Title: Systematic analysis of local flexibility of multiple-structure proteins

Author: Naoto Morikawa

Affiliation: GENOCRIPT

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

E-mail: nmorika@genocript.com

Short title: % 42 chars (limit 45 chars including spaces)
Flexibility of multiple-structure proteins

Keywords: % Ten key words (limit five to ten key words not used in the title)
conformational change; variable region; structural difference; local protein structure; domain-swapped dimer; misfolding disease; protein engineering; drug design; D2 code; discrete differential geometry.

Abstract: % 250 words (limit 250 words)
It is widely accepted that knowledge of protein flexibility is fundamental for an understanding of the mechanism of protein function. Because conformational changes of a protein are attributable to a small fraction of residues within the protein, identification of these regions is important for an understanding of protein dynamics and their function. In this paper, we propose a discrete differential geometrical technique, *D^2 encoding*, for analysis of local protein structures. After assessing the sensitivity and selectivity of the *D^2 code*, we applied the technique to identify regions of 60 multiple-structure proteins in which conformational changes take place. Due to the sensitivity of the D^2 code to the twisting of a Cα trace, the sources of structural differences were successfully pinpointed by comparison of D^2 codes. We found that a multiple-structure protein can be identified as a sequence of rigid subdomains of 14.4 residues on average, connected by variable regions with an average length of 2.0 residues. The variable regions are coupled to each other to induce global structural transitions between two forms of the same protein. We compared the results with those of DSSP state assignment, and rigid and flexible structural alignments. Among the 60 proteins, we were particularly interested in domain-swapped dimers, whose mechanism of formation has implications for the development of amyloid plaques observed in misfolding diseases such as Alzheimer's and Parkinson's disease. The implications of the results for protein engineering and drug design are also considered. The programs used and the data obtained are freely available.

% About 7,700 (6,600 + captions) words + 8 figures + 6 tables + 3 supplements

INTRODUCTION

It is widely accepted that knowledge of protein flexibility is fundamental for an understanding the mechanism of protein function. A substantial amount of research has been focused on integrating protein flexibility considerations into protein engineering and drug design.[1-13] In this paper, we have identified multiple-structure proteins identified by Kosloff and Kolodny[14] in which conformational changes take place. We discuss the implications of the results for protein engineering and drug design. For example, it is important to determine how local conformational changes are coupled to each other to induce global structural transitions in order to understand the mechanism of conformational changes observed in misfolding diseases, such as Alzheimer's, Parkinson's, and mad cow (BSE) disease,[15] as well as the mechanism of drug resistance due to non-active site mutations.

As shown by NMR-based 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 pre-existing conformations of the protein.

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 substates of the ensemble. 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 used successfully to describe the mechanism of communication between ligand-binding sites and the susceptibility of these binding sites to distal mutations.[24–27] In the context of the ensemble-based concept, 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 manifestations of the redistribution of conformational substates that are triggered by the propagation of a fluctuation. For example, allostery is a consequence of the redistribution induced by ligand binding.

One notable implication of this 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 can reach regions far away from the triggering site. Residues can be coupled to each other thermodynamically without any visible connection pathway and they can play a significant role in modulating the cooperative network. Thus, identification and characterization of the residues affected is important for our understanding of—and engineering of—protein functions.

In addition, it should be noted that subtle conformational changes are often essential for protein function. Because proteins undergo changes in the population of conformational

substates during the course of their biological function, energy barriers for transition between substates should be low in order to allow a rapid reaction.[3] For example, ligand binding and catalysis are generally performed on the micro- to millisecond time scale, which means that the collective motions of the Cα atoms involved in the transitions are in the pico- to nanosecond scale. It is also known that proteins can detect a conformational change as small as 1 Å. For example, in the aspartate receptor a conformational change of 1 Å at the ligand binding site is propagated to a cytoplasmic activation site located 100 Å away.[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 an understanding of protein dynamics and their function. In this study, using pairs of X-ray crystallographic structures that have been determined for the same protein and that contain significant structural differences, we examined the degree of local distortion that accounts for the conformational changes induced by various biological activities.

In order to identify local structural differences between X-ray crystallographic coordinates, it is necessary to consider 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. Secondly, we should quantify the structural differences between local backbone conformations.

As shown by the famous controversy concerning the artificial distortion of myoglobin upon CO binding,[34] inaccuracies in crystallographic structures are troubling.[35] According to Rejto and Freer, 25–30% of the surface of a protein is contact with protein molecules belonging to other crystal units.[36] In addition, the coordinate error of Cα atoms at loops and surface regions can be as great as 1.0 Å.[37] Moreover, due to the high solvent content, crystalline proteins are also dynamic and exhibit extensive, discrete conformational substates.[38] Proteins can 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 may be attributed to heterogeneity.

To quantify the structural differences between local backbone conformations, Flocco and Mowbray[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 determine 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 more 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 were computed by the DSSP program[43]. Two distinct conformations were identified if their Cα root-mean-square deviation (RMSD) was below a certain threshold. To remove poorly characterized fragments, they also used 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 because 30% of side chains can exist in multiple conformations and several side chain conformations frequently occur at residues with low B-factor.[38]

Other methods have also been used to quantify the structural differences between local backbone conformations. Because similar secondary structure assignments do not guarantee

structural similarity, and because there are often significant variations in peripheral loops, Cohen and coworkers[45] studied the structural plasticity of hexapeptide fragments. This was based on the virtual dihedral angle joining four consecutive Cα atoms, using not only secondary structure but also the backbone RMSD and a structural classification of loops.[46]

Template-based methods are not usually employed for identification of local structural differences because the template-based approximation of a fragment is not uniquely determined. 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 the method is not useful for our purposes because of the uncertainty introduced by the substitution matrix. As for flexible structural alignment tools such as RAPIDO,[48] FATCAT,[49] and FlexProt,[50] these are also not useful for our purposes because they often ignore subtle local differences between conformations.

In contrast to previous studies, we have not used the position of individual atoms or 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 (that is, we divide the background space into tetrahedrons) and discretize gradient vectors at Cα atoms. It should be noted that we cannot 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 also 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.

MATERIALS AND METHODS

In this study, we considered the Cα trace of a protein and we used five-tetrahedron sequences to identify the local structural features of a protein and to quantify the differences between two protein backbone conformations.

Encoding of a Sequence of Tetrahedrons

-- Figure 1 --

As mentioned in the Introduction, we divide a three-dimensional Euclidean space into a collection of tetrahedron blocks (Figure 1a) and connect adjacent blocks to form a sequence of tetrahedrons (Figure 1b).[54] Each block consists of four short edges and two long edges, where the ratio of their length is $(\sqrt{3})/2$. Of the six ways of connecting three blocks (Figure 1c), we use only four excluding two U-turns (Figure 1c, far right). We call the direction of the edge (bold line) of a block that is not contained in the connected blocks the *gradient* of the

block. In what follows, we have used bold arrows to indicate the gradient vector of a block, as in Figure 1d.

Because we forbid U-turns, there are only two ways to connect a new block to the tail of a tetrahedron sequence. For instance, let us consider a three-tetrahedron sequence (Figure 1d, left) where the tail block is colored white. Shown in the right-hand panel of Figure 1d are the two blocks (colored white) that can be connected to the tail block. It should be noted that a tail block assumes one of two gradient vectors: the same gradient vector as the predecessor or the other. We can thus use a {0, 1}-valued sequence to describe variation in a gradient vector along a tetrahedron sequence.

