nonbijective complexes is further illustrated in fig. S1.

Unlike our simple enumeration model that requires only three types of assembly steps, nonbijective complexes would require other asymmetric fractional transition and nonstoichiometric transition assembly steps. To explore this, we performed an exhaustive enumeration of all possible bijective and nonbijective topologies for complexes with specific stoichiometries. We found that for complexes with 2:2 stoichiometry, there are two possible bijective topologies, compared with seven possible nonbijective topologies (fig. S2). For

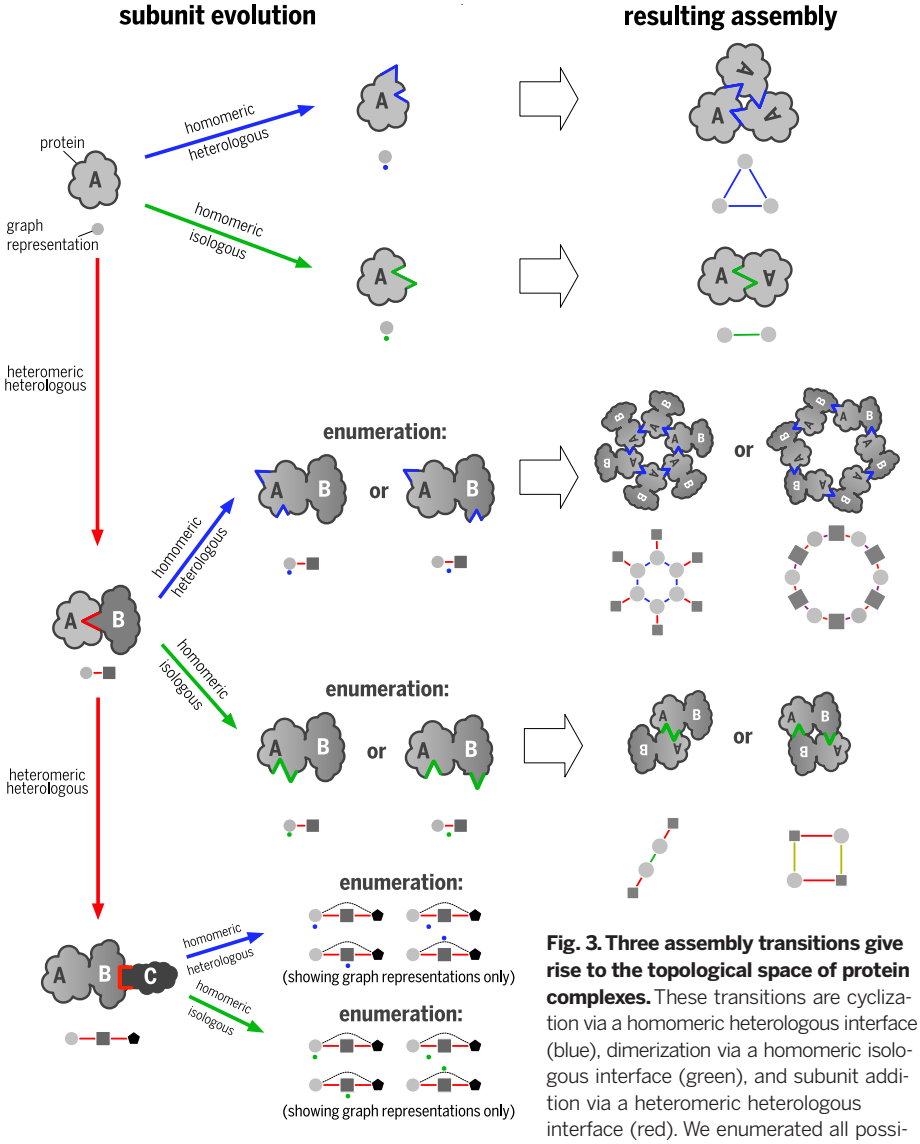


Fig. 3. Three assembly transitions give rise to the topological space of protein complexes. These transitions are cyclization via a homomeric heterologous interface (blue), dimerization via a homomeric isologous interface (green), and subunit addition via a heteromeric heterologous interface (red). We enumerated all possible topologies arising from these steps by calculating all ways in which a cyclic or dihedral interface can be distributed across a heteromer with 1:1 stoichiometry. For heterodimers, there are two such ways for both the cyclization and the dimerization steps. For heterotrimers, there are four such ways for each step. In the graph representation of the enumeration step, the possible locations of the distributed interfaces are indicated by colored dots.

complexes with 3:3 stoichiometry, there are also two possible bijective topologies, whereas the number of possible nonbijective topologies rises sharply to 250 (fig. S3). This illustrates a major benefit of our approach: By limiting our model to only three simple assembly steps, we are able to cover most observed protein complexes with a much smaller set of possible quaternary structure topologies.

To further justify our classification into bijective and nonbijective complexes, we used the fact that the quaternary structure assigned to a protein complex is often incorrect and does not represent the quaternary structure in solution or within the cell (*15*, *16*). Using a database of manually confirmed quaternary structure assignments (*17*), complemented by additional manual assignments of our own, we compared error rates for bijective and nonbijective homomers and heteromers (Fig. 4). We found that, whereas bijective complexes have a low rate of quaternary structure error (~10%), more than half of the nonbijective structures are the result of errors. Thus, most nonbijective protein complex structures are not genuine examples of biological asymmetry but instead are due to artifacts or errors in the structure determination process. This also suggests that a protein complex's nonbijective status could be very useful for identifying likely quaternary structure assignment errors.

The nonbijective complexes with uneven stoichiometry are an exception to the above, with only a 20% quaternary structure error rate. We recently studied these in detail and determined several different structural mechanisms by which they can form, which include varying degrees of pseudosymmetry, steric occlusion, and subunit flexibility leading to conformational differences between identical subunits (*18*). We found that those complexes for which a structural mechanism for uneven stoichiometry could not be ascertained were mostly the result of quaternary structure assignment errors (*18*).

