Interface size drives cotranslational assembly of protein complexes

Mihaly Badonyi[1], Joseph A Marsh[1*]

[1] MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

*correspondence to joseph.marsh [at] ed.ac.uk

Summary

Assembly pathways of protein complexes should be precise and efficient to minimise misfolding and unwanted interactions with other proteins in the cell. One way to achieve this is by seeding complex assembly during translation via nascent chain engagement. Here, we considered the possibility that the propensity of subunits to cotranslationally assemble is ingrained within the interface hierarchy of protein complexes. Using a combination of proteome-specific structure data and assembly-onset positions determined by ribosome profiling, we show that larger interfaces are prioritised in the course of cotranslational assembly. We observe that this effect is not exclusive to homomeric complexes, but appears to drive the assembly of heteromeric subunits, to the extent that interface size differences are detectable between N and C-terminal locations, with the former being larger on average. We provide explanations to this phenomenon and discuss its importance in a biological context.

Introduction

The majority of proteins across all domains of life function as part of multimeric complexes. Although we have a comprehensive understanding of the diverse quaternary structure space occupied by complexes[1], much less is known about where, when, and how their component subunits assemble. Nonetheless, continuing advances in X-ray crystallography, nuclear magnetic resonance spectroscopy, cryo-electron microscopy, mass spectrometry, and genetic interaction mapping[2] are enabling a transition towards a structural view of proteomes[3], which promises to answer these questions. It is difficult to overstate how much the development of ribosome profiling has accelerated this transition by offering quantitative measurements at the level of translation. Adaptations of the technique have revealed the cotranslational action of chaperones[4–6], shed light on the role of collided ribosomes in protein homeostasis[7–9], and supported the view of the ribosome as a signalling hub[10]. To the present work, however, it is of outstanding relevance that ribosome profiling has laid down strong evidence that the assembly of protein complexes often begins on the ribosome[5,11–16].

Two factors appear to be particularly important for cotranslational assembly: the proximity of nascent chains on adjacent (*cis*) or between juxtaposed (*trans*) ribosomes, and the localisation of interface residues towards the N-terminal side of a protein, thus allowing more time for an interaction to occur during translation[16–18]. Recent findings demonstrated that homomers, formed from multiple copies of a single type of polypeptide chain, predominantly assemble on the same transcript via the interaction of side-by-side elongating ribosomes[16]. This process is highly effective because it takes advantage of the fact that homomeric subunits are identical; hence nascent chains are essentially colocalised by the nature of their synthesis. On the other hand, heteromers, products of distinct genes that stably interact, can only employ the *trans* mode of assembly in eukaryotes, providing mechanisms that colocalise their transcripts exist[19–21]. Although

homomers may benefit from polysome-driven assembly, it requires allocation of cellular resources to ensure at least two ribosomes are actively translating at any one time[22]. In contrast, assembly of heteromers may only require a single ribosome on each messenger RNA, which could allow lowly abundant regulatory proteins to cotranslationally assemble[23,24]. Alternate ribosome usage and translation-coupled assembly may explain how cells achieve efficient construction of complexes with uneven stoichiometry, accounting for a substantial fraction of heteromeric complexes[25].

Despite growing evidence supporting the importance of cotranslational assembly, little is known about the properties of the interfaces involved. For example, it has been frequently reported that cotranslationally binding subunits have a tendency to fall out of solution or become degraded by orphan subunit surveillance mechanisms in the absence of their partner subunits[5,15,18,26,27]. These observations may be rationalised under two assumptions: all N-terminal interfaces are aggregation prone *in vivo* due to disturbance of cotranslational folding[28–32], or that cotranslationally forming interfaces possess unique structural properties which, when their nascent chains are devoid of partner subunits, predispose them to aggregation. Whilst there is evidence for the former[18], interfaces involved in nascent complex formation have not been systematically studied before; thus, we cannot exclude the possibility that they have structural attributes that make them more susceptible to a cotranslational route.

The essentiality of hydrophobic surface area has long been recognised in nucleation theory[33–35] and protein folding[36–38], but more importantly, it served as the founding principle of protein-protein recognition theory[39,40]. Interface size, defined as the buried surface area between subunits, shows correspondence to hydrophobic surface area owing to the fractional content of interface core residues[41], and it is a property that is relatively simple to compute from structural data[42–45]. Although the relationship between interface size and binding affinity is non-linear[46–49], interface size shows remarkable correspondence with subunit dissociation energy *in vitro*, and is reflective of the evolutionary history of subunits within complexes[50].

We hypothesised that cotranslational binding may be distinguished from other interactions based upon the areas of the interfaces involved. The size hierarchy of interfaces is well established in protein complexes and it can be used to predict the order in which their subunits assemble, in good agreement with experimental data[50–52]. According to this, the largest interfaces in a complex correspond to the earliest forming subcomplexes within the assembly pathway, irrespective of how binding is initiated. While the presence of specific contacts that increase affinity could introduce compositional biases into the sequence space exerting undue selection pressure on proteomes, variability in interface size can emerge from non-adaptive processes as the organising principle of cotranslational assembly[53–58]. In the present study, we address this idea by analysing assembly-onset positions determined by ribosome profiling of human complexes[16] and by comparing areas of first and last translated interfaces in multi-interface heteromers.

Results

Cotranslationally forming homomers are characterised by large interfaces

A recent study used ribosome profiling to identify cotranslational assembly-onset positions of over 4,000 human proteins[16], the majority of which are homomers that assemble on the same polysome. We therefore first sought to use this data to investigate the importance of interface area in the cotranslational assembly of homomeric complexes. The arrangement of homomeric subunits with respect to one or more rotational axes allows their classification into symmetry groups. The three most abundant groups are the twofold symmetric (C_2), cyclic ($C_{n;\ n>2}$), and dihedral (D_n) complexes, which all have distinct structural and functional characteristics[59–61] and should therefore be considered separately.

We reasoned that if size plays a role in determining which interfaces form cotranslationally, a trend should be discernible among homodimeric complexes. C_2 dimers represent the most highly populated symmetry group and their single isologous interface (*i.e.* symmetric or head-to-head) makes the analysis very simple to perform. In **Figure 1A**, we compare the interface area difference between cotranslationally forming *vs* all other C_2 symmetric human homodimers of known structure. In line with our expectation, dimers that assemble during synthesis are characterised by 21% larger interfaces ($p = 7.6 \times 10^{-7}$, Wilcoxon rank-sum test). While it is possible that larger interfaces are more resilient when subjected to the experimental procedures of ribosome profiling, we observed robustness of the trend to different interface area cutoffs, and found no statistically significant differences between the lengths of the protein sequences in either set (**Figure S1B**). Additionally, the location of these interfaces is N-terminally shifted (**Figure S1C**), suggesting that the interface size trend cannot be ascribed only to the survival bias of stable interfaces under harsh conditions.

Higher-order cyclic complexes are centred on one rotational axis so that every subunit has two interfaces, each with an adjacent protomer. Both interfaces are heterologous (*i.e.* asymmetric or head-to-tail) and approximately the same size. Cyclic symmetry is potentially confounded by its tendency to form ring-like structures, which are ubiquitous components of biological membranes[62]. As a result, membrane-bound complexes are enriched in non-polar amino acids that form the interface with the alkane core of the lipid bilayer. We focused on the analysis of cyclic homomers that do not localise to plasma or endomembrane systems, because of the competing hydrophobic forces exerted by protein-lipid interactions. Cotranslational membrane insertion of complexes is likely an active process requiring, for example, shielding factors[63,64], rather than being solely driven by hydrophobic surface area. Despite the relatively limited number of structures available, we detect a significant difference in the mean subunit-to-subunit interface area among soluble members of the cyclic symmetry group (**Figure 1B**), with a 19% increase in the mean of cotranslationally forming members ($p = 0.03$, Wilcoxon rank-sum test). Membrane-associated cyclic complexes show a non-significant opposite trend (**Figure S1D**). Interestingly, those membrane complexes that cotranslationally assemble expose their first heterologous interface earlier during translation than others (**Figure S1E**), presumably to avoid premature protein-lipid interactions. Both soluble and membrane-associated cyclic complexes that form on the ribosome are characterised by interfaces that are better separated in their linear sequence than those that lack evidence of cotranslational assembly (**Figure S1F**). Although the result is marginally non-significant, we speculate that this may facilitate a single mode of binding on the polysome.

Dihedral symmetry can be thought of as the stacking of a dimeric or cyclic complex via acquisition of a new twofold axis. All dihedral complexes have isologous interfaces, and those with at least six subunits can have both isologous and heterologous interfaces (*e.g.* D_3 dimers of cyclic trimers). We observe that cotranslationally assembling dihedral complexes have on average 29% larger subunit-to-subunit interfaces than those assumed to assemble after their complete synthesis (**Figure 1C**; $p = 4.1 \times 10^{-6}$, Wilcoxon rank-sum test). A dihedral complex is likely to have evolved from a symmetric dimer if the largest interface is isologous and, conversely, when the heterologous interface is the largest, the complex is likely to have arisen from a cyclic intermediate[50,65]. When the complexes are grouped based upon their evolutionary history, whether they evolved via dimeric or cyclic intermediates, the trend is present in both groups (**Figure S1G**).

