Analysis of motion features for molecular dynamics simulation of proteins

Mayumi Kamada a, Mikito Toda b,⇑, Masakazu Sekijima c, Masami Takata d, Kazuki Joe d

⇑ Corresponding author. Fax: +81 742 20 3383.
E-mail addresses: kamada@kucr.kyoto-u.ac.jp (M. Kamada), toda@ki-rin.phys.nara-wu.ac.jp (M. Toda).

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan
b Department of Physics, Faculty of Science, Nara Women’s University, Kitauoyanishi-machi, Nara 630-8506, Japan
c Global Scientific Information and Computing Center, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
d Graduate School of Humanities and Sciences, Nara Women’s University, Kitauoyanishi-machi, Nara 630-8506, Japan

1. Introduction

Fluctuation of proteins is supposed to play an important role for their functions [1,2]. This leads us to investigate the relationship between the dynamics of proteins and their functions. Such studies would enable us to control the functional behavior of the protein and to understand how certain diseases arise.

The dynamics of proteins take place on the potential surface with a large number of local energy minima, and their behavior involves both harmonic and unharmonic motions [3]. When the temperature is low enough, the protein just exhibits small fluctuation within a single energy minimum and its dynamics is almost harmonic. Such fluctuation around local minima can be described by Normal Mode Analysis (NMA), and several methods using NMA have been developed [4–8]. However, with an increasing temperature, unharmonic fluctuation becomes larger and larger, and eventually gives rise to transitions among local minima. When unharmonic motions become dominant, we need new methods which pay due attention to unharmonicity.

Fluctuation of proteins is not just oscillations of small amplitude but involves large conformational changes. For example, the idea of ‘population-shift’ is presented [9], where proteins are supposed to experience conformational changes around multiple conformations even without ligand binding. Based on this idea, the mechanism of the functional behavior of proteins becomes clearer these days [10–14].

The purpose of this Letter is to apply a new method for time series analysis using the wavelet transformation toward molecular dynamics simulation of proteins involving conformational changes. Because of unharmonicity, the amplitude and the frequency of the motions vary along the time evolution. Such transient features of motions can be revealed using the wavelet transformation. In particular, we are interested in the proteins wandering around multiple conformations. From this point, the dynamics of Thermomyces lanuginosus lipase (TLL) is a good example since multiple conformations, either open or closed, are found experimentally.

The content of this Letter is as follows. In Section 2, we explain the target protein of our study TLL, and give a brief review of the method of time series analysis using wavelet transformation [15]. In Section 3, we present the results of our analysis toward molecular dynamics simulation of TLL. In Section 4, we conclude our studies and discuss future prospects including the study toward ‘disordered proteins’.

2. Target and method

2.1. Target molecule

Our target protein is Thermomyces lanuginosus lipase (TLL), and it has 269 residues. Lipase TLL is one of the lipases (EC: 3.1.1.3) which catalyze the hydrolysis of fat, and these lipases have a common α/β-hydrolase fold inside of their conformations. This fold provides with a stable scaffold for the active sites [16,17]. Their active sites are covered by the ‘lid’, which, for TLL, corresponds to the residues from 84 to 93. When the lid is open, the active sites become accessible to the substrate and the lipase shows activity. Hence, its conformation is referred to as the active (open) conformation. When the lid is closed, the lipase does not show its activity, and its conformation is referred to as the inactive (closed) conformation.
Seven conformations are reported for TLL in PDB [22], three open (1DT5, 1EIN and 1GT6) and four closed (1DTE, 1DU4, 1DT3 and 1TIB) ones [23]. In Figure 1, we show the secondary structure of 1TIB [24], which we will use as an initial structure of our MD simulation.

In Figure 2, we show TLL’s B-factors [22] for the seven conformations. The difference of the B-factors reveals how their motions differ. The closed conformations tend to exhibit pronounced peaks and the open conformations 1EIN and 1GT6 show flatter ones, though the locations of their peaks in general coincide. The B-factors of the α/β-hydrolase exhibit smaller values implying their stability, and that of the ‘lid’ for 1TIB and 1DTE reveal pronounced peaks.

