A novel method for identification of local conformational changes in proteins

Naoto Morikawa
Genocript, Zama-shi, 228-0001, Japan

ABSTRACT

Motivation: Proteins are known to undergo conformational changes in the course of their functions. The changes in conformation are often attributable to a small fraction of residues within the protein. Therefore identification of these variable regions is important for an understanding of protein function.

Results: We propose a novel method for identification of local conformational changes in proteins. In our method, backbone conformations are encoded into a sequence of letters from a 16-letter alphabet (called D² codes) to perform structural comparison. Since we do not use clustering analysis to encode local structures, the D² codes not only provides a intuitively understandable description of protein structures, but also covers wide varieties of distortions. This paper shows that the D² codes are better correlated with changes in the dihedral angles than a structural alphabet and a secondary structure description. In the case of the N37S mutant of HIV-1 protease, local conformational changes were captured by the D² coding method more accurately than other methods. The D² coding also provided a reliable representation of the difference between NMR models of an HIV-1 protease mutant.

Availability: Program ProteinEncoder and detailed data are freely available at http://www.genocript.com.

Contact: nmorika@genocript.com

1 INTRODUCTION

Proteins are known to undergo conformational changes in the course of their functions. The changes in conformation may range from small local fluctuations to large domain motions. But they are often attributable to a small fraction of residues within the protein. Therefore an efficient and accurate algorithm for the identification of conformational changes is crucial for protein function analysis.

Over the past decade, numerous methods have been developed for flexible structural alignment. However current flexible alignment methods, such as RAPIDO (Mosca et al., 2008), FATCAT (Ye and Godzik, 2003), FlexProt (Shatsky et al., 2002), and DynDom (Hayward et al., 1997) often ignore subtle local differences between conformations.

To quantify the structural differences between local backbone conformations, a great variety of methods have been proposed. In those studies, the backbone conformation is often described using various geometric descriptors, such as, backbone dihedral angles (ψ, η), dihedral angles of the Cα trace, or distances derived from the positions of Cα atoms (Korn and Rose, 1994; Kuznetsov and Rackovsky, 2003; Miao et al., 2008; Flocco and Mowbray, 1995; Francisco et al., 2007; Kelley et al., 1997; Schneider, 2000; Rackovsky and Scheraga, 1984) Using these methods, one can pinpoint differences between two conformations of the same protein. But their accuracy are limited by coordinate errors associated with crystal structures, distortions due to crystal packing, and others. As for (ψ, η), significant fluctuations of neighbouring dihedral angles do not necessarily mean a change in the Cα trace because of the anticorrelation between them (McCammom et al., 1977).

Another approach for quantification of structural differences is based on secondary structure assignments, where the variable regions of a protein are identified as the regions of variable secondary structure (Chou and Fasman, 1974). However secondary structures in proteins usually have a number of distortions such as bends, twists, and end fraying. In other words, proteins can change their conformation without changing their secondary structure.

To capture local backbone conformations more accurately, "structural alphabet" has been proposed in various studies, where descriptions by the traditional three-state secondary structure are replaced with descriptions by a library of local structure prototypes (de Brevem et al., 2000; de Brevem et al., 2001; Kolodny et al., 2002; Camproux et al., 2004; Wang et al., 2005; Yang and Tung, 2006; Pandini et al., 2010). A structural alphabet is usually obtained by clustering protein backbone fragments based on their conformations, where local protein structures are described differently by different groups. According to Le et al. (2009), 1D representations based on a structural alphabet perform significantly better than the 1D representation based on the secondary-structure for fold recognition purposes, but not enough to become a viable replacement to computationally intensive procedures such as structural alignment tools.

Finally, Chang et al. (2006) proposed a local geometric measure, the “writhing number,” which is obtained from the knot theory. By encoding the distribution of writhing numbers across all the structures in the PDB, protein geometries are represented in a 20-letter alphabet.

In the following we propose a simple encoding method which translates the conformation of the Cα trace into a sequence of letters from a 16-letter alphabet (called D² codes), which is based on the discrete differential geometry of tetrahedron sequences. In our method, five-residue fragments are first approximated by a tetrahedron sequence. Then, the fragments are assigned one of the 16
letters, using the second derivative of the corresponding tetrahedron sequence. Since we do not use clustering analysis to encode local structures, this approach not only provides a intuitively understandable description of protein structures, but also covers wide varieties of distortions. Regarding structural alignment, one can use naive sequence alignment algorithms to compare protein structures. In SHREC’10 Protein Model Classification we achieved results comparable to more sophisticated methods, using the length of the longest common subsequence as the measure of structural similarity (Mavridis et al., 2010).

