The Protein Journal, Vol. 25, No. 5, July 2006 (© 2006)
DOI: 10.1007/s10930-006-9016-5
Published Online: September 7, 2006

Quantitative Analysis of the Conservation of the Tertiary Structure of Protein Segments

Jishou Ruan,[1,*] Ke Chen,[4,#] Jack A. Tuszynski,[2,3,†] and Lukasz A. Kurgan[4,5]

The publication of the crystallographic structure of calmodulin protein has offered an example leading us to believe that it is possible for many protein sequence segments to exhibit multiple 3D structures referred to as multi-structural segments. To this end, this paper presents statistical analysis of uniqueness of the 3D-structure of all possible protein sequence segments stored in the Protein Data Bank (PDB, Jan. of 2003, release 103) that occur at least twice and whose lengths are greater than 10 amino acids (AAs). We refined the set of segments by choosing only those that are not parts of longer segments, which resulted in 9297 segments called a sponge set. By adding 8197 signature segments, which occur uniquely in the PDB, into the sponge set we have generated a benchmark set. Statistical analysis of the sponge set demonstrates that *rotating*, *missing* and *disarranging* operations described in the text, result in the segments becoming multi-structural. It turns out that *missing segments* do not exhibit a change of shape in the 3D-structure of a multi-structural segment. We use the root mean square distance for unit vector sequence (URMSD) as an improved measure to describe the characteristics of *hinge rotations*, *missing*, and *disarranging segments*. We estimated the rate of occurrence for rotating and disarranging segments in the sponge set and divided it by the number of sequences in the benchmark set which is found to be less than 0.85%. Since two of the structure changing operations concern negligible number of segment and the third one is found not to have impact on the structure, we conclude that the 3D-structure of proteins is conserved statistically for more than 98% of the segments. At the same time, the remaining 2% of the sequences may pose problems for the sequence alignment based structure prediction methods.

KEY WORDS: Multi-structural segments; protein structure; protein structure comparison; protein structure conservation; URMSD.

1. INTRODUCTION

In the past, it has been commonly assumed that a protein can only adopt a unique 3D structure that corresponds to the minimal value of the free energy (Anfinsen, 1973). With over twenty thousand 3D protein structures determined, it has become clear that some proteins may adopt multiple equilibrium tertiary structures, at least partially due to the influence of the surrounding environment (Drum *et al.*, 2002; Elhorst *et al.*, 1999; Meador *et al.*, 1992; Schumacher *et al.*, 2004; Yap *et al.*, 2003). In a recent study we

* Jishou Ruan research was supported by Liuhui Center for Applied Mathematics, China-Canada exchange program administered by MITACS and NSFC (10271061).
# Ke Chen and Lukasz A. Kurgan research was partially supported by NSERC Canada.
† Jack A. Tuszynkski research has been supported by MITACS, NSERC Canada and the Allard Foundation.

[1] Chern Institute of Mathematics, College of Mathematical Science & LPMC, Nankai University, Tianjin, 300071. P. R. China.
[2] Department of Physics, University of Alberta, Edmonton, AB, Canada, T6G 2J1.
[3] Department of Experimental Oncology, Cross Cancer Institute, Edmonton, AB, Canada, T6G 2J1.
[4] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada, T6G 2V4.
[5] To whom correspondence should be addressed. E-mail: lkurgan@ece.ualberta.ca

Abbreviations: amino acid, AA; Protein Data Bank, PDB; root mean square distance, RMSD; root mean square distance for unit vector sequence, URMSD; three-dimensional, 3D.

have demonstrated that for the calmodulin protein (Chen *et al.*, 2006). The existence of multiple-structure proteins, with a greater preponderance for specific primary sequences, may be an intrinsic property even though it may be tempting to explain this as a result of the surrounding environment. Prediction methods for the tertiary structures based on the primary sequences alone use the statistical tools where the uniqueness of the 3D structures of proteins is assumed. The present paper attempts to shed light on the problem of the existence of multi-structural elements in proteins which may open new vistas in the area of protein structure prediction in general. The outline of this paper is as follows:

1. In Section 2, we construct the benchmark set and describe three causes for a given protein sequence segment to become multi-structural, i.e. rotating, missing and disarranging.
2. In Section 3, we describe the method in detail that includes
   A. Finding the numerical algorithm to describe the characteristics of the three causes of proteins to become multi-structural.
   B. Finding threshold values for the three characteristics mathematically.
   C. Constructing the indicator sequences for segments, and using it to describe quantitative characteristics of the rotating, missing and disarranging segments.
   D. Estimating the upper bounds on the rate of the rotating and disarranging segments related to the benchmark set.
3. In Section 4, we discuss implications of our findings.

In contrast to the recent study of protein segments by Kihara and Skolnick (Kihara and Skolnick, 2003), which concentrates on the similarity of structures between segments sharing low sequence homology, we discuss the native characteristic of the multi-structural protein segments with respect to the conservation of their 3D structures.

2. CLASSIFICATION OF THE MULTI-STRUCTURAL SEGMENTS

2.1. Searching the Benchmark Set Based on the PDB database

In order to analyze multi-structural segments of proteins, we need to construct the benchmark set. To this end, we use all chains in the PDB database (release # 103) (Berman *et al.*, 2000), which number approximately 53,000. Among these chains 8197 occur in the PDB database only once. We collect all these segments whose lengths are more than 10 AAs and which occur at least in two chains. This is a time-consuming task that requires a huge amount of memory space if we use a naïve method. To save space we use Shen's method (Shen *et al.*, 2004), which is summarized it as follows:

- We search for all of the 10-residue segments that occur at least twice in the PDB database and denote this set by $S(10)$. When searching for the second segment identical to the one selected, we examine any position in a sequence.
- For each segment s in $S(10)$, we randomly add an AA at its end, to generate 20 possible 11-residue segments. If none of the 20 segments can be found in the PDB database more than twice, then we remove s from $S(10)$ to form $S_1(10)$. Otherwise, we delete s from $S(10)$, and for all of these enlarged segments occurring in the PDB database at least twice, we store them in $S(11)$. Continuing this procedure until $S(10)$ becomes empty, we end up having two sets: $S_1(10)$ and $S(11)$. It can be readily verified that $S(11)$ is just the set consisting of all 11-residue segments that occur at least twice in the PDB database.
- With the same argument applied to $S(11)$, we obtain two new sets: $S_1(11)$ and $S(12)$.
- We continue this procedure to get $S_1(m)$ and $S(m + 1)$ for an arbitrary value of m.
- The procedure terminates when $S(m + 1)$ is found to be empty.

Since the longest segments that occur at least twice in PDB database are the D and N chains of 1IW7, with a length of 1524, the set $\bigcup_{m=10}^{1524} S_1(m)$ contains all the segments that occur in the PDB database at least twice. Therefore, all segments drawn from the 44,813 chains are collected in the set and some segments drawn form the 8,197 chains are also collected in the set if they occur twice. However, the actual number of elements contained in the set $\bigcup_{m=10}^{1524} S_1(m)$ is too large for us to check. Note also that there are many redundant segments in the set. For example, segment "LVETRPAGD GTFQ

KWA" and segment "RPAGDGTFQKWA" both belong to $\bigcup_{m=10}^{1524} S_1(m)$, while the latter is a sub-segment of the former. We say that a shorter segment is absorbed by a longer one if we delete the shorter one from $\bigcup_{m=10}^{1524} S_1(m)$. After this additional processing, the total number of the remaining segments in the set $\bigcup_{m=10}^{1524} S_1(m)$ is found to be 9297. We regard the set consisting of these 9297 segments as a block of sponge that can sponge all the segments in $\bigcup_{m=10}^{1524} S(m)$, and denote it by S_p. It is important to note that S_p absorbs all protein sequences occurring at least twice in the PDB database piece-wise. Thus, S_p can sponge all protein segments drawn from the 44,813 chains. We define the signature segment as a segment that uniquely occurs in the PDB database and is more than 10 AAs long. Therefore, if a segment is drawn from the 8197 chains, then S_p will absorb it unless it is a signature. Since the 8197 chains occur in PDB only once, this implies that every protein chain has at least one signature. Thus, it is imperative that we include all signatures drawn from the 8197 chains in S_p together with all the remaining signatures as the benchmark set. Then any protein in the PDB database will be absorbed by the benchmark set in a piece-wise fashion.

