THE NATIONAL UNIVERSITY
of SINGAPORE

School of Computing
Lower Kent Ridge Road, Singapore 119260

TRD1/07

Protein Conformational Flexibility Analysis with Noisy Data□□

Anshul NIGHAM and David HSU

January 2007

Technical Report

Foreword

This technical report contains a research paper, development or tutorial article, which has been submitted for publication in a journal or for consideration by the commissioning organization. The report represents the ideas of its author, and should not be taken as the official views of the School or the University. Any discussion of the content of the report should be sent to the author, at the address shown on the cover.

JAFFAR, Joxan
Dean of School

Protein Conformational Flexibility Analysis with Noisy Data

Anshul Nigham[1] and David Hsu[2]

[1] Singapore–MIT Alliance, Singapore 117576, Singapore
anshulni@comp.nus.edu.sg
[2] National University of Singapore, Singapore 117543, Singapore
dyhsu@comp.nus.edu.sg

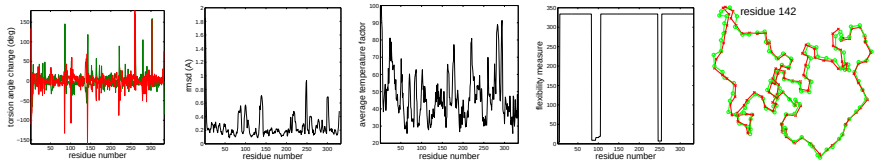
Abstract. Protein conformational changes play a critical role in biological functions such as ligand-protein and protein-protein interactions. Due to the noise in structural data, determining salient conformational changes reliably and efficiently is a challenging problem. This paper presents an efficient algorithm for analyzing protein conformational changes, using noisy data. It applies a statistical flexibility test to all contiguous fragments of a protein and combines the information from these tests to compute a consensus flexibility measure for each residue of the protein. We tested the algorithm, using data from the Protein Data Bank and the Macromolecular Movements Database. The results show that our algorithm can reliably detect different types of salient conformational changes, including well-known examples such as hinge and shear, as well as the flap motion of HIV-1 protease. The software implementing our algorithm is available at `http://motion.comp.nus.edu.sg/projects/proflexana/proflexana.html`.

1 Introduction

Protein structural changes, called conformational changes, play a critical role in vital biological functions such as immune protection, enzymatic catalysis, and cellular locomotion [7]. An example is the "flap" motion of HIV-1 protease, a major inhibitory drug target for AIDS therapy. Conformational changes are a direct consequence of protein structural flexibility and provide insight into the essential link between structure and function.

The structures of an increasing number of proteins have been determined in multiple conformations. In the long term, one may hope to reconstruct, computationally, protein motions from multiple experimentally-determined structures. The motions can then be classified and archived, in order to better understand protein structures and their relationships with protein functions [3]. More immediately, analyses of multiple conformations can help in identifying salient conformational changes, such as hinge or loop motions, as well as in locating active sites in ligand-protein binding [21].

With these goals in mind, our work focuses on analyzing protein conformational changes, an important problem that has received much attention over the years (see Section 2). Specifically, our problem is to identify the flexible and rigid



(a) (b) (c) (d) (e)

Fig. 1. Various methods for detecting flexibility in the N-lobe of lactoferrin. (*a*) Torsion angle differences. (*b*) The minimum RMSD for 5-residue fragments centered at each residue. (*c*) Average temperature factors from X-ray crystallography data. (*d*) Our new algorithm. For (*a*)–(*c*), large absolute values indicate flexible regions. For (*d*), small values indicate flexible regions. (*e*) Superimposition of the two conformations (in red and green, respectively) for the 40-residue fragment centered around residue 142.

regions of a single protein, given its structure, *i.e.*, the 3D coordinates, in two different conformations. An example of such flexible regions is a *hinge*, a consecutive sequence of flexible residues that cause rotational motion between two rigid domains of a protein. This analysis can also be easily extended to more than two conformations through pairwise comparison, if a protein has a relatively small number of distinct conformations that are biologically relevant.

Our problem may appear easily solved by comparing backbone torsion angles, ϕ and ψ. Unfortunately, experimental data obtained through X-ray crystallography or NMR methods are *noisy*. A rigid domain may appear flexible due to noise in the data. Consider Fig. 1*a*, in which the peaks of the curves indicate large differences in torsion angles ϕ and ψ between two conformations of the N-lobe of lactoferrin. The N-lobe of lactoferrin is known to undergo inter-domain motion hinged around residues 90 and 250 [1]. However, the curves appear quite noisy and show many peaks in regions where there are no genuine conformational changes. For example, although there is a sharp peak at residue 142, superimposing the two conformations for the 40-residue protein fragment centered around residue 142 shows no significant conformational change (Fig. 1*e*). Other common methods for detecting conformation changes, such as the root-mean-square distance (RMSD) and the temperature factor, are also susceptible to noise to various degrees (see Fig. 1*b*–*c*).

Key to our problem is to distinguish genuine conformational change from noise. Our algorithm addresses this difficulty at two levels. At the low level, we have developed a reliable statistical test for determining the flexibility of a protein fragment, with noisy data. At the high level, we apply this test to all fragments of a protein and combine information from both short and long fragments to compute a consensus flexibility measure for each residue of the protein. As a result, the algorithm highlights the genuine conformational changes by suppressing the spurious ones due to noise. See Fig. 1*d* for an example, in which our new algorithm unambiguously detects the two main conformations changes in the N-lobe of lactoferrin, despite the noise in the data. Our algorithm takes $O(n^2)$ time for a protein with n backbone atoms. In our tests, it ran at interactive speed even for large proteins with thousands of atoms.

In the following, after a brief review of previous work (Section 2), we first describe our algorithm for protein flexibility analysis under noise (Section 3). We

then give details on efficient implementation of the algorithm and provide a running time analysis (Section 4). Using data from the Protein Data Bank (PDB), we tested our algorithm on proteins that exhibit different types of conformational changes. The results show that our algorithm can reliably detect salient conformational changes (Section 5). We then highlight the main features of the algorithm and address some remaining issues (Section 6). Finally we summarize the results and point out possible future improvements (Section 7).