Simultaneous assembly of heteromers involves large interfaces

We next sought to investigate the existence of an interface size trend in interactions between pairs of heteromeric subunits, formed between the products of different human genes. The assembly-onset positions are derived from a type of ribosome profiling technique that relies on the isolation of disomes[16], which are pairs of ribosomes connected by intertwined nascent chains. Heteromers identified in this manner are therefore undergoing "simultaneous" cotranslational assembly[15], whereby both chains interact while in the process of being translated (**Figure 2A**)**.** Although this mechanism could be less common in heteromers because it may limit the degree of freedom accessible to folding of nascent chains, several lines of evidence support its existence in cells[13,15,19]. Simultaneous assembly contrasts with "sequential" heteromer assembly, in which only one of the interacting subunits is in the process of being translated[5,13,15].

Heterodimers are similar to homodimers in that they have a single interface, enabling a simple comparison. **Figure 2B** shows the contrast between the interface sizes of simultaneously forming heterodimers with those that lack support of cotranslational assembly, with the former being 13% larger on average ($p = 2.3 \times 10^{-3}$, Wilcoxon rank-sum test). The trend, similar to that observed among symmetric homodimers, seems robust to interface size cutoffs, without exhibiting significant differences in subunit length (**Figure S2A**). An exciting example of a simultaneously assembling heterodimer is the complex of cyclin A2 and cyclin dependent kinase 2 (cyclin A2:CDK2; **Figure 2C**), where CDK2 buries more than 1,500 $Å^2$ area of its surface into cyclin A2. The complex is an important cell cycle regulator in metazoa, ensuring steady progression through S phase by phosphorylating various targets after the restriction point in G1 is reached, which is the point of commitment to the cell cycle[66]. Timely synthesis and independence from the diffusional regime are warranted by coordinated translation and simultaneous assembly.

Next, we identified interacting pairs of subunits in larger heteromeric complexes, *i.e.* those that have at least three subunits and therefore at least two different interfaces. Using the approach detailed in **Methods** and illustrated in **Figure S2B**, we mapped interface pairs most likely to form simultaneously in 253 multi-interface heteromers. Some of these belong to complexes that have been shown to use cotranslational assembly routes, such as the proteasome[13] and subunits of the transcription initiation complex[15], but many are not yet described in the literature, *e.g.* the loading of histone 2A-2B dimer onto importin-9[67], which has been reported to act as a storage chaperone while transporting the dimer to the nucleus. Another example is the V-type ATPase (**Figure 2D**), whose catalytic A and B subunits have been probed for their ability to assemble

in the sequential mode with a negative result[5], but the assembly-onset position combined with our structural approach identified the E1 subunit to simultaneously form with the catalytic B subunit. In **Figure 2E**, we show that the sizes of simultaneously forming interfaces are on average 14% larger than other interfaces within the same set of heteromers (p = 0.04, Wilcoxon rank-sum test).

Finally, we wished to put the identified simultaneously forming interfaces into the context of their full complexes. Do simultaneously forming interface pairs represent early forming subcomplexes that form the core of further assembly events? Although the largest interfaces are always predicted to assemble earliest in the assembly pathway, subsequent steps are non-trivial, as the earliest forming subcomplexes will continue to assemble with subunits harbouring interfaces of different sizes[51,52]. We predicted the assembly pathways of heteromeric complexes on the basis of their structures[51], which revealed that simultaneous cotranslational interfaces tend to form significantly earlier than other interfaces in the complexes (**Figure 2F**; $p = 8.4 \times 10^{-3}$, random sampling). Another interpretation is given by classifying assembly steps into "early" and "late", depending on their normalised assembly order[68], which is a 0-to-1 scale indicating the first-to-last steps of a pathway, where early steps have values ≤0.25. According to this, a simultaneously forming interface is 1.8-times more likely to be an early forming interface (97 [38%] of 253 *vs* 1150 [26%] of 4435; $p = 2.77 \times 10^{-5}$, Fisher's exact test).
Thus, notwithstanding the potential noise in the ribosome profiling data and the uncertainty around the determinism of assembly-onset positions, these results support the concept that assembly in cells is seeded tightly following the interface size hierarchy.

Cotranslational assembly influences N-terminal interface size in heteromers

Regardless of whether cotranslational assembly occurs simultaneously or sequentially, the N-terminal regions of proteins are more likely to be involved in interactions with partner subunits, by virtue of the vectorial synthesis on the ribosome. We therefore hypothesised that, in heteromeric subunits with multiple interfaces, N-terminal interfaces, which will be translated first, should tend to be larger than C-terminal interfaces, which will be translated last.

To address this, our approach was to pairwise compare the sizes of the most N-terminal (*i.e*. "first" appearing) and most C-terminal (*i.e*. "last" appearing) interfaces, as illustrated in **Figure 3A**. We defined the first interface as the one that exposes the first interface residue in the linear sequence, and, to treat the termini symmetrically, the last interface was defined as the one that presents the last interface residue. In **Figure 3B**, we show the result of this analysis for humans, as well as for heteromeric complexes from *Saccharomyces cerevisiae* (yeast) and *Escherichia coli*. Consistent with our hypothesis, in *E. coli* and human heteromers, the mean area of the first interface is larger than the last by 21% ($p = 2.0 \times 10^{-3}$) and 7% (p = 0.02, Wilcoxon signed-rank test), respectively. Interestingly, however, we do not observe a significant trend in yeast, where the mean area is in fact slightly larger for the last interface.

A factor that may explain the trend in yeast is that our yeast dataset is heavily skewed towards very large complexes. The median number of subunits in the heteromeric complexes are 6 for *E. coli*, 8 for human and 27 for yeast (full distribution in **Figure S3A**). Therefore, we investigated the interface size trend further by splitting multi-interface heteromers depending on the size of their complexes (**Figure S3D**). When considering only heteromeric complexes with fewer than 10 subunits, the trend for first interfaces to be larger

than last is present in all species (although still not significant in yeast with the smaller dataset). In contrast, none of the species show a significant trend for complexes with ≥10 subunits.

We can speculate about two potential reasons why the first *vs* last interface size trend is less prominent in larger complexes. First, evidence suggests that interface size is less useful for predicting experimental assembly order in large complexes[51], which is unsurprising given the greater number of interactions to consider and steps to predict. Thus, the relative sizes of interfaces may simply be less important in large complexes. Second, in large complexes, many subunits will participate as assembly intermediates and possess large intersubunit interfaces[51] that render the unbound state vulnerable to aggregation and these may be more prone to assemble simultaneously, thus diluting the N-terminal interface size trend otherwise imposed more strongly by the sequential route. Simultaneous assembly in large complexes could arise from random arrangements of colocalised transcripts, but still allow matching of expression levels and some degree of control over complex stoichiometry[26,69]. Of the subunits of human heteromeric complexes identified to simultaneously assemble based upon the ribosome profiling data, 114 are either the first or the last interface by our definition. Although the first interface is used significantly more in this mode of assembly (n = 70 [61%]; p = 0.02, binomial test), it is not exclusively employed. While we have no way of comparing this to the frequency of first interface usage in the sequential mode of assembly, it is worth noting that these 114 simultaneously assembling heteromeric subunits do not follow the first *vs* last interface size pattern, whereas the trend is highly significant in other human subunits not shown to assemble simultaneously (**Figure S3C**).

Lastly, we hypothesised that the distance between translation start points of the first and last interfaces (*i.e.* the earliest emerged interface residues of each) should correlate with the propensity for cotranslational assembly of the first interface. The later the translation of the last interface starts, the higher the chance that assembly of the first interface will take place undisturbed, without competition from other structurally important interfaces. To tackle the effect of large variances in length and interface size, we normalised the translational distance between the first and last interface by expressing it as the percentage of the protein's sequence length, and scaled the area difference by the sum of both interfaces. **Figure 3C** shows the correlation between the separation of translation start points and the area differences of the first and the last interface. As expected, increasing the distance between the translation start points monotonically increases the extent of the area difference, in favour of the first interface, and the observed bias does not appear to be confounded by subunit length (**Figure S3E**). The causal direction, whether this trend reflects that cotranslational assembly tends to happen more when interface separation is pronounced, or that it drives the evolution of interface size and sequence separation, remains to be explored.

The N-terminal size bias is strong in operon-encoded complexes

In prokaryotes, many heteromeric complexes are encoded by operons, where the different subunits will therefore be translated off of the same polycistronic mRNA molecules. Early studies in bacteria indicated that operon gene order may be correlated to physical interactions between the encoded proteins[70,71]. Subsequent analyses laid down theoretical, mechanistic, and evolutionary evidence in support of this[17,51,52,72]. Cotranslational assembly is likely to be particularly common in operon-encoded heteromers, given that the translation of different subunits is inherently colocalised. We therefore hypothesised that the tendency for

N-terminal interfaces to be larger should be stronger in operon-encoded heteromers, compared to non-operonal heteromers, where cotranslational assembly should be less likely to occur.

We used annotations derived from RNA sequencing datasets[73] to group heteromers from *E. coli* according to whether or not they are encoded by operons. One such complex, the RecBCD nuclease, is illustrated in **Figure 4A**. Genes of the subunits are located in adjacent loci encoding transcriptional units for RecC and RecB/D. A study aiming at reconstructing subunit activities noted that purification of RecD is complicated by the formation of inclusion bodies, while the other two subunits remain in the soluble fraction[74]. Moreover, a genetic analysis dissecting its assembly pathway proposed that partially folded RecC and RecD might interact during translation, or that RecC forms a complex with RecB first, onto which RecD is then assembled[75].

The regulatory subunit RecD has two interfaces well separated in the sequence, where the interface with RecC appears first. One might imagine that the nascent chain of RecD forms a complex with mature subunit of RecC, having double the interface area to accommodate RecC than that for RecB. In this scenario, the assembly efficiency is not only maximised by gene order reducing the stochasticity of binding events, but by cotranslational assembly minimising the need for post-translational diffusional association.

