2007－BIO－10 (4)
2007／9／13

タンパク質ヒンジ構造検出のための高速かつ正確なアルゴリズム

渋谷 哲朗

東京大学医科学研究所ヒトゲノム解析センター
〒108-8639 東京都港区白金台4-6-1
E-mail: tshibuya@hgc.jp

概要: タンパク質構造の類似性を測る尺度としてRMSDがあるが、本論文では、これをヒンジ構造部分は異なるが、それ以外の部分は類似するようなタンパク質構造の比較に用いることができるように拡張する。本論文では、それをRMSDhと名づけ、これを計算する高速なアルゴリズムを提案する。このRMSDhはヒンジ構造が1箇所の場合には線形時間で計算することができることを示す。さらにヒンジ構造が高々k箇所の場合には、$O(kn^2)$時間・$O(n)$空間で計算することができることを示す。さらに計算機実験を行い、実際にヒンジ構造を検出することが高速かつ正確にできることを示す。

Fast and Accurate Algorithms for Protein Hinge Detection

Tetsuo Shibuya

Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.
E-mail: tshibuya@hgc.jp

Abstract: Analysis of hinge positions in flexible proteins is one of the keys to the understanding of their functions and interactions. The RMSD (Root Mean Square Deviation) is the most popular measure for comparing two protein structures, but it is only for rigid structures without hinge domains. In this paper, we propose a new measure called RMSDh (Root Mean Square Deviation considering hinges) and its variant RMSDh$^{(k)}$ for comparing two flexible proteins with hinge domains. We also propose efficient algorithms for computing them, which can detect the hinge positions at the same time. The RMSDh is suitable for cases where there is one small hinge domain in each of the two target structures. The algorithm for computing the RMSDh runs in linear time, which is same as the time complexity for computing the RMSD and is faster than any of previous algorithms for hinge detection. The RMSDh$^{(k)}$ is designed for comparing structures with more than one hinge domain. The RMSDh$^{(k)}$ measure considers at most k small hinge domains, *i.e.*, the RMSDh$^{(k)}$ value should be small if the two structures are similar except for at most k hinge domains. To compute the value, we propose an $O(kn^2)$-time and $O(n)$-space algorithm based on dynamic programming. We also test our measures against both flexible protein structures and non-flexible protein structures, and show that the hinge positions can be correctly detected by our algorithms.

1 Introduction

Proteins play enormous variety of roles in living systems. The functions of the proteins are said to be determined by their 3-D structures, and consequently the analysis of protein structures is one of the most important research topics in molecular biology. The analysis of protein structures often starts with comparison of two similar structures, and there have been proposed tremendous number of methods to compare two protein 3-D structures [Eidhammer et al. 2000, Lemmen et al. 2000, Wolfson et al. 2005]. Structure comparison algorithms can be categorized into two types: rigid structure comparison methods and flexible structure comparison methods. The former methods consider protein structures as rigid bodies. But there are not a few proteins which change conformationally. Most of their structures can be divided into several rigid substructures separated by small parts (which often consists of

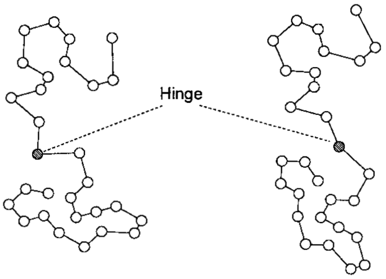

Figure 1: Hinge bending of a protein. A protein sometimes changes its structure by rotating around an atom, which is called a hinge.

only one atom) called hinge domains or just 'hinges' (Figure 1). They change their structures by rotating around the hinge, due to their physical conditions, relations to other proteins, or some point mutations. The hinges sometimes take very important roles for their functions [Wolfson et al. 2005]. The latter flexible structure comparison methods take hinges into consideration when they compare structures.

There are three tasks when we compare two flexible structures. At first we have to find the correspondence between atoms. We next have to find locations of hinges and finally we have to calculate superposition for each rigid domain. But if we have determined the correspondence and the hinge positions, it is not difficult to compute the superposition. Thus, flexible structure comparison methods can be categorized into two types.[1] One is a type of methods that does everything — they find the atom correspondence, the hinge positions, and the superposition simultaneously [Shatsky et al. 2004, Ye et al. 2003]. The other type of methods is dedicated to only hinge detection and calculation of superposition [Boutonnet et al. 1995, Huang et al. 1993, Nigham et al. 2007, Ochagavia et al. 2002, Wriggers et al. 1997], as they consider the atom correspondence is given. The methods of the former type are more general than those of the latter, but they are definitely more difficult. Note that there are many situations in which only the latter methods are needed. For example, we always know the atom correspondence in two structures of the same, or significantly similar proteins. In this paper, we do not deal with how to find the atom correspondence.