Suppose that the previous block is assigned a value of 0 (Figure 1d, left). Then, assign a value of 0 to the tail block if the block has the same gradient vector as the previous block (left in the right panel of Figure 1d). Otherwise, assign a value of 1 to the tail block (right in the right panel). That is, the value is changed if the gradient vector is changed. In the case of the three-tetrahedron sequence of Figure 1d, we obtain a binary sequence of 000 or 001, depending on the gradient of the tail block.

The D^2 encoding of Cα traces

Because we are concerned with local structural features, we encode the conformation of all five-Cα fragments (i.e. fragments of five Cα atoms) of a protein using a five-tetrahedron sequence. First, we represent the conformation of a five-Cα fragment by a five-tetrahedron sequence, where we allow rotation and translation of blocks to absorb the irregularity inherent in actual protein structures. Next, we assign the corresponding {0, 1}-valued sequence of length five, which are denoted as a base-32 number, to the center Cα atom of the fragment. We can then describe the conformation of a protein by arranging base-32 numbers in the order that the corresponding Cα atoms appear in the Cα trace. We call the base-32 number sequence the *D^2 code* of a protein. 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.

-- Figure 2 --

First, we define the gradient vector of the Cα trace at the i-th Cα atom C(i) 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 length of a shorter edge is about one-fifth of the average distance between the Cα atoms. The gradient vector OR of T(i) is the fourth from the left in Figure 1c. 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.

Once the spatial orientation of T(i) is fixed, the position and spatial orientation of the next tetrahedron T(i+1) is also uniquely determined. 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)) × Dir(C(i+1)) until the direction of Grad(T(i+1)) coincides with that of Dir(C(i+1)) (i.e. "turn" 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 uniquely determined. 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 encode the fragment C(i)C(i-1)C(i-2) starting from C(i), and obtain a {0, 1}-valued 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).

{0, 1}-valued sequences of length five are denoted by the corresponding base-32 numbers: *0*, *1*, ..., *9*, *A*, *B*, ..., *V*. For example, 00010 is denoted by "2", 01001 is denoted by "9", 01010 is denoted by "A", and so on.

Longest Common Subsequence and Length of Alignment

One of the simplest measures of sequence similarity is the length of the longest common subsequence (LCS). A *subsequence* of a character 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". Note that there may be more than one LCS between two strings. We use the length of the LCSs of two D^2 codes (D^2 code-LCSs) to quantify the differences between two protein backbone conformations. We call the ratio of "the length of D^2 code-LCS" to "the length of the whole chain minus four" the *D^2 code-LCS ratio*. Two residues at both termini are excluded from the computation because they are not assigned a D^2 code.

On the other hand, the structural similarity of two proteins is usually measured with the root-mean-square deviation (RMSD) of aligned residues after rigid or flexible structural alignment. The alignment length (AL) is the number of residues aligned and provides another measure of structural similarity, although RMSD and AL are dependent on each other and one of them could be improved at the expense of the other.

Datasets and Programs

-- Table I --

After assessing the sensitivity and selectivity of the D^2 code, we applied the D^2 encoding for analysis of the multiple-structure proteins identified by Kosloff and Kolodny.[14] Table 1 summarizes the datasets and programs used for each purpose.

To assess the sensitivity of the D^2 code, we used 66 crystallographic structures (space group $P6_1$) and 28 NMR models of HIV-1 proteases. HIV-1 protease is a homodimeric molecule consisting of two identical 99-residue polypeptide chains. The structures of the two monomers are almost identical and superimposed RSMDs between the two monomers are 0.1-0.6 Å for the $P6_1$ crystals and 0.5-1.2 Å for the NMR models. See supplement A for the PDB IDs of their coordinate files. We compared the D^2 codes of the two monomers of the same molecule with each other. We also compared the DSSP state[43] sequences of the monomers with each other. The D^2 encoding was carried out with the program ProteinEncoder, which computes the D^2 code from a PDB file and outputs the result in a ".code" file. DSSP state assignments were obtained from the PDB database of EMBL-EBI (http://www.ebi.ac.uk/msd/). We also used the DSSPcont server[56] to compute the DSSP state assignment if it was not available from the database. In addition, we examined whether any correlation exists between the number of D^2 code/DSSP state-assignment conflicts and the DaliLite Z-score, which is a measure of alignment-quality used by the DaliLite server.[57] As a general rule, a Z-score above 20 means that the two structures are definitely structurally similar, between 8 and 20 means that the two are probably structurally similar, between 2 and 8 is a gray area, and a Z-score below 2 is not significant.

To assess the selectivity of the D^2 code, we performed retrieval of the ASTRAL SCOP 1.73 (95%) dataset[58] for structurally similar amino acid fragments of three query chains: an HIV-1 protease monomer 2nphA (the $\alpha+\beta$ type, 99 residues) and two chains, d2hkjal (the mainly α type, 78 residues) and d1j7ma_ (the mainly β type, 60 residues) of the ASTRAL dataset. The D^2 code of all the chains of the ASTRAL dataset were computed by ProteinEncoder and saved in a ".code" file (6.2 MB): *target_ASTRAL173.code*. Retrieval of the database was then carried out with program ComSubstruct, which computes the exact length of the LCS of two D^2 codes from ".code" files and gives an example of D^2 code alignment. The top 200 D^2 code-similar fragments of the same length as the query chain are obtained by typing the following command: *ComSubstruct -l -o1 -s -w1.0 -b200 query_chain.code target_ASTRAL173.code*. Because some of the top 200 fragments overlap each other, we manually chose a fragment for each chain contained in the top 200 (or top 150) list. We obtained 50 fragments of 99 residues for 2nphA (top 200), 42 fragments of 78 residues for d2hkja1 (top 200), and 55 fragments of 60 residues for d1j7ma_ (top 150).

We used the DaliLite server to compute rigid structural alignment of a query chain and each of the 50, 42, or 55 fragments. We also used the FATCAT server[59] to compute flexible structural alignment of the pairs. We then examined the correlation between the length of D^2 code-LCS and the DaliLite Z-score as well as the correlation between the length of D^2 code-LCS and the FATCAT P-value. The Z-score is explained above. The P-value is used in the FATCAT server to evaluate the significance of structural similarity. Pairs of structures with a P-value of less than 0.05 are considered to be significantly similar by the server.

-- Table II --

For analysis of multiple-structure proteins, we used 158 pairs of crystal structures solved at a resolution of 2.5 Å 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 superimposed RMSD is greater than 6.0 Å. They are clustered into 60 classes based on amino acid sequence. For the purpose of statistical analysis, we chose a structure pair from each of the 60 clusters to avoid statistical bias caused by proteins with many structures. Table II shows the distribution of the causes for the structural differences observed for the 60 pairs. Computation of D^2 codes, LCSs of two D^2-codes, rigid structural alignment, and flexible structural alignment were performed with ProteinEncode, ComSubstruct, the DaliLite server, and the FACTCAT server, respectively. DSSP state assignments are obtained from the PDB database of EMBL-EBI (http://www.ebi.ac.uk/msd/). We also used the DSSPcont server to compute the DSSP alignment if not available from the database. Figures of protein backbones were prepared with program ProteinViewer.

Programs ProteinEncode, ComSubstruct, and ProteinViewer, and also the data obtained, are available from http://www.genocript.com (see the *PROGRAM* section and the *EXAMPLES (Encoding/Decoding) > Encoding Statistics* subsection).

RESULTS

Basic Properties of the D^2 Code

D^2 encoding of simple hypothetical fragments

-- Figure 3 --

Figure 3a demonstrates the 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 successfully detected bending of loops with the D^2 code. Figure 3b demonstrates the 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)[58] 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%), "B" (6%), "Q" (6%), "G" (5%), and "1" (1%).[59] "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. It should be noted that specific types of global conformational changes cannot be detected by local Cα trace analysis such as the one shown in Figure 3c.