We found that the first interface is significantly larger than the last in complexes that have operon annotations (n = 110, $p = 9.6 \times 10^{-3}$, Wilcoxon signed-rank test; **Figure 4B**), but not in non-operonal *E. coli* complexes. Although the sample sizes are relatively small, bootstrap estimates of differences in mean interface area also indicate a significant difference only in operon-encoded complexes (**Figure S4**). This suggests that cotranslational assembly is likely to be relatively infrequent for non-operonal complexes in bacteria. While cotranslational assembly is clearly very common for non-operonal complexes from eukaryotes, it is plausible that most bacterial heteromers that would benefit from cotranslational assembly have evolved to be encoded by operons.

Discussion

It has long been understood that interface area is important for assembly, but the capacity in which it shapes interface hierarchy in complexes remained elusive. Here, it is shown that cotranslational subunit assembly harnesses differences in interface size, and, owing to the vectorial nature of protein synthesis, imposes constraints on the size of N-terminally located interfaces. The prevalence of cotranslational assembly and the evolutionary selection it puts on interfaces at the sequence level suggest that this mechanism exists as an integral part of biological assembly pathways.

Although our results agree with the principles of self-assembly, they do not answer the more fundamental question: have interfaces become larger and better separated to promote cotranslational assembly, or does assembly just happen to be more energetically favourable under these conditions? In other words, can it increase cellular fitness enough to influence the course of proteome evolution? Cotranslational assembly may well represent a ratchet-like mechanism of constructive neutral evolution[55,58], whereby a drift in interface properties creates ideal conditions for assembly on the ribosome. Reversion to the post-translational route is prevented by the accumulation of mutations that are neutral in the cotranslationally entrenched subunit, but would otherwise be deleterious in the ancestor.

In light of our current knowledge of bacterial operon structure and translation regulation, it is not surprising that we observed a strong trend among *E. coli* heteromers with respect to the size of the interface that first emerges from the exit tunnel. Supposedly, the effect is attributable to the widespread sequential assembly between mature subunits and nascent chains, reflecting the mechanism of effective subunit recruitment by large N-terminal “bait” interfaces. An interesting question for experiments is what fraction of prokaryotic heteromers encoded by polycistronic mRNA assemble simultaneously? Do heteromers, whose genes are transcribed in tandem, assemble in *cis*, providing that the structural organisation of the polysome allows for such a precise coordination[76]?

Attention should be paid to the far-reaching genetic consequences of cotranslational assembly[12]. How much does transcript-specific assembly buffer the dominant-negative effect, and what does it mean in the context of human genetic disease and cancer[77]? The dominant-negative effect requires mutant subunits to be stable enough to assemble into complexes, and thus the impact of these mutations tends to be milder at the structural level than of other pathogenic mutations[78], making them exceptionally difficult to detect using the existing armamentarium of variant effect predictors. Interestingly, however, the phenomenon of *cis* cotranslational assembly should actually make the dominant-negative effect less common, as it will limit the mixing that occurs between wild-type and mutant subunits. It remains to be seen whether dominant-negative mutations are in fact less common in cotranslationally assembling complexes. In addition, coexistence of both wild-type and mutant complexes in the cell may pose a substantial hurdle to cancer, which could be one reason why loss of heterozygosity is observed in certain tumour types, although its relevance is actively debated[79,80].

Nevertheless, many more topics of inquiry remain open for future studies. Analogous to protein folding, cotranslational assembly can be thought of as a hydrophobic collapse, sculpting the quaternary structure of the complex. As with folding in the cell, the involvement of other factors must not be overlooked. A wide

array of ribosome-associated chaperones are vital for nascent chain homeostasis[4,5,81–83] and the degree to which they choreograph assembly steps is yet to be elucidated. Furthermore, because interface size alone does not fully explain the range of affinities observed in a growing number of complexes[46–49], it will be exciting to see whether binding affinity plays a role in colocalising mRNA molecules in eukaryotic cells[13,14,19,21,84,85].

Our results build on evidence from the last decade and emphasise the importance of protein complex assembly at the translatome level. Although evolutionary selection against N-terminal interface contacts was previously found in homomers presumably to avoid misassembly[18], here we report an opposite phenomenon in which proteins that do cotranslationally assemble sustain large N-terminal interfaces in order to ingrain binary interactions into their assembly pathways.

Methods

Protein structural datasets

Starting from the entire set of protein structures in the Protein Data Bank[86] (PDB, rcsb.org) on 2021-02-18, we searched for all polypeptide chains longer than 50 residues with greater than 90% sequence identity to *Homo sapiens*, *Saccharomyces cerevisiae*, and *Escherichia coli* genes. For genes that map to multiple chains, we selected a single chain sorting by sequence identity, then number of unique subunits in the complex, and then the number of atoms present in the chain. Only biological assemblies (pdb1) were used and symmetry assignments were taken directly from the PDB. Heteromeric complexes formed by cleavage were excluded. In the generation of multi-interface heteromer datasets, chains with an at least 70% complete structure were considered and only included if they formed interface pairs >800 $Å^2$ with at least two different subunits.

Computation of interface related properties

Pairwise interfaces were calculated between all pairs of subunits using AREAIMOL from the CCP4 suite[45], with the default surface probe radius of 1.4 Å. The interface was defined as the difference between the solvent accessible surface area of each subunit in isolation and within the context of the full complex. The normalised interface location (interface midpoint), was calculated according to the equation $(i - 1) / (L - 1)$, where i marks the residue at which half the buried surface area of the interface is passed, and L is the sequence length.

The distance between translational start points was calculated as $(f_{\text{first}} - f_{\text{last}}) / L$, where f marks the first interface residue of the given interface, and L is the sequence length. Area differences between the first and the last were scaled according to $(BSA_{\text{first}} - BSA_{\text{last}}) / (BSA_{\text{first}} + BSA_{\text{last}})$, where BSA is the buried surface area of the respective interface. Assembly order was computed by predicting the assembly pathway assuming additivity of pairwise interfaces in each complex[51], implemented with the *assembly-prediction* Perl package[52].

Cotranslational assembly-onset positions

Data was downloaded from the supplementary material of reference [6], doi:10.1126/science.abc7151. For the determination of simultaneously forming heteromeric interface pairs, 30 residues were subtracted from the onset positions to account for the length of the ribosome tunnel.

Mapping simultaneously forming interface pairs

For the mapping of assembly-onset positions in multi-interface heteromers, to combat the probabilistic effect originating from the high number of residues that make up large interfaces, we employed a method (considered in **Figure S2B**) that calls the nearest interface midpoint, the residue at which half the area of the eventually formed interface is reached, to the assembly-onset position. The method discards proteins where the onset maps to a homomeric interface, with the assumption that the homomeric interface is hierarchically higher and undergoes *cis*-assembly on the same polysome.

First *versus* last interface

The “first” interface was defined as the first interface residue from the N terminus, and the “last” interface as the first interface residue from the C-terminal direction. Subunits where these residues mapped to the same interface were not considered for the analysis. These steps are necessary because large interfaces are more

spread out in the sequence, meaning that a unidirectional definition would be biased towards large interfaces. Cases where both the first and the last interface mapped to a heterologous interface, and are therefore likely representing the two halves of a cyclic complex, were also excluded.

Protein subcellular localisation

Protein localisation data was downloaded from the Human Protein Atlas[87] (HPA, proteinatlas.org). Locations were taken from the "Main location" column and those with "Uncertain" reliability were discarded. Localisation of genes lacking HPA data (23%) was predicted with the BUSCA[88] tool (busca.biocomp.unibo.it).

Mapping subunits to operons

Operons were downloaded from OperomeDB[73]. Genes were mapped to UniProt identifiers using data from the UniProt[89] FTP site (ftp.uniprot.org).

Visualisation and statistics

Molecular graphics were performed with UCSF ChimeraX[90] version 1.1. Data and statistical analyses were carried out in RStudio[91] version 1.4.1106 using R version 4.0.3 (r-project.org), in compliance with the tidyverse philosophy[92]. Violin plots in this study display the area-normalised density profiles with the boxes marking the first and third quartiles, the middle lines are the median, the notches show the 95% confidence interval of the median, and the circles represent the mean. In charts where only the boxplot is shown, the whiskers extend up to 1.5× the interquartile range, and the dots represent outliers.

Statistical tests were two-sided and performed at the significance level $\alpha = 0.05$. Although the distribution of interface sizes appear lognormal, in most cases Shapiro-Wilk tests and visual inspection of q-q plots did not support normality, and thus the non-parametric Wilcoxon rank-sum test was used consistently. In Wilcoxon signed-rank tests, the main assumption that the data are symmetric around the median was supported by boxplot distributions. In the regression shown in **Figure 3D**, conditions for statistical inference including linearity of relationship between the variables, the independence of the residuals, the normality of the residuals, and homoscedasticity were met; validations are included in the analysis script.

Data and code availability

The data and code will be deposited on Edinburgh DataShare upon acceptance of this manuscript.

Acknowledgements

MB is supported by the Biotechnology and Biological Sciences Research Council EASTBIO Doctoral Training Partnership (BB/M010996/1), and used resources provided by the Edinburgh Compute and Data Facility (ecdf.ed.ac.uk). JAM was supported by a Medical Research Council Career Development Award (MR/M02122X/1) and is a Lister Institute Research Prize Fellow.

