Boston University

College of Arts and Sciences
Computer Science Department
111 Cummington St., Boston, MA 02155

Loredana Lo Conte
loredana@bu.edu

617/353-4726
fax 617/353-6457

Visible Volume: a Robust Measure for Protein Structure Characterization

Loredana Lo Conte (corresponding author)
loredana@cs.bu.edu, Computer Science Dept. and Molecular Engineering Research Lab, Boston University
and
Temple Smith
Biomolecular Engineering Research Center, Boston University

BU-CS-97-003
Available at *http://www.cs.bu.edu/techreports*

To appear in Journal of Molecular Biology

March 20, 1997
Last revision July 10, 1997

We propose a new characterization of protein structure based on the natural tetrahedral geometry of the β carbon and a new geometric measure of structural similarity, called *visible volume*. In our model, the side-chains are replaced by an ideal tetrahedron, the orientation of which is fixed with respect to the backbone and corresponds to the preferred rotamer directions. Visible volume is a measure of the non-occluded empty space surrounding each residue position after the side-chains have been removed. It is a robust, parameter-free, locally-computed quantity that accounts for many of the spatial constraints that are of relevance to the corresponding position in the native structure. When computing visible volume, we ignore the nature of both the residue observed at each site and the ones surrounding it. We focus instead on the space that, together, these residues could occupy. By doing so, we are able to quantify a new kind of invariance beyond the apparent variations in protein families, namely, the conservation of the physical space available at structurally equivalent positions for side-chain packing. Corresponding positions in native structures are likely to be of interest in protein structure prediction, protein design, and homology modeling.

Visible volume is related to the degree of exposure of a residue position and to the actual rotamers in native proteins. In this article, we discuss the properties of this new measure, namely, its robustness with respect to both crystallographic uncertainties and naturally occurring variations in atomic coordinates, and the remarkable fact that it is essentially independent of the choice of the parameters used in calculating it. We also show how visible volume can be used to align protein structures, to identify structurally equivalent positions that are conserved in a family of proteins, and to single out positions in a protein that are likely to be of biological interest. These properties qualify visible volume as a powerful tool in a variety of applications, from the detailed analysis of protein structure to homology modeling, protein structural alignment, and the definition of better scoring functions for threading purposes.

Key words: visible volume, threading, protein structural alignment, homology modeling, protein design.

Even though the canonical description of a protein is a full list of its atomic coordinates, this is neither a compact representation nor a useful one for recognizing common structural features. From the very beginning, proteins have been described by a number of higher-order properties, such as the set of secondary structure elements of which they are composed. At an even higher level, they have been classified into basic folds, identified as recognizable 3-D packings of the two major types of secondary structure. Richardson (1981) has constructed a taxonomy of protein structures using this heuristic description.

Protein structures have also been characterized using geometric descriptors which are directly related to protein conformation: these include dihedral angles (Ramachandran & Sasisekharan, 1968), accessible surface area (Lee & Richards, 1971), and residue volumes (Richards, 1974). These and other geometric features, mainly inter-atomic distances, have been used in a variety of ways, ranging from protein structural classification to the definition of structural profiles and scoring functions for the inverse folding problem (see Holm & Sander, 1994; Orengo, 1994; Bowie *at al.*, 1991; Bowie & Eisenberg, 1993; Wodak & Rooman, 1993 for reviews). Residue volumes have been extensively used to study volume changes in families of proteins (Lesk & Chothia, 1980; Ptitsyn & Volkenstein, 1986; Gerstein *et al.*, 1994; Kapp *et al.*, 1995) and packing properties of proteins (Richards, 1974; Chothia, 1975; Harpaz *et al.*, 1994; Chothia & Gerstein, 1997). A different approach, described in Pattabiraman *et al.* (1995), quantifies protein packing on the basis of the fraction of atomic surface that is occluded by neighboring atoms.

Our approach naturally falls within this line of research, namely, the identification of geometric descriptors that are useful for our understanding of protein structure. We propose a new geometric representation for proteins, based on backbone atoms and β carbons only, and a new measure, *visible volume*, defined as the volume of the empty space surrounding each residue position and in the line of sight of the corresponding β carbon (after the side-chains have been removed). A related geometric construction appears in Gregoret & Cohen, 1990, in which spheres of different radii are substituted for side-chains and dilated until all the space surrounding a residue position is filled.

Visible volume is a robust, parameter-free, locally-computed quantity that accounts for many of the relevant spatial constraints. It is related to amino acid solvent exposure. However, being side-chain independent, it captures a notion of "exposability" rather than of exposure by identifying positions in the sequence that could be buried or exposed, instead of assigning a degree of exposure to the actual amino acids. This distinction is important for applications in which avoiding sequence memory is both meaningful and desirable, as in statistics for threading purposes or in the definition of structural profiles. We discuss the properties of visible volume and the kind of structural information that it encodes, compared to other related geometric descriptors. We also show its effectiveness in representing structural motifs characteristic of a protein family and in singling out regions of biological interest. Visible volume quantifies one of the most important aspects of protein structure, the physical space available for side-chain packing. It has the unique property of identifying local units that are conserved within a protein family by measuring how much room is available for different combinations of side-chains to fit in. Corresponding positions in native structures that are invariant with respect to this quantity are likely to be of interest in protein structure prediction, protein design, and homology modeling.

In what follows, we discuss the properties of this new measure, namely, its robustness with respect to both crystallographic errors and naturally occurring variations in atomic coordinates, and the remarkable fact that it is essentially independent of the choice of the parameters used in calculating it. We also show how visible volume can be used to align protein structures and to identify structurally equivalent positions that are conserved in a family of proteins. Finally, we discuss its application to the definition of environmental states in the threading approach to the inverse folding problem in proteins.

First, let us briefly introduce our geometric representation and some of its properties. All our geometric descriptors are locally defined and are based on the natural tetrahedral geometry of sp^3 carbon atoms. From any given protein with known 3-D structure, we derive a model using non-hydrogen backbone atoms and β carbons only. The rationale for this choice is twofold: (a) the β carbon is the natural point of view for defining a local environment for each residue position, and (b) it is well known that proteins can have similar folds even in the absence of significant sequence similarity. Therefore, it makes sense to ignore the side-chains if the question is what *could* be found (or what is invariant) in a given position rather than what is actually observed.