When we compare two structures (by either of the two approaches), some appropriate scoring measure is desired. The measure must be mathematically clear and moreover easy to compute. The RMSD (Root Mean Square Deviation) [Arun et al. 1987, Eggert et al. 1997, Kabsch 1976, Kabsch 1978, Schwartz et al. 1987] is the most commonly used measure for comparing two rigid structures (see section 2 for details). It is defined very clearly and can be computed very efficiently (in linear time). But it is designed only for rigid structures. There are no standard measures to be optimized for flexible structure comparison, as it seems very difficult to design a measure that can be efficiently computed.

In this paper, we propose measures for comparing flexible protein structures, which can be very efficiently computed. We first propose a measure called the RMSDh (Root Mean Square Deviation considering hinges), which is a generalization of the RMSD with consideration of a hinge domain. We also propose an algorithm that computes the RMSDh in linear time, which is same as the time complexity for computing the RMSD. Moreover, it detects the hinge position at the same time. It is much faster than any of previous hinge detection algorithms. We then generalize the RMSDh for proteins with at most k hinges, and call the generalized measure the RMSDh$^{(k)}$. We propose an $O(kn^2)$-time and $O(n)$-space algorithm for computing it, where n is the length of the structures to be compared. We will also show that we can detect the hinge positions with the same time and space complexity, by using a divide-and-conquer technique.

The organization of this paper is as follows. In section 2, we present the definition of the RMSD and algorithms for computing it as preliminaries. Then we propose the new RMSDh measure and algorithms for it in section 3. We propose the RMSDh$^{(k)}$ measure and algorithms for it in section 4. In section 5, we show computational experiments. Finally in section 6, we conclude our results and discuss future work.

[1] Similarly, rigid structure comparison methods can also be categorized into two types.

2 Preliminaries

2.1 RMSD: The Root Mean Square Deviation

A protein 3-D structure can be represented by various ways, but the most common way is to represent it by a list of 3-D coordinates of its backbone C_α atoms. The RMSD (root mean square deviation) [Arun et al. 1987, Eggert et al. 1997, Kabsch 1976, Kabsch 1978, Schwartz et al. 1987] is the most common way to compare two lists of 3-D coordinates.

Let the two sets of points (*i.e.*, protein structures) to be compared be $\mathbf{P} = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_n\}$ and $\mathbf{Q} = \{\vec{q}_1, \vec{q}_2, \ldots, \vec{q}_n\}$, where $\vec{p}_i$ and $\vec{q}_i$ are the coordinates of the i-th C_α atoms of $\mathbf{P}$ and $\mathbf{Q}$, respectively. Then the RMSD between $\mathbf{P}$ and $\mathbf{Q}$ is defined as the minimum value of

$$D_{R,\vec{v}}(\mathbf{P},\mathbf{Q}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\vec{p}_i - (R \cdot \vec{q}_i + \vec{v})\|^2}$$

over all the possible rotation matrices R and translation vectors $\vec{v}$, where $\|\cdot\|$ denotes the norm. Let $RMSD(\mathbf{P},\mathbf{Q})$ denote the minimum value, and let $\hat{R}(\mathbf{P},\mathbf{Q})$ and $\hat{\vec{v}}(\mathbf{P},\mathbf{Q})$ denote the R and $\vec{v}$ that satisfy $D_{R,\vec{v}}(\mathbf{P},\mathbf{Q}) = RMSD(\mathbf{P},\mathbf{Q})$.

2.2 How to compute the RMSD

In this section, we briefly describe how to compute the RMSD. Let $R \cdot \mathbf{X}$ denote the structure $\mathbf{X}$ rotated by the rotation matrix R. If the rotation matrix R is fixed, $D_{R,\vec{v}}(\mathbf{P},\mathbf{Q})$ is known to be minimized when the centroid of $R \cdot \mathbf{Q}$ is translated to the centroid of $\mathbf{P}$ by the translation vector $\vec{v}$, regardless of what the rotation matrix R is. Hence, if both $\mathbf{P}$ and $\mathbf{Q}$ are translated so that their centroids are moved to the origin of the coordinates, the RMSD problem is reduced to a problem of finding R (*i.e.*, $\hat{R}(\mathbf{P},\mathbf{Q})$) that minimizes $F_R(\mathbf{P},\mathbf{Q}) = \sum_{i=1}^{n} \|\vec{p}_i - R \cdot \vec{q}_i\|^2$.