Author contributions

JAM conceived the project and generated the structural and assembly order data. MB acquired all other data for use in the study and carried out the analyses. JAM supervised and aided in interpreting the results. MB wrote the manuscript, which was finalised with considerable contributions by JAM.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

1. Ahnert, S. E., Marsh, J. A., Hernández, H., Robinson, C. V. & Teichmann, S. A. Principles of assembly reveal a periodic table of protein complexes. *Science (80-.).* **350**, aaa2245–aaa2245 (2015).

2. Braberg, H. *et al.* Genetic interaction mapping informs integrative structure determination of protein complexes. *Science (80-.).* **370**, (2020).

3. Levy, E. D. & Vogel, C. "Structuromics": another step toward a holistic view of the cell. *Cell* **184**, 301–303 (2021).

4. Oh, E. *et al.* Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell* **147**, 1295–1308 (2011).

5. Shiber, A. *et al.* Cotranslational assembly of protein complexes in eukaryotes revealed by ribosome profiling. *Nature* **561**, 268–272 (2018).

6. Stein, K. C., Kriel, A. & Frydman, J. Nascent Polypeptide Domain Topology and Elongation Rate Direct the Cotranslational Hierarchy of Hsp70 and TRiC/CCT. *Mol. Cell* **75**, 1117-1130.e5 (2019).

7. Han, P. *et al.* Genome-wide Survey of Ribosome Collision. *Cell Rep.* **31**, 107610 (2020).

8. Arpat, A. B. *et al.* Transcriptome-wide sites of collided ribosomes reveal principles of translational pausing. *Genome Res.* **30**, 985–999 (2020).

9. Zhao, T. *et al.* Disome-seq reveals widespread ribosome collisions that promote cotranslational protein folding. *Genome Biol.* **22**, 1–35 (2021).

10. D'Orazio, K. N. & Green, R. Ribosome states signal RNA quality control. *Molecular Cell* vol. 81 1372–1383 (2021).

11. Duncan, C. D. S. & Mata, J. Widespread cotranslational formation of protein complexes. *PLoS Genet.* **7**, e1002398 (2011).

12. Natan, E., Wells, J. N., Teichmann, S. A. & Marsh, J. A. Regulation, evolution and consequences of cotranslational protein complex assembly. *Current Opinion in Structural Biology* vol. 42 90–97 (2017).

13. Panasenko, O. O. *et al.* Co-translational assembly of proteasome subunits in NOT1-containing assemblysomes. *Nat. Struct. Mol. Biol.* **26**, 110–120 (2019).

14. Sepulveda, G. *et al.* Co-translational protein targeting facilitates centrosomal recruitment of PCNT during centrosome maturation in vertebrates. *Elife* **7**, (2018).

15. Kamenova, I. *et al.* Co-translational assembly of mammalian nuclear multisubunit complexes. *Nat. Commun.* **10**, (2019).

16. Bertolini, M. *et al.* Interactions between nascent proteins translated by adjacent ribosomes drive homomer assembly. *Science (80-.).* **371**, 57–64 (2021).

17. Shieh, Y. W. *et al.* Operon structure and cotranslational subunit association direct protein assembly in bacteria. *Science (80-.).* **350**, 678–680 (2015).

18. Natan, E. *et al.* Cotranslational protein assembly imposes evolutionary constraints on homomeric proteins. *Nat. Struct. Mol. Biol.* **25**, 279–288 (2018).

19. Liu, F., Jones, D. K., De Lange, W. J. & Robertson, G. A. Cotranslational association of mRNA encoding subunits of heteromeric ion channels. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 4859–4864 (2016).

20. Pizzinga, M. *et al.* Translation factor mRNA granules direct protein synthetic capacity to regions of polarized growth. *J. Cell Biol.* **218**, 1564–1581 (2019).

21. Wang, G. *et al.* Spatial organization of the transcriptome in individual neurons. *bioRxiv* 2020.12.07.414060 (2020) doi:10.1101/2020.12.07.414060.

22. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* vol. 165 535–550 (2016).

23. Heyer, E. E. & Moore, M. J. Redefining the Translational Status of 80S Monosomes. *Cell* **164**, 757–769 (2016).

24. Biever, A. *et al.* Monosomes actively translate synaptic mRNAs in neuronal processes. *Science (80-.).* **367**, (2020).

25. Marsh, J. A., Rees, H. A., Ahnert, S. E. & Teichmann, S. A. Structural and evolutionary versatility in protein complexes with uneven stoichiometry. *Nat. Commun.* **6**, 1–10 (2015).

26. Juszkiewicz, S. & Hegde, R. S. Quality Control of Orphaned Proteins. *Molecular Cell* vol. 71 443–457 (2018).

27. Choe, Y. J. *et al.* Failure of RQC machinery causes protein aggregation and proteotoxic stress. *Nature* **531**, 191–195 (2016).

28. Zipser, D. & Perrin, D. Complementation on Ribosomes. *Cold Spring Harb. Symp. Quant. Biol.* **28**, 533–537 (1963).

29. Ciryam, P., Morimoto, R. I., Vendruscolo, M., Dobson, C. M. & O'Brien, E. P. In vivo translation rates can substantially delay the cotranslational folding of the Escherichia coli cytosolic proteome. *Proc. Natl. Acad. Sci. U. S. A.* **110**, (2013).

30. Jacobs, W. M. & Shakhnovich, E. I. Evidence of evolutionary selection for cotranslational folding. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 11434–11439 (2017).

31. Kudva, R. *et al.* The shape of the bacterial ribosome exit tunnel affects cotranslational protein folding. *Elife* **7**, (2018).

32. Kramer, G., Shiber, A. & Bukau, B. Mechanisms of cotranslational maturation of newly synthesized proteins. *Annual Review of Biochemistry* vol. 88 337–364 (2019).

33. Hermann, R. B. Theory of hydrophobic bonding. II. The correlation of hydrocarbon solubility in water with solvent cavity surface area. *J. Phys. Chem.* **76**, 2754–2759 (1972).

34. Tanford, C. The hydrophobic effect and the organization of living matter. *Science* vol. 200 1012–1018 (1978).

35. Chandler, D. Interfaces and the driving force of hydrophobic assembly. *Nature* vol. 437 640–647 (2005).

36. Privalov, P. L. & Khechinashvili, N. N. A thermodynamic approach to the problem of stabilization of globular protein structure: A calorimetric study. *J. Mol. Biol.* **86**, 665–684 (1974).

37. Chothia, C. Structural invariants in protein folding. *Nature* **254**, 304–308 (1975).

38. Gething, M. J. & Sambrook, J. Protein folding in the cell. *Nature* vol. 355 33–45 (1992).

39. Kauzmann, W. Some Factors in the Interpretation of Protein Denaturation. *Adv. Protein Chem.* **14**, 1–63 (1959).

40. Chothia, C. & Janin, J. Principles of protein-protein recognition. *Nature* **256**, 705–708 (1975).

41. Levy, E. D. A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution. *J. Mol. Biol.* **403**, 660–670 (2010).

42. Mitternacht, S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research* **5**, (2016).

43. Kleinjung, J. & Fraternali, F. POPSCOMP: An automated interaction analysis of biomolecular complexes. *Nucleic Acids Res.* (2005) doi:10.1093/nar/gki369.

44. Hubbard SJ, T. J. NACCESS. *Department of Biochemistry and Molecular Biology, University College, London.* (1993).

45. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography* vol. 67 235–242 (2011).

46. Brooijmans, N., Sharp, K. A. & Kuntz, I. D. Stability of macromolecular complexes. *Proteins Struct. Funct. Genet.* (2002) doi:10.1002/prot.10139.

47. Eisenberg, D. & Mclachlan, A. D. Solvation energy in protein folding and binding. *Nature* **319**, 199–203 (1986).

48. Horton, N. & Lewis, M. Calculation of the free energy of association for protein complexes. *Protein Sci.* 169–181 (1992).

49. Vangone, A. & Bonvin, A. M. J. J. Contacts-based prediction of binding affinity in protein–protein complexes. *Elife* **4**, (2015).

50. Levy, E. D., Erba, E. B., Robinson, C. V. & Teichmann, S. A. Assembly reflects evolution of protein complexes. *Nature* **453**, 1262–1265 (2008).

51. Marsh, J. A. *et al.* Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell* **153**, 461–470 (2013).

52. Wells, J. N., Bergendahl, L. T. & Marsh, J. A. Operon Gene Order Is Optimized for Ordered Protein Complex Assembly. *Cell Rep.* **14**, 679–685 (2016).

53. Conant, G. C. Neutral evolution on mammalian protein surfaces. *Trends in Genetics* vol. 25 377–381 (2009).

54. Ahnert, S. E., Johnston, I. G., Fink, T. M. A., Doye, J. P. K. & Louis, A. A. Self-assembly, modularity, and

physical complexity. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **82**, 026117 (2010).

55. Gray, M. W., Lukeš, J., Archibald, J. M., Keeling, P. J. & Doolittle, W. F. Irremediable complexity? *Science* vol. 330 920–921 (2010).

56. Lynch, M. Evolutionary diversification of the multimeric states of proteins. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E2821–E2828 (2013).

57. Leonard, A. S. & Ahnert, S. E. Evolution of interface binding strengths in simplified model of protein quaternary structure. *PLoS Comput. Biol.* **15**, e1006886 (2019).

58. Hochberg, G. K. A. *et al.* A hydrophobic ratchet entrenches molecular complexes. *Nature* **588**, 503–508 (2020).

59. Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. *Annual Review of Biophysics and Biomolecular Structure* vol. 29 105–153 (2000).

