STRUCTURAL BIOLOGY

ISSN 2059-7983

Empirical power laws for the radii of gyration of protein oligomers

John J. Tanner*

Departments of Biochemistry and Chemistry, University of Missouri-Columbia, Columbia, MO 65211, USA.
*Correspondence e-mail: tannerjj@missouri.edu

Received 16 April 2016
Accepted 16 August 2016

Edited by T. O. Yeates, University of California, USA

Keywords: radius of gyration; protein oligomerization; structural bioinformatics; small-angle X-ray scattering; structure validation.

Supporting information: this article has supporting information at journals.iucr.org/d

The radius of gyration is a fundamental structural parameter that is particularly useful for describing polymers. It has been known since Flory's seminal work in the mid-20th century that polymers show a power-law dependence, where the radius of gyration is proportional to the number of residues raised to a power. The power-law exponent has been measured experimentally for denatured proteins and derived empirically for folded monomeric proteins using crystal structures. Here, the biological assemblies in the Protein Data Bank are surveyed to derive the power-law parameters for protein oligomers having degrees of oligomerization of 2–6 and 8. The power-law exponents for oligomers span a narrow range of 0.38–0.41, which is close to the value of 0.40 obtained for monomers. This result shows that protein oligomers exhibit essentially the same power-law behavior as monomers. A simple power-law formula is provided for estimating the oligomeric state from an experimental measurement of the radius of gyration. Several proteins in the Protein Data Bank are found to deviate substantially from power-law behavior by having an atypically large radius of gyration. Some of the outliers have highly elongated structures, such as coiled coils. For coiled coils, the radius of gyration does not follow a power law and instead scales linearly with the number of residues in the oligomer. Other outliers are proteins whose oligomeric state or quaternary structure is incorrectly annotated in the Protein Data Bank. The power laws could be used to identify such errors and help prevent them in future depositions.

1. Introduction

The radius of gyration (R_g) is a fundamental parameter of molecular structure. R_g is defined as the mean-square distance from the center of the molecule, as in (1*a*),

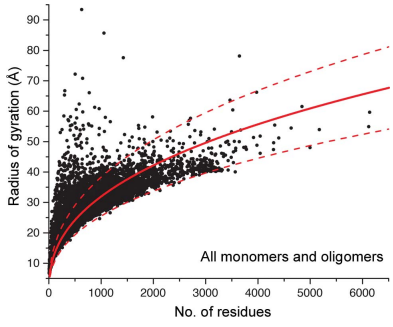
$$R_g^2 = N_{at}^{-1} \sum (\mathbf{r}_i - \mathbf{r}_o)^2. \qquad (1a)$$

In (1*a*), N_{at} is the number of atoms in the molecule, $\mathbf{r}_i$ is the vector from the origin to atom i and $\mathbf{r}_o$ is the vector from the origin to the geometrical center of the molecule (defined as $N_{at}^{-1} \sum \mathbf{r}_i$). Some treatments use a mass-weighted R_g as in (1*b*) (Ivankov *et al.*, 2009),

$$R_g^2 = M^{-1} \sum m_i (\mathbf{r}_i - \mathbf{r}_{com})^2. \qquad (1b)$$

In the mass-weighted version, M is the total mass of the molecule, m_i is the mass of atom i and $\mathbf{r}_{com}$ is the center-of-mass vector, defined as $M^{-1} \sum m_i \mathbf{r}_i$.

R_g is particularly useful for describing polymer structure. Flory showed that the average linear dimension of a real polymer, *i.e.* a three-dimensional chain in which the monomers cannot occupy the same location in space ('excluded volume polymer'), is proportional to the chain length raised to the power 0.6 (Flory, 1949). This seminal result of polymer physics is often expressed as a power-law relationship, such as



© 2016 International Union of Crystallography

$$R_g = R_0 N^{\nu}. \quad (2)$$

In (2), N is the number of residues in the polymer, R_0 is a prefactor and ν is an exponent that approaches the Flory value of 0.6 for long-chain polymers in good solvents, *i.e.* when polymer–solvent interactions are favored over polymer–polymer interactions (Thirumalai, 2000; Wilkins *et al.*, 1999; see chapter 1 of de Gennes, 1979 or Doi, 1996). The exponent ν is a universal constant and is interpreted to be a measure of the compactness of the polymer. The prefactor R_0 depends on the chemical details of the monomer and the bond geometry. R_0 and ν are relatively insensitive to the types of secondary-structure elements (α-helices or β-sheets) in the protein (Hong & Lei, 2009). As described by Hinsen and coworkers, it is possible, in principle, to estimate R_0 from knowledge of the covalent bonding in the polymer, the chemical details of the solvent, and variable physical and chemical factors such as temperature and pH (Hinsen *et al.*, 2013). It is also possible to obtain R_0 experimentally, and such studies have yielded values of R_0 for proteins of approximately 2.0 Å (Wilkins *et al.*, 1999; Hofmann *et al.*, 2012; Kohn *et al.*, 2004).

Decades after Flory's original work, (2) was tested for denatured proteins. The fundamental question was whether the universal exponent of 0.6 applies, considering that denatured proteins are thought to retain considerable residual structure. For example, Kohn and coworkers used small-angle X-ray scattering (SAXS) to show that several small, chemically denatured monomeric proteins exhibit power-law behavior with an exponent of 0.598, which is very close to the Flory value (Kohn *et al.*, 2004). Hofmann and coworkers likewise obtained ν close to 0.6 at high denaturant concentration using single-molecule experiments, but interestingly observed sequence-dependent deviations at lower denaturant concentrations (Hofmann *et al.*, 2012).

Computational scientists realised that R_g could be a powerful constraint in protein structure modeling, given an *a priori* estimate from the sequence. This idea motivated studies of the power-law dependence of monomeric proteins in their native conformations. Several studies showed that the Flory equation applies, but with an exponent smaller than 0.6. The smaller exponent reflects the compaction of the polymer owing to secondary-structure and tertiary-structure elements in the native protein conformation. For example, Skolnick's group reported an exponent of 0.38 for single-domain proteins (Kolinski *et al.*, 1993). Surveys of monomeric proteins in the Protein Data Bank (PDB; Berman *et al.*, 2000) yielded exponents of 0.33–0.34 (Dima & Thirumalai, 2004; Hofmann *et al.*, 2012; Gong *et al.*, 2005). Calculations based on the asymmetric units of crystal structures, which do not necessarily correspond to biologically relevant oligomers, produced exponents of 0.365–0.39 (Hong & Lei, 2009; Hinsen *et al.*, 2013).

Here, (2) is used to study the R_g of protein oligomers. Oligomerization is an important aspect of protein structure. Surveys of the PDB suggest that 30–50% of proteins form homooligomers (Goodsell & Olson, 2000; Levy & Teichmann, 2013). Oligomerization often plays a role in function. For example, substrate-binding and cofactor-binding sites can occur in oligomer interfaces (Marsh & Teichmann, 2015). Some transcriptional repressors, such as ribbon–helix–helix proteins, bind DNA as obligate dimers (Schreiter & Drennan, 2007). Many integral membrane secondary transporter proteins function as oligomers (Veenhoff *et al.*, 2002). Oligomerization is essential to the function of the potassium channel, since the ion-conduction pathway coincides with the fourfold axis of a homotetramer (MacKinnon, 2003). Quaternary structure sometimes contributes to the regulation of protein function. For example, most allosteric proteins are oligomers (Perica *et al.*, 2012), enzyme cooperativity usually requires oligomerization (Cornish-Bowden, 2014), and changes in quaternary structure underlie the regulation of morpheein enzymes (Jaffe, 2005). Oligomerization also contributes to protein stability (Marsh & Teichmann, 2015).

Table 1
Database statistics and power-law parameters.

n	No. of structures	ν†
1	11298	0.405 ± 0.002
2	8319	0.407 ± 0.002
3	1041	0.400 ± 0.010
4	2054	0.394 ± 0.004
4, cyclic	412	0.404 ± 0.007
4, dihedral	1592	0.390 ± 0.003
5	94	0.394 ± 0.001
6	640	0.391 ± 0.006
6, cyclic	122	0.407 ± 0.001
6, dihedral	511	0.387 ± 0.006
8	253	0.385 ± 0.009
All	23699	0.401 ± 0.001

† R_0 fixed at 2.0 Å in (2).

Given the propensity of proteins to form oligomers, and the intimate relationship between quaternary structure and protein function, it is important to understand protein oligomerization and to have robust methods for characterizing oligomeric state and quaternary structure. For the present purpose, the term 'protein oligomer' is defined as a protein–protein complex that can be purified and characterized structurally (Marsh & Teichmann, 2015). The surfaces and interfaces of such oligomers have been extensively characterized. For example, the accessible surface area of protein oligomers is related to the molecular mass through a power law, similar to a relationship that has been described for monomers (Miller *et al.*, 1987), and the composition of protein oligomer interfaces (*e.g.* amino-acid composition, interfacial area, hydrogen boding, ion pairs, complementarity *etc.*) has been cataloged (Janin *et al.*, 1988; Jones & Thornton, 1995, 1996; Keskin *et al.*, 2008). This and other information has been used to identify stable oligomers from crystal structures (Ponstingl *et al.*, 2000, 2003; Henrick & Thornton, 1998; Krissinel & Henrick, 2007). In particular, the *PISA* algorithm, which is based on physical-chemical models of protein interactions and chemical thermodynamics, is used by the PDB to identify the most likely biological assembly of deposited structures (Krissinel & Henrick, 2007).

