PCCP

Dynamic Article Links

Cite this: *Phys. Chem. Chem. Phys.*, 2011, **13**, 17044–17055

www.rsc.org/pccp

PAPER

Why not consider a spherical protein? Implications of backbone hydrogen bonding for protein structure and function†

Michal Brylinski, Mu Gao and Jeffrey Skolnick*

Received 11th April 2011, Accepted 20th May 2011
DOI: 10.1039/c1cp21140d

The intrinsic ability of protein structures to exhibit the geometric features required for molecular function in the absence of evolution is examined in the context of three systems: the reference set of real, single domain protein structures, a library of computationally generated, compact homopolypeptides, artificial structures with protein-like secondary structural elements, and quasi-spherical random proteins packed at the same density as proteins but lacking backbone secondary structure and hydrogen bonding. Without any evolutionary selection, the library of artificial structures has similar backbone hydrogen bonding, global shape, surface to volume ratio and statistically significant structural matches to real protein global structures. Moreover, these artificial structures have native like ligand binding cavities, and a tiny subset has interfacial geometries consistent with native-like protein–protein interactions and DNA binding. In contrast, the quasi-spherical random proteins, being devoid of secondary structure, have a lower surface to volume ratio and lack ligand binding pockets and intermolecular interaction interfaces. Surprisingly, these quasi-spherical random proteins exhibit protein like distributions of virtual bond angles and almost all have a statistically significant structural match to real protein structures. This implies that it is local chain stiffness, even without backbone hydrogen bonding, and compactness that give rise to the likely completeness of the library solved single domain protein structures. These studies also suggest that the packing of secondary structural elements generates the requisite geometry for intermolecular binding. Thus, backbone hydrogen bonding plays an important role not only in protein structure but also in protein function. Such ability to bind biological molecules is an inherent feature of protein structure; if combined with appropriate protein sequences, it could provide the non-zero background probability for low-level function that evolution requires for selection to occur.

1. Introduction

Proteins are dense geometric objects with a variety of different structural properties that emerge on different scales. Locally, due to the requirement that residues be hydrogen bonded,[1] proteins often adopt regular secondary structure with roughly 60% of their residues assigned to helices or β-strands.[2] On a more global level, most single domain proteins are ellipsoidal in shape.[3–5] The structural comparison of the library of experimental single domain protein structures to a library of artificially generated compact, hydrogen bonded polypeptide structures, led to the conclusion that the library of single domain proteins is likely complete.[6] It was also suggested that this completeness requires backbone hydrogen bonds in protein structures. This suggestion was based on the results for the structural comparison of Freely Jointed Chain (FJC) structures to the PDB,[7] whose average TM-score (a measure of structural similarity) = 0.3,[8,9] the value of a pair of randomly related protein structures.[6] Thus, we investigated what would happen if chains devoid of main chain hydrogen bonds (and consequently regular secondary structure) but with protein-like local chain geometry, as opposed to the highly flexible FJCs, were constructed subject to the restraint that they be spherical and packed at protein-like densities. Would such quasi-spherical random structures have significant structural matches to real protein structures? What features, if any, do such highly idealized quasi-spherical random proteins share with native protein structures as well as to artificially generated compact structures that contain protein like secondary structure and backbone hydrogen bonding? More generally, what are the minimum requirements that give rise to the observed structural and geometric properties of proteins? These are the questions we seek to address in this contribution.

It is important to remember that in a cell, proteins perform functional roles that on a molecular level involve interactions

Center for the Study of Systems Biology, Georgia Institute of Technology, 250 14th St NW, Atlanta GA, 30076, USA. E-mail: skolnick@gatech.edu

† This article was submitted as part of a themed collection on the *Physical Foundations of Protein Folding*.

with other molecules. Here, we focus on purely geometric properties and defer discussions of sequence dependent effects to future work. The differentiation of structure and sequences effects is conceptually useful because a necessary requirement is that the local geometry and packing permit the intermolecular interaction to occur. The geometrically allowed interacting complex must then have a favorable free energy in order that the bound pose be significantly populated. For example, proteins often bind small molecule ligands[10–13] or metals[14] in cavities in their structure. Such binding might lead to an allosteric transition[15–17] or might be the first step in catalysis,[18,19] if the protein happens to be an enzyme. Without cavities, ligand binding is far less likely to occur, as an enveloping surface is required to generate a sufficient number of intermolecular interactions with the small molecule to stabilize the complex. For real proteins that are the product of both physics and evolution, it is difficult to ascertain whether cavities must be selected for by evolution or are an inherent feature of protein-like structures. By performing a computer experiment, we can *a priori* eliminate evolutionary effects and focus on purely structural features. An important question is what are the minimum structural properties required to generate cavities? Is it just the packing of side chains independent of the secondary structure of the backbone? If not, what happens when regular secondary structure is included?

Similarly, both protein–protein and protein–DNA interactions occur at interfaces that involve geometrically complementary surfaces. In the case of proteins, their interaction interface has a strong tendency to be planar (see ref. 20 and Results below) because a relatively flat surface provides a sufficient number of complementary interacting residues to provide a favorable free energy of association. Recent work[20] demonstrated that the library of protein interfaces is likely complete and comprised of $\sim$1000 statistically distinguishable interfaces, with the majority ($\sim$83%) recovered on docking artificial monomeric structures generated by the packing of hydrogen bonded, secondary structural elements. Again, the issue is what are the requirements for generating quasi-planar protein–protein interfaces? Does it result from the packing of regular secondary structural elements such as helices and strands or does it emerge from the purely local, residue geometries that reproduce the local ϕ/ψ distribution seen in proteins?

Turning to protein/DNA interactions, while the detailed geometry of DNA/protein interfaces has been analyzed,[21–23] the number of statistically distinct interfaces and the completeness of the space of DNA–protein interfaces are not known. Moreover, while it has been demonstrated that protein–protein interactions mainly involve planar interfaces, the distribution of shapes of DNA–protein surfaces is not as well characterized. Is it possible that artificial homopolypeptide protein structures have the requisite surface geometry (geometries) that is complementary to DNA? If so, this would suggest that the geometric ability to interact with DNA is also an inherent feature of protein structure and does not require evolutionary selection. What about quasi-spherical proteins lacking regular secondary structure? Can they similarly interact? These are important questions that bear on the intrinsic ability to engage in macromolecular interactions without the selection pressure of evolution.

The outline of this paper is as follows: The Materials and Methods section describes the set of real, artificial and quasi-spherical protein structures; how they are generated and analyzed. Then, in the Results and Discussion, for each type of structural property, we compare and contrast the results for quasi-spherical random structures and artificial structures with real protein structures. We begin with an examination of local geometric properties, secondary structure content, and backbone hydrogen bonding. We next focus on the global structural similarity of the quasi-spherical random structures and artificial structures to real protein structures and vice versa. Then, the nature of their internal packing as assessed by their surface/volume ratios and the overall molecular shape are characterized. Subsequently, the size distributions of cavities in the different types of structures are examined. Then, the nature of their protein–protein and protein–DNA interaction interfaces is explored. Finally, the Conclusions highlight the implications of this work.

2. Materials and methods

2.1 Library of crystal structures

The full PROSPECTOR_4 template[24] library, termed PDB contains 13 148 structures between 40 and 1962 residues in length and covers the entire Protein Data Bank[7] at 35% pairwise sequence identity. PDB300 (PDB250) is the subset of PDB composed of proteins $\leq$300 (250) residues in length and contains 9867 (6999) proteins.

2.2 Artificial homopolypeptide library

For each member of a subset of PDB300 comprised of 4968 proteins, following the procedure previously described in ref. 25, TASSER simulations of a polyvaline homopolypeptide with the corresponding secondary structures were undertaken and the top two structural clusters of the resulting compact, hydrogen bonded protein structures were selected. The resulting library, termed artificial300, contains 9935 structures; artificial250, that contains 8011 structures, is the subset of proteins $\leq$250 residues in length.