60. Levy, E. D. & Teichmann, S. Structural, evolutionary, and assembly principles of protein oligomerization. in *Progress in Molecular Biology and Translational Science* vol. 117 25–51 (Elsevier B.V., 2013).

61. Bergendahl, L. T. & Marsh, J. A. Functional determinants of protein assembly into homomeric complexes. *Sci. Rep.* **7**, (2017).

62. Forrest, L. R. Structural Symmetry in Membrane Proteins∗. *Annu. Rev. Biophys.* **44**, 311–337 (2015).

63. Pleiner, T. *et al.* WNK1 is an assembly factor for the human ER membrane protein complex. *Mol. Cell* (2021) doi:10.1016/j.molcel.2021.04.013.

64. Nicolaus, F. *et al.* Residue-by-residue analysis of cotranslational membrane protein integration in vivo. *Elife* **10**, 1–16 (2021).

65. Marsh, J. A. & Teichmann, S. A. Protein Flexibility Facilitates Quaternary Structure Assembly and Evolution. *PLoS Biol.* **12**, e1001870 (2014).

66. De Boer, L. *et al.* Cyclin A/cdk2 coordinates centrosomal and nuclear mitotic events. *Oncogene* **27**, 4261–4268 (2008).

67. Padavannil, A. *et al.* Importin-9 wraps around the H2A-H2B core to act as nuclear importer and histone chaperone. *Elife* **8**, (2019).

68. McShane, E. *et al.* Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell* **167**, 803-815.e21 (2016).

69. Schwarz, A. & Beck, M. The Benefits of Cotranslational Assembly: A Structural Perspective. *Trends in Cell Biology* vol. 29 791–803 (2019).

70. Mushegian, A. R. & Koonin, E. V. Gene order is not conserved in bacterial evolution. *Trends in genetics□: TIG* vol. 12 289–290 (1996).

71. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328 (1998).

72. Sneppen, K., Pedersen, S., Krishna, S., Dodd, I. & Semsey, S. Economy of operon formation: Cotranscription minimizes shortfall in protein complexes. *MBio* **1**, (2010).

73. Chetal, K. & Janga, S. C. OperomeDB: A Database of Condition-Specific Transcription Units in Prokaryotic Genomes. *Biomed Res. Int.* **2015**, (2015).

74. Masterson, C. *et al.* Reconstitution of the activities of the RecBCD holoenzyme of Escherichia coli from the purified subunits. *J. Biol. Chem.* **267**, 13564–13572 (1992).

75. Amundsen, S. K., Taylor, A. F. & Smith, G. R. A domain of RecC required for assembly of the regulatory RecD subunit into the Escherichia coli RecBCD holoenzyme. *Genetics* **161**, 483–492 (2002).

76. Brandt, F. *et al.* The Native 3D Organization of Bacterial Polysomes. *Cell* **136**, 261–271 (2009).

77. Bergendahl, L. T. *et al.* The role of protein complexes in human genetic disease. *Protein Sci.* **28**, 1400–1411 (2019).

78. McEntagart, M. *et al.* A Restricted Repertoire of de Novo Mutations in ITPR1 Cause Gillespie Syndrome with Evidence for Dominant-Negative Effect. *Am. J. Hum. Genet.* **98**, 981–992 (2016).

79. Gencel-Augusto, J. & Lozano, G. p53 tetramerization: At the center of the dominant-negative effect of mutant p53. *Genes and Development* vol. 34 1128–1146 (2020).

80. Shahbandi, A. & Jackson, J. G. Analysis across multiple tumor types provides no evidence that mutant p53

exerts dominant negative activity. *npj Precis. Oncol.* **3**, 1–3 (2019).

81. Döring, K. *et al.* Profiling Ssb-Nascent Chain Interactions Reveals Principles of Hsp70-Assisted Folding. *Cell* **170**, 298-311.e20 (2017).

82. Koldewey, P., Horowitz, S. & Bardwell, J. C. A. Chaperone-client interactions: non-specificity engenders multi-functionality. *J. Biol. Chem.* (2017) doi:10.1074/jbc.R117.796862.

83. Zhang, Y. *et al.* The ribosome-associated complex RAC serves in a relay that directs nascent chains to Ssb. *Nat. Commun.* **11**, 1–12 (2020).

84. Heidenreich, M. *et al.* Designer protein assemblies with tunable phase diagrams in living cells. *Nat. Chem. Biol.* **16**, 939–945 (2020).

85. Dufourt, J. *et al.* Imaging translation dynamics in live embryos reveals spatial heterogeneities. *Science (80-.).* **372**, 840–844 (2021).

86. Berman, H. M. Westbrook, J. Feng, Z. Gilliland, G. Bhat, T. N. Weissig, H. Shindyalov, I. N. Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* (2000).

87. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science (80-.).* **356**, (2017).

88. Savojardo, C., Martelli, P. L., Fariselli, P., Profiti, G. & Casadio, R. BUSCA: An integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res.* **46**, W459–W466 (2018).

89. Bateman, A. *et al.* UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

90. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

91. Rstudio, T. RStudio: Integrated Development for R. *Rstudio Team, PBC, Boston, MA URL http://www.rstudio.com/* (2020) doi:10.1145/3132847.3132886.

92. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

93. Buckland, S. T., Davison, A. C. & Hinkley, D. V. Bootstrap Methods and Their Application. *Biometrics* **54**, (1998).

Figures

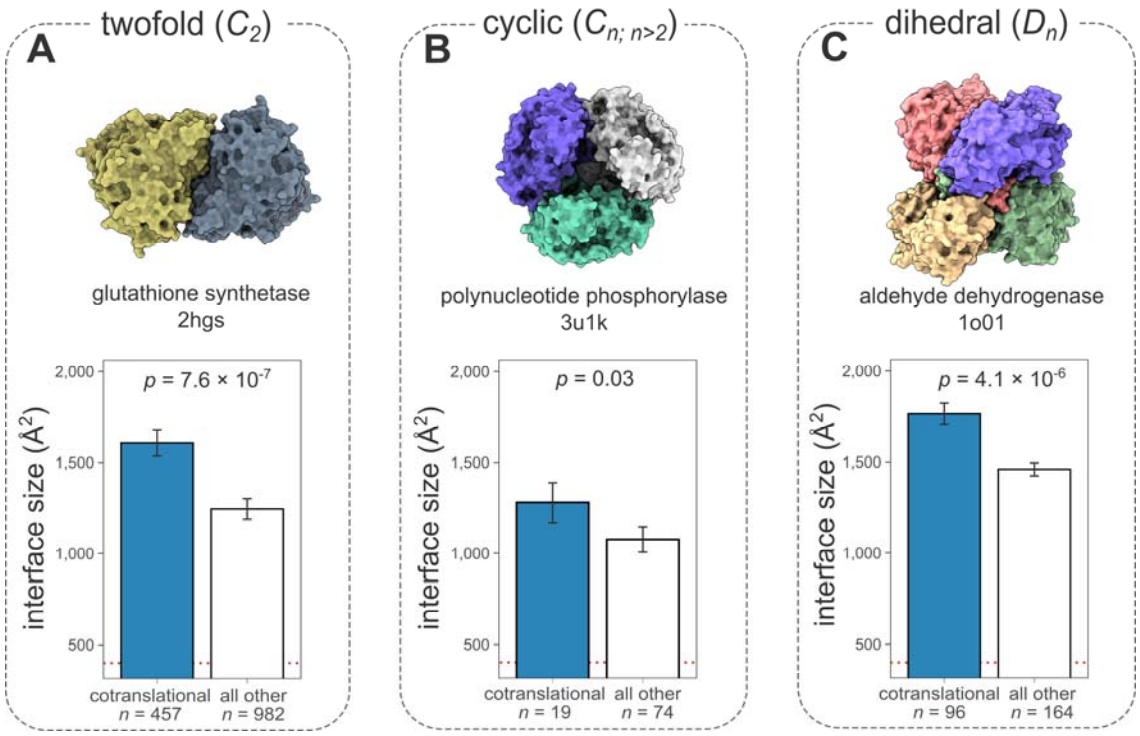

Figure 1 | Interface size differences between cotranslationally assembling human homomers and all other human homomers with known structure.

(**A**) Twofold symmetric homodimers; their full distribution is shown in **Figure S1A**.

(**B**) Cyclic complexes.

(**C**) Dihedral complexes.

For cyclic and dihedral complexes, the average subunit-to-subunit interface size is shown. Representatives of symmetry groups are illustrated above the charts, where the four letter code is the PDB identifier. Error bars represent the standard error of the mean, the red dashed line marks the 400 Å^2 interface size cutoff, and *n* numbers denote the number of proteins in each group. The *p*-values were calculated with two-sided Wilcoxon rank-sum tests.