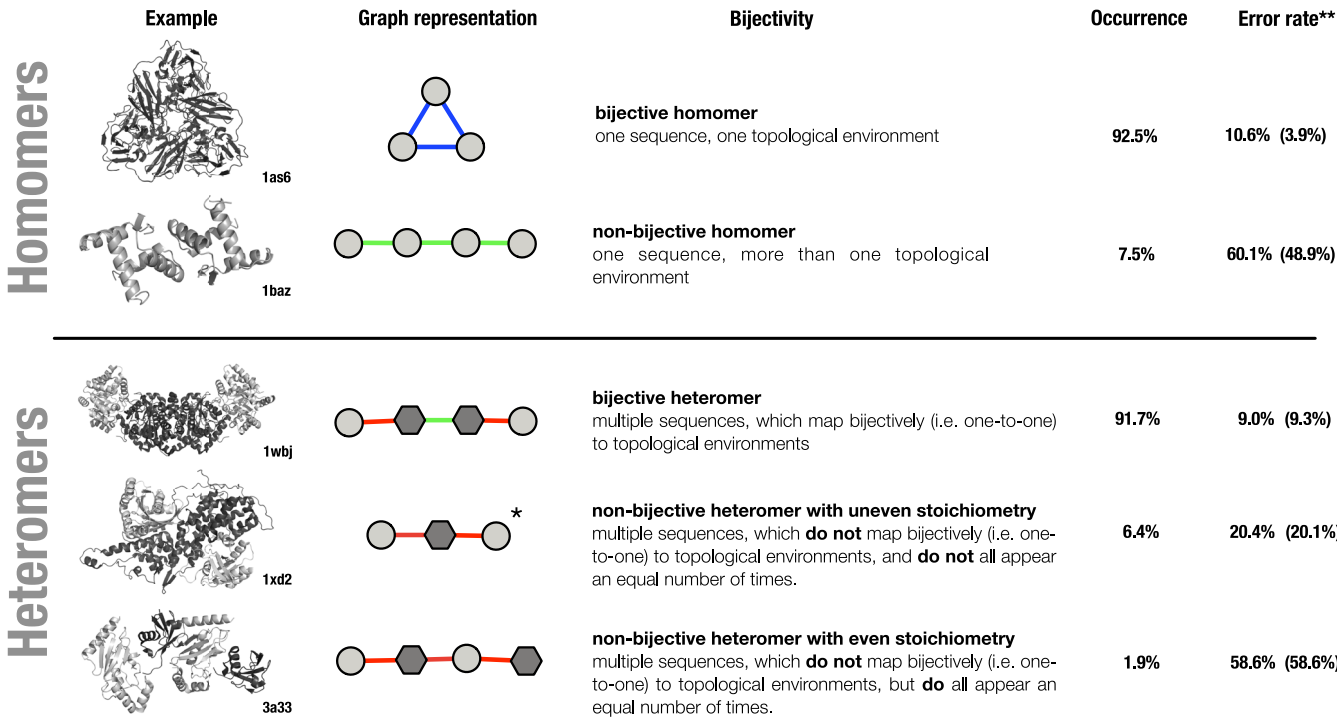
A periodic table of protein complexes

Analysis of the real and enumerated quaternary structure topologies above shows that all bijective heteromers can be related to simpler homomeric topologies. Specifically, if the different subunit types are grouped together as protomers, then the interactions between protomers will be equivalent to a bijective homomer topology or, for cases with no subunit repeats, to a monomer. This suggests a natural classification for protein complexes: in one dimension, by the number of their repeats, and within that, by their equivalent homomeric topologies; and in the other dimension, by the number of unique subunits. Figure 5 illustrates this periodic table of protein complexes for all topologies with ≤12 repeats and ≤4 unique subunits. In this classification, complexes related to the equivalent homomers are contained in the same column of the table, thus allowing the similarities between different heteromeric complexes to be easily recognized.

Most symmetry groups are associated with a single homomeric topology, except dihedral groups for ≥6 subunits and tetrahedral groups (12 subunits), which both have two topologies each. Higher-order symmetries, such as the octahedral and icosahedral groups, only appear for 24 or more subunits and can have three or more topologies each. Although the graph representation of the topology incorporates similarities between binding surfaces based on the identities of interacting residues, it does not require any nonlocal geometric information. Our graph representations therefore inherently include the symmetry group information of a complex, and they constitute the first network representation of complexes to do this.

Figure S4 shows the frequencies of each equivalent homomer symmetry group for complexes with varying numbers of unique subunits. Complexes with different numbers of unique subunits have similar distributions. Thus, homomers and heteromers populate the horizontal axis of the periodic table in a similar manner, although complexes with more unique subunits do tend to have fewer repeats.

The regions of the periodic table that correspond to higher numbers of repeats and subunits are sparsely populated. This can be attributed to two factors. First, there is a considerable bias among structurally characterized protein complexes toward those with smaller numbers of unique subunits, whereas evidence suggests that protein complexes in vivo will tend to have more unique components (*19*, *20*). Second, as shown in fig. S4, topologies toward the right of the periodic table tend to be less common, which suggests that cyclic or dihedral complexes with more repeated subunits may be less stable or more difficult to evolve. These regions can also be expected to be



* In this example the central protein forms two different interfaces with the two outer proteins due to the inherent asymmetry of proteins. The topological environments of the outer proteins therefore differ.

** values in brackets exclude 'probably yes' and 'probably no' error assignments in PiQSi from the analysis

Fig. 4. Frequencies of protein complex types and their quaternary structure assignment error rates. Among nonbijective heteromers, we further distinguished between those with even stoichiometry and those with uneven stoichiometry. The former are much more like to be the result of quaternary structure assignment errors. The latter are more likely to represent a biologically relevant quaternary structure. In the last column, we give alternative error rates in brackets that exclude the PiQSi (*17*) error assignments "probably yes" and "probably no" from the analysis. These error rates follow the same pattern for nonbijective heteromers of both even and uneven stoichiometries.

filled in coming years, at least to a certain extent. Figure 5A shows the rate at which new topologies have been discovered—roughly four per year for the past 20 years, with no signs of slowing. To illustrate the space of possible topologies, the number of discovered topologies versus the number possible as determined through exhaustive enumeration is shown in each cell of the periodic table (an example is shown in Fig. 5B).

