Journal of Structural Biology 196 (2016) 299–308

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Structure based sequence analysis of viral and cellular protein assemblies

Daniel J. Montiel-García [a], Ranjan V. Mannige [b,1], Vijay S. Reddy [b], Mauricio Carrillo-Tripp [a,2,*]

[a] Biomolecular Diversity Laboratory, Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional, Mexico
[b] Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA

ARTICLE INFO

Article history:
Received 27 June 2016
Accepted 18 July 2016
Available online 29 July 2016

Keywords:
Viral capsid proteins
Sequence conservation
Protein-protein interactions
Subunit interface
Bioinformatics
Protein complexes

ABSTRACT

It is well accepted that, in general, protein structural similarity is strongly related to the amino acid sequence identity. To analyze in great detail the correlation, distribution and variation levels of conserved residues in the protein structure, we analyzed all available high-resolution structural data of 5245 cellular complex-forming proteins and 293 spherical virus capsid proteins (VCPs). We categorized and compare them in terms of protein structural regions. In all cases, the buried core residues are the most conserved, followed by the residues at the protein-protein interfaces. The solvent-exposed surface shows greater sequence variations. Our results provide evidence that cellular monomers and VCPs could be two extremes in the quaternary structural space, with cellular dimers and oligomers in between. Moreover, based on statistical analysis, we detected a distinct group of icosahedral virus families whose capsid proteins seem to evolve much slower than the rest of the protein complexes analyzed in this work.

1. Introduction

It has been half a century since early efforts started to describe protein-protein complexes of the cell (Chotia, 1974; Richards, 1974; Chotia and Janin, 1975). Such works showed that the shape complementarity of cellular protein interfaces could be used to characterize the interactions regarding the size of buried surface, paucity of buried water molecules and packing density of interface atoms. The modern view of cellular protein interfaces has produced a rule base of characteristics that include size, residue composition, hydrophobicity, and planarity (Lawrence and Colman, 1993; Jones and Thornton, 1995).

The exponential increase of available structural information has enabled us to revisit and improve on past results. For example, studies of non-redundant datasets of a couple of thousand protein structures have shown that cellular protein interfaces differ in amino acid composition, residue–residue preferences between interactions, and secondary structure, from those of surface and core residues (Ofran and Rost, 2003; Yan et al., 2008). In general, studies use datasets containing complexes from different species. However, focus on a single species was approached by analyzing all the available data on structural complexes from the yeast *Saccharomyces cerevisiae* (Talavera et al., 2011). It was found that, as previously seen, there is a significant contribution of main-chain atoms to protein-protein contacts and the type of interaction seems to depend on both amino acid side chain and secondary structure type involved at the contact. Cellular homo and hetero-complexes showed no clear distinction. Interestingly, there seem to be no significant differences between the interface regions and the rest of the surface from a thermodynamic standpoint regarding the solvation energy.

Just like the cellular proteins that have a structural function, the virus capsid proteins (VCPs) also present intriguing features. In the case of icosahedral viruses, at least 60 copies of a type of VCP must self-assemble into symmetric closed protein complexes in the form of spherical shells (capsids) that encapsulate the viral genome (Cann, 2005). The capsids display a defined size and structural architecture depending on the type of virus (Caspar and Klug, 1962). A detailed description of the molecular specificity, recognition and self-assembly properties of the VCPs remain elusive. Such molecular mechanisms still need to be well understood, in comparison to cellular proteins (Janin et al., 2008). Recently, the geometric and physical-chemical properties of a set of 49 icosahedral virus capsids were analyzed and compared with the interfaces of cellular protein–protein complexes. It seems that small capsid

* Corresponding author.
E-mail addresses: mauricio.carrillo@cinvestav.mx, mauricio.carrillo@cimat.mx (M. Carrillo-Tripp).
[1] Present address: Theory of Nanostructured Materials facility, Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.
[2] Present address: Ciencias de la Computación, Centro de Investigación en Matemáticas, A.C., Guanajuato 36000, Mexico.

http://dx.doi.org/10.1016/j.jsb.2016.07.013
1047-8477/

interfaces are loosely packed, like crystal contacts, whereas the larger interfaces are close-packed, as in cellular homodimers (Bahadur et al., 2007). Also, a statistical analysis of a set of 319 icosahedral viruses showed that VCPs exhibit an apparent segregation in structural fold space (Cheng and Brooks, 2013). It was suggested that the unique folds of VCPs present a favorable geometry to allow adequate packing and assembly into the right architecture. Such structural folds might be under particular constraints during evolution by the requirement of the assembled cage-like structure, as opposed to their surface chemistry. Furthermore, some structural characteristics seem to be also unique to non-capsid viral proteins. When compared to their cellular counterparts, they show lower contact densities, higher occurrence of random coil segments, shorter disordered regions, and less destabilizing effects when mutations happen (Tokuriki et al., 2009).

Even though proteins can diverge beyond the point where there is no detectable sequence similarity, the packing of the tertiary structure can maintain similar folds. Efforts have been made to understand the underlying principles of structural conservation during protein sequence evolution. Earlier works have analyzed the relationship between the divergence of sequence and the three-dimensional structure of cellular proteins (Chotia and Lesk, 1986), and the relation between the sequence identity and structure similarity to the alignment length (Sander and Schneider, 1991; Rost, 1999). An alternative approach came in the form of the classification of the protein fold space. One example is SCOP, an expert-based hierarchical classification of protein structures (Murzin et al., 1995). SCOP groups together those domains that have structural, functional, and sequence evidence for a common evolutionary ancestor at the superfamily (SF) level. In particular, out of the 560 SCOP v1.73 protein domains found in viruses, >10% do not have any structural or evolutionary relatives in modern cellular organisms at the SF level (Abroi and Gough, 2011).

In general, the variations on the level of residue conservation at different locations in the protein structure is still a controversial subject, being inconclusive or even contradictory. For instance, Grishin and Phillips, 1994 concluded that interface and core residues are not well conserved and evolve nearly as rapidly as the overall protein sequence, after analyzing 135 sequences and 16 structures of five cellular oligomeric enzyme families. Valdar and Thornton, 2001 concluded that interface residues are more conserved than expected for a random distribution, after analyzing 195 sequences representative of six cellular homodimer families. Caffrey et al., 2004 found that the interface is rarely more conserved than the surface, after analyzing 64 homologous cellular dimers. Guharoy and Chakrabarti, 2005 concluded that the average conservation at the central region of the protein-protein interface is higher than its surroundings, after analyzing 122 cellular homodimers. In the case of VCPs, Bahadur and Janin, 2008 concluded that the core and interface residues are better conserved than the chain average, after analyzing 32 icosahedral viruses. Subsequently, Chih-Min et al., 2015 concluded that some global patterns derived from the capsid structure, like the residue packing density, are consistent with those present in VCP sequence conservation profiles. Overall, it is a plausible idea that the fact that all previous analyses have been performed on families of homologous proteins using small data sets could bias the results. The particular role of the VCP in a structural and evolutionary context needs further investigation, and an extensive comparison to cellular proteins is in order.

In this work, we further investigate and highlight the differences between cellular and icosahedral capsid proteins. We address three particular questions. First, what is the correlation between the conservation of sequence and the similarity in tertiary and quaternary structures? Second, how are the conserved residues distributed in the protein structure? And third, what is the variation on the level of residue conservation at different locations in the protein structure? We analyzed all the available high-resolution structural data and compared cellular protein *n*-mers with icosahedral VCPs.

2. Materials and methods

2.1. Datasets

In this work, we analyzed all the data available on the three-dimensional structure and sequence of cellular and icosahedral capsid proteins, grouped in four independent datasets. In the case of cellular *n*-mer complexes, we included monomers ($n = 1$. Data not shown), dimers ($n = 2$. Table S1), and higher order oligomers ($n = 3, 4, 5, 6, 8, 10, 12, 22$, and 24. Table S2). In the case of VCPs, we included icosahedral viruses belonging to 36 different genera from 21 different families, according to the classification proposed by the International Committee on Taxonomy of Viruses (ICTV, Fauquet et al., 2005). This dataset spans a broad range of icosahedral triangulation numbers (T = 1, 2, 3, 4, 7d, 7 L, pT3. Table S3).