For calculations that require a specific sequence (see below), a randomized sequence having the same composition as the corresponding native protein is generated,[24] then the all-atom conformation is rebuilt from the Cα trace by Pulchra,[26] and additionally energy-minimized in the CHARMM22 force field[27] using the Jackal modeling package.[28] The list of proteins and corresponding artificial homopolypeptide structures and all atom models may be found at http://cssb.biology.gatech.edu/suppl/quasi_spheres.

2.3 Quasi-spherical random structure library

For each target protein whose length corresponds to one of the PDB300 proteins, we build its quasi-spherical random conformation as follows: First, we construct an ideal sphere, whose volume (V) is estimated from the number of residues (N):

$$V = 133.74 \times N - 524.73 \quad (1)$$

Next, the sphere is randomly populated with N Cα atoms such that the distance between any pair of Cα atoms is >3.8 Å. For a given arrangement of Cα atoms, we construct the shortest Cα trace by solving the traveling salesman problem (TSP) in three dimensions. Here, we use the Concorde TSP solver (http://www.tsp.gatech.edu/concorde/), which is currently applicable to many thousands of points.[29] No explicit restrictions on local chain geometry are imposed; *viz*., ***we do not enforce the virtual bond angles or dihedral angles seen in real protein structures***. The particular protein sequence is a randomized version of the corresponding PDB300 sequence and the same rebuilding procedure as for the artificial structure library was followed, *i.e*. the all-atom conformation was rebuilt from the Cα trace by Pulchra[26] and energy-minimized in the CHARMM22 force field.[27] The set of 8249 (6995) quasi-spherical proteins ≤ 300 (250) residues in length is called quasi-spherical300 (quasi-spherical250). The list of proteins and corresponding all-atom models of quasi-spherical structures can be found at http://cssb.biology.gatech.edu/suppl/quasi_spheres.

2.4 Library of protein–protein and protein–DNA complexes

The library of 1690 dimeric protein–protein complexes was taken from the M-TASSER[30] template library. The complexes were selected such that at most one monomer in one complex can share a global sequence identity $>35\%$ with respect to another monomer from any other complex in the library. The library of 399 DNA-binding domains were taken from the DBD-Threader[31] template library of DNA–protein complexes. The global sequence identity is less than 90% between any two DNA-binding proteins in this library. The list of protein–protein complexes and DNA-binding proteins and their corresponding structures can be found at http://cssb.biology.gatech.edu/suppl/quasi_spheres.

2.5 Structural properties

To characterize protein structures across the different sets, crystal, quasi-spherical random structures and artificial, we use and analyze a variety of local as well as global structural properties that are summarized in Table 1.

2.6 Analysis of protein–protein complexes

A possible native-like protein–protein complex must satisfy the following two conditions: (i) each monomer of the putative complex has a structure significantly similar to a separate monomer from the native dimer complex structure library (ii) the two protein–protein interfaces also have significantly similar structures.

Table 1 Structural properties used to characterize protein structures

Property	Implementation	Ref.
Molecular volume	MSMS	41
Accessible surface area	POPS/NACCESS	33, 42
Mass-weighted principal axes	in-house	43, 44
Phi-Psi distribution	in-house	45
Amino acid flexibility index	in-house	35
Hydrogen bonds (high-resolution)	HBPLUS	1
Hydrogen bonds (low-resolution)	TASSER	46
Pockets and cavities	LIGSITE	38, 39
Backbone knots	KNOT	47
Protein Structural Similarity	TM-score	8
Interface Structural Similarity	iAlign	32
Planarity	SURFNET	34

As the first step in identifying quasi-spherical and artificial proteins that adopt a similar quaternary structure as native proteins, a structural alignment between monomeric spherical and native structures is conducted with the program TM-align;[9] a putative monomer in the complex must have a TM-score ≥ 0.4 to the closest native monomer. Then, using the resulting structural alignment to position the two molecules in the dimer, we assess the structural similarity of the interface to that in the corresponding native pair of proteins. To calculate the side chain contact based, interface similarity, IS-score,[32] an all-atom structure was built with PULCHRA,[26] and a heavy-atom distance cutoff of 4.5 Å is employed to define a protein–protein interfacial contact. A protein–protein interface is defined as the collection of all residues with at least one interfacial contact between protein pairs. Protein–protein interface comparison between putative spherical structures and native protein complexes was conducted with the program iAlign[32] in the sequential alignment mode. To eliminate those complexes with significant clashes, we remove a putative complex if it has more than one interfacial contact within a 1 Å distance cutoff.

To define a surface patch, a seed surface residue is first selected. Then, the patch is enlarged by adding the nearest neighbor, the second nearest neighbor, and so on. The procedure stops when the total accessible surface area, ASA, of the patch reaches a pre-defined threshold value, *e.g*., 1000 Å^2. The ASA was calculated with the program NACCESS.[33] For spherical and artificial structures, an evolved sequence[24] is arbitrarily chosen for the ASA calculations. A protein residue is defined as a surface residue if its relative ASA is larger than 1% according to NACCESS. The planarity of a surface patch is defined as the RMSD of the Cα atoms from the best-fit plane as given by the program SURFNET.[34] The curvature of a surface patch is defined as $1/r$, where r is the radius of the best-fit sphere that minimizes the sum of the squared distance from the patch Cα atoms to the sphere. The least square minimization was implemented with the statistical computing platform R (http://www.R-project.org).

2.7 Analysis of DNA–protein complexes

The procedure for building and analyzing a putative DNA–protein complex is similar to that described above for protein–protein interactions. The comparison of the DNA-binding interface of a pair of proteins is conducted with the program fr-TM-align.[8] In this calculation, unlike for protein–protein interfaces, a larger heavy atom distance cutoff of 10 Å was employed to define a DNA-binding interface.

3. Results and discussion

In what follows, for systems 40–300 residues in length, we compare the structural properties of a representative set of PDB structures with a set of the same length compact, hydrogen bonded homopolypeptide structures and a set of quasi-spherical random structures packed at protein like densities

but lacking hydrogen bonded secondary structural elements. In addition, for each model system, we examine the presence or absence of geometric features necessary for intermolecular interactions with small molecule ligands, protein–protein interactions, and protein–DNA interactions.

3.1 Local structure quality of artificial and quasi-spherical random structures is acceptable

We begin by assessing the local properties of the polypeptide backbones using an analysis of φ and ψ dihedral angles. The Ramachandran maps for crystal, artificial and random structures are shown in Fig. 1. Most residues from crystal structures occupy regions in the Ramachandran plot that correspond to helical and extended secondary structures (Fig. 1A). The higher flexibility of artificial and particularly quasi-spherical random structures results in larger areas in the Ramachandran map being populated by amino acid residues (Fig. 1B and C, respectively). However, in both cases, the local backbone geometry is within an acceptable range so that the structures exhibit acceptable local stereochemical geometry. For quasi-spherical structures, this is an interesting result in that there is no restriction on the φ and ψ dihedral angles that are allowed when the chain is built; the only restriction is to construct a path of minimum length path whose bond lengths and distance of closest approach cannot be smaller than 3.8 Å.

We quantify the overlap between two maps by calculating the Pearson correlation coefficient (*CC*) for amino acid frequencies within 30° grid cells (shown in white in Fig. 1). The *CC* between artificial (quasi-spherical) random structures and crystal structures is 0.54 (0.51), which indicates adequate overlap. Interestingly, the Ramachandran map overlap calculated for individual amino acids correlates with their average flexibilities.[35] This is shown in Fig. 2 for the map overlap between crystal and quasi-spherical random structures. In quasi-spherical random structures, highly flexible residues occupy a larger fraction of the Ramachandran plot; this results in a smaller overlap with the crystal structures, which are strongly biased towards regular secondary structure elements. On the other hand, many residues, whose flexibility is relatively low, tend to cover similar regions as indicated by a high overlap ($CC > 0.5$). Thus, we conclude that in terms of individual residue geometric properties, both the artificial homopolypeptide based and quasi-spherical random structures are protein-like. We again point out that for the quasi-spherical random structures, this is an emergent feature of the calculation that was not built in, but was found when all atom models were built on the random traces within a sphere packed at protein like densities.