Here, the PDB is surveyed to study the dependence of R_g on the number of residues for protein oligomers. The analysis shows that biologically relevant oligomers generally follow

power-law behavior. In contrast, highly extended proteins deviate substantially from power-law dependence, and this deviation can be used to identify proteins with atypical aspect ratios. For proteins that are not highly extended, a simple power-law formula is derived for estimating the oligomeric state from an experimental measurement of the radius of gyration. Deviation from power-law dependence also occurs for proteins whose oligomeric state or quaternary structure is incorrectly annotated in the PDB. The derived power laws may be useful for identifying and correcting misannotated PDB entries.

2. Methods

The advanced search tool of the RCSB PDB was used to obtain data sets for protein homo-oligomers. The data sets were culled from crystal structures with the following characteristics: (i) deposited between 1 January 2000 and 27 February 2016, (ii) a resolution of 2.5 Å or better and (iii) a pairwise sequence identity less than or equal to 90%. The degree of oligomerization (n) was selected using the 'Number of Chains (Biological Assembly)' search criterion. This criterion searches 'BIOMOLECULE 1' of REMARK 350 and retrieves 'Biological Assembly 1' of the PDB entry, which has the file extension `.pdb1`. Homooligomers were selected using the 'Protein Stoichiometry' search criterion,

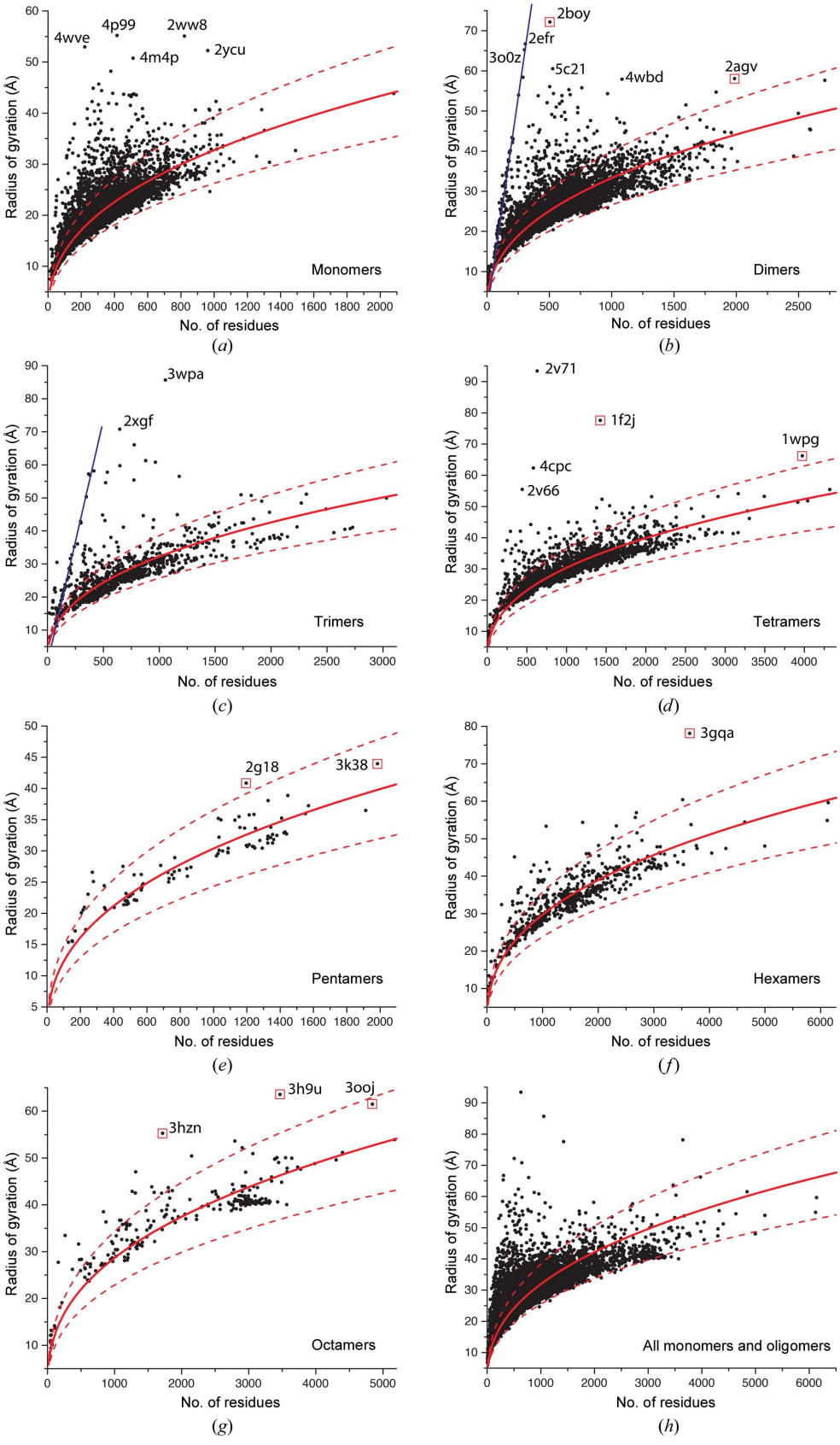


Figure 1
Empirical power laws for protein monomers and oligomers. Plots of R_g *versus* the total number of residues in the particle for (*a*) monomers, (*b*) dimers, (*c*) trimers, (*d*) tetramers, (*e*) pentamers, (*f*) hexamers, (*g*) octamers and (*h*) monomers and all oligomers. Red boxes indicate misannotated oligomers. In each panel, the solid curve represents the best fit to the equation $R_g = R_0N^{\nu}$, with R_0 fixed at 2.0 Å and N equal to the total number of residues in the particle. The dashed curves represent ±20% deviation from the fitted power-law curve. The percentage of structures within the ±20% boundaries are as follows: 94% for monomers, 92% for dimers, 87% for trimers, 92% for tetramers, 90% for pentamers, 92% for hexamers, 88% for octamers and 92% for all structures. The blue lines in (*b*) and (*c*) represent linear fits for dimeric and trimeric coiled coils, respectively.

with 'A' chosen to select monomers and 'A*n*' to select homo-oligomers with degree of oligomerization *n*. Table 1 lists the numbers of structures used in the calculations. The accession codes of the structures are provided in the Supporting Information.

R_g was calculated using *MOLEMAN* v.041001/7.4 (Kleywegt *et al.*, 2001) called by a Linux script (provided in the Supporting Information). The R_g calculated from *MOLEMAN* uses (1*a*), *i.e.* the atomic vectors are not mass-weighted and the center is the geometrical center of the protein, not the center of mass (Gerard Kleywegt, personal communication). Note that for proteins, (1*a*) and (1*b*) yield similar results (typically <1% difference) because the predominant heavy atoms in proteins (N, C, O) have similar masses. The 'Biological Assembly' structure entries (file extension `.pdb1`) were used for all calculations. Only protein atoms were included in the R_g calculations. The calculated R_g values are provided in the Supporting Information.