In our model, the side-chains are replaced by a tetrahedron centered at the β carbon. The planes originating at the β carbon split space into four solid angles, conventionally called window 1, 2, 3, and 4. The orientation of the tetrahedron is fixed with respect to the backbone (Fig. 1).

The bottom window does not contribute any relevant information to the description of the residue's structural environment, since it always points towards the residue's backbone. But the other three windows correspond to the preferred rotamer directions. Discretized side-chain rotamers can be computed by assigning to a residue position the number of the window through which the

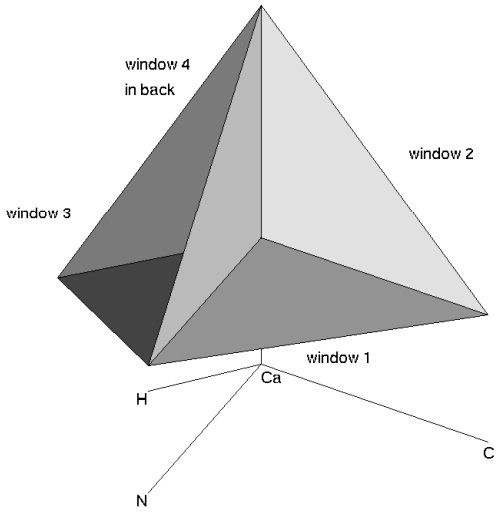

Figure 1: Tetrahedral Geometry. The side chain of each amino acid residue is replaced by a tetrahedron centered at the β carbon. When the β carbon is missing, as in glycine, we imagine it to be at the position it should have been, given the backbone coordinates and the tetrahedral geometry of sp^3 carbon atoms. The β carbon is the origin of a local coordinate system: the C_β–C_α vector determines the direction of the z-axis; the x-axis is defined by the projection of the bisector of the $\widehat{NC_\alpha C}$ angle onto the plane perpendicular to the z-axis; the y-axis is orthogonal to the other two. The β carbon and any two of the tetrahedron vertices define six planes; these planes split space into four solid angles, conventionally called window 1 (bottom view), window 2 (front view), window 3, and window 4 (going clockwise).

atom next to the β carbon extends. In general, this window representation is a suitable framework for a local definition of structural features close to the underlying physico-chemical reality and a good compromise if one wants to avoid going down to atomic detail. Window 2, for example, will almost always be unoccupied in an α helix, since it points towards the helix axis. Indeed, when we computed the likelihood of seeing the side-chains projecting out of each window for a set of all-α proteins, the only positive values for window 2 were those corresponding to SER, THR, and PRO (data not shown). PRO can never point out of window 4, because of its geometry; windows 2 and 3 correspond to its *cis* and *trans* conformations respectively. As for SER and THR, it is well known that their hydroxyl group can form hydrogen bonds back to the backbone (Kendrew, 1962); in order to do so, it must point out of window 2, and our representation captures this kind of information. SER and THR are also often observed at the helix ends, where the structural constraints are somewhat looser and window 2 is more populated.

The tetrahedral geometry representation has been used to define a number of geometric descriptors such as discretized rotamers, number of atoms seen out of each window, and potential contacts among spatially close residues. We focus here on one of these descriptors, the *visible volume*, defined as the empty space surrounding each residue position and in the line of sight of the corresponding β carbon. To visualize it, imagine sitting at the β carbon and looking around. Your horizon is a sphere of which you occupy the center. Visible volume is the measure of the empty space you can see, that is, the space that is not in the shadow cone of any of the surrounding atoms (Fig. 2). In general, in any position, you will see almost nothing but your own backbone when looking down. If you are at the surface, almost all of the surrounding space is empty. If you are deeply buried within the protein, much of the space is full of atoms, and the visible volume associated with your position is reduced accordingly. Visible volume is expressed as a fraction of the total available volume, as defined by the enclosing sphere. By visible volume we mean the total volume out of windows 2, 3, and 4 (disregarding the bottom window), but it can also be computed for each individual window if more detailed information is needed. It is worth noticing that, for each residue position, only the closest atoms contribute to the definition of the corresponding visible volume. Atoms that are in the shadow cone of other atoms do not contribute anything. Also, each atom's contribution is weighted by its distance from the β carbon, since the width of the shadow cone depends on this distance.

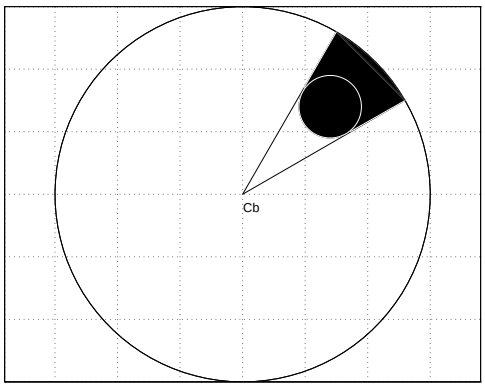

Figure 2: A pictorial 2-D illustration of visible volume. The shadowed region cannot be seen from the β carbon and does not contribute to the visible volume for that position.

Fig. 3 shows visible volume patterns for sperm whale myoglobin and human deoxyhaemoglobin α and β chains. The deep notches correspond to GLYs in the native structures at the close crossing between the B and E helices. For comparison, the visible volume for a protein

of the same length as myoglobin (cytokine interlukin-1 β) is also shown. These apparently random sequences are indeed characteristic of the globin family, a 1-D signature for 3-D structures (see below). The globin family is known to exhibit a wide range of structural variations (Lesk & Chothia, 1980). However, since the overall 3-D fold is conserved within the family, the environment of most of the corresponding residue positions is likely to be somewhat analogous. The issue is to properly encode this environment, in a robust way with respect to tolerated variations. Since individual amino acids are not conserved, a suitable representation should be independent of them. Visible volume is one of the possible representations with this property. When computing it, we ignore the nature of both the residue occurring at a given position and of the ones surrounding it; we focus instead on the space that these residues could occupy. We claim that this space is more or less the same across a protein family. To the extent that this assumption holds, that is, as long as the surroundings of corresponding positions are similar enough (as in the case of conserved contacts or functional sites, for example) visible volume is likely to manifest this structural invariance. In what follows, we investigate both its robustness and the kind of structural information it encodes.