Sensitivity of the D^2 code

-- Figure 4 --

We found that the D^2 code was as sensitive as the DSSP state, and it successfully identified the structural differences between the two monomers of the same HIV-1 protease molecules by comparing their D^2 codes. In total, 284 D^2 code-assignment conflicts were detected, seven of which were related to a pair of visually indistinguishable local structures (false positive). There appears to be a linear correlation between the number of D^2 code-assignment conflicts and the DaliLite Z-score (Figure 4a). As for the DSSP state, a total of 323 DSSP state-assignment conflicts were observed, but the number of DSSP state-assignment conflicts had no clear relationship with the DaliLite Z-score (Figure 4b).

Selectivity of the D^2 code

-- Figure 5 --

We successfully isolated amino acid fragments with similar structure within a few minutes on a notebook computer (2GHz Intel Core 2 Duo and 1GB 667MHz DDR2 SDRAM). Figure 5a, 5b, and 5c show the correlations between the DaliLite Z-score and the length of D^2 code-LCS of a query chain, and the fragments retrieved. By visual inspection, we found that all the pairs with Z-score above eight were structurally similar: 12 fragments for 2nphA, six fragments for d2hkja1, and six fragments for d1j7ma_. Figure 5d, 5e, and 5f show the correlations between the FATCAT P-value and the length of D^2 code-LCS. There were 12 fragments for 2nphA, 19 fragments for d2hkja1, and six fragments for d1j7ma_ that had a P-value of less than 0.05. All but 13 fragments for d2hkja1 had a Z-score above eight, where the 13 fragments had no clear similarity with the query chain. The plots show (1) that all the fragments with similar structure had a D^2 code-LCS ratio above 80%, and (2) that two fragments were structurally similar if their D^2 code-LCS ratio was greater than 85%.

Examples of Multiple-Structure Proteins: Domain-Swapped Dimers

Let us examine concrete examples of multiple-structure proteins before considering statistical analysis of 60 crystallographic structure pairs of the same proteins. We shall look at domain-swapped dimers, the mechanism of formation of which has implications for the development of amyloid plaques observed in misfolding diseases.[60,61] See supplement B for more examples.

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

PpL is a multi-domain protein from the bacterium *Peptostreptococcus magnus*; it is an immunoglobulin-binding protein that 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,[62] where a domain-swapped dimer is a homodimeric molecule that is obtained by replacing a fragment of a monomeric

protein by the same fragment from another. The asymmetric unit of its crystallographic structure (PDB ID 1k50) contains two monomers (chains A and C) and one domain-swapped dimer (chains 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 several human diseases such as Alzheimer's and Parkinson's.

-- Figure 6 --

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 and 43) (Figure 6). The hinge mechanism involves five residues (51–57) that induce a significant conformational change of the second β-hairpin. Changes at residues 42 and 43 are also attributable to the hinge motion. These changes are not thought to be a result of crystal packing because residues 42 and 43 have not moved in the same direction. We also observed a movement of the loop near residue 14, which preserved the D^2 code. The D^2 code of the loop was "00QAB00" and the movement was within the range of D^2 codes "Q" and "B". The distortion was probably caused by crystal packing because residues 13–15 project to the surface of the crystal unit.

Conformations of the hinge regions

Of the 60 structure pairs, nine of the 11 monomeric/homo-multimeric pairs consisted of a monomeric protein and a member of a domain-swapped dimer. Domain swapping has been repeatedly observed in a variety of proteins and is believed to result from destabilization due to mutations—as in the case of the PpL V49A mutant examined above—or changes in environment.[63]

Here we made comparisons between various types of hinge motions. These have implications for the design and evolution of proteins because (1) protein design often incorporates point mutations with increased or reduced stability[63], and (2) the monomer-dimer equilibrium can be controlled by mutations in the hinge region.[60]

-- Figure 7 --

Figure 7a 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 structural differences over the region were successfully quantified by the number of D^2 code-assignment conflicts. The image at the far right in Figure 7a is the hinge region of the V49A mutant inspected above, where the mutation occurred at the N terminus (bottom). All Cα atoms of the fragment except both termini undergo conformational changes upon dimerization. The motion seems to be less efficient and more prone to misfolding than others. On the other hand, the image on the far left of Figure 8a is an example of conformational changes that are not detectable by local Cα-trace analysis, i.e. rotation around a bond (See Figure 3c). The motion is simple and the

monomer-dimer equilibrium could show a larger propensity for dimerization than others. If we 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.

Figure 7b shows the DSSP state assignment of the same fragments. No clear relationship can be seen between the conformational variability and the DSSP state variability. For example, the hinge pair of the image located on the far left has more DSSP state-assignment conflicts than the hinge pair of the image located on the far right. However, the structural differences between the hinge pairs in the image on the far left are smaller than those in the image on the far right. That is, the number of DSSP state-assignment conflicts is not an adequate measure of the structural difference between two protein backbone conformations. The DSSP program actually uses not only geometrical features but also hydrogen bonding patterns and solvent exposure to assign one of eight states to each residue.[43]

Statistical Analysis of Multiple-Structure Proteins

Now let us examine the 60 crystal structure pairs of the same proteins as identified by Kosloff and Kolodny. First, we consider the dissimilarities between the pairs, detected as the difference in D^2 codes. We then consider the similarities between the pairs, captured by the length of D^2 code-LCS. Also analyzed is the frequency distribution of D^2 code-assignment conflicts in the case of variable regions involving one amino acid.

The number of D^2 code-assignment conflicts

-- Figure 8 --

Due to the sensitivity of the D^2 code to the twisting of a Cα trace, the sources of structural differences were successfully pinpointed by comparison of D^2 codes. Figure 8 shows the correlations between the DaliLite Z-score and the length of D^2 code-LCS of the 60 crystallographic structure pairs. All the pairs with a Z-score above eight except one had a D^2 code-LCS ratio above 80%, although there was no clear correspondence between the two values. The plot indicates that large structural differences are often caused by deformation of small regions of a protein, such as hinge motions. As for the FATCAT P-value, all but one had a P-value of below 0.05, where the P-value of the 1sfcD/1xtgB pair was 0.06.

-- Table III --

We divided the 60 structure pairs into five groups based on the D^2 code-LCS ratio and the DaliLite Z-score (Table III). According to the table, deformations upon binding (type E and F) are local (with D^2 code-LCS ratio > 80%) and cause relatively small overall changes of shape (Z-score ≥ 8), although some deformations of type F are global (with D^2 code-LCS ratio < 80%) and result in significant overall changes of shape (Z-score < 8). Deformations upon multimeric protein formation (type A, B, C, and D) are also local, but they could cause

significant overall changes of shape. In the case of hetero-multimeric proteins (type A), the deformation may be global, but it does not always result in significant overall changes of shape. Lipid-bound apolipoproteins (type G) are very flexible and undergo significant conformational changes. The others (type H) are evenly distributed among four groups.

-- Table IV --

Table IVa shows the conflicts in D^2 code assignment between two different conformations of the same proteins. Multiple-structure proteins are as flexible as NMR models of HIV-1 protease monomers, and 12% of the residues are assigned different D^2 codes. They can be (thermodynamically) identified with a sequence of D^2 code-rigid subdomains with an average length of 14.4 residues connected by D^2 code-variable regions with an average length of 2.0 residues. Of the four examples of multiple-structure proteins, Mcg is the nearest to the average. Mlc1p has only half the flexibility of the average. LBP has a similar profile to $P6_1$ crystallographic structures of HIV-1 protease monomers. As for HIV-1 protease monomers, NMR models are approximately twice as flexible as $P6_1$ crystallographic structures.

