2008－BIO－12 (12)
2008／3／4

Flexible Protein Alignment of 3D-Structures Allowing Dynamic Transformation

KOICHI SUEMATSU[†1] and TETSUO SHIBUYA[†1]

In these days attention is focused on the functions of proteins and it is known that when the 3D-structures of proteins are similar the functions of them are often similar. Thus to find the similarity of protein structures is a very important problem. However, most parts of proteins have been treated as rigid ones in the previous papers, though the protein knocks against various molecules in the solvent and is transformed. In this paper, we focus on the local structures of proteins, consist of a sequence of 4 coadjacent amino acids. We investigated the features of Local Structures on two similar proteins, and on two proteins that have no relationship. Then by using the features and the probabilistic argument, we defined a new Similarity Score (SS) between two proteins that can be calculated in the time proportional to the size of the proteins. We show that we can get almost the same results in much shorter time compared to the results that uses $RMSDh$. We also show SS can be used to point out the existence of hinge regions with high probability. Our method is several hundreds times faster than existing methods and by the fact we can make some groups of similar proteins allowing some hinge regions, or to reduce the candidates of similar protein pairs before using some slow algorithms.

1. Introduction

1.1 Background

Now analysis of genes of living things has advanced. Base sequences of human genome is determined in 2003[2], mouse genome in 2002[1], rice plant in 2004[11], etc. Base sequences of many other animals or plants are also being determined. Now analysis of genes is on the way but not only genes but also proteins are playing a great part of living things. Protein has lots of functions, not just the material that our body consists of. Sometimes proteins become the cause of diseases, then to know the aspects of proteins is of help to prevent those diseases. Proteins also control genes by binding to a specific region, then to know the functions of proteins is essential to know the function of genes, that can also be used to prevent diseases or to give a genetic diagnosis. Thus researches are on the way to reveal the functions of proteins. It is well known that when the protein structures look alike, the functions of the proteins are often similar. Then it is very basic and an important problem to calculate the degree of similarity of protein structures from given 3D-structures. Because the number of proteins whose 3D-structures are known is rapidly increasing and thought to keep exponential grow for many years, not only a slow and accurate method but also a fast method is in need.

Because of the background, many kind of methods have been suggested to compare the protein 3D-structures[8),13),16)–19)]. However, most parts of the protein have been treated as rigid ones in the existing technique proposed by present though the protein knocks against various molecules in the solvent, vibrates, and is transformed. So, in this paper, we aim to find out pairs of similar protein pairs among proteins on condition that the structures of two proteins are given and slight transformation of those proteins are allowed. To find similar proteins in short time, we also aim to create a very fast method for calculation with high reliability.

1.2 Outline

Every research on the protein alignment using RMSD was using somehow rigid regions of proteins, and transformations of proteins were treated as special incidents. But proteins are vibrating and changing its shape all the time. Moreover the coordinates given by X-ray crystallography or by NMR have some errors. This assuming that proteins are rigid is not so much a mistake in most cases, but it is not desirable when some hidden similarity exists between proteins.

In this paper, we first investigate the structures of proteins in PDB[3]. In January 2008, there exists more than 48,000 proteins having the coordinates of every atoms in proteins. Moreover, even if the proteins are the same,

†1 University of Tokyo

the protein can take some different structures depending on the environment. In such cases, PDB has more than one set of coordinates of the atoms in proteins. There are so many data, thus some error in coordinates in the database can be ignored statistically. Then by using the knowledge we get, we define the new method to calculate the similarity score. By using the method, we show how to find similar proteins.

2. Related Works

2.1 Protein Comparison

RMSD RMSD (Root Mean Square Deviation)[5] is the most popular measure of the distance between two proteins. The value can be calculated on condition the correspondences of the location of amino acids in two proteins are known a priori. Both proteins can rotate or move in 3D space without any constraint. We call the matrix which declares the rotation and parallel shift T. When we choose T, then every coordinates α_i in α are transformed to α'_i. The distance of two corresponding atoms are $T\alpha_i - \beta_i$, then the RMSD value can be represented as $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(T\alpha_i - \beta_i)^2}$, where n is the number of pairs of amino acids. The original RMSD has an advantage that there exists an $\mathcal{O}(n)$ algorithm to calculate the RMSD value between the sequences with n points, so it is useful when the calculation time is the main concern.