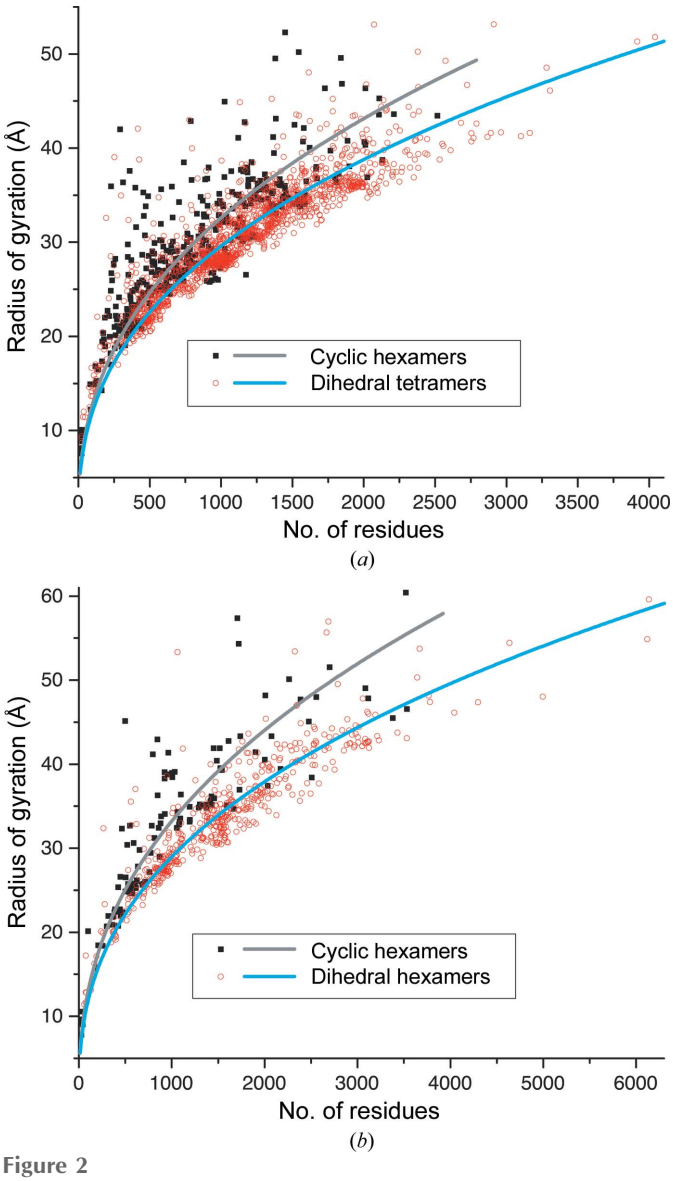


Figure 2
Power-law curves for cyclic and dihedral oligomers. (*a*) The tetramer data set was divided into assemblies with cyclic (black squares) or dihedral (open red circles) symmetry. The associated power-law curves are shown in gray (cyclic) and cyan (dihedral). (*b*) The hexamer data set was divided into assemblies with cyclic (black squares) or dihedral (open red circles) symmetry. The associated power-law curves are shown in gray (cyclic) and cyan (dihedral).

Table 2
Estimation of oligomeric state from SAXS data.

				R_g power law		Volume of correlation†	
Protein‡	L	SAXS R_g‡ (Å)	Known n‡	Predicted n	Error	Predicted n	Error
Phl p 7	77	12.9	1	1.4	0.4	1.0	0.0
ALDH7A1	511	37.7	4	3.0	−1.0	3.2	−0.8
AfUGM	513	47.3	4	5.3	1.3	3.7	−0.3
BhP5CDH	515	31.3	2	1.9	−0.1	1.6	−0.4
TtP5CDH	516	43.4	6	4.3	−1.7	5.4	−0.6
BjPutA	999	56.0	4	4.2	0.2	3.5	−0.5
GsPutA	1005	43.6	2	2.2	0.2	1.7	−0.3
RcPutA	1127	32.3	1	0.9	−0.1	0.9	−0.1
EcPutA	1320	62.8	2	4.2	2.2	1.6	−0.4

† Calculated using *SCÅTTER* 1.0 (Rambo, 2015). ‡ Abbreviations and references for the SAXS R_g and known *n*: Phl p 7, timothy grass polcalcin (Henzl *et al.*, 2013); ALDH7A1, aldehyde dehydrogenase 7A1 (Luo & Tanner, 2015); AfUGM, *Aspergillus fumigatus* UDP-galactopyranose mutase (Dhatwalia *et al.*, 2012); BhP5CDH, *Bacillus halodurans* Δ^1-pyrroline-5-carboxylate dehydrogenase (Luo *et al.*, 2013); TtP5CDH, *Thermus thermophilus* Δ^1-pyrroline-5-carboxylate dehydrogenase (Luo *et al.*, 2013); BjPutA, *Bradyrhizobium japonicum* proline utilization A (Srivastava *et al.*, 2010); GsPutA, *Geobacter sulfurreducens* proline utilization A (Singh *et al.*, 2014); RcPutA, *Rhodobacter capsulatus* proline utilization A (Luo *et al.*, 2014); EcPutA, *Escherichia coli* proline utilization A (Singh *et al.*, 2011).

Power-law parameters were obtained by nonlinear curve fitting to (2) using *Origin* 2016 with R_0 fixed at the experimental value for proteins of 2.0 Å (Wilkins *et al.*, 1999; Hofmann *et al.*, 2012; Kohn *et al.*, 2004). Fixing R_0 avoids the problem of excessive correlation between R_0 and ν that occurs when both parameters are allowed to refine. In all the calculations, N in (2) is the total number of residues in the oligomer, *i.e.* the number of subunits times the number of residues per subunit. N was calculated directly from the atomic coordinates, not from the amino-acid sequence, since residues in the sequence can be absent in the structure.

3. Results

3.1. Power laws for protein oligomers

The radii of gyration of monomers and homooligomers with degree of oligomerization n = 2–6 and 8 were calculated from structures obtained from the PDB. Plots of R_g *versus* the total number of residues in the particle are shown in Fig. 1. Table 1 lists the power-law exponents (ν) calculated with the constraint of R_0 fixed at the experimental value of 2.0 Å. The exponent for monomers (n = 1) is 0.40, which is in the range of 0.34–0.40 found in other studies (Skolnick *et al.*, 1997; Gong *et al.*, 2005; Hinsen *et al.*, 2013; Hong & Lei, 2009). The exponents for oligomers are also near 0.40, spanning the narrow range 0.38–0.41 (Table 1). This result shows that oligomers and monomers follow essentially the same power law, which was not previously appreciated. Combining the monomers and oligomers into one graph yields an exponent of 0.40 (Fig. 1*h*).

Although these calculations show qualitatively that protein oligomers exhibit power-law behavior, note that there is a vertical spread in the data points above and below the fitted curves. To evaluate the spread in the data, boundary curves representing ±20% deviation from the fitted power-law curves are included in Fig. 1 (dashed curves). The fraction of the structures within the ±20% boundary is approximately 90%, and ranges from 87% for trimers to 94% for monomers, which suggests that the power law reasonably captures the dependence of R_g on the number of residues.

Cyclic and dihedral symmetries are possible for homomers with an even number of subunits and a degree of oligomerization greater than 2, and it is possible that these two different types of oligomers have different exponents. To test this idea, the cyclic and dihedral oligomers were fitted separately for the tetramer and hexamer groups (Fig. 2). In both cases, for a given number of residues, cyclic oligomers tend to have a larger R_g than the dihedral oligomers. Accordingly, the cyclic oligomer exponent is slightly larger (by ~5%) than the corresponding dihedral oligomer exponent (Table 1).

3.2. Predicting oligomeric state from an experimental measurement of R_g

The observation that protein oligomers generally follow a power-law dependence suggests a simple method for estimating the oligomeric state (n) when an experimental measurement of R_g is available. (2) can be rearranged to solve for the order of oligomerization, n,

$$n = L^{-1}(R_g/R_0)^{1/\nu}. \qquad (3)$$

In (3), L is the polypeptide chain length for a monomer, which is known from the gene sequence. Inserting the power-law parameters derived from the analysis of all monomers and oligomers (Fig. 1*h*) into (3) provides a formula for estimating the degree of oligomerization from R_g,

$$n = L^{-1}(R_g/2.0)^{2.5}. \qquad (4)$$

To illustrate the method, (4) was used to predict the oligomeric states from the SAXS R_g for several proteins from the author's laboratory (Table 2). For five of the nine proteins, the predicted n is within 0.4 subunits of the known value, meaning that the prediction is essentially correct in these cases. For two other proteins the method is reasonably accurate, predicting the oligomeric state to within one subunit of the correct value (AfUGM and ALDH7A1). The method performed less satisfactorily for only two proteins, where the estimated n value is two subunits away from the known value. The poor prediction of the oligomeric state for *Escherichia coli* proline utilization A is likely to be owing to the unusual shape of this protein. SAXS shape reconstructions show that the dimeric particle is highly elongated, with dimensions of 205 × 85 × 55 Å (Singh *et al.*, 2011). As shown next, highly elongated proteins deviate substantially from power-law behavior.

3.3. Highly elongated proteins deviate from power-law behavior

Several proteins are notable in that they deviate substantially from the power laws by having unusually high R_g values. Some of these outliers are proteins that have very elongated structures.