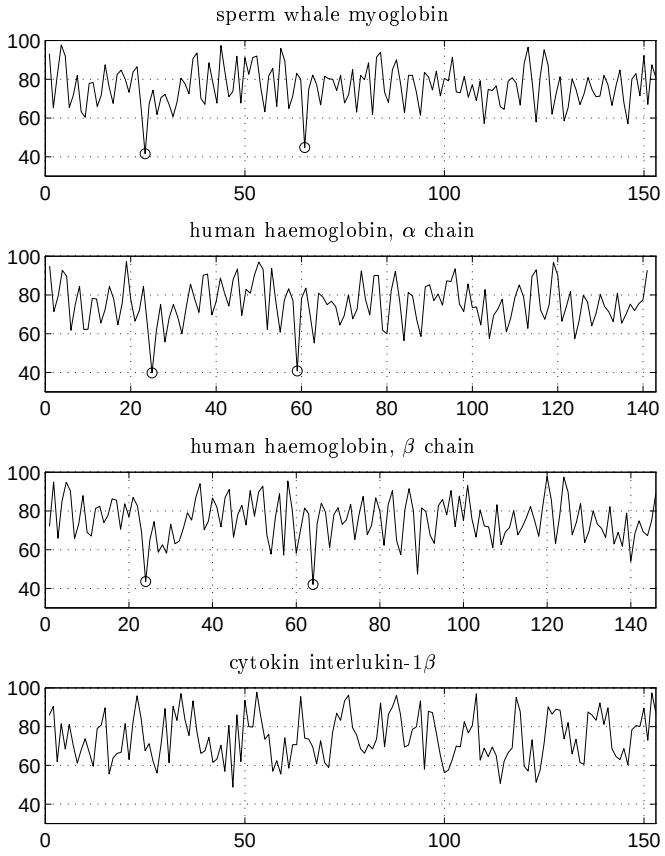

Figure 3: Visible volume patterns for three proteins of the globin family: sperm whale myoglobin, human haemoglobin (α chain), and human haemoglobin (β chain). The two marked notches correspond to GLYs at the close crossing between the B and E helices. For comparison, the pattern for cytokine interlukin-1 β is also shown. Visible volume is the volume available for the side-chains in the native structure out of window 2, 3, and 4, and is expressed as a percentage of the total volume, as defined by an enclosing sphere of 11Å radius.

Being an integral quantity, visible volume is only slightly affected by point misplacements, and is more reliable than other measures based either on inter-atomic distances, or heavily relying on atomic coordinates. To see why, consider an error of 10% in the position of any one of two atoms. For example, the first atom can be at position (0,0,0), the second at position (9,0,0), while the correct position is (10,0,0). The absolute error in computing the distance between the two atoms is 1. The relative error is 1/10, of the same order of magnitude as the original error in the coordinates. Now, let's consider visible volume. The absolute error is the difference between the two shadow cones, one corresponding to the atom in the correct position, and the other corresponding to the misplaced one. But in this case the relative error is much smaller, since the absolute error is divided by the volume of the enclosing sphere, that is, by a much bigger number, of the order of the cube of the distance.

To assess the robustness of visible volume with respect to atom displacements, we jiggled the myoglobin model by applying a random rotation to each of the ϕ, ψ angles and moving the atoms accordingly. For each residue i, the rotation applied to the ϕ angle affects only the carbon and oxygen atoms of residue $i-1$, and the one applied to the ψ angle affects only the oxygen atom of residue i and the nitrogen atom of residue $i+1$. The position of the β carbons is recontructed on the basis of the new backbone coordinates and the ideal tetrahedral geometry. In this way, we generate a local perturbation of all but the C_α atoms in the model structure. This local pertubation is meant to simulate errors in the atomic coordinates. Depending on the entity of the rotations, the sphere surrounding each position will enclose a different set of atoms, at different distances from the β carbon, resulting in a different set of visible volume values. For random rotations of $\pm 10°$, $\pm 20°$, $\pm 30°$, $\pm 50°$, corresponding to a maximum linear displacements for a unit vector of 0.17Å, 0.35Å, 0.5Å, and 0.85Å respectively, the correlation coefficients between the visible volume for the original model and the rotated ones are 0.98, 0.97, 0.94, and 0.84. From this, we can conclude that visible volume is insensitive to crystallographic errors.

Visible volume is a robust measure also in another sense: it is essentially parameter free. The only parameters entering its computation, that is, the atomic radii and the radius of the enclosing sphere, affect the absolute val-

ues, but not the general pattern. Therefore, their choice is not critical, if they are used consistently. We computed the visible volume for all residues in myoglobin using the van der Waals' radii[1] reported by Chothia (1975) and the covalent radii[2] given by Richards (1974). The visible volume for the two sets of atoms vary with the size of the atoms by a constant factor and the overall pattern is absolutely conserved, as indicated by a correlation coefficient of 0.99.

We also investigated how visible volume depends on the radius of the enclosing sphere. A larger sphere corresponds to a wider horizon, and therefore should result in a better characterization of the residue environment, but at the expense of a higher computational burden. Unfortunately, in many situations, relaxing a threshold also means including non-relevant information and therefore adding noise to the representation. This is not the case with visible volume: from its definition it follows that, for any chosen cutoff, only those geometric constraints that are relevant to the current position are taken into account, since only nearby atoms contribute to the shape of the empty space around that position. We computed the visible volume for all residues in myoglobin using a sphere of 15Å radius. Again, the overall pattern does not change: the correlation coefficient between this set of values and the ones corresponding to a sphere of 11Å radius is 0.97. Similar results were obtained for other model structures. Therefore, we can safely conclude that visible volume is insensitive to the choice of parameters. In all the calculations that follow, we used a set of reduced van der Waals' radii (75%) and an 11Å radius for the enclosing sphere.

To evaluate how effective visible volume is in encoding useful structural information we considered the nine globins studied by Lesk and Chothia (1980). This set includes the α and β chains of human and horse haemoglobin, sperm whale myoglobin, the sea lamprey and an annelid worm monomeric haemoglobins, a larval insect erythrocruorin, and a leghaemoglobin. These molecules have very different amino acid sequences, but similar secondary and tertiary structures. Eight helices (A, B, C, D, E, F, G, and H) are common to all of them, except that four structures lack the small D helix. The helices assemble in a common fold that encloses the haeme pocket.