There are many researches using RMSD value, for example *WHAT-IF*, *FlexProt*[16] and *RMSDh(n)*[18].

Geometric Techniques There are also other methods to compare other features to find similar regions using geometric techniques[10),15)]. DALI[10] use the three-dimensional coordinates of each protein to calculate residue-residue (C alpha-C alpha) distance matrices. They decompose the distance matrices and combine them into consistent sets of pairs. Secondary structure elements (SSEs) of proteins are also sometimes used because the secondary structures are thought to be saved even if the 3D-structure of a protein changes dynamically.

Local Structures There are some papers focusing on the local area of proteins[13),20)]. They use angles to define the structures. For example, *DRAGON* creates virtual protein structures randomly that is similar to proteins in the real world. The program is focusing on the local structure of a protein, consists of $4C_\alpha$ atoms. It provides the protein structure satisfying some conditions that often occur in protein structures. TALI[13] also uses the features of rotations that are the same with DRAGON. TALI calculats the distance between sequences by the difference of angles representing local structures, then calculats the alignment score by using gap penalties.

3. Preliminaries

3.1 Protein Data Bank

The fundamental three-dimensional structure description consists of the specification of the coordinates of each atom, as given in the PDB[3], and the data is available on the Internet. We can use the data to compare proteins.

3.2 Protein Representation

The location of C_α atoms are often used as location of the amino acids, because of some reasons; (1) every amino acid has a C_α in it, (2) C_α atoms can be regarded as a backbone of the protein structure, because the C_α atoms is connected to a carboxyl group and an amino group. Because it is almost impossible to treat all the atoms in a given protein, we usually use C_α atom as a representative of an amino acid and explain the protein structures by using them. Thus, a protein can be represented by an array, which contains coordinates of C_α atoms.

3.3 Local Structure of a Protein

Let P_n be the coordinate of the C_α atom of nth amino acid in a protein. In our paper, we pay attention to the local structures of the bonds of amino acids, so in each calculation we just use *Local Structures* as a target and do not pay attention to the global structure of proteins. A Local Structure of a protein P consists of the coordinates of 4 Co-adjacent amino acids, and the location of those amino acids are explained as P_i, P_{i+1}, P_{i+2} and P_{i+3}.

3.4 Features of a Local Structure

We define a simple function to make it easier to understand some expressions as follows. Distance between two coadjacent C_α atoms are the features of Local Structures, but we find that the distances are always almost the same value. So in this paper we treat the distances as a constant value and do not think them as features of a Local Structure.

$$angle(p,q,r) = cos^{-1}\left(\frac{\overrightarrow{qp}\cdot\overrightarrow{qr}}{|\overrightarrow{qp}|\cdot|\overrightarrow{qr}|}\right)$$

In this paper, we just apply the *angle* with coadjacent C_α atoms, so we redefine the value in another way, as $\theta_{P,i} = angle(P_i, P_{i+1}, P_{i+2})$. But to define the whole structure of a Local Structure consists of P_i, P_{i+1}, P_{i+2} and P_{i+3}, function $\theta_{P,i}$ is not enough. We need another parameter $\phi_{P,i} = skew(P_i, P_{i+1}, P_{i+2}, P_{i+3})$, which can be obtained by the following steps.