2.2. Molecular dynamics data

The time-series data of TLL are obtained by MD simulation using SANDER program of AMBER9. The simulation procedure is as follows. Starting from the initial conformation 1TIB taken from PDB, we first perform 2000 minimization cycles of the potential energy using Conjugate Gradient Method. Second, we relax the system from 0 K to 300 K by 5 K every 10 ps with 24 Å cut-off. For setting the temperature, we use AMBER ff99 force field with a small modification [26] and NTP ensemble with a given temperature and pressure. The histidine has a proton bonded at the epsilon position. The solvent molecules are treated implicitly using Generalized Born (GB) model by Hawkins et al. [27]. After these processes, we perform simulation at 300 K. Using GB model of 24 Å cut-off, the calculation is done at 300 K with time-step of 1 fs, and we sample the data every 10000 step, i.e. the time step for sampling is 10 ps. The time-series data for the analysis are the space coordinates of the Cα atoms for 2 ns (200 sampled data), and we adjust the trajectory so that its center of gravity does not move.

2.3. Time series analysis

In this Letter, we apply a new feature extraction method using the wavelet transforms [15]. This method is composed of two steps, the wavelet transforms together with the low-pass filter, and the singular value decomposition (SVD). In the following, we use the Morlet wavelet as the wavelet transformation, and apply the method to time series data of Cα atoms. Suppose that we have a given time-series data \( q_n(t) \) \( (n = 0, \ldots, 3N - 1) \) of \( N \) Cα atoms obtained by MD simulation of the protein having \( N \) residues. First, we apply the wavelet transforms to the data of each degree of freedom \( n \), and obtain their transformations \( q_n(t, \omega_0) \) with \( \omega_0 = \frac{2\pi}{T} \) where \( T \) is the time duration of the data. In order to extract slow motion, we retain only those components of lower frequencies, i.e. \( l = M_1, \ldots, M_2 \) \( (0 < M_1 < M_2 < \frac{T}{2} - 1) \) and \( l = M - M_1, \ldots, M - M_2 \), i.e. complex conjugates of those with \( l = M_1, \ldots, M_2 \). Here, \( M \) is the number of data in each time series of the Cα atom. Using only these
components, we construct the matrix \( A(t) = \{ A_i(t) \} \) with \( A_i(t) = q_i(t, \omega_0) \) for \( i \) in the above ranges. We set \( A_i(t) = 0 \) otherwise. Second, we apply SVD to the matrix \( A(t) \) at each time \( t \), and obtain its left-singular vectors, right-singular vectors and singular values. They describe motion features of the vibrations of Cx atoms in space, those of the fluctuation of frequencies and their amplitudes, respectively. Note that left-singular vectors thus obtained are real vectors.

2.4. Reduction of the degrees of freedom

We apply the wavelet transforms to the trajectory data obtained by MD simulation of TLL. Lipase TLL consists of \( N = 269 \) residues. The total length of the time series is \( 2 \) [ns] with the time step of \( 10 \) [ps]. Thus, the total number \( M \) of the time series is \( M = 200 \). In order to avoid artifacts caused by the finiteness of the time series, we use the data between \( 0.2 \) [ns] and \( 1.8 \) [ns] in the following analysis. In the low-pass filter, we set \( M_1 = 15 \) and \( M_2 = 50 \). This corresponds to the frequency range with its period between \( 133 \) [ps] and \( 40 \) [ps]. Then, the matrix \( A(t) \) has \( 3N = 807 \) rows and \( 2(M_2 - M_1 + 1) = 72 \) non-zero columns. As the result of SVD, we obtain at most \( K = \min(3N, 2(M_2 - M_1 + 1)) \) non-zero singular values and their left-singular and right-singular vectors. We use Octave [28] for the calculation of SVD.

