Systematic analysis of local flexibility of multiple-structure proteins

Naoto Morikawa

October 25, 2008.

Abstract

It is widely accepted that knowledge of protein flexibility is fundamental for understanding the mechanism of protein function. The native state of a protein is considered to be a dynamic ensemble of conformational substates, where each substate is distinguished by locally unfolded regions that may contain only a few amino acids. In addition, functional changes of proteins are induced by redistribution of the substates. Because conformational changes of proteins are attributable to only a fraction of the residues within the protein, identification of these regions is particularly important for understanding protein dynamics and their function.

In this paper, we identify regions of 158 multiple-structure proteins where conformational changes take place using a discrete differential geometrical technique, D^2 *encoding* . In the first step we examine concrete examples to assess the sensitivity of our method. We start with simple hypothetical fragments, such as loop and extended fragments. We then considered domain-swapped dimers, whose formation mechanism has implications for the development of amyloid plaques observed in misfolding diseases such as Alzheimer's and Parkinson's diseases. We also analyzed several other proteins before performing statistical analysis of 60 crystal structure pairs of the same proteins. Finally, we discuss implications of the results for protein engineering and drug design.

The D^2 *code* of a protein is a base-32 number sequence, where each number represents the conformation of the five-Cα fragment centered on a residue of the protein. Roughly speaking, the base-32 number is the " second derivative " of a five-Cα fragment and the D^2 code is obtained by arranging the base-32 numbers of all the five-Cα fragments in the order of appearance. Due to the sensitivity of the D^2 code to the twisting of a Cα trace, the sources of structural differences are successfully pinpointed by comparison of D^2 codes.

We have found that a multiple-structure protein can be thermodynamically identified with a sequence of rigid subdomains of an average length 14.4 Cα atoms connected by variable regions of an average length 2.0 Cα atoms, where deformation of the variable regions are coupled to induce global structural transitions between two forms of the same proteins. As for the location of variable regions, some places, such as the N terminus of a β-strand and the C-cap of an α-helix, seem to be more favored than others.

In regard to implications of the results for protein engineering and drug design, it is suggested that the number of D^2-variable Cα atoms in the hinge region of a domain-swapped dimer can be a good measure of evolution of the dimer. In addition, the existence of long D^2-variable regions may play a role in conformational changes observed in misfolding diseases. Moreover, analysis of the distribution of D^2-variable regions can lead to a more detailed description of the mechanism of multi-drug resistance due to non-active site mutations. All the programs used and the data obtained are available from `http://www.genocript.com`.

Keywords: *conformational change; variable region; structural difference; local protein structure; domain-swapped dimer; misfolding disease; protein engineering; drug design; D2 code; discrete differential geometry*

Affiliation: GENOCRIPT

Address: 27-22-1015, Sagami-ga-oka 1-chome, Zama-shi, Kanagawa 228-0001 Japan.

Email: nmorika@genocript.com

INTRODUCTION

It is widely accepted that knowledge of protein flexibility is fundamental for understanding the mechanism of protein function. Substantial research is focused on integrating protein flexibility into protein engineering and drug design ([1]–[13]). In this paper, we identify regions of 158 multiple-structure proteins identified by Kosloff and Kolodny14 where conformational changes take place. We discuss implications of the results for protein engineering and drug design. For instance, it is important to analyze how local conformational changes are coupled to induce global structural transitions in order to understand the mechanism of conformational changes observed in misfolding diseases, such as Alzheimer 's, Parkinson's, or mad cow (BSE) diseases [15], as well as the mechanism of drug resistance due to non-active site mutations.

As shown by NMR-detected hydrogen exchange experiments ([16],[17]), the native state of a protein is considered to be a dynamic ensemble of conformational substates, where the population of conformational substates is determined by the network of cooperative interactions within the protein ([18]–[21]). In addition, the function of a protein can be altered by redistribution of the substates ([22],[23]). For example, ligand-binding proteins can adopt two different conformations, ligand-free (open) and ligand-bound (closed) forms, even in the absence of ligand [3]. Ligand binding causes a shift in the distribution of the preexisting protein conformations.

Each substate of the ensemble is distinguished by locally unfolded regions that may contain only a few amino acids. Local unfolding events occur independently of each other, and the cooperativity within a protein is a result of thermodynamic coupling between different regions. That is, two regions are positively coupled if both regions are either folded or unfolded in the most probable ensemble substates. The regions are negatively coupled if one is always folded whenever the other is unfolded and the regions are not coupled if they are folded randomly.

The ensemble-based approach has been successfully used to describe the mechanism of communication between ligand-binding sites and the susceptibility of these binding sites to distal mutations ([24]–[27]). Within the context of the ensemble-based description, proteins use intrinsic local conformational fluctuations to perform their functions, such as catalysis, allosterism, and signal transduction. Fluctuations at binding sites are propagated to remote locations via the network of cooperative interactions between local segments ([28]–[30]). We observe a manifestation of the redistribution of conformational substates that is triggered by the propagation of a fluctuation. For example, allostery is a consequence of the redistribution induced by ligand binding.

A notable implication of the approach is non-uniform propagation of the cooperative interactions throughout the entire protein molecule ([31],[32]). That is, not all amino acids are affected equally by the propagation. The cooperative pathways involve only a fraction of residues within the protein even though interactions could reach regions far away from the triggered site. Residues can be thermodynamically coupled without any visible connection pathway and they could play a significant role in modulating the cooperative network. Therefore, identification and characterization of the affected residues is important for understanding and engineering of protein functions.

Moreover, it should be noted that subtle conformational changes are often essential to protein function. Because proteins change the population of the conformational substates during the course of their biological function, energy barriers for transition between substates should be low to allow a quick reaction [3]. For instance, ligand-binding and catalysis are typically performed on the micro- to milli-second timescale, which implies that the collective motions of the Cα atoms involved in the transitions are in the pico- to nano-second regime. In addition, it is known that proteins can detect a conformational change as small as $1\mathring{A}$. For example, a $1\mathring{A}$ conformational change at the ligand-binding site is propagated to a cytoplasmic activation site located $100\mathring{A}$ away in the aspartate receptor [33].

As mentioned above, conformational changes of proteins are attributable to only a fraction of residues within the protein. Thus, identification of these regions is important for understanding protein dynamics and their function. In this study, we have examined the extent of the local distortion that accounts for the conformational changes induced by various biological activities using pairs of X-ray crystal structures that have been determined for the same protein that contain significant structural differences.

In order to identify local structural differences between X-ray crystallographic coordinates, it is necessary to con-

sider the following two problems. First, we should identify only statistically significant differences by assessing the effect of coordinate errors and distortions due to crystal packing. Second, we should quantify the structural difference between local backbone conformations.

The first problem is clearly demonstrated. As shown by the famous controversy concerning the artificial distortion of myoglobin upon CO binding [34], inaccuracies in crystal structures are troubling [35]. According to Rejto and Freer, 25–30% of a protein surface is in contact with protein molecules belonging to other crystal units [36]. In addition, the coordinate error of Cα atoms at loops and surface regions could become as large as $1.0\mathring{A}$ [37]. Moreover, due to the high solvent content, crystalline proteins are also dynamic and exhibit extensive, discrete conformational substates [38]. Proteins could bind ligands reversibly even in the crystalline form [39]. Currently, most protein crystals are solved in a single conformation, and artifacts such as Ramachandran outliers might be attributed to the heterogeneity.

To quantify the structural difference between local backbone conformations, Flocco and Mowbra[40] proposed a method based on dihedral angles defined by four consecutive Cα atoms and Korn and Rose [41] proposed a similar method based on the backbone ϕ and ψ angles. Both of these methods use cutoff values to remove artifacts introduced during structure determination. Considering the uncertainty of the position of side-chain atoms in crystal structures, it is reasonable to analyze the conformation of Cα-traces or backbones only. On the other hand, Kuznetsov and Rackovsky [42] characterized structurally ambivalent fragments of five amino acids or greater in proteins selected from the PDB database, where they used secondary structure to detect differences between conformations of the same fragment. In their work, secondary structures are computed by the DSSP program [43]. Two distinct conformations are identified if their Cα root mean square deviation (RMSD) is below a certain threshold. To remove poorly characterized fragments, they also used the temperature B-factors, which are essentially determined by spatial variations in local packing density [44] and are not a good predictor of heterogeneity in structures. The B-factors are not reliable predictors of heterogeneity because 30% of side chains can exist in multiple conformations and multiple side chain conformations frequently occur at residues with low B-factor [38].

Other methods have also been used to quantify the structural difference between local backbone conformations. Because similar secondary structure assignments do not guarantee structural similarity and there are often significant variations in peripheral loops, Cohen and coworkers [45] studied structural plasticity of hexapeptide fragments based on the virtual dihedral angle joining four consecutive Cα atoms, using not only secondary structure, but also the backbone RMSD and a structural loop classification [46].

