Protein complexes are under evolutionary selection to assemble via ordered pathways

Citation for published version:

Digital Object Identifier (DOI):
10.1016/j.cell.2013.02.044

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Cell

Publisher Rights Statement:
Open Access funded by Medical Research Council
License: http://www.elsevier.com/open-access/userlicense/1.0/

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
SUMMARY

Is the order in which proteins assemble into complexes important for biological function? Here, we seek to address this by searching for evidence of evolutionary selection for ordered protein complex assembly. First, we experimentally characterize the assembly pathways of several heteromeric complexes and show that they can be simply predicted from their three-dimensional structures. Then, by mapping gene fusion events identified from fully sequenced genomes onto protein complex assembly pathways, we demonstrate evolutionary selection for conservation of assembly order. Furthermore, using structural and high-throughput interaction data, we show that fusion tends to optimize assembly by simplifying protein complex topologies. Finally, we observe protein structural constraints on the gene order of fusion that impact the potential for fusion to affect assembly. Together, these results reveal the intimate relationships among protein assembly, quaternary structure, and evolution and demonstrate on a genome-wide scale the biological importance of ordered assembly pathways.

INTRODUCTION

In order to function, most proteins assemble into complexes—either homomers, comprised of self-interacting copies of a single type of subunit, or heteromers, composed of two or more distinct polypeptide chains. Is the order in which protein subunits associate important for the formation and biological function of the final complex? Although protein interactions have been studied extensively (Janin et al., 2007; Shoemaker and Panchenko, 2007) and the misassembly of proteins can have severe biological consequences (Dobson, 2003; Ellis, 2007), the multistep process by which proteins assemble into complexes has received comparatively little attention in recent years. By analogy to Levinthal’s paradox of protein folding (Levinthal, 1969), we can presume that assembly must proceed via energetically favorable intermediate subcomplexes, lest the time required for productive multisubunit complex formation be prohibitively long. Thus, just as proteins preferentially fold via a limited number of energetically favorable folding pathways (Lindorff-Larsen et al., 2011), protein complexes should be expected to assemble via ordered assembly pathways.

Ordered assembly has now been observed experimentally for a number of systems. Classic studies used a variety of techniques to characterize putative assembly intermediates, which in combination with kinetic measurements, allowed the assembly of various homomeric and heteromeric complexes to be characterized (Friedman and Beychok, 1979). In addition, ordered assembly has been seen in larger multisubunit complexes such as the spliceosomal snRNP core (Raker et al., 1996), the preinitiation transcription complex (Baldick et al., 1994), and the 26S proteasome (Gallastegui and Groll, 2010). In recent years, electrospray mass spectrometry (MS) has emerged as an extremely useful method for studying assembly, having the distinct advantage of being able to probe the oligomeric states of multiple subcomplex intermediates simultaneously, thus allowing in vitro ordered assembly pathways to be elucidated in detail (Sobott et al., 2002; Hernández and Robinson, 2007; Levy et al., 2008).

A powerful way to demonstrate the importance of assembly order would be to test whether assembly pathways have been conserved in evolution. A large-scale analysis of simple homomeric complexes suggested that the order of self-assembly for identical subunits recapitulates quaternary structure evolution and is generally conserved (Levy et al., 2008). However, in heteromers, which account for most in vivo protein complexes (Kühner et al., 2009), the relationship between assembly and evolution has not been investigated. Since there are far fewer published structures for heteromers than for homomers (Perica et al., 2012), it is difficult to employ a similar strategy. Fortunately,
however, we have identified a unique evolutionary phenomenon that allows us to test whether heteromer assembly pathways have been conserved: gene fusion.