We can compute $\hat{R}(\mathbf{P},\mathbf{Q})$ in linear time by using the singular value decomposition (SVD) [Arun et al. 1987, Kabsch 1976, Kabsch 1978] as follows. Let $H = \sum_{i=1}^{n} \vec{p}_i \cdot \vec{q}_i^{\,t}$, where $\vec{v}^t$ means the transpose of vector $\vec{v}$. Clearly, H can be computed in $O(n)$ time. Then $F_R(\mathbf{P},\mathbf{Q})$ can be described as $\sum_{i=1}^{n} (\vec{p}_i^{\,t}\vec{p}_i + \vec{q}_i^{\,t}\vec{q}_i) - trace(R \cdot H)$, and $trace(RH)$ is maximized when $R = VU^T$, where $U\Lambda V$ is the SVD of H and A^T means the transpose of matrix A. Thus $\hat{R}(\mathbf{P},\mathbf{Q})$ can be obtained from H in constant time, as H is a 3×3 matrix and the SVD can be computed in $O(d^3)$ time for a $d \times d$ matrix [Golub et al. 1996]. Note that there are rare degenerate cases where $det(VU^T) = -1$, which means that VU^T is a reflection matrix. We ignore the degenerate cases in this paper. We can compute the RMSD value in linear time once we have obtained $\hat{R}(\mathbf{P},\mathbf{Q})$. In total, we can compute the RMSD value in $O(n)$ time.

Let $\mathbf{S}[i..j]$ denote the substructure of $\mathbf{S}$ from the i-th atom to the j-th atom (*e.g.*, $\mathbf{P}[i..j] = \{\vec{p}_i, \vec{p}_{i+1}, \ldots, \vec{p}_j\}$). According to [Shibuya 2007], the RMSD and corresponding superposition between two substructures $\mathbf{P}[i..j]$ and $\mathbf{Q}[i..j]$ can be computed in constant time for any i and j, after linear-time preprocessing, as follows: $RMSD(\mathbf{P}[i..j], \mathbf{Q}[i..j])$, $\hat{R}(\mathbf{P}[i..j], \mathbf{Q}[i..j])$ and $\hat{\vec{v}}(\mathbf{P}[i..j], \mathbf{Q}[i..j])$ can be computed in $O(1)$ time if we are given $\sum_{k=i}^{j} \vec{p}_k$, $\sum_{k=i}^{j} \vec{p}_k^{\,t}\vec{p}_k$, $\sum_{k=i}^{j} \vec{q}_k$, $\sum_{k=i}^{j} \vec{q}_k^{\,t}\vec{q}_k$, and $\sum_{k=i}^{j} \vec{p}_k\vec{q}_k^{\,t}$. These values can be computed also in constant time, if we compute the following values in advance: $\sum_{k=1}^{\ell} \vec{p}_k$, $\sum_{k=1}^{\ell} \vec{p}_k^{\,t}\vec{p}_k$, $\sum_{k=1}^{\ell} \vec{q}_k$, $\sum_{k=1}^{\ell} \vec{q}_k^{\,t}\vec{q}_k$, and $\sum_{k=1}^{\ell} \vec{p}_k\vec{q}_k^{\,t}$, for all ℓ ($1 \le \ell \le n$). It is easy to see that all of these values can be computed in $O(n)$ time in total. Thus we conclude that the RMSD and corresponding superposition between $\mathbf{P}[i..j]$, and $\mathbf{Q}[i..j]$ can be computed in $O(1)$ time after linear-time preprocessing.

3 RMSDh: A Linear-Time Computable Measure for Hinge Detection

3.1 Definition of the RMSDh

In this section, we consider a new measure to compare two flexible protein 3-D structures that are very similar except for one small hinge domain. We consider that the hinge domain is so small that it can be considered as only a single backbone atom.[2] Let the two structures to be compared be $\mathbf{P} = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_n\}$ and $\mathbf{Q} = \{\vec{q}_1, \vec{q}_2, \ldots, \vec{q}_n\}$, and consider that the

[2] On the backbone, there are atoms other than the C_α atoms, and the hinge can be located at any of them.