Template-based methods are usually not employed for identification of local structural differences because the template-based approximation of a fragment is not determined uniquely. One of a few examples is the PBE-ALIGN method [47], which uses 16 short structural templates to encode protein structures. They align two template sequences using a derived substitution matrix and simple dynamic programming algorithm. However, their main purpose is large database mining for similar structures and their method is not useful for our purposes because of the uncertainty induced by the substitution matrix. As for flexible structural alignment tools, such as RAPIDO [48], FATCAT [49], and FlexProt [50], they are also not useful for our purposes because they often ignore subtle local differences between conformations.

In contrast to the previous studies, we do not use the position of individual atoms nor secondary structure to quantify the local structural features and differences between backbone conformations. Instead, we consider the "second derivative" of the Cα trace of a protein, where the gradient vector at the i-th Cα atom $C(i)$ is defined as the direction from the position of $C(i-1)$ to that of $C(i+1)$ as defined by Rackovsky and Scheraga [51]. To identify only statistically significant features, we quantify the background space (i.e., we divide the background space into tetrahedrons) and discretize gradient vectors at Cα atoms. It should be noted that we can not deal with variation in a gradient vector along a Cα trace in a "differential geometrical" setting without quantization of space. For example, Louie and Somorjai [52] and Montalvao and coworkers [53] applied the differential geometry of curves to the analysis of Cα traces, although they did not consider the relationship between the gradient vectors of consecutive Cα atoms.

Table I: Frequency distribution of causes that account for the structural differences observed for the 60 crystal structure pairs used in this study. The monomeric/hetero-multileric pairs consist of a monomeric protein and a member of a hetero-multimeric protein. The monomeric/homo-multileric pairs consist of a monomeric protein and a member of a homo-multimeric protein. The members-of-dimmer pair consist of the members of a homo-dimeric protein. The members-of-oligomer pair consist of two members of a homo- or hetero-oligomeric protein. The change-upon-ligand-binding pairs consist of a ligand-free form and a ligand-bound form of the same protein. The complex-with-different-partners pairs consist of two conformations of the same protein from two different protein-ligand complexes. The lipid-bound-apolipoproteins pairs consist of two different lipid-bound forms of the same apolipoprotein. With regard to the monomeric/homo-multimeric pairs, nine of the 11 pairs consist of a monomeric protein and a member of a domain-swapped dimer.

Type	#
Monomeric / Hetero-multimeric	12
Monomeric / Homo-multimeric	11
Members of dimer	8
Members of oligomer	8
Change upon ligand-binding	7
Complex with different partners	5
Lipid-bound apolipoproteins	3
Else	6
Total	60

MATERIAL AND METHOD

In this study, we have used 158 pairs of crystal structures solved at a resolution of $2.5\mathring{A}$ or better that were identified by Kosloff and Kolodny [14]. These data were chosen for our study because the protein structures were aligned based solely on sequence information. The 158 pairs are 100% identical in sequence and the sequence-based superpositioning RMSD is greater than $6.0\mathring{A}$. Moreover, they are clustered into 60 classes based on amino acid sequence. For statistical analysis purposes, we have chosen a pair from each of the 60 clusters to avoid statistical bias caused by proteins with many structures. Table I shows the distribution of the causes for the structural differences observed for the 60 pairs.

Discrete differential geometry of tetrahedrons

As mentioned above, we divide a three-dimensional Euclidean space into tetrahedrons, on which we construct a discrete version of differential geometry. Each tetrahedron consists of four short edges and two long edges, where the ratio of their length is $\sqrt{3}/2$. Note that it is one of the identical six tetrahedrons which are obtained when a facet of a four-dimensional unit cube is projected diagonally onto a three-dimensional hyperplane in a four-dimensional Euclidean space. The division of a three-dimensional Euclidean space is obtained by projecting unit cubes of a four-dimensional integer lattice diagonally onto a three-dimensional hyperplane [54].

Space curves are uniquely determined by curvature and torsion, where curvature is a measure of the deviation from a straight line (i.e., "turn") and torsion is a measure of the deviation from a plane (i.e., "twist"). In our version of differential geometry, each tetrahedron can assume one of four gradient vectors which are the direction of the four short edges of the tetrahedron (Figure 1A), and each tetrahedron trajectory goes along the gradient vector. Curvature of a space curve can then assume one of two values, "turn" or "not turn" (Figure 1B right) because curves are not differentiable in the other two cases (Figure 1B left). That is, the current tetrahedron (colored white) would be assigned more than one gradient vectors in the two cases of Figure 1B left. As for the torsion value (or "twist"), it is uniquely determined once the curvature is given because of the rigidity of the tetrahedron blocks.

Because curvature is binary, we can describe variation in a gradient vector (i.e., the "second derivative") along a tetrahedron curve as a binary sequence of 0 and 1. If the

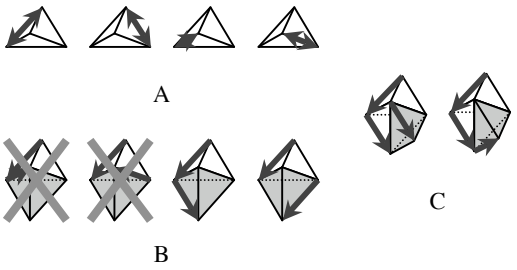

Figure 1: Flow of tetrahedron blocks. (A) Four gradient vectors (i.e., the direction of the four short edges) of a tetrahedron, where the arrows indicate the direction of flow. (B) Permitted and prohibited combination of gradient vectors of consecutive tetrahedron blocks. Suppose that the current block (white) is assigned the gradient vector shown on the far-left panel of A and the flow goes downward. The next block (gray) can then assume only two of the four gradient vectors shown in A. (C) Examples of smooth curves of length three. Once the gradient vector of the first two blocks (white) are given, a pair of permissible gradient vectors of the third block (gray) is uniquely determined: "not turn" (left) and "turn" (right). Variation in a gradient vector is 011 (or 100) for the left case and 010 (or 101) for the right case.

current tetrahedron is assigned a value of 0 (or 1), then assign a value of 0 (or 1) to the next tetrahedron if the gradient vectors of the two consecutive tetrahedrons are the same, otherwise assign a value of 1 (or 0, respectively) to the next tetrahedron. That is, the value is changed if the gradient vector is changed. For example, the space curve shown on the left panel of Figure 1C is assigned the binary sequence 011 or 100 (going downward) and the space curve shown on the right panel of Figure 1C is assigned the binary sequence 010 or 101 (going downward), depending on the initial value assignment. In the following, we always use the gradient vector shown on the far-left panel of Figure 1A as the initial gradient vector, as in the case of Figure 1C, and assign a value of 0 to the initial tetrahedron.

D^2 encoding of Cα traces

In this study, we apply the discrete differential geometry of tetrahedrons to quantify the local structural features and differences between protein backbone conformations. We use a Cα trace to represent the conformation of a protein and consider variation in a gradient vector along a Cα trace, i.e., the "second derivative" of a Cα trace. Because we are concerned with local structural feature, we consider all five-Cα fragments of the Cα trace of a protein.

First, we assign a $0, 1$-valued sequence for five amino acids to the center Cα atom of every five-Cα fragments as explained in detail below, where $0, 1$-valued sequences for five amino acids are denoted by the corresponding base-32 numbers: $0, 1, \ldots, 9, A, B, \ldots, V$. For example, '2' is assigned to the center Cα atom if the variation in a gradient vector along a fragment is 00010, '9' is assigned to the center Cα atom if the variation in a gradient vector along a fragment is 01001, 'A' is assigned if 01010, and so on.

We can then describe the conformation of a protein by arranging the base-32 numbers in the order as the corresponding Cα atoms appear in the protein. We call the base-32 number sequence the D^2 code of a protein because each number represents the "second derivative" of the Cα trace around the corresponding Cα atom. (Previously, the base-32 number sequence was called the 5-tile code.) In this study, encoding was carried out with program *ProteinEncoder* and figures are prepared with program *ProteinViewer*.

Upon computation of the $0, 1$-valued sequence of a five-Cα fragment, we allowed rotation and translation of tetrahedrons to absorb the irregularity inherent in actual protein structures. The protocol used to encode the Cα fragment $C(i-1)C(i)C(i+1)C(i+2)$ in Figure 2A is described here. In the following, we denote the vector from point A to point B by AB.

First, the gradient vector of a Cα trace at the i-th Cα atom $C(i)$ is defined as the direction from the position of $C(i-1)$ to that of $C(i+1)$ as defined by Rackovsky and Scheraga [51]. The initial tetrahedron $T(i)$ (defined by four vertices O, P, Q, and R) is then aligned with the Cα atom $C(i)$ and given a value of 0 (Figure 2B). The gradient vector OR of $T(i)$ is shown on the far-left in Figure 1A and the length of the vector OR is about one fifth of the average distance between the Cα atoms. $T(i)$ is aligned with $C(i)$ in such a way that (1), the direction of the vector OR and the vector $C(i-1)C(i+1)$ coincide and (2), the direction of the vector $OS + OP$ and the vector $C(i)C(i-1) + C(i)C(i+1)$ coincide, where S $(= O + RQ)$ is a vertex of an adjacent tetrahedron.