The following results appear in Lesk & Chothia (1980): once aligned, there are 116 positions common to all nine globins, with 16 to 88% side-chain identity; buried residues vary both in amino acid identity and size, the mean change in side-chain volume at any position being 56Å^3; the volume of the residues that form the interfaces between homologous helices also varies by up to 57%; the shift in relative position and orientation of homologous pairs of helices may be as much as 7Å and 30°. Moreover, one of the helices is missing in four of the nine structures. Although almost nothing is locally conserved, the nine globins clearly show an overall structural similarity. Therefore, this set is a challenging testbed for any method that aims to capture this global structural invariance on the basis of features that are locally computed. Moreover, there is a well defined sequence alignment based on the equivalence of structural positions (Lesk & Chothia, 1980) with which to compare other results.

We asked the following question: How much structural information, and which kind of information, is encoded by standard geometric descriptors vis-a-vis visible volume? In other words: If there is a common structural pattern, which of these features tell us that this is so? Since these descriptors associate a single value to each position in a chain, they can be treated as 1-D signals, or *space* series, and compared in a very simple and general way, based on correlation, without introducing assumptions or parameters that could bias the results. Pairwise properties that associate a value to two positions (like inter-atomic distances) express relations between these two positions instead of attributes of a single one and are not easily comparable. We limit our analysis to visible volume, accessible surface area, and dihedral angles. The last two correspond to the degree of burial and secondary structure, and are expected to contain useful information for identifying structurally homologous regions. Each of these descriptors has different properties and possible usages, for which the others may not even be defined. It is for the role that they play in protein structural characterization that we consider them here (see Holm & Sander, 1994; Orengo, 1994; Bowie *at al.*, 1991; Bowie & Eisenberg, 1993; Wodak & Rooman, 1993 for reviews, and Orengo & Taylor, 1990, for a discussion of the effectiveness of different descriptors in selecting subsets of residues for an initial structural alignment of proteins). In what follows, ϕ, ψ angles are taken from DSSP files (Kabsh & Sander, 1983). Accessible surface area values (ASA) are also from DSSP files, normalized in order to minimize the effect of side-chain size. Residue volumes are from Lesk & Chothia (1980). Visible volume is computed as described above (unless otherwise stated, by visible volume we mean the total volume out of windows 2, 3 and 4). Visible volume and ASA have been calculated for isolated subunits in the case of multiple chains.

We consider the α and β chains of human haemoglobin as models (the first one does not have the small D helix, the second one does). Our structural patterns are the individual helices as coded by the corresponding set of values for each of the geometric descriptors. For each descriptor, and for each helix in the model, we compute

[1] Cα,Cβ: 1.87 Å, carbonyl carbon: 1.76 Å, carbonyl oxygen: 1.40 Å, amide nitrogen: 1.6 Å

[2] Cα,Cβ: 0.77 Å, carbonyl carbon: 0.77 Å, carbonyl oxygen: 0.66 Å, amide nitrogen: 0.7 Å

the cross-correlation function between the values for that helix (pattern) and the whole set of values for each of the globins (target) in the set. That is, we slide the pattern along the target, one position at a time, and compute for each position the correlation coefficient between the two, therefore obtaining a function whose maximum corresponds to the best alignment of the helix pattern to the target. The operation is repeated independently for each helix and without imposing any constraints; aligned patterns can end up in the target in any order, and they can even partially or totally overlap. No knowledge is assumed about the target, such as the kind of secondary structure or the order in which the helices should appear. This is the most unsophisticated alignment algorithm one can think of. Its only purpose is to exhibit the amount of structural information encoded by each of the descriptors without adding anything of its own, therefore permitting a fair comparison. Our aim in this case is not to compare different structural alignment algorithms or propose a new one; rather, we would like to compare the information content of visible volume to that of other commonly used geometric features, using a well-studied example of difficult-to-detect structural similarity.

For similarity measure, we took the fraction of correctly aligned helices (the reference alignment and helix definition being the ones given in Lesk & Chothia, 1980). The B and C helices are lumped together, since they are contiguous. However, we consider the small D helix separately, because it is of interest to see whether or not it can be correctly identified. With the α chain of human haemoglobin as a model, 35 out of 48 (73%) helices are correctly aligned using ASA and 42 (88%) using visible volume. These fractions drastically drop if ϕ and ψ angles are used instead (25% and 46% respectively). Table 1 shows detailed results for ASA and visible volume, the model this time being the human haemoglobin β chain. Each entry indicates the difference in residue number between the correct alignment for the pattern and the position found by our method. A zero entry means that the pattern has been correctly aligned. The ϕ, ψ angle representation does not perform any better with respect to the α chain model, with 25% and 44% of correctly aligned helices (details not shown).

Using ASA, 35 out of 52 helices (67%) find their correct position, 43 (83%) using visible volume. The mean correlation coefficient for correctly aligned helices is 0.83 for ASA, 0.89 for visible volume. Only one of the eight A helices is correctly aligned using ASA, versus all of them using visible volume. Three of the four small D helices are correctly aligned with visible volume, none with ASA. For all three, the correlation coefficient is 0.97. The detection of the D helix is particularly tricky, because it is only 5 to 7 residues long and contiguous to the E helix. For both ASA and visible volume the closest globin (i.e., the one for which the total correlation coefficient is higher than for any other in the set) is the other β chain haemoglobin; however, two of the helices of horse haemoglobin β chain are misplaced by ASA.

(a)	HHbα	EHbα	EHbβ	SWMb	LHb	GHb	CEr	LgHb
A	0	0	0	0	0	0	0	0
BC	0	0	0	0	0	0	92	0
D			0	0	0		−15	
E	0	0	0	0	0	0	0	−57
F	0	0	0	0	1	39	0	0
G	0	0	0	0	0	0	−92	0
H	0	0	0	0	−8	0	−1	−73

(b)	HHbα	EHbα	EHbβ	SWMb	LHb	GHb	CEr	LgHb
A	0	130	78	130	57	126	78	130
BC	0	0	0	0	0	0	−6	0
D			−30	75	−30		83	
E	0	0	0	0	0	0	0	0
F	0	0	0	0	−87	−83	−87	0
G	0	0	0	0	0	0	0	0
H	0	0	0	0	−137	0	−1	0

Table 1: Alignment using human haemoglobin β chain. Each entry indicates the difference in residue number between the helix and its aligned target using visible volume (a) and ASA (b). A zero entry means that the helix has been correctly aligned. An empty entry means that the helix is missing in the native structure. Abbreviations used: HHbα (human haemoglobin, β chain, pdb locus 4hhb), EHbα, EHbβ (horse haemoglobin, α and β chains, pdb locus 2mhb), SWMb (sperm whale myoglobin, pdb locus 1mbd), LHb (sea lamprey monomeric haemoglobin, pdb locus 2lhb), GHb (annelid worm monomeric haemoglobin, pdb locus 2hbg), CEr (larval insect erythrocruorin, pdb locus 1ecd), LgHb (leghaemoglobin, pdb locus 2lh6). A, BC, D, E, F, G, H: helices.