2 METHODS

2.1 The D^2 coding of backbone conformations

In contrast to previous studies, we don’t use the position of individual atoms nor secondary structure to identify local structural features. Instead, we approximate the backbone conformation by a folded tetrahedron sequence and consider the second derivative of the tetrahedron sequence.

2.1.1 Folded tetrahedron sequence To encode local protein structure, we use space filling tetrahedra. Each of the tetrahedra consists of four short edges and two long edges, where the ratio of their length is √3:2. Connecting tetrahedra which share a face one after another, we obtain a folded tetrahedron sequence (Figure 1A). Although there are six ways to connect three tetrahedra (Figure 1B), we forbid U-turns (Figure 1B, bottom). As a consequence, there are only two ways to add a new tetrahedron to the tail of a tetrahedron sequence. For instance, let’s consider a three-tetrahedron sequence in Figure 1C (left), where the tail tetrahedron is colored white. Shown in the right-hand panel of Figure 1C are the two tetrahedra (colored white) that can be added to the tail.

To encode the conformation of a tetrahedron sequence, we consider the “gradient” of a tetrahedron, which is the direction of the edge of the tetrahedron that is not shared with the adjacent tetrahedra (Bald lines in Figure 1B). Note that a “current” tetrahedron assumes one of two gradients: the same gradient as the “predecessor” or the other. We use bold arrows to indicate the gradient of a tetrahedron, as in Figure 1C.

Now we can describe variation in a gradient (i.e., the "second derivative") along a tetrahedron sequence by a \{0, 1\}-valued sequence. That is, change the value if the gradient is changed. Suppose that the previous tetrahedron is assigned a value of 0 (Figure 1C, left). Then, assign a value of 0 to the current tetrahedron if the tetrahedron has the same gradient as the predecessor (left in the right panel of Figure 1C). Otherwise, assign a value of 1 to the current tetrahedron (right in the right panel). In the case of the three-tetrahedron sequence of Figure 1C, we obtain a binary sequence of 000 or 001, depending on the gradient of the tail tetrahedron.

2.1.2 Encoding algorithm In the first place, each of the five-residue fragments in a protein is approximated by a five-tetrahedron sequence. To absorb the irregularity inherent in actual protein structures, we allow translation and rotation of tetrahedra in the process as described below. Next, we compute the second derivative of the tetrahedron sequences to obtain \{0, 1\}-valued sequences of length five. In this study \{0, 1\}-valued sequences are denoted as a base-32 number 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. Then, we assign the base-32 numbers to the center Ca atom of the corresponding five-residue fragment. By arranging the base-32 numbers in the order they appear in the Cα trace, we obtain a 1D representation of protein structures. We call the base-32 number sequence the “D^2 code” sequence of a protein.

As an illustration, let’s consider the Cα fragment C(i)C(i+1)C(i+2) shown in Figure 2A. The gradient of the Cα fragment 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+2) (Rackovsky and Scheraga, 1978). In the following, we denote the vector from point A to point B by AB.

The initial tetrahedron T(1), OPQR in Figure 2b, is given as follows. T(1) is aligned with C(i) in such a way that the gradient of T(1) is parallel to the gradient of the Cα trace at C(i), and (2) the direction of the vector OS + OP and the vector C(i)C(i+1) + C(i+1)C(i+2) coincide, where S (= O + RQ) is a vertex of an adjacent tetrahedron. By definition T(1) is always assigned a base-32 number because of T(1)C(i+2). Encoding the fragment C(i)C(i+1)C(i+2) of Figure 2C is closer to C(i+1) than T(i+2) in Figure 2D. Since the gradient 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 moved along the Cα trace to absorb structural irregularities of proteins: T(i+1) is translated to the position of C(i+1) (Figure 2F) and rotated there (Figure 2G). To be more precise, T(i+1) is rotated around the cross-product Grad(T(i+1)) × Grad(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 of T(i+1) and Grad(C(i+1)) is the gradient 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+1) are determined uniquely. On the other hand, we have two choices for the gradient of T(i+1), which are shown in Figure 2C and 2D. Then the gradient of T(i+1) is determined based on the distance between T(i+2) and C(i+1). In the current case, we choose the gradient shown in Figure 2C (Figure 2E) because T(i+2) in Figure 2C is closer to C(i+1) than T(i+2) in Figure 2D. Since the gradient 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 moved along the Cα trace to absorb structural irregularities of proteins: T(i+1) is translated to the position of C(i+1) (Figure 2F) and rotated there (Figure 2G). To be more precise, T(i+1) is rotated around the cross-product Grad(T(i+1)) × Grad(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 of T(i+1) and Grad(C(i+1)) is the gradient 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+1) are uniquely determined. But we have two choices for the gradient of T(i+1). The gradient of T(i+2) is determined based on the distance between T(i+2) and C(i+1) (Figure 2H). As the gradient T(i+2) is different from that of T(i+1), a value of 0 is assigned to T(i+2).

