This chapter was originally published in the book *Progress in Molecular Biology and Translational Science, Vol. 117*, published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues who know you, and providing a copy to your institution’s administrator.

All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution’s website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at: [http://www.elsevier.com/locate/permissionusematerial](http://www.elsevier.com/locate/permissionusematerial)


ISBN: 978-0-12-386931-9

© Copyright 2013 Elsevier Inc.

Academic Press
CHAPTER TWO

Structural, Evolutionary, and Assembly Principles of Protein Oligomerization

Emmanuel D. Levy*, Sarah Teichmann†,‡

*Department of Structural Biology, Weizmann Institute of Science, Rehovot, Israel
†MRC Laboratory of Molecular Biology, Cambridge, United Kingdom
‡Wellcome Trust Genome Campus, Cambridge, United Kingdom

Contents

1. Introduction: From a Crystal to a Quaternary Structure 26
   1.1 Increase in structural data over time 26
   1.2 The crystalline lattice and the problem of nonbiological contacts 26
   1.3 Distinguishing biological from nonbiological protein contacts 27
   1.4 Inferring biological assemblies 29
2. Homomer Morphology 29
   2.1 Size of homomers 29
   2.2 Symmetry types 31
   2.3 Emergence of symmetric structures 33
   2.4 Classification of homomers to facilitate their study 33
3. Interfaces in Homomers 35
   3.1 Defining the protein surface and the interface 35
   3.2 Interface properties: Focusing on the amino acid composition 36
4. Creating Interfaces in Homomers 38
   4.1 Are point mutations likely to create new interfaces? 38
   4.2 Molecular mechanisms associated to changes in oligomeric state 39
5. Evolution of Homomer Geometry and Diversity Through Gene Duplication 40
   5.1 Probable oligomeric state transitions 40
   5.2 Evolution through gene duplication 41
6. Assembly of Homomers 43
   6.1 Assembling into dimers 43
   6.2 Assembling into higher-order oligomers 44
   6.3 When assembly recapitulates evolution 45
7. Conclusion 45
Acknowledgments 46
References 46
Abstract

In the protein universe, 30–50% of proteins self-assemble to form symmetrical complexes consisting of multiple copies of themselves, called homomers. The prevalence of homomers motivates us to review many of their properties. In Section 1, we describe the methods and challenges associated with quaternary structure inference—these methods are indeed at the basis of any analysis on homomers. In Section 2, we describe the morphological properties of homomers, as well as the database 3DComplex, which provides a taxonomy for both homomeric and heteromeric protein complexes. In Section 3, we review interface properties of homomeric complexes. In Section 4, we then present recent findings on the evolution of homomer interfaces, which we link in Section 5 to the evolution of homomers as entire entities. In Section 6, we discuss mechanisms involved in their assembly and how these mechanisms can be linked to evolution.

1. INTRODUCTION: FROM A CRYSTAL TO A QUATERNARY STRUCTURE

1.1. Increase in structural data over time

A homomer is formed by self-interacting copies of a protein unit. As noted by Janin et al.,¹ the characterization of many homomers began long before protein structures could be solved by X-ray crystallography. At the time, the method of choice was analytical centrifugation. By 1975, this led to the characterization of the quaternary state for hundreds of oligomeric assemblies.² At the same time, there were less than 13 structures deposited in the Protein Data Bank. Structural data have, however, grown steadily since then,³ particularly over the past two decades. For example, 829 structures were recorded in 1992, 16,322 structures were recorded in 2002, and this number rose to 74,937 as of October 2012. Note that these numbers reflect only protein structures solved by X-ray crystallography, which is the most commonly used method for solving the structure of proteins.

1.2. The crystalline lattice and the problem of nonbiological contacts

This abundance of structural data means that today, the largest body of information available on protein quaternary structure comes from X-ray crystallography. Importantly, however, a prerequisite to solving a protein’s structure by X-ray crystallography is to obtain a crystal, that is, to induce the protein (or protein complex) into self-assembling with other copies of itself in a regular fashion to form a lattice. The structure solved by crystallographers
corresponds to the asymmetric unit (ASU), which is the minimal unit from which the whole crystal can be reconstructed using symmetry operations. Thus, a protein crystal can be seen as a mosaic, whose repeating unit is the ASU. This is illustrated in Fig. 2.1, with a “fish” being an ASU.

Importantly, the ASU does not necessarily reflect the biological state of a protein. For example, an ASU may contain a dimer of a protein that exists as a monomer in vivo. Conversely, the ASU may contain a single copy of a protein that actually forms a dimer or a tetramer in vivo, as illustrated in Fig. 2.1.

In sum, due to the nature of X-ray crystallography, artificial crystal contacts must be formed between proteins, resulting in many assemblies that can potentially be considered as biological. A task of fundamental importance is therefore to distinguish biological protein interfaces from nonbiological interfaces in the crystal.