The concept of the sponge segment is a key aspect of this paper so we elucidate it further using Fig. 1 showing the distribution of segment lengths.

Four Thousand six hundred and forty five segments in S_p are 10–29 AAs in length, and they can be located at the tail end of a protein. The diversity of their 3D-structures is largely due to the fact that

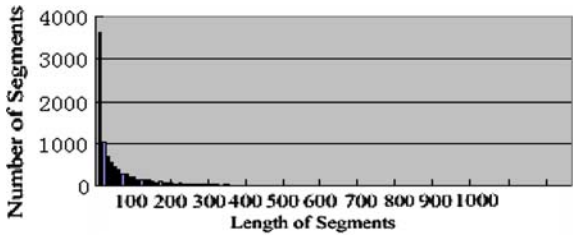

Fig. 1. Histogram showing the number of segments for a given length in the sponge set. Amongst the 9297 segments, there are 4645 segments whose lengths are between 10 and 29 residues; 2790 segments whose lengths are between 30 and 99 residues; 1550 segments whose lengths range between 100 and 299 residues; 249 segments whose lengths range between 300 and 499; and only 63 segments whose lengths are more than 500 residues, the longest segment consists of 1524 residues.

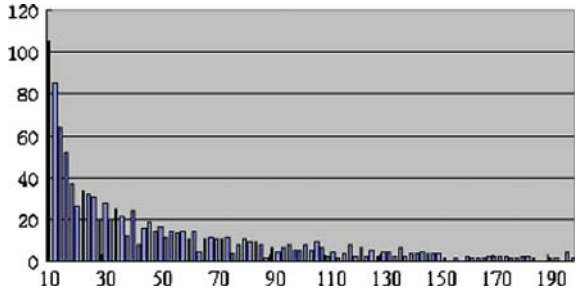

Fig. 2. Histogram showing the distribution of the tail segments according to their length in the sponge set. The particular numbers are 483 for the tail segments with 10–29 AAs, 407 for the tail segments with 30–99 AAs, and 133 for the tail segments with 30–99 AAs.

it is not always possible to exactly determine the coordinates of the AAs at the head (N-terminus) and tail (C-terminus) ends of protein. Hence the location of the segments in S_p is quite important. In Fig. 2 we show the distribution of the segments that constitute the tail ends of proteins.

2.2. The Causes of the Emergence of Multi-structural Segments

The differences between the individual structures of a multiple-structure protein are due to the presence of special segments with multiple-structures. A protein chain can be also seen as a segment of the protein. By aligning the high-level structures of segments in S_p, we find numerous segments occurring naturally, except the tail segments with various structures resulting often from human error. While the diversity of the structures of segments is natural, it arises only as a result of the following three causes.

First, a large number of multiple-structure proteins are domain-swapped dimers that are linked differently and which exhibit a rotation around the hinge region (Barrientos *et al.*, 2002). Typically, as is the case with calmodulin, different structures are found by making a rotation at its hinge region, while the structure of every single domain is invariant (Toyoshima and Nomura, 2002; Xu *et al.*, 2002) as is shown as Fig. 3(a–c).

For each of these segments, its secondary structure is invariant, but the differences between pairs of 3D structures of the multi-structural segment are very large. Generally, all multi-structural segments occurring for this reason are called *rotating* segments.

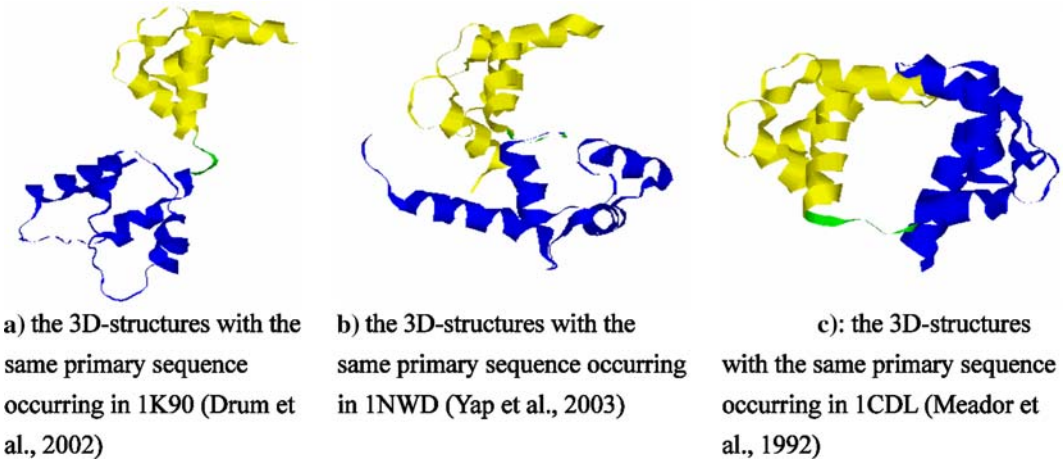

Fig. 3. Example domain-swapped dimmers; domains 1 (residues 1–75; in blue) and 2 (residues 81–147; in yellow), which are connected by the link (residues 76–80; in green) are invariant in the three cases shown even though the shapes of the corresponding 3D-structures are quite different.

Second, some segments often lose a helix or a sheet, which may change their 3D-structures to some degree. That is, a given segment occurs in at least two proteins; it forms a helix or a sheet in one protein but forms no particular motif (other than a random coil) in another protein. In this case, we may conclude that the latter protein misses a secondary structure. For example, the segment (LEU214 to LEU279) occurs in both proteins 2HMI (Ding *et al.*, 1998) and 1IKX (Lingberg, *et al.*, 2002). Observing the blue (dark) part in Fig. 4(a, b), we may find that 2HMI misses a sheet.

Again, the segment (VAL33 to LYS76) occurs in both 1I87 (Falzone *et al.*, 2001) and 1WDB (Muskett and Whitford, to be published) proteins. Comparing Fig. 5(a, b) we find that it misses three helices in 1I87.

In general, all of these segments that miss their secondary structure are called *missing* segments.

Third, segments are called *disarranging* when their 3D-conformations are similar and secondary structures are the same, but different residues in different regions form similar conformations. For example, the two loops of the same segment (GLU47 to VAL80) that occur in both 1IW7 (Berman *et al.*, 2000) and 1L9U (Chew *et al.*, 1999) are similar, but are formed by different residue pieces. For illustration compare the blue, green and yellow parts in Fig. 6(a, b).

In addition to the three main sources above, we found only seven multi-structural segments, i.e. 1FJI, 3ITS, 1DL9, 1LQN, 1ITN, 1B29 and 1B61 (Bamborough *et al.*, 1994; Brody *et al.*, 1999; Hansson *et al.*, 1997; Tiraboshi *et al.*, 1999; Veerapandian, 1992), which arise due to other causes. The differences between their 3D structures are due to the fact that each of these seven segments has a pair of 3D structures, one is given by a theoretical model

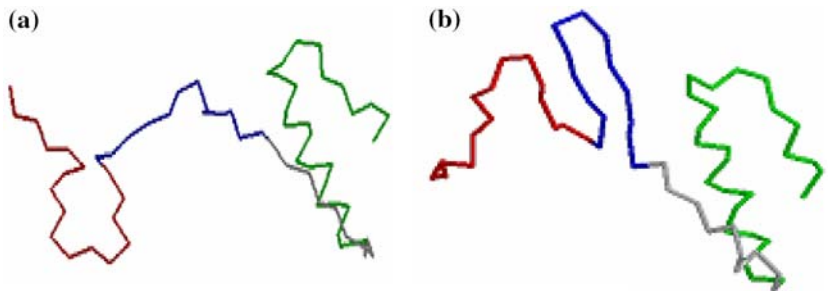

Fig. 4. The LEU214 to LEU279 segment from (a) 2HMI and (b) 1IKX.