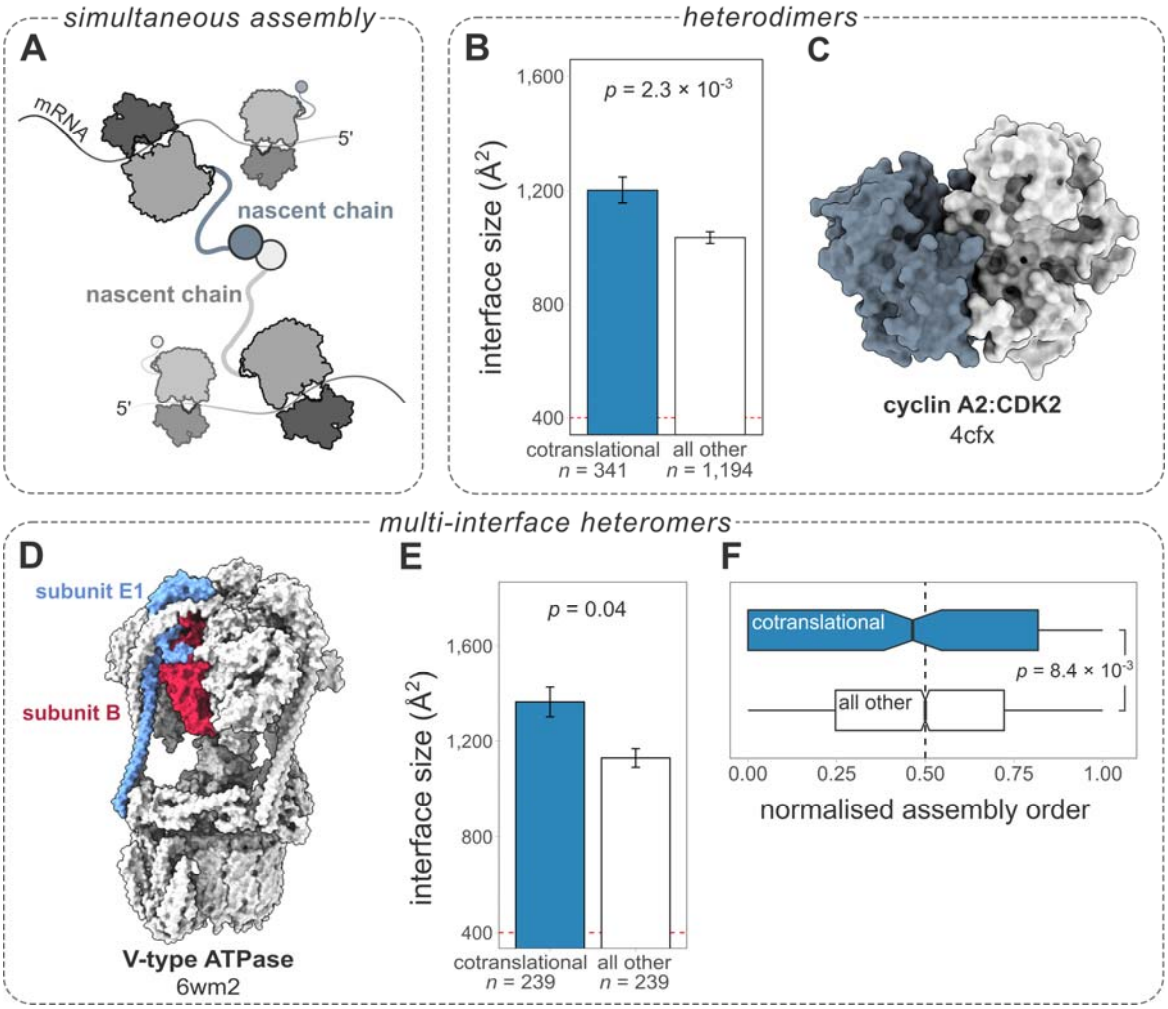

Figure 2 | Simultaneous cotranslational assembly in heteromers.

(**A**) Schematic representation of a simultaneous cotranslational assembly event: both nascent chains, encoded by transcripts of different genes, interact during synthesis on juxtaposed ribosomes.

(**B**) Comparison of interface areas between simultaneously forming and all other human heterodimeric interfaces. The *p*-value was calculated with a two-sided Wilcoxon rank-sum test.

(**C**) An example of simultaneous assembly in heterodimers is that between cyclin A2 and cyclin dependent kinase 2.

(**D**) An example in multi-interface heteromers is that between subunits B and E1 of the V-type ATPase.

(**E**) Pairwise comparison of interface areas in multi-interface heteromers between simultaneously forming interfaces and the mean of all other heteromeric interfaces. Subunits with only a homomeric interface in addition to the simultaneously forming heteromeric interface were excluded. The *p*-value was calculated with a Wilcoxon signed-rank test.

(**F**) Comparison of the normalised assembly order (0-to-1 indicate the first-to-last steps of the assembly pathway) of simultaneously assembling interface pairs (n = 253) and all other interfaces in the complexes (n = 4,435). The *p*-value was calculated from 10,000 resamples of randomly drawn interfaces from each structure by determining the fraction of point estimates (mean assembly order) smaller or equal to the mean assembly order of simultaneously forming interfaces, with correction for finite sampling[93].

Four letter codes are PDB identifiers. Error bars represent the standard error of the mean, the red dashed line marks the 400 $Å^2$ area cutoff and *n* numbers denote the number of proteins.

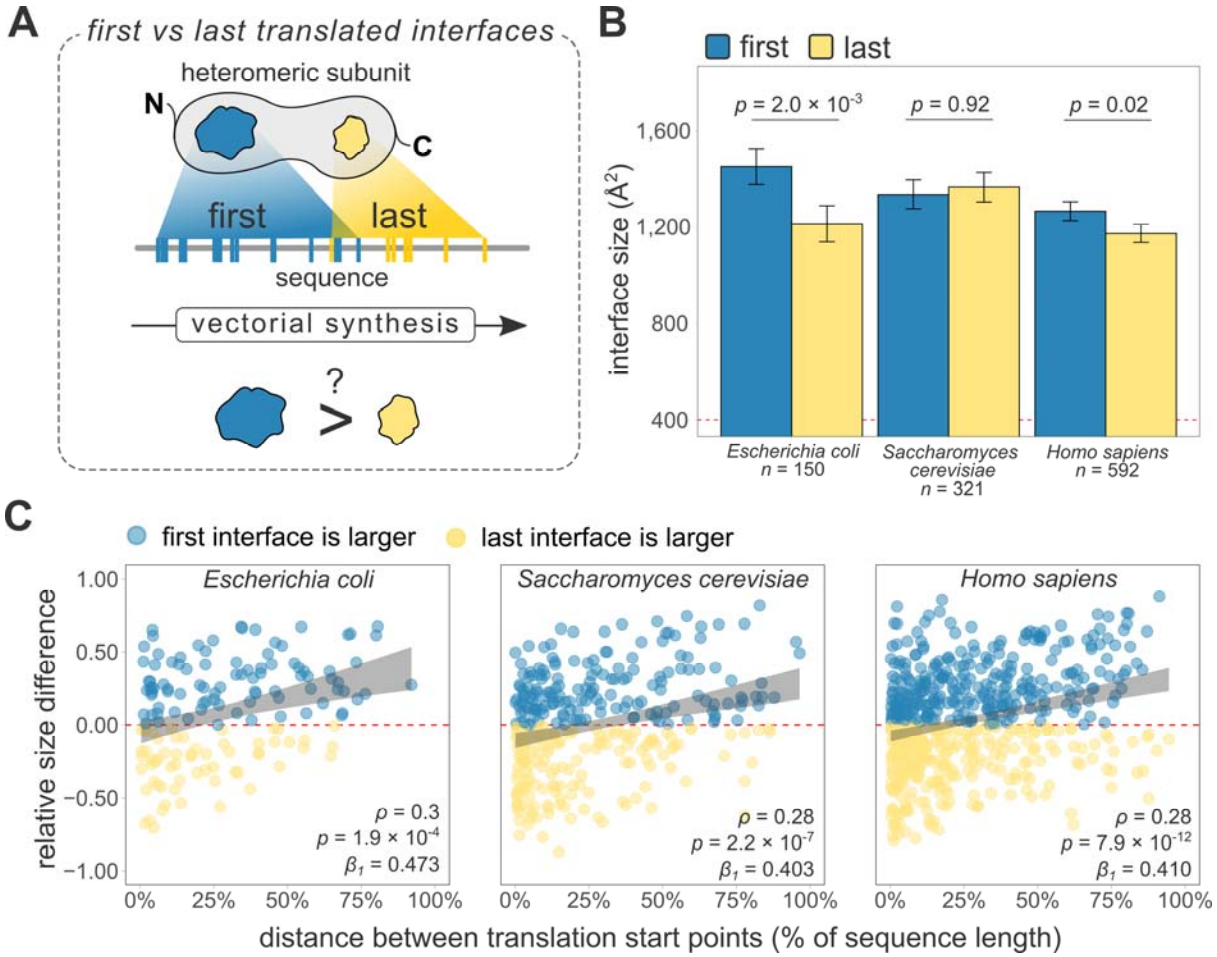

Figure 3 | Cotranslational assembly influences N terminal interface size in heteromers.

(**A**) The basis of our hypothesis testing is to see whether the first translated interface has a larger area than the last translated interface. This concept is schematically represented here, where the grey round shape symbolises the subunit of a complex, with the differently coloured patches representing the first and last translated interfaces (blue and yellow). The interface order is determined by projection of the interface residues onto the protein's linear sequence.
The first and last interface is the one that presents the first and the last interface residue in the sequence, respectively.
(**B**) Pairwise comparison of the areas of first and last translated interfaces in multi-interface heteromers in the different species. Error bars represent the standard error of the mean, the red dashed line marks the 400 $Å^2$ area cutoff, and *n* numbers denote the number of protein structures analysed in each proteome. The *p*-value was calculated with a Wilcoxon signed-rank test. **Figure S3B** shows the full distribution of interface sizes.
(**C**) Relationship between the area difference of the first and the last translated interface and the distance between their translation start points, expressed as the fraction of sequence length (non-scaled scatter plot in **Figure S3F**). Dots are coloured according to the sign of the difference (blue – larger first, yellow – larger last interface). The *rho*-values denote the Spearman rank-order correlation with the associated *p*-values. The β_1 parameter is the estimate for the population regression slope, representing the change in interface size difference for every increase of one unit of distance, *i.e.* the full sequence length. The shaded areas correspond to 95% confidence intervals of the regression line.