1.3. Distinguishing biological from nonbiological protein contacts

We all know that strong friendships require time to develop and that people rarely become “best friends” in a day. Similarly, a pair of proteins is unlikely to exhibit a large interaction interface due to chance. In other words, a large
interface is likely to be the result of an evolutionary process. Consistent with
this idea, the buried surface area of crystal contacts, which occur by chance
and are not selected for, is on average significantly smaller than that of bio-
logical interfaces. Typically, biological interfaces bury more than 400 Å²
per subunit (800 Å² total), while crystal contacts bury on average less than
400 Å² per subunit. It can happen, however, that crystal contacts bury
400–1000 Å², and in some instances even more. Thus, a number of
properties must be considered to discriminate between biological and
nonbiological interfaces as accurately as possible. These properties can be
broadly classified into one of two types: (i) structural properties, deduced
directly from the protein structure or from the crystal lattice, and (ii) com-
parative properties, which make use of external data not present in the struc-
ture itself.

Among structural properties, interface amino acid composition has prob-
ably been the most frequently described, showing that crystal contacts are
more polar on average than biological interfaces. Another property
described by Pal et al. is the “number of polypeptide segments” at interfaces,
with biological interfaces involving, on average, a smaller number of longer
segments. Other promising descriptors are the fraction of atoms that are
fully buried at the interface and the associated packing density at the inter-
face, indicating that biological interfaces tend to exhibit better packing and
involve a larger fraction of fully buried atoms. An interesting observation is
also that amino acids at crystal contacts have on average lower confor-
mational entropies than amino acids found at the surface. This may prove
a useful feature to consider in classifiers developed in the future. Such exis-
ting classifiers include NOXclass and IPAC, which are based on a com-
bination of structural parameters coupled to a machine-learning framework.

In terms of comparative properties, conservation has been the most
widely used descriptor. A premise in using conservation information is that
functional elements are on average more conserved than less functional ele-
ments. Amino acids at biological interfaces, which have a specific function
(to maintain or regulate the interaction), are thus expected to be more con-
served than surface ones, which may or may not be involved in specific func-
tions. A number of studies have indeed confirmed that interface amino acids
are more evolutionarily conserved than surface ones. Rather than con-
sidering evolutionary conservation, Xu and Dunbrack have assessed the
conservation of interfaces in different crystal forms. Since proteins are

---

Methods used to measure protein surfaces and volumes are discussed in Section 3.
frequently crystallized in multiple conditions (e.g., with different ligands),
it is possible to discriminate contacts that are recurrent across crystal
forms from those that are unique to a particular crystallization condition.
As expected, recurrent contacts tend to be biological.

1.4. Inferring biological assemblies

Historically, the ASU was the default information downloadable from the
Protein Data Bank. This prompted the development of protein quaternary
structure (PQS) (1998), which was the first automatic approach to infer qua-
ternary structures from crystal lattices. Later, the PDB started to provide
information about biological quaternary structures, called “Biological
Units” at the time and “Biological Assemblies” nowadays. The PDB relies
on three sources for inferring Biological Assemblies: the authors of the struc-
ture; the PQS server; and the protein interfaces, surfaces, and assemblies
(PISA) server. Nevertheless, errors persist due to the difficult nature of
the problem. The database ProtBuD in fact provides a side-by-side compar-
ison between the biological units provided by PQS and PDB. PQS, which
is hosted at the European Bioinformatics Institute, is however going to be
discontinued soon, as another server, PISA, replaces it. Interestingly, PISA
uses a different approach from those described above. It aims at computing
the $\Delta G$ of association between proteins, which in theory should integrate all
the structural features in the best possible way. PISA provides accurate results
in over 80% of cases. Yet, another approach is taken with our PiQSi server,
which makes quaternary structure annotation community based. Anyone
can register with PiQSi and annotate quaternary structures with the help of
the interactive interface to explore, for example, homologous complexes.
Annotations are then reviewed internally before being added to the database.

2. HOMOMER MORPHOLOGY

2.1. Size of homomers

The most basic property that can be used to describe a homomer is its num-
ber of subunits. In a thorough review on structural symmetry and protein
function, Goodsell and Olson record the occurrence of different oligomeric
states for proteins found in *Escherichia coli*. Not considering hetero-
complexes, they observe 23% of monomers and 77% of homomers. Cer-
tainly, these numbers should be critically assessed as they may be biased
to some extent. For example, a scientist may be more likely to report an
oligomeric state when a dimer is observed than when a monomer is observed. In other words, oligomers might be more “noticeable” compared to monomers. This prompts us to assess their prevalence in crystal structures, which represents another independent type of data.