- Rotate a rigid Local Structure to satisfy the following conditions; (1) $\overrightarrow{P'_{i+1}P'_{i+2}}$ is parallel to z-axis. (2)$\overrightarrow{P'_iP'_{i+1}}$ is on the xz-plane. We define the 3×3 rotation matrix that satisfies the condition as $\Sigma_{P,i}$ and P'_{i+k} be $\Sigma_{P,i}P_{i+k}(k = 0, 1, 2)$.
- Get the projected coordinates of $P'_{i+k}(k = 0, 1, 2, 3)$ on xz-plane. Call the projected coordinates as $P''_{i+k}(k = 0, 1, 2, 3)$. Notice that the coordinates of P''_{i+1}, P''_{i+2} will be the same location.
- We define $angle(P''_i, P''_{i+1}, P''_{i+3})$ as $\phi_{P,i} = skew(P_i, P_{i+1}, P_{i+2}, P_{i+3})$ that defines the rotation around the axis $\overrightarrow{P_{i+1}P_{i+2}}$.

Now we have obtained every parameter to define the Local Structure listed below. In the following chapters, we will use the parameters as features of a Local Structure.

- $\theta_{i+k} = angle(P_{i+k}, P_{i+k+1}, P_{i+k+2})$ $(k = 0, 1)$.
- $\phi_i = skew(P_i, P_{i+1}, P_{i+2}, P_{i+3})$.

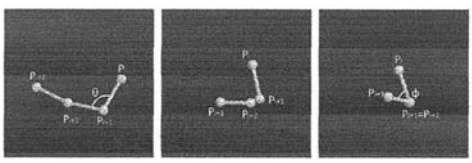

Fig. 1 Definition of features

Figure 1 shows the definition of θ and ϕ visually. The three images in the figure show the same rigid local structure consists of 4 atoms from different point of view. From θ and ϕ and the distance between atoms, we can reconstruct the whole Local Structure, so we call these values as features of a Local Structure.

4. Similarity of Two Proteins

In this chapter, we focus on the *Local Structures* of proteins. Here we define a Local Structure as a region consists of four contiguous C_α atoms in a protein. We first examine every feature that a Local Structure has, to know which kind of features are suitable for representing the structure. After that, we define the Similarity Score by using the features and probablistic arguments.

4.1 Analysis of Proteins

To find similar proteins, we need to know which type of transformation is conceivable and which type of transformation is thought to be impossible. Then we first investigated some features of proteins. Let the protein to be investigated be P, we drew a histogram of the following features; (1) Distribution of $\theta_{P,i}$ (2) Distribution of $\phi_{P,i}$.

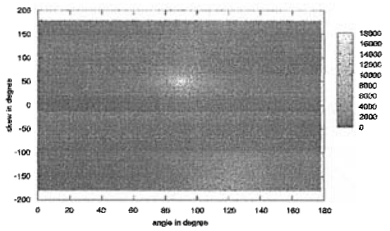

Fig. 2 Distribution of $\theta_{P,i}$ and $\phi_{P,i}$

We also drew the distribution of $\theta_{P,i}$ and $\phi_{P,i}$ in Figure 2. In the figure, we can see that the two variables have some relationships. There are two peaks in the graph, one is around $\theta_{P,i} =$ 90-degree, $\phi_{P,i} =$ 50-degree, and the other is around $\theta_{P,i} =$ 120-degree, $\phi_{P,i} = -165$-degree. By the figure, we can say that protein structures tend to have this two patterns.

4.2 Analysis of Isomeric Protein Pairs

To examine the appropriate criteria of the similarity allowing transformation, we investigated some isomeric protein pairs on some features; (1) distribution of $\theta_{P,i} - \theta_{Q,i}$, (2) distribution of $\phi_{P,i} - \phi_{Q,i}$. P and Q are the pair of proteins and each of P_i and Q_i are the corresponding amino acids.

In the former figures, it is shown that proteins tend to have two particular Local Structures. But Figure 3 shows that in most cases the protein structure keeps its Local Structure and rarely have the other structure. It is a remarkable result.

4.3 Analysis of Random Protein Pairs

In this section, we investigated the features between two Local Structures which are randomly selected.