Gene fusion occurs when two previously distinct genes become fused into a single open reading frame. A considerable number of studies have focused on understanding gene fusion as an evolutionary mechanism at the DNA sequence and protein domain levels. In fact, evolutionary reconstructions suggest that gene fusion is the most common mechanism by which multidomain proteins acquire new domains in both bacteria and higher eukaryotes (Björklund et al., 2005; Pasek et al., 2006; Buljan et al., 2010). Gene fusion has received extensive attention since it was shown that evolutionary fusion events could be used to predict protein interactions on a genomic scale (Enright et al., 1999; Marcotte et al., 1999a, 1999b). Essentially, the idea is that proteins that are encoded by different genes in one organism but fused together in another are very likely to physically interact, or at least be functionally related, when expressed as separate gene products. This has been supported by comprehensive analyses (Enright and Ouzounis, 2001; Yanai et al., 2001; Marcotte and Marcotte, 2002; Kamburov et al., 2007; Reid et al., 2010).

Because gene fusion forces the permanent, covalent association of two protein subunits, it provides a mechanism by which protein complex assembly pathways can be either conserved or modified in evolution. As illustrated in Figure 1, a fusion event can be compatible with and conserve the existing assembly pathway if it mimics the first step of assembly. Alternatively, a fusion-induced linkage can disrupt the order of assembly. Therefore, if careful examination of the evolutionary record were to reveal a significant tendency for gene fusion events that conserve rather than modify existing protein-complex assembly pathways, this would strongly support the importance of ordered assembly for the formation of functional protein complexes.

Here, we exploit the large number of fully sequenced genomes and protein complex structures that are now available in order to identify evolutionary gene fusion events that have occurred between genes encoding the subunits of heteromeric complexes. First, by experimentally determining the assembly pathways of several of these complexes, we show that assembly can be predicted on a large scale from crystal structures. This allows us to demonstrate significant evolutionary selection for gene fusion events that conserve the existing order of subunit assembly. In addition, we observe a tendency for fusion to optimize assembly by maximally reducing the interfaces in protein complexes and discrete interactions in protein interaction networks. Finally, we show protein structural constraints on the gene order of fusion, which arise from a preference for optimally positioned N and C termini and influence the potential for fusion to affect assembly. Overall, these results demonstrate the role of protein complex assembly in evolution and provide fundamental insight into the biophysics and biological importance of ordered assembly pathways.

RESULTS

Prediction of Heteromer Assembly Pathways and Characterization by Nanoelectrospray Ionization MS

We first searched the Protein Data Bank (Berman et al., 2000) for heteromeric complexes for which there is genomic evidence of fusion occurring between subunits in the STRING database (Szklarczyk et al., 2011). In each of these complexes, a pair of subunits is encoded by two separate genes that are known to become fused in another species. We refer to these as “prefusion” complexes because they are likely to be similar to the ancestral complexes that existed prior to the evolutionary gene fusion event. In total, we identified 94 nonredundant pairs of heteromeric subunits associated with fusion events (Table S1). Thus, if we knew the assembly pathways of these complexes, we could assess whether the evolutionary fusion events were compatible with the existing order of assembly and would have conserved that order.

Previously, we showed that one can predict the assembly of homomeric complexes by invoking a simple model in which the strength of each interface is assumed to be proportional to the surface area buried between the two subunits, as calculated from the crystal structure (Levy et al., 2008). However, we were uncertain whether a similar phenomenon would hold true for heteromeric complexes, especially considering that interface size generally shows weak correlation with binding affinity in heteromers (Brooijmans et al., 2002), and that heteromeric subunits are...
often more flexible in isolation and tend to undergo larger conformational changes upon binding (Marsh and Teichmann, 2011; Marsh et al., 2012). Furthermore, the presence of multiple distinct subunits means that heteromers have far more potential routes of assembly, which could complicate predictions.

To test the association between interface size and assembly, we performed nanoelectrospray ionization (nESI)-MS experiments (Sobott et al., 2002; Hernández and Robinson, 2007) on five of the prefusion complexes identified above in order to determine their reversible in vitro disassembly pathways. Representative mass spectra are shown in Figures 2A and S1. Although the process of disassembly is different from that of assembly, the two processes are generally reversible in homomeric complexes (Levy et al., 2008). To further support this notion, we show that the prefusion complexes studied here can be reassembled from their dissociated states without the formation of off-pathway subcomplexes, thus demonstrating the reversibility of assembly and disassembly in heteromers (Figure S2). Therefore, we refer to “(dis)assembly” as this reversible process we can probe in solution.