2 Related Work

Many approaches have been proposed to study protein conformational flexibility. At one extreme, some methods use none or a single experimentally determined protein conformation [11, 14]. In particular, it has been suggested that temperature factors obtained from X-ray crystallography may be correlated with protein flexibility [24]. However, temperature factors reflect mainly the thermal motion and disorder of atoms, and are not reliable for detecting salient conformational changes (see, *e.g.*, Fig. 1*c*). At the other extreme, one may exploit the huge number of different conformations generated by molecular dynamics simulation and infer coordinated motion involving many residues of a protein [22].

Most methods, however, compare two or a small number of experimentally determined conformations, because, despite the rapid growth of protein structural data, the number of known conformations for any particular protein usually remains small. These methods differ in the similarity metric used for comparing protein conformations. They also differ in how they search for flexible and rigid regions of a protein. Below, we briefly review some of them.

Backbone torsion angles are used in several methods to determine the similarity of protein conformations [12, 13, 16]. As mentioned earlier, torsion angles are highly sensitive to noise: small changes in atom coordinates may cause drastic changes in torsion angles. These methods are useful, only if the noise level is extremely low. A better similarity metric makes use of the pairwise distance matrix, in which every entry is the distance between two atoms of a protein [10, 17]. The most commonly used similarity metric is probably the minimum RMSD between the backbone atoms or the C_α atoms of a protein. Other related metrics have been suggested as well [2]. As shown in Fig. 1*b*, RMSD is less sensitive to noise than torsion angles, but not immune to it.

To search for flexible or rigid regions of a protein, the sieve-fit method chooses a rigid core of atoms to align two protein conformations and iteratively improves the alignment until a user-selected threshold is reached [4, 15, 25]. The results are sensitive to the initial choice of the rigid core. A different method uses a heuristic measure of protein flexibility, called the deformation index [10], to locate hinge regions. The fit-all method of Gerstein and Chothia [6] systematically computes the RMSD of all contiguous fragments of a protein between two conformations. It treats the resulting RMSD values as a function of two variables and uses optimization methods to search for the function's inflection points, which indicate hinge regions. However, the search may get stuck locally if not restricted to a suitable domain, which must be chosen manually.

To our knowledge, few methods systematically take into consideration noise in the data when comparing protein structures.

The problem addressed here is related to that of protein structure alignment, which tries to find structural similarities in arbitrary proteins [19, 26]. Our problem is more constrained. We focus on the same protein in different conformations and require no alignment. This simplifies the problem and allows us to develop a more efficient and robust algorithm. However, an important issue common to both problems is to compare the structural similarity of protein fragments. The statistical test that we have developed is thus useful in both problems.

3 Methods

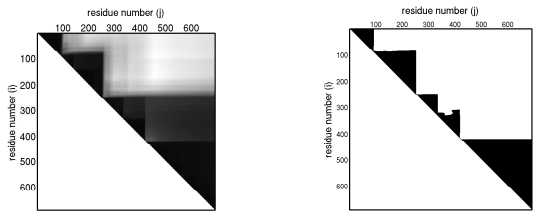
To detect protein conformational flexibility accurately and reliably under noise, we check the flexibility of all contiguous fragments of a protein between two conformations. We then extract a set of "minimal flexible fragments" and use them to compute a flexibility measure for each residue of the protein (Section 3.1). An important element of our algorithm is a statistical test for determining the flexibility of a protein fragment based on the similarity of its structures in two conformations (Section 3.2). The details are described below.

3.1 An all-fragment analysis of protein flexibility

Our algorithm aims to identify flexible and rigid regions of a protein. A flexible protein fragment changes its shape between two conformations, while a rigid fragment remains the same. To distinguish flexible and rigid fragments, we need a measure of similarity between two conformations of a protein fragment. We have chosen the minimum RMSD, a commonly used similarity metric. Intuitively, the minimum RMSD tries to superimpose two conformations of a protein fragment as well as possible, using translations and rotations.

Ideally, the minimum RMSD is 0 if the fragment is rigid and increases as the fragment becomes more flexible. With noisy data, RMSD is unlikely to be 0, even if two conformations are the same. To decide whether a fragment is flexible or not, we use a statistical test to set a threshold for the RMSD. The exact threshold values depend on the amount of noise in the data, the required confidence level, and the length of the fragment tested. In particular, the threshold values are higher for shorter fragments (see Fig. 3). It is thus more difficult to detect small conformational changes in shorter fragments. We defer the detailed discussion until Section 3.2.

Given a suitable threshold, we can compute the minimum RMSD for a fragment of a protein and test its flexibility. However, we still must choose which fragments of the protein to test. If we test short fragments and they turn out to be flexible, we can localize the flexible residues better. On the other hand, short fragments may fail to reveal small conformational changes, which could be masked as noise. We must then rely on longer fragments. To identify all flexible residues accurately and reliably, we examine *all* contiguous fragments and derive a *consistent* interpretation of the information from them. The main advantage of this approach is that it collates information from both long and short fragments and is thus more robust against noise.



(a) (b)

Fig. 2. (a) The RMSD matrix $\hat{\mathbf{R}}$ for lactoferrin. Darker colours indicate smaller RMSD values. (b) The corresponding matrix $\mathbf{T}$. Black indicates 0 (rigid). White indicates 1 (flexible).

RMSD matrices. We represent a protein as a sequence of backbone atoms. Let $F(i, j)$ denote the protein fragment between backbone atoms i and j. The *length* of $F(i, j)$ is the number of backbone atoms contained in it.

We start by computing the minimum RMSD of every contiguous fragment of a protein and storing the results in an upper triangular matrix $\hat{\mathbf{R}}$. For $i < j$, the entry $\hat{R}(i, j)$ of $\hat{\mathbf{R}}$ is the minimum RMSD of the fragment $F(i, j)$ between the two given conformations. For example, Fig. $2a$ shows the pseudo-colored minimum RMSD matrix for lactoferrin, with darker colors indicating smaller RMSD values. The dark triangular regions along the matrix's main diagonal correspond to relatively rigid protein fragments.