In all cases, the basic criteria used to choose structures was to have available data determined by X-ray crystallography (resolution $\leqslant 4$ Å), consistent polypeptide chain sequence (no missing loops or fragment miss-annotations), and long chains (>65 residues). Atomic coordinates were obtained from the Protein Data Bank (Berman et al., 2000), in the case of cellular proteins, and from the Virus Particle Explorer database (Carrillo-Tripp et al., 2009), in the case of VCPs. Following (Valdar and Thornton, 2001), the term *protomer* denotes a unique polypeptide chain of a multimeric complex. Hence, homomers will be represented by one protomer, whereas heteromers will be represented by two or more protomers. Hence, our datasets consisted of 5087 cellular monomers, 51 cellular dimers (represented by 57 protomers), 65 cellular oligomers (represented by 101 protomers), and 212 VCPs (represented by 293 protomers).

2.2. Structure similarity

The root mean square deviation (RMSD) has been the standard way to measure structural similarity. However, other metrics provide a better quantification, like the TM-score (Zhang and Skolnick, 2004). The TM-score presents several advantages over the RMSD.

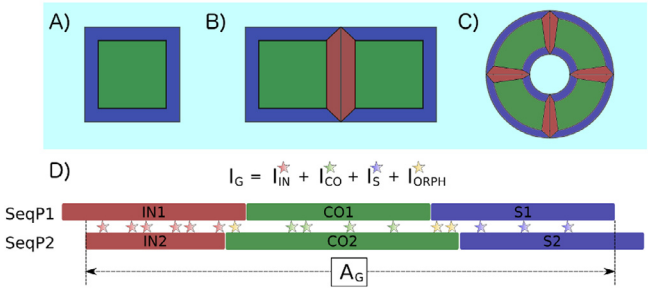

Fig. 1. Simplified diagrams depicting the three different locations of amino acids in the protein tertiary and quaternary structure. Amino acids are located at the protein's solvent accessible surface (S, in blue), at the protein core (CO, in green), or at the protein-protein interface (IN, in red). Schematics shown for a monomer (A), a dimer (B), and an oligomer or virus capsid proteins (C), immersed in a solvent (light blue). In order to study the distribution and level of conservation of amino acids in these locations, a sequence alignment is derived from a 3D structural alignment for a pair of proteins (P1 vs. P2). Then, conserved residues (identical amino acids, indicated by stars) are identified and labeled according to their location in the protein structure (D). In the case where no location correspondence is found between the two proteins for a conserved residue, those are categorized as orphans (ORPH, in yellow). *A* represents the number of aligned residues, and *I* is the number of identical residues.

TM-score values are bound to the interval (0, 1], with 1 being two identical structures (equivalent to an RMSD value of 0). The TM-score is independent of protein size, and it weighs a close match stronger than a distant one. Based on Bayesian theory, it was shown that a TM-score value >0.5 indicates that the two structures have the same fold/topology (Xu and Zhang, 2010). The TM-align tool identifies the best structural superposition between a pair of proteins and calculates the TM-score (Zhang and Skolnick, 2005). Based on the optimal structural superposition, the amino acid sequence alignment between the two proteins was derived.

2.3. Sequence identity

The sequence identity (S_G) considers the fraction of identical residues (I_G) from the total number of aligned residues (A_G) in a sequence alignment, i.e., $S_G = I_G/A_G$. To distinguish the location of the conserved residues in the protein structure, we use three different categories: protein-protein interface (IN), protein core (CO), and solvent accessible surface (S), depicted in Fig. 1. We define a sequence identity index per location category, $S_k^* = I_k/I_G$ (k = IN, CO or S), to quantify the relative percentage of amino acid conservation found in the different regions of the protein. Based on the protein structure superposition, regions specific to each location category were identified in the amino acid sequence alignment. Hence, I_k is the number of conserved residues found in each category region. This procedure excludes certain conserved residues that do not match a common structural location in both sequences based on the above classification, as can be seen from Fig. 1D. We define such residues as orphans (ORPH) in the context of this work.

2.4. Structural categories

The structural classification takes into account the tertiary and quaternary complex structure, i.e., whether the residues are at the protein-protein interface, the core, or on the solvent accessible surface (Fig. 1). We used the same criteria for cellular and capsid proteins. Interface residues are those having at least one close contact with a neighboring protein. The residue type specific cut-offs method is a proper strategy to identify contacting residues between two closely interacting molecules, AB. In the case of protein-protein interactions, the definition of contacts we used (Damodaran et al., 2002) provides a true description of the presence/absence of inter-residue interactions at the AB interface. In the cut-offs method, one calculates the distance between every residue of protein A versus every residue in protein B. Those residue pairs $R_i^A - R_j^B$ being at a shorter distance than the corresponding residue type specific cut-off value are identified as interface residues in the protein complex AB. This approach works well when the atomic positions of both A and B are known. Because the solvent molecules are missing in the data, the lack of atomic information prevents the use of this method to distinguish core from surface residues. The next best approach is to consider the solvent accessible surface area (SASA). Core residues are buried in the protein, whereas solvent accessible residues can interact directly with the environment. It makes sense to use the level of exposure to the solvent as the criteria to group the non-interface residues into core or surface. The SASA method is useful to distinguish which residues have enough area accessible to the solvent to be considered to be at the surface of the protein. A comparison we made between the two methods showed that the SASA approach underestimates the number of interface residues by 10%, on average, when compared to the distance based approach. We performed an extensive examination looking into the distribution of SASA values per residue in proteins, independently done for each dataset (Fig. S1). We found a peak at the [0, 5] interval of the relative accessible surface area (%SASA) in both cellular and capsid proteins. We assume residues in this range are the protein's core residues. Correspondingly, residues with %SASA > 5% are the surface residues. SASA values were computed using the PDBASA library (Shrake and Rupley, 1973).

2.5. Sequence conservation

The sequence conservation is related to the residue variability at each sequence position i of a polypeptide chain, measured by the Shannon entropy,

$$S(i) = -\Sigma_k p_k \ln p_k$$

where $p_k = n_k/N$ is the frequency of residue type k, and n_k is the fraction of sequences having the residue type k at position i on a