The monomer power-law plot reveals five proteins that have R_g > 50 Å (Fig. 1*a*). This group includes bacterial adhesion

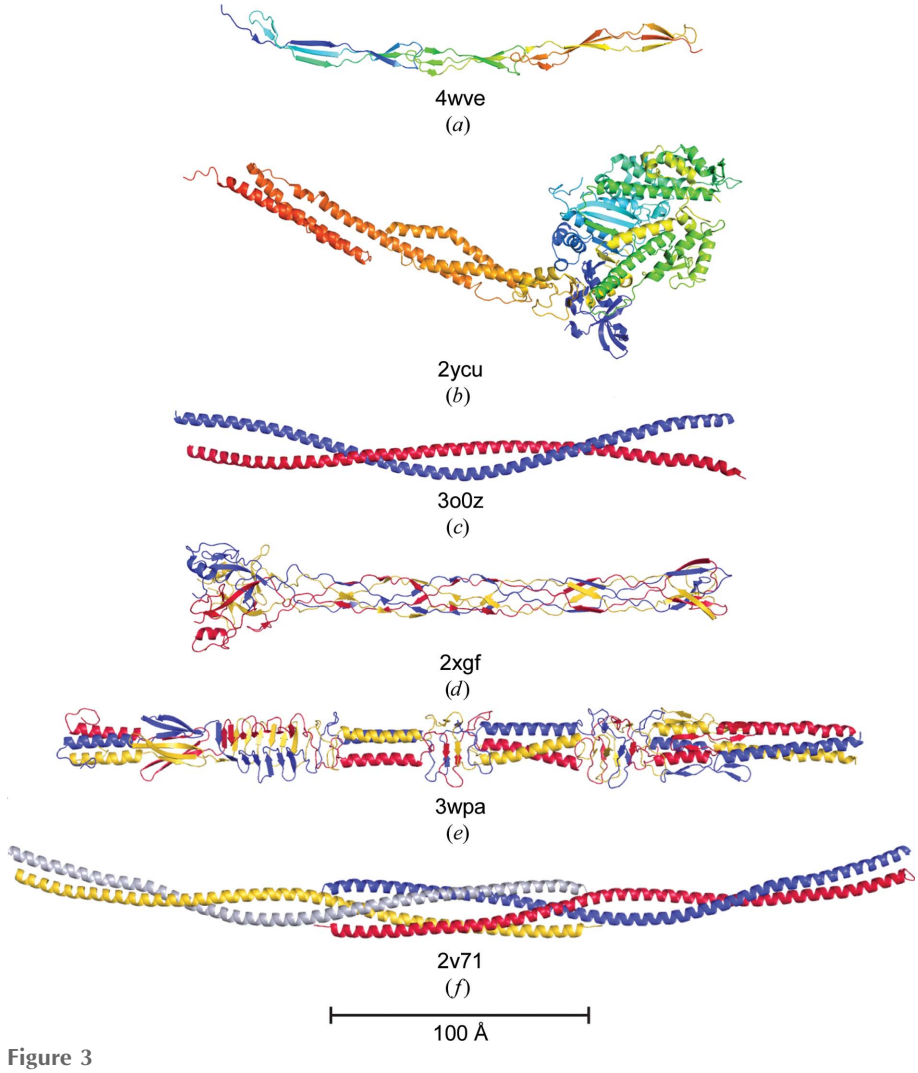


Figure 3
Examples of highly elongated proteins which deviate from power-law behavior. The PDB code is listed for each structure. In the monomers (*a*, *b*), the chain is colored in a rainbow scheme with blue at the N-terminus and red at the C-terminus. In the oligomers (*c*–*f*), each protomer is in a different color. The structures are displayed on a common length scale (see the scale bar at the bottom of the figure).

proteins [PDB entries 4p99 (Vance *et al.*, 2014), 4wve (Gruszka *et al.*, 2015) and 2ww8 (Izoré *et al.*, 2010)], a fragment of nonmuscle myosin 2C (PDB entry 2ycu) and the EphA4 ectodomain (PDB entry 4m4p; Xu *et al.*, 2013). The adhesion proteins share a common feature of tandem domain repeats, as shown for PDB entry 4wve (Fig. 3*a*). The myosin protein is over 200 Å in length because of an extended α-helical domain (Fig. 3*b*). Analysis of the protein–protein interfaces within these crystal structures using *PDBePISA* (Krissinel & Henrick, 2007) indicates no stable quaternary structures, thus their annotation in the PDB as monomers is reasonable. Nevertheless, it is likely that some of these proteins form higher order homoassemblies or heteroassemblies when they are in their respective biological complexes.

Highly elongated dimers also appear as outliers in the power-law plots. Notable dimers include a tropomyocin fragment (PDB entry 2efr; S. Minakata, Y. Nitanai, K. Maeda, N. Oda, K. Wakabayashi & Y. Maeda, unpublished work) and the central domain of a Rho-associated kinase (PDB entry 3o0z; Tu *et al.*, 2011). These proteins have $R_g > 65$ Å, while other dimers with a similar number of residues have an R_g of 18–30 Å (Fig. 1*b*). Both proteins form long α-helical coiled-coil structures that span over 200 Å, as shown for PDB entry 3o0z (Fig. 3*c*). Tropomyocins are known to dimerize (Gimona, 2008), and dimerization of the kinase has been confirmed by multi-angle light scattering (Tu *et al.*, 2011).

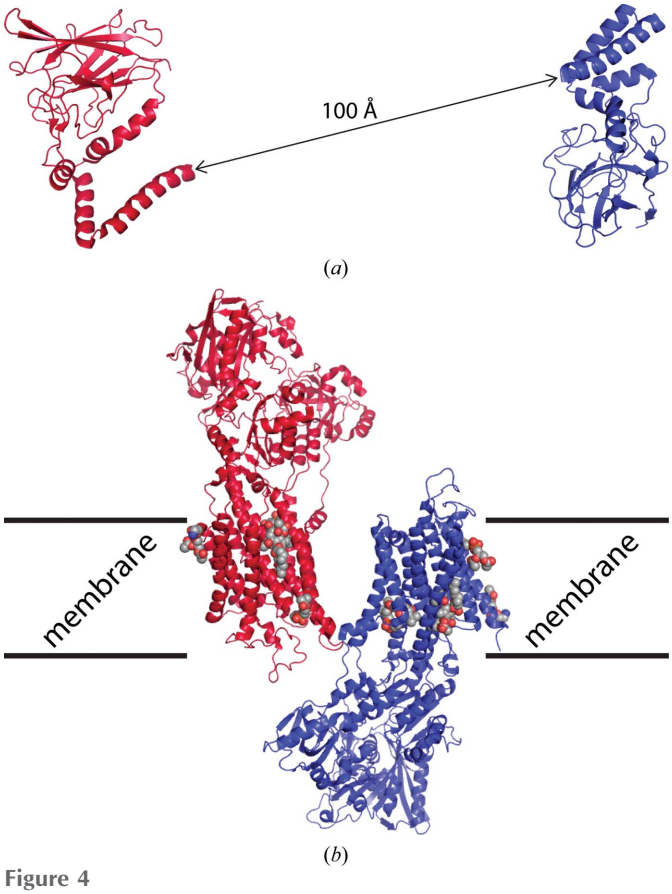


Figure 4
Two misannotated dimers. (*a*) Biological Assembly 1 of PDB entry 2boy. (*b*) Biological Assembly 1 of PDB entry 2agv. Detergent molecules are drawn as spheres. In both panels, each protomer of the assembly is in a different color.

Elongated trimers are seen as outliers from the power-law curve (Fig. 1*c*). For example, the receptor-binding tip of the bacteriophage T4 long tail fiber (PDB entry 2xgf) forms a interwoven trimer of 210 Å in length (Fig. 3*d*; Bartual *et al.*, 2010). The longest trimer in the data set is the *Acinetobacter* autotransporter adhesin (PDB entry 3wpa), which forms a remarkable particle of 310 Å in length (Fig. 3*e*; Koiwai *et al.*, 2016).

Interestingly, the highly elongated coiled-coil oligomers form a line of points in the power-law plots. This feature is prominent in the dimer and trimer plots, where one observes a group of structures forming a line with a high positive slope (blue lines in Figs. 1*b* and 1*c*). Linear regression yields slopes of 0.21 Å per residue for dimeric coiled coils and 0.15 Å per residue for trimeric coiled coils.

A very long tetramer is notable. The coiled-coil domain (residues 58–169) of Ndel1 (formerly NudEL; PDB entry 2v71) forms an α-helical structure of 350 Å in length (Fig. 3*f*). Its R_g of 93.4 Å is the largest of the entire data set (Fig. 1*d*). Although this protein is primarily dimeric, analytical ultracentrifugation provides evidence of a low-affinity tetramer in solution (Derewenda *et al.*, 2007); the four-body assembly in the crystal (Fig. 3*f*) could represent this tetramer.

In summary, proteins with rare aspect ratios deviate from the power laws. These proteins have truly remarkable shapes (Fig. 3). Proteins that closely follow power-law behavior have ordinary aspect ratios. Thus, the power laws can be used to identify atypically long proteins and provide an objective basis for using terms such as 'remarkably extended' and 'highly elongated' to describe protein structures.

3.4. Misannotated oligomers deviate from power-law behavior

Several outliers from the power laws are proteins whose oligomeric state is incorrectly annotated in the PDB entry. A structure of 3-chlorocatechol 1,2-dioxygenase (PDB entry 2boy; Ferraroni *et al.*, 2006) provides an example of an egregious misannotation. The space group is *P*1 with eight molecules in the asymmetric unit. The PDB Biological Assembly 1 has an R_g of 72.1 Å, which appears as an extreme outlier from the power law (Fig. 1*b*). Note that this assembly has the largest R_g of the dimer data set. The assembly consists of chains *A* and *E* of the asymmetric unit, which are separated by over 100 Å (Fig. 4*a*). Obviously, this assembly has no biological significance.