In Fig. 2.2, we show the distribution of the number of subunits in homomers as obtained from structural data. Importantly, these data may not suffer from the same “reporting” bias but are likely to suffer from other biases related to, for example, the probability of crystallization of different types of proteins. Yet, it also shows that homomers are widespread. Similar to the data from Goodsell and Olson, we observe only 35% of monomers in E. coli. In Homo sapiens, this number goes up to 54% (whether this difference is biological or results from different target choices by crystallographers would need to be investigated further—we also note an important difference between both species: E. coli structures often span full length proteins, while H. sapiens structures often correspond to protein fragments). Taken together, these data suggest that among proteins that are not part of a stable heterocomplex (e.g., such as the ribosome or the proteasome), homomers are formed at frequencies ranging from ~50% to ~70%.

Considering the frequencies for the different types of homomers, we note a general decrease as the number of subunits increases in Fig. 2.2. However, it is also clear that homomers with even numbers of subunits are favored over those with odd numbers. For example, in E. coli, there are 4.0% homomers with three subunits and 8.6% with four. The reason for this effect is that two types of symmetry (discussed below) can make up

![Figure 2.2](http://www.weizmann.ac.il/Structural_Biology/faculty_pages/ELevy/intDef/interface_def.html)  
**Figure 2.2** Number of quaternary structures per symmetry type. The frequency of different symmetry types obtained using the E. coli and the H. sapiens data set from Ref. 33 available at [http://www.weizmann.ac.il/Structural_Biology/faculty_pages/ELevy/intDef/interface_def.html](http://www.weizmann.ac.il/Structural_Biology/faculty_pages/ELevy/intDef/interface_def.html).
even-numbered homomers, while a single type can lead to odd-numbered ones.\textsuperscript{34} This is also illustrated in Fig. 2.2, where the corresponding symmetry type is indicated under each number of subunits. There is indeed not a single “C4” tetramer in the \textit{E. coli} data set, but there are 8.6% of “D2” tetramers. The decrease observed with increasing numbers of subunits should thus be considered independently for both symmetry types (C type or cyclic, and D type or dihedral).

### 2.2. Symmetry types

Cyclic and dihedral symmetries are the two main types encountered in homomers. In a cyclic symmetry, the subunits are related by a single \( n \)-fold symmetry axis, where \( n \) denotes the symmetry type (i.e., \( C_n \)). In a \( C_2 \) dimer, for example, the subunits are related by a twofold symmetry axis, meaning that rotating the complex by 180° around the axis results in the subunits being mutually superposed. In a \( C_3 \) trimer, the same occurs after a 120° rotation (Fig. 2.3).

Importantly, among all cyclic symmetries, \( C_2 \) is the only one where the contact between subunits is mediated by the same surface patch on both subunits—Monod termed this type of contact “\textit{iso}logous,” where subunits contact each other in a “face-to-face” fashion. In contrast, when considering a higher-order symmetry like \( C_4 \), the subunits contact each other in a “face-to-back” fashion, so that contacts are now “\textit{hetero}logous”. In other words,
interfaces associated with a twofold symmetry are symmetric themselves (the axis of symmetry passes through the center of mass of the interface), while interfaces associated with higher-order symmetries are not symmetric (and the axis of symmetry does not coincide with the center of the interface).

Dihedral symmetries (D2–D5 in Fig. 2.3) can be seen as two cyclic structures stacked onto each other and related by twofold symmetry axes. The resulting geometry has important functional implications in terms of, for example, directionality. In contrast to cyclic complexes, dihedral complexes are nondirectional and are not found in membranes (a notable exception is the nuclear pore complex, which exhibits dihedral pseudosymmetry but spans two lipid bilayers).

A cyclic complex can therefore acquire a new isologous interface to become dihedral. A dihedral complex, however, cannot acquire a new isologous interface as this would lead to the formation of an infinite polymer. This is illustrated in Fig. 2.4, with the example of hemoglobin. An allelic variant of the β chain (an aspartic acid mutated to a valine) induces the formation of a new protein–protein interaction interface. Because of the symmetry of a hemoglobin tetramer, this interface appears twice per tetramer (once on each copy of the β chain), so that one tetramer interacts with another, which interacts with another, etc. This allelic variant thus leads to the formation of fibrils in red blood cell, conferring on them an elongated

Figure 2.4 Open symmetries can be triggered by single point mutations as in sickle cell anemia. The left side of the image shows a schematic representation of hemoglobin, while the right side shows a molecular surface. Each representation shows free (wild-type) hemoglobin and an allelic variant associated to sickle cell anemia, which forms fibrils. Note that the fibrils have an illustrative purpose only and may not be accurate with respect to the orientation of the subunits.
shape characteristic of the sickle cell disease. Such “open” symmetries, which appear dangerous due to their potential to form polymers, are uncommon. Their prevalence is shown in the histogram Fig. 2.2 in the category “NCS” (no closed symmetry).

2.3. Emergence of symmetric structures

Whether the axis of symmetry is within or outside of the interface introduces a fundamental difference in terms of the energy of interaction. If inside, as put by Monod et al. “mutation of one residue, conferring upon it the capacity to form a bond with its partner, would result in two new bonds being achieved in the dimer.”37 The fact that one mutation has double the effects in a C2 association indeed has important energetic consequences. Let us consider a very simple model where a random mutation $i$ favors or disfavors the creation of a new interaction by $\Delta G_i$, and let us assume that $\Delta G_i$ is normally distributed.