This table is not "periodic" in the same sense as the periodic table of the elements, because it is in principle open-ended, as opposed to periodic with respect to atomic number. There are no theoretical limitations to quaternary structure topology space in either dimension, although the vast majority of known structures can be placed on the table in Fig. 5. In fig. S5, we have provided an expanded version of the periodic table, where complexes with up to 14 unique subunits and 48 subunit repeats can be visualized. We believe that the analogy to the periodic table of the elements is useful, because it provides a means of organizing quaternary structure topologies and visualizing similarities. Furthermore, just as the periodic table of the elements has successfully predicted many new chemical elements, our periodic table of proteins has considerable predictive power by revealing the regions of quaternary structure space that remain to be populated.

We showed above that the majority of nonbijective complexes are the result of quaternary structure assignment errors. The exception to this is complexes with uneven stoichiometry, most of which represent genuine cases of biological asymmetry. Therefore, we sought to reconcile uneven stoichiometry with our periodic table. We found that if we consider the periodic table at the level of local topological environments, rather than at the level of subunits, then two sequence-identical subunits can play different roles within the graph representing the complex. Examination of the topologies of nonbjiective complexes revealed that many of them were equivalent to the same symmetric homomer topologies observed for the bijective periodic table. Figure S6 illustrates this with a periodic table made for nonbijective heteromers with 2:1 subunit stoichiometry. For these cases, the 2:1 protomer can be considered analogous to a heterotrimer with three unique subunits. The only difference between 2:1 heteromers here and 1:1:1 heteromers from the main periodic table (the third row in Fig. 5) is that in the 2:1 heteromers, sequence-identical subunits sometimes can still form isologous interfaces, despite existing in different local environments. Thus, the results of our quaternary structure enumeration can be easily applied to complexes with uneven stoichiometry, if the repeated subunits from the protomer are considered to be different subunit types in our enumeration model.

Predicting likely yet unobserved quaternary structure topologies

The exhaustive enumeration allows us to determine what quaternary structure topologies are possible, but it does not tell us which are most likely or should be most abundant in nature. To address this, we adapted our enumeration procedure to produce topologies according to the observed distribution in the periodic table. We know that each cell on the periodic table can be defined by a specific set of assembly steps needed to build the topologies within that cell. Combining the steps in different ways can produce all the topologies compatible with a given cell. Therefore, we sampled cells of the periodic table according to the observed distribution in real complexes, each time randomly combining the assembly steps associated with each cell. This was repeated 3×10^7 times (details are provided in the Methods).

All of the quaternary structure topologies present on the periodic table were observed at least once in our calculations. In addition to the previously observed quaternary structure topologies, our model also predicted 579 topologies that were not seen in any of the complexes in our data set. To independently validate this result, we