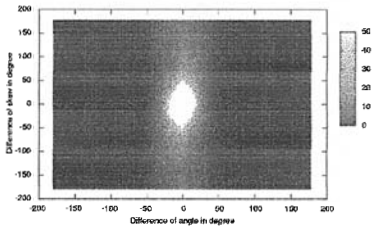

Fig. 3 Distribution of $\theta_{P,i} - \theta_{Q,i}$ and $\phi_{P,i} - \phi_{Q,i}$

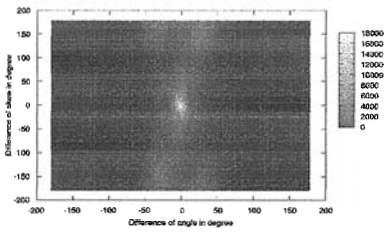

Fig. 4 Distribution of $\theta_{P,i} - \theta_{Q,i}$ and $\phi_{P,i} - \phi_{Q,i}$ when P and Q are selected at random

Figure 4 is a 2-dimensional graph of $\theta_{P,i} - \theta_{Q,i}$ and $\phi_{P,i} - \phi_{Q,i}$, on condition P and Q are randomly selected. Compared to Figure 3, this graph has wide distribution and we can find two peaks in the graph though Figure 3 only has a single peak. Note that the four peaks out of five we can find in the figure are caused by the same reason. Because the top of the graph and the bottom of the graph are showing the same value, we can find there are only three peaks in the graph. Moreover, the graph should be point-symmetric. Because of those reasons, the number of peaks having some meaning is 2. The two peaks can be explained by the existence of two peaks in Figure 2.

4.4 Similarity Score

By using the facts shown in the last two sections, we design a similarity score between two different protein structures. According to Bayes' theorem, we can obtain the posterior probability that the given two Local Structures are similar by using the histograms.

4.4.1 Bayesian Decision Theory

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. We here note that the probability of finding a pattern that is in category ω_i and has feature value $\mathbf{x}$ can be expressed in two ways: $P(\omega_i, \mathbf{x}) = P(\omega_i|\mathbf{x})P(\mathbf{x}) = P(\mathbf{x}|\omega_i)P(\omega_i)$. Arrange the equation and we get a formula that is called *Bayes formula*:

$$P(\omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})}$$

4.4.2 Similarity Probability

In the last section we detect that θ and ϕ are likely to be similar when two proteins are similar. So we use two features $d\theta = \theta_{P,i} - \theta_{Q,i}$ and $d\phi = \phi_{P,i} - \phi_{Q,i}$ as a feature vector $\mathbf{x}$. Then we get the expression described below.

$$P(\omega_i|d\theta, d\phi) = \frac{P(d\theta, d\phi|\omega_i) \cdot P(\omega_i)}{P(d\theta, d\phi)} \quad (1)$$

When we use the theorem to compare two proteins, the proteins we are looking at never change its structures, so $P(\omega_i)$ is a constant value. Thus on condition that $\mathbf{x}$ is known, the probability that the given two proteins are similar or not is proportional to $\frac{P(\mathbf{x}|\omega_i)}{P(\mathbf{x})}$. Because we do not know the exact value of $P(\omega_i)$ we cannot calculate the exact probability of $P(\omega_i|\mathbf{x})$ but we can compare the ratio of two probabilities.

By using the histograms we got in the last section, we can estimate the value of $P(\mathbf{x}|\omega_i)$ and $P(\mathbf{x})$. Then we can get the approximate value of $P(\omega_i|\mathbf{x})$ for every i and $\mathbf{x}$. Note that our definition of $P(\omega_i|\mathbf{x})$ is based on the observed structures of the known proteins. Because of the fact, even if a feature vector was not observed, it does not directly mean such transformation never occurs between similar proteins. Though on such situations, we can say the kind of transformations rarely occurs. So we define the values called *Similarity Probability (SP)*, which is the probability mainly based on $P(\omega_i|\mathbf{x})$. $SP(d\theta, d\phi)$ represents the Similarity Probability on condition that the difference of *angle* is $d\theta$ and the difference of *skew* is $d\phi$.

$$SP(d\theta, d\phi) = max(P(\omega_{sim}|d\theta, d\phi), \epsilon) \quad (2)$$

ϵ represents the probability that ω_{sim} occurs even if the feature vector was not observed in the experiment, satisfying $0 < \epsilon < P(\omega_{sim}|d\theta, d\phi)(\forall(d\theta, d\phi), P(\omega_{sim}|d\theta, d\phi) > 0)$. In this paper we just multiply SP, so we define SS as the logarithmic value of Similarity Probability that we call Similarity Score; $SS(d\theta, d\phi) = \log SP(d\theta, d\phi)$.