Table IVb shows the DSSP state-assignment conflicts between the same structure pairs as in Table IVa. DSSP state-variable regions are distributed more coarsely than D^2 code-variable regions, although the total numbers of conflicts are almost equal, i.e. 13% and 12% of residues respectively. NMR models of HIV-1 protease monomers are more flexible than multiple-structure proteins with respect to the DSSP state, and almost three times as flexible as $P6_1$ crystallographic structures. These data can be explained by the fact that DSSP states are assigned based on hydrogen bonding pattern, solvent exposure, and geometrical features.[43] In contrast, the D^2 code assignment is based solely on the Cα trace.

The lengths of D^2 code-LCSs

-- Table V --

Table V shows the lengths of alignments of the same structure pairs as in Table IV, given by three programs: DaliLite (rigid structural alignment), ComSubstruct (D^2 code alignment), and FATCAT (flexible structural alignment). The length of alignment increases in the order DaliLite, ComSubstruct, and FATCAT for the 60 structure pairs on average, although the ratios of alignment lengths of DaliLite and FATCAT are concentrated around 80% and 100%, and do not exist around 90%. The D^2 code-LCS ratio is more evenly distributed than the others, and it gives the most efficient measure of structural similarity.

Frequency distribution of D^2 code-assignment conflicts

-- Table VI --

Table VI shows the distribution of deformation types of variable regions involving 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 fragments of five Cα atoms centered on the residue. Thus, the table is concerned with deformation of the five-Cα fragment centered on a residue.

With regard to the 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 VIa. Note that both of these cases exhibit a kind of asymmetry along a Cα trace. In the case of deformation of extended strands, 29.1% are bent in the middle of an extended strand, i.e., a conversion between "0" and "R" (= **11**0**11**), 20.4% are bent at the N-terminal end, i.e., a conversion between "0" and "G" (= **1**0000) or between "0" and "O" (= **11**000), and 6.2% are bent at the C-terminal end, i.e., a conversion between "0" and "1" (= 0000**1**) or between "0" and "3" (= 000**11**). Thus, a bend at the N-terminal end occurs three times or more as frequently as a bend at the C-terminal end. In the case of deformation of helices, 4.9% are a conversion between "A" (= 01010) and C-cap "B" (= 010**11**), 3.4% are a conversion between "A" and "H" (= **10001**) which implies a kink in the middle of a helix, and 2.3% are a conversion between "A" and N-cap "Q" (=**11**010). That is, folding at the C-terminal end of helices occurs twice or more as frequently as folding at the N-terminal end. See supplement C for examples of the two cases: bend at the N-terminal end of an extended strand, and folding at the C-terminal end of a helix.

With respect to the DSSP state transition, 66% of the deformations involve no assigned structure (i.e. ".") and 20% involve turns (i.e. "T") as shown in Table VIb. In the case of deformation of no assigned structures, 40.3% are converted to a bend "S", 21.2% to a strand "E", and 4.1% to a bridge "B". However, no detailed information on the structure can be obtained from these figures.

DISCUSSION

Here we consider four issues: (1) sensitivity of the D^2 code, (2) artifacts due to crystal packing, (3) structural evolution of proteins, and (4) implications for protein engineering and drug design.

As shown in Figure 7a, we can use the D^2 code to pinpoint the very residues that induce twists in a protein backbone because the D^2 code is sensitive to the twisting of Cα traces, by definition. We actually detected even small twists in the flanking regions, as in the case of residues 42–43 of PpL (Figure 6), as well as the flanking regions of the CDR2 segment of Mcg (Supplement B). With 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 (Supplement B).

With respect to potential artifacts due to crystal packing, the influence of the latter on sensitivity appears to be low because small movements related to crystal contacts are often without twist, as in the case of residues 13–15 of PpL (Figure 6). 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 (Supplement B).

Thirdly, recall the domain-swapped dimmers inspected above. There are various types of hinges with D^2 code-assignment conflicts ranging from 0 Cα atom to 7 Cα atoms (Figure 5). The monomer-dimer equilibrium can be affected by the number of D^2 code-assignment conflicts: the smaller the number of conflicts, the more stable a domain-swapped dimer is. The number of D^2 code-variable Cα atoms in the hinge region thus gives a good indication of the stages of evolution of domain-swapped dimers.

Finally, let us consider the 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 selection of mutation points, i.e. D^2 code-variable regions more than two residues in length (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 in the hinge region if the mutation reduces the number of D^2-variable Cα atoms (Figure 7a). The D^2 code analysis could be also useful for rational drug design. For instance, the existence of long D^2 code-variable regions may play a role in conformational changes observed in misfolding diseases such as Alzheimer's and Parkinson's disease. Analysis of the distribution of D^2 code-variable regions may also lead to a more detailed description of the mechanism of multidrug resistance due to non-active site mutations.

CONCLUSION

We propose a discrete differential geometric technique, D^2 encoding, for analysis of local protein structures. After assessing the sensitivity and selectivity of the D^2 code, we applied the technique to identification of regions of 60 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 were successfully pinpointed by comparison of D^2 codes. We found that a multiple-structure proteins can be (thermodynamically) identified with a sequence of rigid subdomains 14.4 residues in length on average connected by variable regions 2.0 residues in length on average. The variable regions are coupled to each other to induce global structural transitions between two forms of the same protein. Regarding the location of variable regions, some locations—such as the N terminus of a β-strand or the C-cap of an α-helix—appear to be more favored than others.

From this work, there are several implications for protein engineering and drug design. The number of D^2 code-variable Cα atoms in the hinge region of a domain-swapped dimer can be a good measure of evolution of the dimer. The existence of long D^2 code-variable regions may play a role in conformational changes observed in misfolding diseases. In addition, D^2 code analysis of the structure of mutants may provide new insights into the problem of drug resistance due to non-active site mutations, and may lead to new ideas on how to design 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. Cα-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. Carter P, Andersen CA, Rost B. DSSPcont: continuous secondary structure assignments for proteins. Nucleic Acids Res 2003;31:3293-5.
57. Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinformatics 2000;16:566-567.
58. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL compendium in 2004. Nucleic Acids Res 2004;32:D189-D192
59. Ye Y, Godzik A. FATCAT: a web server for flexible structure comparison and structure similarity searching. Nucleic Acid Res 2004;32:W582-585.
58. 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.
59. Morikawa N. Number sequence representation of protein structures based on the second derivative of a folded tetrahedron sequence. ArXiv: q-bio.BM/0610017; 2006.
60. Jaskolski M. 3D domain swapping, protein oligomerization, and amyloid formation. Acta Biochim Pol 2001;48:807-27.
61. Yamasaki M, Li W, Johnson DJD, Huntington JA. Crystal structure of a stable dimmer reveals the molecular basis of serpin polymerization. Nature 2008;455:1255-1258.
62. 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.
63. Ding F, Prutzman KC, Campbell SL, Dokholyan NV. Topological determinants of protein domain swapping. Structure 2006;14:5-14.