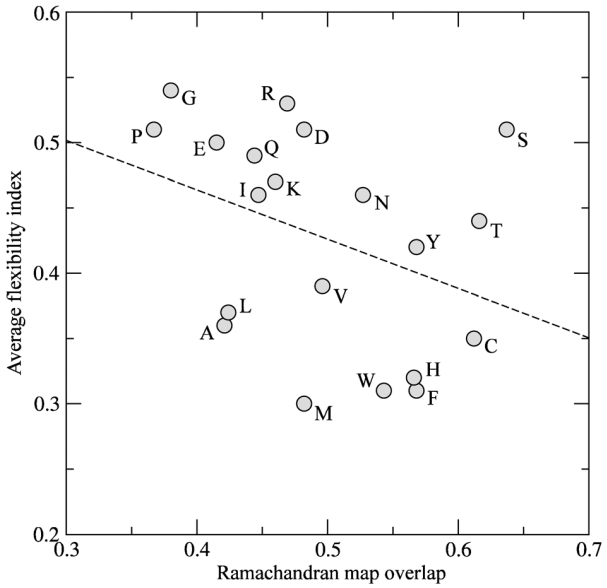

Fig. 2 For quasi-spherical and native protein structures, plot of the average flexibility index *versus* correlation coefficient of their Ramachandran maps, the "Ramachandaran map overlap".

3.2 Random structures lack regular secondary structure elements

Next, we calculate the overall regular secondary structure content across the sets of crystal, artificial and quasi-spherical random structures. Here, we use two secondary structure assignment procedures: a high-resolution, 7-state assignment by STRIDE[2] and a low-resolution, 3-state assignment according to the TASSER force field.[36] As shown in Table 2, according to STRIDE, crystal structures have >60% of their residues assigned to one of four regular structural elements: α-helix, π-helix, 3–10 helix or β-strand. For artificial structures, considerably fewer residues (29%) form secondary structural elements; this is caused by their significantly reduced β-structure content (<5%). Quasi-spherical random structures lack almost any secondary structure elements, with 99% of their residues assigned to either coil or turn conformations.

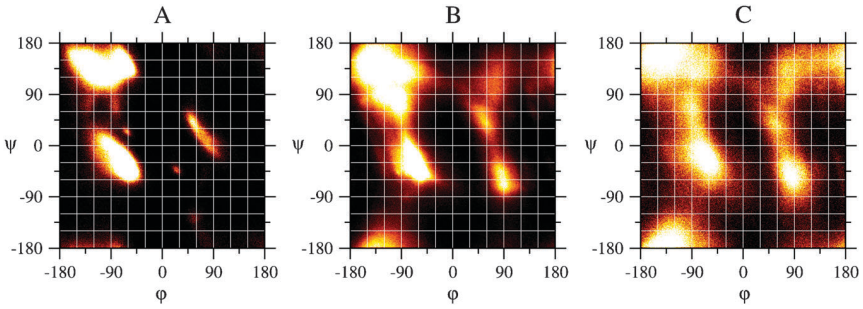

Fig. 1 Ramachandran maps calculated for different datasets: (A) crystal structures, (B) artificial structures and (C) quasi-spherical random structures.

Table 2 Secondary structure content as the percentage of residues assigned to each secondary structure class across the datasets

	Target structure		
Secondary structure	Crystal	Artificial	Random
STRIDE (high-resolution)			
3–10 Helix	3.5	0.5	0.2
α-Helix	33.4	23.4	0.013
Bridge	1.0	1.9	0.865
Coil	18.5	37.5	37.1
π-Helix	0.006	0.003	0.003
β-Strand	23.2	4.7	0.073
Turn	20.4	32.1	61.7
TASSER (low-resolution)			
α-Helix	29.8	22.5	1.8
β-Strand	23.1	26.5	11.4
Coil	47.1	51.0	86.8

However, these results are somewhat misleading in that the high-resolution secondary structure assignment is highly sensitive to the location of the explicit hydrogen bonds calculated from the positions of the backbone heavy atoms. In the case of artificial and quasi-spherical random structures, where heavy atom coordinates are rebuilt from the Cα trace of the chain, many hydrogen bonds are missing due to minor local structural distortions. This reduces the β-structure content as detected by STRIDE. Therefore, we also assess the results using a low-resolution model consistent with the TASSER force field. Here, the secondary structure assignment for artificial structures closely follows that for the crystal structures (Table 2), with the small difference of 3–7% per class. Consistent with the STRIDE assignment, quasi-spherical random structures lack secondary structure elements; ~87% of the residues form coil structures.

Secondary structure content can be explained by a detailed analysis of their hydrogen bond patterns. In Fig. 3, we show the number of hydrogen bonds per residue calculated for the crystal, artificial as well as quasi-spherical random structures, calculated using a high-resolution model consistent with STRIDE as well as a low-resolution model based on Cα packing preferences, consistent with the TASSER force field. If one uses the high-resolution assignment (Fig. 3A), then the average number of hydrogen bonds between backbone atoms drops from crystal (0.57) to artificial (0.25) to random structures (0.07); this correlates very well with the overall secondary structure content. Significantly fewer hydrogen bonds are formed between side chain (0.09, 0.02 and 0.06) as well as backbone/side chain (0.17, 0.08 and 0.17) atoms for crystal, artificial, and quasi-spherical random structures, respectively.

As indicated above, the minor local distortions present in the all-atom models of artificial structures reconstructed from their Cα coordinates are mainly responsible for the reduced number of backbone hydrogen bonds. Therefore, in Fig. 3B, we also assess the hydrogen bond pattern using a low-resolution assignment, which approximates the location of hydrogen bonds from Cα packing preferences rather than the explicit positions of the peptide bond atoms. Here, the number of hydrogen bonds per residue is 0.43, 0.35 and 0.02 for crystal, artificial and quasi-spherical random structures, respectively. Thus, the differences in secondary structure content track the differences in hydrogen bonding, with the artificial compact structures resembling crystal structures at low resolution, whereas the quasi-spherical structures are entirely devoid of regular secondary structure. We next examine the global structural and functional consequences of this difference.

3.3 Most random/artificial structures have statistically significant matches to real protein structures and *vice versa*

For artificial and quasi-spherical systems up to 250 residues in length, we generated structural alignments to the corresponding full set of PDB structures as well as the set of PDB structures up to 300 residues in length, PDB300 (see Methods). In the limit of weak structural similarity, for a small subset, there is the need to have template structures that are somewhat longer than the target in order to detect structurally similarity. Analogously, the PDB250 set (the subset of proteins in PDB300 whose length is $\leq$250 residues), were only aligned to the quasi-spherical300 and artificial300 set. We note that a structural alignment with a TM-score $\geq$ 0.4 is statistically significant and can be used to generate a useful length model.[9,37] In Fig. 4, the cumulative fraction of proteins whose TM-score $\geq$ abscissa is shown for quasi-spherical random, artificial structures to the corresponding full set of PDB templates as are PDB structures to the full PDB.