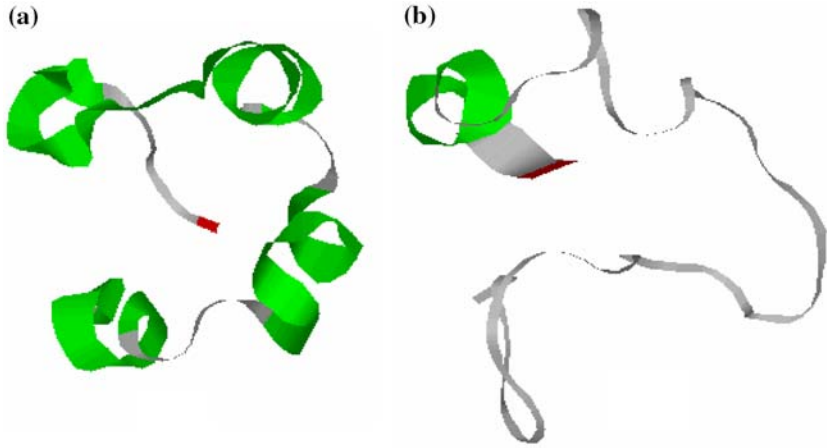

Fig. 5. The VAL33 to LYS76 segment from (a) 1WDB and (b) 1I87.

while the other is found empirically. The references given above indicate that these theoretical models may be fraught with some problems hence we ignore these cases as we have reason to believe that the differences are due to a human error. That is, we only consider the 9290 segments in the sponge set.

3. METHODS

3.1. The Measurement Method

In order to compare two distinct protein structures, we must ignore the rigid translational and rotational transforms that leave the structure invariant. Thus, we translate each 3D-structure of a multi-structural segment into a sequence consisting of unit vectors in R^3. An ideal measure for this is the so-called URMSD which stands for the Root Mean Square Distance for Unit Vectors proposed by Chew *et al.* (1999) for detecting the common geometric substructure in proteins.

URMSD is a variant of the RMS distance, in which instead of comparing residue positions we compare the corresponding unit vectors. The tertiary protein structure is represented by a set of coordinates of the C_α-atoms, $\{C_\alpha(i)$, which become inputs for the root mean square distance (RMSD) (Kabsch, 1978). Given two n residues long structures, A and B, and the corresponding two sets of coordinates $\{C_\alpha(i)\}$ denoted as $(x_A(i), y_A(i), z_A(i))$ and $(x_B(i), y_B(i), z_B(i))$, the RMSD is defined as

$$\mathrm{RMSD}(A,B) = \mathrm{RMS}(\{x_A(i), y_A(i), z_A(i)\}, \{x_B(i), y_B(i), z_B(i)\}; i = 1, 2, \ldots, n)$$

URMSD is defined as the minimal RMS distance between the corresponding unit vectors. Let V_i and W_i be the vectors from atom $C_\alpha(i+1)$ to $C_\alpha(i)$ in structure A and B, respectively:

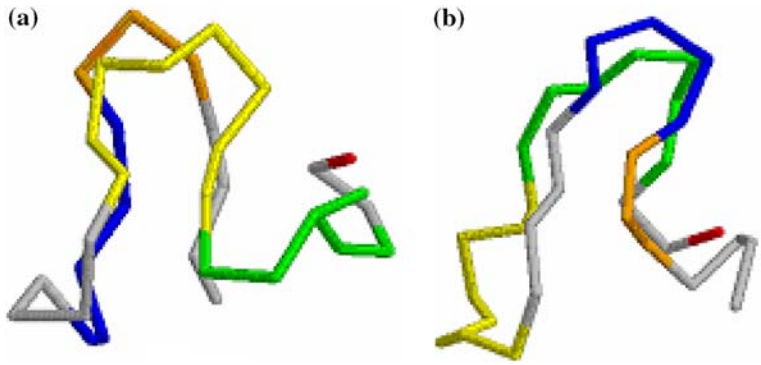

Fig. 6. The GLU47 to VAL80 segment from (a) 1IW7 and (b) 1L9U.

$$V_i = (v_{i1}, v_{i2}, v_{i3}) = (x_A(i+1), y_A(i+1), z_A(i+1)) - (x_A(i), y_A(i), z_A(i))$$
$$W_i = (w_{i1}, w_{i2}, w_{i3}) = (x_B(i+1), y_B(i+1), z_B(i+1)) - (x_B(i), y_B(i), z_B(i))$$

The sets of the two vectors $\{V_i\}$ and $\{W_i\}$, $i = 1, 2, ..., n$, can be used to represent the structures of A and B, respectively. For proteins, $\{V_i\}$ and $\{W_i\}$ can be normalized to "unit vector" since their lengths have almost the same value of $3.8\dot{A}$.

$$T_i = (t_{i1}, t_{i2}, t_{i3}) = \frac{V_i}{\|V_i\|}, \quad U_i = (u_{i1}, u_{i2}, u_{i3}) = \frac{W_i}{\|W_i\|}$$

where $\|V_i\| = \|W_i\| \approx 3.8$. The URMSD of structure A and B is defined as:

$$\text{URMSD}(A,B) = \text{RMS}(\{t_{i1}, t_{i2}, t_{i3}\}, \{u_{i1}, u_{i2}, u_{i3}\}; i = 1, 2, \ldots n-1)$$

The above formula shows that URMSD is computed in the same way as the RMSD, but it uses different inputs, i.e. unit vectors. Therefore, URMSD is defined as follows:

$$\text{URMSD}(A,B) = \sqrt{\frac{2n - 2 * \text{trace}(svd\ C)}{n}}$$

where C is the autocorrelation matrix between the unit vectors for A and B, *svd C* denotes diagonal matrix in the singular value decomposition of C, and trace denotes the sum of the diagonal terms of the *svd C* matrix.

Figure 7 shows an example calculation of URMSD between the 4-residue segment from Gln28 to Lys31 in chain A of 1F0V and chain E of 1RTA (we follow up on this example later in the text):

Step1: Transform the C_α-atoms coordinates into vectors $\{V_i\}$ and $\{W_i\}$.
Step2: Normalize the vectors to unit vectors $\{T_i\}$ and $\{U_i\}$.
Step3: Calculate the autocorrelation matrix C between the two groups of unite vectors; cell values in column p and row q are defined as $c_{pq} = \sum_{i=1}^{n-1} t_{ip} u_{iq}$.
Step4: Compute singular value decomposition of C denoted as *svd C*.
Step5: Calculate $URMSD(A,B) = \sqrt{\frac{2n-2*\text{trace}(svd\ A)}{n}}$.

URMSD is superior to the RMSD since:

- URMSD is not overly sensitive to the change of the protein's outliers. The structural differences of just a few AAs in a long sequence can induce a large value of RMSD, which is not the case for URMSD. For example, chain A of 1F0V and chain E of 1RTA share a sequence segment ranging between GLN28 and SER 123. The RMSD distance between the two structures equals $9.635\dot{A}$, while the URMSD equals $0.435\dot{A}$. In fact, almost 90% of the two structures are essentially the same, see Figure 8, and larger RMSD is due to the tail of segment in 1FOV which protrudes from the center of the structure and forms an outlier. At the same time, two structures of about 100 AAs are assumed different when the corresponding RMSD is about $9\dot{A}$. Therefore, RMSD

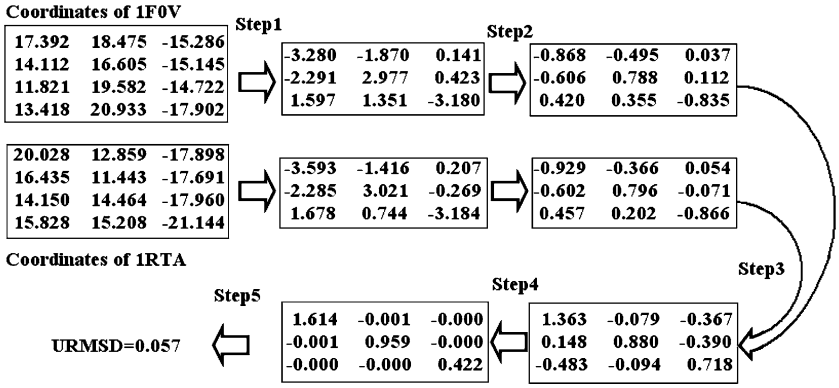

Fig. 7. An example calculation of the URMSD from coordinates for 4-residue segment from Gln28 to Lys31 in chain A of 1F0V and chain E of 1RTA.