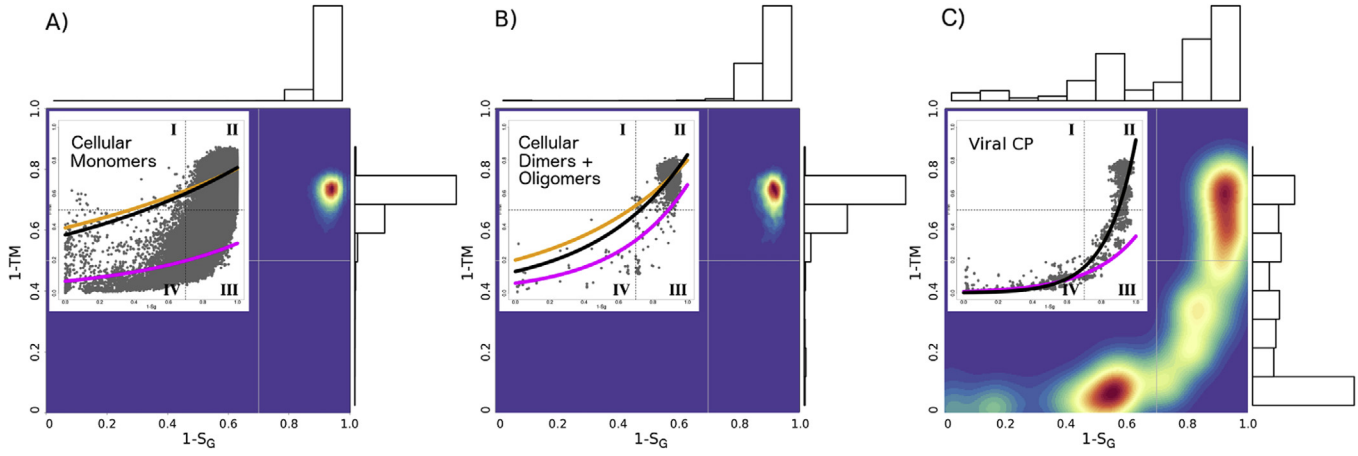

Fig. 2. Correlation between the sequence identity and the structure similarity in proteins. Equivalently, shown is the fraction of mutated residues ($1 - S_G$) and the structure divergence ($1 - TM$). All-vs-all protein pairs analysis of cellular monomers (A), cellular dimers plus oligomers (B), and icosahedral virus capsid proteins (C). Two-dimensional axis-aligned bivariate normal kernel density estimation, evaluated on a square grid of 300 points in each direction. Low density regions are shown in purple, whereas high density regions are red. Probability densities of ($1 - S_G$) and (1-TM) are shown in horizontal and vertical histograms, respectively. Insets: each point in the cloud represents a unique protein pair (gray). Fits to the exponential model are shown for the set of pairs with ($1 - S_G$) > 0 (black), ($1 - S_G$) < 0.7 (magenta), and ($1 - S_G$) > 0.7 (orange). Thresholds are defined as: ($1 - TM$) < 0.5 are pairs with the same protein fold, and ($1 - S_G$) < 0.7 are pairs of homologous proteins. These thresholds produce four sectors of the Cartesian plane, indicated by the roman numerals I, II, III, and IV.

multiple sequence alignment (MSA) of N sequences. $S(i)$ varies between 0 at positions fully conserved, and approximately 3 at positions where all residue types are equally found in the MSA. A normalized entropy can be defined as $s(i) = S(i)/\langle S \rangle$, where $\langle S \rangle$ is the average value of $S(i)$ over all the residues of the polypeptide chain. Values of $S(i)$ for each protomer in our datasets were extracted from the HSSP database (Touw et al., 2015). All statistical tests and plots were carried out using the R package and libraries therein.

3. Results

3.1. Relation between sequence identity and structure similarity

Chotia and Lesk, 1986 analyzed 32 pairs of homologous cellular protein structures. They found that the extent of the structural changes was directly related to the extent of the sequence changes. Following the same strategy, we performed a pair-wise comparative analysis of the structure and sequence of all protomers in each of our datasets. Following their work, we plot the structure divergence (1 − TM-score) as a function of the fraction of mutated residues ($1 - S_G$), where S_G is the global sequence identity, shown in Fig. 2 (insets). Thresholds on the TM-score and S_G produce four sectors in the Cartesian plane. Sector I contains pairs of homologous proteins having a different tertiary fold. Sector II contains pairs of non-homologous proteins with a different tertiary fold. Sector III contains pairs of non-homologous proteins having the same tertiary fold. Sector IV contains pairs of homologous proteins having the same tertiary fold. The total number of protomer pairs analyzed in this work are listed in Table 1. At first glance, it appears that the cloud of points behaves the same in all cases. However, a density analysis reveals differences among all datasets, although more contrasting between cellular proteins and VCPs. Most pairs lie in sector II in the former case (Sander and Schneider, 1991; Rost, 1999), whereas pairs are similarly distributed between sectors II, III, and IV in the later case. Furthermore, ∼30% of the characterized VCPs pairs have TM-score values >0.9, notwithstanding the sequence identity value [20–99%].

Even though the percentage of pairs in sector III is low in the case of cellular proteins (<1%), there are several examples where

Table 3
Nonlinear weighted least-squares estimates for the parameters of the exponential model (1-TM-score) ∼ f*exp(k*(1-S_G)) using the Gauss-Newton algorithm. Fits to each dataset for scenarios (1-S_G) > 0 (black), (1-S_G) < 0.7 (magenta), and (1-S_G) > 0.7 (orange) are shown in Fig. 2.

	(1-S_G) > 0		(1-S_G) < 0.7		(1-S_G) > 0.7	
	f	k	f	k	f	k
Monomer	0.351	0.768	0.072	1.429	0.393	0.647
Dimers + Oligomers	0.130	1.854	0.060	2.392	0.990	1.390
VCP	0.003	5.912	0.009	3.612	0.003	5.900

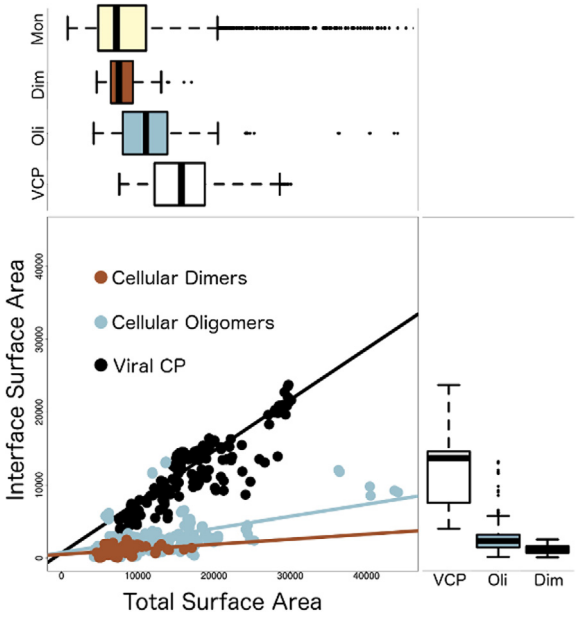

Fig. 3. Correlation between interface surface area and total surface area in proteins, independently estimated for cellular monomers (Mon), cellular dimers (Dim), cellular oligomers (Oli), and icosahedral virus capsid proteins (VCP). In the case of cellular monomers, only the total surface area is analyzed. Linear regression is shown for each set. Statistics summaries (median, first - third quartiles, minimum - maximum values, and outliers) are depicted with boxplots for the interface (right) and the total surface area (top). Area values are in units of Å^2.

Table 1
Count of total protein pairs in each dataset, dissected into sectors (as defined in Fig. 2). Percentage of pairs in each sector in parentheses. Homologous proteins (pairs with a sequence identity S_G > 0.3) are found in sectors I and IV.

	Total pairs	Sector I	Sector II	Sector III	Sector IV
Monomer	12,936,241	103 (0.001%)	12,822,459 (99.10%)	105,459 (0.82%)	8220 (0.064%)
MonomerRND	38,809	0 (0.0%)	38,480 (99.15%)	304 (0.78%)	25 (0.064%)
Dimers + Oligomers	1596 + 5050 = 8290	8 (0.097%)	8037 (97.00%)	133 (1.60%)	112 (1.35%)
VCP	42,778	0 (0.0%)	14,454 (33.79%)	12,348 (28.87%)	15,976 (37.35%)

Table 2
Non-parametric two-sample Kolmogorov-Smirnov test D value (p-value), independently estimated for the total sequence identity (S_G, above the diagonal) and for the structure similarity (*TM-score*, below the diagonal) distributions of each analyzed dataset. A p-value < 0.05 means that the compared datasets do not come from the same distribution.

TM-score/S_G	Monomer	MonomerRND	Dimers + Oligomers	VCP
Monomer	–	0.002 (0.816)*	0.377 (<2.2e−16)	0.569 (<2.2e−16)
MonomerRND	0.002 (0.601)*	–	0.378 (<2.2e−16)	0.569 (<2.2e−16)
Dimers + Oligomers	0.102 (<2.2e−16)	0.102 (<2.2e−16)	–	0.526 (<2.2e−16)
VCP	0.756 (<2.2e−16)	0.756 (<2.2e−16)	0.692 (<2.2e−16)	–

* p-value ⩾ 0.05.

the protein fold is extremely conserved, in spite of a high sequence divergence and different function and organism source, as previously observed (e. g. Holm and Sander, 1994). Surprisingly, the extreme cases found in this study do not involve the ubiquitous Beta Barrel fold (CATH classification 2.40), but the Mainly Beta 7 Propeller and Trefoil fold (CATH classification 2.130 and 2.80 respectively), as shown in Table S4 and Fig. S2.