The purpose of SVD is to reduce the degrees of freedom. In order to see to what extent the reduction is possible, we estimate the weight \( W_j(t) \) of the \( j \)th singular value as follows,

\[
W_j(t) = \frac{(\lambda_j(t))^2}{\sum_{i=1}^N (\lambda_i(t))^2},
\]

where \( \lambda_i(t) \) is the \( i \)th singular value at time \( t \). In Figure 3, we show how the weights \( W_1(t) \) and \( W_2(t) \) of the first and the second singular values vary as time evolves, respectively. We also indicate time evolution of their sum \( W(t) = W_1(t) + W_2(t) \). There, we can see that the first singular value alone accounts for about 30% of the dynamics. When we include both the first and the second singular values, about 50% of dynamics is retained. In the following, we focus our attention to the first singular vector as the first step of our analysis.

3. Analysis

Here, we present results of our analysis concerning motion features using the first left singular vector \( e_{i-1}(t) \). The left singular vectors describe spatial features of the dynamics. First, in Section 3.1, we characterize collective motion of Cx atoms. Second, in Section 3.2, we quantify collective motion of the secondary structures and investigate correlation of the motion between the secondary structures.

3.1. Collective motion of the protein

3.1.1. Indexes characterizing collective motion

We define ‘collective motion’ of the protein as a kind of motion when neighboring Cx atoms oscillate along similar directions. We characterize ‘collective motion’ around the \( p \)th Cx atom as shown in Figure 4. There, the three-dimensional vector \( u_p(t) \) is defined using the first left singular vector by \( e_{i-1}(t) = (u_1(t), \ldots, u_3(t), \ldots, u_N(t)) \). We call \( u_p(t) \) the oscillation vector of the \( p \)th Cx atom at time \( t \). Then, similarity of the oscillation vectors can be captured by either their inner product or the cosine of the angle between them. Neighborhood of the \( p \)th Cx atom can be taken either along the sequence of the protein or within its three-dimensional conformation. Thus, we can introduce four indexes \( x^i_p(t) \) \((i = 1, \ldots, 4)\) as is summarized in Table 1. In their definitions, \( n \) is the number of Cx atoms from the \( p \)th Cx atom, and \( r \) is the distance from the \( p \)th Cx atom in space.

3.1.2. Time evolution of collective motion

In Figure 5, time evolution of the four indexes \( x^i_p(t) \) \((i = 1, \ldots, 4)\) are shown. Here, we set \( n = 10 \) and \( r = 10 \) [Å]. There, bright color indicates that the collective motion characterized by the index \( x^i_p(t) \) is stronger. The results using the inner product show that pronounced motion appears near the 90th Cx atom, i.e. the ‘lid’, during the time interval around \( t = 0.45 \) [ns]. For other time ranges, different types of collective motion are exhibited. The results using the cosine reveal much broader excitation of collective motion. On the other hand, the definition of the neighborhood does not give rise to different features. In the following section, we apply clustering algorithm to classify different types of collective motion.

3.1.3. Clustering of collective motion

In the previous section, we show that collective motion exhibits different features along time evolution. Here, we apply clustering algorithm to classify these features, and investigate correlation between these features and the \( B \)-factors shown in Section 2.1.  

3.1.3.1. Clustering. Here, we classify collective motion at different times using Affinity Propagation (AP) [29]. In this method, we can change the number of clusters by the parameter called ‘preference’, and these clusters are represented by their cluster centers called ‘exemplars’. We define ‘similarity’ \( s(t_i, t_j) \) of collective motion between different times \( t_i \) and \( t_j \) as

\[
s(t_i, t_j) = -\sum_{p=1}^{N} \left[ x_p^2(t_i) - x_p^2(t_j) \right]^2,
\]

where we use the index \( x_p^2(t) \) as the characteristics of collective motion. We choose the value of ‘preference’ so that we obtain four clusters. Their cluster centers are at times 0.42 [ns], 0.57 [ns], 1.34 [ns] and 1.54 [ns], respectively. They are indicated by the red arrows in Figure 5.