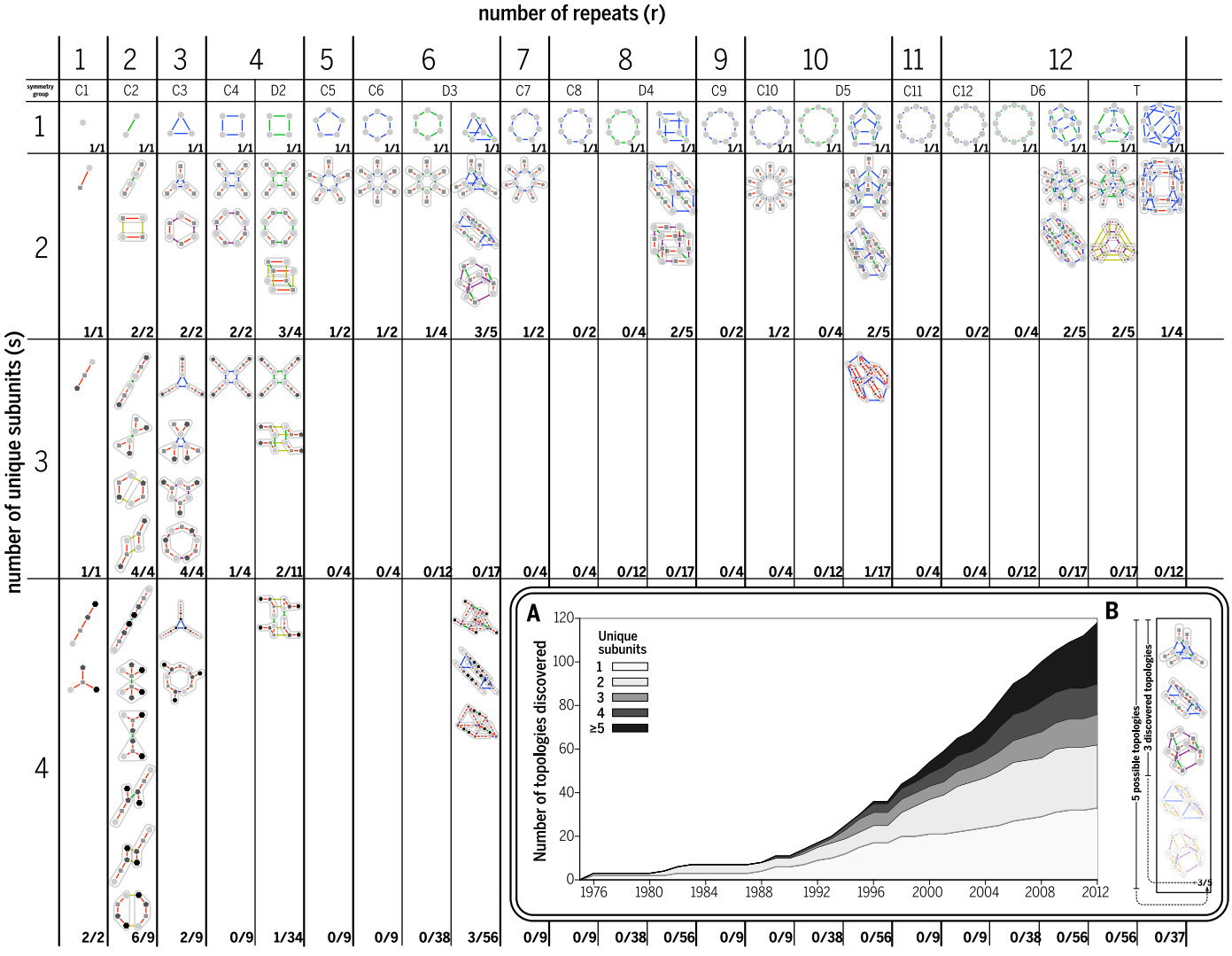


Fig. 5. Periodic table of protein complexes. All bijective protein complex topologies can be arranged according to the number of different subunit types (*s*) and the number of times these subunits are repeated (*r*). Isologous interfaces between the same subunits (dihedral interfaces) are shown in green, and heterologous interfaces between subunits of the same types (cyclic interfaces) are shown in blue. Heteromeric interfaces are shown in red, apart from those that correspond to a symmetric dimerization (yellow) or to higher-order cyclization (purple). The topologies in the $s = 1$ row are the equivalent homomers of the heteromeric structures in the $s > 1$ rows. To clarify this equivalence, subunits in the heteromers are grouped according to the repeated subcomplexes. In addition, the yellow and purple interfaces of the heteromeric complexes highlight interfaces that are dihedral (green) and cyclic (blue) in the equivalent homomers. The ratio in the bottom right of each cell indicates the number of topologies that have been observed and the total number of possible topologies of this type. The table shown here is an excerpt ($s < 5$; $r < 13$) of the full table. An interactive version of this table with information on the structures represented by each topology can be found at http://www.periodicproteincomplexes.org/. [Inset (A)] Number of discovered topologies as a function of time, which has been steadily increasing at a rate of about four topologies per year for the past two decades. [Inset (B)] An illustration of observed topologies versus all possible topologies with six repeats and two subunits ($r = 6$; $s = 2$). Three of the possible five topologies have been observed thus far.

compiled an extended set of heteromeric complexes not present in our original data set because they were published more recently, because they were determined with electron microscopy (which we did not initially include), or because they were originally excluded based on structural criteria (see Methods).

The extended set of heteromers contained 53 different quaternary structure topologies, 14 of which were not present in the main data set. These 14 tended to be among the most highly predicted topologies in our model. For example, six of them were observed among the top 20 most likely predicted topologies (Fig. 6), out of a total of 579 predicted ($P = 2 \times 10^{-6}$; Fisher's exact test). Figure S7 illustrates how the observed topologies cluster within the most highly predicted rankings, thus supporting the predictive utility of our model.

We also used a complementary approach for the prediction of the relative abundances of topologies within a given cell, which makes fewer assumptions but also yields less specific predictions. We considered the number of distributed interfaces (that is, single interfaces that are spread across two subunits) and the number of topological equivalents (marked by red crosses in fig. S8) of a given interface distribution. We compared topologies pairwise within cells of the periodic table with ≤4 unique subunits and ≤12 subunit repeats, and we counted the instances in which topology A had fewer distributed interfaces and equal or more topological equivalents than topology B, or fewer or equal distributed interfaces and more topological equivalents. Out of the 30 such instances, topology A was more abundant that topology B 21 times (70% of instances). This is because distributed interfaces restrict the order in which evolutionary steps can happen, making topologies

with more such interfaces rarer. Larger numbers of topological equivalents, on the other hand, make topologies more common, because there are more ways in which such complexes can evolve.

Finally, to further validate our predictive model, we compared the predicted frequencies of heteromeric quaternary structure topologies with those observed in both the main and extended data sets (fig. S9). Overall, the correlations are strong, with the predictions recapitulating the observed frequencies. Although the predictions are partially fitted to the frequencies of topologies observed in the main data set, the high correlation with the extended data set provides strong independent validation of our model.

Conclusion

In this study, we have shown that the assembly of protein complexes is dominated by three main transition types, which in combination can explain most observed quaternary structure topologies. This also allows a natural organization of protein complexes in a periodic table, in which heteromeric protein complexes are grouped according to their equivalent homomeric quaternary structure topologies. The periodic table illustrates both the variety of observed protein complexes and the space of possible topologies through exhaustive enumeration, analogous to previous strategies of investigating network topologies (*21*, *22*). Given that new topologies have been discovered at a fairly constant rate of four per year over the past two decades (Fig. 5A), we can expect new additions to the unfilled or partially filled cells of the periodic table in the near future. These unfilled or partially filled cells constrain the total space of expected protein complexes, similar to the proposed upper bound of 10,000 total types of interacting domain pairs (*23*).

A major practical application of the periodic table will be in predicting and modeling the quaternary structure of protein complexes. Specifically, our results show that bijective quaternary structure topologies are far more likely to occur than nonbijective topologies, despite the fact that there are far more possible nonbijective topologies. We also provide predictions for the relative likelihoods of different bijective topologies. This knowledge can inform the interpretation of high-throughput interaction experiments (*24*) or structure-based interaction predictions (*25*) by highlighting the quaternary structure topologies that are possible and most likely to occur. Homology information can aid these quaternary structure predictions and give further insight into the evolution and assembly of complexes, because subcomplexes often arise as evolutionary precursors and assembly intermediates (*12*). Similarly, the periodic table can tell us which evolutionary precursor topologies are likely to have given rise to a specific complex. The periodic table can also provide constraints for multi-subunit docking or modeling, both on the relative arrangements of subunits and on the overall complex symmetry (*26*–*29*). Such constraints could be further integrated into hybrid methods that combine different experimental measurements (*30*), such as electrospray (*31*) or cross-linking (*32*) mass spectrometry.

This work could be of substantial utility in the bioengineering of protein complexes. The self-assembly formalism introduced here allows specification of exactly which essential interfaces would need to be engineered to form a protein complex of a given topology. This could facilitate de novo engineering of oligomeric assemblies (*33*–*35*) and allow for directed modulation of existing quaternary structures, stepping either across or down the periodic table in an incremental manner.