hinge is located at a backbone atom between the ℓ-th C_α atom and the $(\ell+1)$-th C_α atom, or at the ℓ-th C_α atom. Then $\mathbf{P}[1..\ell]$ and $\mathbf{Q}[1..\ell]$ should be similar to each other, and $\mathbf{P}[\ell+1..n]$ and $\mathbf{Q}[\ell+1..n]$ should also be similar to each other. Thus if the two rigid parts of $\mathbf{Q}$ are rotated and translated appropriately with different rotation matrices and translation vectors, $\mathbf{P}$ and the transformed $\mathbf{Q}$ should be very similar to each other, and consequently should have a small RMSD value. It means that

$$G_\ell(\mathbf{P},\mathbf{Q}) = \min_{R_1,R_2,\vec{v}_1,\vec{v}_2} \sqrt{\frac{1}{n}\{g^{left} + g^{right}\}}$$

must be a very small value, where R_1 and R_2 are (possibly different) rotation matrices, $\vec{v}_1$ and $\vec{v}_2$ are (also possibly different) translation vectors, and g^{left} and g^{right} are defined as follows:

$$g^{left} = \sum_{i=1}^{\ell} \|\vec{p}_i - (R_1 \cdot \vec{q}_i + \vec{v}_1)\|^2$$

$$g^{right} = \sum_{i=\ell+1}^{n} \|\vec{p}_i - (R_2 \cdot \vec{q}_i + \vec{v}_2)\|^2$$

$G_\ell(\mathbf{P},\mathbf{Q})$ can be used as the similarity measure between $\mathbf{P}$ and $\mathbf{Q}$, if the hinge is at or around the ℓ-th atom. Note that this value is same as the RMSD if $R_1 = R_2$ and $\vec{v}_1 = \vec{v}_2$ when they are optimized.

But we do not know the actual hinge position in most cases when we compare two structures. Hence, we consider the minimum value of $G_\ell(\mathbf{P},\mathbf{Q})$ over all the possible hinge positions ℓ, *i.e.* $\min_{1\le\ell<n} G_\ell(\mathbf{P},\mathbf{Q})$, as the measure to compare a pair of flexible structures with one hinge. We call it the RMSDh (Root Mean Square Deviation considering hinges), and let $RMSDh(\mathbf{P},\mathbf{Q})$ denote this value. Note that the RMSDh is always smaller than or equal to the RMSD.

3.2 How to compute the RMSDh

The problem of computing $RMSDh(\mathbf{P},\mathbf{Q})$ can be reduced to the problem computing

$$\min_{1\le\ell<n} L_{1,\ell}(\mathbf{P},\mathbf{Q}) + L_{\ell+1,n}(\mathbf{P},\mathbf{Q}),$$

where

$$L_{i,j}(\mathbf{P},\mathbf{Q}) = \min_{R,v} \sum_{x=i}^{j} \|\vec{p}_x - (R \cdot \vec{q}_x + \vec{v})\|^2.$$

Notice that $L_{i,j}(\mathbf{P},\mathbf{Q}) = n \cdot (RMSD(\mathbf{P}[i..j],\mathbf{Q}[i..j]))^2$, which means that we can compute the RMSDh value by computing $2n-2$ RMSD values, *i.e.*, $RMSD(\mathbf{P}[1..\ell],\mathbf{Q}[1..\ell])$ and $RMSD(\mathbf{P}[\ell+1..n],\mathbf{Q}[\ell+1..n])$ for all ℓ $(1 \le \ell < n)$. According to section 2.2, the computation of each RMSD can be done in constant time after linear-time preprocessing.[3] Hence, the RMSDh value can be computed in $O(n)$ time, including the preprocessing phase. Moreover, we can detect the corresponding hinge position at the same time.

4 RMSDh$^{(k)}$: More Flexible Measures

4.1 Definition of the RMSDh$^{(k)}$

In the previous section, we considered only one hinge, but many flexible protein structures are known to have more than one hinge. In this section, we consider that the target structures have k hinges at most, which means they can be divided into $k+1$ rigid domains. Again, let $\mathbf{P} = \{\vec{p}_1,\vec{p}_2,\ldots,\vec{p}_n\}$ and $\mathbf{Q} = \{\vec{q}_1,\vec{q}_2,\ldots,\vec{q}_n\}$ be the flexible structures to be compared, and let the positions of the hinges be $\ell_1,\ell_2,\ldots,\ell_k$. To ease discussion, let $\ell_0 = 1$ and $\ell_{k+1} = n+1$. Let $H_{\ell_1,\ldots,\ell_k}(\mathbf{P},\mathbf{Q})$ be

$$\min_{R_0,\ldots,R_k,\vec{v}_0,\ldots\vec{v}_k} \sqrt{\frac{1}{n}\sum_{j=0}^{k}\sum_{i=\ell_j}^{\ell_{j+1}-1} \|\vec{p}_i - (R_j \cdot \vec{q}_i + \vec{v}_j)\|^2}$$

where $R_0, R_1, \ldots, R_k$ are rotation matrices, and $\vec{v}_0, \vec{v}_1, \ldots, \vec{v}_k$ are translation vectors. Then, with discussion similar to section 3.1, the value should be a very small value,