In addition to the MS experiments, we also identified four prefusion complexes in which (dis)assembly pathways could be

---

**Figure 2. Experimentally Characterized (Dis)Assembly Pathways of Heteromeric Prefusion Complexes**

(A) (Dis)assembly pathways of complexes characterized by nESI-MS as well as representative mass spectra. See Table S2 for a full list of subcomplexes identified under different solution conditions.

(B) (Dis)assembly pathways of complexes identified from previously published experiments. In the graph representations of protein complexes, interfaces that undergo fusion are shown in orange.

See also Figure S1.
not there was actually any genomic evidence for fusion occurring complexes, we assessed the effects of a hypothetical fusion each heteromeric pair of subunits in a large set of nonredundant either conserving or modifying (dis)assembly pathways. For gene fusion events associated with prefusion complexes. can then investigate in detail the tendency for assembly to be large scale for all protein complexes of known structure. We structures enables us to simulate (dis)assembly pathways on a The ability to confidently predict (dis)assembly from crystal inferred from previously published literature (Evans et al., 1974; Poulsen et al., 1993; Payne et al., 1997; Durand and Merrick, 2006). The full (dis)assembly pathways for all nine complexes are shown in Figure 2 and detailed descriptions are provided in the Extended Experimental Procedures. We found excellent agreement between interface sizes and (dis)assembly, with seven out of nine complexes (23/27 total steps) agreeing perfectly with predictions (Table 1). This strongly demonstrates that the (dis)assembly of both homomeric and heteromeric complexes is primarily determined by the sizes of their interfaces and can therefore be easily predicted.

It is interesting to note the two complexes that show some deviations from the assembly predictions. These are both related urease complexes, representing two separate fusion events. In each case, the first few (dis)assembly steps proceed exactly as predicted, followed by a split into parallel pathways that is not predicted. We hypothesize that, for these large complexes, the loss of some subunits may lead to tertiary and/or quaternary structural rearrangements, which could change the relative interface sizes. Thus, the interface model might still hold in these cases, if only we knew the conformational rearrangements that occur upon subunit loss.

### Evolutionary Selection for Conservation of Protein Complex Assembly Pathways upon Gene Fusion

The ability to confidently predict (dis)assembly from crystal structures enables us to simulate (dis)assembly pathways on a large scale for all protein complexes of known structure. We can then investigate in detail the tendency for assembly to be conserved or modified by the 94 nonredundant evolutionary gene fusion events associated with prefusion complexes.

We first considered the intrinsic likelihood of subunit fusions either conserving or modifying (dis)assembly pathways. For each heteromeric pair of subunits in a large set of nonredundant complexes, we assessed the effects of a hypothetical fusion event on the (dis)assembly pathway, regardless of whether or not there was actually any genomic evidence for fusion occurring between them. Of the 1,487 hypothetical fusions that could occur between nonredundant subunit pairs, only 201 (13.5%) would conserve (dis)assembly, and the remainder would disrupt existing (dis)assembly pathways (Figure 3A). Thus, we can immediately see that if fusion were to occur randomly between the subunits of heteromeric complexes (i.e., without evolutionary selection), assembly-conserving fusion events would be quite rare.

Next we looked at how frequently actual evolutionary gene fusion events have occurred in these two groups. Whereas 24/201 (11.9%) subunit pairs that would conserve (dis)assembly pathways actually have evolutionary evidence for fusion occurring between them in some other species, this is true for only 48/1,286 (3.7%) pairs that would modify (dis)assembly (p = 8 × 10⁻⁶, Fisher’s exact test; Figure 3A). Thus, although the large majority of heteromeric subunit pairs show no evidence of fusion, a fusion event is far more likely to occur if it conserves the existing assembly pathway.