One can visualize intuitively that considering the “double effect” at C2 interfaces results in a larger variance for the distribution of $\Delta G_i$.36 This suggests that the impact of mutations should have more often extreme energetic effects (both positive and negative) on isologous interfaces when compared to heterologous interfaces. Several studies have investigated these notions in detail—from the statistical energetic effects associated with self-association,38 to their implication in proteins using a coarse-grained model,39 an atom-based model40 or an evolutionary-oriented model.41 These studies all point to the fact that isologous (C2) interfaces are energetically more favorable than heterologous interfaces.

Considering the above arguments helps understand why cyclic symmetries with three and more subunits are rare (as these do not involve isologous but heterologous interfaces). On top of this, geometry can also help explain their scarcity. Considering a monomer, a new isologous interface naturally yields a closed dimer. However, when considering a new heterologous interface, the two surface patches on the monomer must be precisely oriented not to yield an open screw-type symmetry (i.e., involving rotations plus translations). In sum, the evolution of new cyclic complexes is limited by two factors: the type and the orientation of the interfaces that need to be created.

2.4. Classification of homomers to facilitate their study

Hemoglobin was the first protein oligomer whose structure was solved.42 Since then, many variants in different forms (e.g., with or without oxygen bound to the heme) or from different species have been crystallized, so that
there are over 200 hemoglobin entries archived in the PDB. Importantly, however, for any analysis relating to quaternary structures, it is usually necessary to remove redundant structures (e.g., to estimate the frequency of different symmetries). This was a driving force for the development of 3DComplex, which is a classification of protein complexes of known structure. 3DComplex organizes protein complexes according to many structural features, the three main of which are (i) the three-dimensional structure of

---

**Figure 2.5** The 3DComplex hierarchical classification. QS stands for “quaternary structure,” which refers to the spatial arrangement of subunits in the complex. The names of different levels of the hierarchy are indicated on the left, and the corresponding number of groups at each level is given on the right. The classification is hierarchical with one root at the top and 21,037 leaves at the bottom. At the top most level, lenient similarity is sufficient to group complexes, and these groups are split into subgroups toward the bottom of the hierarchy, where similarity of various detailed features is required. This scheme is similar to the taxonomic classification of species, where human beings are in the same group as baker’s yeast at the top of the hierarchy, but are separated at lower levels.

---

A

<table>
<thead>
<tr>
<th>Total</th>
<th>Terminology</th>
<th>Same graph topology</th>
<th>Subunits with same structure (SCOP superfamily)</th>
<th>Within-complex subunit sequence identity</th>
<th>Across-complex subunit sequence similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>QS topologies</td>
<td>✅</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>3785</td>
<td>QS families</td>
<td>✅</td>
<td>✅</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>3849</td>
<td>QS</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✗</td>
</tr>
<tr>
<td>6441  to QS30</td>
<td></td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✗</td>
</tr>
<tr>
<td>15,270 to QS100</td>
<td></td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✗</td>
</tr>
<tr>
<td>30,001</td>
<td>All data</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✗</td>
</tr>
</tbody>
</table>

B

Displays a nonredundant set of 207 tetramers with “square” topology

<table>
<thead>
<tr>
<th>Code</th>
<th>Picture</th>
<th>Function</th>
<th>No. chains</th>
<th>Symmetry</th>
<th>QS</th>
<th>QS30</th>
<th>QS70</th>
<th>QS100</th>
<th>All data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1j40_1</td>
<td><img src="image1.png" alt="Image" /></td>
<td>Oxygen transport</td>
<td>4</td>
<td>D2</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>63</td>
<td>136</td>
</tr>
<tr>
<td>1luq</td>
<td><img src="image2.png" alt="Image" /></td>
<td>Biotin binding</td>
<td>4</td>
<td>D2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>23</td>
<td>123</td>
</tr>
<tr>
<td>4crx</td>
<td><img src="image3.png" alt="Image" /></td>
<td>Lyase</td>
<td>4</td>
<td>C4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td>1jkj_2</td>
<td><img src="image4.png" alt="Image" /></td>
<td>Succinyl CoA synthetase</td>
<td>4</td>
<td>C2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Displays homologous chains from the PDB in a new window

Click (requires firefox)
the constituent subunits in terms of domains and protein families, (ii) the pattern of binary interactions between the subunits, and (iii) the symmetry (if any) adopted by the subunits.

Figure 2.5 depicts the data accessible from the Web site www.3Dcomplex.org, where complexes are organized into a hierarchy. At the topmost level, the similarity of simple features is sufficient to group complexes, and these groups are split into subgroups toward the bottom of the hierarchy, where similarity of more detailed features (such as minimal sequence similarity of subunits) is required. The classification can be used to subselect specific groups of complexes such as: “tetramers where each subunit contacts two others, and where all subunits have a globin fold and exhibit D2 symmetry.”