Next, we apply our statistical test (see Section 3.2) to threshold each entry of $\hat{\mathbf{R}}$ and compute a new matrix $\mathbf{T}$, which tentatively classifies every contiguous protein fragment as flexible or rigid. If $\hat{R}(i, j)$ is greater than the threshold, the entry $T(i, j)$ of $\mathbf{T}$ is 1, and the fragment $F(i, j)$ is considered flexible. Otherwise, $T(i, j)$ is 0, and $F(i, j)$ is considered rigid. See Fig. $2b$ for an example.

Minimal flexible fragments. The matrix $\mathbf{T}$ contains a wealth of information on the flexibility of a protein, but requires careful interpretation. Suppose that $T(i, j) = 1$, which indicates that the fragment $F(i, j)$ is flexible. Shall we then consider every residue within $F(i, j)$ flexible? The answer is no. Possibly, $F(i, j)$ contains two sub-fragments, one flexible and one rigid. It is thus inaccurate to declare the whole fragment flexible. Also, what shall we do if we have two overlapping fragments, both of which are classified as flexible according to $\mathbf{T}$?

To give a consistent interpretation of the information in $\mathbf{T}$, we introduce the notion of *minimal flexible fragment* (MFF). An MFF is a flexible fragment that contains no proper sub-fragment that is also flexible. In other words, all proper sub-fragments of an MFF are rigid. Two remarks can be made about an MFF $F(i, j)$. First, $F(i, j)$ is flexible, based on the evidence from data. Second, there is no further evidence to attribute the flexibility to any sub-fragment of $F(i, j)$. Therefore, an MFF identifies a flexible region of a protein as accurately as possible, given all the evidence in $\mathbf{T}$.

The definition of MFF implies that a fragment $F(i, j)$ is an MFF if and only if $T(i, j) = 1$ and $T(i', j') = 0$ for $i \leq i' < j' \leq j$ in the upper triangular part of $\mathbf{T}$. This leads to an efficient dynamic programming algorithm for computing the set $\mathcal{L}$ of all MFFs (see Section 4).

The flexibility measure. The length of an MFF $F \in \mathcal{L}$ is correlated with the magnitude of conformational change. As mentioned earlier, the threshold for declaring a fragment flexible is higher for shorter fragments. Thus, small MFF length indicates that conformational changes are large, as they are detectable even in a short fragment. Such conformational changes are often observed in hinge regions, which cause large rotational motion between two rigid domains of a protein and create open and closed conformations. In comparison, large MFF length indicates that statistically significant conformational changes are only detectable in long fragments, which implies that the conformational changes are relatively small. This type of conformational changes include intra-domain motions such as "induced fit", which involves gradual, directed displacements around the binding site of a protein in order to accommodate ligand binding.

The above discussion suggests that the length of an MFF is a good indicator of conformational flexibility, and we use it assign a flexibility measure $f(i)$ to each shortest fragment $F(i, i+1)$, for $1 \leq i < n$. For a given i, let $\mathcal{L}'$ be the subset of $\mathcal{L}$ such that every fragment in $\mathcal{L}'$ contains $F(i, i+1)$ as a sub-fragment. The flexibility measure $f(i)$ for $F(i, i+1)$ is the length of the shortest fragment in $\mathcal{L}'$. Smaller f values indicate higher flexibility. If $\mathcal{L}'$ is empty, we set $f = n+1$ by convention to indicate that $F(i, i+1)$ is rigid.

In practice, we almost always use the standard kinematic model of protein motion. It assumes that bond lengths and bond angles remain fixed during conformational change. In addition, we are usually more interested in determining conformation change at the level of residues rather than atoms. Due to these restrictions, we only need to consider the fragments $F(3i, 3j)$ for $1 \leq i, j \leq n/3$ and assign the flexibility measure to $F(3i, 3i+3)$. In this case, i corresponds to the residue number, and the flexibility measure is assigned on a per residue basis. Our algorithm also applies, with little change, if only the C_α atoms, instead of all the backbone atoms, are used.

Interpretation of the results. The final output of our algorithm is a flexibility measure $f(i)$ for each residue i (see Fig. 6 for examples). For example, $f(50) = 15$ means that to detect conformational change due to residue 50, we require a fragment of at least 15 atoms. Since the length of an MFF is correlated with the magnitude of conformational change, $f(i)$ gives an indication of conformational flexibility at residue i. As another example, if $f_{90} = n+1$, where n is the length of a protein, then no MFF contains residue 90. Any flexible fragment that contains residue 90 must have sub-fragments that are also flexible. Hence the flexibility cannot be reliably attributed to residue 90. It is thus considered rigid.

Instead of giving a binary classification of each residue as either flexible or rigid, our flexibility measure provides a richer description by indicating the degree of flexibility. Based on our experiments, the conformational changes reported in the literature usually have values less than 30 in our measure.

3.2 A statistical test for protein flexibility

To test a fragment F for flexibility, we make the null hypothesis that F is rigid. We then compare the minimum RMSD $\hat{R}$ of F between two given conformations with a threshold r. If $\hat{R} > r$, we reject the hypothesis and consider F flexible;

otherwise, we consider F rigid. The key issue here is to choose a suitable threshold r that is robust against the noise in the data. These thresholds are used to convert the minimum RMSD matrix $\hat{\mathbf{R}}$ to the matrix $\mathbf{T}$, as described in the previous section.

The noise model. Our flexibility test uses the minimum RMSD as the similarity metric. Let (x_i, y_i, z_i) and (x'_i, y'_i, z'_i) for $1 \leq i \leq n$ be the backbone atom coordinates of respectively two conformations q and q' of a protein fragment F. The RMSD is given by $R = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((x_i - x'_i)^2 + (y_i - y'_i)^2 + (z_i - z'_i)^2)}$. The similarity between q and q' is defined as the minimum RMSD $\hat{R}$, which minimizes R over all possible translations and rotations of the two conformations.

Let us now analyze the effect of noise on the distribution of RMSD values. We assume that the noise at each coordinate of each atom of a protein is independently and identically distributed (i.i.d.) according to the normal distribution with zero mean and a given variance. Although this is a simple model, it allows us perform principled statistical analysis and has led to good results in our work (see Section 5) and in related previous work [25].