An alternate means of testing for assembly conservation is to compare the frequency with which (dis)assembly pathways are conserved in our set of evolutionary gene fusion events with the frequency we would expect based upon the intrinsic topologies of the complexes. We implemented a simple null model in which the quaternary structure topology of each complex was retained but random weights were assigned to each unique interface type. We then predicted the (dis)assembly pathway for each randomly reweighted complex and assessed the conservation frequency, and repeated this process many times in order to calculate the intrinsic probability of assembly conservation. The observed frequency with which real evolutionary gene fusion events conserve (dis)assembly is 33.3% (24/72), which is nearly double the intrinsic expectation for complexes with the same topologies according to this model (17.3%, p = 1 × 10⁻⁶; Figure 3B). In fact, a marginal level of significance is retained even when only the nine experimentally characterized complexes are considered (44.4% [4/9] conserved versus 19.5% expected [p = 0.05]).

Finally, to investigate the evolutionary selection for assembly-conserving gene fusion events more directly, we considered only heteromeric complexes with more than two unique subunits. In these complexes, multiple fusion events are hypothetically possible, which allows us to assess the probability of assembly conservation if fusion occurred randomly (e.g., for a complex with three unique subunits, as in Figure 1, each fusion would have a one in three chance of occurring). We observe that 38.9% (14/36) evolutionary fusion events in these complexes conserve (dis)assembly, compared with only 14.9% expected if fusions had randomly occurred within the same complexes (p = 7 × 10⁻⁵; Figure 3C). Therefore, given the set of fusion events that are hypothetically possible within a heteromeric complex, evolution appears to have strongly preferred those that mimic and thus conserve existing assembly pathways.

Above we have shown that (dis)assembly in heteromers is primarily driven by the sizes of the intersubunit interfaces. Large interfaces have been noted as characteristic of obligate complexes, in which the subunits are permanently associated within the cell (Nooren and Thornton, 2003). In Figure S3 and the Extended Experimental Procedures, we present multiple

<table>
<thead>
<tr>
<th>Complex Name</th>
<th>PDB ID</th>
<th>Correctly Predicted Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbamoyl phosphate synthase</td>
<td>1BXR</td>
<td>2/2</td>
</tr>
<tr>
<td>Tryptophan synthase</td>
<td>1WBJ</td>
<td>2/2</td>
</tr>
<tr>
<td>Acetyl-CoA carboxylase</td>
<td>2F9Y</td>
<td>1/1</td>
</tr>
<tr>
<td>Klebsiella aerogenes urease</td>
<td>1KRA</td>
<td>4/6</td>
</tr>
<tr>
<td>Helicobacter mustelae urease</td>
<td>3QGA</td>
<td>9/11</td>
</tr>
<tr>
<td>Nitrile hydratase</td>
<td>3QXE</td>
<td>1/1</td>
</tr>
<tr>
<td>AmtB-GlnK</td>
<td>2NS1</td>
<td>1/1</td>
</tr>
<tr>
<td>Aspartate transcarbamoylase</td>
<td>1D09</td>
<td>2/2</td>
</tr>
<tr>
<td>Anthranilate synthase</td>
<td>2NS1</td>
<td>1/1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>23/27</td>
</tr>
</tbody>
</table>

See also Figure S2.
lines of evidence that fusion occurs preferentially in obligate complexes, including a lower tendency for fusing subunits to be observed in isolation and a much higher propensity for correlated messenger RNA (mRNA) expression. Importantly, we show that the observed assembly conservation does not arise from a tendency for fusion to occur in obligate complexes.

Taken together, our results provide robust evidence of evolutionary selection for assembly-conserving gene fusion events. Importantly, we emphasize that this is not an absolute rule, and that a slight majority of fusions do in fact disrupt assembly. However, one must consider that random subunit fusions would conserve (dis)assembly in only a very small fraction of cases and thus the evolutionary frequency of (dis)assembly-conserving fusions is far higher than would be expected by chance.

Optimization of Assembly upon Fusion through Simplification of Protein Complex Topologies

Despite the strong selection for assembly conservation, it is clear that many evolutionary fusion events have modulated existing assembly pathways. Thus, we hypothesized that there may have been further evolutionary selection for fusion events that optimize assembly. For instance, although any fusion event between subunits will reduce the number of assembly steps by at least one, greater simplification will occur if the fusion involves two subunits that both share other interaction partners, as this will result in fewer intermolecular interfaces in the fused complex (Figure 4A).