3. INTERFACES IN HOMOMERS

3.1. Defining the protein surface and the interface

In a homomer, protein subunits are held together by interaction interfaces. To study these regions of contact between subunits, we first need to define them (i.e., their boundaries). There are several approaches for this purpose, and a first group of approaches involves the calculation of a “protein surface,” while a second is based on amino acid contacts (described in the last paragraph of this section). In terms of the first approach, several methodologies exist to calculate protein surfaces, a thorough description of which can be found at http://www.netsci.org/Science/Compchem/feature14.html. Below we briefly outline three methodologies.

One of the first methods to calculate protein surfaces was that of Lee and Richards.44 It consists of “rolling a ball” of arbitrary radius (a typical value is the radius of a water molecule, 1.4 Å) over the protein of interest. The surface accessible to the “rolling ball” is then defined as the solvent-accessible surface area (ASA). An implementation of this algorithm is available in the programs Naccess, b MSMS, c as well as in the program SURFACE that is part of the CCP4 package. d Another implementation from Michael Connolly e is also available.

A conceptually similar method consists in placing equidistant points on a sphere around each atom and then considering only those points that are not

b http://www.bioinf.manchester.ac.uk/naccess/.
c http://mgltools.scripps.edu/packages/MSMS.
d http://www.ccp4.ac.uk.
included in another sphere. This method was described in Ref. 48 and is implemented in the program AREAIMOL, also part of the CCP4 package, in the program DSSP, where the area is calculated by “geodesic sphere integration,” as well as in SurfRace.50

Another method is based on Voronoi and Voronoi-related tessellations. Such methods have been thoroughly reviewed in Ref. 51. An advantage of Voronoi-related tessellations is that they provide access to a variety of descriptors of the geometry of the surface, as well as of amino acid contacts at the interface. Bernauer et al. used such descriptors in combination with support vector machines to distinguish crystal contacts from biological interfaces. A recent software called “Intervor” enables the computation of such surfaces.52

Considering the first two methods, calculating the solvent-accessible areas of a complex does not directly yield information on the interaction interface. In order to obtain this information, one needs to measure the ASA of the complex (ASAcomplex) and that of each subunit individually (ASAsub1, ASAsub2). Thus, the surface buried at the interface can be obtained by taking ASAburied = (ASAsub1 + ASAsub2) - ASAcomplex. In order to measure the involvement of a residue at the interface, the same principle can be applied, that is, measure the ASA of the residue in the subunit alone and in complex. Depending on the goal of the study, it can be important to consider the relative ASA rather than the absolute ASA, as not all amino acids have the same size (i.e., a fully buried glycine is likely to bury less surface than a partially buried tryptophan).

Finally, protein–protein interfaces can be defined based on contacts between subunits rather than based on ASA. The definition can be loose, for instance, any pair of residues is considered in contact when their alpha or beta carbons are closer than a given cut-off value. It can be more detailed by taking into account side chains (e.g., van der Waals radii plus 0.5 Å), and even more detailed by taking into account salt bridges, hydrogen bonds, solvent molecules, and ions.54

3.2. Interface properties: Focusing on the amino acid composition

The literature on the general properties of protein–protein interfaces is so extensive that it cannot be thoroughly reviewed here. We thus refer the reader to the following books and reviews. In this section, we focus specifically on the amino acid composition of interfaces and how it

http://swift.cmbi.ru.nl/gv/dssp/.
varies across different regions within the interface. In particular, we will try to measure the extent to which interface and surface differ.

Protein–protein interfaces, in general, tend to be more hydrophobic than protein surfaces, and homomers have frequently been described as having more hydrophobic interfaces than hetero-oligomers. The same is true of obligate homomers, the interfaces of which appear more hydrophobic than those of transient homomers. To some extent, this is linked to differences in interface sizes, because polar amino acids tend to be exposed to the solvent, while hydrophobic amino acids tend to be buried. Considering a simplified elliptic interface, we note that the ratio “perimeter divided by surface” decreases as the surface increases. In other words, the perimeter, which is exposed to the solvent and therefore hydrophilic, represents a smaller fraction of larger interfaces. This “perimeter/surface” dichotomy is reflected in the core–rim model of protein–protein interfaces, where the rim is composed of residues having all of their atoms accessible to the solvent and the core comprises residues with at least one fully buried atom.

The core–rim model was extended with a third category, called the support. The support, rim, and core are illustrated in Fig. 2.6, which also depicts how they are defined. This definition relies first on the dichotomy of structures into two classes: the “surface” (relative ASA > 25%) and the “interior” (relative ASA < 25%). The value of 25% is chosen because it maximizes the difference in composition between these two regions. Based on this first dichotomy
- a rim residue is a surface residue in the structure alone AND in the complex
- a support residue is an interior residue in the structure alone AND in the complex
- a core residue is a surface residue in the structure alone and it becomes an interior residue in the complex.