If the fragment F is rigid, then q and q' actually represent the same conformation. We can apply suitable translation and rotation to the coordinates (x'_i, y'_i, z'_i) so that the resulting new coordinates (x''_i, y''_i, z''_i) are the same as (x, y, z), except for the noise. More precisely, let σ^2 and σ'^2 be the variances of the coordinate noise for q and q', respectively. We have

$$(x_i - x''_i) \sim N(0, \sigma^2 + \sigma'^2), \tag{1}$$

where N denotes a normal random variable, because x_i and x''_i both follow the normal distribution and the sum of normal random variables is again a normal random variable. Thus,

$$\frac{x_i - x''_i}{\sqrt{\sigma^2 + \sigma'^2}} \sim N(0, 1), \tag{2}$$

i.e., a standard normal random variable with mean 0 and variance 1. The same holds for $(y_i - y''_i)/\sqrt{\sigma^2 + \sigma'^2}$ and $(z_i - z''_i)/\sqrt{\sigma^2 + \sigma'^2}$.

Now, let R be the RMSD between (x_i, y_i, z_i) and (x''_i, y''_i, z''_i) for $1 \leq i \leq n$. Consider

$$S = \frac{nR^2}{\sigma^2 + \sigma'^2} = \sum_{i=1}^{n}\left((\frac{x_i - x''_i}{\sqrt{\sigma^2 + \sigma'^2}})^2 + (\frac{y_i - y''_i}{\sqrt{\sigma^2 + \sigma'^2}})^2 + (\frac{z_i - z''_i}{\sqrt{\sigma^2 + \sigma'^2}})^2\right) \tag{3}$$

According to (2), each term in the above sum is a squared standard normal random variable. By definition, S is then a Chi-square (χ^2) random variable with $3n$ degrees of freedom, and R^2 is a scaled χ^2 random variable.

The threshold for the minimum RMSD. To choose the threshold r, we need to bound the probability $\Pr(\hat{R} > r)$. Since $\hat{R} \leq R$, we have $\Pr(\hat{R} > r) \leq \Pr(R > r)$. We thus calculate $\Pr(R > r)$:

$$\Pr(R > r) = \Pr(R^2 > r^2) = \Pr(\frac{nR^2}{\sigma^2 + \sigma'^2} > \frac{nr^2}{\sigma^2 + \sigma'^2}) = 1 - F_{\chi^2}(\frac{nr^2}{\sigma^2 + \sigma'^2}) \tag{4}$$

which follows from (3) and F_{χ^2} denotes the cumulative distribution function of a χ^2 random variable. Given a desired bound p on $\Pr(\hat{R} > r)$, we can calculate the threshold r from (4), which shows that r depends on the noise level in the data (σ and σ'), the p-value, and the length of the protein fragment (n). In particular, r increases with decreasing n (see Fig. 3).

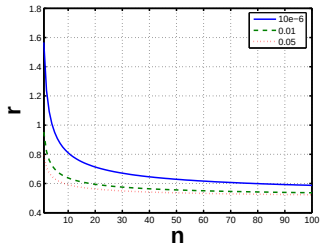


Fig. 3. The change of threshold r as a function of fragment length n, with p-value set to 1×10^{-6}. Each curve corresponds to a different noise level.

The threshold r implies that if F is rigid, then $\hat{R} > r$ with probability at most p. So, the p-value represents the confidence level of our statistical flexibility test. For example, suppose that $p = 0.01$. If $\hat{R} > r$, then F is rigid with probability at most $p = 0.01$, or equivalently, F is flexible with probability at least $1-p = 0.99$.

Choosing the p-value requires some additional thought. Suppose that the probability of incorrectly assigning *any* fragment of a protein to be flexible should be at most γ. The obvious choice of $p = \gamma$ is incorrect, because we encounter the *multiple testing* problem. For a protein with n backbone atoms, our algorithm applies the statistical test once for each contiguous fragment of the protein, resulting in $n(n-1)/2$ tests in total. Let E denote the event that any of the tests gives the incorrect result and E' denote the event that a particular test gives the incorrect result. Then,

$$\Pr(E) \leq \frac{n(n-1)}{2}\Pr(E') \leq \frac{n(n-1)}{2}p.$$

Since we want $\Pr(E) \leq \gamma$, we must choose $p \leq 2\gamma/(n(n-1))$. As an example, for $\gamma = 0.05$ and $n = 300$, p should be smaller than 1.1×10^{-6}.

Although the thresholds calculated this way may appear overly conservative, it is justified due to the presence of noise in the data. Also, they are only used to generate intermediate results stored in $\mathbf{T}$. These results are further synthesized to generate the final output. Tests of our algorithm on PDB data show that it is reliable and does not miss salient conformational changes (Section 5).

4 Computational Efficiency

We now show that our algorithm runs efficiently in $O(n^2)$ time, where n is the number of backbone atoms in a protein. Our algorithm consists of four

main steps: (i) computing the minimum RMSD matrix $\hat{\mathbf{R}}$, (ii) converting $\hat{\mathbf{R}}$ into the matrix $\mathbf{T}$, (iii) extracting the set $\mathcal{L}$ of minimal flexible fragments, and (iv) computing the flexibility measure $f(i)$ for each residue i. To achieve the $O(n^2)$ running time, the most challenging step is to compute the matrix $\hat{\mathbf{R}}$ efficiently. Since $\hat{\mathbf{R}}$ contains $O(n^2)$ entries, we must compute the minimum RMSD for each protein fragment in *constant time* on the average.