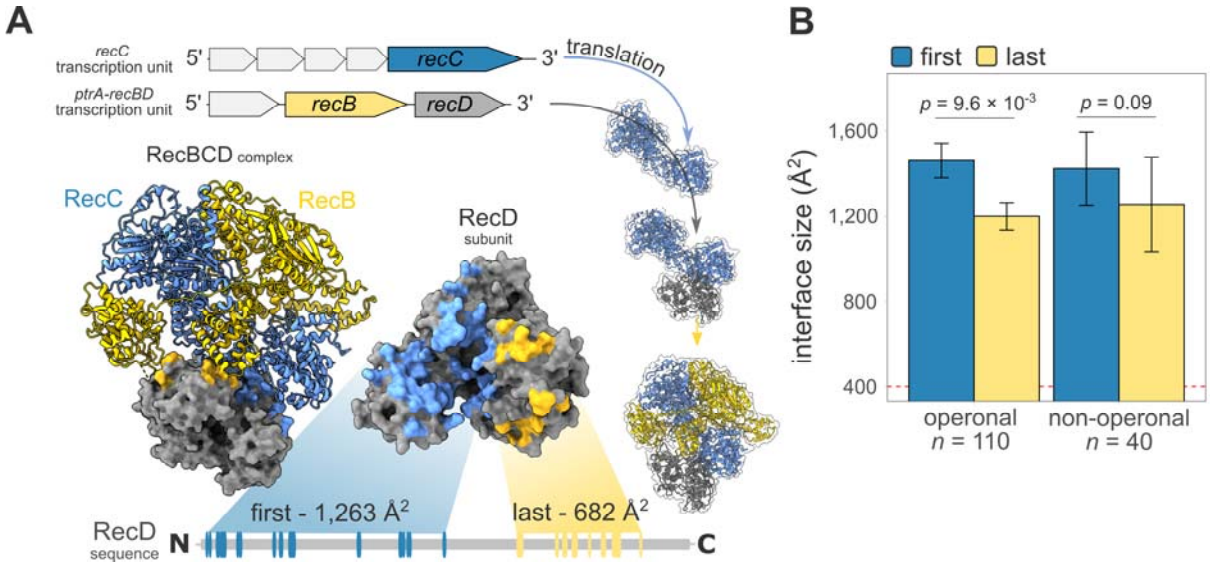

Figure 4 | The N-terminal interface size bias is strong in operon-encoded complexes.

(**A**) An example of an operon-encoded complex is the RecBCD nuclease (PDB ID: 5ld2), involved in homologous recombination in bacteria. The RecB and RecD subunits of the complex are encoded within the same operon and adjacent to this locus is the recC operon, encoding the third subunit of the complex. When interface residues of the RecD subunit are projected onto its linear sequence (*bottom*), we see that the first appearing interface is formed with RecC and the last with RecB. The RecC interface exposes twice the buried surface area of RecB, which can be explained by a model in which RecC is translated first and then cotranslationally forms with the nascent chain of RecD.

(**B**) Area differences between the first and last translated interfaces in *E. coli* operon-encoded and non-operonal complexes. Error bars represent the standard error of the mean, the red dashed line marks the 400 $Å^2$ area cutoff, and *n* numbers denote the number of proteins. The *p*-values were calculated with a Wilcoxon signed-rank test.

Supplementary figures

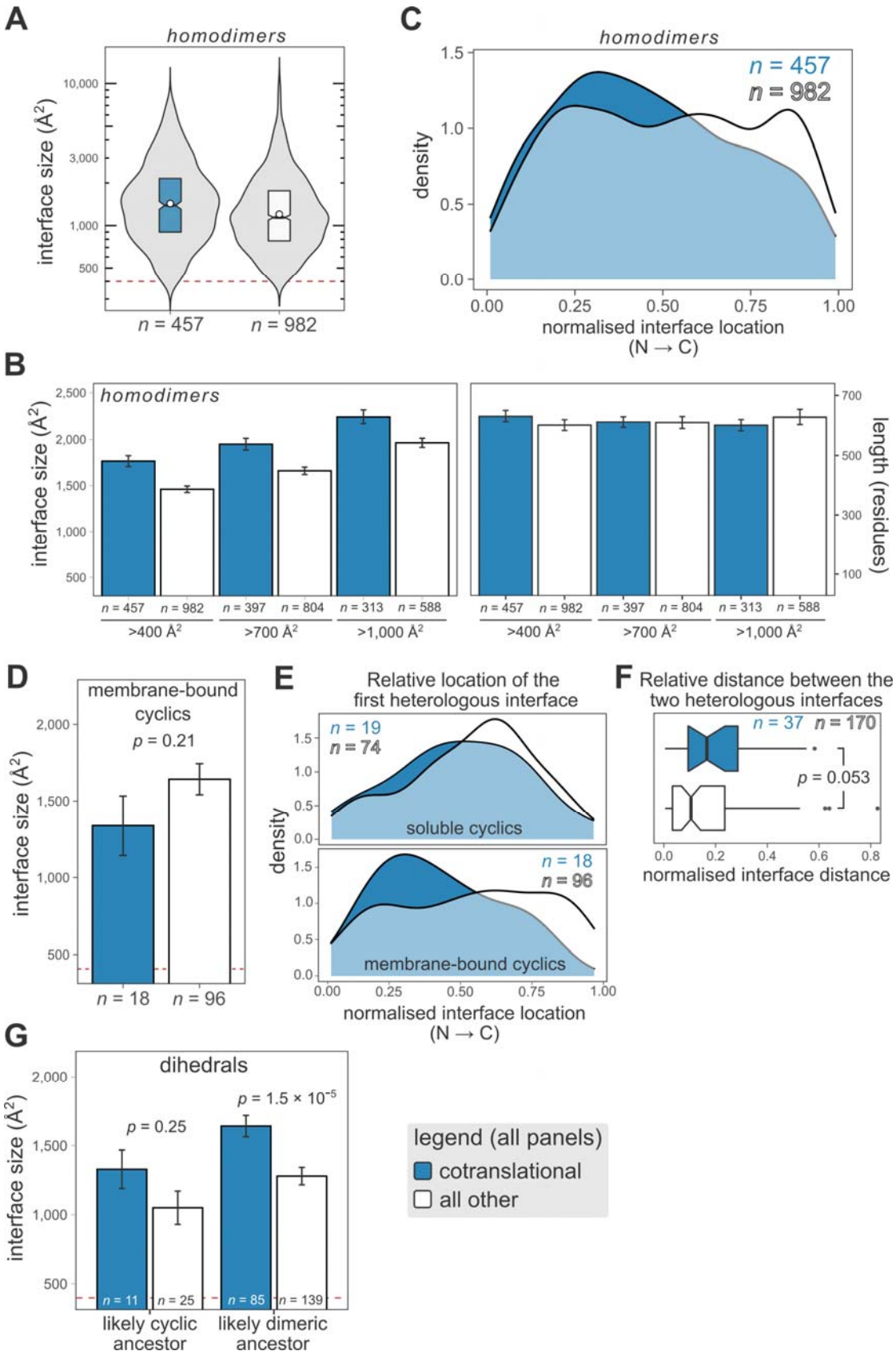

Figure S1 | Supplementary figures to Figure 1.

(**A**) Interface area distributions of cotranslationally assembling and all other human homodimers of known structure.

(**B**) Comparison of interface size differences in homodimers at increasing area cutoffs. On the right, the means of protein lengths are shown, demonstrating that the interface size trend is not attributable to longer polypeptide chains.

(**C**) Distribution of interfaces (>400 $Å^2$) along the length of homodimer sequences. The normalised interface location is calculated by determining the residue at which half the area of the interface is passed and expressing it as a fraction of the sequence length (therefore 0 and 1 are the N and C terminus, respectively).

(**D**) Comparison of mean subunit-to-subunit interface size differences of membrane-localised cotranslationally assembling and all other membrane-bound cyclic complexes.

(**E**) Normalised location of the first heterologous interface (the most N-terminal interface in cyclic homomers), grouped by soluble and membrane-bound cyclic complexes and whether they cotranslationally assemble.

(**F**) Normalised interface distance, which is a measure of separation, between the first and the second heterologous interfaces in cyclic complexes.

(**G**) Comparison of average subunit-to-subunit interface area differences in dihedral complexes, split by whether they originate from a cyclic or a dimeric ancestor.

On all figures, *n* numbers denote the number of proteins analysed in each group. Error bars on bar charts represent the standard error of the mean, and the red dashed line marks the 400 $Å^2$ area cutoff. The *p*-values were calculated with two-sided Wilcoxon rank-sum tests.