To test for a sampling imbalance, we constructed a fifth dataset, MonomerRND, by randomly selecting a small percentage of pairs from the cellular monomers dataset. The size of this new dataset is similar to that of the VCPs dataset. Table 2 shows the Kolmogorov-Smirnov test results when comparing all datasets, independently done for the TM-score and S_G. As expected, the test indicates that the MonomerRND dataset follows the same distribution as the cellular monomers dataset. However, all other datasets follow different distributions, even when comparing cellular monomers with higher order cellular *n*-mers.

Chotia and Lesk, 1986 proposed an exponential model to describe the relation between sequence identity and structure similarity. Here, we fit our data to such model, expressed as

$$(1 - \text{TM-score}) \sim f^* \exp(k^*(1 - S_G))$$

Table 3 shows the values of the computed coefficients of proportionality f and k, considering three different scenarios, i.e., homologous proteins, non-homologous proteins, and the whole S_G range. The differences between sequence scenarios and datasets are also illustrated in Fig. 2.

Table 4
Average protein surface area <*SA*>, in units of Å^2 [percentage of total], with standard deviation *SD*, estimated for the cellular monomers, cellular dimers, cellular oligomers, and icosahedral virus capsid proteins (VCPs) datasets. Values estimated for the total surface and the protein-protein interface with Pearson's correlation *r*. Estimation of the average surface density <*σ*>, in units of residues per 1000 Å^2, for the protein-protein interface and the solvent accessible surface (SAS).

	Total		Interfaces		Correlation	<σ>	
	<SA>	SD	<SA>	SD	r	Interface	SAS
Monomers	9169	7110	0 [00%]	0	–	0	16
Dimers	8318	2597	1073 [13%]	583	0.31	12	12
Oligomers	11,670	5411	2680 [23%]	2141	0.42	13	13
VCPs	16,535	5823	12,260 [74%]	4551	0.90	13	17

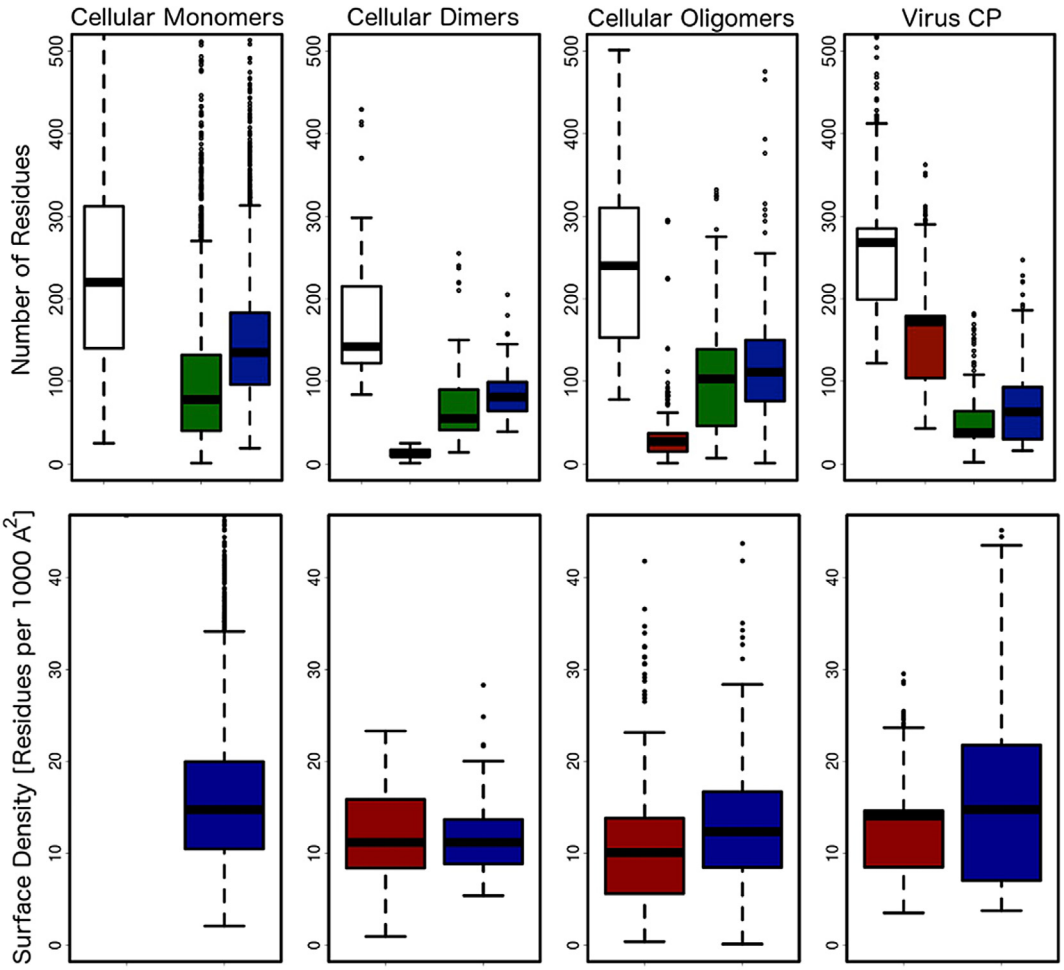

Fig. 4. Distribution of the number of residues (top row) and surface residue density (bottom row) in proteins, independently estimated for cellular monomers (first column), cellular dimers (second column), cellular oligomers (third column), and icosahedral virus capsid proteins (fourth column). Statistics summaries (median, first - third quartiles, minimum - maximum values, and outliers) are depicted with boxplots for total counts (empty), protein-protein interface (red), protein core (green), and proteiń s solvent accessible surface (blue).

3.2. Distribution of conserved residues

We estimated the protomer's surface area of cellular proteins and VCPs. Fig. 3 shows the distribution of values in each dataset and the correlation between the interface and the total surface area. We found a weak positive correlation in the case of cellular proteins, as opposed to VCPs, for which the correlation is strong. On average, cellular proteins tend to have a smaller total and interface surface area than VCPs (Table 4 and S5). A remarkable difference is that the interface region of VCPs takes 74% of the total surface, whereas for cellular dimers is only around 10%, and around 20% for higher *n*-mers. This observation is directly related to the number of residues comprising the structural categories. On average, cellular monomers and oligomers have the same total number of residues as VCPs (Fig. 4). This value is significantly smaller for cellular dimers (Table 5 and S6). More than half of the total residues make the interfaces of VCPs, whereas a small percentage are interface residues in cellular oligomers. In the case of cellular monomers, 40% of the total residues form the core of the protein. This amount is conserved in cellular dimers and oligomers but is reduced by half in VCPs. The other 60% of the total residues are exposed at the surface of cellular monomers. This amount is reduced by ~10% in the case of cellular dimers and oligomers in order to make the interface region. VCPs present a significantly smaller number of solvent exposed residues, being only about 26% of the total residues on average. Having the values of the surface area and the number of exposed and interface residues, we estimated the surface residue density distribution, σ. The residue density at the interface regions is the same, on average, in both cellular proteins and VCPs (Table 4 and S7). However, σ is significantly higher at the solvent accessible region for cellular monomers than it is for dimers and oligomers. Interestingly enough, VCPs have the same σ value at the solvent accessible area as cellular monomers.

S_G only gives an overall similarity between two amino acid sequences, estimating the relative percentage of residues that are identical in type of amino acid and position in the global alignment. To further investigate how the conserved residues are localized in the protein tertiary structure, we analyzed the distribution of the conserved residues on the different structural categories by estimating the sequence identity index per location category, S^*_k, in all datasets. Fig. 5 shows the correlation and distribution of S^*_k as a function of S_G, independently calculated for the interface, core and surface protein regions. The average values are reported in Table 6. Knowing how the size of the different structural categories contrast between cellular proteins and VCPs, it is not surprising to find that more than half of the conserved residues are in the interface region in the later case. The amount of conserved residues at the interface is low in cellular *n*-mers, given that this is a small region in such proteins. Interestingly, even though the number of residues is higher at the solvent accessible surface compared with

Table 5
Average number of residues <*NR*> [percentage of total], with standard deviation *SD*, in proteins at different structure locations independently estimated for cellular monomers, cellular dimers, cellular oligomers, and icosahedral virus capsid proteins (VCPs).