4.1 Computing the minimum RMSD matrix

To compute $\hat{R}$ between two given conformations q and q' of a fragment, we apply the eigenvalue algorithm of Horn [9]. First, we compute $(\bar{x}, \bar{y}, \bar{z})$, the centroid of the atom positions for conformation q:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \text{and} \quad \bar{z} = \frac{1}{z}\sum_{i=1}^{n} z_i. \tag{5}$$

The centroid of atoms positions for q', $(\bar{x}', \bar{y}', \bar{z}')$, is computed similarly. Next, we compute the covariances between the coordinates in the two conformations, $(S_{xx}, S_{xy}, S_{xz}, S_{yy}, S_{yz}, S_{zz})$, where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})(x_i' - \bar{x}')$ and the others are computed similarly. Finally, we assemble the matrix $\mathbf{M}$ shown in Fig. 4.

$$\mathbf{M} = \begin{pmatrix} S_{xx}+S_{yy}+S_{zz} & S_{yz}-S_{zy} & S_{zx}-S_{xz} & S_{xy}-S_{yx} \\ S_{yz}-S_{zy} & S_{xx}-S_{yy}-S_{zz} & S_{xy}+S_{yx} & S_{xz}+S_{zx} \\ S_{zx}-S_{xz} & S_{xy}+S_{yx} & -S_{xx}+S_{yy}-S_{zz} & S_{yz}+S_{zy} \\ S_{xy}-S_{yx} & S_{xz}+S_{zx} & S_{yz}+S_{zy} & -S_{xx}-S_{yy}+S_{zz} \end{pmatrix}$$

Fig. 4. The matrix whose largest eigenvalue gives the minimum RMSD.

The minimum RMSD $\hat{R}$ is given by the largest eigenvalue of $\mathbf{M}$ [9]. Computing the centroids and the covariances takes $O(n)$ time for a fragment of length n. Computing the largest eigenvalue takes $O(1)$ time, because $\mathbf{M}$ has constant size. Hence, the minimum RMSD between two conformations of a protein fragment can be computed in $O(n)$ time.

For a protein with n backbone atoms, there are $O(n^2)$ contiguous fragments. By applying the above method to each fragment, the minimum RMSD matrix $\hat{\mathbf{R}}$ can be computed in $O(n^3)$ time.

We now show an incremental algorithm that runs in $O(n^2)$ time, which is asymptotically optimal. A similar algorithm was reported recently in [20]. Recall that the minimum RMSD of all the contiguous fragments can be stored in an upper triangular matrix $\hat{\mathbf{R}}$, in which the entry $\hat{R}(i,j)$ gives the minimum RMSD of $F(i,j)$, for $1 \le i < j \le n$. We compute $\hat{\mathbf{R}}$ incrementally, row by row. Assuming that $\hat{R}(i,j)$ has been computed, we now compute $\hat{R}(i,j+1)$. The key observation is that each centroid coordinate in (5) is a scaled sum of the corresponding atom coordinates. By maintaining a partial sum of the form $\sum_i x_i$ for each

centroid coordinate, we can obtain the new centroid coordinates for $F(i, j+1)$ in $O(1)$ time. Doing the same for the covariances requires some simple algebraic manipulation, which we illustrate for S_{xx}:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} (x_i x_i' - x_i \bar{x}' - \bar{x} x_i' + \bar{x}\bar{x}') \\ &= \sum_{i=1}^{n} x_i x_i' - n\bar{x}\bar{x}' - n\bar{x}\bar{x}' + n\bar{x}\bar{x}' \\ &= \sum_{i=1}^{n} x_i x_i' - n\bar{x}\bar{x}' \end{aligned}$$

By maintaining partial sums of the form $\sum_i x_i x_i'$, we can update the covariances for $F(i, j+1)$, again in $O(1)$ time. Then, the matrix $\mathbf{M}$ is constructed and solved for $\hat{R}$, all in $O(1)$ time. Hence, the minimum RMSD matrix $\hat{\mathbf{R}}$ can be computed in $O(1)$ per entry, or in $O(n^2)$ time in total.

Lemma 1. *For a protein with n backbone atoms, it takes $O(n^2)$ time to compute the minimum RMSD for all contiguous fragments of the protein between two given conformations.*

4.2 Running time analysis

With the minimum RMSD matrix $\hat{\mathbf{R}}$ computed, the remaining three steps are relatively straightforward.

Computing the matrix T Given the RMSD value $\hat{R}$ for a protein fragment, we apply our flexibility test by computing the threshold value r and comparing it to $\hat{R}$. The threshold value r can be computed in $O(1)$ time. Hence, each entry in the matrix $\mathbf{T}$ can be computed in $O(1)$ time from the corresponding entry in $\hat{\mathbf{R}}$. Since $\mathbf{T}$ has $O(n^2)$ entries, computing it requires $O(n^2)$ time.

Extracting minimal flexible fragments. The set $\mathcal{L}$ of MFFs can be extracted from from $\mathbf{T}$, using dynamic programming. To do this, we construct another binary matrix $\mathbf{T}'$ based on $\mathbf{T}$ and go through $\mathbf{T}'$ diagonal by diagonal. We start with the first off-diagonal of T' and set $T'(i, i+1) = T(i, i+1)$ for $1 \leq i < n$. We then move to the next off-diagonal and iterate. If $T(i, j) = 1$, $T'(i+1, j) = 0$, and $T'(i, j-1) = 0$, then the corresponding fragment $F(i, j)$ is an MFF by definition. We add it to $\mathcal{L}$ and set $T'(i, j) = 1$ to indicate that $F(i, j)$ contains a flexible sub-fragment, in this case, itself. If $T(i, j) = 0$, $T'(i+1, j) = 0$, and $T'(i, j-1) = 0$, then $F(i, j)$ is rigid, and all its sub-segments are rigid. We set $T'(i, j) = 0$. Otherwise, we set $T'(i, j) = 1$, because $F(i, j)$ must contain a proper sub-segment that is flexible. Since $\mathbf{T}'$ contains $O(n^2)$ entries, the algorithm completes in $O(n^2)$ time.

Furthermore, the set $\mathcal{L}$ contains at most n fragments. To see this, consider any two fragments $F(i, j)$ and $F(i', j')$ in $\mathcal{L}$. We must have $i \neq i'$. Otherwise,

one fragment would be a sub-fragment of the other. This is impossible, as both are MFFs. Since every fragment in $\mathcal{L}$ must have a distinct i value and $1 \leq i \leq n$, $\mathcal{L}$ contains at most n fragments.

Computing the flexibility measure. Since $\mathcal{L}$ has length at most n, it takes $O(n)$ time to compute the flexibility measure $f(i)$ for each fragment $F(i, i+1)$ and $O(n^2)$ time to compute f for all such fragments in a protein.