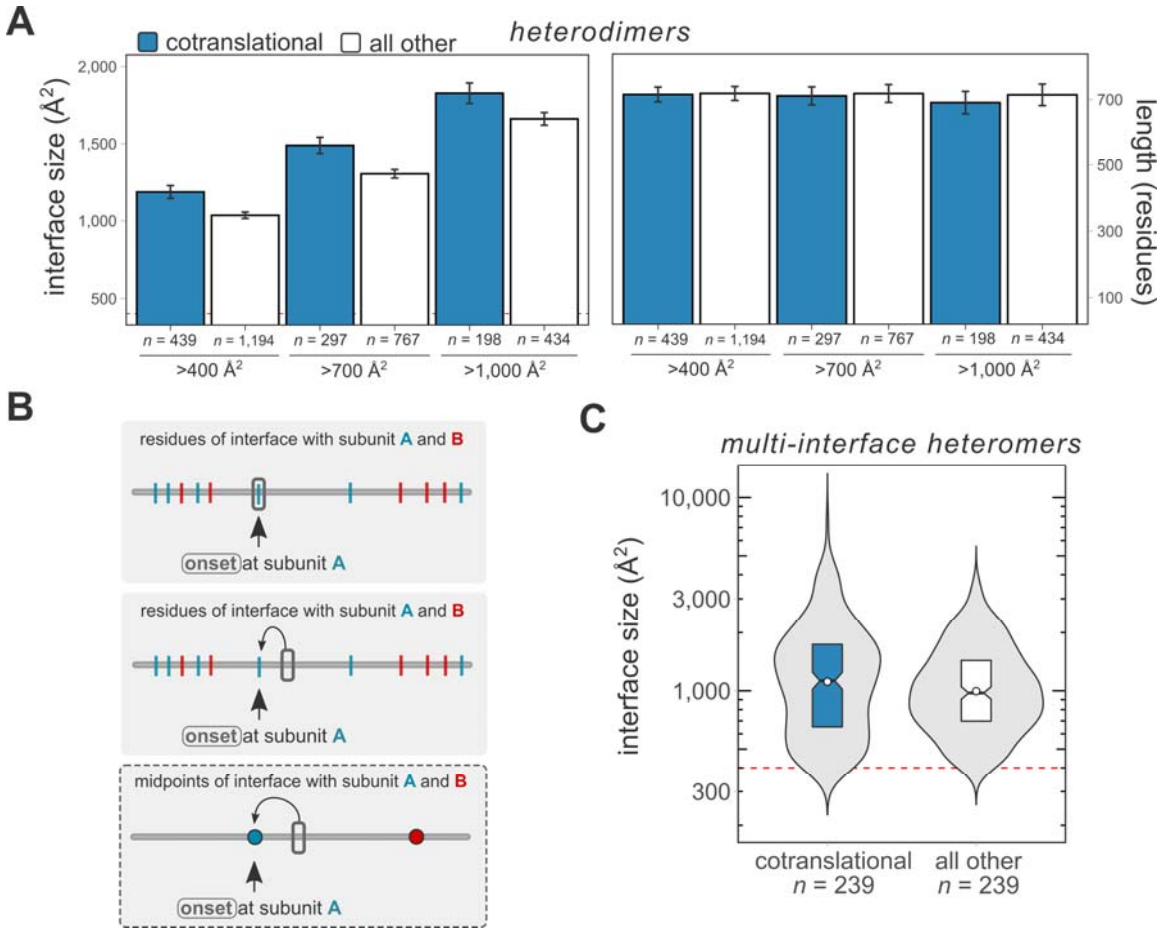

Figure S2 | Supplementary figures to Figure 2.

(**A**) Area distributions of simultaneously assembling interfaces (blue) and all other interfaces (white) on human multi-interface heteromers. On the right, the means of subunit lengths are shown.

(**B**) Methods considered for mapping cotranslational assembly-onset positions to interfaces in human multi-interface heteromers. *Top*: taking the onset position (a single residue) as is. This results in a reduction in the number of structures available for analysis because the mapped residue is often not directly at an interface. *Middle*: in cases where the onset position does not map to an interface residue, the last occurring interface residue is taken. *Bottom*: in the applied method, the assembly-onset is mapped to the nearest interface midpoint, which is the residue at which half the area of the interface is reached.

(**C**) Interface size distributions of simultaneously assembling interfaces and the mean of all other interfaces in multi-interface heteromers (related to **Figure 2B**).

Where applicable, *n* numbers denote the number of proteins.

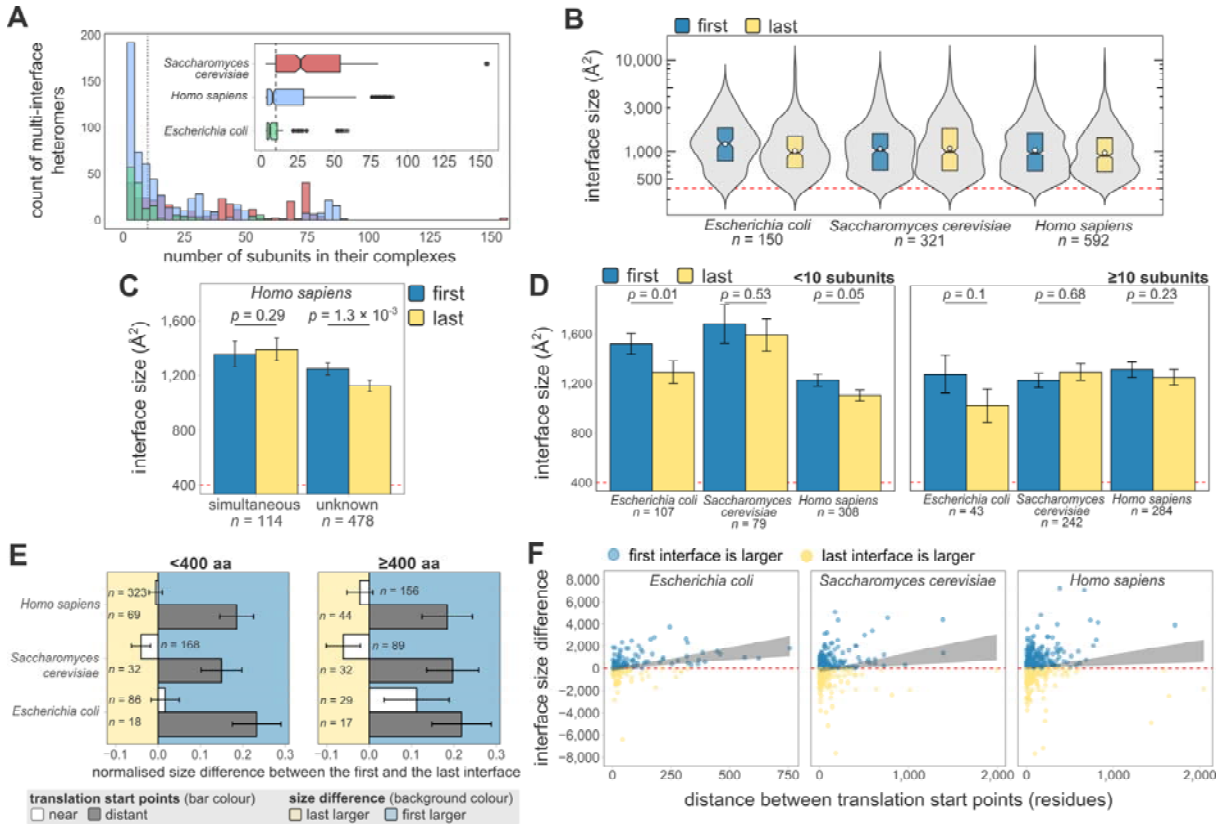

Figure S3 | Supplementary figures to Figure 3.

(**A**) Histogram (main) and boxplot (inset) of the number of subunits in the analysed structures, where multi-interface heteromers were selected from. The dashed line represents *n* of subunits = 10.

(**B**) Comparison of area distributions of first and last interfaces (blue and yellow) in *E. coli*, yeast, and human multi-interface heteromers.

(**C**) First *vs* last interface trend in human multi-interface heteromers where either the first or the last interface has been identified to simultaneously assemble and all other multi-interface heteromers with "unknown" mode of assembly.

(**D**) Results of **Figure 3B** grouped into <10 and ≥10 subunit containing complexes. Error bars represent the standard error of the mean, the red dashed line marks the 400 $Å^2$ area cutoff.

(**E**) The relationship between the interface area difference of the first and last translated interface and the distance between their translation start points is not confounded by subunit length. The distance between translation start points was classified into "near" and "distant" based on the midpoint of the interval, which slightly varies within species depending on the minima and maxima of the ranges (*E. coli*: (0.00105,0.453], (0.453,0.905]; yeast: (0.00115,0.477], (0.477,0.954]; human: (0.00114,0.473], (0.473,0.945]). The subunit lengths were further split into <400 or ≥400 residues.

(**F**) The results of **Figure 3C** with absolute interface size difference and translation distance values.

Where shown, *n* numbers denote the number of proteins, and the *p*-values were calculated with Wilcoxon signed-rank tests.

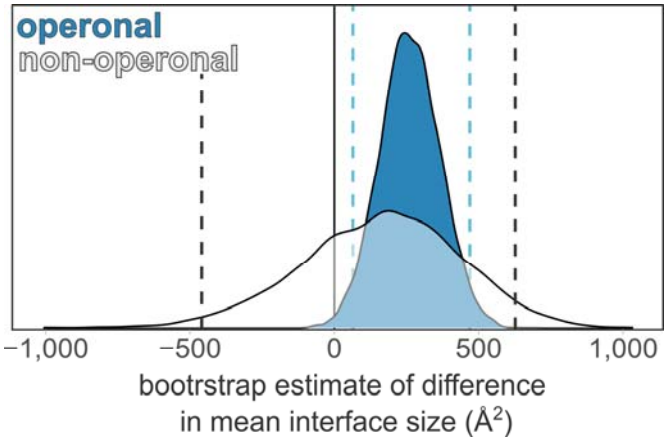

Figure S4 | Supplementary figure to Figure 4.

Bootstrap estimates of the difference in mean area between the first and last interfaces in operon-encoded complexes and those without annotation (blue and white). Dashed lines represent bias-corrected accelerated confidence intervals (α = 0.05), calculated from 10,000 bootstrap replicates.