	Total		Interface		Core		Surface	
	<NR>	SD	<NR>	SD	<NR>	SD	<NR>	SD
Monomers	245	135	–	–	97 [40%]	77	148 [60%]	70
Dimers	178	83	13 [07%]	7	78 [44%]	55	87 [49%]	33
Oligomers	253	120	35 [14%]	35	102 [40%]	65	116 [46%]	65
VCPs	281	119	158 [56%]	63	51 [18%]	36	72 [26%]	47

Table 6
Average percentage of conserved residues at different structure locations <S^*_k>, with standard deviation *SD* and Pearson's correlation *r* with respect to the total sequence identity (S_G). Independent estimations made by the analysis of *n* pairs in the cellular dimers plus oligomers and icosahedral virus capsid proteins datasets, for $S_G > 0$, $S_G > 0.3$, or $S_G < 0.3$.

	Cellular dimers + Oligomers				Viral CP			
	n	<S^*_k>	SD	*r*	*n*	<S^*_k>	SD	*r*
	Interface							
$S_G < 30\%$	3832	14	6	−0.64	2433	42	0.23	0.67
$S_G > 30\%$	59	9	4	0.13	1453	66	0.12	−0.34
$S_G > 0$	3891	13	6	−0.42	3886	52	0.22	0.5
	CORE							
$S_G < 30\%$	3832	41	18	−0.04	2433	23	13	−0.34
$S_G > 30\%$	59	32	1	−0.43	1453	15	6	0.13
$S_G > 0$	3891	41	18	−0.08	3886	20	11	−0.34
	Solvent accessible surface							
$S_G < 30\%$	3832	30	15	−0.06	2433	20	11	−0.36
$S_G > 30\%$	59	38	14	0.32	1453	8	6	0.55
$S_G > 0$	3891	30	15	0.05	3886	12	12	−0.16
	Orphans							
$S_G < 30\%$	3832	15	–	–	2433	–	–	–
$S_G > 30\%$	59	21	–	–	1453	11	–	–
$S_G > 0$	3891	16	–	–	3886	16	–	–

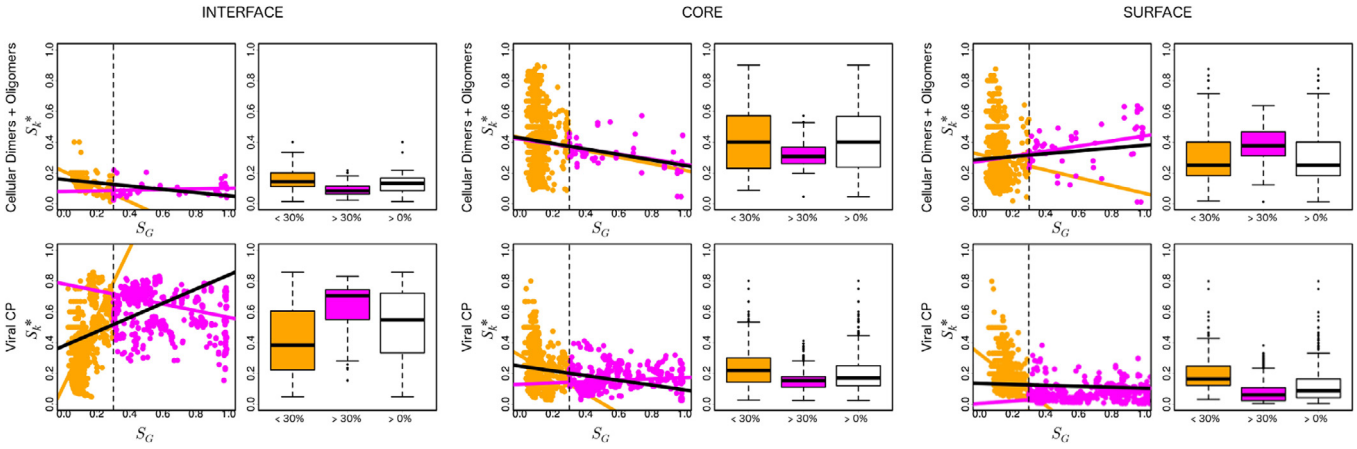

Fig. 5. Correlation between the specific location of conserved residues (S^*_k) and the total sequence identity (S_G) in proteins, at the protein-protein interface, protein core, and proteiń solvent accessible surface, independently estimated for cellular dimers plus oligomers (top rows), and icosahedral virus capsid proteins (bottom rows). Each point in the cloud represents a pair of polypeptide chains. Linear regression and statistical analysis are shown for pairs with $S_G > 0$ (black, white), $S_G > 0.3$ (magenta), and $S_G < 0.3$ (orange). Statistics (median, first – third quartiles, minimum – maximum values, and outliers) are summarized with boxplots.

those at the proteiń s buried core (Fig. 4), the amount of conserved residues is significantly higher at the later in both cellular proteins and VCPs.

In addition to considering the whole S_G range to perform the analyses, we also used a 30% sequence identity arbitrary threshold to compare non-homologous vs. homologous protein scenarios. The correlation of S_k^* with respect to S_G is weak to nonlinear in all cases, except for the interface region of non-homologous cellular proteins, which have a strong negative correlation, and non-homologous VCPs, having a strong positive correlation. There is a variation in the average amount of conserved residues at all structure categories for cellular proteins and VCPs, although the relative behavior is different. In the case of cellular *n*-mers, S_k^* is higher for non-homologous proteins at the interface and core than homologous proteins. The opposite is seen at the solvent exposed surface. In the case of VCPs, S_k^* is higher for non-homologous proteins at the core and the solvent exposed surface than homologous proteins. The opposite is seen at the interface. We found another interesting difference between cellular *n*-mers and VCPs. Whereas the amount of the percentage of conserved residues in the orphan category increases for homologous proteins with respect to non-homologous in cellular *n*-mers, the opposite is seen in VCPs (Table 6).

3.3. Residue conservation in the structural categories

We estimated the entropy-based conservation by residue $S(i)$ for each protomer in our datasets. Table 7 shows the average residue conservation <*S*> and the normalized value <*s*> calculated by structural category. In the case of cellular proteins, the normalized residue conservation does not change much from the non-normalized residue conservation since <*S*> for the whole polypeptide chain is approximately 1. In general, the core and interface regions are significantly more conserved than the solvent exposed surface, with some preference for the buried volume, specially at cellular oligomers (Table S8).

On the other hand, we found that the probability distribution of <*S*> for the VCPs dataset is bimodal, as illustrated in Fig. 6. We

Table 7
Average entropy-based residue conservation estimation per polypeptide chain, <S>, and normalized residue conservation, <s>, in proteins, independently estimated for cellular monomers, cellular dimers, cellular oligomers, and icosahedral virus capsid proteins (VCPs). Average residue conservation estimated for the protein-protein interface, protein core, and proteiń s solvent accessible surface (SAS). Standard deviation indicated in parentheses.

	Monomers		Dimers		Oligomers		VCPs		
	<s>	<S>	<s>	<S>	<s>	<S>	<s>†	$\langle S\rangle_{G1}$††	$\langle S\rangle_{G2}$††
Interface	–	–	0.91 (0.27)	0.98 (0.44)	0.94 (0.14)	0.92 (0.34)	0.96 (0.13)	0.41 (0.17)	0.99 (0.16)
Core	0.78 (0.10)	0.83 (0.31)	0.83 (0.15)	0.84 (0.26)	0.77 (0.09)	0.75 (0.26)	0.75 (0.12)	0.28 (0.09)	0.84 (0.10)
SAS	1.23 (0.10)	1.29 (0.41)	1.26 (0.18)	1.28 (0.37)	1.29 (0.13)	1.23 (0.34)	1.29 (0.12)	0.56 (0.21)	1.24 (0.18)

† Values considering one single probability distribution.
†† Values assuming two independent probability distributions.