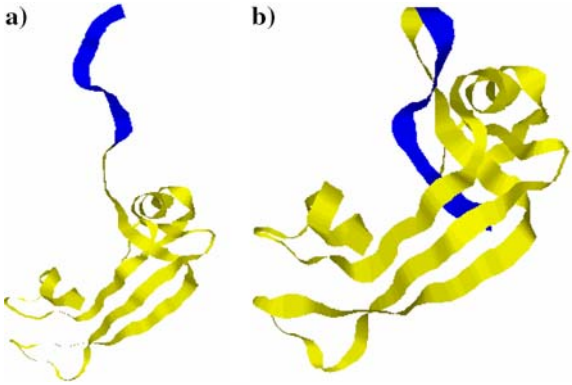

Fig. 8. The GLN28 to SER123 segment from (a) chain A of 1F0V and (b) chain E of 1RTA. The yellow segment (GLN28 to ASN113) corresponds to common structure in the two chains, while the blue segment (PRO114 to SER 123) correspond to the different structure, which results in RMSD $= 9.635\dot{A}$. The common, yellow segment covers 89.6% of the two structures.

should not be used to compare structures that have outliers.

- URMSD weighs all portions of the protein equally, while in RMSD portions far from the center of mass are weighted more heavily.
- RMSD usually increases with the length of the segment while URMSD has an upper bound of $2.0\dot{A}$ and is not sensitive to the increase of the length of the segment. For every pair of random unit sequences with the same length n, the expected upper bound is about $\sqrt{2-\frac{2.82}{\sqrt{n}}}$.
- Finally, URMSD has the same computational complexity as the RMSD.

3.2. The Threshold

RMSD is a popular measure used to determine structural similarity. It is generally believed that two structures can be regarded as the same if their RMSD is less than $3\dot{A}$. On the other hand, URMSD is not as popular, but we believe that it is more appropriate in case of our investigation. In a recent study, Yona and Kedem suggested URMSD threshold equal to 0.6 to differentiate between similar and dissimilar structures for 8-residue segments (Yona and Kedem, 2005). Using thresholds of 0.6 and 0.5, more than 80% and 85% of the fragment pairs from the sponge set will be assumed to have the same structure, respectively. Although there is no direct mapping that would transfer a RMSD threshold into the corresponding URMSD value, we show that the URMSD threshold of 0.5 should be used by comparing it to using RMSD of $3\dot{A}$.

Assuming that a given sequence segment has n different 3D-structures we need to compute URMSD for each pair of the different 3D-structures. Therefore, $n(n-1)/2$ URMSD values must be computed to consider all pairs and the maximal URMSD is the maximum among the $n(n-1)/2$ URMSD values. Among 9297 segments from the sponge set, 8486 segments have maximal URMSD $\leq 0.5\dot{A}$ and the remaining 811 segments have maximal URMSD$>0.5\dot{A}$. The distribution of the segments as a function of the maximal URMSD is illustrated in Fig. 9.

One hundred and forty seven of the 9297 segments from the sponge set have maximal URMSD < 0.5 and corresponding maximal RMSD $> 3\dot{A}$. Thus, only about 1.58% of the segments would be incorrectly classified as structurally similar based on URMSD, when compared to the classification using RMSD. At the same time, 257 of the 9297 segments have maximal URMSD > 0.5 and maximal RMSD $< 3\dot{A}$. Again, only about 2.76% of the segments that URMSD would classify as dissimilar are in fact similar according to the RMSD. Similarly, for the 8,486 segments we found that only about 1.78% of them have maximal RMSDs $< 3\dot{A}$. The maximal RMSDs for most of the remaining 811 segments are greater than $3\dot{A}$, which is shown in detail later in the paper. Assuming that 811 segments are in fact structurally dissimilar, the probability of an error due to using URMSD instead of RMSD can be estimated as $1.78 \times \frac{8486}{9297} = 1.62\%$. The above discussion implies that 0.5 is a suitable threshold for URMSD to decide if a given segment is structurally conserved.

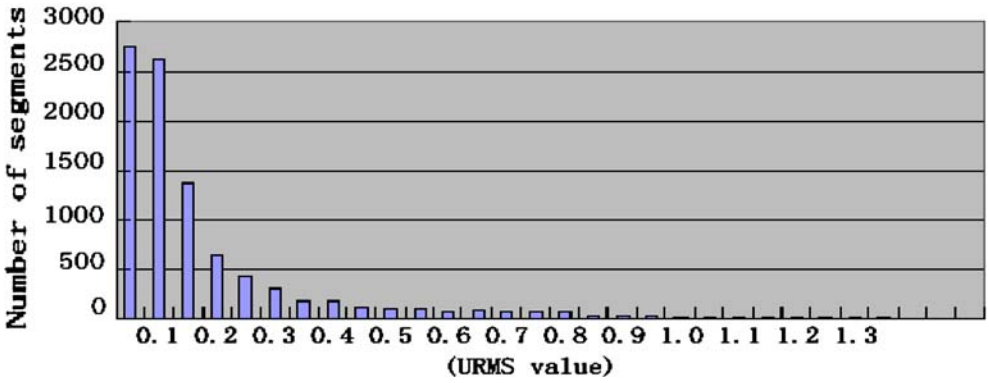

Fig. 9. The distribution of the maximal URMSD distances of the 9297 segments from the sponge set.

Generally, RMSD increases with the increase of segment's length. For segments with more than 200 residues, $3\dot{A}$ threshold is too strict to detect similarity between two structures. The RMSD threshold can reach even $6\dot{A}$ value (Kihara and Skolnick, 2003; Reva *et al*. 1998). At the same time, as argued above, the URMSD value is not sensitive with respect to the sequence length, and thus a single threshold value can be used for sequences of different length. To further illustrate this fact we computed average RMSD value for sequence of varying length, for which URMSD equals 0.5:

- for segments of 150–160 residues, URMSD of 0.5 on average equals RMSD of $3\dot{A}$
- for segments of about 200 residues, URMSD of 0.5 on average equals RMSD of $3.7\dot{A}$
- for longer segments, we do not have sufficient number of samples to calculate the average, but for some segments of about 400 residues that have a maximal URMSD of about 0.5, the corresponding RMSD equals about $5\dot{A}$– $6\dot{A}$.

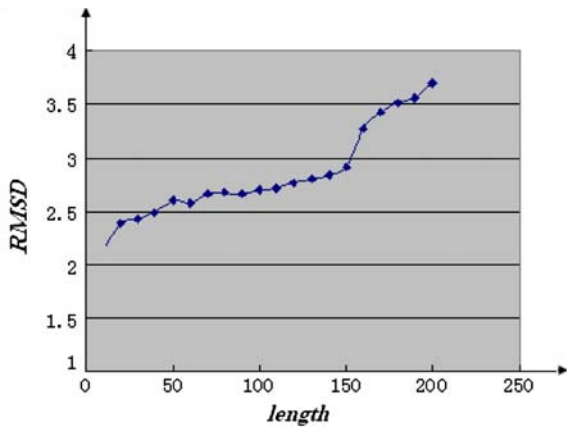

Fig. 10. Average RMSD value for sequences for which URMSD equals 0.5. The value of the RMSD increases with the increase of the length of the segments.

The detailed relation between sequence of different length for which URMSD equals 0.5 and the corresponding value of RMSD is shown in Fig. 10. This demonstrates that in contrast to RMSD, the fixed URMSD threshold can be successfully used to decide about structural conservation for sequences of varying length.