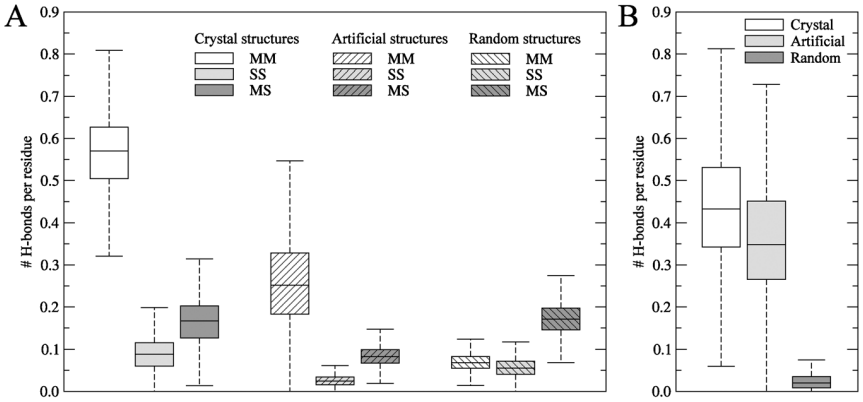

Fig. 3 Hydrogen bond patterns across the datasets. (A) high-resolution model (MM—main chain/main chain, SS—side chain/side chain, MS—main chain/side chain) and (B) low-resolution model consistent with the TASSER force field.

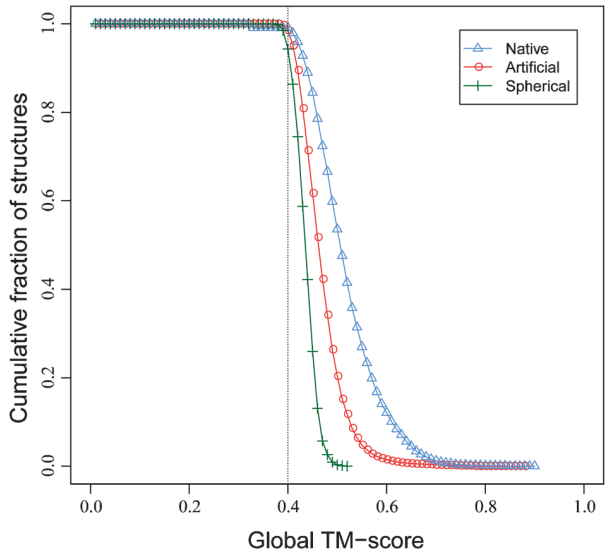

Fig. 4 Cumulative fraction of artificial250, quasi-spherical250 and PDB250 proteins that have a TM-score $\geq$ abscissa to the full PDB. For the comparison of PDB250 to the full PDB, templates whose sequence identity to the target is >3% are excluded.

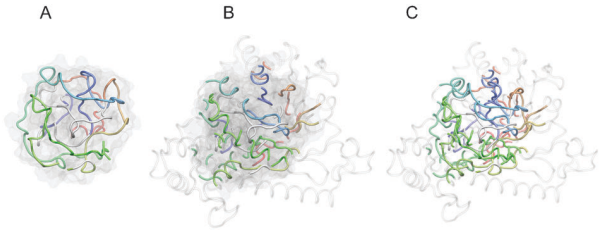

Fig. 5 Representative structural alignment of a 154 residue quasi-spherical target structure to the closest PDB structure. (A) The full-length quasi-spherical structure. The tube represents the backbone of the structure, and the surface representation shown in the transparent mode illustrates the shape of the structure. In the tube representation, the residues aligned with the native template are colored in the RGB scheme from the N- to C-terminal, whereas unaligned region is shown in white. (B) The closest native template (PDB code 1rw8, chain A) has a TM-score of 0.43, 124 residues aligned and an RMSD of 5.3 Å. The regions aligned to the spherical structure are shown in the same representations as in A using the same color scheme, whereas dimmed tubes represent unaligned regions. (C) The two structures were superimposed according to the optimal structural alignment. Aligned regions in both structures are shown in solid colors with the Cα atoms shown in spheres. Molecular images were obtained with the program VMD.[48]

As summarized in Table 3, what is remarkable is that 94% of the quasi-spherical structures have a structurally related protein (TM-score $\geq$ 0.4) in the PDB with a mean TM-score = 0.43, despite that fact that they are entirely lacking in secondary structure. This just reflects that we are looking at the spatial arrangement (as measured by the chain contour) of geometric objects. A typical example is shown Fig. 5A–C with a TM-score of 0.43 and 79% alignment coverage. The structure has significant local distortions and gaps but the global fold or topology is recovered.

There is a key difference between the structures considered here and those of the FJCs,[6] whose average TM-score to native is 0.30. These quasi-spherical structures have local chain dihedral angle preferences that introduce local chain rigidity that sufficiently restricts conformational space so that their overall folds roughly resemble real protein structures. Thus, we conclude that local chain rigidity and global compactness result in the space of structures sampled by real proteins.

On average, quasi-spherical random structures have 3.6 gaps for the best structural alignment of the target structure to the best template. We only consider gaps that are at least 4 residues in length (a threshold introduced to ignore local effects such as when a helix is aligned to a β-strand and vice versa[25]). The average gap length in the target protein is 7.1 residues. The lengths of the target and template gaps are defined as follows: If target residues i and $i + k$ are aligned to template residues m and $m + j$ respectively, with no intervening residues equivalenced between these target and the template residues, then the lengths of this gap in the target and template are k and j.

Conversely, for the PDB250 set, 71% of crystal structures have a significant structural alignment to the quasi-spherical random structures. Their behavior is similar to that when quasi-spherical structures are aligned to the PDB300 set. Roughly 25% of targets require significantly larger templates to generate a significant structural alignment from which a physical model can built. All cases have 76–78% of their target residues aligned.

Turning to the case of the artificial library of compact, hydrogen bonded homopolypeptide structures, 99% of the targets have a significant template match to the full PDB with

Table 3 Properties of global structural alignments for quasi-spherical random, artificial and real protein structures

Target	Template	Fraction of targets with TM-score $\geq$ 0.4	Average coverage[a,c]	Average TM-score	Average number of gaps per target[b,c]	Average gap length per target[c]	Average number of gaps per template[b,c]	Average gap length per template[b,c]
Quasi-spherical250	PDB[d]	0.94	0.77	0.43	3.6	7.1	11.7	22.5
Quasi-spherical250	PDB300	0.69	0.76	0.42	3.2	7.0	8.3	14.1
PDB250	Quasi-spherical300	0.71	0.78	0.42	2.3	6.4	10.1	13.8
Artificial250	PDB[d]	0.99	0.77	0.47	2.9	8.8	10.8	25.9
Artificial250	PDB300	0.77	0.77	0.45	2.4	8.5	6.9	15.6
PDB250	Artificial300	0.77	0.74	0.44	2.6	8.5	6.2	16.0
PDB250	PDB[d,e]	0.99	0.78	0.51	2.5	8.0	7.7	24.0
PDB250	PDB300[e]	00.90	0.75	0.46	2.5	8.5	5.2	15.2

[a] Fraction of residues in the target sequence that are part of the best structural alignment. [b] Only gaps whose lengths are >3 residues are considered where the gap length is defined in the text. [c] Only templates with a TM-score $\geq$ 0.4 are considered. [d] Structural alignments to the entire PDB library without chain length restrictions. [e] All template structures with a sequence identity >3% to the target are excluded.

a mean TM-score of 0.47, which is somewhat higher than the mean TM-score of the quasi-spherical structures to the full PDB. Since the local geometries now contain a significant number of regular secondary structural elements, the average number of gaps/target protein is reduced, but the average length of a gap in the target proteins increases from 7.1 to 8.8 residues. Similar coverage effects as for the quasi-spherical random case are seen when structural alignments are restricted to the PDB300 set. Thus, our conclusion about the likely completeness of the PDB[6] is now increased from proteins that are 200 residues to 250 residues in length. Indeed, the PDB250 structural alignments to the artificial300 structure library (with proteins up to 300 residues in length) have almost identical behavior as the artificial250 library does to PDB300. The differences in the behavior of PDB250 alignments to artificial300 and quasi-spherical300 reflect the complex interplay of the number of targets with good alignments which is lower for quasi-spherical proteins (71% *vs.* 77%), the surface to volume ratio which is lower in quasi-spherical proteins, see Fig. 6 below, (hence there are more internal local points with which to align than in artificial structures), and the local geometric fidelity of secondary structure which is higher in the artificial300 structures, as shown in Fig. 3.