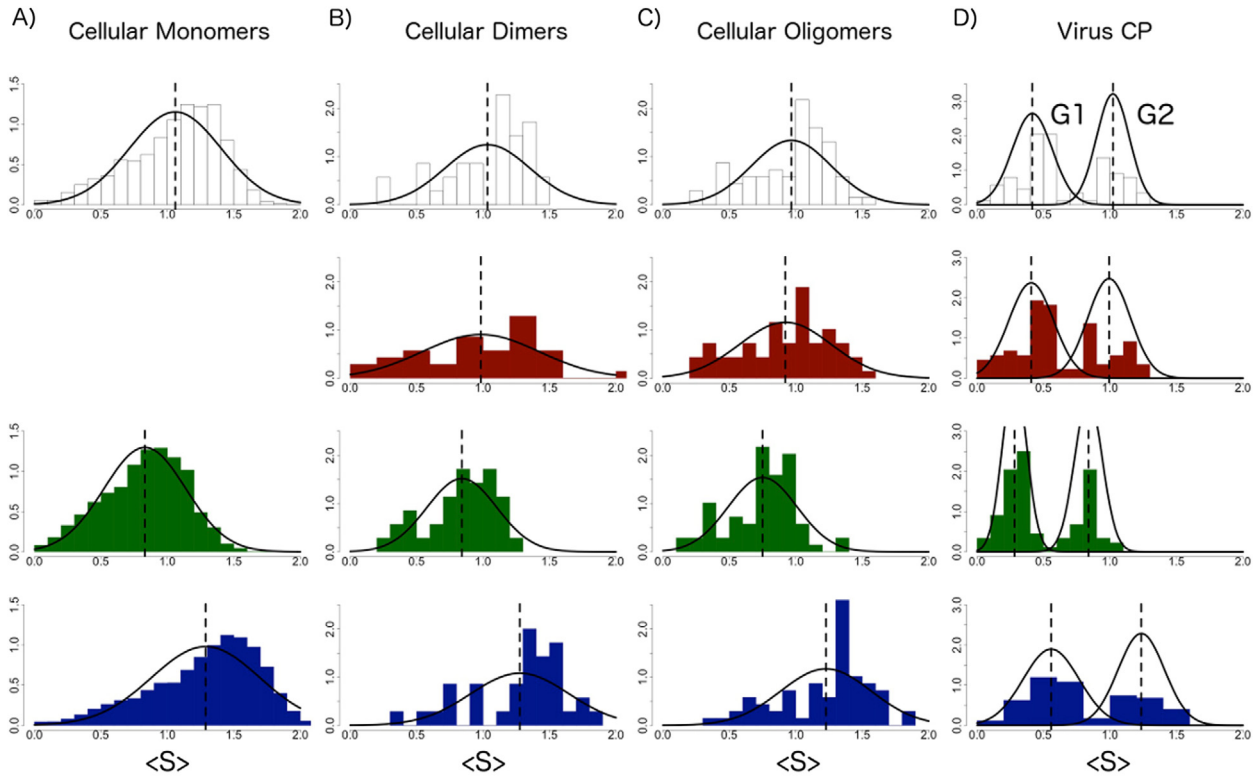

Fig. 6. Probability distribution of the average entropy-based residue conservation estimation per polypeptide chain, <S>, in proteins, independently estimated for cellular monomers (A), cellular dimers (B), cellular oligomers (C), and icosahedral virus capsid proteins (D). Average residue conservation estimated for the whole polypeptide chain (white), protein-protein interface (red), protein core (green), and protein's solvent accessible surface (blue). Normal distribution with the same mean (vertical dashed line) and standard deviation of the probability distribution is shown for each case.

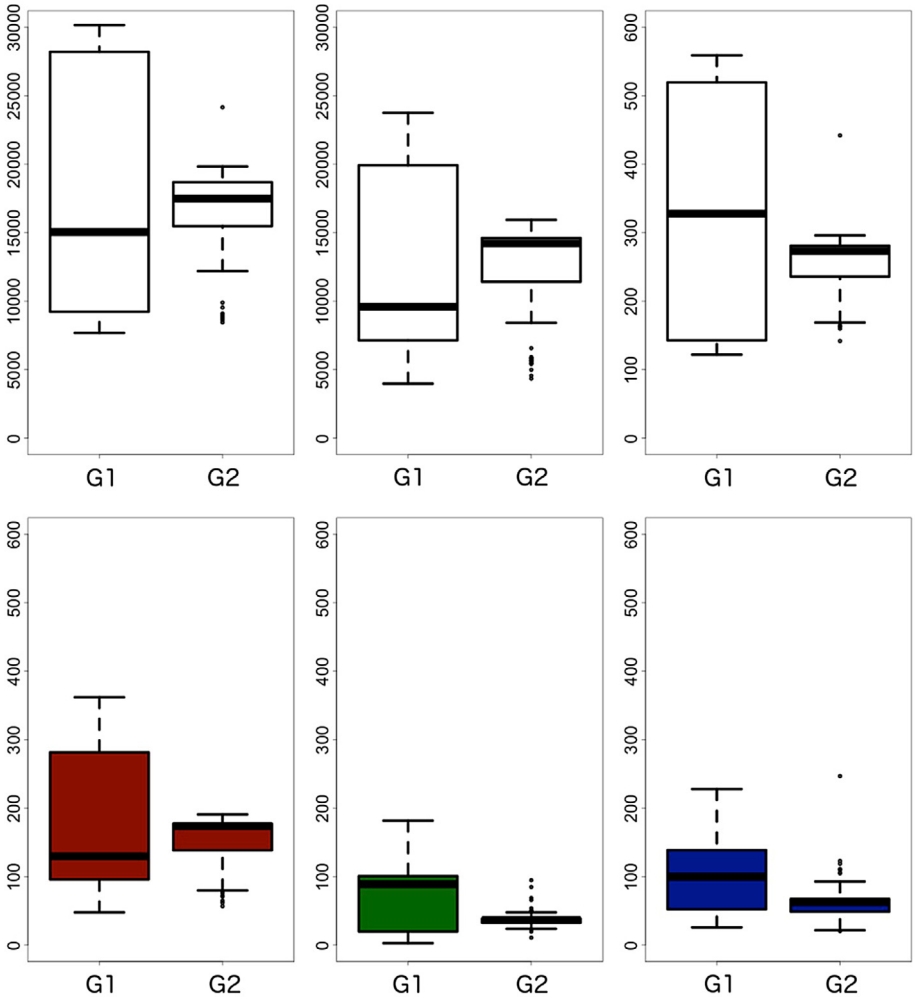

Fig. 7. Statistical comparison between virus families group 1 and group 2 (G1 and G2 respectively). Characterization of (top row, from left to right) total solvent accessible surface area (SASA, in Å^2), interface SASA, total number of residues per polypeptide chain, (bottom row, from left to right) number of residues at the protein-protein interface (red), protein core (green), proteiń s solvent accessible surface (blue). Statistics (median, first – third quartiles, minimum – maximum values, and outliers) are summarized with boxplots.

labeled the distinct distributions as G1 and G2. It is interesting that the average residue conservation per structure category of G2 behaves as the cellular proteins. However, G1 seems to be evolving much slower. We identified the virus families that belong to each group as: Adenoviridae (T = 1), Birnaviridae (T = 1), Bromoviridae (T = 3), Caliciviridae (T = 3), Comoviridae (T = pT3), Hepadnaviridae (T = 4), Hepeviridae (T = 1), Leviviridae (T = 3), Microviridae (T = 1), Nodaviridae (T = 3), Parvoviridae (T = 1), Polyomaviridae (T = 7d), Sobemoviridae (T = 3), and Tetraviridae (T = 4) in G1, and Dicistroviridae (T = p3), Picornaviridae (T = pT3), Siphoviridae (T = 7 L), Sobemoviridae (T = 3), Togaviridae (T = 4), Tombusviridae (T = 3), and Tymoviridae (T = 3) in G2. There is no obvious correlation with the T number. Both groups have, on average, the same total and interface solvent accessible surface area, and the same number of residues making the interface region. However, one significant difference is in the number of residues comprising the core and the exposed surface, with G1 having almost twice as many as G2 (Fig. 7 and Table S9).