FIGURES

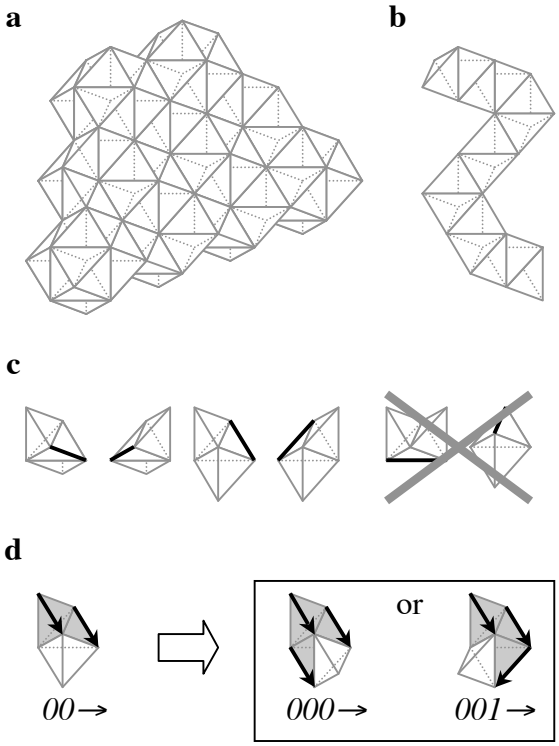

Figure 1: Sequence of tetrahedron blocks.
a. Division of a three-dimensional Euclidean space into a collection of tetrahedron blocks. **b.** Example of tetrahedron sequence. **c.** Six ways of connecting three blocks. We use only four of them, excluding two U-turns (far right). The direction of the bold edge specifies the gradient of the middle block. **d.** Encoding of variation in a gradient vector along a tetrahedron sequence. Bold arrows indicate the gradient vector of blocks. The binary sequences (shown underneath) describe the variation in a gradient vector of the gray blocks.

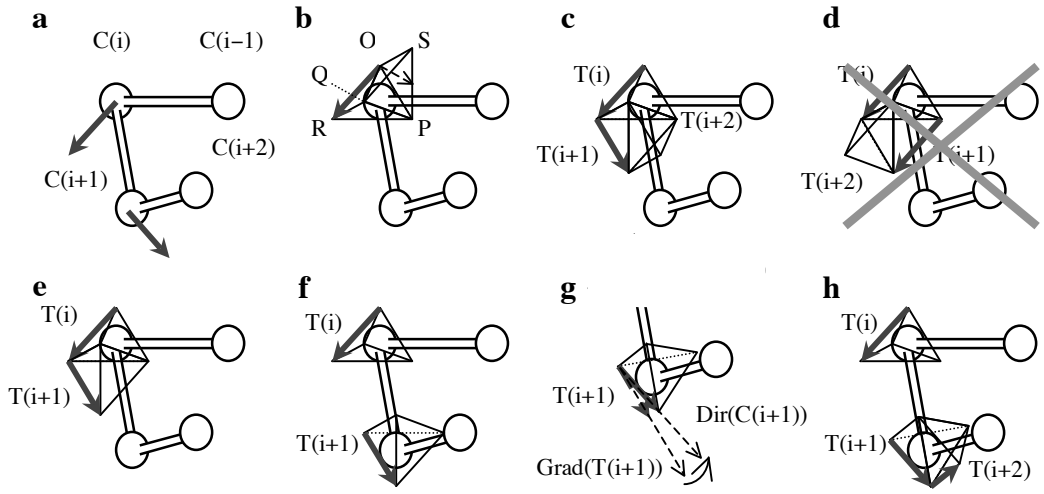

Figure 2: The D^2 encoding algorithm.
a. The Cα trace of a protein to be encoded. The arrows indicate the direction of a gradient vector of the Cα trace. **b.** Spatial orientation of the initial tetrahedron. **c**, **d.** Two permitted values of the gradient vector of T(i+1). **e.** The gradient vector 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.

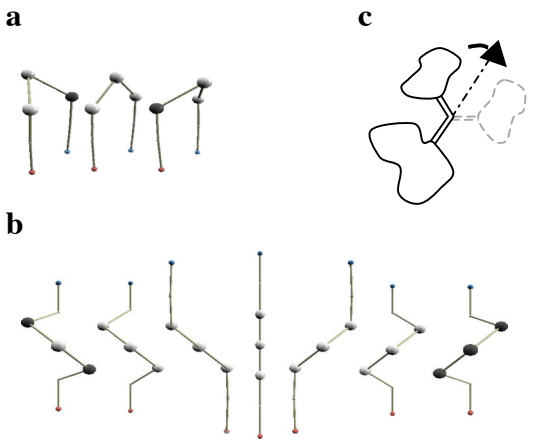

Figure 3: The D^2 encoding of simple hypothetical fragments.
a. Three loop fragments, where spheres indicate the positions of Cα atoms. From the left: a loop fragment whose curved region (i.e. the tip of the loop) is bent to the left side, a loop fragment that resides in a virtual flat plane, and a loop 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 positions of Cα atoms. From the 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 positions of the Cα atoms that have a D^2 code value other than "0". **c.** Example of global conformational changes that cannot be detected by local Cα-trace analysis. The rotation of the upper part around a bond cannot be detected by local Cα-trace analysis.

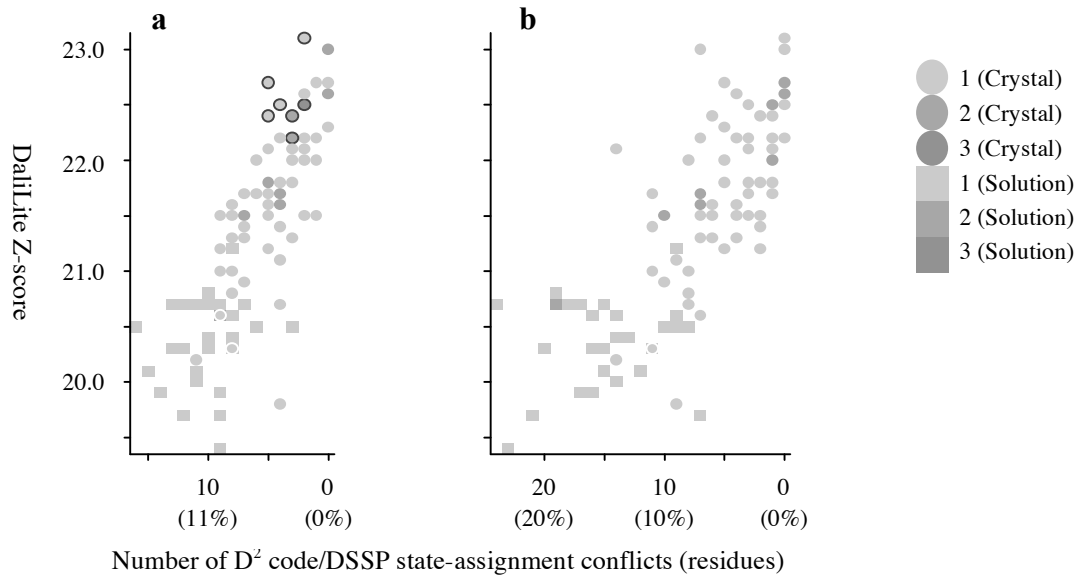

Figure 4: DaliLite Z-scores and D^2 code/DSSP state-assignment conflicts between the monomer pairs of 94 HIV-1 proteases. The plots show the correlations between the DaliLite Z-score and the number of **a.** D^2 code-assignment conflicts, and **b.** DSSP state-assignment conflicts. Circles represent the 66 crystallographic structures. Squares represent the 28 NMR models. Seven spheres encircled by a black line in the left figure indicate the positions of the HIV-1 monomer pairs with a D^2 code-assignment conflict that is related to a pair of visually indistinguishable local structures.