We compared the reduction of intersubunit interfaces in protein complexes upon fusion with what would be expected if fusion occurred randomly between subunits (essentially as in Figure 3C). Interestingly, we observed that gene fusion events tended to reduce the number of interfaces by considerably more than would be expected by chance (2.90 versus 2.21, \( p = 1 \times 10^{-4} \); Figure 4B). This strongly implies evolutionary selection for fusions that maximally reduce the number of interfaces in a protein complex, thereby simplifying their topologies and assembly pathways. We suggest that having fewer intersubunit interfaces would both lower the risk of misassembly and increase the speed of assembly.

We investigated this phenomenon further by searching high-throughput interaction data for interacting proteins with evidence of fusion occurring between them. Each binding partner shared by a pair of proteins will further reduce the number of distinct protein-protein interactions by one upon fusion (Figure 4C). Pairs of proteins from Escherichia coli that undergo fusion share a mean of 19.2% of their binding partners, compared with 13.2% expected for random fusions within the interaction network (\( p = 3 \times 10^{-4} \); Figure 4D). Similar trends are also seen in yeast (14.7% versus 7.1%, \( p = 0.008 \)), humans (23.2% versus 16.4%, \( p = 0.04 \)), and a large number of other species (Table S3). Contrary to our structure-based analysis, if two proteins share a binding partner in these high-throughput data, it does not necessarily mean that they are interacting simultaneously (Kim et al., 2006a). Nevertheless, these results imply evolutionary selection for fusion events that optimize network topology by reducing the number of discrete protein interactions, in analogy to the simplification of assembly.

Protein Structural Constraints on Fusion

Because gene fusion essentially forces a pair of proteins to interact permanently with each other, the influence of fusion on assembly may be limited by protein structural constraints dictating whether or not a fusion event is likely to occur. Upon fusion of two proteins, the C terminus of the first will become covalently linked to the N terminus of the second. If these termini are far apart in the prefusion complex, fusion would require either the addition of a lengthy linker or a major disruption of the intersubunit interface. However, if these termini are close in space, fusion would be more likely to conserve the existing quaternary structure (Figure 5A).
To illustrate this, we consider the case of the prefusion complex *Klebsiella aerogenes* urease (Jabri and Karplus, 1996), where fusion is known to occur between genes corresponding to the γ and β subunits. Because the γ subunit fuses upstream of the β subunit, fusion will result in a linkage between the C terminus of the γ subunit and the N terminus of the β subunit. Examination of the complex crystal structure reveals that these termini are in fact quite close, separated by only 16 Å (Figure 5B). We will refer to this as the “fusion distance.” The “reverse distance” (if fusion were to occur in the opposite gene order [i.e., β upstream of γ]) is much greater (66 Å).

We systematically compared the fusion and reverse distances of all prefusion complexes in our data set in which the subunits correspond closely to the full-length genes (Figure 5C). We observe that for cases in which fusion has occurred in only a single gene order, the fusion distances are shorter than the reverse distances in 35/47 (74.5%) fusion events (p = 0.001, binomial test). Furthermore, the mean fusion distance is 14.1 Å shorter than the mean reverse distance (p = 0.001, Wilcoxon signed-rank test). Importantly, this tendency for fusion to occur between the closer termini is not related to the (dis)assembly conservation demonstrated earlier (see Extended Experimental Procedures). Therefore, the order of gene fusion is closely related to the structure of protein complexes, with significant evolutionary selection for fusion events that link more proximal termini. This is consistent with a previous study in which pairs of domains that were observed to interact both inter- and intramolecularly, which included several fusions, were shown to conserve their binding orientations in most cases (Kim et al., 2006b).

**DISCUSSION**

By comparing the identities of assembly intermediates observed in nESI-MS experiments with the structures of protein complexes, we were able to gain a fundamental mechanistic insight into protein assembly. Essentially, assembly in both homomeric and heteromeric complexes is driven by the hierarchy of interface sizes within a protein complex, such that assembly intermediates will tend to possess larger intersubunit interfaces. By taking advantage of Nature’s grand protein engineering experiment, i.e., the large number of gene fusion events that have occurred throughout evolutionary history, we show that these assembly intermediates are under evolutionary selection. This suggests that modifying existing assembly pathways has a significant tendency to lower an organism’s evolutionary fitness.