Despite its strong predictive power, the basic periodic table model does not account for ~8% of known protein complex structures. More than half of these exceptions arise as a result of quaternary structure assignment errors. A benefit of this approach is that it highlights likely quaternary structure misassignments, particularly by identifying

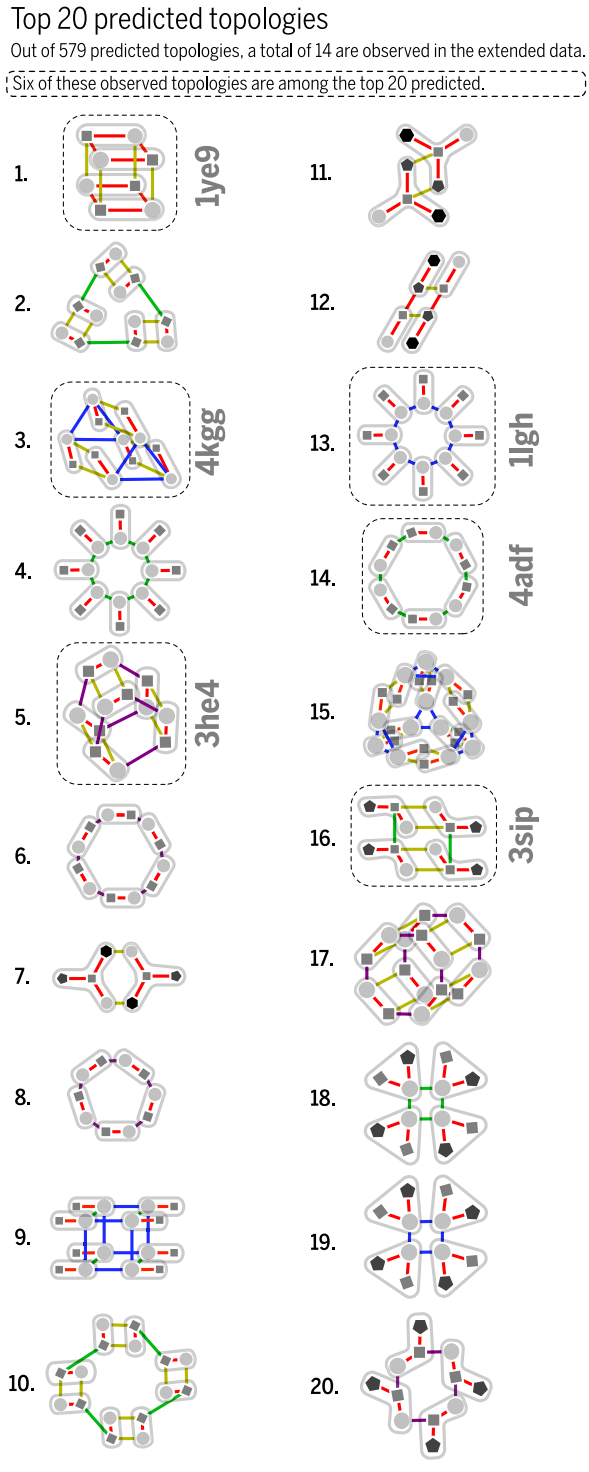


Fig. 6. The top 20 most likely quaternary structure topologies from our model that are not observed in the main data set. Of these top 20, six are observed in the extended data set, validating the power of the model ($P = 2 \times 10^{-6}$). The other 14 topologies in the top 20 are also expected to occur relatively frequently in nature and thus to be observed soon in experimentally determined structures. The distribution of all new topologies observed in the extended data set compared with the expected frequencies of all predicted topologies is shown in fig. S7.

nonbijective complexes with even subunit stoichiometry. However, this still leaves ~4% of known structures that are correct but are not compatible with the periodic table.

Although nonbijective complexes are possible, they are rare, and this should be given consideration in any protein modeling or engineering attempts. Related to this, it would be particularly interesting to see whether chaperones are more frequently involved in the assembly of nonbijective complexes to stabilize the required asymmetric transitions. To model the nonbijective protein complexes, additional assembly steps involving fractional and nonstoichiometric transitions are needed. However, as we showed, this would also greatly expand the number of possible quaternary structure topologies. Therefore, we consider the periodic table in its current implementation to be a reasonable compromise, allowing the vast majority of existing quaternary structure topologies to be explained and the most likely unobserved topologies to be predicted.

Methods

Mass spectrometry experiments

The complexes were kindly donated as follows: *Saccharomyces cerevisiae* SAGA deubiquitinating module [PDB accession number (ID), 3MHH; C. Wolberger, John Hopkins University School of Medicine]; *Thermus thermophilus* MglA/MglB complex (PDB ID, 3T1Q; A. Wittinghofer, Max Planck Institute for Molecular Physiology); *Geobacillus stearothermophilus* PDH E1 subunit (PDB ID, 3DVA; B. Luisi, University of Cambridge); *Streptococcus pyogenes* toxin-antitoxin complex (PDB ID, 3Q8X; A. Meinhart, Max Planck Institute for Medical Research); *Saccharomyces cerevisiae* phosphofructokinase (PDB ID, 3O8O; T. Schöneberg, University of Leipzig); *Ruegeria pomeroyi* propionyl-CoA carboxylase (PDB ID, 3N6R; L. Tong, Columbia University); *Homo sapiens* MSL1-MSL2 complex (PDB ID, 4B7Y; J. Kadlec, European Molecular Biology Laboratory); *Synechococcus elongatus* acetylglutamate kinase/PII complex (PDB ID, 2V5H; V.Rubio, Instituto de Biomedicina de Valencia); and *Pseudomonas aeruginosa* 3-methylcrotonyl-CoA carboxylase (PDB IDs, 3U9T and 3U9S; L. Tong, Columbia University).

The nano-electrospray ionization mass spectrometry experiments on complex disassembly and reassembly were performed as previously described using quadrupole time-of-flight mass spectrometers modified for high–mass/charge ratio operation (*11*). A list of all (sub)complexes observed under various experimental conditions is provided in table S1.

Protein complex data sets

The full set of protein x-ray crystal structures was taken from the PDB on 8 August 2012. Only polypeptide chains with at least 30 residues were considered. Backbone-only models were ignored, as were structures containing nucleic acids or >10% non-water heteroatoms. Heteromeric protein complexes formed by polypeptide cleavage were also ignored. In addition, this approach had the effect of removing complexes with protein chains that lack unique "db_id" sequence identifiers in the PDB. Complexes with >59 subunits or that were split over multiple PDB entries were excluded. In total, the final data set contained 30,469 monomers, 28,935 homomers, and 5543 heteromers. Manual quaternary structure assignments came from PiQSi entries with errors assigned as "probably yes" or "probably no" (*17*) and from additional manual searching of the literature.

The size of the interface between each pair of subunits from all protein complexes was calculated with AREAIMOL (*36*). For each complex, all interfaces of the same type were identified by calculating the correlation between atom-specific buried surface areas for each pair of interfaces. Only interfaces >200 $Å^2$ were considered. Two interfaces were considered to be of the same type if the Pearson correlation between the buried surface areas (in terms of equivalent amino acids) was >0.7. Interface sizes were averaged over all interfaces of the same type. Similarly, homomeric interfaces were classified as isologous if the correlation between the residue-specific buried surface area for each subunit was >0.7.