Finally, we consider structure alignments of real protein structures in PDB250 to other PDB structures, subject to the constraint that the sequence identity between the target-template pair must be $\leq 3\%$. This is done to remove obvious (and not so obvious) evolutionary similarities between the pairs of aligned proteins. Not surprisingly, the space of real crystal structures is somewhat denser, with 90% of PDB250 proteins finding a structural similar partner among members of PDB300, but their coverage, 0.75, is comparable to that of the artificial library, and their average TM-score is 0.46. For the best target-template structural alignment, the average number of gaps/target is 2.5, with an average gap length of 8.5 residues. These numbers are similar to the results of the artificial structures. On the basis of these results, we conclude that in the limit of little, if any, detectible evolutionary relationship of target and templates, the structural space of real single domain proteins and artificial structures are very similar.

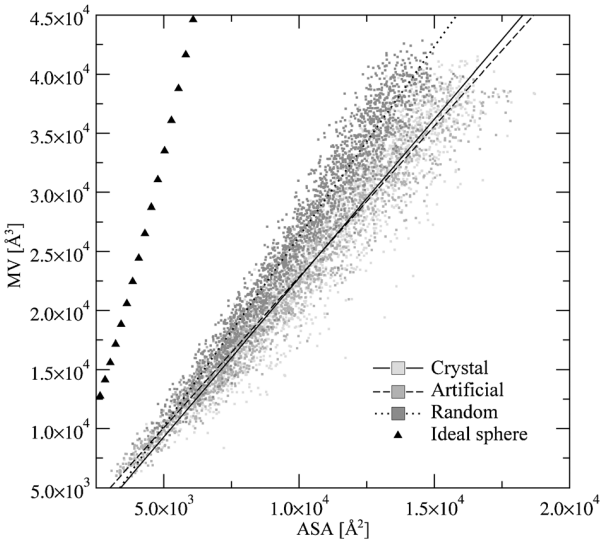

Fig. 6 Accessible surface area (ASA) *vs.* molecular volume (MV) for the dataset proteins compared to ideal spheres (triangles).

Taking all the above results into consideration, we find that the space of protein structures is strongly dictated by the requirement of dense packing of locally semi stiff chains and much to our surprise does not require backbone hydrogen bonding. In other words, it is an inherent feature of densely packed, quasi-spherical objects comprised of residues with single residue, protein like local geometries. Once secondary structure is allowed, the local geometric fidelity to real structures improves as does the global structural similarity, but this effect is not dramatic.

3.4 Random/artificial structures have similar internal packing as crystal structures

The relationship of accessible surface area (ASA) and molecular volume (MV) calculated across the set of artificial structures closely follows that for the crystal structures (Fig. 6). By design, since the quasi-spherical random structures are built to be close to spheres, they occupy a slightly higher volume than the crystal and artificial structures at the same surface area, with the ASA/MV relation shifted toward that of ideal spheres (Fig. 6). In other words, at the same molecular volume, random structures have less solvent-accessible surface, which may reduce their functional capabilities, since molecular functions typically take place on a protein's surface. Nevertheless, they still have a larger solvent accessible surface area than the corresponding perfect sphere. These deviations from a perfect sphere are caused by the requirement that the local geometry be protein like, as was demonstrated in Fig. 1 and 2.

Hydrogen bonding, which allows for the creation of secondary structure elements, also shapes the global structure of a polypeptide chain. This is shown in Fig. 7, where we compare the length of mass-weighted principal axes calculated across crystal, artificial and quasi-spherical random sets of protein structures. The overall global shapes of crystal and artificial structures are comparable, with a principal axes $X:Y:Z$ ratio of $1:0.75:0.61$ and $1:0.82:0.70$, respectively. By design, the quasi-spherical random structures are highly spherical, with an $X:Y:Z$ ratio of $1:0.96:0.92$.

Interestingly, as indicated by Table 4, artificial and particularly quasi-spherical random structures contain more knots than crystal structures, as on an intermediate distance scale quasi-spherical structure are more flexible. The KNOT algorithm characterizes backbone knots in proteins by the number of residues that must be removed from each end to eliminate the knot (see ref. 47 for details). Considering a criterion that at least 10 residues that must be removed to abolish the knot, then 2.5% of random structures contain knots; this is significantly more than the fraction of knotted crystal (0.18%) as well as in artificial structures (0.53%).

3.5 Geometrical criteria for molecular function

Protein function often emerges from the capability to bind other molecular species present in a cell. Binding events take place at specific locations on the protein's surface such as binding pockets and interfaces, whose geometrical features are

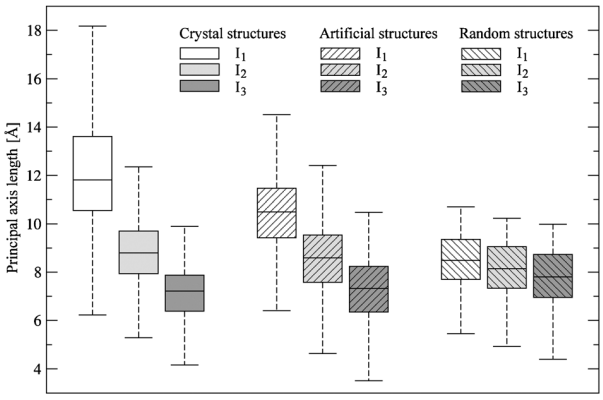

Fig. 7 Distribution of the lengths of mass-weighted principal axes for the dataset proteins.

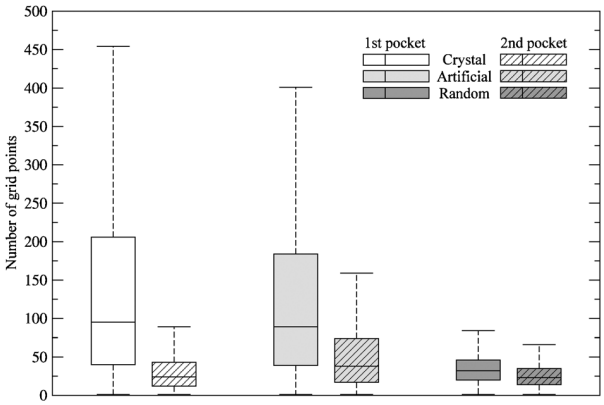

Fig. 8 Distribution of the number of grid points assigned to the largest (1st) and the second largest (2nd) pockets detected in the sets of crystal, artificial and quasi-spherical random structures.

Table 4 Percentage of knotted proteins for the sets of protein structures

	Target structure		
Knot length[a]	Crystal	Artificial	Random
≥ 1	0.65	3.10	9.27
≥ 10	0.18	0.53	2.52

[a] Number of knotted residues reported by KNOT.[47]

often well-defined and different from non-binding surface patches. Below, using purely geometrical criteria, we compare crystal structures to artificial and quasi-spherical random structures to examine their capability to bind small organic compounds, other proteins and DNA.

3.5.1 Quasi-spherical random structures have much smaller pockets. In proteins, rigid secondary structure elements interact with each other to create a compact object. Their spatial arrangements typically result in significant irregularities in the surface geometry and the formation of pockets and cavities. In Fig. 8, we measure the average number of grid points assigned by LIGSITE[38,39] to the largest as well as the second largest pockets present in the set of crystal, artificial and quasi-spherical random structures. Interestingly, the sizes of the largest (2nd largest) pockets in crystal and artificial structures are very similar: 95 (24) and 89 (38) grid points, respectively. Quasi-spherical random structures, which have practically no secondary structure elements, form very small cavities on their surfaces. The average size of the largest (2nd largest) pocket is only 32 (23). Since the binding of small organic compounds requires a specific microenvironment, which is typically formed by a concave protein surface, quasi-spherical random structures have significantly reduced binding capabilities. On the other hand, hydrogen-bonded artificial structures, that contain rigid secondary structure elements, fully satisfy the geometrical criteria for binding of small organic compounds. Thus, the presence of protein-like cavities that are necessary for small molecule ligand binding is an inherent feature of the packing of regular secondary structural elements.