3.1.3.2. Correlation between collective motion and TLL structures. In Table 2(a), we show the correlation between the index \( x_p^2(t_i) \) at four cluster centers \( t_j \) \((j = 1, \ldots, 4)\) and \( B \)-factors of seven TLL structures. The center at \( t_j = 0.42 \) [ns] reveals large correlation with the \( B \)-factor of 1TIB. This is reasonable considering the facts
that the ‘lid’ exhibits pronounced motion around this time range and that the ‘lid’ reveals the characteristic peak in the $B$-factor of 1TIB. For other cluster centers, larger correlations with multiple structures are exhibited. For example, the center at 1.54 [ns] has largest correlation with the structure of 1GT6, and relatively larger ones with the structures of 1DT5, 1EIN, 1DU4 and 1TIB.

Table 1

<table>
<thead>
<tr>
<th></th>
<th>Inner product</th>
<th>Cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Dimension</td>
<td>$x_p^{(1)}(t) = \frac{1}{2n} \sum_{</td>
<td>j</td>
</tr>
<tr>
<td>3 Dimension</td>
<td>$x_p^{(3)}(t) = \frac{1}{b_r(t)} \sum_{</td>
<td>j</td>
</tr>
</tbody>
</table>

that the ‘lid’ exhibits pronounced motion around this time range and that the ‘lid’ reveals the characteristic peak in the $B$-factor of 1TIB. For other cluster centers, larger correlations with multiple structures are exhibited. For example, the center at 1.54 [ns] has largest correlation with the structure of 1GT6, and relatively larger ones with the structures of 1DT5, 1EIN, 1DU4 and 1TIB. To estimate...
structural similarity, we calculate RMSD by GASH algorithm [30] between each TLL structure and the structures of the members of each cluster, and values of RMSD are averaged over the members of the cluster. In Table 2(b), we show the averaged values of RMSD between the TLL structures and the clusters. The cluster 1 with its exemplar \( t_c 1 = 0.42 \text{ ns} \) has the smallest RMSD value with the 1TIB structure. For other clusters, they are structurally closer to other TLL structures than 1TIB. Such transient features imply that the dynamics is not limited in a single potential well but occurs around multiple potential wells of different conformations.

### Table 2

(a) The coefficients of the correlation between the index \( x^{ij}_p(t_c) \) at four cluster centers \( t_c (j = 1 \ldots 4) \) called ‘exemplars’ and \( B \)-factors of seven TLL structures. The coefficients with * are judged as insignificant at the 5% significance level. (b) The averaged RMSD values between each TLL structure and the structures of the members of each cluster, where the average is done over the members of the cluster. The value of RMSD is calculated using GASH program [30].

<table>
<thead>
<tr>
<th></th>
<th>Open</th>
<th>Closed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1DT5</td>
<td>1EIN</td>
</tr>
<tr>
<td>(a) Correlation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exemplar1 (0.42 ns)</td>
<td>0.15899</td>
<td>0.02312*</td>
</tr>
<tr>
<td>Exemplar2 (0.97 ns)</td>
<td>0.17992</td>
<td>0.27449</td>
</tr>
<tr>
<td>Exemplar3 (1.34 ns)</td>
<td>0.40421</td>
<td>0.36016</td>
</tr>
<tr>
<td>Exemplar4 (1.54 ns)</td>
<td>0.27327</td>
<td>0.24812</td>
</tr>
<tr>
<td>(b) RMSD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster1 (0.42 ns)</td>
<td>1.4284</td>
<td>1.4063</td>
</tr>
<tr>
<td>Cluster2 (0.97 ns)</td>
<td>1.4996</td>
<td>1.4419</td>
</tr>
<tr>
<td>Cluster3 (1.34 ns)</td>
<td>1.4856</td>
<td>1.4191</td>
</tr>
<tr>
<td>Cluster4 (1.54 ns)</td>
<td>1.5080</td>
<td>1.4271</td>
</tr>
</tbody>
</table>