![A definition of structural regions for residues at interfaces.](image-url)
An interesting property of this definition is that it allows one to compare structurally equivalent regions. This can be useful within the structure, for example, to compare the evolutionary conservation of the support with that of the interior, and the evolutionary conservation of the rim or core with that of the surface. It can also be useful across structures, because the definition yields regions that are stable in their composition even across interfaces of different sizes and of different types (Fig. 2.7). In other words, while we know that the global composition of interfaces changes with size, it seems that the breakdown into support, core, and rim makes individual regions more comparable to each other, even across different types of complexes.

### 4. CREATING INTERFACES IN HOMOMERS

#### 4.1. Are point mutations likely to create new interfaces?

The dichotomy of interfaces into the regions described above also enables us to ask the following question: from a simple amino acid composition perspective, “how many mutations are required to turn an average 1000 Å² surface patch into an average 1000 Å² interface patch?” Surprisingly, it was found that as little as two amino acid changes are sufficient. This illustrates that protein surfaces are in fact quite close to interfaces in terms of their...
chemical composition. This is also reflected in the fact that it is difficult to distinguish interface from surface based on amino acid composition alone. In other words, protein surfaces may be more “sticky” than commonly assumed.

This character suggests that it may be easier than commonly thought to create new protein interfaces. Surely, a number of factors such as shape and physicochemical complementarity must be considered to infer if two specific proteins are likely to interact, but the above result tells us about average stickiness properties. Interestingly, we saw in Section 2.2 that symmetric (isologous) interfaces are expected to be more frequently favorable energetically than heterologous interfaces. Consistent with this notion, there are several examples where a single or a couple of point mutations are associated with a dimerization event, but to our knowledge, there is no such example for heterologous interfaces.

The chemical proximity between surface and interface also suggests that functional, nonfunctional, and deleterious interactions are likely explored during the random course of evolution. One such example is illustrated by sickle cell anemia, where a single amino acid substitution is responsible for the formation of fibrils. Other examples are reflected in homologous proteins forming different types of dimers, that is, where the dimeric interface is situated in different regions of the structure on both homologues, as in, for example, sulfhydryl oxidases.

4.2. Molecular mechanisms associated to changes in oligomeric state

Analyses of homologous proteins forming different quaternary structures can reveal molecular mechanisms associated with evolutionary changes in homomeric state. Several molecular mechanisms have been observed, and we describe them below starting from the most indirect effects (i.e., where a mutation is away from the interface) to the most direct effects (i.e., where the mutations take place at the interface).

Perica et al. recently estimated that up to 50% of changes in oligomeric state might originate from allosteric effects, that is, they result from changes in sequence away from the interface. The PyrR family illustrates such a case, where a rotation of the subunits at the dimeric interface creates a geometric incompatibility at the tetrameric interface.

A distinct mechanism by which a mutation may induce a change in oligomeric state is through domain swapping. In this case, two or more proteins mutually exchange a part of their structure, usually at the N- or
C-terminus, with each other. An interesting case was recently described where a domain-swapping event rescued the stability of a monomeric protein. This suggests that homomers may sometime appear in evolution simply as a mechanism to rescue stability, and may not necessarily be driven by function per se.83

Several studies88–91 have carried out comprehensive comparisons of homologous protein pairs forming monomers and dimers. They focused on the interface region in the dimer and a structurally equivalent region in the monomer to identify differences that may explain changes in oligomeric state. Protein loops were found to have an important impact, both in promoting (enabling loops) and avoiding (disabling loops) dimerization.89 These loop regions have been characterized in terms of several properties including their length (mostly less than 10 amino acids) and their structure (mostly devoid of secondary structure).88 Another study focusing on the family of glycosyltransferases arrived at similar conclusions.92

5. EVOLUTION OF HOMOMER GEOMETRY AND DIVERSITY THROUGH GENE DUPLICATION

5.1. Probable oligomeric state transitions

In the previous section, we discussed molecular mechanisms that can explain evolutionary changes in oligomeric state. Below we ask first, how frequently do such changes take place during the course of evolution? And when they do, which oligomeric transitions are most common?

In recent years, a significant body of work has focused on “interface” conservation and “interface geometry” conservation,93–96 as exemplified by the many databases making this information available.27,97–99 These allow one to predict whether two proteins are likely to interact with each other and if they do, which geometry is adopted. Such information is very important for the structural modeling of protein interactions and to analyze the evolutionary conservation of individual interfaces. Importantly, however, information on binary interactions between domains is not sufficient to assess the conservation of homomers, which may involve two or more distinct interfaces. An estimate of the conservation of quaternary structure during evolution has been made based on the symmetry of homomers. This shows that at sequence identity levels of 90%, 50%, and 30%, there are 93%, 81%, and 70% of proteins that exhibit identical symmetry types, respectively.100 However, this estimate does not take domain geometry into account, so that two homologous proteins forming two C2 dimers may be
counted as having an identical quaternary structure although they exhibit distinct interfaces. In the future, it will therefore be important to systematically combine symmetry information with domain-interaction-geometry information in order to refine these numbers.