3.5.2 Quasi-spherical random structures lack geometrically suitable interfaces needed for protein–protein interactions. Next, we examine whether the quasi-spherical or artificial structures can form native-like protein–protein complexes as assessed by the geometric similarity to the structures of real (native) protein–protein complexes. We first investigate how similar the global structures of quasi-spherical or artificial proteins are compared to the structures of native proteins taken from a representative set of 1690 nonredundant native dimeric protein complexes.[30,32] In this calculation, we randomly selected 1988 spherical and artificial structures, and performed structural comparison to 1690 monomeric native structures by arbitrarily taking one monomer from each native complex. The TM-score distributions of these all-against-all comparisons are shown in Fig. 9. Consistent with the results of section 3.3, only a tiny fraction, 0.38% of all pairwise comparisons has a significant TM-score > 0.4 for the quasi-spherical structures. The fraction is over ten times lower than that (4.8%) for artificial structures, and about 36 times lower than that (14%) for the set of all native monomers against each other. In the later case, in order to remove homologs, we excluded any hit if the sequence identity is higher than 3% in the aligned region. For each target, the mean numbers of significant hits (TM-score > 0.4) are 13/160/369 for quasi-spherical/artificial/native structures.

We then ask the question: can quasi-spherical or artificial structures provide a surface patch with structural features similar to native protein–protein interfaces? We first conducted a planarity analysis, whereby we search for the most planar surface patch within a solvent accessible area of 1000 Å^2 (a typical interface area per protein in a protein complex[40]) for a set of 1051 pairs of quasi-spherical random/artificial structures. The selection of these pairs is random, except that each pair of quasi-spherical random and artificial structures has the same number of residues in order to eliminate potential size-effects. For comparison, we also calculated the most planar patch formed by interfacial residues from the 1690 native protein–protein interfaces. The planarity of a surface patch is defined as the RMSD of the C_α atoms of the best-fit plane through the patch. As shown in Fig. 10A, the distribution of the minimal planarity values of the artificial structures has a much greater overlap with the distribution of native interfacial patches than that of the spherical structures. The mean planarity of the native/artificial structures is 1.23/1.35 Å,

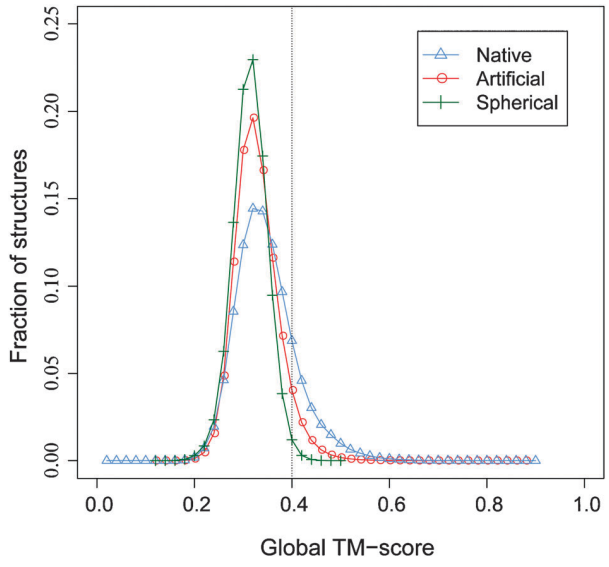

Fig. 9 Global structural similarity among quasi-spherical, artificial, and native protein structures taken from protein–protein complexes. Histograms represent the distributions of global TM-scores calculated from an all-against-all structural comparison of quasi-spherical *vs.* native, artificial *vs.* native, and native *vs.* native structures. The TM-score is normalized by the length of the shorter structure in each pair of compared structures. A vertical dashed line is located at a significant TM-score threshold of 0.4.

compared to 1.61 Å for quasi-spherical structures. Statistically, these three sets of structures are significantly different in their distributions of planarity (T-test $< 2.2 \times 10^{-16}$ between each other), and it is clear that the native and the artificial structures are more planar than the quasi-spherical structures. We further calculate these patches' curvature, which is defined by $1/r$, where r is the radius of the best-fit sphere. A positive curvature value indicates a convex patch, whereas a negative one indicates a concave patch. Consistent with the planarity analysis, as shown in Fig. 10 B, on average, about 68/47% of the native/artificial patches have a very low curvature, with absolute values <0.01 Å^{-1}. In contrast, 32% of quasi-spherical structures are found at the same curvature threshold. These results indicate quasi-spherical structures are less likely than the artificial structures to have a surface patch suitable for protein–protein interactions. Interestingly, a small percent

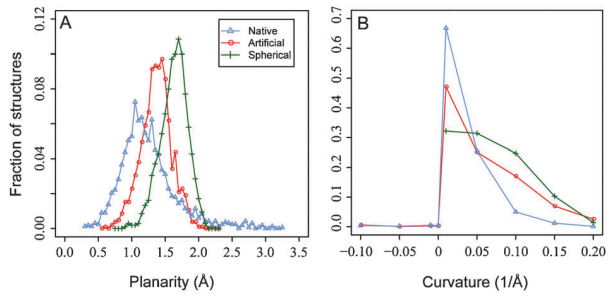

Fig. 10 Statistics of the most planar surface patches on quasi-spherical and artificial structures and in real protein–protein interfacial structures. (A) Histograms of the minimal planarity of the patch in each structure. (B) Curvature of the same patches as in (A).

(1.4/1.2%) of native/artificial surface patches are concave, but no spherical patches have a concave shape, consistent with the previous analysis that the quasi-spherical structured lack pockets of significant size.

Using significant (TM-score > 0.4) global structural alignments between quasi-spherical/artificial structures and native protein monomer structures taken from protein–protein complexes, we further built putative complexes by superimposing individual quasi-spherical random structures onto their corresponding aligned monomers from the native templates. We consider all-against-all alignments of 1988 quasi-spherical structures, and 30 000 randomly selected pairs of artificial structures. Each structure has 80 possible evolved sequences.[24] After removing structures with steric clashes, we compare the remaining putative protein–protein interfaces against the real, native protein–protein interface of the corresponding dimeric template; one example is illustrated in Fig. 11.

From the statistical analysis shown in Fig. 12, one can immediate recognize that putative interfaces formed by quasi-spherical structures generally lack structural similarity to the native protein–protein interfaces. On searching about 12.6 trillion quasi-spherical random structure pairs, none have an interfacial TM-score (iTM-score) > 0.4 or an IS-score > 0.3. Moreover, only 8/1 spherical complex structures have a significant iTM-score/IS-score at $P < 1 \times 10^{-3}$. The former metric considers the geometric similarity of backbone Cα atoms, while the latter evaluates interfacial contact similarity in addition to the geometric similarity. By comparison, at the same P-value thresholds, 53 662/51 096 pairs were found among only 192 million artificial structure pairs. The chance of finding a putative, structurally native-like interface with a significant iTM-score at $P = 1 \times 10^{-3}$ is about 2.8×10^{-4}, about four million times higher than that found in spherical structures. Consistent with these results, the iTM-score and IS-scores of the interface of a typical quasi-spherical structure dimer structure to 1em8C shown in Fig. 11 are 0.28 ($P = 0.04$) and 0.18 ($P = 0.09$) respectively.