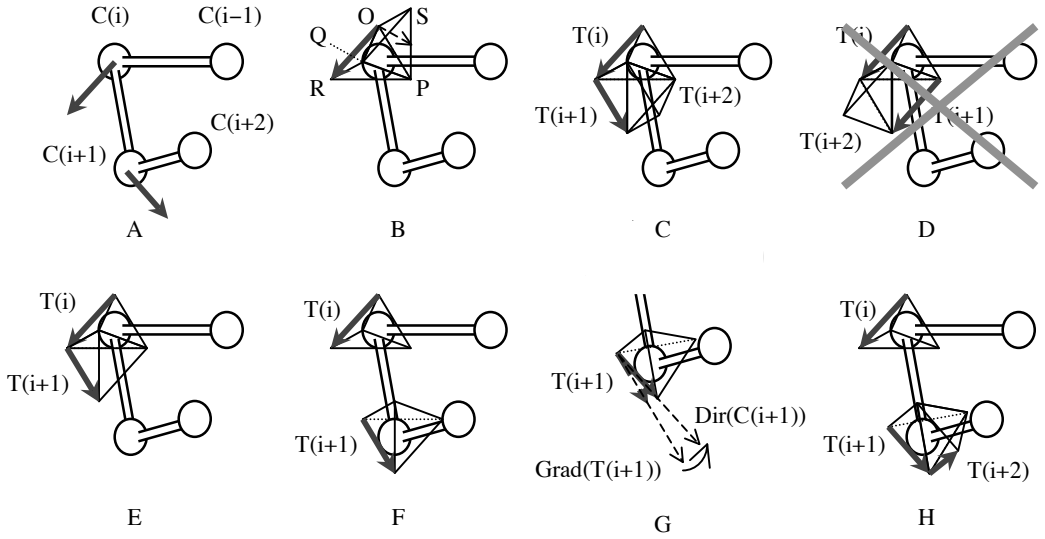

Figure 2: Algorithm of D^2 encoding. (A) The Cα trace of a protein to be encoded. The arrows indicate the direction of a gradient vector along the Cα trace. (B) Spatial orientation of the initial tetrahedron. (C) and (D) Two permitted values of the gradient vector of $T(i+1)$. (E) The gradient vector of the first two tetrahedrons. (F) Translation of $T(i+1)$ to the position of $C(i+1)$. (G) Rotation of $T(i+1)$ in the position of $C(i+1)$ (H) The gradient vector of the three tetrahedrons.

Once the spatial orientation of $T(i)$ is fixed, the position and spatial orientation of the next tetrahedron $T(i+1)$ is also determined uniquely. $T(i+1)$ can then assume one of the two gradient vectors shown in Figure 2C and 2D. The gradient vector of $T(i+1)$ is chosen from the two vectors based on the distance between $T(i+2)$ and $C(i+1)$. In the current case, $T(i+2)$ in Figure 2C is closer to $C(i+1)$ than $T(i+2)$ in Figure 2D. Thus, the gradient vector shown in Figure 2C is assigned to $T(i+1)$ (Figure 2E). Because the gradient vector of $T(i+1)$ is different from that of $T(i)$, a value of 1 is assigned to $T(i+1)$.

Next, $T(i+1)$ is translated to the position of $C(i+1)$ and rotated in that position to absorb irregularity of the Cα trace (Figure 2F and 2G). $T(i+1)$ is rotated around the cross product $Grad(T(i+1)) \times Dir(C(i+1))$ until the direction of $Grad(T(i+1))$ coincides with that of $Dir(C(i+1))$ (i.e., "tur" without "twist"), where $Grad(T(i+1))$ is the gradient vector of $T(i+1)$ and $Dir(C(i+1))$ is the gradient vector of the Cα trace at $C(i+1)$. Once the spatial orientation of $T(i+1)$ is fixed, the position and spatial orientation of the next tetrahedron $T(i+2)$ is also determined uniquely. The gradient vector of $T(i+2)$ is then chosen from two candidate vectors based on the distance between $T(i+2)$ and $C(i+2)$ (Figure 2H). Because the gradient vector of $T(i+2)$ is different from that of $T(i+1)$, a value of 0 is assigned to $T(i+2)$ and we obtain the binary sequence 010 which describes the variation in a gradient vector along the fragment $C(i)C(i+1)C(i+2)$.

In the same way, we encoded the fragment $C(i)C(i-1)C(i-2)$ starting from $C(i)$, and obtain a binary sequence of length five, which describes the variation in a gradient vector along the five-Cα fragment $C(i-2)C(i-1)C(i)C(i+1)C(i+2)$. Note that the D^2 code is sensitive to the twisting of Cα traces by definition.

Longest common subsequence and alignment length

One of the simplest measures of sequence similarity is the length of the longest common subsequence (LCS).

We quantified the differences between backbone conformations by the length of the LCS of their D^2 codes. A subsequence of a string is an ordered sequence of characters (not necessarily consecutive) from the string. A common subsequence of two strings is a subsequence of both of them [55]. For example, "$QA0$" is a subsequence of "$QAAB0$", and "QB" is a common subsequence of "$R0QB$" and "$QAAB$".

On the other hand, the structural similarity of two proteins is usually measured with the root mean square deviation (RMSD) of aligned residues after geometrical superposition. The alignment length (AL) of a superposition is the number of aligned residues and provides another measure of structural similarity, even though RMSD and AL are depending on each other and one of them could be improved at the expense of the other.

We used program *ComSubstruct* to find the LCS of two D^2 codes, and the *DaliLite* server [56] to compute a geometrical superposition of two proteins.

RESULTS

First, we examined concrete examples of multiple-structure proteins to assess the sensitivity of the D^2 code. We start with simple hypothetical fragments, such as loop and extended fragments. We then considered domain-swapped dimers, whose formation mechanism has implications for the development of amyloid plaques observed in misfolding diseases such as Alzheimer's and Parkinson's diseases. We also analyzed several other proteins before performing statistical analysis of 60 crystal structure pairs of the same proteins.

We used the DSSP assignments available from the PDB database of EMBL-EBI (`http://www.ebi.ac.uk/msd/`). All the programs used and the data obtained are available from `http://www.genocript.com`.

Simple hypothetical fragments dimers

Figure 3A demonstrates examples of D^2 encoding of loop fragments. Detection of loop movements is indispensable for our purposes because they often play an important role in protein function. We detected bending of loops successfully by the D^2 code. On the other hand, Figure 3B demonstrates D^2 encoding of helical and extended fragments, which shows a feature of the coding algorithm. That is, all the fragments except both ends are identical with respect to the D^2 code. Approximately 40% of the five-residue fragments contained in a representative set of the SCOP family (1.69 release) [57] are assigned a D^2 code of '0', where the top seven D^2 codes of five-residue fragments are '0' (40%), 'A' (31%), 'R' (10%),

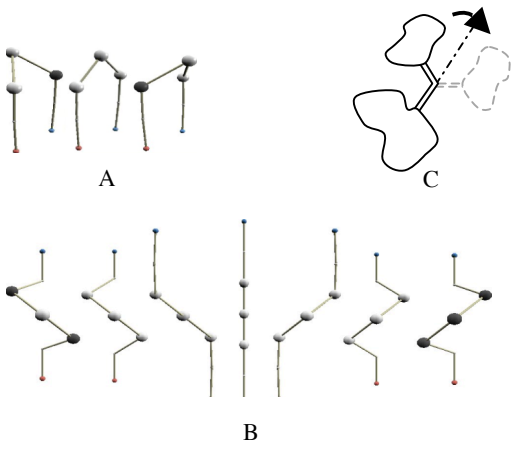

Figure 3: D^2 encoding of simple hypothetical fragments. (A) Three loop fragments, where spheres indicate the position of Cα atoms. From left, a loop fragment whose curved region (i.e., the tip of the loop) is bent to the left side, a loop fragment which resides in a virtual flat plane, and a loop fragment whose curved region is bent to the right side. Their D^2 codes are "$0RG$", "$0R0$", and "$1R0$" respectively, where fragments are encoded from this side to the other side. The black spheres show the position of the Cα atoms that have a D^2 code value other than '0'. (B) Five helical and extended fragments, where spheres indicate the position of Cα atoms. From left, three left-helical, a straight, and three right-helical fragments. Their D^2 codes are "$O03$", "000", "000", "000", "000", "000", and "$OH3$" respectively, where fragments are encoded from bottom to top. The black spheres show the position of the Cα atoms that have a D^2 code value other than '0'. (C) Example of global conformational changes that can not be detected by local Cα trace analysis. The rotation of the upper part around a bond can not be detected by local Cα trace analysis.