Combining the results for the four steps gives the theorem below:

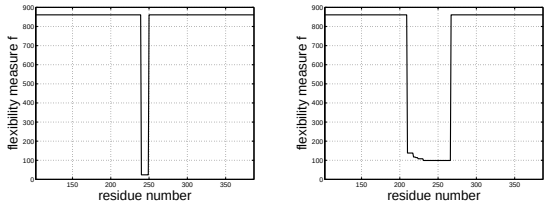
Theorem 1. *For a protein with n backbone atoms, it takes $O(n^2)$ time to compute the flexibility measure f for all fragments $F(i, i+1), 1 \leq i < n$.*

5 Results

We tested our algorithm on both synthetic data, in which case we know the ground truth, and experimental data from the PDB.

5.1 Synthetic data

We took the PDB data for the TBSV coat protein (PDB code 2tbv, residues 102–387) and artificially changed a single torsion angle at residue 245 by 50°. We then added noise with mean 0 and standard deviation 0.2 to the atom coordinates. After creating the new structure, we compared it with the original structure using our algorithm. The results (Fig. 5) show that residues 240–249 are flexible and the rest are rigid. The f-values for residues 240–249 are 24. We then varied the same torsion angle by a smaller amount, 10°, and repeated the test. This time, the algorithm reported a much larger flexible region, residues 212–260, and the f-values for these residues range from 100 to 139. The larger f-values in the second test clearly indicate that the conformational change is smaller than that in the first test, and the residues involved are thus less flexible.



$\theta = 50°$ $\theta = 10°$

Fig. 5. Results for synthesized conformational change. A single torsion angle in residue 245 of the TBSV coat protein is changed by an angle of θ, and Gaussian noise is added to all atom coordinates.

The single torsion angle that was changed is not identified because given the noise in the coordinates, there is not sufficient statistical evidence in the data

Table 1. Test proteins

Protein	Num. Res.	PDB code	σ	Motion
TBSV coat protein	286	2tbv, A	0.2	inter-domain,
(residues 102–387)		2tbv, C	0.2	hinge
adenosylcobinamide	180	1cbu, B	0.2	intra-domain,
kinase		1c9k, B	0.2	shear
lactoferrin	691	1lfg	0.2	inter-domain,
		1lfh	0.2	hinge
HIV-1 protease	99	3hvp	0.1	induced-fit
		4hvp	0.1	
lactate	329	1ldm	0.1	intra- and
dehydrogenase		6ldh	0.1	inter-domain
aspartate trans-	310	5at1, A	0.1	intra- and
carbamoylase		8atc, A	0.1	inter-domain
control	310	1rab, A	0.1	none
		1rac, A	0.1	

that allows this. If the conformational change is small, we must examine a large fragment in order to differentiate genuine conformational change from noise, with confidence. Thus, the smaller the conformational change, the less precisely we can identify the region of flexibility.

5.2 Protein structures

Using PDB data, we tested our algorithm on proteins exhibiting a wide range of conformational changes. Our data set (see Table 1) consists of all the proteins used in [21] to test similar algorithms. It also includes two additional proteins: adenosylcobinamide kinase, which undergoes shear motion, and HIV-1 protease, which undergoes a gradual, induced-fit type of motion. We performed tests on other proteins as well, but cannot report all the results here for lack of space; the readers are encouraged to use our software, which is freely available, to test other proteins of interest.

Tomato Bushy Stunt Virus (TBSV) coat protein. The results on this viral coat protein (Fig. 6*a*) show small f-values for residues 267–276 and very large f-values for the rest of the protein, which suggests a small region of high conformational flexibility with the rest of the protein being rigid. This closely matches the experimental evidence for the conformational change in this protein, which is reported to exhibit rigid-body closure about a single hinge at residues 266–272 [8].

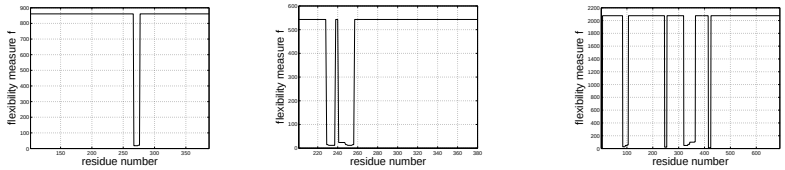
Adenosylcobinamide kinase. Conformational change in adenosylcobinamide kinase is limited to a small fragment and involves the shearing of a helix (chain B, residues 233–249) effected by the residues at both ends of the helix. [23]. This is clearly shown in a morph of two conformations available from the Macromolecular Movements Database [3]. Our results (Fig. 6*b*) agree well with this interpretation. Residues near the ends of the helix (residues 229–236 and 241–256) are identified as the flexible regions. The middle of the helix is not much affected by the shearing. The rest of the protein is rigid.

Lactoferrin. Lactoferrin is responsible for the reversible binding and transport of ferric iron. It contains multiple hinges and is folded into two similar lobes, the N-lobe (residues 1–333) and the C-lobe (residues 345–691), each of which binds with a cation. The domain closure is effected by local changes in two β-strands centered around residues 90 and 250 in the N-lobe [1]. Our results (Fig. 6*c*) correctly identify the two flexible β-strands, residues 90–95 and residues 249–252, which separate the N-lobe into three regions. See also Fig. 1 for improvement of our algorithm over some common existing ones. The C-lobe, which also binds with a cation, may also exhibit conformational flexibility. However, according to earlier work [5], "(The C-lobe) ... shows no appreciable conformational change... The absence of changes in the C-lobe is not completely understood, but could arise from crystal-packing effects." Our plot of the flexibility measure shows a definitive flexible region (residues 415–425) between two rigid regions, which indicates the presence of the suspected conformational change. Movements of the N-lobe relative to the C-lobe is also detected through a flexible region between residue 321 and 362.