But, as in section 3.1, we do not know the actual hinge positions in most cases. Thus we propose to use the minimum value of the above expression over all the possible sets of k hinge positions $\{\ell_1,\ell_2,\ldots,\ell_k\}$, *i.e.*,

$$\min_{1\le\ell_1<\ell_2<\cdots<\ell_k\le n} H_{\ell_1,\ldots,\ell_k}(\mathbf{P},\mathbf{Q}),$$

as the measure for comparing such flexible proteins. We call it RMSDh$^{(k)}$, and let $RMSDh^{(k)}(\mathbf{P},\mathbf{Q})$ denote the value. Note that

[3] One can do the same in the same time complexity by using the incremental RMSD computation technique described in [Shibuya 2006].

the RMSDh$^{(1)}$ is same as the RMSDh. Note also that the RMSDh$^{(k)}$ is always smaller than or equal to the RMSDh$^{(k-1)}$, for any k.

4.2 How to Compute the RMSDh$^{(k)}$

As in the case of computing the RMSDh, the problem of computing the RMSDh$^{(k)}$ can be reduced to the problem of computing

$$\min_{1 \leq \ell_1 < \ell_2 < \cdots < \ell_k \leq n} \sum_{i=0}^{k} L_{\ell_i, \ell_{i+1}-1}(\mathbf{P}, \mathbf{Q}),$$

where $L_{i,j}(\mathbf{P}, \mathbf{Q})$ is the same expression as defined in section 3.2. There are ${}_{n-1}C_k$ possible set of k hinge positions, which means that we might need $O((k+1) \cdot {}_{n-1}C_k)$ time or more if we naively compute it. But, in the following, we propose an algorithm based on dynamic programming (DP), which compute the RMSDh$^{(k)}$ in $O(kn^2)$ time and $O(n)$ space.

Let $I_{i,r}$ be

$$\min_{1 \leq \ell_1 < \ell_2 < \cdots < \ell_r \leq i} \sum_{j=0}^{r} L_{\ell_j, \ell_{j+1}-1}(\mathbf{P}[1..i], \mathbf{Q}[1..i]),$$

where we let $\ell_0 = 1$ and $\ell_{r+1} = i + 1$ to ease discussion. We utilize this value to compute the RMSDh$^{(k)}$ as follows. The RMSDh$^{(k)}$ is described as $(I_{n,k}/n)^{\frac{1}{2}}$. Note that $I_{i,r}$ is defined only when $0 \leq r < i \leq n$. In case $r = 0$, it can be easily seen from the definition that $I_{i,0} = L_{1,i}(\mathbf{P}, \mathbf{Q})$ for any i. In addition, the following equation holds when $r \geq 1$:

$$I_{i,r} = \min_{r \leq j < i} \{I_{j,r-1} + L_{j+1,i}(\mathbf{P}, \mathbf{Q})\}.$$

The above equation represents a DP algorithm for computing $I_{n,k}$ and consequently the RMSDh$^{(k)}$ (*i.e.*, $(I_{n,k}/n)^{\frac{1}{2}}$). During the DP procedure, we compute $I_{n,r}$ for all r $(1 \leq r \leq k)$, from which we can immediately obtain the RMSDh$^{(r)}$ values $(1 \leq r \leq k)$ too.

Recall that the $L_{i,j}(\mathbf{P}, \mathbf{Q})$ can be computed in constant time after linear-time ($O(n)$) preprocessing (see section 3.2). Thus the values $I_{i,0}$ for all i can be computed in $O(n)$ time in total. Moreover, in case $r > 0$, the value $I_{i,r}$ can also be computed in $O(i - r)$ time by using the values of $I_{j,r-1}$ $(j < i)$. It means that the overall computation time required for computing $I_{n,k}$ (and consequently the RMSDh$^{(k)}$) is $O(kn^2)$. The space required for computing the RMSDh$^{(k)}$ is only $O(n)$, because we only need the information of $I_{j,r-1}$ values (for all j such that $j < i$) to compute the $I_{i,r}$ values for any i.

To compute the positions of the corresponding hinges, we can use the ordinary tracing back technique for DP algorithms, without increasing the time complexity of the overall algorithm. But the space requirement increases to $O(nk)$ space, if we do it naively by using a table of $O(nk)$ size for tracing back. It can be reduced to $O(n)$ space by using a divide-and-conquer technique similar to the Hirschberg algorithm for sequence alignment [Hirschberg 1975], as follows.