Although numerous functional benefits arise from the formation of multisubunit complexes, the increased complexity is associated with a greater risk of misassembly. Our results suggest that evolution has selected for protein complexes that assemble via well-defined, ordered pathways. Presumably, this leads to faster and more efficient formation of the functional complexes. If these assembly pathways become modified in evolution, the identities of the assembly intermediates will change, potentially increasing their susceptibility to misassembly or aggregation. Thus, the evolutionary conservation and optimization of assembly pathways revealed here provide a potential means of minimizing these risks while maintaining the advantages of complex formation. Furthermore, our results have practical implications in that the identities of assembly intermediates can now be predicted from the three-dimensional structures of protein complexes. This may provide clues as to how misassembly occurs and how it might be prevented.

The assembly and quaternary structure of protein complexes are highly important for determining which gene fusion events...
are selected. Since the vast majority of hypothetical fusion events would modify existing assembly pathways, this helps to rationalize why most protein interactions are not predicted by fusion-based methods (e.g., only 3.7% of the nonredundant subunit pairs in our data set are associated with evolutionary fusion events). In addition, we demonstrated further selective pressure upon fusion related to assembly optimization and the requirement for covalent linkage of termini.

These findings provide a more detailed, structural understanding of fusion that should allow one to better interpret and utilize fusion-based predictions. Furthermore, fusion-based strategies have been gaining prominence in the field of protein engineering (Padilla et al., 2001; Sinclair et al., 2011; Lai et al., 2012). Our insights can also potentially guide future protein engineering approaches: if covalent fusion of subunits is desired in order to stabilize a complex, success is most likely to be achieved with engineered fusions that conserve existing assembly pathways and in which the gene order is chosen to best match the existing quaternary structure.

This work also reveals an evolutionary connection between protein and genome structure. In 13% of the cases we examined, fusion occurred in both orders (i.e., AB and BA), in similarity to previous work showing that the vast majority (~92%) of domain pairs occur in only a single order (Apic et al., 2001). It has been suggested that the order of domain combinations in multidomain proteins is due primarily to historical chance, as domain pairs with the same structure and function can occur in both orders given the presence of a long interdomain linker (Bashton and Chothia, 2002; Vogel et al., 2004). Thus, multidomain proteins are highly versatile and a short interterminal fusion distance is not a strict requirement. However, our results suggest that the formation of a long linker (as required to preserve the quaternary interaction) can be a limiting factor, because we observe a strong preference for fusions in the order corresponding to the shorter interterminal distance. Therefore, our work implies that, rather than being an evolutionary artifact, the order in which genes fuse can be directly related to the structural features of the proteins they encode, thus demonstrating a simple way in which protein structure can influence genomic organization.

Finally, our results highlight a fascinating connection between evolutionary processes, which act over millions of years, and assembly, which occurs on the order of seconds. Although the assembly pathways of homomeric complexes were previously found to reflect their evolutionary histories (Levy et al., 2008), here we observed an opposite phenomenon in which the evolutionary process of gene fusion mimics heteromer assembly in order to conserve the existing assembly pathway.

EXPERIMENTAL PROCEDURES

Structural Data Sets

We started with the full set of heteromeric biological units from protein crystal structures in the RCSB Protein Data Bank (Berman et al., 2000). We filtered heteromers formed by polypeptide cleavage by identifying different chains with the same external database reference identifier (db_id, which generally corresponds to the UniProt sequence) but with a sequence identity of <90%. Only subunits with at least 50 residues were considered. Protein complexes containing nucleic acids were ignored because we have no way of reliably predicting (dis)assembly for these cases.

We filtered subunit pairs from the protein complexes for redundancy, first by grouping them by their SUPERFAMILY domain assignments (Gough et al., 2001) and then by calculating the sequence identities between all pairs in each group. If both subunits from a pair had >70% sequence identity to another pair, only the pair from the higher-resolution crystal structure was kept. After the sequence redundancy filtering was completed, we had a total of 2,544 nonredundant heteromeric subunit pairs. All subunit pairs used in this study, along with their various relevant properties, are provided in Table S1.