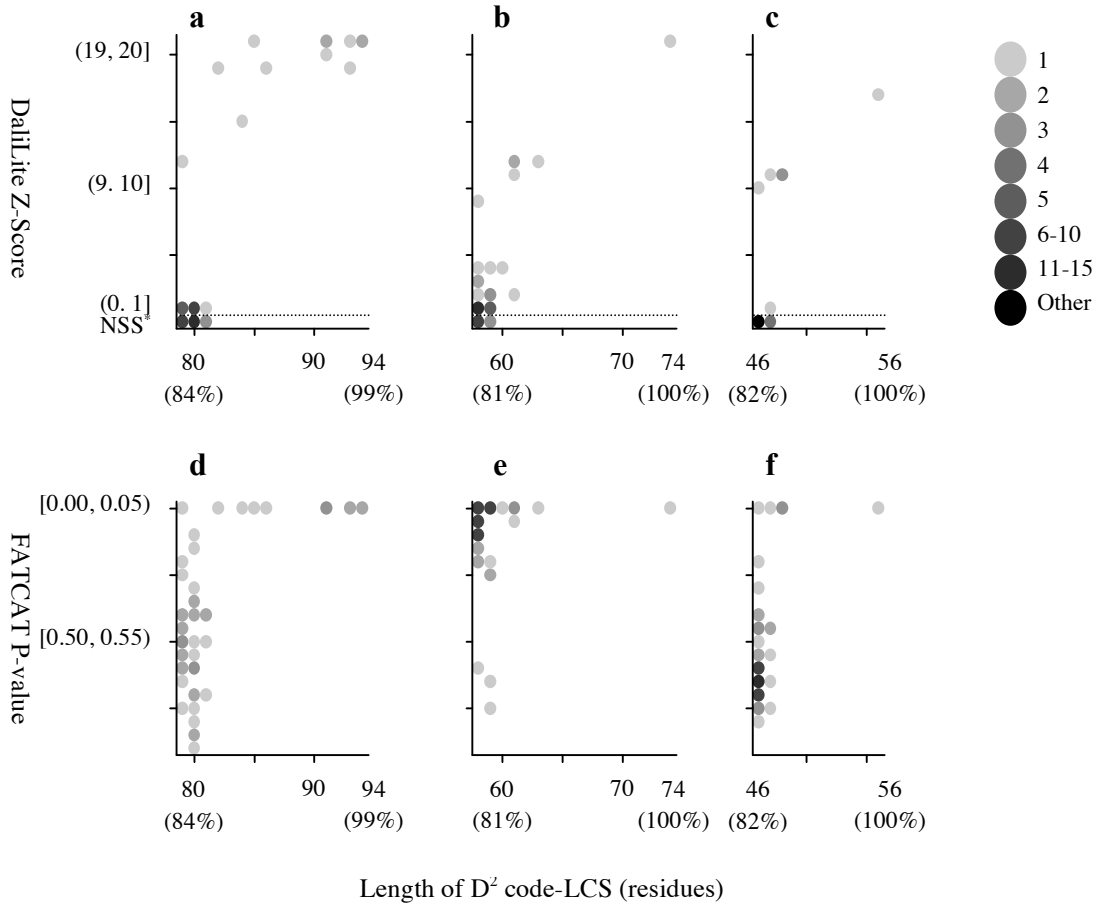

Figure 5: DaliLite Z-scores/FATCAT P-values and the length of D^2 code-LCS of the query chain and the retrieved fragments of the ASTRAL dataset. The top plots show the correlations between the DaliLite Z-score and the length of D^2 code-LCS of **a.** 2nphA (α+β) and 50 fragments; **b.** d2hkja1 (mainly α) and 42 fragments; and **c.** d1j7ma_ (mainly β) and 55 fragments. The bottom plots show the correlations between the FATCAT P-value and the length of D^2 code-LCS between **d.** 2nphA and 50 fragments; **e.** d2hkja1 and 42 fragments; and **f.** d1j7ma_ and 55 fragments. *NSS stands for Not Significant Similarity.

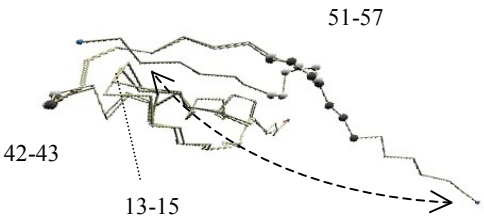

Figure 6: Superimposition of the Cα traces of monomeric (compact) and homodimeric (extended) PpLs. Black (extended) and white (compact) spheres indicate the position of the Cα atoms of residues 42–43 and 51–57, which assume two different D^2 codes. The peripheral loop movement at residues 13–15 does not change the D^2 code of the loop.

Figure 7: Cα traces of the hinge region of domain-swapped dimers.
Monomeric (compact) and homodimeric (extended) forms are superimposed manually, where the N terminus is at the bottom. Their PDB IDs are (from left to right): 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 positions 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 positions of the Cα atoms that assume two different DSSP states, whose numbers are 5, 3, 3, 8, 5, and 4 (from left).

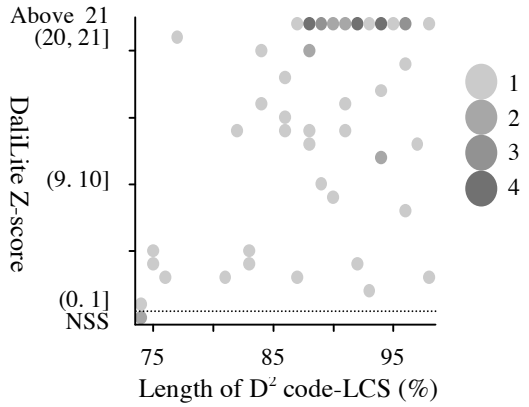

Figure 8: DaliLite Z-scores and the length of D^2 code-LCS of 60 structure pairs of mutiple-structure proteins. The plot shows the correlations between the DaliLite Z-score and the length of D^2 code-LCS.

TABLES

Table I: Datasets and programs used in the study

	Dataset	Program / Server				
		PrEnc[1] *(D^2 code)*	EBI[2] *(DSSP)*	ComSS[3] *(D^2 LCS)*	Dali[4] *(Z-score)*	FTCT[5] *(P-value)*
Sensitivity analysis	66+28 pairs of HIV-1 PR ($P6_1$+NMR)	√	√			
Selectivity analysis	2nph A, d2hkja1, d1j7ma ASTRAL (1.73 95%)	√		√	√	√
Multi-struct analysis	60 pairs identified by Kosloff & Kolodny	√	√	√	√	√

[1]ProteinEncoder.
[2]The PDB database of the EMBL-EBI or the DSSPcont server.
[3]ComSubstruct.
[4]The DaliLite server.
[5]The FATCAT server.

Table II: Frequency distribution of causes that account for the structural differences observed for the 60 crystallographic structure pairs used in this study

Type	#
A: Monomeric/Hetero-multimeric[1]	12
B: Monomeric/Homo-multimeric[2]	11
C: Members of dimer[3]	8
D: Members of oligomer[4]	8
E: Change upon ligand-binding[5]	7
F: Complex with different partners[6]	5
G: Lipid-bound apolipoproteins[7]	3
H: Other	6
Total	60

[1]Pairs of a monomeric protein and a member of a hetero-multimeric protein.
[2]Pairs of a monomeric protein and a member of a homo-multimeric protein.
[3]Pairs of the members of a homodimeric protein.
[4]Pairs of two members of a homo- or hetero-oligomeric protein.
[5]Pairs of a ligand-free form and a ligand-bound form of the same protein.
[6]Pairs of two conformations of the same protein from two different protein-ligand complexes.
[7]Pairs of two different lipid-bound forms of the same apolipoprotein.

Table III: Classification of 60 pairs of multiple-structure proteins. 60 structure pairs are divided into five classes based on the D^2 code-LCS ratio and the DaliLite Z-score (see Figure 8). The table shows frequency distribution for each type of pair defined in Table II

	D^2 code-LCS ratio below 80%	*D^2 code-LCS ratio above 80%*
Z-score above 20	A:1*	A:5, B:4, C:5, D:4, E:6, F:1, H:1
Z-score [8, 20]	None	A:5, B:5, C:2, D:2, E:1, F:2, H:2
Z-score below 8	A:1, F:2, G:2, H:1	B:2, C:1, D:2, G:1, H:2

*The pair of 1a1qA and 1jxpA.