Figure 5 shows the similarity score at every $d\theta$ and $d\phi$.

4.5 Similarity Score of a Protein Pair

We describe our definition of similarity be-

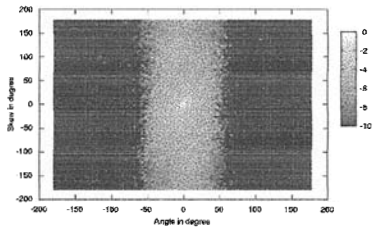

Fig. 5 Similarity Score at every type of transformations

tween two proteins on condition that features of every Local Structures are independent of those of other Local Structures. This assumption is probably not true, but later we will compare the result we get from our method to the results from existing methods, then describe the assumption does not affect the result. By using the assumption, we define the *average Similarity Score* between two proteins, that can be easily calculated on condition that the corresponding Local Structures are known a priori.

When we call the value as $SS(P,Q)$ that means *average Similarity Score of Proteins*, the value is: $SS(P,Q) = \log P(\omega_{sim}) + \frac{1}{n-3}\sum_{k=1}^{n-3} SS(\theta_{P,k} - \theta_{Q,k}, \phi_{P,k} - \phi_{Q,k})$. The definition enables us to compare the Similarity Scores between multiple protein pairs that have different size.

The value depends on the prior probability $P(\omega_{sim})$. In usual we cannot get the exact value of $P(\omega_{sim})$, so we treat the value as a constant one, and when comparing the Scores we just use the term $\frac{1}{n-3}\sum_{k=1}^{n-3} SS(\theta_{P,k} - \theta_{Q,k}, \phi_{P,k} - \phi_{Q,k})$. We cannot get the exact probability by the method, though the value can be used to compare Similarity Scores.

4.5.1 Expected Similarity Score between Two Proteins

To give a rough standard to *SS*, we calculated the expected value and the variance of *SS* between two proteins on two conditions, one is the two proteins are similar, and the other is the two proteins has no relationship.

We calculated the value from the figures shown above by using the hypothesis that every Local Structures are independent of the other Local Structures, and get some values described below.

- When two proteins are similar, the expected value of *SS* at a Local Structure is -2.14 and the variance is 1.56.
- When two proteins are not similar, the expected value of *SS* at a Local Structure is -5.27 and the variance is 1.93.

The values are calculated at every Local Structure, so *SS* of two proteins has different distribution. *SS* of a pair of proteins is the mean of every pair of Local Structures, so the mean of *SS* is the same as that of each Local Structure. Thus when two proteins are similar, the expected value of *SS* is -2.14 and when they are not similar, the expected value is -5.27. But the variance of mean differs according to the number of the pair of Local Structures. When there is t times more pair of Local Structures, the standard variation becomes $\frac{1}{\sqrt{t}}$ times according to the statistical theory. For example, when there is 100 Local Structures, the standard variation becomes $\frac{1}{\sqrt{100}}$ times, that is 0.1 times. Thus the case with comparing two structures with length 103, the standard variance of *SS* between two proteins is 0.156 when two proteins are similar and is 0.193 when no prior information is given about the two proteins.

5. Experiments

We implemented our method to show that the method can be used as an alternative method of former methods. We used the machine with the following specifications. CPU: AMD AthlonX2 BE-2350, Memory: 4.0GB and HDD: 500GB.

5.1 Calculation Time

We first comapred the time for calculation. When measureing the time for computing, we exclude the time for reading the pdb format file, because the size of a file is independent of the size of a protein.