The extended set of quaternary structure topologies, used to validate our predictive model, was taken from a more recent set of protein complex structures from the PDB on 16 December 2014. In addition to the crystal structures used in the main set, electron microscopy structures were also considered here. The constraints on the main data set were loosened, so that complexes formed via cleavage, as well as complexes containing nucleic acids or other heteroatoms, were also included (although only protein chains were considered). In total, this extended set possessed 4214 bijective heteromers (with ≤4 unique subunits and ≤12 subunit repeats) not present in the main data set. Within this set, there were 53 different quaternary structure topologies, 14 of which are new.

For certain analyses, we used nonredundant subsets of the full and extended data sets. Proteins were filtered for redundancy at the 50% sequence identity level, essentially following a previous approach (*18*, *20*). The nonredundant sets were used for the histogram in fig. S4, for the comparison between observed and predicted results in fig. S9, and for fitting the expected distribution of periodic table quaternary structures in our model.

Both the main set and the extended set of quaternary structure topologies used in this study are provided in table S2, in the form of pairwise interfaces formed between subunits.

Determination of assembly pathways

All of the mass spectrometry and literature-identified (dis)assembly pathways involved protein complexes of known structure for which the subunit composition of at least one subcomplex intermediate could be identified. To complement this, we also performed a structural analysis in which we identified similar protein complexes with different quaternary structures. Starting from the full set of homomeric and heteromeric complexes in our main data set, we searched for complexes for which another complex could be considered as a subset of the full complex. For example, a homodimer was considered to be a subset of a homotetramer if the subunits shared >90% sequence identity. All of these subset complexes were considered to be similar to the experimentally identified subcomplexes for the purpose of defining assembly transitions. All of the experimentally identified subcomplexes and structural subsets are provided in table S3.

Although we observed the subcomplexes formed during (dis)assembly, we did not directly observe the assembly or disassembly steps that occur in solution. However, we inferred these from the subcomplexes identified. For every subcomplex and full complex, we identified the largest subcomplex that can be considered to be a subset of that (sub) complex. If no subcomplex was observed, then (dis)assembly was assumed to occur via free monomers. We then assigned the transition between the two states into one of the five categories from Fig. 2. All (dis)assembly transitions are provided in table S4.

Interface cutting procedure

To distinguish subunit interfaces that are circumstantial from those that are essential to the self-assembly process, we removed interfaces from the weighted subunit contact graph of a given complex, in increasing order of interface size. We skipped interfaces if cutting them would result in the complex becoming disconnected, and we stopped when no further interfaces could be cut.

This procedure does not necessarily result in a tree-like graph, because the same interface may appear several times in the same complex. For example, a cyclic ring will not be cut further, because all interfaces, which are of the same size, would have to be cut at once.

The advantages of this approach are that (i) it greatly simplifies the number of possible topologies and unifies similar topologies that would be treated differently if all interfaces were taken into account; (ii) it is less arbitrary than conventional thresholds of interface size, and therefore it is a more fundamental description of a complex; and (iii) symmetry emerges directly from the graph topology through the number of distinct interfaces of a subunit.

Topological enumeration procedure

The combinations of different interface types that can be present in a homomeric structure include one symmetric interface (C_2 symmetry), one asymmetric interface (cyclic symmetry with more than two subunits), one symmetric and one asymmetric interface (dihedral or tetrahedral symmetry), two symmetric interfaces (dihedral symmetry), and two asymmetric interfaces (only in tetrahedral symmetry). Larger numbers of interface types are not possible in homomers, because of the constraints placed on the symmetry of the complex by these interface types.

As explained above, any heteromer that is formed using a combination of cyclization, dimerization, and subunit addition can be represented as a homomer of multiple copies of the same heteromeric module, in which each subunit

type appears exactly once. We therefore enumerated all possible heteromeric topologies by carrying out the following procedure (illustrated in fig. S8).

1) We enumerated all trees of s unique subunits. These are the heteromeric modules, multiple copies of which are joined together into "homomers." The reason for restricting ourselves to trees instead of all possible graphs is that we aimed to consider only the most important interfaces in the original complexes, by using the interface cutting procedure outlined above. This will always lead to tree-like structures as the repeated heteromeric modules.

2) For each of the five possible combinations of interface types in homomers discussed above, we considered all topologically distinct ways in which the interfaces can be distributed across the set of subunits and pairs of subunits of the tree. Here, topologically distinct means that we cannot convert two distributions of interfaces into each other by only using symmetry operations of the tree.

3) We constructed the topologies of the complexes from these distributions of interfaces across the tree. For this, we also needed to consider the possible number of repetitions for each cyclic interface (for example, in the case of the ring-like complexes of varying size). In complexes with 12 or more subunits, there can be more than one pair of divisors with at least one even divisor (e.g., in the case of 12 subunits, three and four, and two and six), which leads to several possible topologies for the same total number of subunits (for example, D_6 and T in the r = 12 column of the s = 1 row of the periodic table).

4) We distinguished isomorphic topologies. In some cases, different distributions of interfaces in the enumeration procedure led to isomorphic topologies. These were easily identified using an isomorphism check, giving us the final enumeration of topologies.

The numbers in the bottom right of each cell in the periodic table (Fig. 5) give the observed and total numbers of different topologies for that particular symmetry group and number of subunits s.

Enumeration of all possible bijective and nonbijective topologies

We enumerated all possible bijective and nonbijective topologies with 2:2 stoichiometry, and all possible bijective and nonbijective topologies with 3:3 stoichiometry and up to six interfaces. We did this by considering all possible four-node graphs with three to six edges (2:2), shown in fig. S2, and all six-node graphs with five or six edges (3:3), shown in fig. S3. For these, we considered all distributions of equal numbers of two node colors (corresponding to the two protein species). In addition, for all edges between nodes of the same color, we considered two possibilities, corresponding to isologous and heterologous interfaces. Edges can thus have three colorings (heteromeric, homomeric-isologous, and homomeric-heterologous). We also considered different interface sizes by considering all possible relative size ranks (including equal ranks) of the different edges and subsequently cutting interfaces according to the same procedure followed in the periodic table. An isomorphism check (which includes the colorings of nodes and edges) was then used to identify distinct topologies in this enumeration. Two additional constraints were (i) that an equal interface size can appear more than once only for one type of subunit pair, and (ii) that the same interface size can only appear once on each subunit for homomeric-isologous and heteromeric interfaces and once or twice for homomeric-heterologous interfaces.