Table IV: Assignment conflicts for the structural pairs inspected. The tables show the length of a protein, the percentage of assignment conflicts (D^2 codes (**a**) or DSSP states (**b**)), and the average lengths of rigid spans (spans without assignment conflicts) and variable spans (spans with assignment conflicts).

a

	Length (res.)	#Conf of D^2 (%)	Span length (res.)	
			Rigid	Var.
60 pairs[1]	212	12	14.4	2.0
PpL	63	14	18.0	4.5
Mcg[2]	216	13	9.8	1.6
Mlc1p[2]	147	5	17.4	1.1
LBP[2]	345	4	25.5	1.2
PR ($P6_1$)[3]	99	5	25.1	1.2
PR (NMR)[4]	99	11	11.3	1.4

b

	Length (res.)	#Conf of DSSP (%)	Span length (res.)	
			Rigid	Var.
60 pairs[1]	212	13	16.3	2.5
PpL	63	6	29.5	4.0
Mcg[2]	216	19	7.3	1.7
Mlc1p[2]	147	5	23.3	1.4
LBP[2]	345	4	41.5	1.9
PR ($P6_1$)[3]	99	5	33.2	1.7
PR (NMR)[4]	99	15	12.6	2.3

[1] The average of 60 structure pairs of multiple-structure proteins.
[2] Examples of the 60 structure pairs. See supplement B for detailed analysis of the proteins.
[3] The average of 66 $P6_1$ crystal structure pairs of HIV-1 protease monomers.
[4] The average of 28 NMR model pairs of HIV-1 protease monomers.

Table V: Alignment length of the structural pairs inspected. Shown are alignment lengths of three programs; the length of D^2 code-LCSs for ComSubstruct, and AL for DaliLite and FATCAT

	DaliLite (%)	ComSub (%)[5]	FATCAT (%)
60 pairs[1]	81	90	97
PpL	84	90	84
Mcg[2]	82	88	100
Mlc1p[2]	79	94	97
LBP[2]	100	96	100
PR ($P6_1$)[3]	100	95	100
PR (NMR)[4]	100	89	100

[1, 2, 3, 4] See the caption to Table IV.

[5]Two residues at both termini are removed from the computation because they are not assigned a D^2 code.

Table VI: Deformation types of variable regions that contain one amino acid.
Distribution of the D^2 code transition types (left) and the DSSP state transition types (right) observed in 60 crystallographic structure pairs of multiple-structure proteins. 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)
other	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)
other	37 (11.6)
all	320 (100)

[SUPPLEMENT A]

THE PDB FILES USED IN THE STUDY

66 $P6_1$ crystal structures of HIV-1 proteases

1a8g	1aaq	1axa	1bdl	1bdq	1bdr	1bv7	1bv9	1bwa	1bwb
1dmp	1fqx	1gnn	1gno	1hbv	1hos	1hps	1hpv	1htf	1hvh
1hvr	1hwr	1hxb	1izi	1lzq	1m0b	1mer	1mes	1met	1meu
1mui	1odx	1ody	1pro	1qbr	1qbs	1qbt	1qbu	1rl8	1sbg
1sgu	1sh9	1u8g	1vij	1zlf	1zpk	1zsf	1zsr	2aqu	2b60
2b7z	2fle	2nph	2p3a	2p3c	2p3d	2pym	2pyn	2q63	2q64
2qak	2qhc	2rkf	2z54	3d3t	9hvp				

28 NMR models of HIV-1 proteases

1bve (28 models)

60 structure pairs of multiple-structure proteins

#	PDB files	#	PDB files	#	PDB files	#	PDB files
1	1a1qA/1jxpA	16	1fguA/1jmcA	31	2ou1/1l6l6	46	1qvcA/1qvcB
2	1a8e_/1bp5C	17	1fm6A/1g1uB	32	2ou1/1l6l2	47	1qz3A/1u4nA
3	1akeA/4akeA	18	1fsgA/1qk3A	33	1l6lD/1l6lK	48	1sfcD/1xtgB
4	1aojA/1i0cA	19	1go4E/1go4F	34	1l9bM/1rg5M	49	1st0B/1st0A
5	1bka_/1cb6A	20	1gv2A/1h89C	35	1lyaA/1lywA	50	1su4A/1wpgA
6	1bmbA/1fyrB	21	1he7A/1wwwX	36	1m1gA/1m1gB	51	1usgA/1usiA
7	1brsB/1yvs_	22	1htmD/2viuB	37	1m46A/1n2dA	52	1vr4A/1vr4D
8	1c0bA/1f0vA	23	1hufA/1k46A	38	1m7gC/1m7hC	53	1y50A/1y51A
9	1cm1A/1g4yR	24	1ihgA/1iipA	39	1mi7R/1p6zR	54	1ygyA/1ygyB
10	1d2zC/1ik7A	25	1ihrB/1u07A	40	1mkmA/1mkmB	55	1zmeC/1zmeD
11	1dclA/1dclB	26	1iykA/1nmtA	41	1mxeA/1oojA	56	2beqD/2beqF
12	1ddt_/1mdtA	27	1iz1A/1iz1B	42	1oaoC/1oaoD	57	1bjmA/1bjmB
13	1esgB/2bamA	28	1k04A/1ow6B	43	1oc3A/1oc3C	58	1jvkA/1jvkB
14	1etsL/1ucyL	29	1k50A/1k50B	44	1opkA/2abl_	59	1sk4A/1twqA
15	1fdjA/1fdjC	30	1l5bA/3ezmA	45	1qexA/1s2eA	60	2mcg1/2mcg2

[**SUPPLEMENT B]**

SOME EXAMPLES OF MULTIPLE-STRUCTURE PROTEINS

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.[B1,B2]

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.[B3] 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".

-- Figure B1 --

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 B1a). The two domains are rather stable under the conformational change (RMSD 1.6 Å for the V domain and 1.0 Å 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 B1b). 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.[B4] 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.

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.[B5] 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.

-- Figure B2 --

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 B2 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 Å 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 B2 bottom), and cause a bend of the α-helix and a movement of a flanking loop in order to adjust the width of a channel that accommodates the IQ motifs (RMSD 0.9 Å 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).[B6]

Upon opening and closing, the two domains remain rather rigid (RMSD 0.6 Å for domain 1 and 0.5 Å 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.

-- Figure B3 --

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 B3a). 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 B3b). 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 of the paper). 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 Å, 3.8 Å, and 2.5 Å respectively by DaliLite (Figure B3b). As for the deformation of connection III, it is the type of movement shown in Figure 3c of the paper, 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.)

REFERENCES FOR SUPPLEMENT B

B1. 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.
B2. 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.
B3. Branden C, Tooze J. Introduction to Protein Structure, 2nd ed. New York: Garland Publishing; 1999.
B4. Schiffer M. Possible distortion of antibody binding site of the Mcg Bence-Jones protein by lattice forces. Biophys J 1980;32:230-232.
B5. 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.
B6. 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.