For each complex, we calculated the size of the interfaces between all pairs of subunits using AREAIMOL (Collaborative Computational Project, Number 4, 1994). In complexes containing more than one copy of each subunit, there can be more than one interface for a given pair of subunit types (e.g., the two different α-β interfaces in 2F9Y; see Figure 2A). Therefore, in compiling our nonredundant set of subunit pairs, we only considered the largest interface.
for a given pair of subunit types from each complex (e.g., only the largest α-β
interface in 2F9Y). Pairs of subunits were considered to be directly interacting if
they buried >200 Å² of intermolecular interface area.

For each pair of subunits, we searched the STRING v9.0 database (Szklarczyk et al., 2011) for evidence of fusion occurring between the genes encoding
those subunits. This was defined as two proteins with a STRING fusion evi-
dence score > 0.3, and each having >50% sequence identity to one of the in-
teracting subunits. Note that STRING uses stringent criteria for identifying
gene fusion events based upon orthology to nonfused genes, thus avoiding
the requirement to filter putative fusion events involving promiscuous domains,
as arises with homology-based approaches (Marcotte et al., 1999a). The sig-
nificance of all of our results remains robust to the choice of STRING evidence
score (see Extended Experimental Procedures). Subunit pairs were thus
divided into fusion pairs (having evidence of fusion between them) and nonfu-
sion pairs (no evidence of fusion). For some complexes, multiple distinct fusion
pairs were identified. In a few of these cases, STRING also identified indirect
fusions. For example, in K. aerogenes urease, γ fuses with β and β fuses with α,
but STRING also identified a γ-α fusion due to the indirect linkage via β.
We manually identified these indirect fusion pairs in STRING and moved
them to the nonfusion set. In total, 94 (3.7%) of the nonredundant heteromeric
subunit pairs were associated with evolutionary gene fusion events.

In this study, we identified gene fusion events as cases in which two sepa-
rate genes became joined. However, it is possible that some of these cases
resulted from gene fission events (i.e., a prefusion complex was really a post-
fusion complex). Although this could potentially have some implications for
our results, there is strong evidence that gene fusion is both the most dominant
mechanism behind the evolution of multidomain proteins (Pasek et al., 2006;
Buljan et al., 2010) and is much more common than gene fission (Kummerfeld
and Teichmann, 2005; Fong et al., 2007). This suggests that any contribution of
fusion to our data set must be minimal and therefore unable to account for the
strong trends we observed.

High-Throughput Protein Interaction Data
Just as we identified the subunit pairs from crystal structures, we compiled
analogous data sets from high-throughput protein-protein interaction data.
Instead of using crystal structures, we identified interacting pairs of proteins
as those with evidence of interaction in the STRING database (experimental
evidence score > 0.3). We could then directly split these interacting pairs
into fusion and nonfusion pairs using the STRING fusion evidence score.

nESI-MS Experiments
The complexes were kindly donated as follows: Salmonella typhimurium tryp-
thon synthase (Protein Data Bank [PDB] ID: 1WBJ; I. Schlichting, Max
Planck Institute for Medical Research, Heidelberg); E. coli acetyl coA carbox-
ylase carboxyltransferase (PDB ID: 2F9Y; G. Waldrop, Louisiana State Univer-
sity); E. coli carbamoyl phosphate synthetase (PDB ID: 1BXR; F. Raushel,
Texas A&M University); and K. aerogenes and Helicobacter mustelae ureases
(PDB ID: 1KRA and 3QGA, respectively; R. Hausinger, Michigan State Univer-
sity). Complexes were buffer exchanged from their purification buffers to
ammonium acetate at near-neutral pH, and further diluted with ammonium ac-
etate to give solutions containing 0.5–8 μM complex in 60–250 mM ammonium acetate.
Concentrations were adjusted for each complex to yield spectra of
the same complex and ammonium acetate concentration as a starting point.

Our aim was to explore a range of voltage and pressure conditions in order to detect sub-
complexes between the m/z values of the intact complex and free subunits
(Hernández and Robinson, 2007). Subcomplex identities were confirmed by
MS/MS spectra.