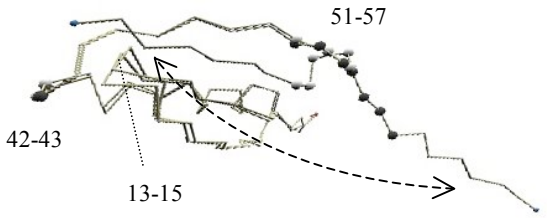

Figure 4: Superposition of the Cα traces of monomeric (compact) and homo-dimeric (extended) PpLs. Black (extended) and white (compact) spheres indicate the position of the Cα atoms of residues 42–43 and 51–57, which assume two different D^2 codes. The peripheral loop movement at residues 13–15 does not change the D^2 code of the loop.

'B' (6%), 'Q' (6%), 'G' (5%), and '1' (1%) [58]. 'A' corresponds to helices, '0' to extended strands, 'B' to the C-caps of helices, 'Q' to the N-caps of helices, and 'R'/'G'/'1' to turns, where the C-cap and N-cap of a helix are the last and the first residue within the helix respectively. Finally, it should be noted that specific types of global conformational changes can not be detected by local Cα trace analysis, such as the one shown in Figure 3C.

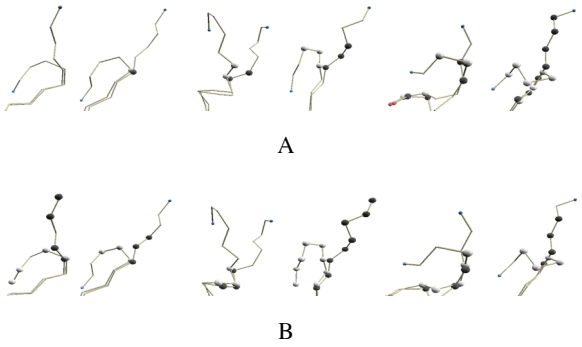

Figure 5: Cα traces of the hinge region of domain-swapped dimers. Monomeric (compact) and homo-dimeric (extended) forms are superimposed manually, where the N terminus is at the bottom. Their PDB IDs are (from left): 1bmbA/1fyrB, 1hufA/1k46A, 1y51/1y50, 1c0bA/1f0vA, 1i0cA/1aojA, and 1k50A/1k50B (compact/extended). (A) D^2 code assignments. White and black spheres show the position of the Cα atoms that assume two different D^2 codes, whose numbers are 0, 1, 2, 3, 4, and 7 (from left). (B) DSSP state assignments. White and black spheres show the position of the Cα atoms that assume two different DSSP states, whose numbers are 5, 3, 3, 8, 5, and 4 (from left).

Domain-swapped dimers

Monomeric/Homo-multimeric: PpL (1k50 A/1k50 B)

PpL is a multi-domain protein from Peptostreptococcus magnus (bacteria) that is an immunoglobulin-binding protein and has been used for the isolation and purification of immunoglobulins. We analyzed a single mutation (V94A) that triggers a population shift of the conformational ensemble towards the domain-swapped dimer [59]. The asymmetric unit of its crystal structure (PDB ID 1k50) contains two monomers (chain A and C) and one domain-swapped dimer (chain B and D). As mentioned above, the mechanism of domain swapping has been suggested to be similar to that of protein aggregation, which is closely associated with numerous human diseases, such as Alzheimer's and Parkinson 's diseases [60].

The V49A mutant consists of two β-hairpins, which are linked by an α-helix. The C terminal β-strand of the second β-hairpin (residues 44–64) is involved in the domain swapping of the dimer. There are nine residues that assume different D^2 codes: seven in the hinge region (51–57) and two around the C-cap of the α-helix (42, 43) (Figure 4). The hinge mechanism involves five residues (51–57) that induce a significant conformational change of the second β-hairpin. Changes at residue 42 and 43 are also attributable to the hinge motion. These changes are not supposed to be a result of crystal packing because they have not moved in the same direction. Also observed was a movement of the loop near residue 14, which has preserved the D^2 code. The D^2 code of the loop is "00QAB00" and the movement is within the range of D^2 codes 'Q' and 'B'. The distortion is probably caused by crystal packing because residues 13–15 project to the surface of the crystal unit.

Conformations of the hinge regions

In the dataset, nine of the 11 monomeric/homomultimeric pairs consist of a monomeric protein and a member of a domain-swapped dimmer, where a domain-swapped dimer is obtained by replacing a fragment of a monomeric protein by the same fragment from another. Domain swapping has been repeatedly observed in a variety of proteins and is believed to result from destabilization due to mutations or changes in environment as seen in the PpL V49A mutant [61].

Here, we make comparisons between various types of hinge motions that can have implications for the design and evolution of proteins because (1) protein design often incorporates point mutations with increased or decreased stability [61] and (2) the monomer-dimer equilibrium could be controlled by mutations in the hinge region [60].

Figure 5A shows the hinge region of representative examples of the nine domain-swapped dimers. From left to right, they are arranged in increasing order of the number of Cα atoms that assume two different D^2 codes: 0, 1, 2, 3, 4, and 7. The number of differences between D^2 codes seems to have successfully quantified the structural differences over the region. The far-right image in Figure 5A is the hinge region of the V49A mutant, where the mutation occurred at the N terminus (bottom). All Cα atoms of the fragment except both termini undergo conformational changes upon dimerization. In particular, the motion seems to be less efficient and might be more prone to misfolding than others. On the other hand, the far-left image in Figure 5A is an example of conformational changes that are not detectable by local Cα trace analysis, i.e., rotation around a bond (See also Figure 3C). The motion is simple and the monomer-dimer equilibrium could show a larger propensity for dimer than others. If we could identify point mutations in the hinge region that increase the number of Cα atoms with different D^2 codes, we could destabilize the dimer and inhibit dimer formation to some degree by the mutations.

Finally, Figure reffig5B shows the DSSP state assignment of the same fragments. Instead of Cα traces, the DSSP program uses hydrogen bonding patterns, solvent exposure, and geometrical features to assign one of eight states to each residue [43]. No clear relationship is observed between conformational and DSSP state variabilities in the figure. For example, the hinge pair of the image located on the far left has more residues that assume two distinct DSSP states than the hinge pair of the image located on the far right. However, the structural differences between the hinge pairs in the far-left image are smaller than those in the far-right image. That is, DSSP states are not an adequate measure of the structural difference between proteins in this case.

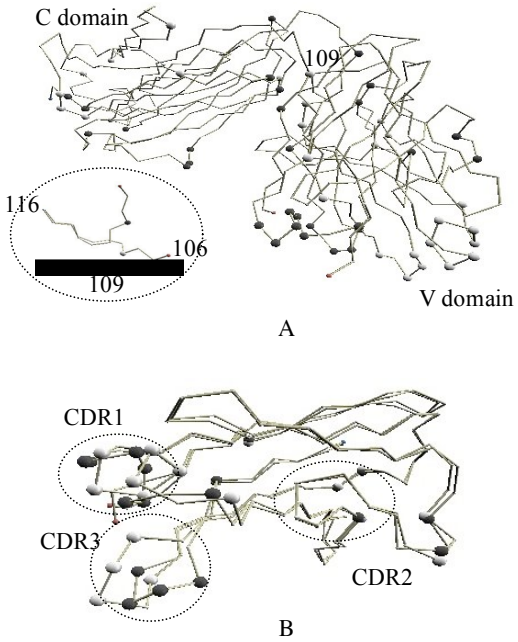

Figure 6: Cα traces of the chains of Bence-Jones protein Mcg. (A) Superposition of chain A and B of Mcg. Encircled with a dashed line is a superposition of the linker region. (B) Superposition of the V domains of chain A and B. Black (chain A) and white (chain B) spheres show the position of the Cα atoms which assume two different D^2 codes.

Members of dimer: Bence-Jones protein Mcg (1dcl A/1dcl B)

Bence-Jones proteins are monoclonal globulin proteins, that are commonly found in the urine of patients with multiple myeloma and often used in the diagnosis of this disease. We analyzed a crystal structure (PDB ID 1dcl) of a

lambda type Bence-Jones protein ([62],[63]).

Bence-Jones proteins are dimers of two identical light chains, A and B, of an immunoglobulin. The two chains adopt different conformations in the Mcg dimer. They are composed of two globular domains, variable (V) and constant (C), where the V domain contains three antigen-binding sites, CDR1 (residues 26–34), CDR2 (52–58), and CDR3 (91–100), which are highly variable among different immunoglobulins [64]. The structural differences between the chains of the Mcg dimer are mainly due to the flexibility of the linker region (109–111) between the V and C domains, which exhibits different "elbow bends."