FIGURES OF SUPPLEMENT B

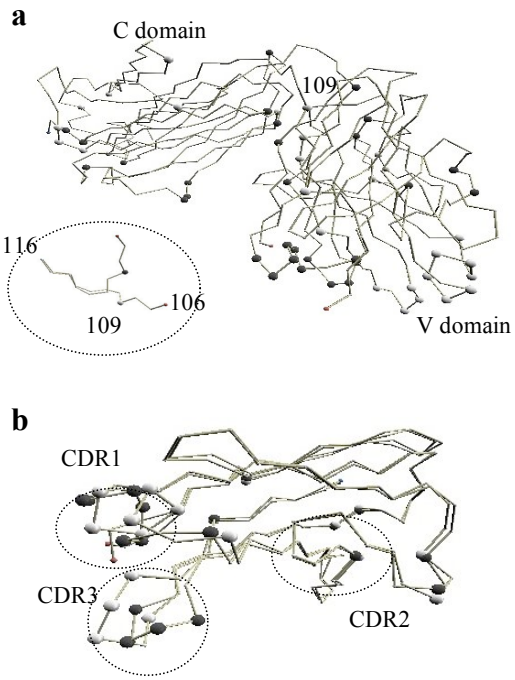

Figure B1: Cα traces of the chains of Bence-Jones protein Mcg.
a, Superimposition of chain A and B of Mcg. Encircled with a dashed line is a superimposition of the linker region. **b**, Superimposition of the V domains of chain A and B. Black (chain A) and white (chain B) spheres show the position of the Cα atoms which assume two different D^2 codes.

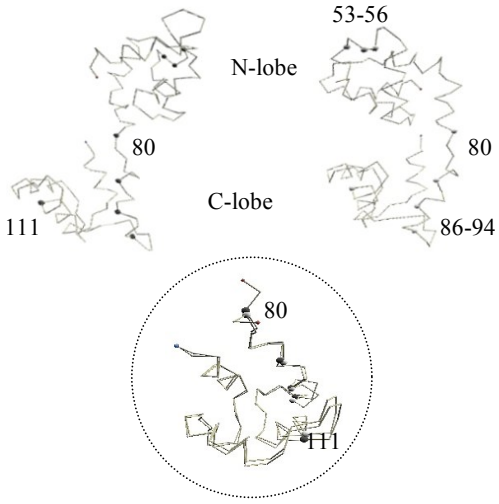

Figure B2: Cα traces of Mlc1p-IQ4 (top-left) and Mlc1p-IQ2 (top-right) protein-ligand complexes.
Neither IQ4 nor IQ2 is shown in the figure. Encircled with a dashed line is a superimposition of the two conformations of the C-lobe of Mlc1p. Black (Mlc1p-IQ4, extended) and white (Mlc1p-IQ2, compact) spheres show the position of the Cα atoms that assume two different D^2 codes.

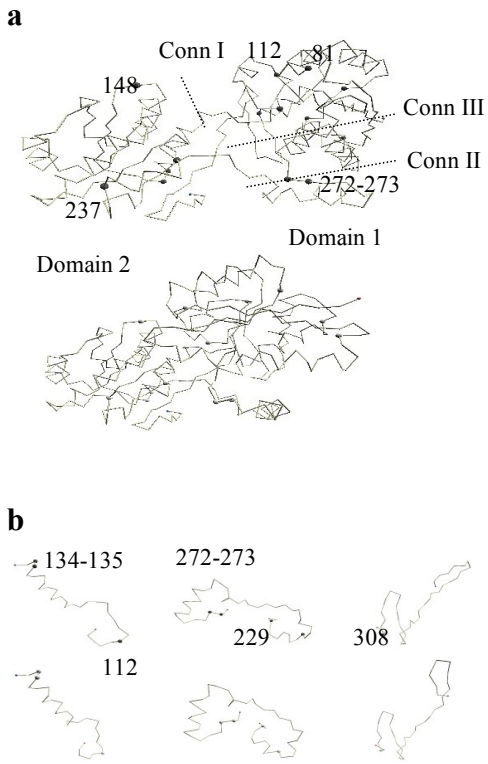

Figure B3: Cα traces of the open and closed forms of leucine-binding protein (LBP). **a**, The open (top) and closed (bottom) forms of LBP. **b**, The open (top) and closed (bottom) forms of the three connections between the two domains (From left, connection I, II, and III). Black (open) and white (closed) spheres show the position of the Cα atoms that assume two different D^2 codes.

[SUPPLEMENT C]

EXAMPLES OF D^2 CODE-ASSIGNMENT CONFLICTS

NMR models of a HIV-1 protease give examples of (1) bend on the N-terminal end of an extended strand, and (2) folding at the C-terminal end of a helix.

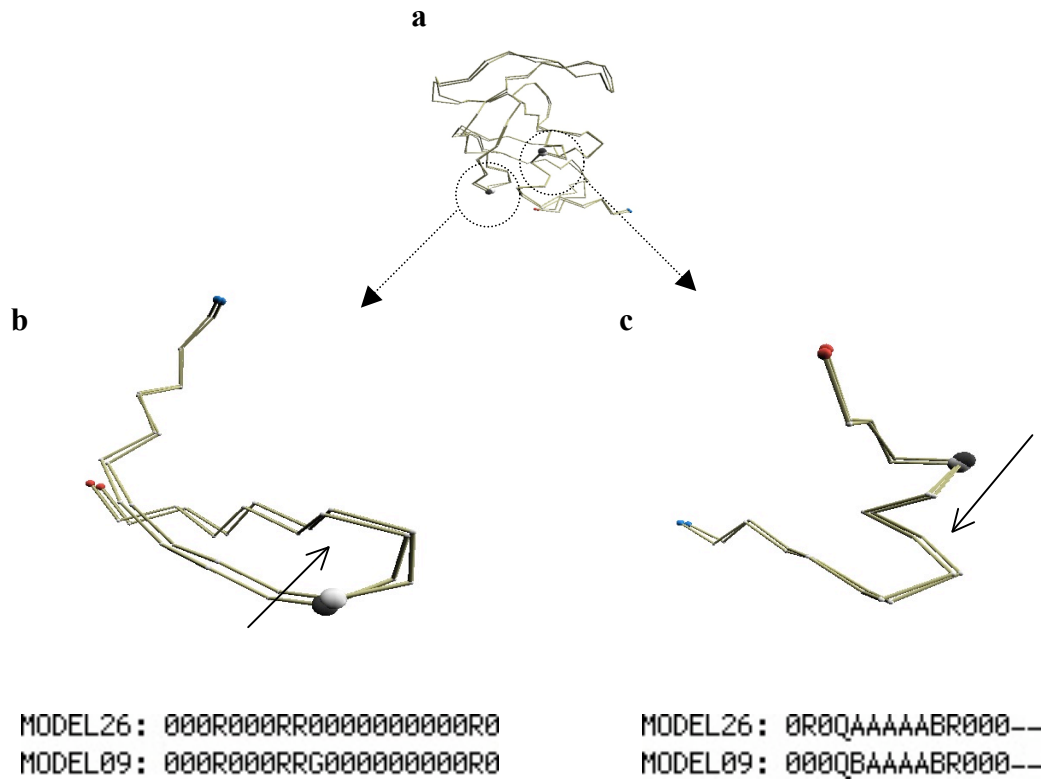

Figure C1: Bend on the N-terminal end of an extended strand and folding at the C-terminal end of a helix. Superimposition of Cα traces of NMR models of a HIV-1 protease monomer (1bveA model 9 and 26): whole structure (**a**); residues 60-80 and their D^2 codes (**b**); residues 84-99 and their D^2 codes (**c**). **b**, Black and white spheres indicate the position of residue 69 of model 26 and 9, respectively. Residue 69 is assigned two different D^2 codes: '0' (model26) and 'G' (model9). **c**, Black and white spheres indicate the position of residue 88 of model 26 and 9, respectively. Residue 88 is assigned two different D^2 codes: 'A' (model26) and 'B' (model9). The arrows indicate the direction of Cα atom movement from Model26 to Model9.

END of the file