Prediction of expected frequencies of quaternary structure topologies

To predict the expected frequencies of quaternary structure topologies, we attempted to recapitulate the observed distribution of complexes within cells on the periodic table, considering complexes with ≤12 subunit repeats and ≤4 unique subunits. This prediction procedure can be divided into three parts.

1) Selecting a cell from the periodic table. We first randomly selected a structure from the nonredundant set of complexes and monomers. The row of the periodic table (the number of unique subunits s) was directly taken from this randomly selected structure. However, the column of the periodic table was not taken directly from this structure. This is because the sampling of cells on the periodic table is sparse for the lower rows, and thus we would have missed cells with no current structures. Instead, each structure was classified into one of three groups: first column (monomeric or no repeated subunits), cyclic (including C_2), or dihedral or tetrahedral. Then, another structure was randomly selected from those in the first row of the table (a monomer or homomer) that belong to the same group. This structure was used to define the column of the periodic table. Thus, the distribution of homomers defines the distribution of predicted heteromers in the horizontal axis, with a correction for the fact that complexes with more subunit repeats tend to be less likely to have cyclic or dihedral subunit repeats.

2) Defining the assembly steps. Each cell of the periodic table is associated with a specific set of subunit addition, dimerization, and/or cyclization assembly steps required to get to it from a monomer. Therefore, to generate a random quaternary structure topology that was compatible with a given cell, we first randomized the order of the assembly steps. There are two exceptions to this: (i) for dihdedral topologies where the homomer has at least six subunits and only isologous interfaces (e.g., a trimer of dimers rather than a dimer of trimers), the dimerization step must occur before the cyclization step; and ii) for tetrahedral complexes, the cyclic trimerization must occur before the tetramerization.

3) Constructing a quaternary structure topology, given a defined set of assembly steps. A new interface was added to the quaternary structure topology for each assembly step. When the subcomplex is heteromeric, there are multiple ways an interface could be formed. In these cases, the subunit(s) to be involved were randomly selected. For example, if a dimerization step was applied to an A-B subcomplex, then a new isologous interface could be formed between two A subunits or two B subunits, or a pair of identical heteromeric interfaces could be formed between each pair of A and B subunits. In the latter case, the isologous interface is distributed across two subunits. In fig. S10, we illustrate the random construction process of the quaternary structure topology for a single cell of the periodic table.

All predicted quaternary structure topologies are provided in table S5.

Enumeration-based shorthand notation of protein complex topologies

The representation of interface distributions across the subunits, as shown in Fig. 3 and fig. S9, allows for a natural shorthand notation of topologies. The subunits are labeled A, B, C, etc., and cyclic, dihedral, and heteromeric interfaces are denoted c, d, and h. The location of an interface follows the type, and in the case of distributed interfaces, both subunits are given. A cyclic homomer therefore is cA, the first cyclic structure with two subunits in Fig. 3 is cA hAB, and the last structure with three subunits in the same figure is dAC hAB hBC.

REFERENCES AND NOTES

1. D. S. Goodsell, A. J. Olson, Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153 (2000). doi: 10.1146/annurev.biophys.29.1.105; pmid: 10940245
2. J. Janin, R. P. Bahadur, P. Chakrabarti, Protein-protein interaction and quaternary structure. *Q. Rev. Biophys.* **41**, 133–180 (2008). doi: 10.1017/S0033583508004708; pmid: 18812015
3. J. A. Marsh, S. A. Teichmann, Structure, dynamics, assembly, and evolution of protein complexes. *Annu. Rev. Biochem.* **84**, 551–575 (2015). doi: 10.1146/annurev-biochem-060614-034142; pmid: 25494300
4. M. Levitt, C. Chothia, Structural patterns in globular proteins. *Nature* **261**, 552–558 (1976). doi: 10.1038/261552a0; pmid: 934293
5. C. A. Orengo, D. T. Jones, J. M. Thornton, Protein superfamilies and domain superfolds. *Nature* **372**, 631–634 (1994). doi: 10.1038/372631a0; pmid: 7990952
6. W. R. Taylor, A 'periodic table' for protein structures. *Nature* **416**, 657–660 (2002). doi: 10.1038/416657a; pmid: 11948354
7. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995). doi: 10.1016/S0022-2836(05)80134-2; pmid: 7723011
8. C. A. Orengo *et al.*, CATH—a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1109 (1997). doi: 10.1016/S0969-2126(97)00260-8; pmid: 9309224
9. E. D. Levy, J. B. Pereira-Leal, C. Chothia, S. A. Teichmann, 3D Complex: A structural classification of protein complexes. *PLOS Comput. Biol.* **2**, e155 (2006). doi: 10.1371/journal.pcbi.0020155; pmid: 17112313
10. E. D. Levy, E. Boeri Erba, C. V. Robinson, S. A. Teichmann, Assembly reflects evolution of protein complexes. *Nature* **453**, 1262–1265 (2008). doi: 10.1038/nature06942; pmid: 18563089
11. J. A. Marsh *et al.*, Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell* **153**, 461–470 (2013). doi: 10.1016/j.cell.2013.02.044; pmid: 23582331
12. J. A. Marsh, S. A. Teichmann, Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays* **36**, 209–218 (2014). doi: 10.1002/bies.201300134; pmid: 24272815
13. Z. Hall, A. Politis, C. V. Robinson, Structural modeling of heteromeric protein complexes from disassembly pathways and ion mobility-mass spectrometry. *Structure* **20**, 1596–1609 (2012). doi: 10.1016/j.str.2012.07.001; pmid: 22841294
14. J. Monod, J. Wyman, J.-P. Changeux, On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12**, 88–118