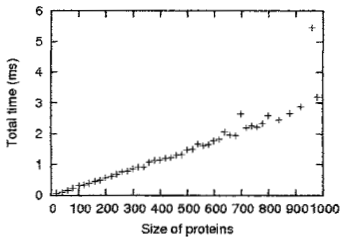

Fig. 6 Time for calculating SS

Figure 5.1 shows the time for calculating *SS* between two proteins. Figure 5.1 shows the

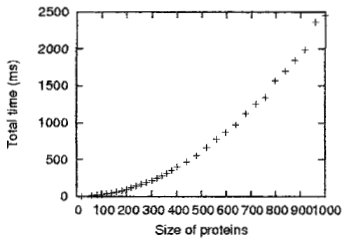

Fig. 7 Time for calculating RMSDh(3)

time for calculating the RMSDh(3) value. The time also exclude the time for reading a file. Figure 5.1 consumes more than 1 second when the size of proteins is large because the time complexity of the algorithm is $\mathcal{O}(n^2)$. Compared to the figure, our result shown in Figure 5.1 requires very short time and every calculation is done within several milliseconds. We can also say that the calculation time is linear to the size of proteins. The fact means our program is of great use especially when long proteins are compared.

5.2 Accuracy

Our method is fast but we did not show that the result we get from our algorithm can be used as the measure of similarity, because of the assumption that there is no relationship between two coadjacent Local Structures. In this section, we compare *SS* with RMSDh that is defined in the former paper, and discuss the relationship between these two values.

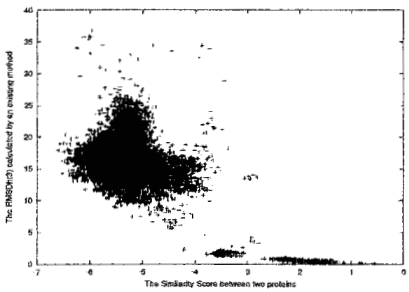

Fig. 8 The relationship between *SS* and the score calculated by a former algorithm

Figure 8 shows the relationship between *SS* and the RMSDh(3) value. We plotted just 20000 points in the figure to show the relationship because calculating tens of thousands of RMSD(3) values takes very long time. We did not show the comparing method of two proteins allowing gaps yet, so the comparison was done without thinking about the existence of gaps. In the experiment, the size of proteins is fixed to 100, because the variance of RMSD or RMSDh values differs depending on the size of proteins.

In the figure, we can find out a big cluster on the left side, and a small cluster or a line on the right side. The center of the large cluster seems to be between $-5.0 < SS < -5.5$. In the last chapter, we showed that the expected value of *SS* between two non-similar proteins is -5.27. From the fact, the large cluster is thought to be the groups showing non-similar protein pairs by our method. The fact that the RMSD values of points in the cluster is larger than the values found in the smaller cluster supports our concept that *SS* can be used to distinguish similar protein pairs from non-similar protein pairs.

On the bottom of the figure, there is a long and thin cluster. Every member of the cluster has relatively low RMSDh value and relatively high *SS*, compared to the elements in the large cluster.

From the figure, it seems that we can figure out a given pair of proteins belongs to which cluster by *SS*.

5.2.1 Proper Threshold for Making a Decision

We give the statistically trustable indicator to the threshold of *SS* to separate similar protein pairs from non-similar protein pairs. We want to know whether given two proteins are similar of not only from *SS*.

When we make a decision, there occur 4 cases as shown in Table 5.2.1. *TN*(True Negatibe) and *FP*(False Positive) are a mistake, so high ratio of *TP*(True Positive) and *FN*(False Negative) is desired.

Prediction \ Real State	Similar	Different
Similar	TP	FP
Different	TN	FN

Table 1 Classification of the Real State and Predicted Result

There are many kinds of methods for evaluation. *Precision* is calculated by $TP/(TP+FP)$ and *Recall* is calculated by $TP/(TP+TN)$. High precision means when a program says two proteins are similar, they are similar with high probability. We get high precision by setting high threshold of *SS*. High recall means the set of protein pairs a program says contains most

of the similar protein pairs. We get high recall by setting low threshold of *SS*. Both indicators are important but they can be achieved by opposite way thus they cannot occur in the same time. Then *F-measure* is sometimes used instead of these values, considering the trade-off. F-measure is calculated by the following expression; $\frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$.