#### 3.2. Collective motion of the secondary structures

Here, we study collective motion of the secondary structures. In Figure 6, we show correspondence between the secondary structures and the indexes \( x^{ij}_p(t) (i = 1 \ldots 4) \) at a given time \( t \). There, the horizontal axis is the residue number \( p \) and the vertical axis is \( x^{ij}_p(t) (i = 1 \ldots 4) \) at time \( t = 0.6 \text{ ns} \). We also indicate the secondary structures using the colors, i.e. pink, blue and white indicate \( \alpha \)-helix, \( \beta \)-strand and loop, respectively. The red arrows show that the indexes using the inner products reveal pronounced peaks which correspond to collective motion of loops and parts of \( \alpha \)-helices nearby loops. On the other hand, the figures using the cosine exhibit more peaks which correspond to \( \alpha \)-helices and \( \beta \)-strands. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this Letter.)

![Figure 6](image-url)
peaks which correspond to collective motion of loops and parts of α-helixes nearby loops. On the other hand, the figures using the cosine exhibit more peaks which correspond α-helixes and β-strands.

These results suggest that correspondence exists between collective motions and the secondary structures. In order to evaluate the correspondence quantitatively, we introduce new indexes \( R_{k}^{(l)}(t) \) of collective motion for the \( k \)th secondary structure as follows,

\[
R_{k}^{(l)}(t) = \sum_{p=1}^{N_{k}} v_{kp} x_{p}^{(l)}(t). \tag{3}
\]

The structure vector \( \mathbf{v}_{k} = (v_{kp}) \) of the \( k \)th secondary structure is defined in such a way that \( v_{kp} = 1/N_{k} \) when the \( p \)th Ca atom belongs to the \( k \)th secondary structure and \( v_{kp} = 0 \) otherwise. Here, \( N_{k} \) is the number of Ca atoms which belong to the \( k \)th secondary structure. In the following, we use the index \( R_{k}^{(l)}(t) \) since the indexes using the cosine capture collective motion of not only loops but also α-helixes and β-strands.

3.2.1. Correlation of collective motion between secondary structures

Correlation of the index \( R_{k}^{(l)}(t) \) between different secondary structures reveal collective motion beyond the level of each secondary structure. In Figure 7, we compare the correlation of \( R_{k}^{(l)}(t) \) of different secondary structures with the time-averaged distance of Ca atoms. The upper left of Figure 7 displays the correlation of \( R_{3}^{(-4)}(t) \) and \( R_{3}^{(-3)}(t) \) between 3rd and 4th secondary structures. There, red, blue and white describe positive, negative and no-correlation, respectively. The lower right of Figure 7 displays the time average \( \langle |x_{p}(t) - \bar{x}_{p}| \rangle \) of the distance between \( p \)th and

**Figure 7.** Comparison of the correlation of \( R_{k}^{(l)}(t) \) of different secondary structures with the time-averaged distance of Ca atoms. The upper left displays the correlation \( R_{k}^{(l)}(t) \) between \( k \)th and \( l \)th secondary structures. There, red, blue and white describe positive, negative and no-correlation, respectively. The lower right displays the time average \( \langle |x_{p}(t) - \bar{x}_{p}| \rangle \) of the distance between \( p \)th and \( q \)th Ca atoms. There, the brighter the color of \((p,q)\) is, the closer their distance is. The horizontal and vertical axes indicate the Ca number \( p \) with the colors, pink, blue and white, describing α-helix, β-strand and loop, respectively. The secondary structure number \( k \) is the same as that shown in Figure 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this Letter.)