3.3. Construction of the Indicator Sequence for a Pair of 3D Structures

We notice that URMSD sometimes is not sensitive to small changes; conversely it sometimes overestimates the effect of minor local changes. For example, in the segments that lose either a helix or a sheet, the URMSD is not sensitive enough to detect that. On the other hand, in rotating segments, common minor changes in the hinge may result in a large URMSD. To repair this shortcoming, we introduce an indicator sequence that reflects the local change more reliably. We divide one segment into smaller pieces of fixed length so that URMSD is sensitive enough to catch the local changes taking place at more than two residues simultaneously being able to overcome the overestimate of the effects of rotation. The fixed length is selected based on the following facts.

Following Lemma 2 (see (Shen *et al*., 2004), p. 9), for a pair of five-unit-vectors, we state that

- If two unit vectors are different, then the expected URMSD is $\sqrt{2-\frac{2.82}{\sqrt{2}}} \approx 0\dot{A}$.

- If three vectors are different, the expected URMSD is $\sqrt{2-\frac{2.82}{\sqrt{3}}} = 0.61\dot{A}$.
- If four vectors are different, then the expected URMSD is $\sqrt{2-\frac{2.82}{\sqrt{4}}} = 0.77\dot{A}$.
- If all five vectors are random variables, then the expected URMSD is $\sqrt{2-\frac{2.82}{\sqrt{5}}} = 0.86\dot{A}$.

Therefore, the threshold of $0.5\dot{A}$ is sensitive enough to catch these changes at each location where at least three vectors in each five-unit-vector are changed. On the other hand, for every pair of 3D-structures corresponding to a multi-structural segment, we have a pair of C_α-atom sequences.

- If 2 coordinates of the C_α-atoms are different, then at least 3 unit vectors in the pair of the corresponding unit vector sequences are different.
- If 3 coordinates of the C_α-atoms are different, then at least 4 unit vectors are different.
- If 4 coordinates of the C_α-atoms are different, then at least 5 unit vectors are different.

Since the average length of a strand is about 2.3 AAs and the average length of a helix is 3.6 AAs, it implies that this will result in at least three vectors being changed in a five-unit-vector. Therefore, $0.5\dot{A}$ used as the threshold for the URMSD offers sufficient sensitivity to catch these changes for a strand or a helix but it will allow only one coordinate of a C_α-atom to be given incorrectly. Thus, choosing the fixed length of 6 AAs is appropriate.

In general, a segment that consists of N residues has $N-6$ six-residue pieces:

$$1\text{st to } 6\text{th } C_\alpha, 2\text{nd to } 7\text{th } C_\alpha, \ldots, \text{and } N-5\text{th to } N\text{th } C_\alpha.$$

They correspond to $N-6$ five-unit-vectors. If more than two coordinates of C_α-atoms are different in every pair of 6-residue pieces, then in the corresponding pair of five-unit-vectors at least three vectors are not the same. This difference will be detected by URMSD with the threshold of $0.5\dot{A}$.

A piece is called active if its maximal URMSD is greater than $0.5\dot{A}$, and we assign its corresponding region a label 'a'. Otherwise, we say it is conserved, and assign its corresponding region a label 'c'. Of course, the state of the last 4 portions corresponding to the front five-unit-vector will be covered by the state of the first 4 portions corresponding to the next five-unit-vector. Finally, we have an indicator sequence that indicates the state of activity or conservation, which is illustrated in Fig. 11.

If for a given residue the result is a, we say the nth residue is "a"-type, then there is at least one pair of five-unit vectors, say, $v_n, v_{n+1}, \ldots, v_{n+4}$ and $w_n, w_{n+1}, \ldots, w_{n+4}$ such that the URMSD between $v_n, v_{n+1}, \ldots, v_{n+4}$ and $w_n, w_{n+1}, \ldots, w_{n+4}$ is greater than $0.5\dot{A}$. Using the coordinates of C_α-atoms to describe this difference, the RMSD between the coordinates of C_α-atoms, say, $C_{\alpha,n-1}$, $C_{\alpha,n}$, ..., $C_{\alpha,n+4}$ and $Cv'_{\alpha,n-1}$, $Cv'_{\alpha,n}$, ..., $Cv'_{\alpha,n+4}$, would be greater than $3\dot{A}$. If the nth residue is "a"-type alone, then it means the difference between $C_{\alpha,n-1}$, $C_{\alpha,n}$ and $Cv'_{\alpha,n-1}$, $Cv'_{\alpha,n}$ is very large, but that difference between

$$C_{\alpha,n+1}, \ldots, C_{\alpha,n+4}, \ldots \text{ and } C'_{\alpha,n+1}, \ldots, C'_{\alpha,n+4}, \ldots$$

is small. In general, there is no "a"-type residue standing alone in a sequence.

3.4. Classification of the Benchmark Set

Combining with the indicator sequence, we classify the segments in the benchmark set as follows. We regard a segment as

1. *invariant*, if it has a unique 3D-structure.
2. *absolutely conserved*, if its indicator sequence consists purely of 'c'-type residues.
3. *conserved*, if its maximal URMSD is less than $0.5\dot{A}$.
4. *missing*, if it is conserved but not absolutely conserved.
5. *rotating*, if it is not conserved and there is a small region in the middle of the segment, in which, the indicator sequence is of 'a'-type.
6. *disarranging*, if it not conserved and there is at least one bigger region, in which, almost all the indicator states are of 'a'-type, but if we allow a shift, then the number of 'a'-type residues is lower.

Primary sequence	E I P P R S E K A N V L Q I A L Q T K I K L F Y R P A A I K
Indicator sequence	a a a a a a a a a a a a a a a a a c c c c c c c c c c c c c c

Fig. 11. Example indicator sequence; the first row is the sequence of AAs represented by single-letters. The second row is the indicator sequence for conservation ('a' means active and 'c' means conserved).

Note that the definitions of rotating, missing and disarranging segments are the same as those given in Section 2.2. However, this definition does not cover all segments in the benchmark set. Typically, for all 9290 segments in the sponge set studied, there are 652 segments that cannot be included in any of the above six classes due to the fact that $0.5\dot{A}$ gives too strict a condition. However, if we raise the threshold from $0.5\dot{A}$ to $0.6\dot{A}$ or $0.7\dot{A}$, then the 645 segments will uniquely belong to the "generalized conservation class". This is discussed in Section 4.3.

4. RESULTS

The above classification was applied to the entire benchmark set and the results are summarized in this section. We start by clarifying the relation between the missing, rotating and disarranging segments and the segment conservation, and next we estimate number of segments in the benchmark set that are not conserved.

It is clear that an invariant segment must be absolutely conserved. In the Appendix 1 a mathematical proof is given that an absolutely conserved segment must be conserved. That is, for a segment, if its indicator sequence is purely a "c"-sequence, then its maximal URMSD must be less than $0.5\dot{A}$. But the inverse is not true. In fact, we have pointed out that 8486 segments in the sponge set are conserved. The remaining 811 segments include the rotating, disarranging and other 645 segments (see Sections 4.1 and 4.2). However, amongst the 8486 segments, only 6222 segments are absolutely conserved, and 2264 segments are not absolutely conserved.

Intuitively, the missing segments should belong to the set consisting of all conserved segments because the operation of segment removal does not change the shape of the corresponding 3D-structure. But these must not be absolutely conserved segments because the missing helices or strands surely result in the corresponding five-unit-vector being changed largely since the average length of the strands is 2.3 AAs while that of the helices 3.6 AAs. So, the indicator sequence will catch the missing ones. Since the missing ones mainly affect the secondary structure of the protein rather than its 3D-shape, we can ignore this case if we predict the coordinates of the C_α-atoms.

It is clear that a rotating segment should not be conserved because rotation causes the hinge to change, which usually increases the URMSD of the whole segment by more than $0.5\dot{A}$ while the indicator sequence in the hinge region is also greater than $0.5\dot{A}$ because every five-unit-vector in this region changes significantly. For the disarranging segment, both the global URMSD and local URMSD are larger unless we compare them accounting for a shift. Hence the AAs are active in the indicator sequence and almost all of them are of the 'a' type. However, in this region, if we permit each piece to shift freely, every piece will find its best match.