4. Discussion

Our findings reveal several differences between cellular protein oligomers (*n*-mers) and icosahedral viral capsid proteins regarding the location, amount, and level of conservation of residues in the tertiary structure. We analyzed and compared four datasets, namely, cellular monomers, cellular dimers, cellular oligomers, and VCPs. Overall, cellular monomers and VCPs seem to be two extremes in the quaternary structural diversity found in nature, with the cellular dimers and oligomers as an intermediate state. Our main results are summarized in Fig. 8.

The correlation between the sequence identity and the tertiary structure conservation seems to follow an exponential model in all cases (Chotia and Lesk, 1986). However, their distribution in [S_G, TM-score] space is quite different (Fig. 2). Cellular monomers are concentrated in a high density, non-homologous, different fold region. On the other hand, VCPs are evenly distributed in three main regions along the whole range of sequence homology and tertiary structure similarity. Cellular dimers and oligomers, although similar to cellular monomers, show a distinct distribution that seems to be in between cellular monomers and VCPs.

We performed a deeper examination of the correlation between protein global sequence identity (S_G) and structural similarity (TM-score) for the case of VCPs. A large majority of pairs having TM-score values >0.7 are capsid proteins that belong to the same virus family. Likewise, most pairs having TM-score values <0.7 are capsid proteins that belong to different virus families. Of note, ICTV taxonomic classification of viruses does not explicitly include structural information as criteria to group viruses. Interestingly,

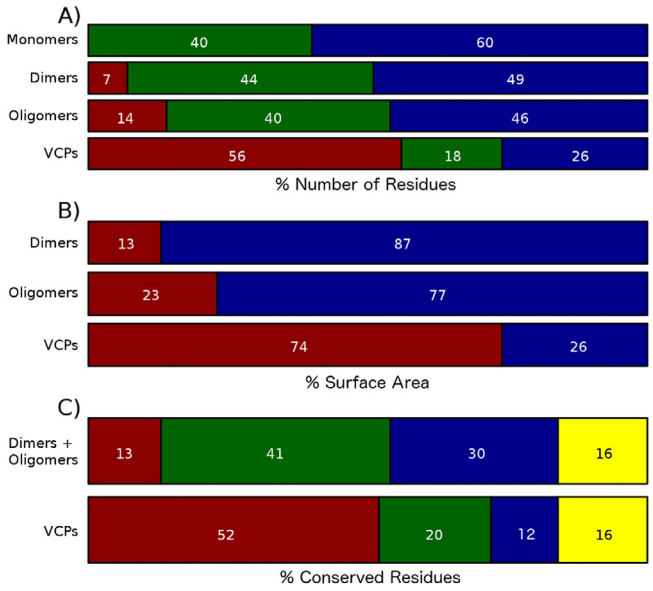

Fig. 8. Average values for the percentage of number of residues (A), the percentage of surface area (B), and the percentage of conserved residues (C) in proteins, at the protein-protein interface (red), protein core (green), protein's solvent accessible surface (blue), and orphans (yellow), independently estimated for cellular monomers, cellular dimers, cellular oligomers, and icosahedral virus capsid proteins (VCPs).

there is a considerable proportion of VCP pairs in the TM-score range between 0.5 and 0.7. They are inter-family viruses that have a global sequence identity below 20%. This observation suggests a substantial sequence divergence during virus evolution. We speculate that virus families have evolved from a few common ancestors, a hypothesis also proposed recently by Nasir and Caetano-Anollés, 2015. Even though some of the comparisons were made between homologous VCPs within the same family, sometimes with sequence identities reaching 90%, these viruses are distinct and display unique virological/serological properties. Importantly, different viruses within the same family can cause distinct diseases (e.g., polio vs. common cold). Hence, we believe that the comparisons made in this study between homologous VCPs are appropriate.

It is important to remark that despite this sequence divergence, the protein fold (e.g., jelly-roll β-barrel) persists and that the majority of conserved residues remain at the protein-protein interfaces. This observation is in agreement with previous conclusions drawn from different analysis approaches regarding evolution and structure conservation. Abroi and Gough, 2011, suggested that the virosphere could be an engine for the genesis of cellular protein structures. Cheng and Brooks, 2013, mention that the structural, but not functional, close relationship found between some classes of modern cellular proteins and VCPs resulted from ancient genetic interactions between viruses and their hosts. In this sense, our results are yet another indication that VCPs are a good model for further investigation on sequence divergence due to pressures rendered by the host immune system.

The distribution of conserved residues in the protein tertiary structure is also different between cellular proteins and VCPs. On average, 41% of the conserved residues in cellular dimers and oligomers are found at the buried core, followed by 30% at the exposed surface, and 13% at the interface region. However, more than half of the total conserved residues of VCPs are found at the interface, followed by 20% at the core, and only 12% at the surface. About 15% of the conserved residues could not be matched to a common structural category and were classified as orphans (ORPH). We found that the distribution of conserved residues is different in cellular proteins compared to VCPs. This observation is correlated to the number of residues making the different protein structure categories. This result suggests that the conserved residues are evenly distributed over the whole protein structure in all cases. Again, we can see that the cellular dimers and oligomers have intermediate values between those of cellular monomers and VCPs (Fig. 4).

Because icosahedral VCPs self-assemble with multiple neighbors to form capsid shells, three-quarters of the protein surface is used to make the protein-protein interfaces, with half of the total conserved residues at this region. This finding is in agreement with evidence showing that mutations in virus capsids mostly take place at the solvent accessible surface, as a way of countering/evading host immune responses (e.g. Jameson et al., 1985; Kanda et al., 1986; Yang et al., 2005; Vitiello et al., 2005) in an evolutionary positive selection manner (Esteves et al., 2008). Of note, a perfect correlation can be seen between the residue structure classification and conservation analysis done in this work with the level of conservation and structural features of spherical capsids recently reported by Chih-Min et al., 2015 (Fig. S3).

The distribution of the variation on the level of residue conservation in different structural categories in the protein is similar in all cellular protein *n*-mers. On average, the buried core is the most conserved, closely followed by the interface region (Grishin and Phillips, 1994; Valdar and Thornton, 2001; Guharoy and Chakrabarti, 2005). There is a greater sequence variation at the solvent exposed surface in all cellular proteins (Caffrey et al., 2004). This relation is also true for some VCPs. We clearly identified two distinct groups of virus families that behave differently concerning sequence variation (see *Residue conservation in the structural categories* in the Results section). One of these groups, G2, has residue conservation average values very similar to those of cellular proteins. However, the second group, G1, seems to have significantly lower residue variations (Table S10), although the relative differences between structure categories remain the same as in all other cases. We found that a difference between G1 and G2 is the number of residues making the buried core and the solvent exposed surface, with the later having significantly lower values. Bahadur and Janin, 2008 analyzed the residue conservation of 32 icosahedral viruses and reported normalized values, $\langle s \rangle$, for the interface, core, and surface of 0.9, 0.7 and 1.6, respectively. We can reproduce those results if we assume a single probability distribution of $\langle S \rangle$, as can be seen in Table 7 and Fig. S4. Their approach and small dataset precluded the realization that their results were the average of two distinct distributions. The reason and the meaning of the existence of these two groups of virus families deserve further investigation.