There are 29 residues that assume different D^2 codes: 18 in the V domain, ten in the C domain, and one (109) is involved in the elbow bend mechanism (Figure 6A). The two domains are rather stable under the conformational change (RMSD $1.6\mathring{A}$ for the V domain and $1.0\mathring{A}$ for the C domain by *DaliLite*). As for the V domain, most of the changes occur in the CRD1 and CDR3 regions (26–31, 33, 94–96, 98) (Figure 6B). In particular, the CRD1 segment forms a left-handed helical segment in chain A and a right-handed helical segment in chain B, which is a result of interference between adjacent protein molecules in the crystal. The CRD1 region of chain A could not form the same conformation as they have in chain B because of a space limitation [65]. On the other hand, the C domain is rather rigid and differences are due to changes in the surface loop regions, where ten Cα atoms with different D^2 codes are evenly distributed over the loops.

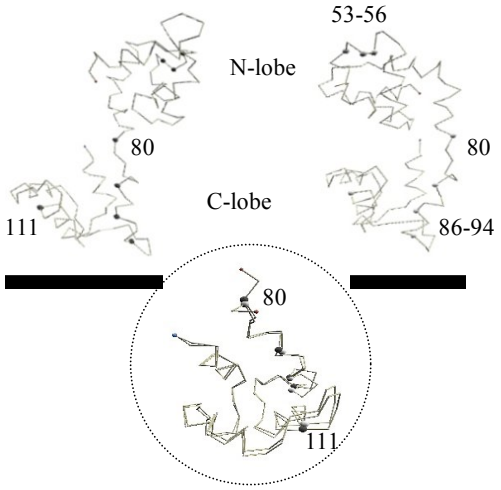

Figure 7: Cα traces of Mlc1p-IQ4 (top-left) and Mlc1p-IQ2 (top-right) complexes. Neither IQ4 nor IQ2 is shown in the figure. Encircled with a dashed line is a superposition of the two conformations of the C-lobe of Mlc1p. Black (Mlc1p-IQ4, extended) and white (Mlc1p-IQ2, compact) spheres show the position of the Cα atoms that assume two different D^2 codes.

Protein-ligand complexes with different partners: Mlc1p (1m46 A/1n2d A)

Mlc1p is a protein from the budding yeast Saccharomyces cerevisiae that binds to IQ motifs of a class V myosin family member and plays a role in polarized growth and cytokinesis. IQ motifs are approximately 25-residue fragments that are folded as an uninterrupted α-helix. Mlc1p recognizes subtle differences between IQ motifs, such as IQ2, IQ3, and IQ4, to assume markedly different conformations [66]. Here we consider the difference between the crystal structures of Mlc1p bound to IQ2 (1n2d) and Mlc1p bound to IQ4 (1m46).

Mlc1p is a dumbbell-shaped molecule where two homologous domains, the N- and C-lobes, are connected by a flexible linker loop (residues 80–82). Depending on the sequence of the IQ motifs, Mlc1p adopts either a compact conformation using both lobes (IQ2) or an extended conformation using the C-lobe alone (IQ4). When bound to Mlc1p, IQ2 interacts with the N-lobe mainly through electrostatic contacts and interacts with the C-lobe mainly through hydrophobic contacts. On the other hand, IQ4 interacts with the C-lobe only and leaves the N-lobe available for other interactions, resulting in the extended conformation of Mlc1p.

There are eight residues that assume different D^2 codes: three for the N-lobe (residues 53, 54, 56), four for the C-lobe (86, 90, 94, 111), and one for the linker (80) (Figure 7 top). The three residues of the N-lobe are located in a loop region between α-helices and do not significantly affect the conformation of the N-lobe (RMSD $0.6\mathring{A}$ by *DaliLite*). In contrast, the four residues of the C-lobe are located either in an α-helix (86, 90, 94) or on the edge of another α-helix (111) (Figure 7 bottom), and cause a bend of the α-helix and a movement of a flanking

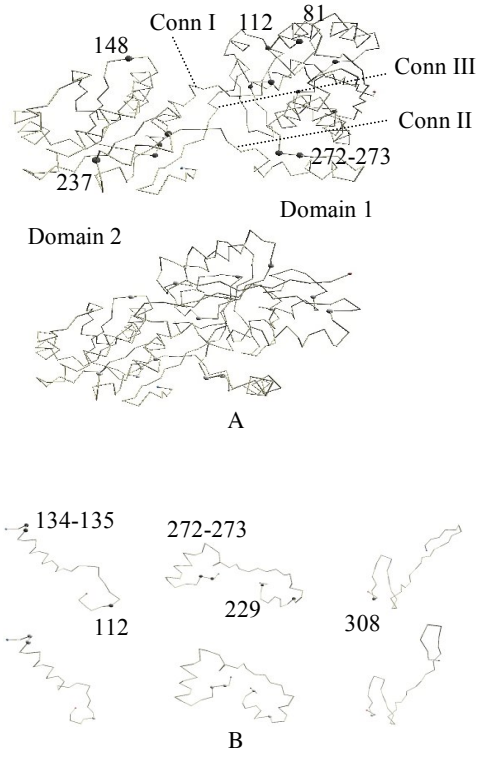

Figure 8: Cα traces of the open and closed forms of LBP. (A) The open (top) and closed (bottom) forms of LBP. (B) The open (top) and closed (bottom) forms of the three connections between the two domains (From left, connection I, II, and III). Black (open) and white (closed) spheres show the position of the Cα atoms that assume two different D^2 codes.

loop in order to adjust the width of a channel that accommodates the IQ motifs (RMSD $0.9\mathring{A}$ by *DaliLite*). Finally, the linker loop undergoes a large deformation that is caused by a distortion around residue 80. (Note that the conformation of the linker of the Mlc1p-IQ4 complex is probably irrelevant because the N-lobe could move freely if it were not in the crystal.)

Changes upon ligand-binding: LBP (1usg A/1usi A)

LBP is a leucine-binding protein from Escherichia coli, which is the primary receptor for the leucine transport system, and binds to leucine and phenylalanine. Here we consider the difference between the crystal structures of a ligand-free (open) form (1usg) and a phenylalanine-bound (closed) form (1usi).

LBP is comprised of two domains, domain 1 and domain 2, connected by a three-stranded hinge. A phenylalanine molecule binds to LBP in a cleft that is formed between the domains by both hydrogen bonding (residues 79, 100, 102, 202, 226) and non-polar interactions (18, 150, 202, 276). In the following, we call the three hinge segments, connection I (117–121), connection II (248–252), and connection III (325–331) [67].

Upon opening and closing, the two domains remain rather rigid (RMSD $0.6\mathring{A}$ for domain 1 and $0.5\mathring{A}$ for domain 2 by *DaliLite*) and most of the conformational changes occur in the flanking regions of the connections. In the closed form, connection I is pushed into the flanking helix (121–133), the flanking extended strand (244–247) of connection II curves, and connection III undergoes a rotation around the virtual bond between Cα atom 329 and 330. As for the ligand-binding sites, the positions of residues 79, 100, and 276 of domain 1 are affected and the side chain of residue 18 (Trp) changes its conformation. A few deformations due to crystal packing are also observed.

There are 14 residues that assume different D^2 codes: nine for domain 1 (39, 71, 81, 100, 112, 272, 273, 294, 308), five for domain 2 (134, 135, 148, 229, and 237), and none for the connections (Figure 8A). Five (112, 134, 135, 272, 273) of them are caused by the distortion around connections, four (81, 100, 272, 273) are caused by the distortion around ligand-binding residues, and eight (39, 294, 308, 71, 112, 229, 237, 148) are due to crystal packing.

Residues 112, 134, 135, 272, and 273 are in the flanking regions of connections I and II, where they absorb the distortion of connections (Figure 8B). However, the deformations over the connections are not detected by the D^2 code because of the biased frequency distribution of the occurrence of D^2 code mentioned above. That is, about 40% of five-residue fragments of a representative set of

the SCOP family are assigned a D^2 code of '0' (See also Figure 3B). The deformation of the extended strands in the connections are too modest to be captured by the D^2 code, although RMSD of 31-residue fragments centered on connection I, II, and III are $3.0\mathring{A}$, $3.8\mathring{A}$, and $2.5\mathring{A}$ respectively by *DaliLite* (Figure 8B). As for the deformation of connection III, it is the type of movement shown in Figure 3C, which can not be captured by local Cα trace analysis. On the other hand, residues 79 and 100 are located in regions directly involved in ligand binding, and the distortion at residues 272 and 273 caused a movement of residue 276 to make room for the ligand. In regard to residue 18, no backbone deformation is observed because it induces side chain movement only.

Also influenced are some fragments distant from the ligand-binding sites, which are probably explained by crystal packing. In the closed form, fragment 296–308 on the surface of domain 1 is pressed uniformly from the outside, and residues 39, 71, and 112 are pushed away by the fragment. With respect to domain 2, residue 148 and fragment 229–237 might be also affected by crystal packing. (Note that the conformation of the open form is also stabilized by the crystal packing, as domain 1 from one molecule is placed between the domains in an adjacent molecule, so preventing the protein from closing.)