**Figure 8.** Important sites of TLL. (A) The ‘lid’ and active sites in 1TIB. The ‘lid’ corresponds the structure v15 in Figure 1, and active sites Ser146, Asp201 and His 258 are included in the structures v23, v30 and v40, respectively. (B) Those sites which change their forms between open and closed conformations. Green and pink colors describe the open and closed conformations, respectively. One is the ‘lid’(I). Another is the structure from residues 36 to 40. While these residues form α-helix in the closed conformations, they unwind and form loop in the open ones (II). (C) The correlation coefficients involving important sites. The horizontal axis is the structure number \( k \) and the vertical axis is the correlation coefficient. Red, blue and grey dots represent positive, negative and no-correlation, respectively. (I) Correlation involving the structure v15, ‘lid’. (II) Correlation involving the structure v06. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this Letter.)
qth Ca atoms. There, the brighter the color of (p, q) is, the closer their distance is. In the above definition, the bracket indicates time averaging. The horizontal and vertical axes indicate the Ca number p with the colors, pink, blue and white, describing α-helix, β-strand and loop, respectively. In Figure 7, positive correlation exists between the secondary structures located near in space. On the other hand, we see negative correlation between the structures located on opposite sides of the protein.

3.2.2. Correlation of collective motion involving important sites

Here, we focus our attention to the correlation involving important sites of the protein. The ‘lid’ plays an important role in the conformational change of TLL. In Figure 8A, we show the ‘lid’ and the active sites of TLL. The ‘lid’ corresponds to the secondary structure v15 in Figure 1. The structure v15 varies its form as shown in Figure 8B-I. In addition to the ‘lid’, the secondary structure v06 changes its form in the conformational change as shown in Figure 8B-II. In both figures, green and pink colors describe the open and closed conformations, respectively. It is known that while v06 forms loop in open conformations, it does α-helix in the closed conformations.

In Figure 8C-I and II, we show correlation of collective motion involving either v15 (the ‘lid’) or v06, respectively. There, the horizontal axis is the secondary structure number k, and the vertical axis is the correlation coefficients. The coefficients are judged as significant at the 5% significance level, and red, blue and grey dots indicate positive, negative and no-correlation, respectively.

There, it is noticeable that the structures v15 (the ‘lid’) and v06 have positive correlation with each other. This implies that collective motion involving these two exists in the conformation change. It is remarkable that both v15 (the ‘lid’) and v06 have no-correlation with eight β-strands constituting a[β]-hydrolase fold. It is also interesting that the structure v15 (the ‘lid’) has relatively poor positive correlation with the structures including the active sites v23, v30 and v40, and the structure v06 has no-correlation with them. These results suggest that the conformational motion involving the structures v06 and v15 does not affect much the functionally important parts of the protein such as the active sites. It is reasonable since the protein would lose its function as an enzyme if these parts vary their forms in the conformation change.

4. Conclusions and future prospects

We have applied a new method for time series analysis [15] to molecular dynamics simulation of lipase TLL. Introducing indexes to characterize collective motion of the protein, we have obtained the following two results. First, time evolution of the collective motion reveals that vibrational motion is not limited to a single potential well but takes place involving multiple conformations. This implies that the protein wanders around multiple conformations during its time evolution. Second, correlation of the collective motion between secondary structures shows that collective motion involving multiple secondary structures exists. Furthermore, collective motion involving the conformational change does not affect much the structures of the functionally important sites.

Our results indicate that the method developed in [15] has applicable to reveal collective motion beyond near harmonic motion. Here, as the first step of our analysis, we have paid our attention to the first singular vector. We will develop our method to include more singular vectors. As a future target, ‘disordered proteins’ are important subjects [31]. These proteins exhibit large conformational changes, and such changes are supposed to play an important role in their functions. Then, new methodology is necessary which is applicable to large unharmonic motion. Since our method is applicable to such unharmonic motion, we expect that the method becomes an important tool to understand the dynamical behavior of ‘disordered proteins’.

Acknowledgements

M.K. and M.T. would like to thank Ms. N. Sakurai for her contribution in developing the method of time series analysis using the wavelet transformation, Ms. S. Kimura for her contribution in the initial stage of this work. M.T. is supported by Grant-in-Aid for Scientific Research (Grant No. 21540413) JSPS, Grant-in-Aid for Research on Priority Area ‘Real Molecular Theory’ MEXT, and Nara Women’s University Intramural Grant for Project Research.

References