The dimer power-law plot also reveals a Ca^{2+}-ATPase (PDB entry 2agv) that is incorrectly annotated as a dimer (Obara *et al.*, 2005). Biological Assembly 1 of PDB entry 2agv has an R_g of 58.0 Å (Fig. 1*b*) and consists of the two molecules in the asymmetric unit (Fig. 4*b*). This assembly contradicts our basic understanding of P-type ATPases (Bublitz *et al.*, 2011). Ca^{2+}-ATPases are integral membrane proteins that transfer two Ca^{2+} ions from the cytoplasm into the lumen of the sarcoplasmic reticulum per ATP hydrolyzed. The soluble domains

of Ca^{2+}-ATPases are always located on the cytoplasmic face of the membrane. Biological Assembly 1 is obviously wrong because the cytoplasmic domains of the two protomers are on opposite sides of the implied membrane (Fig. 4*b*).

An aldolase structure provides a good example of a tetramer whose quaternary structure is incorrectly assigned in the PDB. Biological Assembly 1 of PDB entry 1f2j (Chudzik *et al.*, 2000) has an R_g of 77.5 Å (Fig. 1*d*), which is the second largest R_g of the tetramer data set. Biological Assembly 1 of PDB entry 1f2j consists of two dimers separated by over 100 Å (Fig. 5*a*). Although the aldolase is indeed tetrameric, this assembly is obviously irrelevant. PDB Biological Assembly 2 provides the correct tetramer, which is a dimer of dimers with an R_g of 34.4 Å (Fig. 5*b*).

A Ca^{2+}-ATPase (PDB entry 1wpg) is misannotated as a tetramer. As described above for PDB entry 2agv, the Biological Assembly 1 of PDB entry 1wpg (Toyoshima *et al.*, 2004) is an artifact of crystallization because the cytoplasmic domains of the protomers are not on the same face of the membrane (not shown).

The pentamer power-law plot reveals a neuraminidase that is incorrectly listed as a pentamer by the PDB (PDB entry 3k38; Fig. 1*e*). It is known that neuraminidase assembles into a box-shaped tetramer of dimensions 100 × 100 × 60 Å (Oakley *et al.*, 2010). Nevertheless, Biological Assembly 1 of PDB entry 3k38 (R_g = 43.9) consists of the box-shaped tetramer plus a satellite monomer (Fig. 6*a*). The R_g of this assembly is the largest of all the pentamers in the data set used (Fig. 1*e*). In

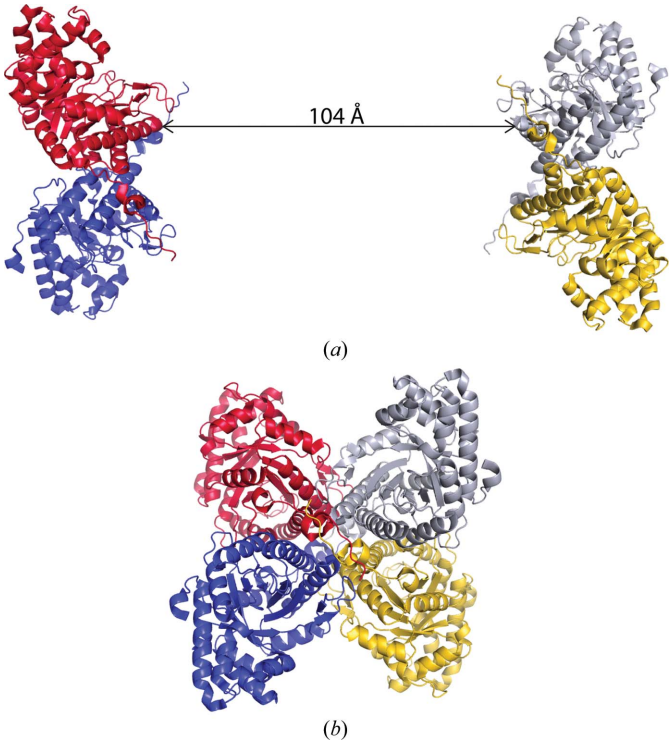


Figure 5
A misannotated aldolase (PDB entry 1f2j). (*a*) Biological Assembly 1 of PDB entry 1f2j. (*b*) The correct aldolase tetramer. In both panels, each protomer of the assembly is in a different color.

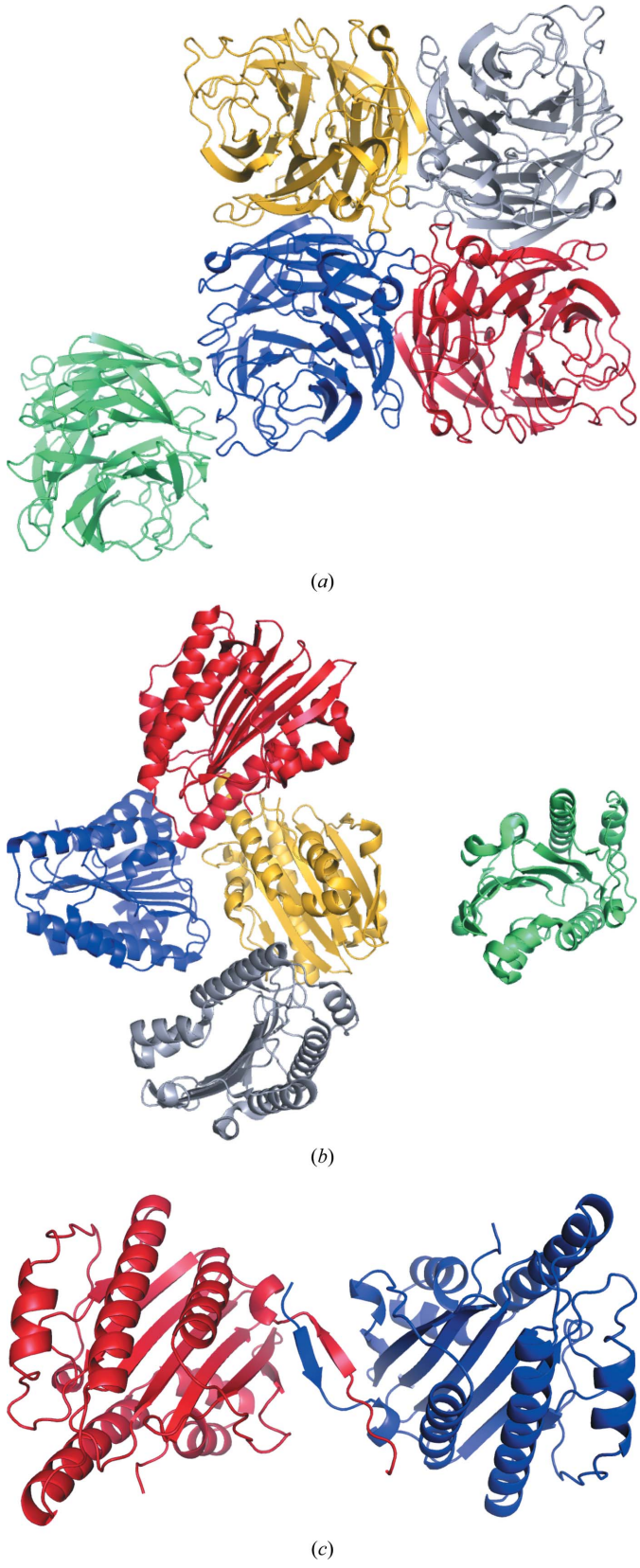


Figure 6
Misannotations revealed by the pentamer power law. (*a*) Biological Assembly 1 of a neuraminidase (PDB entry 3k38). (*b*) Biological Assembly 1 of PDB entry 2g18. (*c*) The dimer of PDB entry 2g18 predicted by *PDBePISA*. In all panels, each protomer of the assembly is in a different color.

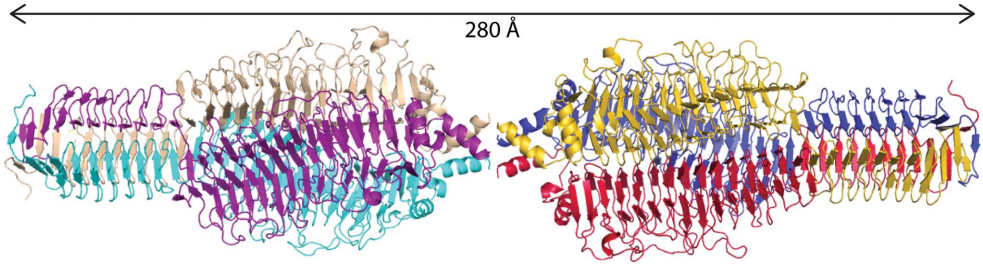


Figure 7
Biological Assembly 1 of PDB entry 3gqa, which is a dimer of trimers. The trimer is biologically relevant. Each protomer of the assembly is in a different color.

contrast, the R_g of the correct tetramer (34.2 Å) is within 2 Å of the value predicted from the tetramer power law (36 Å).