Let $J_{i,r}$ be

$$\min_{i \leq \ell_1 < \ell_2 < \cdots < \ell_r \leq n} \sum_{j=0}^{r} L_{\ell_j, \ell_{j+1}-1}(\mathbf{P}[i..n], \mathbf{Q}[i..n]),$$

where we let $\ell_0 = i$ and $\ell_{r+1} = n + 1$ to ease discussion. $J_{i,r}$ can also be computed by DP, as the following equation holds:

$$J_{i,r} = \min_{i \leq j < n} \{L_{i,j}(\mathbf{P}, \mathbf{Q}) + J_{j+1,r-1}\}.$$

Moreover the $I_{n,k}$ can be described as follows,[4] letting $k' = \lfloor k/2 \rfloor$ and $k'' = k - k' - 1$:

$$\min_{k' < i < n - k''} I_{i,k'} + J_{i+1,k''}.$$

The i that minimizes this value is the position of the $(k' + 1)$-th hinge. Let the position be p. To compute it, we need $O(kn^2)$ time and $O(n)$ space.

Similarly, we can next compute the position of the $(\lfloor k'/2 \rfloor + 1)$-th hinge by computing the RMSDh$^{(k')}$ between $\mathbf{P}[1..p-1]$ and $\mathbf{Q}[1..p-1]$ in $O(k'p^2)$ time and $O(p)$ space. Similarly, we can also compute the position of the $(k' + 1 + \lfloor k''/2 \rfloor)$-th hinge by computing the RMSDh$^{(k'')}$ between $\mathbf{P}[p+1, n]$ and $\mathbf{Q}[p+1, n]$ in $O(k''(n-p)^2)$ time and $O(n-p)$ space. Notice that $k'p^2 + k''(n-p)^2 < kn^2/2$, and $kn^2 + kn^2/2 + kn^2/2^2 + \cdots < 2kn^2$. It means that we can compute all the hinge positions in $O(kn^2)$ time and $O(n)$ space by repeating the above until we obtain all of them.

[4] Once we get the $I_{n,k}$, we can immediately compute the RMSDh$^{(k)}$ value, as was described before, *i.e.*, $RMSDh^{(k)}(\mathbf{P}, \mathbf{Q}) = (I_{n,k}/n)^{\frac{1}{2}}$.

5 Experimental Results

In this section, we test the proposed measures against the proteins shown in Table 1. They include 3 pairs of flexible proteins (AK, HIV, and LDH), a pair of rigid proteins (AT), and a pair of non-related proteins (MR). In the experiments, we compare these protein structures without considering gaps (*i.e.* we do not use any alignment programs before comparing them). Table 2 shows the RMSD/RMSDh/RMSDh$^{(k)}$ values for each protein set, for which we describe more details below.

5.1 Adenosylcobinamide kinase

The adenosylcobinamide kinase [Thompson et al. 1998] is known to be a flexible protein with shearing movement. According to our experiment, the RMSDh value is a little smaller than the RMSD value, and it seems still a large value (more than 2.4Å).[5] The hinge position that corresponds to the RMSDh value is between #52 and #53 atoms, but we consider that we should not take it seriously because the RMSDh value is not so small.

The RMSDh$^{(2)}$ drops far beyond the RMSDh, and is less than 1Å. It implies that there seems to be at least two hinges in this protein. The hinge positions obtained in the RMSDh$^{(2)}$ computation are 34–35 (*i.e.*, between #34 and #35 C_α atoms) and 51–52. The hinge position 34–35 is also reported in [Nigham et al. 2007] and [Thompson et al. 1998], while the other hinge could be a new discovery.

5.2 HIV-1 protease

The HIV-1 protease [Perryman et al. 2004] is the major drug target against the AIDS (acquired immunodeficiency syndrome). It is also a flexible protein and its flexibility is said to affect the effectiveness of drugs. In this experiment, the RMSDh$^{(k)}$ drops when $k = 2$, for which the corresponding hinge positions are 44-45 and 56–57. Our result does not contradict literature [Jacobs et al. 2001], which reports a highly flexible region from the 34th residue to the 55th residue.

5.3 Lactate dehydrogenase (LDH)

The lactate dehydrogenase (LDH) [Gerstein et al. 1991] is also known to be a flexible protein. It moves very dynamically, *i.e.*, more than 10Å. The experiment shows that the RMSDh$^{(k)}$ value drops down when $k = 2$, but it is still over 1Å. When $k = 3$, it goes down under 1Å. When $k = 2$, the corresponding hinge positions are 97–98 and 109-110. Both of the two hinge positions are the same as those reported in [Gerstein et al. 1991]. When $k = 3$, the corresponding rigid regions are 97–98, 109-110, and 324–325. In this case, two of the hinges are the same as those obtained when $k = 2$, and an additional hinge is found at 324–325. This hinge is reported in literature [Nigham et al. 2007].