Statistical analysis

In this section, we performed a statistical analysis of 60 crystal structure pairs of the same proteins identified by Kosloff and Kolodny [14]. First, we considered the dissimilarities between the pairs captured by the D^2 code and characterized them via comparison with DSSP state assignments. Then, we considered the similarities between the pairs captured by the D^2 code and characterized them via comparison with three-dimensional structural superpositions. Finally, we analyzed the distribution of deformation types in the case of variable regions which contain one amino acid.

Dissimilarities between crystal structure pairs

Figure 9A shows the distribution of the lengths of variable regions observed in the 60 proteins with respect to both the D^2 code and the DSSP state. Recall that D^2 codes are computed based on Cα traces, and DSSP states are assigned based on hydrogen bonding pattern, solvent exposure, and geometrical features [43]. As for D^2 encoding, approximately 61% of variable regions contain one amino acid, 17% contain two amino acids, and 93% contain less than five amino acids. The length of three longest regions are 17, 21, and 38 amino acids, all of which are associated with deformation of α-helix. On average, variable regions contain 2.0 amino acids. Regarding the DSSP assignment, only 47% of variable regions contain one amino acid, 22% contain two amino acids, and 89% contain less than five amino acids. The length of the three longest regions are 20, 21, and 38 amino acids, all of which are also associated with deformation of α-helix. The average length is 2.5 amino acids, which is longer than that of D^2 code.

Table II: Structural profiles of the four crystal structure pairs inspected above. Table shows the length of proteins, the percentage of Cα atoms that assume different D^2 codes (and DSSP states) to all Cα atoms, and the percentages of LCS length (and AL) to the entire sequence. Top row shows the average over the 60 pairs.

	Length	*Difference (D^2/DSSP)*	*LCS / AL*
Ave.	212 residues	12% / 13%	88% / 81%
PpL	63	14 / 6	84 / 84
Mcg	216	13 / 19	87 / 82
Mlc1p	147	5 / 5	92 / 79
LBP	345	4 / 4	95 / 100

Figure reffig9B shows the distribution of the number of variable regions of the 60 proteins with respect to the D^2 code and the DSSP state. As for D^2 encoding, 31 proteins have less than 11 variable regions. On one hand, three proteins (of 59, 70, and 288 amino acids) have only one variable region. On the other hand, the top three proteins have 34, 45, and 67 variable regions, which contain 728, 688, and 994 amino acids respectively. The average protein contains 211.8 amino acids, and the number

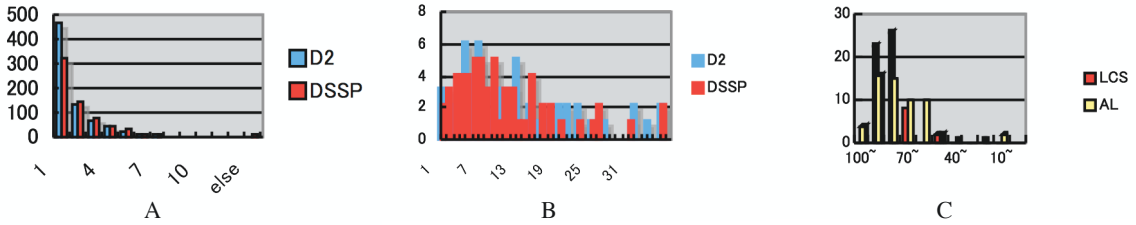

Figure 9: Statistical analysis of 60 crystal structure pairs of the same proteins. (A) The distribution of length of variable regions of the 60 proteins with respect to the D^2 code and the DSSP state. The average lengths are 2.0 Cα atoms (D^2 code) and 2.5 Cα atoms (DSSP state). (B) The distribution of the number of variable regions of the 60 proteins with respect to D^2 code and DSSP state. The average numbers are 12.9 regions (D^2 code) and 11.3 regions (DSSP state) per protein. (C) The distribution of the percentage of LCS length (and AL) to the entire sequence of the 60 crystal structure pairs. The average percentages are 85% (LCS) and 79% (AL). On the other hand, the percentages of the average lengths of LCS and AL to the average length of the entire sequence are 88% and 81% respectively, as shown in Tables II.

of variable regions in a protein is 12.9. That is, about 12% of Cα atoms in a protein are variable with respect to the D^2 code because the average length of the variable regions is 2.0 amino acids. In regard to the DSSP assignment, 30 proteins have less than nine variable regions. Two proteins (of 59 and 63 amino acids) have only one variable region. In addition, the top three proteins have 31, 45, and 54 variable regions, which contain 527, 688, and 994 amino acids respectively. On average, the number of variable regions is 11.3 and about 13% of Cα atoms are variable with respect to the DSSP state.

In conclusion, D^2-variable regions are distributed more sparsely than DSSP-variable regions although the total amount are almost equal, i.e., 12% and 13% of Cα atoms respectively. Table II shows structural profiles of the examples inspected above. It is evident that the profile of Mcg is almost equal to the average, and Mcg is approximately twice as flexible as Mlc1p and LBP.

Similarities between crystal structure pairs

Figure 9C shows the distribution of the length of LCS (Longest Common Subsequence) between the D^2 code sequences of the 60 crystal structure pairs, and the AL (alignment length) of structural superpositions of the same 60 pairs, where LCS lengths and ALs were computed with *ComSubstruct* and the *DaliLite* server respectively. Regarding the percentage of LCS length to the entire sequence, the LCS percentages of 26 pairs are in the range from 80% to 89%, the LCS percentages of 23 pairs are in the range from 90% to 99%, and no structure pair is 100% identical. As for the percentage of AL to the entire sequence, the AL percentages of 51 pairs are rather evenly distributed between 60% and 99% and four pairs are 100% aligned, including LBP analyzed above, where the RMSD of their alignments are 5.5Å, 7.1Å, 7.1Å, and 8.3Å. The LCS percentages of the four pairs are 85%, 84%, 95%, and 88% respectively.

On average, the LCS percentage is 85%, which is 6% larger than the AL percentage (the LCS length is 10% longer than the AL length). This result is reasonable because the *DaliLite* server doesn't take flexibility of proteins into account.

Distribution of deformation types

Table III shows the distribution of deformation types of variable regions which contain one amino acid with respect to both the D^2 code and the DSSP state. Recall that the D^2 code of a residue represents the conformation of the five-Cα fragments centered on the residue, where '0' (= 00000) corresponds to extended strands, 'A' (= 01010) to helices, and so on. That is, the table is concerned with isolated deformation of the five-Cα fragment

Table III: Deformation types of variable regions which contain one amino acid. Distribution of D^2 code transition types (left) and DSSP state transition types (right) observed in the 60 crystal structure pairs. (Regarding the DSSP state, 'S' is bent, 'E' is extended strand, 'H' is helix, 'T' is turn, 'B' is bridge, and '.' denotes no assigned structure.)

Type	*Occurrence*
0↔R	137 (29.1%)
0↔G	78 (16.6)
B↔G	36 (7.6)
A↔B	23 (4.9)
0↔O	18 (3.8)
O↔R	17 (3.6)
0↔1	16 (3.4)
A↔H	16 (3.4)
0↔3	13 (2.8)
0↔B	11 (2.3)
A↔Q	11 (2.3)
else	95 (20.2)
all	471 (100)

Type	*Occurrence*
S↔.	129 (40.3%)
E↔.	68 (21.2)
H↔T	53 (16.6)
B↔.	13 (4.1)
S↔T	10 (3.1)
E↔S	10 (3.1)
else	37 (11.6)
all	320 (100)

centered on a residue.

In regard to D^2 code transition, 58% of the deformations involve extended strands (i.e., '0') and 11% involve helices (i.e., 'A') as shown in Table III left. Note that both of these cases exhibit a kind of asymmetry along a Cα trace. In the case of deformation of extended strands, 50% are bent in the middle of an extended strand, i.e., a conversion between '0' and 'R' (= **11**011), 35% are bent on the N-terminal side, i.e., a conversion between '0' and 'G' (= **1**0000)/'O' (= **11**000), and 11% are bent on the C-terminal side, i.e., a conversion between '0' and '1' (= 0000**1**)/'3' (= 000**11**). That is, a bend on the N-terminal side occurs three times or more as frequently as a bend on the C-terminal side. In the case of deformation of helices, 46% are a conversion between 'A' (= 01010) and C-cap 'B' (= 010**11**), 32% are a conversion between 'A' and 'H' (= **10001**) which implies a kink in the middle of a helix, and 22% are a conversion between 'A' and N-cap 'Q' (= **11**010). That is, unfolding at the C-terminal end of helices occurs two times or more as frequently as unfolding at the N-terminal end.