The rather poor performance of dihedral angles on this set of structures characterized by a very high variability confirms their well-known sensitivity to minor variations in atomic coordinates. Moreover, dihedral angles tend to encode positions in similar secondary structures in a similar way, regardless of whether or not these positions play equivalent roles in two proteins. This means that a pattern of dihedral angles corresponding to an α helix is likely to align with any other pattern corresponding to another α helix, not just with the one that is structurally equivalent to the model — as also observed by Orengo and Taylor (1990). This is not surprising: after all, dihedral angles were originally intended for secondary structure characterization. In the case of ferredoxin (pdb locus 1fca), a striking example of gene elongation by duplication and internal symmetry with an apparent homology between the two halves of the amino acid sequence, the correlation coefficients between the values corresponding to the two halves for ϕ, ψ, ASA, and visible volume are 0.95, 0.89, 0.87, and 0.97 respectively. This shows that dihedral angles do

not necessarily perform poorly; however, low sequence homology and consequent high variability in atomic coordinates can reduce dramatically their effectiveness in correctly identifying structurally equivalent positions. It also shows that visible volume is able to identify internal symmetries; in the case of ferredoxin, the sequence conservation is particular high and visible volume patterns for the two halves are basically identical (data not shown).

One of the anonymous referees raised some interesting issues: 1) Does the fact that window 2 is almost always unoccupied in a largely α helical protein affect the analysis done on globins? Is visible volume still superior to dihedral angles and ASA in the case of all-β structures and $\alpha\beta$ structures? 2) What if the structure has internal symmetry, as in the case of $\alpha\beta$ barrels?

To answer these questions, we applied the same correlation coefficient method to the structure of vitelline membrane outer layer protein I and to three $\alpha\beta$ barrel structures. Vitelline membrane outer layer I (pdb locus 1vmo) is an all-β protein. The β-sheets have clear sequence similarities and are related by nearly perfect three-fold symmetry (Chothia & Murzin, 1993). The pdb file includes the coordinates for two chains. Their structures are similar but not identical and constitute an ideal test-case for answering both of the above questions. We aligned all 5 residue-long segments of chain B to chain A (i.e. we aligned visible volume patterns corresponding to residues 1 to 5, 2 to 6 and so on), and considered as correct the alignment for which the correlation coefficient between the pattern and the aligned target was maximum. The fractions of correctly aligned segments using visible volume, ASA, ϕ, and ψ angles are 65%, 60%, 35%, and 40% respectively. We repeated the same experiment, this time with patterns of length 10 instead of 5. The fractions are 97%, 83%, 68%, 60% for visible volume, ASA, ϕ, and ψ angles. To exhaust the set of most common folds, and further explore the effect of internal symmetry, we applied the same procedure to the pairwise alignment of the A chains of three $\alpha\beta$ barrel structures (chicken, human, and yeast triose phosphate isomerase, pdb loci 1tim, 1hti, 7tim). In all cases, for both 5 residue-long and 10 residue-long segments and for all three $\alpha\beta$ barrels, visible volume outperforms the others descriptors. We report only the best and worst results for ASA and dihedral angles, and the corresponding results for visible volume. ASA: 71%, 14% (visible volume 89%, 32%); ϕ angles, 43%, 9% (visible volume 80%, 35%); ψ angles, 58%, 10% (visible volume 80%, 32%).

Results obtained with ASA and visible volume are clearly related (see Table 1). When they fail, they almost always fail together, for the same helices. However, ASA consistently fails more often, even using different models (data not shown). To help understand why ASA performs so poorly in the case of the A helix, we plotted ASA and visible volume corresponding to that helix for all globins (Fig. 4). ASA patterns show a much higher variability. In general, volume-based measures yield more information than surface-based ones, simply because they consider the whole three-dimensional space and not a two-dimensional manifold. In the particular case of ASA, the distribution of values is not smoothed: all buried residues collapse to a unique exposure value (zero). On the other hand, the discrimination for partially exposed residues is too fine. Of 24,520 residues from 99 non-homologous proteins 3,537 have a zero ASA, 5,656 have an ASA greater than zero and less than 10%, and the remaining 15,327 form a long tail distribution of residues with an ASA between 10% and 100%. On the other hand, visible volume is never exactly zero. Even for buried residues, what it accounts for is their spatial environment, which varies. Therefore, there are no abrupt changes and patterns are more consistently represented.

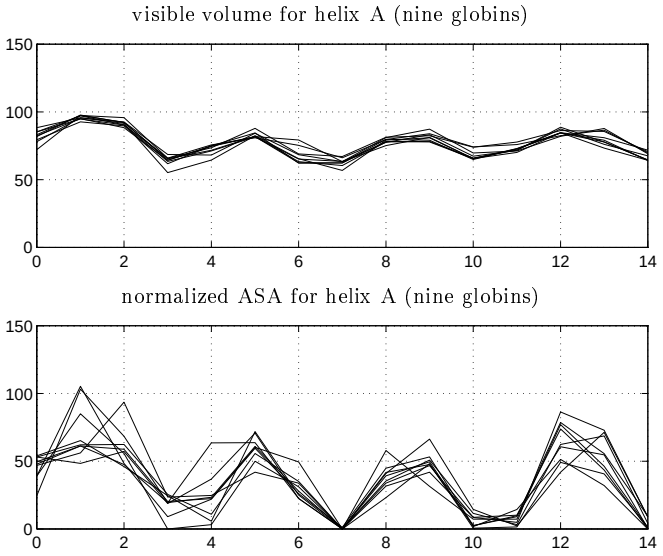

Figure 4: Patterns for the A helix in all nine globins for visible volume and normalized ASA. The two sets of patterns are clearly related, but ASA exhibits a much more pronounced variability.