The pentamer plot also uncovers a bizarre annotation for *Nostoc* sp. 7120 phycocyanobilin:ferredoxin oxidoreductase (PDB entry 2g18; Fig. 1*e*). Biological Assembly 1 of PDB entry 2g18 (Tu *et al.*, 2007) is a cluster of five proteins with an R_g of 40.8 Å (Fig. 6*b*). An unusual aspect of this assembly is that one of the protomers does not contact the others. Analysis of the interfaces in the $P2_1$ lattice of PDB entry 2g18 with *PDBePISA* suggests a domain-swapped two-body assembly with an R_g of 28.4 Å that may be stable in solution (Fig. 6*c*). In this assembly, the N-termini of the protomers form an antiparallel β-sheet. Eight of the 12 molecules in the asymmetric unit form this assembly, while the N-terminus is disordered in the other four chains. Although additional studies are needed to determine the true oligomeric state of this protein in solution, it is unlikely that Biological Assembly 1 depicts the functional enzyme.

The hexamer power-law plot shows an extreme outlier with R_g = 78.1 Å (PDB entry 3gqa; Fig. 1*f*). This assembly derives from a structure of an N-terminal fragment of the tailed bacteriophage phi29 appendage protein gp12 (ΔNgp12N). The hexamer consists of two ΔNgp12N trimers packed in a linear arrangement so that the N-termini interact (Fig. 7). It should be noted that *PDBePISA* generated the hexameric assembly from the crystal lattice, whereas the depositors of PDB entry 3gqa provided strong evidence that the trimer is the correct biological assembly (Xiang *et al.*, 2009). The interface between the two trimers in the crystal consists of only four residues from each chain and buries only 118 Å^2 of surface area. In contrast, the trimer interface involves 123 residues and buries 4354 Å^2 of surface area. It is also noteworthy that the protein used for crystallization is an N-terminal deletion mutant, so it is conceivable that the hexamer, although not found in the virus, is present in solution *in vitro*. Additional studies would be needed to verify the existence of the hexamer in solution.

The octamer power-law plot has three main outliers, and the oligomeric state is incorrectly assigned in all three cases (Fig. 1*g*). Biological Assembly 1 of *S*-adenosylhomocysteine hydrolase from *Trypanosoma brucei* (PDB entry 3h9u; Structural Genomics Consortium, unpublished work) consists of a tetramer flanked by two dimers (R_g = 63.5 Å; Fig. 8*a*). This annotation is curious because this enzyme is known to be tetrameric (Turner *et al.*, 2000). The flanking dimers make few interactions with the tetramer, which casts doubt on the relevance of the eight-molecule assembly. Analysis of the *C*2 lattice with *PDBePISA* suggests that the tetramer is stable in solution. This tetramer is also formed in space group *I*222 by human *S*-adenosylhomocysteine hydrolase (PDB entry 4pfj), a protein that is suggested to be tetrameric in solution by size-exclusion chromatography (Wang *et al.*, 2014). The occurrence of an assembly in different crystal forms is strong evidence that it represents the oligomer in solution. Oddly, the conserved tetramer is not listed among the Biological Assemblies of PDB entry 3h9u.

The C1A mutant of *E. coli* glucosamine-6-phosphate synthase is similarly misannotated as an octamer (PDB entry 3ooj). This enzyme is subject to morpheein-type allosteric regulation involving an equilibrium between an active dimer and an inactive trimer-of-dimers hexamer (Mouilleron *et al.*, 2012). Biological Assembly 1 of PDB entry 3ooj (R_g = 61.5 Å) consists of the hexamer and an additional dimer (Fig. 8*b*). It seems unlikely that the dimer and hexamer interact as suggested in the octamer.

Finally, the octameric Biological Assembly 1 of *Salmonella typhimurium* NfnB dihydropteridine reductase (PDB entry 3hzn; Center for Structural Genomics of Infectious Diseases, unpublished work) is a linear arrangement of four dimers with an R_g of 55.3 Å (Fig. 8*c*). The dimer appears to be the correct oligomer, based on analysis with *PDBePISA* and comparison to homologous (88% identical) nitroreductases (*e.g.* PDB entries 1yki and 1nec; Race *et al.*, 2005; H.-J. Hecht, C. Bryant, H. Erdmann, H. Pelletier & R. Sawaya, unpublished work). The dimer is Biological Assembly 3 in the PDB.

4. Discussion

4.1. A simple, intuitive method for predicting oligomeric state from the SAXS R_g

A potential application of oligomer power laws is the prediction of oligomeric state from an experimental measurement of R_g. Simplicity is a strength of the method, since it requires only an estimate of R_g and the number of residues in a monomer of the protein. Also, unlike more sophisticated methods (Rambo & Tainer, 2011, 2013), extensive knowledge of SAXS theory is not required to understand the basis of the R_g power-law method. The method was found to be reasonably accurate. The prediction of oligomeric state was correct for five of the nine test proteins and was within one subunit for two other cases (Table 2). The largest error in n was two subunits.

The R_g power-law method can be compared with Rambo and Tainer's volume-of-correlation method, which is the gold

standard for determining molecular mass, and hence oligomeric state, from a SAXS curve collected on a relative intensity scale (Rambo & Tainer, 2013). The volume-of-correlation method correctly predicted the oligomeric state to within one subunit for all of the test cases used here (Table 2). The R_g power-law method performed as well or better than the volume-of-correlation method for six out of the nine cases, which is surprisingly good considering the simple basis of the R_g method. Although the R_g power-law formula (4) is not likely to replace the volume-of-correlation method, it nevertheless provides a simple and intuitive tool for confirming the oligomeric state from SAXS data.

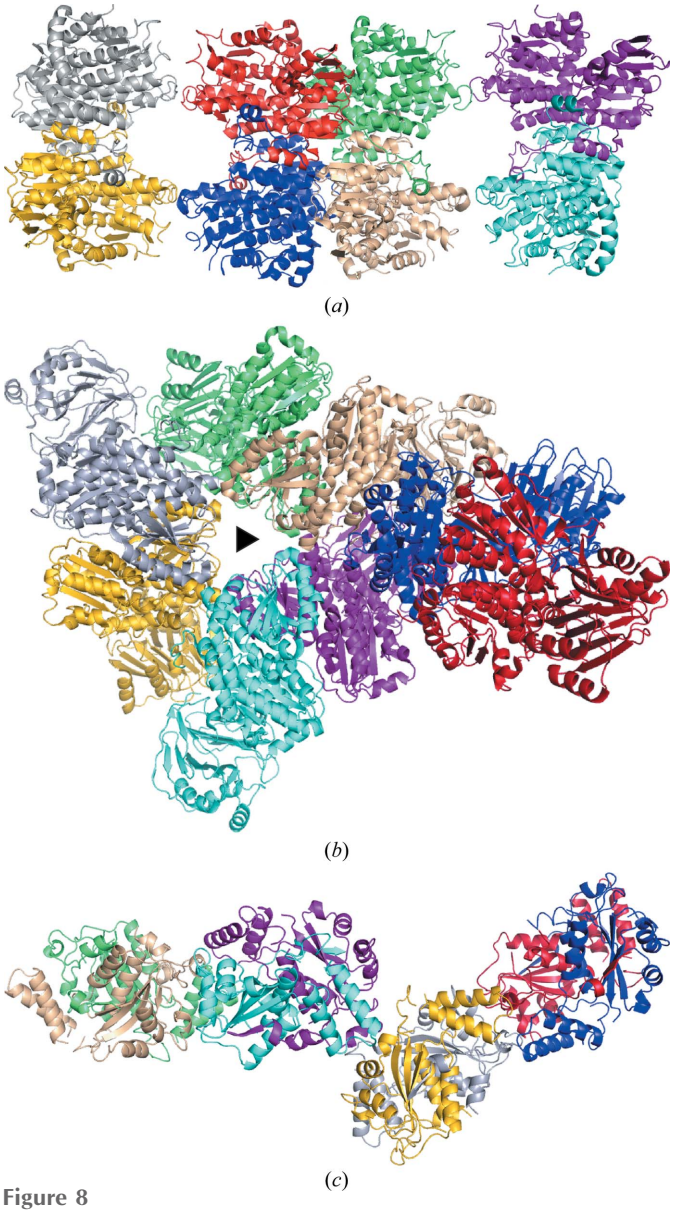


Figure 8
Misannotated octamers. (*a*) Biological Assembly 1 of an *S*-adenosylhomocysteine hydrolase (PDB entry 3h9u). (*b*) Biological Assembly 1 of glucosamine-6-phosphate synthase (PDB entry 3ooj). The assembly is oriented so that the threefold axis of the hexamer is perpendicular to the page. The hexamer consists of the chains colored green, wheat, purple, cyan, gold and silver. (*c*) Biological Assembly 1 of a dihydropteridine reductase (PDB entry 3hzn).