With respect to DSSP state transition, 66% of the deformations involve no assigned structure (i.e., '.') and 20% involve turns (i.e., 'T') as shown in Table III right. In the case of deformation of no assigned structures, 61% are converted to a bend 'S', 32% to a strand 'E', and 6% to a bridge 'B'. However, no detailed information of the structure is obtained from these figures.

DISCUSSION

First, let's consider sensitivity of the D^2 code. We have seen that the ratio, 12%, of residues with variable D^2 code is almost the same as that of residues with variable DSSP state. The most meaningful conformational differences between two forms of the same proteins seem to be covered by D^2-variable residues, as shown in the examples above.

Because the D^2 code is sensitive to the twisting of Cα traces by definition, we can use the D^2 code to pinpoint the very residues that induce twists in a backbone, as shown in Figure 5A. Actually, we have detected even small twists in the flanking regions as in the case of residues 42–43 of PpL, as well as the flanking regions of the CDR2 segment of Mcg. In regard to the insensitivity of the D^2 code to differences between extended fragments (Figure 3B), deformation of an extended fragment often induces twists in the flanking regions, which are detectable for the D^2 code as in the case of connection I and II of LBP (Figure 8B).

With respect to potential artifacts due to crystal packing, its influence on sensitivity seems low, because small movements related to crystal contacts are often without twist, as in the case of residues 13–15 of PpL (Figure 4). The discrete nature of the algorithm also circumvents the effect of coordinate errors and facilitates detection of significant differences between backbone conformations. However, it might be difficult to distinguish meaningful distortions from artifacts that are induced indirectly by

crystal packing, such as residues 39, 71, and 112 of LBP.

Secondly, let's consider local flexibility of proteins. We have seen that the average length of variable regions with respect to the D^2 code is 2.0 Cα atoms. However, there are various types of hinges, as shown in Figure 5, ranging in length from 0 Cα atom to 7 Cα atoms. The monomer-dimer equilibrium could be affected by the number of Cα atoms which assume different D^2 codes depending on the form, compact or extended. That is, the smaller the number of variable Cα atoms, the more stable a domain-swapped dimer is. In particular, the number of variable Cα atoms in the hinge region may be a good indicator for the stages of evolution of a domain-swapped dimer.

On the other hand, statistics shows that each fragment of length 16.4 ($= 211.8/12.9$) Cα atoms has a variable region. In other words, we can thermodynamically identify a multiple-structure protein with a sequence of rigid subdomains of length 14.4 Cα atoms connected by variable regions (or springs) of length 2.0 Cα atoms. The deformation of the variable regions are then coupled to form global structural transitions. As for the location of variable regions, some places, such as the N terminus of a β-strand and the C-cap of an α-helix, are more favored than others.

Finally, let's consider implications for protein engineering and drug design. When engineering proteins, point mutations are often introduced to enhance the stability of a target protein. The D^2 code analysis can provide valuable information for mutation points selection, i.e., D^2-variable regions of length more than two except α-helices are good candidates. For example, the monomer-dimer equilibrium of a domain-swapped protein could be fine-tuned by introducing a mutation on the hinge region if the mutation reduces the number of D^2-variable Cα atoms (Figure 5A).

The D^2 code analysis can be also useful in rational drug design. For instance, the existence of long D^2-variable regions may play a role in conformational changes observed in misfolding diseases. Additionally, analysis of the distribution of D^2-variable regions may lead to a more detailed description of the mechanism of multi-drug resistance due to non-active site mutations.

As an example, let's consider mutations in human immunodeficiency virus type 1 (HIV-1) protease which produce resistance to HIV-1 protease inhibitors, one of the major anti-HIV-1 drug targets, in the therapy of HIV-1 infection [68]. Because of the short life cycle and the high mutation rate of HIV-1, every mutation of HIV-1 protease is created thousands of time each day in each patient [69]. As a result, HIV-1 protease exists within a patient as a mixture of genetically related but distinguishable variants often referred to as a "swarm" or "quasi-species" [70]. Drug-resistant HIV-1 strains are then developed under the selective pressure of protease inhibitor therapy.

More than 60 mutations are currently associated with protease inhibitor resistance [71], which could be classified as active site or non-active site mutations depending on their location within the protease molecule. The mechanism of resistance due to non-active site mutations is not immediately apparent unlike the case of active site mutations, and extensive studies have been made on this topic ([72]–[76]), although the detailed mechanisms of drug resistance are yet to be clarified. The D^2 code analysis of a large collection of crystal structures of HIV-1 protease variants can provide a new perspective to the problem of drug resistance and lead to new ideas on designing more efficient drugs.

CONCLUSION

In this study, we have applied a discrete differential geometrical technique, D^2 encoding, to identify regions of 158 multiple-structure proteins where conformational changes take place. Due to the sensitivity of the D^2 code to the twisting of Cα traces, the sources of structural differences are successfully pinpointed by comparison of D^2 codes. In particular, we have found that a multiple-structure protein can be thermodynamically identified with a sequence of rigid subdomains of average length 14.4 Cα atoms connected by variable regions of average length 2.0 Cα atoms, where deformation of the variable regions are coupled to induce global structural transitions between two forms of the same proteins. As for the location of variable regions, some places, such as the N terminus of a β-strand and the C-cap of an α-helix, seem to be more favored than others.

We have suggested several implications of these results for protein engineering and drug design. For example, the number of D^2-variable Cα atoms in the hinge region of a domain-swapped dimer can be a good measure of evolution of the dimer. In addition, the existence of long

D^2-variable regions may play a role in conformational changes observed in misfolding diseases. Moreover, the D^2 code analysis of the structure of mutants can provide a new perspective to the problem of drug resistance due to non-active site mutations and lead to new ideas on designing more efficient drugs.

References

[1] Carlson HA, McCammon JA. Accommodating protein flexibility in computational drug design. Mol Pharm 2000;57:213-218.

[2] Carlson HA. Protein flexibility is an important component of structure-based drug discovery. Curr Pharm Design 2002;8:1571-1578.

[3] Teague SJ. Implications of protein flexibility for drug discovery. Nat Rev Drug Discov 2003;2:527-541.

[4] Teodoro ML, Kavraki LE. Conformational flexibility models for the receptor in structure based drug design. Curr Pharm Design 2003;9:1635-1648.

[5] Meagher KL, Carlson HA. Incorporating protein flexibility in structure-based drug discovery: using HIV-1 protease as a test case. J Am Chem Soc 2004;126;13276-13281.

[6] Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005;33:W306-W310.

[7] Blundell TL, Sibanda BL, Montalvao RW, Brewerton S, Chelliah V, Worth CL, Harmer NJ, Davies O, Burke D. Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. Phil Trans R Soc B 2006;361:413-423.

[8] Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations Nucleic Acids Res 2006;34:W239-W242.

[9] Barril X, Fradera X. Incorporating protein flexibility into docking and structure-based drug design. Expert Opin Drug Disco 2006;1:335-349.

[10] Ambroggio XI, and Kuhlman B. Design of protein conformational switches. Curr Opin Struct Biol 2006;16:525-530.

[11] Ahmed A, Kazemi S, Gohlke H. Protein flexibility and mobility in structure-based drug design. Frontiers in Drug Design & Discovery: Structure-Based Drug Design in the 21st Century 2007;3:455-476.

[12] Gunasekaran K, Nussinov R. How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. J Mol Biol. 2007 Jan 5;365(1):257-73.

[13] Chockalingam K, Blenner M, Banta S. Design and application of stimulus-responsive peptide systems. Protein Eng Des Sel 2007;20:155-216.

[14] Kosloff M, Kolodny R. Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins. 2008;71:891-902.

[15] Lesk AM. Introduction to protein science: architecture, function, and genomics, Oxford, UK: Oxford University Press; 2004.

[16] Englander JJ, Mar CD, Li W, Englander SW, Kim JS, Stranz DD, Hamuro Y, Woods VL Jr. Protein structure change studied by hydrogen-deuterium exchange, functional labeling, and mass spectrometry. Proc Natl Acad Sci USA 2003;100;7057-7062.

[17] Maity H, Lim WK, Rumbley JN, Englander SW. Protein hydrogen exchange mechanism: local fluctuations. Protein Sci 2003;12:153-160.

[18] Hilser VJ, Dowdy D, Oas TG, Freire E. The structural distribution of cooperative interactions in proteins: analysis of the native state ensemble. Proc Natl Acad Sci USA 1998;95:9903-9908.

[19] Tsai C-J, Kumar S, Ma B, Nussinov R. Folding funnels, binding funnels, and protein function. Protein Sci 1999;8:1181-1190.

[20] Freire E. Can allosteric regulation be predicted from structure? Proc Natl Acad Sci USA 2000;97:11680-11682.

[21] Luque I, Leavitt SA, Freire E. The linkage between protein folding and functional cooperativity: two sides of the same coin? Annu Rev Biophys Biomol Struct. 2002;31:235-56. Epub 2001 Oct 25.