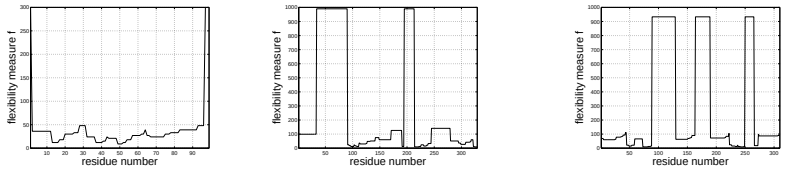
HIV-1 protease. HIV-1 protease plays a critical role in the maturation of HIV-1 virus and is a major inhibitory drug target. Its conformational flexibility affects the effectiveness of various inhibitors [18]. We applied our algorithm to two of the many known conformations of HIV-1 protease. The results (Fig. 6*d*) show that most of the residues have moderately low f-values. Thus, almost the entire protein is flexible to some degree. This reflects that the ligand binding process in HIV-1 protease fits the induced-fit model, in which many small movements of the receptor occur during the binding process. The results also show three regions of high conformational flexibility (residues 14–16, 38–40, and 50–53), and they are consistent with results reported in the literature [11].

Lactate dehydrogenase (LDH). The binding of LDH with the cofactor of nicotinamide adenine dinucleotide (NAD) induces major conformational changes as well as several smaller intra-domain changes. Our results (Fig. 6*e*) indicate regions of maximum flexibility in residues 91–114, 191–196, 214–235 and 322–329, and larger regions of moderate flexibility in residues 1–30, 115–190 and 235–320. These results agree well with the conformational changes suggested by Gerstein and Chothia [6]. The differences occur in only two regions. In [6], residues 1–8 are designated as static, and residues 191–196 have no designation.

Aspartate transcarbamoylase. Aspartate transcarbamoylase, from *E. coli.*, exhibits a complex combination of inter-domain and intra-domain conformational changes. The enzyme is found in two states often referred to as the tense (T) and the relaxed (R) states. Our results show high flexibility in three regions (residues 45–55, 75–90, and 230–246), which correspond to regions of intra-domain conformational changes found in earlier work [21]. Our results also show conformational flexibility in residues 130–155 and residues 260–270, located near the boundaries of known domains. They correspond to inter-domain conformational changes. Several other regions of moderate flexibility within the domains are also detected.



(*a*) TBSV coat protein (*b*) adenosylcobinamide kinase (*c*) lactoferrin



(*d*) HIV-1 protease (*e*) lactate dehydrogenase (*f*) aspartate transcarbamoylase

Fig. 6. Computed protein flexibility measure f, based on PDB data.

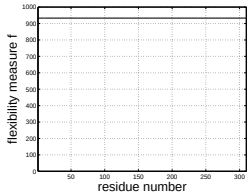


Fig. 7. Control experiment

Control. As a control experiment, we applied our method to two independently-determined structures of aspartate transcarbamoylase in the T state (PDB codes 1rab and 1rac). As expected, no flexibility is detected even at $\sigma = 0.1$ (Fig. 7).

The above results show that our algorithm works well for both inter-domain and intra-domain motions, including well-known examples such as hinge and shear. These results also show that despite its simplicity, the Gaussian model of coordinate noise is adequate, as an approximation, for accurate detection of conformational flexibility.

6 Discussion

An important feature of our algorithm is the assignment of a continuous per-residue flexibility measure, which allows it to handle sharp conformational changes as well as smaller, more gradual ones. Our algorithm does not presuppose the existence of a particular type of conformational change and, as a result, is able to identify a wide range of conformational changes. This is clearly illustrated in the HIV-1 protease example, in which the induced-fit motion is correctly identified. Some alignment algorithms (*e.g.*, [19, 26]) which presuppose the existence of hinges separating rigid domains detect only a single hinge in this protein. This is clearly an incomplete picture of the conformational change in HIV-1 protease.

Our algorithm gives more accurate results than a number of commonly used approaches, as shown earlier in Fig. 1. We believe this is primarily due to the principled treatment of noisy data through the all-fragment analysis at the high level and the statistical flexibility test at the low level.

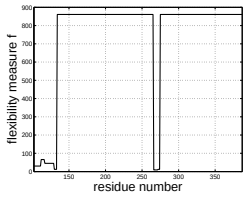


Fig. 8. Applying our algorithm to TBSV with $\sigma = 0.1$.

One issue that affects the accuracy of our algorithm is the setting for σ^2, the variance of the noise in the input protein structure coordinates. Sometimes, σ is readily available from multiple structure determination experiments. Other times, we can get a rough estimate from standard parameters in crystallographic data, such as temperature factors, but getting an accurate estimate may be difficult, as the relationship between these parameters and σ is complex and not easy to establish quantitatively. In such cases, we have found out from our experiments with PDB data that σ values between 0.1 and 0.2 Å work well.

Let us now consider what happens if we over- or under-estimate σ. Essentially, σ controls the sensitivity of our algorithm. For larger σ values, the sensitivity of our algorithm decreases. It detects only more significant conformational changes and may miss some subtle ones, which are masked as noise. For smaller σ values, the sensitivity of our algorithm increases. It is more likely to detect subtle conformational changes, but may also generate some false positives due to noise. For example, we set $\sigma = 0.1$ and re-ran our algorithm on the TBSV coat protein. The result (Fig. 8) is consistent with that for $\sigma = 0.2$ (Fig. 6*a*). The single hinge is detected in both cases. However, the result for $\sigma = 0.1$ shows an additional flexible region at one end of the protein (residues 102–133). This is likely due to increased noise in the structural data at the ends of a protein, rather than genuine conformational change.

Thus, when it is difficult to get an accurate estimate of σ, we can run the algorithm multiple times. We start with a relatively large σ value (say, 0.2) and gradually reduce σ. The conformational changes detected at high σ values are more reliable. With reduced σ, additional, more subtle conformational changes can be detected, but some false positives may also occur.

7 Conclusion

We have developed an efficient algorithm for analyzing conformational changes of a protein. It applies a statistical flexibility test to all contiguous fragments of a protein and combines the information to compute a consensus flexibility measure for each residue of the protein.