4.1. Estimating the Number of Rotating Segments

Rotating segments are those segments, in which backbones in the hinge regions are rotated, while the domains are structure-conserving. These rotations will make the URMSD of a whole segment large and its indicator sequence will be of 'a'-type in the middle region corresponding to a hinge. On the one hand, a rotating segment usually has a large URMSD as a whole, which is usually larger than $0.8\dot{A}$. For example, the longest rotating segment is GLN238 to GLU993 as found in chain A of 1EUL (Toyoshima *et al.*, 2000) and chain B of 1IWO (Toyoshima and Nomura, 2002). Figure 12(a, b) shows the conformations of the same segment GLY88 to ARG537 found in chain A and chain B of 1UAA (Korolev *et al.*, 1997); the URMSD is $0.84\dot{A}$. On the other hand, the rate defined as the length of hinge regions divided by the length of the segment, is much smaller than 20% AAs. Moreover, all 'a'-type sequences should be located in the middle region.

We can estimate the upper bound on the total number of rotating segments using the following procedure:

- First, calculate the URMSD between every pair of 3D-structures of one n-length segment and denote its maximal URMSD by D.
- Second, make the indicator sequence of the segment.
- Third, if $D > 0.5$ and the rate of 'a'-type sequences in its indicator sequence is less than 20%, we regard this segment as a possible rotating segment.

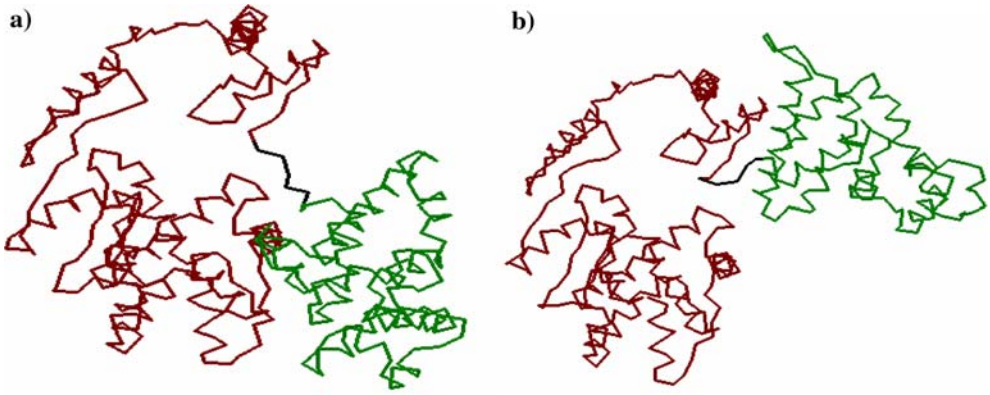

Fig. 12. The 3D-structures of the same segment occurring in (a) chain A of 1UAA, and (b) chain B of 1UAA; the segments consist of two structure-conserved parts (residues 88–371 in red and residues 378–537 in green), and link part (residues 373–377 in black).

After checking the segments acquired by the above steps, among all of the 9290 segments, only 66 segments have been found to be rotating segments. That is, at most 0.71% of the sponge set would be rotating segments.

4.2. Estimating the Number of Disarranging Segments

For most of these segments, amongst all of its 3D-comformations, the maximal similarity is not as good as for the conserved segments. Moreover, in its indicator sequence, the rate of 'a'-type sequences is very high. To detect disarranging characteristics, we use a method that is similar to the cycle-convolution used in Signal Processing. We let $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ be unit vector sequences that are calculated from a pair of 3D-structures of a segment, for a shift $l = 0, 1, 2, \ldots, n-1$. We then have orderly permutation sequences induced by $v_1, \ldots, v_n$ as shown in Table 1.

Then, $z = (z_1, \ldots, z_n)$, where $z_l = \sum_{m=1}^{n} u_m \bullet \pi_{l-1}(\{v_k\})(m)$ and $u_m \bullet \pi_{l-1}(\{v_k\})m)$ is the point product of two unit vectors. We formally say it is the cycle-convolution of $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$. If there is an $l > 1$ such that $|z_l| > |z_1|$, we conclude that the segment is a disarranging segment. In short, the procedure is as follows:

- If there is a pair of unit vector sequences $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ such that the URMSD greater than $0.5\dot{A}$
- If the indicator sequence produced by $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ has a region that the rate of 'a'-type residues is greater than 50%.
- In the region, the corresponding pair of unit vector sub-sequences is given by $u_m, \ldots, u_{m+p}$ and $v_m, \ldots, v_{m+p}$, If there is a shift $l > 1$ such that the similarity coefficient between $\{v_i(l)\}$ and $u_1, \ldots, u_n$ is strictly larger than that between $\{v_i(0)\}$ and $u_1, \ldots, u_n$.

Then we regard this segment as a disarranging segment.

Using this procedure, we obtained 93 segments among all of the 9290 segments as possible disarranging segments, which is about 1% of the size of the sponge set.

Table 1. Permutation Sequences Where $\pi_l(\{v_k\})(m)$ Corresponds to the Unit Vector in the ith row and mth Vector

Shift	Abbreviation	Complete form
$l = 0$	$\pi_0(\{v_i\})$	$v_1, v_2, v_3, \ldots, v_n$
$l = 1$	$\pi_1(\{v_i\})$	$v_2, v_3, v_4, \ldots, v_n, v_1$
$l = 2$	$\pi_2(\{v_i\})$	$v_3, v_4, v_5, \ldots, v_n, v_1, v_2$
...	...	...
$l = n-1$	$\pi_{n-1}(\{v_i\})$	$v_n, v_1, v_2, v_3, v_4, v_5, \ldots, v_{n-1}$

4.3. Analysis of the Remaining 652 Segments

The remaining 652 segments, which cannot be included in any of the six classes defined in the Section 3.4 due to the fact that $0.5\dot{A}$ gives too strict a condition, have been checked manually and the result is as follows:

- 3D-structures of each segment within the 645 segment set are almost the same, even though its whole URMSD is greater than $0.5\dot{A}$ (the corresponding RMSD may be

Table 2. The Statistical Correspondence between URMSD and RMSD for the Set of 652 Segments

The set for which URMSD < the given threshold	0.6Å	0.7Å	0.8Å
The mean RMSD for the given set	2.913Å	3.405Å	4.696Å
The maximum RMSD for the given set	11.065Å	15.366Å	19.851Å
The number the segments that RMSD > 4Å	20	32	36

greater than 3Å). Since the threshold (3Å) in terms of the RMSD is too strict, it is relaxed somewhat in practice. There is a statistical relationship between the URMSD and RMSD for the 652 segment as can be seen in Table 2. If we relax the threshold slightly, then most of the 645 segments will be classified as conserved segments, which agrees with our manual inspection.

- There are 7 segments: 1FJI, 3ITS, 1DL9, 1LQN, 1ITN, 1B29 and 1B61 (see Section 2.2) which are exceptions, each of which has a pair of 3D-structures that are quite different. However, all of these exceptions are due to one given by a theoretical model and the other given by an experimental method. So we suspect that these exceptions are not reliable and ignore them.

The missing segments are included in the set that has conserved but not absolutely conserved sequences, and their total number is less than 2264. Since the 3D-configurations of each missing segment show no significant differences, we do not present it as an independent subsection.

5. SUMMARY AND CONCLUSIONS

Below, we give the summary of our findings as follows:

- For a possible multi-structural segment that belongs to one of the invariant segments, absolutely conserved segments, or conserved segments, its 3D structures always seem to be unique structurally.
- For each segment that is not conserved and also not the rotating and the disarranging segment, its 3D-structures have very small structural differences. Therefore these segments can be regarded as structurally conserved.
- Only for all rotating segments and disarranging segments, their maximal RMSD may be significant indicating variability of their structure. However, the number of the rotating segments divided by the size of the sponge set that is less than 0.71%, and the same number for the disarranging segments is less than 1%. Furthermore, the total number of all possible rotating segments and disarranging segments divided by the number of sequences in the benchmark set, which represents sequences in PDB, is less than 0.085%.