[22] Freire E. Statistical thermodynamic linkage between conformational and binding equilibria. Adv Protein Chem 1998;51:255-279.

[23] Gunasekaran K, Ma B, Nussinov R. Is allostery an intrinsic property of all dynamic proteins? Proteins 2004;57:433-443.

[24] Freire E. The propagation of binding interactions to remote sites in proteins: analysis of the binding of the monoclonal antibody D1.3 to lysozyme. Proc Natl Acad Sci USA 1999;96:10118-10122.

[25] Pan H, Lee JC, Hilser VJ. Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble. Proc Natl Acad Sci USA 2000;97:12020-12025.

[26] Yu EW, Koshland DE Jr. Propagating conformational changes over long (and short) distances in proteins. Proc Natl Acad USA 2001;98:9517-9520.

[27] Saen-Oon S, Ghanem M, Schramm VL, Schwartz SD. Remote mutations and active site dynamics correlate with catalytic properties of purine nucleoside Phosphorylase. Biophys J 2008;94:4078-4088.

[28] Kern D, Zuiderweg ER. The role of dynamics in allosteric regulation. Curr Opin Struct Biol. 2003 Dec;13:748-57.

[29] Whitten ST, Garca-Moreno BE, Hilser VJ. Local conformational fluctuations can modulate the coupling between proton binding and global structural transitions in proteins. Proc Natl Acad Sci USA 2005;102:4282-4287.

[30] Meinhold L, Smith JC. Fluctuations and correlations in crystalline protein dynamics: a simulation analysis of staphylococcal nuclease. Biophys J. 2005;88:2554-2563.

[31] Liu T, Whitten ST, Hilser VJ. Functional residues serve a dominant role in mediating the cooperativity of the protein ensemble. Proc Natl Acad USA 2007;104:4347-4352.

[32] Gandhi PS, Chen Z, Mathews FS, Di Cera E. Structural identification of the pathway of long-range communication in an allosteric enzyme. Proc Natl Acad USA 2008;105:1832-1837.

[33] Ottemann KM, Xiao W, Shin YK, Koshland DE Jr. A piston model for transmembrane signaling of the aspartate receptor. Science 1999;285:1751-1754.

[34] Stec B, Phillips GN Jr. 2001. How the CO in myoglobin acquired its bend: lessons in interpretation of crystallographic data. Acta Cryst 2001;D57:751-754.

[35] Acharya KR, Lloyd MD. The advantages and limitations of protein crystal structures. Trends Pharmacol Sci 2005;26:10-14.

[36] Rejto PA, Freer ST. Protein conformational substates from X-ray crystallography. Prog Biophys Mol Biol 1996;66:167-196.

[37] Eyal E, Gerzon S, Potapov V, Edelman M, Sobolev V. The limit of accuracy of protein modeling: influence of crystal packing on protein structure. J Mol Biol 2005;351:431-42.

[38] DePristo MA, de Bakker PI, Blundell TL. Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure 2004;12:831-838.

[39] Petsko GA. Not just your average structures. Nat Struct Biol 1996;3:565-566.

[40] Flocco MM, Mowbray SL. C α -based torsion angles: a simple tool to analyze protein conformational changes. Protein Sci 1995;4:2118-2122.

[41] Korn AP, Rose DR. Torsion angle differences as a means of pinpointing local polypeptide chain trajectory changes for identical proteins in different conformational states. Protein Eng 1994;7:961-967.

[42] Kuznetsov IB, Rackovsky S. On the properties and sequence context of structurally ambivalent fragments in proteins. Protein Sci 2003;12:2420-2433.

[43] Kabsch W, Sander,C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577-2637.

[44] Halle B. Flexibility and packing in proteins. Proc Natl Acad Sci USA 2002;99:1274-1279.

[45] Cohen BI, Presnell SR, Cohen FE. Origins of structural diversity within sequentially identical hexapeptides. Protein Sci 1993;2:2134-2145.

[46] Ring CS, Kneller DG, Langridge R, Cohen FE. Taxonomy and conformational analysis of loops in proteins. J Mol Biol 1992;224:685-699.

[47] Tyagi M, de Brevern AG, Srinivasan N, Offmann B. Protein structure comparison using a structural alphabet. Proteins 2008;71:920-37.

[48] Mosca R, Schneider TR. RAPIDO: a web server for the alignment of protein structures in the presence of conformational changes. Nucleic Acids Res 2008;36:W42-W46.

[49] Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics 2003;19:ii246-ii255.

[50] Shatsky M, Wolfson HJ, Nussinov R. Flexible protein alignment and hinge detection. Proteins 2002;48:242-256.

[51] Rackovsky S, Scheraga HA. Differential geometry and polymer conformation. 1. Comparison of protein conformations. macromolecules 1978;11:1168-1174.

[52] Louie AH, Somorjai RL. Differential geometry of proteins: a structural and dynamical representation of patterns. J Theor Biol 1982;98:189-209.

[53] Montalvao RW, Smith RE, Lovell SC, Blundell TL. CHORAL: a differential geometry approach to the prediction of the cores of protein structures. Bioinformatics 2005;21:3719-3725.

[54] Morikawa N. Discrete differential geometry of tetrahedrons and encoding of local protein structure. ArXiv: 0710.4596; 2007.

[55] Jones NC, Pevzner PA. An introduction to bioinformatics algorithms. Cambridge, MA: MIT press; 2004.

[56] Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinformatics 2000;16:566-567.

[57] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536-540.

[58] Morikawa N. Number sequence representation of protein structures based on the second derivative of a folded tetrahedron sequence. ArXiv: q-bio.BM/0610017; 2006.

[59] O'Neill JW, Kim DE, Johnsen K, Baker D, Zhang KY. Single-site mutations induce 3D domain swapping in the B1 domain of protein L from Peptostreptococcus magnus. Structure 2001;9:1017-27.

[60] Jaskolski M. 3D domain swapping, protein oligomerization, and amyloid formation. Acta Biochim Pol 2001;48:807-27.

[61] Ding F, Prutzman KC, Campbell SL, Dokholyan NV. Topological determinants of protein domain swapping. Structure 2006;14:5-14.

[62] Ely KR, Herron JN, Harker M, Edmundson AB. Three-dimensional structure of a light chain dimer crystallized in water. Conformational flexibility of a molecule in two crystal forms. J Mol Biol 1989;210:601-15.

[63] Hanson BL, Bunick GJ, Harp JM, Edmundson AB. Mcg in 2030: new techniques for atomic position determination of immune complexes. J Mol Recognit. 2002;15:297-305.

[64] Branden C, Tooze J. Introduction to Protein Structure, 2nd ed. New York: Garland Publishing; 1999.

[65] Schiffer M. Possible distortion of antibody binding site of the Mcg Bence-Jones protein by lattice forces. Biophys J 1980;32:230-232.

[66] Terrak M, Wu G, Stafford WF, Lu RC, Dominguez R. Two distinct myosin light chain structures are induced by specific variations within the bound IQ motifs-functional implications. EMBO J 2003;22:362-371.

[67] Magnusson U, Salopek-Sondi B, Luck LA, Mowbray SL. X-ray structures of the leucine-binding protein illustrate conformational changes and the basis of ligand specificity. J Biol Chem 2004;279:8747-8752.

[68] Wlodawer A, Vondrasek J. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. Annu Rev Biophys Biomol Struct 1998;27:249?84.

[69] Shafera RW, Rheea S-Y, Pillayb D, Millerc V, Sandstromd P, Schapiroa JM, Kuritzkese DR, Bennettf D. HIV-1protease and reverse transcriptase mutations for drug resistance surveillance. AIDS 2007;21:215?223.

[70] Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, Shafer RW. Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res 2003;31:298-303.

[71] Shafer RW, Schapiro JM. HIV-1 drug resistance mutations: an updated framework for the second decade of HAART, AIDS Rev. 2008;10:67-84.

[72] Piana S, Carloni P, Rothlisberger U. Drug resistance in HIV-1 protease: flexibility-assisted mechanism of compensatory mutations. Protein Sci 2002;11: 2393?2402.

[73] Zoete V, Michielin O, Karplus M. Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. J Mol Biol 2002;315:21-52.

[74] Ohtaka H, Sch?n A, Freire E. Multidrug resistance to HIV-1 protease inhibition requires cooperative coupling between distal mutations. Biochemistry 2003;42:13659-66.

[75] Prabu-Jeyabalan M, Nalivaika EA, King NM, Schiffer CA. Viability of a drug-resistant human immunodeficiency virus type 1 protease variant: structural insights for better antiviral therapy. Virol 2003;77:1306-15.

[76] Ode H, Neya S, Hata M, Sugiura W, Hoshino T. Computational simulations of HIV-1 proteases-multi-drug resistance due to nonactive site mutation L90M. J Am Chem Soc 2006;128:7887 -7895.