4.2. Using power laws to identify, correct and prevent misannotations in the PDB

The educational portal of the PDB, known as PDB-101 (Rose *et al.*, 2013), defines the Biological Assemblies provided for each entry as follows:

> The biological assembly (also sometimes referred to as the biological unit) is the macromolecular assembly that has either been shown to be or is believed to be the functional form of the molecule.

The analysis presented here shows that several of the assemblies offered by the PDB are wrong. This result is consistent with earlier studies reporting 14–15% error rates in the quaternary-structure assignment of protein homooligomers in the PDB (Ponstingl *et al.*, 2003; Levy, 2007). Correcting these errors is important not only for individual users of the PDB, but also for servers that use the PDB's biological assemblies, such as BioAssemblyModeler (Shapovalov *et al.*, 2014) and the Protein Biological Unit Database (Xu *et al.*, 2006).

More generally, there is increasing concern about incorrect annotations in databases. Lundin and coworkers compared their manually curated ribonucleotide reductase database with GenBank and found that only 23% of the ribonucleotide reductase entries in GenBank were annotated correctly with regard to class, role and function (Lundin *et al.*, 2009). Schnoes and coworkers performed a wider evaluation of misannotation levels for molecular function of six superfamilies in four public protein-sequence databases (Schnoes *et al.*, 2009). The misannotation levels averaged 5–63% across the superfamilies studied and reached as high as 80% for some cases. Green and Karp reported errors in the Enzyme Commission numbers assigned to entries in several databases (Green & Karp, 2005). Misannotations can mislead individual users, especially novices such as students, corrupt data-mining studies and negatively impact the training and validation of new bioinformatics algorithms. Thus, computational methods for identifying errors in databases are welcome.

The power laws described here may be useful for identifying and correcting misannotated oligomers. The use of power laws for this purpose has precedent. For example, the accessible surface area of monomeric proteins follows a power law of the molecular weight raised to the power 0.7–0.8 (Miller *et al.*, 1987; Marsh, 2013). Marsh showed that deviation from this behavior could be used to identify proteins that were erroneously classified as monomers rather than oligomers (Marsh, 2013). Similarly, it seems possible to identify misannotations of oligomeric state or quaternary structure by analyzing the biological assemblies in the PDB for deviation from R_g power-law behavior. The outlying assemblies could then be analyzed with more robust computational methods, such as *PDBePISA* (Krissinel, 2015; Krissinel & Henrick, 2005, 2007).

This strategy could be implemented as part of the PDB deposition procedure. For example, the deposition validation report could include the relevant oligomer scatter plot showing the current PDB holdings and the fitted power-law curve, along with symbols indicating the depositor's assembly and those calculated from *PDBePISA*. An example of such a

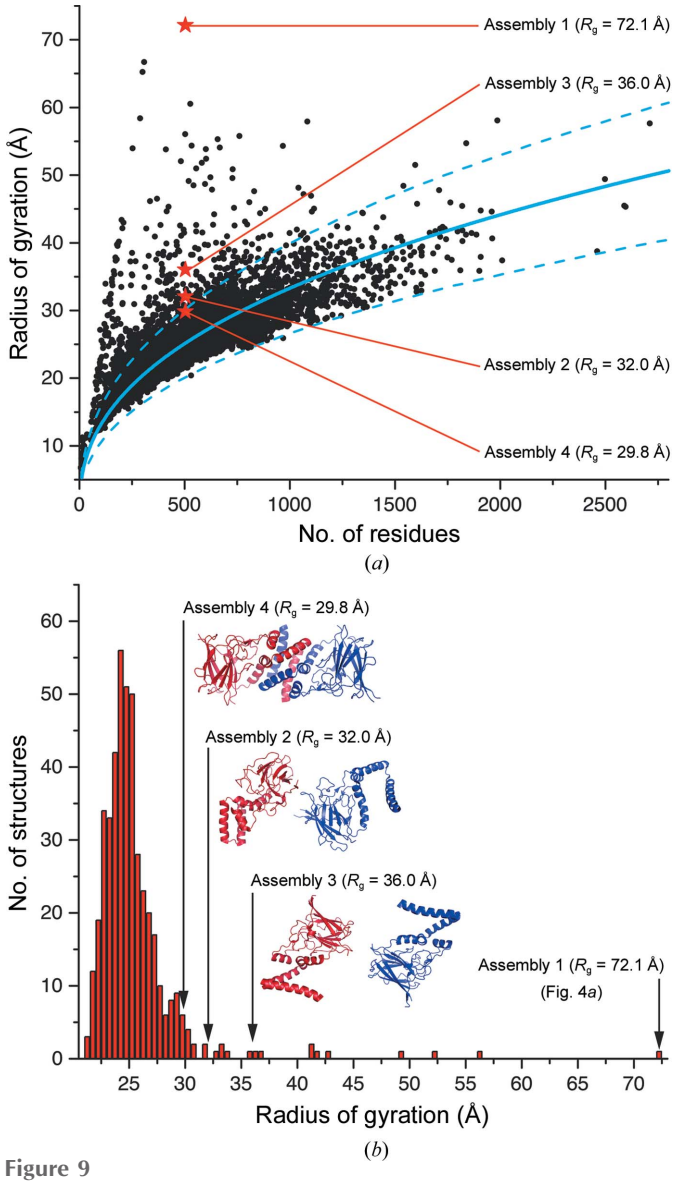


Figure 9
Validation of the assemblies of PDB entry 2boy using R_g data. (*a*) The power-law plot for dimers, with the four assemblies currently assigned for PDB entry 2boy indicated by red stars. The solid cyan curve is the fitted power law. The dashed cyan curves represent ±20% deviation from the fitted power-law curve. (*b*) A histogram showing the frequency of R_g for dimers with 480–520 residues. The R_g values and structures of the four assemblies of PDB entry 2boy are indicated.

validation figure is shown for PDB entry 2boy (Fig. 9*a*). Alternatively, this information could be conveyed by a histogram showing the frequency of R_g (Fig. 9*b*). In this example, note that only one of the four assemblies has a significant dimer interface and occupies the highly populated region of R_g space (assembly 4); this assembly is the one suggested by *PDBePISA*. It should also be possible to provide this information with a percentile slider bar, as is currently performed for R_{free} and other global validation metrics. Such figures would provide a visual representation of how the depositor's assembly compares with others in the PDB. Consideration of these ideas would help improve the premier database of structural biology.

Acknowledgements

The research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award Nos. R01GM065546, R01GM061068 and R01GM093123.

References

Bartual, S. G., Otero, J. M., Garcia-Doval, C., Llamas-Saiz, A. L., Kahn, R., Fox, G. C. & van Raaij, M. J. (2010). *Proc. Natl Acad. Sci. USA*, **107**, 20287–20292.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bublitz, M., Morth, J. P. & Nissen, P. (2011). *J. Cell Sci.* **124**, 2515–2519.
Chudzik, D. M., Michels, P. A., de Walque, S. & Hol, W. G. J. (2000). *J. Mol. Biol.* **300**, 697–707.
Cornish-Bowden, A. (2014). *FEBS J.* **281**, 621–632.
Derewenda, U., Tarricone, C., Choi, W. C., Cooper, D. R., Lukasik, S., Perrina, F., Tripathy, A., Kim, M. H., Cafiso, D. S., Musacchio, A. & Derewenda, Z. S. (2007). *Structure*, **15**, 1467–1481.
Dhatwalia, R., Singh, H., Oppenheimer, M., Karr, D. B., Nix, J. C., Sobrado, P. & Tanner, J. J. (2012). *J. Biol. Chem.* **287**, 9041–9051.
Dima, R. I. & Thirumalai, D. (2004). *J. Phys. Chem. B*, **108**, 6564–6570.
Doi, M. (1996). *Introduction to Polymer Physics.* Oxford: Clarendon Press.
Ferraroni, M., Kolomytseva, M. P., Solyanikova, I. P., Scozzafava, A., Golovleva, L. A. & Briganti, F. (2006). *J. Mol. Biol.* **360**, 788–799.
Flory, P. J. (1949). *J. Chem. Phys.* **17**, 303–310.
Gennes, P. G. de (1979). *Scaling Concepts in Polymer Physics.* Ithaca: Cornell University Press.
Gimona, M. (2008). *Adv. Exp. Med. Biol.* **644**, 73–84.
Gong, H., Fleming, P. J. & Rose, G. D. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 16227–16232.
Goodsell, D. S. & Olson, A. J. (2000). *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153.
Green, M. L. & Karp, P. D. (2005). *Nucleic Acids Res.* **33**, 4035–4039.
Gruszka, D. T., Whelan, F., Farrance, O. E., Fung, H. K. H., Paci, E., Jeffries, C. M., Svergun, D. I., Baldock, C., Baumann, C. G., Brockwell, D. J., Potts, J. R. & Clarke, J. (2015). *Nature Commun.* **6**, 7271.
Henrick, K. & Thornton, J. M. (1998). *Trends Biochem. Sci.* **23**, 358–361.
Henzl, M. T., Sirianni, A. G., Wycoff, W. G., Tan, A. & Tanner, J. J. (2013). *Proteins*, **81**, 300–315.
Hinsen, K., Hu, S., Kneller, G. R. & Niemi, A. J. (2013). *J. Chem. Phys.* **139**, 124115.
Hofmann, H., Soranno, A., Borgia, A., Gast, K., Nettels, D. & Schuler, B. (2012). *Proc. Natl Acad. Sci. USA*, **109**, 16155–16160.
Hong, L. & Lei, J. (2009). *J. Polym. Sci. B Polym. Phys.* **47**, 207–214.
Ivankov, D. N., Bogatyreva, N. S., Lobanov, M. Y. & Galzitskaya, O. V. (2009). *PLoS One*, **4**, e6476.
Izoré, T., Contreras-Martel, C., El Mortaji, L., Manzano, C., Terrasse, R., Vernet, T., Di Guilmi, A. M. & Dessen, A. (2010). *Structure*, **18**, 106–115.
Jaffe, E. K. (2005). *Trends Biochem. Sci.* **30**, 490–497.
Janin, J., Miller, S. & Chothia, C. (1988). *J. Mol. Biol.* **204**, 155–164.
Jones, S. & Thornton, J. M. (1995). *Prog. Biophys. Mol. Biol.* **63**, 31–65.
Jones, S. & Thornton, J. M. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
Keskin, O., Gursoy, A., Ma, B. & Nussinov, R. (2008). *Chem. Rev.* **108**, 1225–1244.
Kleywegt, G. J., Zou, J.-Y., Kjeldgaard, M. & Jones, T. A. (2001). *International Tables for Crystallography*, Vol. *F*, edited by M. G.