(1965). doi: 10.1016/S0022-2836(65)80285-6; pmid: 14343300
15. K. Henrick, J. M. Thornton, PQS: A protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358–361 (1998). doi: 10.1016/S0968-0004(98)01253-5; pmid: 9787643
16. E. Krissinel, K. Henrick, Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007). doi: 10.1016/j.jmb.2007.05.022; pmid: 17681537
17. E. D. Levy, PiQSi: Protein quaternary structure investigation. *Structure* **15**, 1364–1367 (2007). doi: 10.1016/j.str.2007.09.019; pmid: 17997962
18. J. A. Marsh, H. A. Rees, S. E. Ahnert, S. A. Teichmann, Structural and evolutionary versatility in protein complexes with uneven stoichiometry. *Nat. Commun.* **6**, 6394 (2015). doi: 10.1038/ncomms7394; pmid: 25775164
19. T. Perica *et al*., The emergence of protein complexes: Quaternary structure, dynamics and allostery. *Biochem. Soc. Trans.* **40**, 475–491 (2012). doi: 10.1042/BST20120056; pmid: 22616857
20. J. A. Marsh, S. A. Teichmann, Protein flexibility facilitates quaternary structure assembly and evolution. *PLOS Biol.* **12**, e1001870 (2014). doi: 10.1371/journal.pbio.1001870; pmid: 24866000
21. S. S. Shen-Orr, R. Milo, S. Mangan, U. Alon, Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**, 64–68 (2002). doi: 10.1038/ng881; pmid: 11967538
22. W. Ma, A. Trusina, H. El-Samad, W. A. Lim, C. Tang, Defining network topologies that can achieve biochemical adaptation. *Cell* **138**, 760–773 (2009). doi: 10.1016/j.cell.2009.06.013; pmid: 19703401
23. P. Aloy, R. B. Russell, Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.* **22**, 1317–1321 (2004). doi: 10.1038/nbt1018; pmid: 15470473
24. P. C. Havugimana *et al*., A census of human soluble protein complexes. *Cell* **150**, 1068–1081 (2012). doi: 10.1016/j.cell.2012.08.011; pmid: 22939629
25. Q. C. Zhang *et al*., Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**, 556–560 (2012). doi: 10.1038/nature11503; pmid: 23023127
26. Y. Inbar, H. Benyamini, R. Nussinov, H. J. Wolfson, Prediction of multimolecular assemblies by multiple docking. *J. Mol. Biol.* **349**, 435–447 (2005). doi: 10.1016/j.jmb.2005.03.039; pmid: 15890207
27. F. DiMaio, A. Leaver-Fay, P. Bradley, D. Baker, I. André, Modeling symmetric macromolecular structures in Rosetta3. *PLOS ONE* **6**, e20450 (2011). doi: 10.1371/journal.pone.0020450; pmid: 21731614
28. J. Esquivel-Rodríguez, D. Kihara, Evaluation of multiple protein docking structures using correctly predicted pairwise subunits. *BMC Bioinformatics* **13**, S6 (2012). doi: 10.1186/1471-2105-13-S2-S6; pmid: 22536869
29. B. G. Pierce *et al*., ZDOCK server: Interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* **30**, 1771–1773 (2014). doi: 10.1093/bioinformatics/btu097; pmid: 24532726
30. F. Alber, F. Förster, D. Korkin, M. Topf, A. Sali, Integrating diverse data for structure determination of macromolecular assemblies. *Annu. Rev. Biochem.* **77**, 443–477 (2008). doi: 10.1146/annurev.biochem.77.060407.135530; pmid: 18318657
31. H. Hernández, C. V. Robinson, Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nat. Protoc.* **2**, 715–726 (2007). doi: 10.1038/nprot.2007.73; pmid: 17406634
32. J. Rappsilber, The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J. Struct. Biol.* **173**, 530–540 (2011). doi: 10.1016/j.jsb.2010.10.014; pmid: 21029779
33. N. P. King *et al*., Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **336**, 1171–1174 (2012). doi: 10.1126/science.1219364; pmid: 22654060
34. Y.-T. Lai, D. Cascio, T. O. Yeates, Structure of a 16-nm cage designed by using protein oligomers. *Science* **336**, 1129 (2012). doi: 10.1126/science.1219351; pmid: 22654051
35. J. Zhang, F. Zheng, G. Grigoryan, Design and designability of protein-based assemblies. *Curr. Opin. Struct. Biol.* **27**, 79–86 (2014). doi: 10.1016/j.sbi.2014.05.009; pmid: 24952313
36. M. D. Winn *et al*., Overview of the *CCP*4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011). doi: 10.1107/S0907444910045749; pmid: 21460441

ACKNOWLEDGMENTS

We thank C. Wolberger, A. Wittinghofer, B. Luisi, A. Meinhart, T. Schöneberg, L. Tong, J. Kadlec, and V. Rubio for providing protein complex samples; H. Rees for assistance with manual quaternary structure assignments and identification of literature-derived assembly pathways; and P. Beltrao, S. Edelstein, T. Flock, D. Gfeller, M. Hein, F. Krueger, R. Laskowski, E. Levy, S. MacKinnon, I. Moal, E. Natan, T. Perica, B. Stauch, S. Velankar, S. Wodak, and X. Zhang for helpful discussions and comments on the manuscript. This work was supported by the Royal Society (S.E.A. and C.V.R.), the Human Frontier Science Program (J.A.M.), the Medical Research Council (grant G1000819 to H.H. and C.V.R.), and the Lister Institute for Preventative Medicine (S.A.T.).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/350/6266/aaa2245/suppl/DC1
Figs. S1 to S10
Tables S1 to S5

17 March 2015; accepted 29 October 2015
10.1126/science.aaa2245

Principles of assembly reveal a periodic table of protein complexes

Sebastian E. Ahnert, Joseph A. Marsh, Helena Hernández, Carol V. Robinson and Sarah A. Teichmann (December 10, 2015)

Science **350** (6266), . [doi: 10.1126/science.aaa2245]

Editor's Summary

The principles of protein assembly

A knowledge of protein structure greatly enhances our understanding of protein function. In many cases, function depends on oligomerization. Ahnert *et al.* used mass spectrometry data together with a large-scale analysis of structures of protein complexes to examine the fundamental steps of protein assembly. Systematically combining assembly steps revealed a large set of quaternary topologies that were organized into a periodic table. Based on this table, the authors accurately predicted the expected frequencies of quaternary structure topologies.

Science, this issue p. 10.1126/science.aaa2245

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.