Fig. 4 also shows that visible volume is related to amino acid exposure. Peaks and valleys in visible volume correspond to peaks and valleys in ASA. Fig. 5 shows the visible volume distribution for hydrophobic and hydrophilic residues from a set of non-homologous monomeric proteins. Mean values for residues belonging to the same class are remarkably close, with about 10% difference between the two classes. Fig. 5 also shows that for ALA and GLY the distribution is bimodal, as expected, since these two residues can be either buried or exposed. SER and THR have a similar, even though

less markedly bimodal, distribution (data not shown). Visible volume encodes the hydrophobicity pattern, one of the most important features of protein structures. Being designed to be side-chain independent, it captures a notion of "exposability" by identifying positions in the sequence that could be buried or exposed, instead of assigning a degree of exposure to the amino acids actually present. We stress this fact because of its relevance in modeling, prediction, and in all applications in which avoiding sequence memory is both meaningful and desirable, as in the case of statistics for threading purposes, or in the definition of structural profiles.

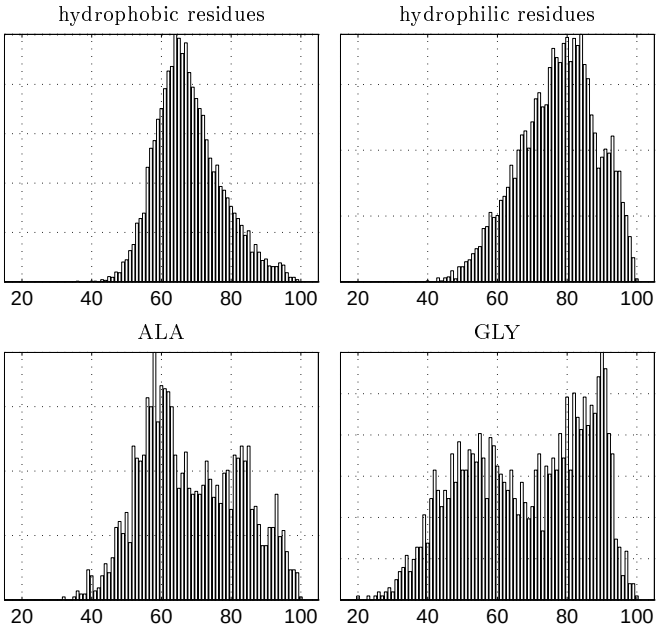

Figure 5: Normalized visible volume distribution for hydrophobic residues (PHE, ILE, LEU, MET, VAL), hydrophilic residues (ASP, GLU, LYS, ARG, ASN), ALA residues, and GLY residues from a set of 99 non-homologous monomeric proteins. The visible volume distributions for hydrophobic and hydrophilic residues are clearly distinct, with a mean value of 68% and 77.5% respectively and a standard deviation of about 10%. ALA and GLY can be either hydrophobic or hydrophilic, as reflected by their bimodal distribution.

We now turn to a different question. Lesk and Chothia (1980) studied the variation in the volumes of individual amino acids found in homologous positions in nine distantly related globins, and the total volumes of interacting sets of amino acids. In what follows, we show how visible volume compares with residue volume. We also discuss the different kind of structural information encoded by these two measures. The reader interested in the analysis of residue volume variation in the globin family is referred to Lesk & Chothia, 1980; Ptitsyn & Volkenstein, 1986; Gerstein *et al.*, 1994; Kapp *et al.*, 1995. We will use the globin residue numbering scheme adopted by Lesk & Chothia (1980).

According to Lesk & Chothia (1980), 59 out of 116 common residues are involved in helix-to-helix or helix-to-haeme contacts in seven or more globins. Of these, 31 are buried and 28 are exposed. Their results show that at the local level the volume of individual residues, even those involved in conserved contacts, is not generally conserved. If we now ask whether the visible volume for different classes of residues is conserved, we get a similar answer: in general, it is not. If we consider the difference between maximum and minimum visible volume at each of the 31 buried contacts, the mean value is 14.4%, the standard deviation 8.4%. Very similar figures are obtained for surface contacts (14.1% and 7.1%), and for 57 positions not involved in contacts (15.3% and 7.9%). However, by taking the visible volume average of the 31, 28 and 57 residues for all globins, the mean values of the averages are 65.8%, 73.4%, and 79.4% respectively, with standard deviations varying from 0.9% to 1.8%. Thus, the three classes separate in conformity with their definition. Lim and Ptitsyn (1970) showed that the total volume of 31 residues that form the hydrophobic core in different globins remains almost constant at 3180Å^3, with a root-mean-square deviation of 15Å^3. From our results, the same persistence is observed if we consider the total visible volume instead of the total residue volume, and this is true for both buried and exposed contact positions.

Residue volume and visible volume are related but non-equivalent quantities. The former is a measure of the volume of a residue in a given position. The latter is, in some sense, complementary: it measures the space surrounding a residue position that could be occupied by the side-chains of the protein from which the model was derived or by any other combination of side-chains and water molecules satisfying all steric, chemical, and physical constraints. It turns out that visible volume is conserved at structurally equivalent positions in a protein family even though amino acid residues (and residue volumes) are not.

The minimum visible volume variation at a single position in all nine globins is 2.7%, the maximum 43.3%. Minimum and maximum variations for residue volume are 0Å^3 and 172Å^3. Of the four positions containing side-chains with the same volume in all nine globins, three show a variation of visible volume of less than 7%, but the variation for the fourth is as much as 13.4%. There are even more striking examples of discordance: the largest variation in residue volume is 172Å^3, for a position in the E helix in the nine globins occupied by one GLY, four ALAs, one LYS, one TRP, one ASP, and one THR (index 74 using the numbering scheme in Lesk & Chothia, 1980) but the corresponding maximum variation in visible volume is only 6.8%. According to Lesk and Chothia (1980), this position is not making contacts, but the next one is involved in both helix-to-helix and helix-to-haeme contacts

in all globins, and the previous one in helix-to-helix contacts in all but one globins. It is likely that position 74, which is caught in the middle, is forced to be where it is, and to have a similar environment. Being on the surface, it exhibits a high variability in the amino acids sitting there and, as a consequence, in residue volume. A similar situation occurs at a position in the G helix (index 120), occupied by two VALs, four GLUs, one LYS, one ASN, and one ALA. The maximum variation in residue volume is 75Å^3 but in visible volume is only 4.1%. One can go back to the native structure and ask what is peculiar about this position. It turns out that the two VALs in the α chain of human and horse haemoglobin are facing the internal cavity and participate in the $\alpha_1\beta_2$ interchain contacts (Perutz *et al.*, 1968). This explains the two hydrophobic residues in an otherwise exposed position, with a mean visible volume of 92.3%: the two VALs are actually buried after the formation of the tetramer. None of the features related more or less directly to the amino acids sitting there, neither residue volume, nor the chemical homology of amino acids, nor ASA, whose maximum variation is 40%, would tell us that this position is likely to be of biological interest.