3.5.3 Quasi-spherical random structures lack geometrically suitable interfaces needed for DNA binding. A similar procedure was followed to examine whether one can find a quasi-spherical

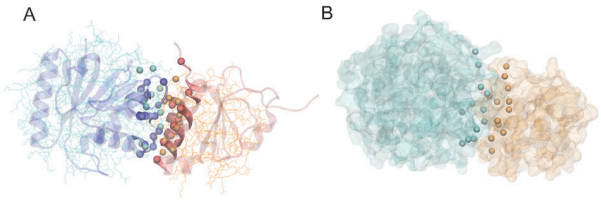

Fig. 11 Example of a putative protein–protein complex. The complex is built by superimposing two quasi-spherical random structures (cyan and orange) onto a native dimeric template (PDB code 1em8, chain C and D, colored in blue and red). (A) The interface alignment according to iAlign.[32] The spherical structure is shown in a line representation, and the native template is shown in a cartoon representation. The Cα atoms of aligned interfacial residues are shown as Van der Waals spheres. (B) Surface representation of the putative complex of the pair of quasi-spherical random at the same orientation as in (A). The iTM- and IS-scores to 1em8C are 0.28($P = 0.04$) and 0.18 ($P = 0.09$) respectively.

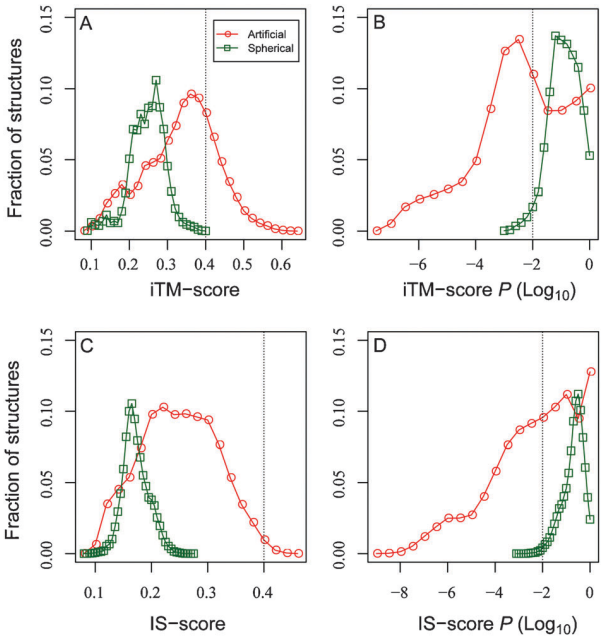

Fig. 12 Interfacial similarity of putative protein–protein interfaces built with quasi-spherical or artificial structures compared to native protein–protein interfaces. Histograms represent the distributions of (A) interfacial TM-scores and (B) their *P*-values, and of (C) IS-scores and (D) their *P*-values. Each score is normalized by the length of the native template.

random structure whose surface is complementary to DNA. The same data set of 1988 quasi-spherical/artificial structures used above was examined. We first conducted an all-against-all structural comparison between individual quasi-spherical/artificial structures and a representative set of 399 native DNA-binding protein domains from 1,350 experimentally determined protein/DNA complexes.[31] As expected, only a small fraction (0.66%) of all pairwise comparisons have a significant TM-score > 0.4 (see Fig. 13A), compared to 12% of all comparisons of artificial structures to native structures, and 26% of native against native structures where the maximum allowed sequence identity is 3%.

Analysis of the most planar surface patch from native DNA-binding interfaces suggests that the distribution of DNA-binding patch curvature is more diverse than those of protein–protein interfaces (see Fig. 10). As shown in Fig. 13B, over 50% of native DNA-binding surface patches have a rather planar shape with an absolute curvature of <0.01 Å^{-1}. Interestingly, about 7% of native patches are concave, resulting from wrapping of the protein around DNA. Obviously, these planar or concave patches are difficult, if not impossible, to find in quasi-spherical structures that lack regular secondary structure. Moreover, about 52% of native DNA-binding residues have either an α-helix or β-strand secondary structure; such geometries are absent in the quasi-spherical random structures.

We further consider those spherical/artificial structures aligned by global structure comparison to more than 50% of DNA-binding protein residues in their corresponding native structure. For each structure, we built 80 all-atom structural models and superimposed them onto the native protein/DNA complex according to the optimal structural alignment. After discarding those with significant steric clashes, we obtain 7289 quasi-spherical/native and 449 992 artificial/native pairs. The putative DNA-binding interfaces are structurally compared to their corresponding DNA-binding interfaces from native proteins, as shown in Fig. 13B and C. Similar to the results of protein–protein interface comparison, the putative DNA-binding interface regions of quasi-spherical random structures generally lack similarity to their corresponding native DNA-binding interfaces. Only 12 putative DNA-binding interfaces from the quasi-spherical random structures have a statistically significant interfacial TM-score > 0.4, compared to 306 649 putative DNA-binding interfaces from the artificial structures under the same criterion. Therefore, it is much less likely to find a surface suitable for DNA-binding in the quasi-spherical structures than in the artificial structures.

4. Conclusions

In this study, we have examined the set of conditions necessary and sufficient to generate the local and global structural properties of single domain proteins as well as typical interaction sites with small molecule ligands, other proteins and DNA. With respect to the distribution of backbone dihedral angles, both artificial and quasi-spherical random protein structures have similar local geometries as in real proteins. This local rigidity is sufficient to restrict the space of structures so that almost all quasi-spherical proteins have a related structure in the PDB. It is interesting to note that the shortest trace for Cα atoms, which are randomly distributed within a sphere, creates a polypeptide chain whose local stereochemical quality is protein-like. However, the lack of backbone hydrogen bonding in such quasi-spherical random proteins results in the absence of regular secondary structural elements such as helices and β-strands. Lacking these regular secondary structural elements effectively eliminates a number of geometric features that are essential for protein function. Since the proteins are better packed than native structures, they have cavities that are too small for small molecule ligand binding. These proteins also lack the planar interfaces needed for protein–protein interactions and the presence of secondary structures and flat interfaces as well as concave surfaces required for DNA binding. In other words, because they lack the requisite geometric features, they are unable to engage in molecular functions typical of proteins.

In contrast, the artificial set of proteins have very similar secondary structures as real native proteins with a comparable number of hydrogen bonds when assessed by hydrogen bonding schemes comparable to their backbone resolution. The packing of the resulting regular secondary structural elements increases the structural fidelity of these artificially generated structures to real proteins and generates a similar global ellipsoidal shape. Even more interesting is that the packing of secondary structural elements yields surface cavities that closely resemble those in real proteins. Moreover, the faces of the regular secondary structural elements yield surfaces that resemble protein–protein and protein–DNA interfaces. Thus, hydrogen bonding by generating secondary structural elements that when driven by hydrophobic interactions

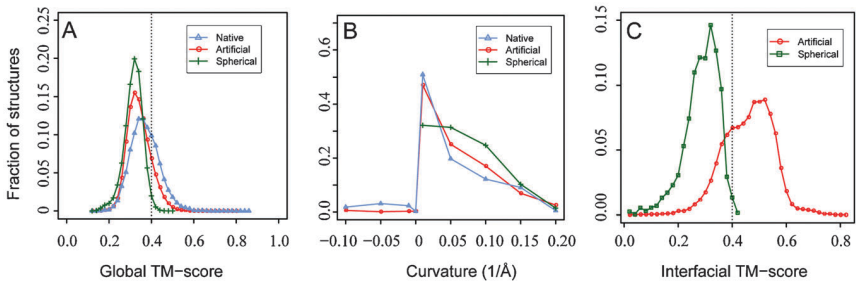

Fig. 13 Global and interfacial structural similarity among quasi-spherical, artificial and native protein structures taken from DNA–protein complexes. (A) Histograms represent the distributions of global TM-scores calculated from all-against-all structural comparison of spherical *vs.* native, artificial *vs.* native, and native *vs.* native DNA-binding domain structures. The TM-score is normalized by the length of the shorter structure. (B) The curvature of the most planar, 1000 $Å^2$ surface patches from native DNA-binding interfaces, compared to the most planar surface patches from artificial or quasi-spherical structures. (C) Histograms represent the distributions of the interfacial TM-scores of putative DNA-binding interfaces built with quasi-spherical or artificial structures compared to native DNA-binding protein interfaces.

to form compact structures gives rise to all the geometric features required for intermolecular interactions as typified in native protein structures.