A high concentration (4–7 μM) of the complex was used to investigate the
extent of reassembly after the addition of acetic acid, ammonium solution, or
organic solvents. Aliquots of the concentrated disassembly solution were
diluted to the same complex concentration with either the buffer/solvent mix
or ammonium acetate alone. A control solution was also prepared from the
complex in ammonium acetate buffer to obtain solution conditions identical
to those of the reassembly solution. Spectra from the three solutions were ac-
quired using identical MS conditions.

In Silico (Dis)assembly
(Dis)assembly pathways were predicted for all heteromeric complexes with
more than three total subunits. We used a simple model based upon interface
size in which a complex was iteratively dissociated so that each step required
the disruption of the smallest total interface area.

For each pair of subunits associated with a fusion event, a heteromeric pair
of subunits from the same complex was randomly selected, giving 36 fusion
pairs and 36 randomly selected pairs. The mean value of the property of inter-
est for the fusion pairs (e.g., conservation of [dis]assembly or reduction of in-
faces upon fusion) was compared with the mean value from the randomly
selected pairs. The procedure was repeated 10000 times, allowing the p value
to be calculated as the frequency with which the random pairs had a mean
value less than or equal to that of the fusion pairs (i.e., the chance that the
mean value could be observed if fusions occurred randomly in the same com-
plex). A Perl script for performing this analysis is provided in the Extended
Experimental Procedures.

We also performed a similar comparison of shared interaction partners from
the protein-protein interaction data. Instead of comparing fusion pairs with
random pairs from the same complex, we compared them with random pairs
from the same interaction network. For example, given a fusion pair, A and B,
we also considered all of the interactions involving A or B, as well as the inter-
actions between proteins that both interacted with A or B. To calculate the p
values, we repeated the process 10⁴ times, and determined the likelihood
that the observed value could have been seen by chance. A Perl script for per-
forming this calculation is provided in the Extended Experimental Procedures.
This analysis was performed for all of the STRING “core” species (Table S3).

Terminal Distance Calculations
The distance between the N and C termini of different chains was calculated as
the distance between the Cα atoms of their terminal residues. Since the N and
C termini present in crystal structures may not represent the actual biologically
relevant termini, for this analysis we used only full-length proteins, and filtered
out fusion events in which any of the termini were missing (e.g., due to disorder
or the expression construct). We did this by identifying subunits in which the 20
N- or C-terminal residues from the full-length protein were missing. We iden-
tified the sequences of the full-length proteins by performing a blastp (Altschul
et al., 1997) search against all proteins in the STRING database and selecting
the sequence with the lowest E value. We determined the order of gene fusion
(i.e., AB or BA) by manually noting the order in which the genes are fused in
the STRING web interface. All of the fusion and reverse distances are provided in
Table S4.

ACCESSION NUMBERS
The protein interactions from this publication have been submitted to the
IMEx consortium (http://www.imexconsortium.org) through InAct (PMID:
19850723) and assigned the identifier IM-19876.

SUPPLEMENTAL INFORMATION
Supplemental Information includes Extended Experimental Procedures, three
figures, and four tables and can be found with this article online at http://dx.doi.
org/10.1016/j.cell.2013.02.044.

ACKNOWLEDGMENTS
We thank I. Schlichting (Max Planck Institute for Medical Research, Heidel-
berg), G. Waldrop (Louisiana State University), F. Raushel (Texas A&M Univer-
sity), and R. Hausinger (Michigan State University) for providing protein
complex samples. We acknowledge E. Boeri-Erba and K. Wright for assistance with the manuscript. J.A.M., T.P., and S.A.T. were supported by the Medical Research Council (MRC file reference number U105161047). J.A.M. was supported by a long-term fellowship from the Human Frontier Science Program Organization. H.H. was supported by an MRC Programme Grant. S.E.A. was supported by the Royal Society. C.V.R. was supported by an ERC Advanced Grant and a Royal Society Professorship.

Received: December 18, 2012
Revised: February 5, 2013
Accepted: February 21, 2013
Published: April 11, 2013

REFERENCES