Then, we obtain the binary sequence of 010, which describes the variation in a gradient along the fragment C(i)C(i+1)C(i+2). Encoding the fragment C(i)C(i+1)C(i+2) starting from C(i) in the same way, we obtain a \{0, 1\}-valued sequence of length five.
The ANGL regions give the residue positions with variable dihedral angles. The ANGL_SUPP regions are the “support” of the variable region, i.e., all the variable positions should be contained in the region. And the ANGL_CORE regions are the “core” of the variable region.

2.2.2 Benchmark datasets Three datasets are employed for the comparison study. The first is the set of 60 multiple-structure proteins identified by Kosloff and Kolodny (Kosloff and Kolodny, 2008). The dataset was chosen because the protein structures were aligned based solely on sequence information. The second is a set of 72 crystal structures of the N37S mutant of HIV-1 protease (PR), which is a homodimeric molecule, consisted of two identical 99-residue chains. (See supplement 1 for the PDBIDs of the 72 structures). HIV-1 PR is one of the major anti-HIV-1 drug targets (Wlodawer and Vondrask, 1998) and a large collection of crystal structures of its variants are available in the PDB. The third is the 28 NMR models of an HIV-1 PR mutant [PDB:1bve]. Note that NMR structures are more prone to local distortions than X-ray structures because they correspond to proteins in solution. Therefore, they provide interesting cases for local structure encoding methods.

2.2.3 Performance measures To measure the performance of each method, we calculated accuracy (ACC), sensitivity (SN), specificity (SP), and Matthews correlation coefficient (MCC) using the following formulas:

\[
\text{ACC} = \frac{TP + TN}{TP + FP + TN + FN}
\]

\[
\text{SN} = \frac{TP}{TP + FN}
\]

\[
\text{SP} = \frac{TN}{TN + FP}
\]

\[
\text{MCC} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}
\]

To measure the performance of each encoding method, we calculated accuracy (ACC), sensitivity (SN), specificity (SP), and Matthews correlation coefficient (MCC). For the ACC, SN, and MCC, we used the following formulas:

\[
\text{ACC} = \frac{TP + TN}{TP + FP + TN + FN}
\]

\[
\text{SN} = \frac{TP}{TP + FN}
\]

\[
\text{SP} = \frac{TN}{TN + FP}
\]

\[
\text{MCC} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}
\]

3 RESULTS AND DISCUSSION

3.1 Conformationally variable regions

First the D^2 coding method was compared with the PB coding and the DSSP coding, using the 60 structure pairs identified by Kosloff and Kolodny and the three gold standards.

On average, 53.8% of all the residue positions in a protein undergoes changes in the dihedral angles (the ANGL_SUPP regions). Significant changes in the dihedral angles are observed in 17.8% of all the positions (the ANGL regions). Very large changes in the dihedral angles are observed in 10.2% of all the positions (the ANGL_CORE regions). On the other hand, 15.4% of all the residue positions assume multiple D^2 codes, 18.0% assume multiple DSSP codes, and 19.8% assume multiple PB codes.

The D^2 coding method is most specific with respect to the ANGL_SUPP regions. The PB coding method is most sensitive with respect to the ANGL_CORE regions. And the D^2 coding is most accurate with respect to the ANGL regions. Concerning MCC, the PB codes correlated better with the ANGL_CORE regions than others. The D^2 codes correlated better with the ANGL regions than others. But none of the three codes correlated well with the ANGL_SUPP regions. See supplement 2 for the ACC, SN, SP, and MCC computed against each of the three gold standards.

Recall that large changes in the dihedral angles do not necessarily mean an obvious change in the Ca trace because of the anti-correlation between \(\phi\) and \(\psi\). In particular, parts of the ANGL_CORE regions are not included in the D^2, PB, and DSSP...
variable regions. Therefore the values of MCC against the ANGL_CORE regions are not high. On the other hand, a residue position without any change in the dihedral angles may assume multiple D$^2$ or PB codes because they are computed from the conformation of a fragment centered on the residue. This is reflected in their low MCCs with respect to the ANGL_SUPP regions.