We want to know if we can distinguish the cluster by using *SS*, so we defined two thresholds. One is the threshold for *SS* that two proteins are thought to be similar and the other is almost the same with RMSDh(3).

When the evaluation method is applied to Figure 8, the maximum value of F-measure was 0.973. In the case, the threshold for *SS* was -3.65 and the threshold for RMSDh(3) was 2.69Åthat achieves 0.972 of Precision and 0.974 of Recall. These values are very high level, so we can say there is a strong relationship between *SS* and RMSDh(3). Usually, the RMSD value that two protein structures are decided as similar is around 3.0Å, and the result of this experiment shows almost the same value as a threshold for RMSD value on protein pairs.

5.3 Detecting the Existence of Hinge Regions

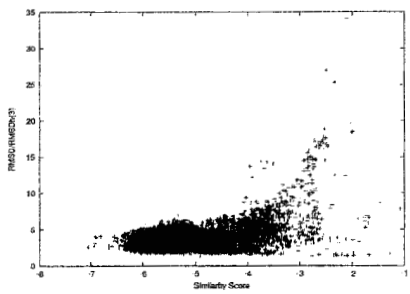

Fig. 9 The relationship between *SS* and RMSD divided by RMSDh(3) when the protein size is 50

When there are hinge regions, the RMSD value is not so small and the RMSDh value is very small. Comparing these two values to detect the existence of a hinge region is a direct and very simple approach. Figure 9 and 10 show the relationship between *SS* and $\frac{\text{RMSD}}{\text{RMSDh(3)}}$. Each figure is drawn with 20000 pairs of proteins that are randomly selected. When there are hinge regions in a pair of proteins, the value $\frac{\text{RMSD}}{\text{RMSDh(3)}}$ is thought to have

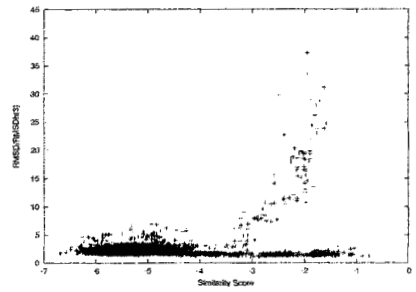

Fig. 10 The relationship between *SS* and RMSD divided by RMSDh(3) when the protein size is 124

large value, then by comparing the value with *SS*, we can test whether our method can be used to find some hinge regions.

In each graph, we can find some points showing high ratio of RMSD to RMSDh(3). We can also find that these points are on the right side of the figure, that is, when *SS* is high, the ratio has relatively high probability of having large value. There are also the cases that *SS* is high but the ratio is low. Those points are representing the very similar proteins, and their RMSD value is very small. Because the RMSD value is very small, the ratio stays small.

5.4 Example of Detecting Similar Protein Pairs Having Hinge Regions

Proteins	SS	RMSD	RMSDh(3)	FlexProt
1b01(A) 11fh	-1.98	6.39	0.807	1.43
1a2w(B) 1wbu(A)	-1.79	15.03	0.372	1.37

Table 2 Similar protein pairs found by our method, having hinge regions

Table 5.4 shows the scores by many methods for several protein pairs found to be similar by using our method. These pairs have large RMSD values and small RMSDh(3) values that indicate they have hinge regions. By using FlexProt, we got more precise result. The result shows that our program can also be used to detect the similar protein pairs having hinge regions without treating the hinge regions as a special event.

6. Conclusion

In this paper, we first defined features of the Local Structures, and took statistics of the features, then showed those features differ depend-

ing on the condition that the pair of proteins is similar or not. By using the difference of the features, we defined the Similarity Score (SS) between Local Structures. After that, we defined SS of a pair of proteins. The time for calculating SS between proteins is linear to the length of proteins, though FlexProt takes $\mathcal{O}(n^6)$ time and RMSDh(n) takes $\mathcal{O}(n^2)$ time.