We tested the algorithm with PDB data. The results show that our algorithm reliably detects a broad range of protein conformational changes, including both inter-domain and intra-domain ones. Furthermore, this algorithm is fully automated. The user only needs to provide an estimate of the level of noise in the input protein structural data and the required confidence level of the results. In contrast to some earlier algorithms, the algorithm does not require the user to know the type of motion (*e.g.*, hinge or shear) in advance. Neither does it ask the user to select an arbitrary threshold for determining flexible protein fragments. Instead, our algorithm chooses such thresholds automatically based on principled statistical analysis. Our algorithm is efficient. It takes $O(n^2)$ for a protein with n backbone atoms and runs at interactive speed on a desktop PC even for large proteins with thousands of atoms.

Currently, our statistical test assumes that the coordinate noise in each atom is i.i.d., and the basis for identifying genuine conformational change is the magnitude of displacements in atom positions. An interesting extension is to explore the correlation among displacements. This may help improve our algorithm's accuracy in detecting coordinated motion involving many atoms.

While our work focuses on finding the flexible regions of a single protein in different conformations, the principle used by our statistical test for noise analysis applies to many other structural comparison problems. An example is to compare a set of different proteins in order to identify a common domain. When the effect of noise is significant, the statistical test may improve the performance of many algorithms for such problems.

Acknowledgement

We thank Jean-Claude Latombe of Stanford, Tomás Lozano-Pérez of MIT, and Lisa Tucker-Kellogg for insightful discussions, and Jack Snoeyink of UNC at Chapel Hill for supporting the initial phase of the work. This research was funded in part by NUS research grant R252-000-145-112.

References

1. B.F. Anderson, H.M. Baker, G.E. Norris, S.V. Rumball, and E.N. Baker. Apolactoferrin structure demonstrates ligand-induced conformational change in transferrins. *Nature*, 344:784–787, 1990.
2. L.P. Chew, D. Huttenlocher, K. Kedem, and J. Kleinberg. Fast detection of common geometric substructure in proteins. In *Proc. ACM Int. Conf. on Computational Biology (RECOMB)*, pages 104–113, 1999.
3. N. Echols, D. Milburn, and M. Gerstein. MolMovDB: Analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Res.*, 31(1):478–482, 2003. `http://molmovdb.mbb.yale.edu/molmovdb/`.
4. M. Gerstein and R.B. Altman. Average core structures and variability measures for protein families: Application to the immunoglobins. *J. Mol. Biol.*, 251:161–175, 1995.
5. M. Gerstein, B.F. Anderson, G.E. Norris, E.N. Baker, A.M. Lesk, and C. Chothia. Domain closure in lactoferrin. *J. Mol. Biol.*, 234:357–372, 1993.

6. M. Gerstein and C. Chothia. Analysis of protein loop closure: Two types of hinges produce one motion in lactate dehydrogenase. *J. Mol. Biol.*, 220(1):133–149, 1991.
7. M. Gerstein, R. Jansen, T. Johnson, J. Tsai, and W. Krebs. Motions in a database framework: from structure to sequence. In M.F. Thorpe and P.M. Duxbury, editors, *Rigidity Theory and Applications*, pages 401–442. Kluwer Academic/Plenum Publishers, 1999.
8. P. Hopper, S.C. Harrison, and R.T. Sauer. Structure of tomato bushy stunt virus. V. Coat protein sequence determinations and its structural implications. *J. Mol. Biol.*, 177(4):701–713, 1984.
9. B.K.P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society A*, 4(4):629–642, 1987.
10. E.S. Huang, E.P. Rock, and S. Subbiah. Automatic and accurate method for analysis of proteins that undergo hinge-mediated domain and loop movements. *Curr Biol.*, 3(11):740–748, 1993.
11. D.J. Jacobs, A.J. Rader, L.A. Kuhn, and M.F. Thorpe. Protein flexibility predications using graph theory. *Proteins: Structure, Function, and Genetics*, 44(2):150–165, 2001.
12. D. Joseph, G. A. Petsko, and M. Karplus. Anatomy of a conformational change: Hinged "lid" motion of the triose phosphate isomerase loop. *Science*, 249:1425–1428, 1990.
13. A.P. Korn and D.R. Rose. Torsion angle differences as a means of pinpointing local polypeptide chain trajectory changes for identical proteins in different conformational states. *Protein Engineering*, 7:961–967, 1994.
14. W.G. Krebs, V. Alexandrov, C.A. Wilson, N. Echols, H. Yu, and M. Gerstein. Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic. *Proteins: Structure, Function, and Genetics*, 48(4):682–695, 2002.
15. A.M. Lesk. *Protein Architecture: A Practical Approach*. Oxford University Press, 1991.
16. M. Levine, D. Stuart, and J. Williams. A method for systematic comparison of the three-dimensional structures of proteins and some results. *Acta Crystallography*, A40:600–610, 1984.
17. A. Nishikawa, T. Ooi, Y. Isogai, and N. Saito. Tertiary structure of proteins. *J Phys. Soc. Jpn.*, 32:1333–1337, 1972.
18. A.L. Perryman, J. Lin, and A. McCammon. 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, 2004.
19. M. Shatsky, H.J. Wolfson, and R. Nussinov. Flexible protein alignment and hinge detection. *Proteins: Structure, Function and Genetics*, 48:242–256, 2002.
20. T. Shibuya. Geometric suffix tree: A new index structure for protein 3-d structures. In *Combinatorial Pattern Matching*, LNCS 4009, pages 84–93, 2006.
21. S. Subbiah. *Protein Motions*. Chapman & Hall, 1996.
22. M. Teodoro, G.N. Jr. Phillips, and L.E. Kavraki. A dimensionality reduction approach to modeling protein flexibility. In *Proc. ACM Int. Conf. on Computational Biology (RECOMB)*, pages 299–308, 2002.
23. T.B. Thompson, M.G. Thomas, J.C. Escalante-Semerena, and I. Rayment. Three-dimensional structure of adenosylcobinamide kinase/adenosylcobinamide phosphate guanylyltransferase from salmonella typhimurium determined to 2.3 a resolution,. *Biochemistry*, 37(21):7686–7695, 1998.
24. M. Vihinen, E. Torkkila, and P. Riikonen. Accuracy of protein flexibility predictions. *Proteins*, 19(2):141–149, 1994.
25. W. Wriggers and K. Schulten. Protein domain movements: Detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. *Proteins: Structure, Function, and Genetics*, 29:1–14, 1997.

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