Since the D$^2$ encoding is based on the discrete differential geometry of tetrahedron sequences, each of 16 D$^2$ codes has an intuitive meaning and they cover wide varieties of distortions. That is, "0" (=00000) is extended strand, "A" (=01010) is helix, "B" (=01011) is helix C-cap, "Q" (=11010) is helix N-cap, "R" (=11011) is turn, and the others correspond to varieties of distortions. In contrast, the letters in structural alphabets, such as the PB variable region, often lack intuitive meaning and distortion. That is because they are obtained by clustering fragments of protein backbone based on their conformations. In particular, both helices and extended strands correspond to a number of codes. Therefore one should compute a scoring matrix for substitutions to align structural alphabet sequences. Figure 3 shows the frequency distribution of the D$^2$, PB, and DSSP codes.

Consequently the D$^2$ variable region is usually much smaller than the PB variable region (15.4% < 17.8% (ANGL) < 19.8%). But it was shown that the D$^2$ codes correlated better with the ANGL regions than the PB codes (44.9% > 43.6%). As a result, one can perform D$^2$ code-sequence alignment without a substitution matrix. In SHREC’10 Protein Model Classification (Mavridis, 2010), we achieved results comparable to more sophisticated methods, using the length of the longest common subsequence (LCS) as the measure of structural similarity.

3.2 Difference between mutants

Next the D$^2$ coding method was compared with the PB coding and a ($\phi$, $\psi$) dihedral angle-based method, using 72 crystal structures (142 chains) of the N37S mutant.

We first computed the major D$^2$ (and PB) codes for each residue position, which describe a typical local conformation of the 142 chains. Then we considered the spatial distribution of the minor D$^2$ (and PB) codes along the backbone (Figure 4).

On average, there are 2.5 minor D$^2$ codes, 7.1 minor PB codes, and 7.3 residues with large deviation (>30°) in $\phi$ or $\psi$ per chain. About 40% of the minor D$^2$ codes are assigned at residue 50 or 51, and about 15% are assigned at residue 5. The deformation at residue 50 and 51 is explained by variation of the size of the ligands bound. The deformation at residue 4 is provably explained by crystal packing. As for the PB coding, not only around residue 50, but also residue 30, 40, and 71-80 often assume a minor PB code. On the other hand, the $\phi$ or $\psi$ angle of residues 40, 50, 51, 73, 78, and 86 are almost always deviated from the average more than 30°.

As examples, let’s consider the local structure of 1ajx chain A and chain B around residue 40 and residue 73 (Figure 5). The $\phi$ angles of residue 40 of 1ajxA and 1ajxB are 133.7° and 101.5°, respectively. That is, the difference of their $\phi$ angles is 32.2°. The difference of their $\psi$ angles is 15.0°. According to the alignment computed by the DALI server, the fragments from residue 25 to residue 55 of 1ajxA and 1ajxB are in very similar conformations (RMSD 0.5Angstrom). Therefore the change in dihedral angles does not represent a change in the Cu backbone.

The two chains have an identical D$^2$ code sequence (“00000”), but different PB code assignments (“EHIAD” and “EHIJAC”) around residue 40.

The $\phi$ angles of residue 73 of 1ajxA and 1ajxB are 177.6° and 170.2° respectively. Because of periodicity, the difference between them is 12.2°. According to the alignment computed by the DALI server, the fragments from residue 58 to residue 88 of 1ajxA and 1ajxB are in also very similar conformations (RMSD 0.3Angstrom). The fragments have an identical D$^2$ code sequence (“00000”), but different PB code assignments (“EHIJAC” and “DFBLC”) around residue 73.

In both cases, the D$^2$ code sequence captured the local conformation more accurately than the others. Regarding the ($\phi$, $\psi$) dihedral angle-based method, it did not provide a reliable representation of protein local structures.
3.3 Difference between NMR models

Lastly the Δ² coding method was compared with the PB coding and the (ψ, ψ) dihedral angle-based method used above, using 28 NMR models (56 chains) of a liganded HIV-1 protease.

On average, there are 8.6 minor Δ² codes, 19.2 minor PB codes, and 26.8 residues with large deviation (>30°) in ψ or ψ per chain. In particular, more than one-fourth of all the residues undergo large deviation in the dihedral angles. Therefore it is difficult to characterize the difference between the NMR models in terms of the dihedral angles.