Nevertheless, these numbers are informative as they represent upper boundaries. This means that two sequences sharing 50% identity have similar quaternary structures in at most 81% of cases. What then happens in the remaining cases? As we have seen in Section 2.2, some symmetry types are geometrically closer to each other than others. In other words, while the transition from a C2 homomer to a D2 homomer can be achieved by creation of a single interface, the transition between a C2 and a C4 homomer would require many more events—namely, breaking the C2 interface and creating two complementary interfaces yielding a C4 symmetry. As a result, and as expected, we observe that geometrically favorable transitions occur more frequently than nonfavorable ones. One example is illustrated by the enzyme phosphoglycerate mutase, which exists as a dimer in *H. sapiens* and as a tetramer in *S. cerevisiae* (Fig. 2.8). We can see, however, that the *S. cerevisiae* tetramer is made of two dimers that are structurally almost identical to that found in *H. sapiens*. This notion is summarized in Fig. 2.9, where we highlight the geometrically most favorable transitions.

### 5.2. Evolution through gene duplication

Interestingly, changing oligomeric state is not the only option a homomer has to evolve. Another route involving gene duplication is frequently taken. Such an event turns a homomer into a “pseudohetero” protein complex, made of homologous proteins. This mechanism has been observed to occur frequently in protein networks, whereby homologous proteins tend to interact more frequently than would be expected by chance.101,102

---

**The phosphoglycerate mutase changed oligomeric state during evolution**

*B. stearothermophilus* | *H. sapiens* | *S. cerevisiae*
---|---|---
C1 | C2 | D2

*Figure 2.8* Changes in oligomeric state among homologues are often related geometrically.
As a matter of fact, many protein complexes are the product of such a scenario. Well-known examples are hemoglobin, RNA polymerases, the photosystem I, the CCT complex, the 20S proteasome, and several transcription factor families such as the Myc/Max factors or NF-kB family which form heterodimers of homologous subunits. All these examples illustrate that symmetry was present at the origin of very fundamental molecular machines and led to “pseudosymmetry,” after gene duplication occurred. Symmetry breaking by gene duplication enables each subunit to evolve independently, to regulate the system and be regulated in an independent fashion. The photosystem I complex illustrates this well, where each duplicate of the “ancestral homodimer” binds different sets of proteins.

We summarize this section with a hypothetical pathway of how the proteasome may have evolved from a single domain into a large and complex cellular machine (Fig. 2.10). All subunits of the 20S proteasome share a similar structure, which belongs to the “N-terminal nucleophile aminohydrolase domain” family, suggesting that they descend from a common ancestral protein. The scenario starts with a single domain (pdb: 1xfg, glutaminase domain\textsuperscript{103}), which may have formed a dimer or a hexameric ring (no such intermediates are known to our knowledge). Six copies of the dimer (6 \times C2) or two copies of the hexamer (2 \times C6) could have led to a structure similar to the Heat Shock Locus V protein complex (D6, pdb: 1ned\textsuperscript{104}). A gene duplication event would have allowed two D6 complexes to be stacked on top of each other to form a “D6 dimer.” (Note that

---

**Figure 2.9** Model of quaternary structure evolution in homomers, as proposed in Ref. 100.
without gene duplication, this would not be possible as it would yield a fibril.) Such a dimerization event and an “opening” of the ring to accommodate one additional subunit yield a structure similar to the archaeal proteasome (pdb: 1pma\textsuperscript{105}). Finally, successive gene duplications yield the eukaryotic 20S proteasome characterized in yeast (pdb: 1ryp\textsuperscript{106}).

6. ASSEMBLY OF HOMOMERS

6.1. Assembling into dimers

After being synthesized by the ribosome, polypeptide chains must find each other and assemble in order to form a homomer. To describe this process, we will distinguish two types of assembly events: (i) “binding steps,” which we use to refer to two subunits or groups of subunits binding each other, and (ii) “assembly pathway,” which we use to describe the sequence of binding steps followed in the assembly of a homomer. Considering a dimer, a single binding step (and therefore a single pathway) can be followed. The binding step can nevertheless take different forms. Schematically, the polypeptide chains can undergo folding and binding either simultaneously (two-state binding) or sequentially (three-state binding). Such a dichotomy is also reflected in the distinction frequently found in the literature between “obligate interactions” (where subunits are not stable on their own) and “non-obligate interactions” (where subunits can exist in their folded state in the absence of their partner).\textsuperscript{63} Of course, many cases are situated in between these two extreme scenarios.
Interestingly, the native-state topology of the protein can in fact be used to infer the extent to which folding and binding are likely to occur simultaneously.\textsuperscript{107–110} Simply put, the more the interface contributes to the stability of the native state, the more it is likely to involve simultaneous folding and binding. In other words, the larger the area buried at the interface relative to the area buried within each subunit, the more likely it is that the binding is coupled to folding.\textsuperscript{111}

It is tempting to link these assembly mechanisms to evolution. One could indeed expect that a stable monomer evolving and forming a dimer is likely to assemble through a “three-state” mechanism where the folded intermediate corresponds to the ancestral monomer.\textsuperscript{112,113} In contrast, dimers burying large interfaces and assembling through a two-state binding may be less likely to exist (or have existed) as monomers in other species. Little is known about this notion, which will be worth investigating in the future.