We investigated the relationships between SS our method gives and the score some existing methods give. The results are, (1) our program is very fast that enables to compare thousands of pairs of proteins in a second, (2) there is a strong relationship between SS and the RMSDh value, and we can predict whether RMSDh value is small enough or not with 0.973 of F-measure, (3) our program can be used to detect the existence of hinge regions. They support that our method is of use.

We conclude that we can detect whether given two proteins are similar or not in very short time with high accuracy by using the new method. The method can also be used to reduce the possibly-similar protein pairs to highly-possibly-similar protein pairs even if there exist some hinge regions. When we treat the existence of some hinge regions, our program is several thousand times fast compared to the existing methods when the size of proteins are large. The results show that our method is suitable for exhaustive similarity search among proteins.

References

1) : A physical map of the mouse genome., *Nature*, Vol.418, No.6899, pp.743–50 (2002).

2) : Finishing the euchromatic sequence of the human genome., *Nature*, Vol.431, No.7011, pp. 931–945 (2004).

3) Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. and Bourne, P.: The protein data bank (2000).

4) Bundschuh, R.: Rapid significance estimation in local sequence alignment with gaps, *RECOMB '01: Proceedings of the fifth annual international conference on Computational biology*, New York, NY, USA, ACM, pp. 77–85 (2001).

5) Chew, L.P., Huttenlocher, D.P., Kedem, K. and Kleinberg, J.M.: Fast Detection of Common Geometric Substructure in Proteins, *Journal of Computational Biology*, Vol. 6, No. 3/4 (1999).

6) Dayhoff, M., Schwartz, R. and Orcutt, B.: *Atlas of protein sequence and structure*, Vol. 5, chapterA model of evolutionary change in proteins., pp. 345–352, National Biomedical Research Foundation (1978).

7) Eidhammer, I., Jonassen, I. and Taylor, W. R.: *PROTEIN BIOINFORMATICS*, chaptera, John Wiley and Sons, Ltd (2004).

8) Gerstein, M. and Hegyi, H.: Comparing genomes in terms of protein structure: surveys of a finite parts list (1998).

9) Henikoff, S. and Henicoff, J.: Amino acid substitution matrices from protein blocks, *Proc Natl Acad Sci U S A*, Vol.89, No.22, pp.10915–10919 (1992).

10) Holm, L. and Sander, C.: Protein structure comparison by alignment of distance matrices., *J Mol Biol*, Vol.233, No.1, pp.123–138 (1993).

11) International: The map-based sequence of the rice genome, *Nature*, Vol.436, No.7052, pp.793–800 (2005).

12) Maiorov, V.N. and Crippen, G.M.: Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins., *J Mol Biol*, Vol. 235, No. 2, pp. 625–634 (1994).

13) Miao, X., Bryson, M. and Valafar, H.: TALI: Protein Structure Alignment Using Backbone Torsion Angles, *BIOCOMP*, pp.3–9 (2006).

14) Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures., *J. Mol. Biol.*, Vol.247, pp.536–540 (1995).

15) Nussinov, R. and Wolfson, H.J.: Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques., *Proceedings of the National Academy of Sciences of the USA*, Vol.88, No.23, pp.10495–10499 (1991).

16) Shatsky, M., Nussinov, R. and Wolfson, H.J.: FlexProt: alignment of flexible protein structures without a predefinition of hinge regions., *J Comput Biol*, Vol.11, No.1, pp.83–106 (2004).

17) Shatsky, M., Wolfson, H.J. and Nussinov, R.: Flexible protein alignment and hinge detection, *Proteins: Structure, Function, and Genetics 48* (2002).

18) Shibuya, T.: Fast and Accurate Algorithms for Protein Hinge Detection, *IPSJ SIG Notes SIGBIO*, Vol.10, No.4, pp.25–32 (2007).

19) Verbitsky, G., Nussinov, R. and Wolfson, H.: Structural comparison allowing hinge bending (1999).

20) Ye, J., Janardan, R. and Liu, S.: Pairwise Protein Structure Alignment Based on an Orientation-Independent Representation of the Backbone Geometry.