Rossmann & E. Arnold, pp. 353–356. Dordrecht: Kluwer Academic Publishers.
Kohn, J. E., Millett, I. S., Jacob, J., Zagrovic, B., Dillon, T. M., Cingel, N., Dothager, R. S., Seifert, S., Thiyagarajan, P., Sosnick, T. R., Hasan, M. Z., Pande, V. S., Ruczinski, I., Doniach, S. & Plaxco, K. W. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 12491–12496.
Koiwai, K., Hartmann, M. D., Linke, D., Lupas, A. N. & Hori, K. (2016). *J. Biol. Chem.* **291**, 3705–3724.
Kolinski, A., Godzik, A. & Skolnick, J. (1993). *J. Chem. Phys.* **98**, 7420–7433.
Krissinel, E. (2015). *Nucleic Acids Res.* **43**, W314–W319.
Krissinel, E. & Henrick, K. (2005). *Computational Life Sciences*, edited by M. R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher & I. Fischer, pp. 163–174. Berlin, Heidelberg: Springer-Verlag.
Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.
Levy, E. D. (2007). *Structure*, **15**, 1364–1367.
Levy, E. D. & Teichmann, S. (2013). *Prog. Mol. Biol. Transl. Sci.* **117**, 25–51.
Lundin, D., Torrents, E., Poole, A. M. & Sjöberg, B. M. (2009). *BMC Genomics*, **10**, 589.
Luo, M., Christgen, S., Sanyal, N., Arentson, B. W., Becker, D. F. & Tanner, J. J. (2014). *Biochemistry*, **53**, 5661–5673.
Luo, M., Singh, R. K. & Tanner, J. J. (2013). *J. Mol. Biol.* **425**, 3106–3120.
Luo, M. & Tanner, J. J. (2015). *Biochemistry*, **54**, 5513–5522.
MacKinnon, R. (2003). *FEBS Lett.* **555**, 62–65.
Marsh, J. A. (2013). *J. Mol. Biol.* **425**, 3250–3263.
Marsh, J. A. & Teichmann, S. A. (2015). *Ann. Rev. Biochem.* **84**, 551–575.
Miller, S., Lesk, A. M., Janin, J. & Chothia, C. (1987). *Nature (London)*, **328**, 834–836.
Mouilleron, S., Badet-Denisot, M. A., Pecqueur, L., Madiona, K., Assrir, N., Badet, B. & Golinelli-Pimpaneau, B. (2012). *J. Biol. Chem.* **287**, 34533–34546.
Oakley, A. J., Barrett, S., Peat, T. S., Newman, J., Streltsov, V. A., Waddington, L., Saito, T., Tashiro, M. & McKimm-Breschkin, J. L. (2010). *J. Med. Chem.* **53**, 6421–6431.
Obara, K., Miyashita, N., Xu, C., Toyoshima, I., Sugita, Y., Inesi, G. & Toyoshima, C. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 14489–14496.
Perica, T., Marsh, J. A., Sousa, F. L., Natan, E., Colwell, L. J., Ahnert, S. E. & Teichmann, S. A. (2012). *Biochem. Soc. Trans.* **40**, 475–491.
Ponstingl, H., Henrick, K. & Thornton, J. M. (2000). *Proteins*, **41**, 47–57.
Ponstingl, H., Kabir, T. & Thornton, J. M. (2003). *J. Appl. Cryst.* **36**, 1116–1122.
Race, P. R., Lovering, A. L., Green, R. M., Ossor, A., White, S. A., Searle, P. F., Wrighton, C. J. & Hyde, E. I. (2005). *J. Biol. Chem.* **280**, 13256–13264.
Rambo, R. P. (2015). *SCÅTTER*. http://www.bioisis.net/tutorial/9.
Rambo, R. P. & Tainer, J. A. (2011). *Biopolymers*, **95**, 559–571.
Rambo, R. P. & Tainer, J. A. (2013). *Nature (London)*, **496**, 477–481.
Rose, P. W. *et al.* (2013). *Nucleic Acids Res.* **41**, D475–D482.
Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. (2009). *PLoS Comput. Biol.* **5**, e1000605.
Schreiter, E. R. & Drennan, C. L. (2007). *Nature Rev. Microbiol.* **5**, 710–720.
Shapovalov, M. V., Wang, Q., Xu, Q., Andrake, M. & Dunbrack, R. L. Jr (2014). *PLoS One*, **9**, e98309.
Singh, H., Arentson, B. W., Becker, D. F. & Tanner, J. J. (2014). *Proc. Nat. Acad. Sci. USA*, **111**, 3389–3394.
Singh, R. K., Larson, J. D., Zhu, W., Rambo, R. P., Hura, G. L., Becker, D. F. & Tanner, J. J. (2011). *J. Biol. Chem.* **286**, 43144–43153.
Skolnick, J., Kolinski, A. & Ortiz, A. R. (1997). *J. Mol. Biol.* **265**, 217–241.
Srivastava, D., Schuermann, J. P., White, T. A., Krishnan, N., Sanyal, N., Hura, G. L., Tan, A., Henzl, M. T., Becker, D. F. & Tanner, J. J. (2010). *Proc. Natl Acad. Sci. USA*, **107**, 2878–2883.
Thirumalai, D. (2000). *Theor. Chem. Acc.* **103**, 292–293.
Toyoshima, C., Nomura, H. & Tsuda, T. (2004). *Nature (London)*, **432**, 361–368.
Tu, D., Li, Y., Song, H. K., Toms, A. V., Gould, C. J., Ficarro, S. B., Marto, J. A., Goode, B. L. & Eck, M. J. (2011). *PLoS One*, **6**, e18080.
Tu, S.-L., Rockwell, N. C., Lagarias, J. C. & Fisher, A. J. (2007). *Biochemistry*, **46**, 1484–1494.
Turner, M. A., Yang, X., Yin, D., Kuczera, K., Borchardt, R. T. & Howell, P. L. (2000). *Cell Biochem. Biophys.* **33**, 101–125.
Vance, T. D., Olijve, L. L., Campbell, R. L., Voets, I. K., Davies, P. L. & Guo, S. (2014). *Biosci. Rep.* **34**, e00121.
Veenhoff, L. M., Heuberger, E. H. & Poolman, B. (2002). *Trends Biochem. Sci.* **27**, 242–249.
Wang, Y., Kavran, J. M., Chen, Z., Karukurichi, K. R., Leahy, D. J. & Cole, P. A. (2014). *J. Biol. Chem.* **289**, 31361–31372.
Wilkins, D. K., Grimshaw, S. B., Receveur, V., Dobson, C. M., Jones, J. A. & Smith, L. J. (1999). *Biochemistry*, **38**, 16424–16431.
Xiang, Y., Leiman, P. G., Li, L., Grimes, S., Anderson, D. L. & Rossmann, M. G. (2009). *Mol. Cell*, **34**, 375–386.
Xu, K., Tzvetkova-Robev, D., Xu, Y., Goldgur, Y., Chan, Y.-P., Himanen, J. P. & Nikolov, D. B. (2013). *Proc. Natl Acad. Sci. USA*, **110**, 14634–14639.
Xu, Q., Canutescu, A., Obradovic, Z. & Dunbrack, R. L. Jr (2006). *Bioinformatics*, **22**, 2876–2882.