5.4 Other experiments

In experiment AT, we computed the RMSD/ RMSDh/RMSDh$^{(k)}$ values for two independently determined structures of the same protein in the same state. Note that the same set is used in [Nigham et al. 2007] for a test on a pair of almost-the-same proteins. Of course all the values are very small, including the RMSD value. We consider that we do not have to compute RMSDh/RMSDh$^{(k)}$ values if the RMSD is very small like in this case. In contrary, we compared two totally different structures in experiment MR (where a myoglobin and a rhodopsin is compared). In this case, all the values are far larger than those in any other experiments. These experiments show that our measure is effective for discriminating flexible proteins from other normal rigid proteins.

5.5 Computation time

Table 3 shows the time for computing the RMSDh$^{(k)}$ values and corresponding hinge positions for all k ($1 \leq k < n$), using a single 3.2 GHz Pentium D processor with 2 GB memory. It shows that the computation time is very reasonable (less than a second), even though the computation time is $O(n^3)$ (as $k = n - 1$ in these experiments).

[5] In case of rigid proteins, we often claim two structures are similar to each other only when the RMSD between them is smaller than or around 1.0Å. Moreover, the RMSDh/RMSDh$^{(k)}$ values are always smaller than or equal to the RMSD.

Table 1: Protein structures used in our experiments.

Set	Proteins	PDB IDs	#residues
AK	Adenosylcobinamide kinase	1CBU(B),1C9K(B)	180
HIV	HIV-1 protease	1LFG, 1LFH	97
LDH	Lactate dehydrogenase (LDH)	1LDM, 6LDH	329
AT	Asparatate transcarbamoylase	1RAB(A), 1RAC(A)	310
MR	Myoglobin / Rhodopsin (residues:1–154)	101M, 1AYR(A)	154

Table 2: Computation of RMSD/RMSDh/RMSDh$^{(k)}$.

Set	RMSD	RMSDh	RMSDh$^{(2)}$	RMSDh$^{(3)}$	RMSDh$^{(4)}$	RMSDh$^{(5)}$
AK	3.1092	2.4417	0.9772	0.7470	0.5386	0.4755
HIV	1.2451	1.1064	0.7267	0.6483	0.5795	0.5359
LDH	1.7886	1.6160	1.1436	0.8902	0.7234	0.6496
AT	0.1714	0.1672	0.1555	0.1508	0.1430	0.1380
MR	19.3841	15.9145	13.4793	10.7877	9.3194	8.3340

Table 3: Time (sec) for computing RMSDh$^{(k)}$ for all k $(1 \leq k < n)$.

	AK	HIV	LDH	AT	MR
Time	0.219	0.047	0.640	0.610	0.156

6 Conclusions and Future Work

We proposed two new measures for comparing two flexible proteins, which can be very efficiently computed. The first measure, RMSDh, can be computed in linear time, while the other measure, RMSDh$^{(k)}$ can be computed in $O(kn^2)$ time, where k is the maximum number of hinges, and n is the length of the structures. Moreover, we can detect the hinge positions while we compute the measures. Both of the measures are tested on actual flexible proteins to show the correctness of our measures.

To compute the RMSDh$^{(k)}$ value, we need to set appropriate k, but we do not know the actual number of hinges in advance in many cases. For the number, we can use the k where the RMSDh$^{(k)}$ value drops, or the value drops below some threshold (say 1Å), as we did in section 5. But it remains as an open problem how to predict the appropriate k more theoretically and/or efficiently. Moreover, some flexible proteins have large flexible domains. Our measure considers only small flexible domains, and some more flexible measure might be desired for structures with larger flexible domains.

Our algorithms suppose that we know the correspondence between residues of two proteins in advance, but we do not in many cases. Thus, we should develop as future work a flexible structure alignment algorithm that finds the residue correspondence minimizing the RMSDh/RMSDh$^{(k)}$.

References

[Arun et al. 1987] Arun, K. S., Huang, T. S., and Blostein, S. D. 1987. Least-squares fitting of two 3-D point sets. *IEEE Trans Pattern Anal. Machine Intell.*, 9, 698–700.

[Boutonnet et al. 1995] Boutonnet, N. S., Rooman, M. J., and Wodak, S. J. 1995. Automatic analysis of protein conformational changes by multiple linkage clustering, *J. Mol. Biol.*, 253, 633–647.