6.2. Assembling into higher-order oligomers

In 1979, the assembly pathway for many homomers and hetero-oligomers had already been characterized,\textsuperscript{114} and in the following two decades, many more were characterized.\textsuperscript{115,116} Today, the assembly pathway best characterized might be that of the tetramerization domain of p53.\textsuperscript{117,118} It takes place in two binding steps: the first is a two-state binding of two monomers into a dimer, and the second is the binding of two folded dimers to form a tetramer.

This assembly pathway is frequently observed among tetramers, that is, where a dimer forms first and two dimers subsequently assemble into a tetramer. This can be expected among D2 tetramers, which often involve two interfaces of different strengths. The assembly pathway observed for p53 results from a “hierarchy” between interfaces, with a strong and a weak interface. Such a hierarchy in the affinity between the two possible dimers has the advantage of favoring a single pathway and minimizing complexes trapped in non-native states.\textsuperscript{119} Assembly through a well-defined pathway becomes even more essential if considering large homomers such as virus capsids. For example, considering the assembly of an icosahedron formed by 60 subunits, the number of possible assembly pathways reduces drastically if considering solely pre-assembled C5 pentamers, 12 of which may form a complete capsid.\textsuperscript{120} Although this is a simplification of a highly complex process,\textsuperscript{121} it illustrates that introducing a hierarchy of interface strength can favor a specific assembly pathway.
6.3. When assembly recapitulates evolution

We have seen in Section 6.1 how the assembly pathway of a dimer can be reflected in its evolution. Regarding higher-order homomers, a similarity between “ontogeny” and “phylogeny” can also be inferred, although based on a different rationale.

The rationale is that a hierarchy of interface sizes is naturally expected. The expected area buried within an oligomer follows a simple power law relationship: \( \text{ASA} = 5.3 \cdot M^{0.76} \), where ASA is the accessible surface area in \( \text{Å}^2 \) and \( M \) the molecular weight in daltons. The expected ASA of a 30-kD protein is therefore \( \sim 13,400 \text{ Å}^2 \). If that protein forms dimers and tetramers, their expected ASA will be 22,700 and 38,400 \( \text{Å}^2 \), respectively. We can thus calculate the expected area buried in the formation of a dimer \( (2 \times 13,400 - 22,700 = 4100 \text{ Å}^2) \) and a tetramer \( (2 \times 22,700 - 38,400 = 6950 \text{ Å}^2) \). Importantly, in the formation of the tetramer, two interfaces (and up to four) are formed, so that the size obtained must be divided by two (or four). In sum, a 30-kD monomer that evolves into a dimer is expected to bury 4100 \( \text{Å}^2 \) at the dimeric interface, and the same dimer that evolves to a tetramer is expected to bury from 1740 to 3745 \( \text{Å}^2 \) at the tetrameric interface(s). This illustrates why, on average, interfaces associated with higher oligomeric transitions \((C_n \leftrightarrow D_n)\) are expected to be smaller than those associated to lower oligomeric transitions \((C_1 \leftrightarrow C_n)\)—as depicted in Fig. 2.9B.

An implication of this observation, which was confirmed in \( \sim 80\% \) of homomers with different symmetry, is that large interfaces in dihedral homomers are likely to be conserved in cyclic homologues. In addition, considering that interface size is indicative of “affinity,” it suggests that the predicted evolutionary intermediate in a complex is also likely to represent an assembly intermediate.

7. CONCLUSION

We have seen that homomers are frequent, with up to 50% of proteins forming such structures. Importantly, however, most proteins do not act alone in cells. In other words, most of the homomers observed in structural data are expected to form higher-order complexes in cells. This is illustrated in a recent study of protein complexes in *Mycoplasma pneumonia*, where it was estimated that up to 47% of soluble proteins have stoichiometries larger than one, that is, more than one copy of the protein is observed in the
complex they form. This suggests that in future studies, it will be particularly interesting to consider the properties of homomers in the broader context of complexes made up of hetero-interactions as well. This will likely reveal new functional and biophysical properties associated with the cellular machinery.

ACKNOWLEDGMENTS

We thank Joël Janin, Peter Rose, Tina Terica, Joseph Marsh, Eviatar Natan, and Ezra Peisach for their comments on this chapter.

REFERENCES