The ''uniqueness'' (conservation) of the 3D-structures of segments is a strong property that allows to predict 3D-structures based on alignment of the primary sequences given that sufficient level of sequence homology is present. At the same time, 1.71% of segments, which account for the rotating and disarranging cases, exhibit variability in the structure despite having identical sequences. These segments may results in inaccuracies for the sequences alignment based methods for both 3D-structure and secondary structure predictions. We also conclude that the prediction accuracy of the secondary structure will suffer bigger loses in comparison with the 3D structure. The 2264 conserved but not absolutely conserved segments, which constitute about 25% of the sponge set, may lead to wrong secondary structure prediction result.

The indicator sequence, which is proposed in this paper, plays an important role in distinguishing rotating, missing and disarranging segments. This knowledge can be used to find the active and the conserved regions for a segment. The absolutely conserved segments of length more than 10 residues form the most useful class, in which each segment is conserved if we ignore the change of the coordinate of a single C_α-atom per 6-residue segment. The mathematical proof given in the Appendix 1 shows that for any multi-structure segment of length greater than 10, if it is an absolutely conserved segment, then it must be a conserved segment. The inverse is not true.

The sponge set S_p defined in this paper is an unexpected result. All chains in the PDB database can be covered by the 17,494 segments drawn from the sponge set consisting of the 9297 protein segments and the signature set that includes 8197 protein chains. This shows that space of valid AA sequences that can be used to assemble the structural information is significantly smaller than the source protein set of 53,000 collected from PDB and the set of all possible AA

sequences, i.e., assuming the segment length of 10, which was the minimum length of segments in the sponge set, there are $20^{10} = 10,240,000,000,000$ possible sequences.

APPENDIX 1

For proving that an absolutely conserved segment must be a conserved segment, we first consider the definition of *URMSD*. We need to prove the following statement mathematically.

Let

$$d(\{v_i\},\{w_i\}) = \min_{\phi}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}||v_i - \phi(w_i)||^2}\right\}$$

be the URMSD between the two unit vector sequences $\{v_i\}_{i=1}^n$ and $\{w_i\}_{i=1}^n$, and let d_i be the URMSD between the pair of $v_{i+1},\ldots,v_{i+5}$ and $\phi(w_{i+1}),\ldots,\phi(w_{i+5})$.

Then $d(\{v_i\},\{w_i\}) \le \max\{d_i \mid i = 0,1,2,\ldots, n-5\}$ for all $n \ge 10$.

Proof. It is easily followed that

$$\begin{aligned}
&\min_{\phi}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}||v_i - \phi(w_i)||^2}\right\}\\
&\overset{y_i=\phi(w_i)}{=}\min_{\{\phi\}}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}||v_i - \phi(w_i)||^2}\right\}\\
&=\min_{\{\phi\}}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}||v_i||^2 + ||\phi(w_i)||^2 - 2(v_i,\phi(w_i))}\right\}\\
&=\min_{\{\phi\}}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}[2 - 2(v_i,\phi(w_i))]}\right\}\\
&=\min_{\{\phi\}}\left\{\sqrt{2 - \frac{2}{n}\sum_{i=1}^{n}(v_i,\phi(w_i))}\right\}
\end{aligned}$$

We can regard that $\sum_{i=1}^{n}(v_i,\phi(w_i))$ as the trace of the correlation matrix $R(n)$, where

$$R(n,\phi) = \begin{pmatrix} (v_1,\phi(w_1)) & (v_1,\phi(w_2)) & \ldots & (v_1,\phi(w_n)) \\ (v_2,\phi(w_1)) & (v_2,\phi(w_2)) & \ldots & (v_2,\phi(w_n)) \\ & \ldots\ \ldots & & \\ (v_n,\phi(w_1)) & (v_n,\phi(w_2)) & \ldots & (v_n,\phi(w_n)) \end{pmatrix}.$$

That is, we have

$$d(\{v_i\},\{w_i\}) = \min_{\{\phi\}}\sqrt{\frac{2n - 2\times\mathrm{trace}(R(n,\phi))}{n}}$$

For a fixed ϕ and every pair of five-unit-vector $v_{i+1},\ldots,v_{i+5}$ and $\phi(w_{i+1}),\ldots,\phi(w_{i+5})$, we have a correlation matrix

$$R_i(5) = \begin{pmatrix} (v_{i+1},\phi(w_{i+1}))(v_{i+1},\phi(w_{i+2}))\ldots(v_{i+1},\phi(w_{i+5})) \\ (v_{i+2},\phi(w_{i+1}))(v_{i+2},\phi(w_{i+2}))\ldots(v_{i+2},\phi(w_{i+5})) \\ \ldots\ldots\ldots \\ (v_{i+5},\phi(w_{i+1}))(v_{i+5},\phi(w_{i+2}))\ldots(v_{i+5},\phi(w_{i+5})) \end{pmatrix}$$

Let $d_i = \sqrt{\mathrm{trace}(R_i(5)'R_i(5))}$ for $i = 0,1,2,\ldots,n-5$. For convenience, then we may assume $d_0 = \max\{d_i \mid i = 0,1,2,\ldots,n-5\}$, then

$$\sqrt{\sum_{j=1}^{5}\sum_{j=1}^{5}(v_i,\phi(w_j))^2} \ge \sqrt{\sum_{j=1}^{5}\sum_{j=1}^{5}(v_{i+k},\phi(w_{j+k}))^2}$$

for all $k \ge 1$.

Considering the relationship

$$\begin{aligned}
\sqrt{\mathrm{tr}(R(n)'R(n))} &= \sqrt{\sum_{j=1}^{n}\sum_{i=1}^{n}(v_i,\phi(w_j))^2} = \sqrt{\sum_{j=1}^{n}(\sum_{i=1}^{5}(v_i,\phi(w_j))^2 + \sum_{i=5}^{n}(v_i,\phi(w_j))^2)}\\
&= \sqrt{\sum_{j=1}^{5}\sum_{i=1}^{5}(v_i,\phi(w_j))^2 + \sum_{j=6}^{n}\sum_{i=1}^{5}(v_i,\phi(w_j))^2 + \sum_{j=1}^{5}\sum_{i=6}^{n}(v_i,\phi(w_j))^2 + \sum_{j=6}^{n}\sum_{i=6}^{n}(v_i,\phi(w_j))^2}
\end{aligned}$$

By ordinary fact of structure of protein: for most AAs, the state at site i is not more frequently correlated to the state at site j if the distance between the two sites is greater than 5 AAs. That is, we may assume that $R(n)'R(n)$ has the following relations mathematically:

- The number of the set $\left\{k \mid \sum_{j=1}^{5}\sum_{i=1}^{5}(v_i,\phi(w_j))^2 < \sum_{j=1}^{5}\sum_{i=1}^{5}(v_{i+k},\phi(w_j))^2\right\}$ related to $n-5$ is very small.
- The number of the set $\left\{k \mid \sum_{i=1}^{5}\sum_{j=1}^{5}(v_i,\phi(w_j))^2 < \sum_{i=1}^{5}\sum_{j=1}^{5}(v_i,\phi(w_{j+k}))^2\right\}$ is also very small related to $n-5$.
- $(v_i,\phi(w_j))^2 > (v_{i+k},\phi(w_j))^2$ and $(v_i,\phi(w_j))^2 > (v_i,\phi(w_{j+k}))^2$ for all $i,j \le 5$ and almost all $k > 6$.