[Eggert et al. 1997] Eggert, D. W., Lorusso A., and Fisher, R. B. 1997. Estimating 3-D rigid body transformations: a comparison of four major algorithms. *Machine Vision and Applications*, 9, 272–290.

[Eidhammer et al. 2000] Eidhammer, I., Jonassen, I., and Taylor, W. R. 2000. Structure Comparison and Structure Patterns, *J. Comput. Biol.*, 7(5), 685–716.

[Gerstein et al. 1991] Gerstein M., and Chothia, C. 1991. Analysis of protein loop closure, two types of hinges produce one motion in lactate dehydrogenase, *J. Mol. Biol.*, 220, 133–149.

[Golub et al. 1996] Golub, G. H., and Van Loan, C. F. 1996. *Matrix Computation.* 3rd eds., John Hopkins University Press,

[Hirschberg 1975] Hirschberg, D. S., A linear space algorithm for computing maximal common subsequences, 1975. *Commun. ACM*, 18, 341–343.

[Huang et al. 1993] Huang, E. S., Rock, E. P., and Subbiah, S. 1993. Automatic and accurate method for analysis of proteins that undergo hinge-mediated domain and loop movements, *Curr. Biol.*, 3(11), 740–748.

[Jacobs et al. 2001] Jacobs, D. J., Rader, A. J., Kuhn, L. A., and Thorpe, M. F. 2001. Protein flexibility predictions using graph theory, *Proteins: Structure, Function, and Genetics*, 44. 150–165.

[Kabsch 1976] Kabsch, W. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, A32, 922-923.

[Kabsch 1978] Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. 1978. *Acta Cryst.*, A34, 827-828,

[Lemmen et al. 2000] Lemmen, C., and Lengauer, T. Computational methods for the structural alignment of molecules, 2000. *J. Computer-Aided Molecular Design*, 14, 215–232.

[Nigham et al. 2007] Nigham A., and Hsu, D. 2007. Protein Conformational Flexibility Analysis with Noisy Data, *International Conference on Research in Computational Molecular Biology (RECOMB)*,

[Ochagavia et al. 2002] Ochagavia, M. E., Richeele, J., and Wodak, S. J. 2002. Advanced pairwise structure alignments of proteins and analysis of conformational changes, *Bioinformatics*, 18(4), 637-640.

[Perryman et al. 2004] Perryman, A. L., Lin, J., and McCammon. A. 2004. Hiv-1 protease molecular dynamics of a wild-type and of the v82f/i84v mutant: Possible contributions to drug resistance and a potential new target site for drugs, *Protein Science*, 13, 1108–1123.

[Schwartz et al. 1987] Schwartz J. T., and Sharir, M. 1987. Identification of partially obscured objects in two and three dimensions by matching noisy characteristic curves. *Intl. J. of Robotics Res.*, 6, 29–44.

[Shatsky et al. 2004] Shatsky, M., Nussinov, R., and Wolfson, H. J. 2004. FlexProt: an Algorithm for Alignment of Flexible Protein Structures, *J. Comput. Biol.*, 11(1), 83–106.

[Shibuya 2006] Shibuya, T. 2006. Geometric Suffix Tree: A New Index Structure for Protein 3-D Structures, *Combinatorial Pattern Matching 2006 (CPM 2006), LNCS 4009*, 84–93.

[Shibuya 2007] Shibuya, T. 2007. Efficient Substructure RMSD Query Algorithms, *J. Comput. Biol*, to appear.

[Thompson et al. 1998] Thompson, T. B., Thomas, M. G., Escalante-Semerena, J. C., and Rayment, I. 1998. Three-dimensional structure of adenosylcobinamide kinase/adenosylcobinamide phosphate guanylyltransferase from salmonella typhimurium determined to 2.3 a resolution, *Biochemistry*, 37(21), 7686–7695.

[Wolfson et al. 2005] Wolfson, H. J., Shatsky, M., Scheneidman-Duhovny, D., Dror, O., Shulman-Peleg, A., Ma, B., and Nussinov, R. 2005. From Structure to Function: Methods and Applications, *Current Protein and Peptide Science*, 6, 171–183.

[Wriggers et al. 1997] Wriggers, W., and Schulten, K. 1997. Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates, *Proteins: Structure, Function, and Genetics*, 29, 1–14.

[Ye et al. 2003] Ye, Y, and Godzik, A. 2003. Flexible structure alignment by chaining aligned fragment pairs allowing twists, *Bioinformatics*, 19, Suppl. 2, ii246–ii255.