Concerning the Δ² coding, about 22% are assigned at the loop around residue 40, 9% are assigned at the loop around residue 7, and 9% are assigned at the loop around residue 64. The deformations are provably explained by collision with another molecule. Note that no minor Δ² code is assigned at the loop around residue 50, which is reasonable because all the NMR models have the same ligand. On the other hand, minority PB codes are assigned not only at the loops around residue 40 and around residue 7, but also at residue 51 and 52.

4 CONCLUSIONS

We proposed a novel method for identification of local conformational changes in proteins. In our method, protein backbone conformations are encoded into a sequence of letters from a 16-letter alphabet (called Δ² codes) to perform structural comparison. Since we do not use clustering analysis to encode local structures, the Δ² codes not only provide an intuitively understandable description of protein structures, but also covers wide varieties of distortions.

Three datasets are employed to compare the performance of our method with three existing coding methods: a structural alphabet (the PB blocks), a secondary structure description (the DSSP assignments), and a dihedral angle-based method. The first dataset is 60 multiple-structure proteins identified by Kosloff and Kolodny (2008). The second is 72 crystal structures of the N37S mutant of HIV-1 protease (PR). And the third is 28 NMR models of a HIV-1 PR mutant.

The results show that the Δ² codes are better correlated with changes in the dihedral angles than the PB blocks and the DSSP assignments although the Δ² coding method identifies much smaller variable regions than the other methods. In the case of the N37S mutant of HIV-1 PR, the local conformational changes were captured by the Δ² coding more accurately than the others. The Δ² coding method also provided a reliable representation of the difference between the dynamic NMR models of a HIV-1 PR mutant.

Conflict of Interest: none declared.

REFERENCES

Camproux,A.C. et al. (2004) A Hidden Markov Model derived structural alphabet for proteins. J. Mol. Biol., 339, 591-605.

Carter,P. et al. (2003) DSSPpoint: continuous secondary structure assignments for proteins. Nucleic Acids Res., 31, 3293-3295.

Chang,P.L. et al. (2006) Structure alignment based on coding of local geometric measures. BMC Bioinform., 7, 346.

Chou,P.Y. and Fasman,G.D. (1974) Prediction of Protein Conformation Biochemistry, 13, 222-245.

de Brevemm,A.G. et al. (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins, 41, 271-287.

dde Brevemm,A.G. et al. (2001) Protein structural alphabets: beyond the secondary structure description. Recent Adv. In Prot. Eng., 1, 319-331.

Flocco,M.M. and Mowbray,S.L. (1995) C-alpha-based torsion angles: a simple tool to analyze protein conformational changes. Protein Sci., 4, 2118-2122.

Francisco,S. et al. (2007) Conformational analysis of alternative protein structures. Bioinformatics, 23, 3131-3138.

Hayward,S. et al. (1997) Model-Free Methods of Analyzing Domain Motions in Proteins from Simulation: A Comparison of Normal Mode Analysis and Molecular Dynamics Simulation of Lysozyme. Proteins, 27, 425-437.

Kabish,W. and Sunderman,F. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22, 2577-2637.

Kelley,L.A. et al. (1997) An automated approach for defining core atoms and domains in an ensemble of NMR-derived protein structures. Protein Eng., 10, 737-741.

Kolodny,R. et al. (2002) Small libraries of protein fragments model native protein structures accurately. J. Mol. Biol., 323, 297-307.

Korn,A.P. and Rose,D.R. (1994) Torsion angle differences as a measure of similarity between protein structures. J. Mol. Biol., 323, 297-307.

Kosloff,M. and Kolodny,R. (2008) Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins, 71, 891-902.

Kuznetsovb,J. et al. (2008) Conformational change in the protein backbone: prediction of conformationally variable positions from sequence and low-resolution structural data. Proteins, 72, 74-87.

Kuznetsovb,J. and Rackovsky,S. (2003) On the properties and sequence context of structurally ambivalent fragments in proteins. Protein Sci., 12, 2420-33.
Schneider, T.R. (2000) Objective comparison of protein structures: error-scaled difference distance matrices. Acta Cryst. D56, 714-721.

Shatsky, M et al. (2002) Flexible protein alignment and hinge detection. Proteins, 48, 242-256.

Tyagi, M et al. (2006) Protein Block Expert (PBE). A web-based protein structure analysis server using a structural alphabet. Nucleic Acids Res., 34, W119-W123.

Wlodawer, A and Vondrasek, J. (1998) Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. Annu. Rev. Biophys. Biomol. Struct., 27, 249-284.

Yang, J.M. and Tung, C.H. (2006) Protein structure database search and evolutionary classification. Nucleic Acids Res., 34, 3646-3659.

Ye, Y. and Godzik, A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics, 19(Suppl.2), ii246-ii255.