There are 12 positions out of the 116 common residues for which the variation in visible volume is less than 6% (13, 43, 49, 50, 70, 75, 78, 82, 117, 120, 151, 160.) Of these, two (75 and 160) correspond to conserved residue volumes. For the remaining 10, the difference between maximum and minimum residue volumes at aligned positions vary from 27Å^3 to 146Å^3. Position 13, 49, and 70 are on the surface in all nine globins. A conserved surface position is even easier to detect, since almost all of the space is empty, with very small variations in visible volume. Positions 50, 75, 78, 82, and 117 are all involved in helix-to-helix and/or helix-to-haeme contacts (Lesk & Chothia, 1980). We have already discussed the role of position 120. Let us look in detail at position 151, populated by ALA, LYS, and ILE. Fig. 6 shows the shape of the space available in the case of LYS and ALA. Visible volumes for LYS and ALA are the same (71.4% and 71.8%) and the two shapes are remarkably similar. Residue volumes are 171Å^3 for LYS and 92Å^3 for ALA. Considering visible volumes for individual windows, we can get further insight. The values for window 2, 3, and 4 for LYS are 50.2%, 75.5%, and 88.6%; for ALA 46.3%, 80.7%, and 88.4%. While visible volume out of window 4 is absolutely conserved, there is a 5% difference between windows 2 and 3, but these differences compensate, resulting in a negligible variation of 0.4% for the total volume out of the three windows. As a historical note, we quote a comment about position 151 (labeled H9) that appears in Dickerson & Geis (1969) in the discussion of conserved amino acids in the small set of globin sequences available at that time:

> But the purpose of Lys H9, which appears to point out away from the molecule into its surroundings, is a mystery. What importance should it have that it should be preserved so carefully through 500 million years of evolution?

Since then, this position has lost its character of absolute conservation, even though only few substitions have been observed. Bashford *et al.* (1987) report the occurrence of 195 LYSs in 207 globin sequences. We still don't know why this LYS is almost always there, and the striking similarity of the two shapes of the space available for side-chain packing when ALA is substituted for LYS could be just a coincidence. Perhaps it is only by chance that position 151 is so peculiar. Perhaps there is a good reason. Visible volume by itself cannot provide a definite answer to this kind of questions, but it can point to regions whose invariance could be biologically relevant. A similar situation occurs at position 43, where a PHE in sea lamprey haemoglobin has 64.3% of the space available, and a VAL in leghaemoglobin has the same amount, 64.8%, but residue volumes are 203Å^3 and 142Å^3 respectively.

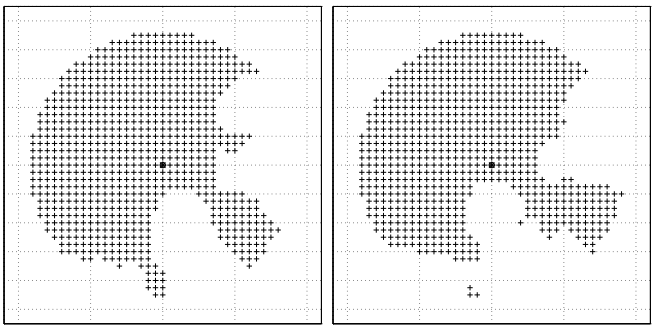

Figure 6: Projection along the z-axis (z=6) of the shape of the space surrounding LYS 127 in horse haemoglobin α chain (left) and ALA 123 in larval insect erythrocruorin (right). The two shapes are remarkably similar for all other z-levels as well.

These examples show that visible volume can help identifying local units of biological interest by measuring how much room is available for different combinations of side-chains to fit in. By ignoring the nature of both the residue at a site and of the ones surrounding it, and focusing instead on the space that they can occupy, we were able to quantify a new kind of invariance beyond the apparent variations in a protein family, namely, the conservation of the space available at corresponding residue positions for 3-D side-chain packing, independently of the amino acids sitting there. Gassner *et al.* (1996) have recently shown that the structure of a variant of T4 lysozyme, in which seven METs have been substituted for corresponding core residues, is similar to the wild type and maintains a well-ordered core. This implies that protein

structure can adapt to changes in the shape of residues, and many different combinations of hydrophobic residues in the core can result in a structurally stable (and partially active) protein (Chothia & Gerstein, 1997). By its definition, visible volume is a suitable tool for exploring a protein structure from this point of view. By asking simple questions, we were able to single out local units within a protein family in cases where the amino acids were not conserved, but the local spatial arrangement of the main chain and, consequently, visible volume, were basically the same. This information can be used for protein structure prediction, protein design, and homology modeling.

The fact that visible volume encodes this kind of structural information also explains why most of the helices in the globin family were correctly aligned, despite the lack of conservation of any other attribute. Even though the average visible volume variation at the aligned positions which are common to all nine molecules is of the order of 15%, the amount of space that is available at most of these positions is more or less conserved across distantly related members of the globin family. The assumption that the local spatial environments do not change much is a valid one, and visible volume properly encodes this information in the most simple and natural way. When used as a means for characterizing protein structures, visible volume is superior to ASA and dihedral angles not only in the case of globins, but also for other kinds of folds, like all-β and $\alpha\beta$ structures. Protein structural alignment and classification appears to be one of the most promising applications for this new measure.

The use of windows adds directionality to our representation — a novel feature that is lacking in all descriptors that we are aware of. A visible volume can be computed for each window, and we can ask questions such as, Which of the windows is buried or exposed? or, Out of which window is the visible volume (i.e. the space available for side-chains) maximum or minimum? These questions naturally lead us to what is perhaps the most natural application of visible volume, namely, the threading approach to the inverse protein folding problem. The intuition, confirmed by calculations, is that a side-chain does not extend out of a given window if there is not enough room for it. In both α and β structures there is a correlation between the rotamers as observed in real proteins and the window through which the visible volume is maximum (Smith *et al.*, 1997a,b).