Then let $y_i = \phi(w_i)$, we have

- $\sum_{j=1}^{5}\sum_{i=1}^{5}(v_i,y_j)^2 \ge \sum_{j=5}^{5}\sum_{i=1}^{5}(v_{i+k},y_j)^2$,
- $\sum_{i=1}^{5}\sum_{j=1}^{5}(v_i,y_j)^2 \ge \sum_{i=1}^{5}\sum_{j=1}^{5}(v_i,y_{j+k})^2$
- $\sum_{i=1}^{5}\sum_{j=1}^{5}(v_i,y_j)^2 \ge \sum_{i=1}^{5}\sum_{j=1}^{5}(v_{i+k},y_{j+k})^2$

for all $k > 6$. Without lost the generality, we may assume that $n = 0 \pmod 5$, and then we have

That is, we have proved that

$$\frac{\sqrt{\sigma_1(n)^2+\sigma_2(n)^2+\cdots+\sigma_n(n)^2}}{n} < \frac{\sqrt{\sigma_1(5)^2+\sigma_2(5)^2+\cdots+\sigma_5(5)^2}}{5} \quad \text{if } n \ge 10.$$

where $\sigma_i(j)$ for $j = 5, n$ and $i \le j$, are the singular value of $R(5)$ and $R(n)$ respectively. Replacing $R(n)$ and $R(5)$ by their "squared root":

$$R^{\frac{1}{2}}(n) = \begin{pmatrix} \sqrt{|(v_1,y_1)|} & \sqrt{|(v_1,y_2)|} & \cdots & \sqrt{|(v_1,y_n)|} \\ \sqrt{|(v_2,y_1)|} & \sqrt{|(v_2,y_2)|} & \cdots & \sqrt{|(v_2,y_n)|} \\ & & \cdots\ \cdots & \\ \sqrt{|(v_n,y_1)|} & \sqrt{|(v_n,y_2)|} & \cdots & \sqrt{|(v_n,y_n)|} \end{pmatrix}$$

and with the same argument, we have

$$\frac{\sigma_1(n)+\sigma_2(n)+\cdots+\sigma_n(n)}{n} < \frac{\sigma_1(5)+\sigma_2(5)+\cdots+\sigma_5(5)}{5} \quad \text{if } n \ge 10.$$

That is, $\frac{\text{trace}(\text{svd } R(5))}{5} > \frac{\text{trace}(\text{svd } R(n))}{n}$ for $n \ge 10$.

Therefore, the maximal URMD among the all five-unit-vectors is greater than the URMSD for the whole segment. This ends the proof.

$$\begin{aligned}
\tfrac{1}{n}\sqrt{\text{tr}(R(n)'R(n))} &= \sqrt{\tfrac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(v_i,y_j)^2} \\
&= \sqrt{\tfrac{1}{n^2}\left(\sum_{j=1}^{5}\sum_{i=1}^{5}(v_i,y_j)^2+\sum_{j=6}^{n}\sum_{i=1}^{5}(v_i,y_j)^2+\sum_{j=1}^{5}\sum_{i=6}^{n}(v_i,y_j)^2+\sum_{j=6}^{n}\sum_{i=6}^{n}(v_i,y_j)^2\right)} \\
&= \sqrt{\tfrac{1}{n^2}\left(\tfrac{n}{5}\sum_{j=1}^{5}\sum_{i=1}^{5}(v_i,y_j)^2+\sum_{j=6}^{n}\sum_{i=1}^{5}(v_i,y_j)^2+\sum_{j=1}^{5}\sum_{i=6}^{n}(v_i,y_j)^2\right)} \\
&\le \sqrt{\tfrac{1}{n^2}\left(\tfrac{n}{5}\sum_{j=1}^{5}\sum_{i=1}^{5}(v_i,y_j)^2+\tfrac{n-5}{5}\sum_{i=1}^{5}(v_i,y_j)^2+\tfrac{n-5}{5}\sum_{j=1}^{5}(v_i,y_j)^2\right)} \\
&\le \sqrt{\tfrac{1}{n^2}\left(\tfrac{n}{5}\sum_{j=1}^{5}\sum_{i=1}^{5}(v_i,y_j)^2+\tfrac{n-5}{5}\sum_{i=1}^{5}\sum_{j=1}^{5}(v_i,y_j)^2\right)} \\
&= \sqrt{\tfrac{2n-5}{5n^2}\left(\sum_{j=1}^{5}\sum_{i=1}^{5}(v_i,y_j)^2\right)} < \sqrt{\tfrac{2}{5n}\left(\sum_{j=1}^{5}\sum_{i=1}^{5}(v_i,y_j)^2\right)} \; hop \overset{n\ge 10}{\le} \tfrac{1}{5}\sqrt{\text{trace}(R(5)'R(5))}
\end{aligned}$$

REFERENCES

Anfinsen, C. B. (1973) *Science* **81:** 223–233.

Barrientos, L. G., Louis, J. M., Botos, I., Mori, T., Han, Z., O'Keefe, B. R., Boyd, M. R., Wlodawer, A., and Gronenborn, A. M. (2002). *Structure* **10:**(5), 673–686.

Bamborough, P., Duncan, D., and Richards, W. G. (1994). *Protein Eng*. **7:**(9), 1077–1082.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). *Nucleic Acids Res*. **28:** 235–242.

Brody, S. S., Gough, S. P., and Kannangara, C. G. (1999). *Proteins* **37:**(3), 485–493.

Chen, K., Ruan, J., and Kurgan, L. A. (2006). *The Protein J*. **25:**(1), 57–70.

Chew, L. P., Huttenlocher, D., Kedem, K., and Kleinberg, J. (1999). *J. Comput. Biol*. **6:**(3–4), 313–325.

Ding, J., Das, K., Hsiou, Y., Sarafianos, S. G., Clark, A. D., Jacobo-Molina, A., Tantillo, C., Hughes, S. H., and Arnold, E. (1998). *J. Mol. Biol*. **284:**(4), 1095–1111.

Drum, C. L., Yan, S.-Z., Bard, J., Shen, Y.-Q., Lu, D., Soelaiman, S., Grabarek, Z., Bohm, A., and Tang, W. J. (2002). *Nature* **415:** 396–402.

Elshorst, B., Hennig, M., Forsterling, H., Diener, A., Maurer, M., Schulte, P., Schwalbe, H., Griesinger, C., Krebs, J., Schmid, H., Vorherr, T., and Carafoli, E. (1999). *Biochemistry* **38:**(38), 12320–12332.

Falzone, C. J., Wang, Y., Vu, B. C., Scott, N. L., Bhattacharya, S., and Lecomte, J. T. (2001). *Biochemistry* **40:** 4879–4891.

Hansson, M., Gough, S. P., and Brody, S. S. (1997). *Proteins* **27:**(4), 517–522.

Kabsch, W. (1978) *Acta Crystallogr*. **A34:** 827–828.

Kihara, D., and Skolnick, J. (2003). *J. Mol. Biol*. **334:** 793–802.

Korolev, S., Hsieh, J., Gauss, G. H., Lohman, T. M., and Waksman, G. (1997). *Cell* **90:**(4), 635–647.

Lindberg, J., Sigurdsson, S., Lowgren, S., Andersson, H. O., Sahlberg, C., Noreen, R., Fridborg, K., Zhang, H., and Unge, T. (2002). *Eur. J. Biochem*. **269:**(6), 1670–1677.

Meador, W. E., Means, A. R., and Quiocho, F. A. (1992). *Science* **257:**(5074), 1251–1255.

Reva, B. A., Finkelstein, A. V., and Skolnick, J. (1998). *Fold Des*. **3:**(2), 141–147.

Schumacher, M. A., Crum, M., and Miller, M. C. (2004). *Structure (Camb)* **12:**(5), 849–860.

Shen, S. Y., Yu, T., Kai, B., Ruan, J. S. (2004). J. Eng. Math. **21:**(6), 862–870 (in Chinese).

Tiraboschi, G., Jullian, N., Thery, V., Antonczak, S., Fournie-Zaluski, M. C., and Roques, B. P. (1999). *Protein Eng*. **12:**(2), 141–149.

Toyoshima, C., Nakasako, M., Nomura, H., and Ogawa, H. (2000). *Nature* **405:**(6787), 647–655.

Toyoshima, C., and Nomura, H. (2002). *Nature* **418:**(6898), 605–611.

Veerapandian, B. (1992) *Biophys. J*. **62:**(1), 112–115.

Xu, C., Rice, W. J., He, W., and Stokes, D. L. (2002). *J. Mol. Biol*. **316:**(1), 201–211.

Yap, K. L., Yuan, T., Mal, T. K., Vogel, H. J., and Ikura, M. (2003). *J. Mol. Biol*. **328:**(1), 193–204.

Yona, G., and Kedem, K. (2005). *J. Comput. Biol*. **12:**(1), 12–32.