The high-resolution cutoff criteria used for assembling our datasets provides high confidence in the reported results. Hence, these datasets are a good representative sample of nature's protein diversity. Our analyses can readily be extended later as more structural data becomes available. In this work, we have included all the icosahedral virus structures available to date. Other viral capsid topologies, such as helicoidal viruses, were not included due to a limited small number of structures available, but will be considered in future studies.

5. Conclusions

Our work extends and complements results previously reported. We find a general agreement regarding sequence variations occurring at different regions of the tertiary structure of cellular proteins and spherical virus capsid proteins. However, we could detect two distinct virus family groups with seemingly

different evolution rates. The robust statistical analyses performed on high-resolution structural datasets had the power to highlight important differences between cellular proteins and spherical VCPs. Our results provide evidence that cellular monomers and VCPs are two extremes in the quaternary space of protein complexes, with the cellular dimers and oligomers being an intermediate state.

Acknowledgements

We acknowledge the thoughtful suggestions of the editor and reviewers which greatly improved this article. V.S.R and M.C.T. would also like to acknowledge the discussions with Professor Charles L. Brooks III at the inception of this work. M.C.T. would like to thank Dr. Johan Van Horebeek from the Computer Science Department, CIMAT, México, for his helpful advice on the statistical methodology. This work was supported by the USA National Institutes of Health (NIH) to the center of Multi-scale modeling tools for structural biology (MMTSB) grant number RR012255 to V.S.R., the Mexican Consejo Nacional de Ciencia y Tecnología (Conacyt) grant number 132376 to M.C.T., and the 2013 Fulbright-García Robles funding to D.J.M.-G and M.C.T. by the USA J. William Fulbright Scholarship Board.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jsb.2016.07.013.

References

Abroi, A., Gough, J., 2011. Are viruses a source of new protein folds for organisms? – virosphere structure space and evolution. BioEssays 33, 626–635.
Bahadur, R.P., Janin, J., 2008. Residue conservation in viral capsid assembly. Proteins 71, 407–414.
Bahadur, R.P., Rodier, F., Janin, J., 2007. A dissection of the protein-protein interfaces in icosahedral virus capsids. J. Mol. Biol. 367, 574–590.
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G.T., Bhat, N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.
Caffrey, D.R., Somaroo, S., Huges, J.D., Mintseris, J., Huang, E.S., 2004. Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci. 13, 190–202.
Cann, J.A., 2005. Principles of Molecular Virology, fourth ed. Elsevier Academic Press.
Carrillo-Tripp, M., Shepherd, C.M., Borelli, I.A., Venkataraman, S., Natarajan, P., Johnson, J.E., Brooks, C.L., Reddy, V.S., 2009. VIPERdb2: an enhanced and web API enabled relational database for structural virology. Nucleic Acids Res. 37, D436–D442.
Caspar, D.T., Klug, A., 1962. Physical principles in the construction of regular viruses, first ed., 27. Press, Cold Spring Harbor Laboratory.
Cheng, S., Brooks III, C.L., 2013. Viral capsid proteins are segregated in structural fold space. PLoS Comput. Biol. 9, e1002905.
Chih-Min, C., Yu-Wen, H., Tsun-Tsao, H., Chung-Shiuan, S., Jenn-Kang, H., 2015. Sequence conservation, radial distance and packing density in spherical viral capsids. PLoS ONE 10, e0132234.
Chotia, C., 1974. Hydrophobic bonding and accessible surface area in proteins. Nature 254, 338–339.
Chotia, C., Janin, J., 1975. Principles of protein-protein recognition. Nature 256, 705–708.
Chotia, C., Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. EMBO J. 5, 823–826.
Damodaran, K., Reddy, V.S., Johnson, J.E., Brooks III, C.L., 2002. A general method to quantify quasi-equivalence in icosahedral viruses. Mol. Biol. 324, 723–737.
Esteves, P.J., Abrantes, J., Carneiro, M., Müller, A., Thompson, G., van der Loo, W., 2008. Detection of positive selection in the major capsid protein VP60 of the rabbit haemorrhagic disease virus (RHDV). Virus Res. 137, 253–256.
Fauquet, C., Mayo, M.A., Maniloff, J., Desselberger, U., Ball, L.A., 2005. Virus Taxonomy: Eighth Report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press.
Grishin, N.V., Phillips, M.A., 1994. The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. Protein Sci. 3, 2455–2458.
Guharoy, M., Chakrabarti, P., 2005. Conservation and relative importance of residues across protein-protein interfaces. PNAS 102, 15447–15452.
Holm, L., Sander, C., 1994. Structural similarity of plant chitinase and lysozymes from animals and phage. An evolutionary connection. FEBS Lett. 340, 129–132.
Jameson, B.A., Bonin, J., Wimmer, E., Kew, O.M., 1985. Natural variants of the Sabin type 1 vaccine strain of poliovirus and correlation with a poliovirus neutralization site. Virology 143, 337–341.
Janin, J., Bahadur, R.P., Chakrabarti, P., 2008. Protein-protein interaction and quaternary structure. Quat. Rev. Biophys. 41, 133–180.
Jones, S., Thornton, J.M., 1995. Protein-protein interactions: a review of protein dimer structures. Prog. Biophys. Mol. Biol. 63, 31–65.
Kanda, T., Furuno, A., Yoshiike, K., 1986. Mutation in the VP-1 gene is responsible for the extended host range of a monkey B-lymphotropic papovavirus mutant capable of growing in T-lymphoblastoid cells. J. Virol. 59, 531–534.
Lawrence, M.C., Colman, P.M., 1993. Shape complementarity at protein/protein interfaces. J. Mol. Biol. 234, 946–950.
Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.
Nasir, A., Caetano-Anollés, G., 2015. A phylogenomic data-driven exploration of viral origins and evolution. Sci. Adv. 1, e1500527.
Ofran, Y., Rost, B., 2003. Analysing six types of protein-protein interfaces. J. Mol. Biol. 325, 377–387.
Richards, F.M., 1974. The interpretation of protein structures: total volume, group volume distributions and packing density. J. Mol. Biol. 82, 1–14.
Rost, B., 1999. Twilight zone of proteins sequence alignments. Protein Eng. 12, 85–94.
Sander, C., Schneider, R., 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins: structure. Funct. Genet. 9, 56–68.
Shrake, A., Rupley, J.A., 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J. Mol. Biol. 79, 351–364.
Talavera, D., Robertson, D.L., Lovell, S.C., 2011. Characterization of protein–protein interaction interfaces from a single species. PLoS One 6, e21053.
Tokuriki, N., Oldfield, C.J., Uversky, V.N., Berezovsky, I.G., Tawfik, D.S., 2009. Do viral proteins possess unique biophysical features? Trends Biochem. Sci. 34, 53–59.
Touw, W.G., Baakman, C., Black, J., te Beek, T.A.H., Krieger, E., Joosten, R.P., Vriend, G., 2015. A series of PDB related databases for everyday needs. Nucleic Acids Res. 43 (Database issue), D364–D368.
Valdar, W.S.J., Thornton, J.M., 2001. Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins Struct. Funct. Genet. 42, 108–124.
Vitiello, C.L., Merril, C.R., Adhya, S., 2005. An amino acid substitution in a capsid protein enhances phage survival in mouse circulatory system more than a 1000-fold. Virus Res. 114, 101–103.
Xu, J., Zhang, Y., 2010. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 26, 889–895.
Yan, C., Wu, F., Dobbs, D., Honabar, V., 2008. Characterization of protein–protein interfaces. Protein 27, 59–70.
Yang, R., Wheeler, C.M., Chen, X., Uematsu, S., Takeda, K., Akira, S., Pastrana, D.V., Viscidi, R.P., Roden, R.B.S., 2005. Papillomavirus capsid mutation to escape dendritic cell-dependent innate immunity in cervical cancer. J. Virol. 79, 6741–6750.
Zhang, Y., Skolnick, J., 2004. Scoring function for automated assessment of protein structure template quality. Proteins: Struct. Funct. Bioinf. 57, 702–710.
Zhang, Y., Skolnick, J., 2005. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302–2309.