In other words, the plethora of geometric features seen in native proteins, the likely completeness of structural space, their global shape, and their interaction surfaces, can be rationalized by the requirements of local chain stiffness, main chain hydrogen bonding and compaction without invoking evolution. Evolution undoubtedly takes advantage of these inherent protein features by selecting for sequences with stable native structures and favorable interaction free energies. But, the background probability on which evolution selects is based entirely on protein physics and is the result of very fundamental physical/geometric properties of proteins. In future work, we shall examine if the thermodynamic stability of native like structures is sufficient to give rise to sequences with the capacity to bind small molecule ligands, proteins and DNA or if explicit evolutionary selection for function is required.

Overall, we conclude that while densely packed, quasi-spherical random structures have a similar local rigidity and global fold as real proteins (but the number of such global matches per structure is less than those of artificial proteins with regular secondary structure), they lack an essential element needed to reproduce the properties of real proteins, namely backbone hydrogen bonding. In essence, it is hydrogen bonding that underlies the capacity of proteins to perform molecular function. With it, there are deviations from a perfect sphere that generate the cavities and the interfacial surfaces needed for intermolecular interactions and molecular function. This is perhaps why nature did not employ a completely spherical protein devoid of helices and strands during the course of evolution.

Acknowledgements

This research was supported in part by NIH grant GM-48835 of the Division of General Medical Sciences of the National Institutes of Health.

References

1 I. K. McDonald and J. M. Thornton, *J. Mol. Biol.*, 1994, **238**, 777–793.
2 D. Frishman and P. Argos, *Proteins*, 1995, **23**, 566–579.
3 R. H. Austin, J. Karohl and T. M. Jovin, *Biochemistry*, 1983, **22**, 3082–3090.
4 U. A. Bommer, G. Lutsch, J. Behlke, J. Stahl, N. Nesytova, A. Henske and H. Bielka, *Eur. J. Biochem.*, 1988, **172**, 653–662.
5 J. E. Brunet, V. Vargas, E. Gratton and D. M. Jameson, *Biophys. J.*, 1994, **66**, 446–453.
6 Y. Zhang, I. A. Hubner, A. K. Arakaki, E. Shakhnovich and J. Skolnick, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 2605–2610.
7 P. W. Rose, B. Beran, C. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, J. D. Westbrook, J. Young, B. Yukich, C. Zardecki, H. M. Berman and P. E. Bourne, *Nucleic Acids Res.*, 2010, **39**, D392–401.
8 S. B. Pandit and J. Skolnick, *BMC Bioinformatics*, 2008, **9**, 531.
9 Y. Zhang and J. Skolnick, *Nucleic Acids Res.*, 2005, **33**, 2302–2309.
10 E. Moreno and K. Leon, *Proteins*, 2002, **47**, 1–13.
11 N. D. Gold and R. M. Jackson, *J. Chem. Inf. Model.*, 2006, **46**, 736–742.
12 N. D. Gold and R. M. Jackson, *Nucleic Acids Res.*, 2006, **34**, D231–234.
13 A. S. Reddy, H. S. Amarnath, R. S. Bapi, G. M. Sastry and G. N. Sastry, *Comput. Biol. Chem.*, 2008, **32**, 387–390.
14 M. Brylinski and J. Skolnick, *Proteins*, 2010.
15 Q. H. Gibson and R. L. Nagel, *J. Biol. Chem.*, 1974, **249**, 7255–7259.
16 E. Horjales, M. M. Altamirano, M. L. Calcagno, R. C. Garratt and G. Oliva, *Structure*, 1999, **7**, 527–537.
17 J. Y. Liang, Y. Zhang, S. Huang and W. N. Lipscomb, *Proc. Natl. Acad. Sci. U. S. A.*, 1993, **90**, 2132–2136.
18 B. Ma and R. Nussinov, *Curr. Opin. Chem. Biol.*, 2010, **14**, 652–659.
19 W. R. Cannon, S. F. Singleton and S. J. Benkovic, *Nat. Struct. Biol.*, 1996, **3**, 821–833.
20 M. Gao and J. Skolnick, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 22517–22522.
21 K. Nadassy, S. J. Wodak and J. Janin, *Biochemistry*, 1999, **38**, 1999–2017.
22 N. M. Luscombe, R. A. Laskowski, D. R. Westhead, D. Milburn, S. Jones, M. Karmirantzou and J. M. Thornton, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 1998, **54**, 1132–1138.
23 J. Janin, F. Rodier, P. Chakrabarti and R. P. Bahadur, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2006, **63**, 1–8.
24 S. Y. Lee and J. Skolnick, *Biophys. J.*, 2010, **99**, 3066–3075.
25 J. Skolnick, A. K. Arakaki, S. Y. Lee and M. Brylinski, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 15690–15695.
26 P. Rotkiewicz and J. Skolnick, *J. Comput. Chem.*, 2008, **29**, 1460–1465.
27 A. D. MacKerell, D. Bashford, Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus, *J. Phys. Chem. B*, 1998, **102**, 3586–3616.
28 Z. Xiang and B. Honig, *J. Mol. Biol.*, 2001, **311**, 421–430.

29 D. L. Applegate, R. E. Bixby, V. Chvatal, W. Cook, D. G. Espinoza, M. Goycoolea and K. Helsgaun, *Operations Res. Lett*., 2009, **37**, 11–15.
30 H. L. Chen and J. Skolnick, *Biophys. J*., 2008, **94**, 918–928.
31 M. Gao and J. Skolnick, *PLoS Comput. Biol*., 2009, **5**, e1000567.
32 M. Gao and J. Skolnick, *Bioinformatics*, 2010, **26**, 2259–2265.
33 S. J. Hubbard and J. M. Thornton, University College London, 1993.
34 R. A. Laskowski, *J. Mol. Graphics*, 1995, **13**, 323–330, 307–328.
35 R. Bhaskaran and P. K. Ponnuswamy, *Int. J. Pept. Protein Res*., 1998, **32**, 242–255.
36 Y. Zhang, A. Kolinski and J. Skolnick, *Biophys. J*., 2003, **85**, 1145–1164.
37 Y. Zhang and J. Skolnick, *Proc. Natl. Acad. Sci. U. S. A*., 2005, **102**, 1029–1034.
38 M. Hendlich, F. Rippmann and G. Barnickel, *J. Mol. Graphics Modell*., 1997, **15**, 359–363, 389.
39 B. Huang and M. Schroeder, *BMC Struct. Biol*., 2006, **6**, 19.
40 J. Janin, R. P. Bahadur and P. Chakrabarti, *Q. Rev. Biophys*., 2008, **41**, 133–180.
41 M. F. Sanner, A. J. Olson and J. C. Spehner, *Biopolymers*, 1996, **38**, 305–320.
42 L. Cavallo, J. Kleinjung and F. Fraternali, *Nucleic Acids Res*., 2003, **31**, 3364–3366.
43 M. F. Browner, E. B. Fauman and R. J. Fletterick, *Biochemistry*, 1992, **31**, 11297–11304.
44 M. Brylinski and J. Skolnick, *Proteins*, 2007, **70**, 363–377.
45 G. N. Ramachandran, C. Ramakrishnan and V. Sasisekharan, *J. Mol. Biol*., 1963, **7**, 95–99.
46 Y. Zhang and J. Skolnick, *Biophys. J*., 2004, **87**, 2647–2655.
47 W. R. Taylor, *Nature*, 2000, **406**, 916–919.
48 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graphics*, 1996, **14**, 33–38, 27–38.