The ability to quantify two of most of the important features of protein structures, namely, the amount of space available for side-chain packing and the degree of exposure of a residue position, together with the correlation of visible volume with the actual rotamers in native proteins, qualify this new measure as a powerful tool in a variety of applications, from the detailed analysis of protein structure to homology modeling, protein structural alignment, and the definition of better scoring functions for threading purposes.

Acknowledgement

The authors wish to express their particular thanks to Jean Garnier, Tom Toffoli, Sandor Vajda, and Jim White for helpful suggestions; to their colleagues for critical reading of this paper; and to one of the anonymous referees for raising interesting questions about the generalization of the alignment results to different kind of folds. Loredana Lo Conte is the recipient of an IBM Cooperative Fellowship. This work was supported in part by grant LM05205-13 from the National Library of Medicine.

References

[1] Bashford, D, Chothia, C., Lesk, A. M. (1987) Determinants of a Protein Fold Unique Features of the Globin Amino Acid Sequences, *J. Mol. Biol.*, **196**, 199–216.

[2] Berbstein, F. C., Koetzle, .T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O, Shimanouchi, T., Tasumi, M. (1977) The protein data bank: a computer-based archival file for macromolecular structures, *J. Mol. Biol.*, **112**, 535–542.

[3] Bowie, J. U., Lüthy, R., Eisenberg, D. (1991) A Method to Identify Protein Sequences That Fold into a Known Three–Dimensional Structure, *Science*, **253**, 164–170.

[4] Bowie, J. U. & Eisenberg, D. (1993) Inverted protein structure prediction, *Curr. Opin. Struc. Biol.*, **3**, 437–444.

[5] Chothia, C. (1975) Structural invariants in protein folding, *Nature*, **254**, 304–307.

[6] Chothia, C. & Gerstein, M. (1997) How far can sequence diverge?, *Nature*, **385**, 579–581.

[7] Chothia C. & Murzin, A. G. (1993) New folds for all-β proteins, *Structure*, **1**, 217–222.

[8] Dickerson, R. E. & Geis, I. (1969) *The structure and action of proteins*, p. 53, Harper & Row, New York.

[9] Gassner, N. C., Baase, W. A., Matthews, B. W. (1996) A test of the "Jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme, *Proc.. Nat. Acad. Sci., USA*, **93**, 12155–12158.

[10] Gerstein, M., Sonnhammer, E. L, L., Chothia, C. (1994) Volume Changes in Protein Evolution, *J. Mol. Biol.*, **236**, 1067–1078.

[11] Gregoret, L. & Cohen, F. E. (1990) Novel Method for the Rapid Evaluation of Packing in Protein Structures, *J. Mol. Biol.*, **211**, 959–974.

[12] Harpaz, Y., Gerstein, M., Chothia, C. (1994) Volume changes on protein folding, *Structure*, **2**, 641–649.

[13] Holm, L. & Sander C. (1994) Searching Protein Structure Databases Has Come of Age, *Proteins: Struc., Func., Genet.*, **136**, 225–270.

[14] Kabsch, W. & Sander, C. (1983) Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features, *Byopolymers*, **22**, 2577–2637.

[15] Kapp, O. M., Moens, L., Vanfleteren, J., Trotman, C. N. A., Suzuky, T., Vinogradov, S. N. (1995) Alignment of 700 globin sequences: Extent of amino acid substitution and its correlation with variation in volume, *Prot. Science*, **4**, 2179–2190.

[16] Kendrew, J. C. (1962) Side-Chain Interactions in Myoglobin, *Brookhaven Symp. Biol.*, **15**, 216–228.

[17] Lee, B. & Richards, F. M. (1971) The Interpretation of Protein Structures: Estimation of Static Accessibility, *J. Mol. Biol.*, **55**, 379–400.

[18] Lesk, A. M. & Chothia, C. (1980) How Different Amino Acid Sequences Determine Similar Protein Structures: The Structure and Evolutionary Dynamics of the Globins, *J. Mol. Biol.*, **136**, 225–270.

[19] Lim, V. I. & Ptisyn, O. B. (1970) On the constancy of the hydrophobic nucleus volume in molecules of myoglobin and hemoglobin, *Mol. Biol. (U.S.R.R.)*, **4**, 372–382.

[20] Orengo, C. (1994) Classification of protein folds, *Curr. Opin. Struc. Biol.*, **4**, 429–440.

[21] Orengo, C. & Taylor, W. R. (1990) A Rapid Method of Protein Structure Alignment, *J. Theor. Biol.*, **147**, 517–551.

[22] Pattabiraman, N., Ward, K. B., Fleming, P. J. (1995) Occluded Molecular Surface: Analysis of Protein Packing, *J. Mol. Recog.*, **8**, 334–344.

[23] Perutz, M. F., Muirhead, H., Cox, J. M., Goaman, L. C. G. (1968) Three-dimensional Fourier Synthesis of Horse Oxyhaemoglobin at 2.8Å Resolution: The Atomic Model, *Nature*, **219**, 131–139.

[24] Ptitsyn, O. B. & Volkenstein, M. V. (1986) Protein structures and the neutral theory of evolution, *J. Biol. Struct. Dynam*, **4**, 137–156.

[25] Ramachandran, G. N. & Sasisekharan, V. (1968) Conformation of polypeptides and proteins, *Adv. Protein Chem.*, **23**, 283–437.

[26] Richards, F. M. (1974) The Interpretation of Protein Structures: Total Volume, Group Volume Distributions and Packing Density, *J. Mol. Biol*, **82**, 1–14.

[27] Richardson, J. S. (1981) The anatomy and taxonomy of protein structure, *Adv. Prot. Chem.*, **34**, 167–339.

[28] Smith, T. F., Lo Conte, L., Bienkowska, J., Roger, B., Gaitatzes, C, Lathrop, R. (1997a) The Threading Approach to the Inverse Protein Folding Problem, *Proc. First Annual Conf. Comput. Mol. Biol. RECOMB97*, ACM press, 287–292.

[29] Smith, T. F., Lo Conte, L., Bienkowska, J., Roger, B., Gaitatzes, C, Lathrop, R. (1997b) Current limitations to protein threading approaches, *J. Comp. Biol.*, in press.

[30] Wodak, S. J. & Rooman, M., J. (1993) Generating and testing protein folds, *Curr. Opin. Struc. Biol.*, **3**, 247–259.