Time series analysis of molecular dynamics simulation using wavelet

Mikito Toda

Citation: AIP Conference Proceedings **1468**, 367 (2012); doi: 10.1063/1.4745595
View online: http://dx.doi.org/10.1063/1.4745595
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1468?ver=pdfcov
Published by the AIP Publishing

Time series analysis of molecular dynamics simulation using wavelet

Mikito Toda

Nara Women's University,
Nara, 630-8506, Japan
E-mail: toda@ki-rin.phys.nara-wu.ac.jp

Abstract. A new method is presented to extract nonstationary features of slow collective motion toward time series data of molecular dynamics simulation for proteins. The method consists of the following two steps: (1) the wavelet transformation and (2) the singular value decomposition (SVD). The wavelet transformation enables us to characterize time varying features of oscillatory motions and SVD enables us to reduce the degrees of freedom of the movement. We apply the method to molecular dynamics simulation of various proteins such as Adenylate Kinase from *Escherichia coli* (AKE) and *Thermomyces lanuginosa* lipase (TLL). Moreover, we introduce indexes to characterize collective motion of proteins. These indexes provide us with information of nonstationary deformation of protein structures. We discuss future prospects of our study involving "intrinsically disordered proteins".

Keywords: molecular dynamics, wavelet transformation, singular value decomposition, proteins, biophysics, molecular functions, intrinsically disordered proteins
PACS: 87.15.A-,87.15.ap,87.15.B-,87.15.H-,87.15.Ya

INTRODUCTION

Dynamical properties of proteins offer a crucial clue to understand how proteins perform their functions [1, 2, 3]. In particular, slow collective motions involve large conformational changes of proteins, and play an important role in various aspects such as ligand binding and signal transduction [4]. In order to investigate such dynamical behavior, molecular dynamics simulations involving all atoms are performed to obtain time series data of how proteins move [5]. These data provide us with a detailed information concerning motions of individual atoms which constitute the protein.

In order to capture large conformational changes, we need a method to extract slow collective behavior from time series of individual atoms. Principal Component Analysis (PCA) is one of the most frequently used methods for this purpose [6, 7, 8, 9]. However, PCA is not suitable for extracting dynamical information, since PCA only pays attention to the static properties of the distribution. This leads us to develop new methods for characterizing slow collective movement [10, 11, 12].

Importance of these methods can be readily seen by considering a gap of time scales between the all-atom simulation and functional behavior of proteins. We should also note a large gap of time scales between the simulation and experiments such as single molecule spectroscopy [13, 14]. Thus, direct comparison of these experiments with the simulation is still difficult. This gap propels us to construct coarse grained models which only take into account collective degrees of freedom [15, 16, 17, 18]. If we

Let's Face Chaos through Nonlinear Dynamics
AIP Conf. Proc. 1468, 367-374 (2012); doi: 10.1063/1.4745595

establish definite methods to extract slow collective motions from time series of the all-atom simulation, such information provides us with a guideline for constructing coarse grained models.

In extracting slow collective motions, nonstationary features of the dynamics are also of interest. It is known that the dynamics of proteins involves a wide range of time scales [1]. This comes from hierarchical structures of energy landscapes with many local minima [19]. Wandering around such a rugged energy landscape, the protein changes its tertiary conformation. Then, slow collective motions would vary depending on where the system moves around on the landscape. Such time dependent features of collective motions will be captured only by developing methodology which can treat nonstationary time series. However, the methodology has not yet been fully developed which reveals nonstationary features of slow collective motions.

The purpose of our study is to present a new method to extract nonstationary features of coarse grained behavior from time series data of molecular dynamics simulation [20]. Our method consists of two steps: (1) the wavelet transformation and (2) the singular value decomposition (SVD). The wavelet transformation enables us to characterize time varying features in frequency components [21, 22, 23, 24, 25, 26, 27], and SVD enables us to reduce the degrees of freedom of the data [28]. Combining these two, we can extract nonstationary features of coarse grained behavior for proteins. Moreover, we introduce indexes to characterize collective motion of proteins. These indexes provide us with information of nonstationary deformation of protein structures. In this article, we will focus our attention to our methodology, thereby leaving detailed discussion of our application to our papers concerning Adenylate Kinase from *Escherichia coli* (AKE) [20] and *Thermomyces lanuginosa* lipase (TLL) [29].

HOW FUNCTIONAL MOTION IS EXHIBITED BY PROTEINS

Here, we give a brief explanation of the present ideas on how proteins exhibit their functions [3]. Concerning the relationship between conformational changes of proteins and their functions, there exists two ideas: (1) "induced-fit" and (2) "population shift". Suppose that the protein changes its conformation to a closed structure binding the ligand. According to the idea of "induced fit", it is supposed that ligand binding induces collective motions of the protein leading toward the closed structure. On the other hand, based on the idea of "population shift", it is proposed that the protein exhibits large conformational motions even without ligands binding, exploring those conformations near the closed structure. Ligand binding only shifts the population of these conformations in favor of closed structures. This ideas is also called the "conformational selection".

Recently, a single molecule experiment using fluorescence resonance energy transfer (FRET) reveals that a protein which works as an enzyme, Adenylate Kinase from *Escherichia coli* (AKE), actually explores those conformations near the closed structure even without ligands binding [13]. Their experiment has shown that ligand binding increases the population of those conformations near the closed structure. Their results indicate that the idea of "population shift" is relevant for AKE [30].

The "population shift" model implies that collective behavior exhibit transient features as the system exhibits different conformations. This leads us to realise importance

of time series analysis which reveals nonstationary aspects of the dynamics. In particular, it implies that functional motion of proteins can be extracted by analysing dynamical behavior of these proteins even without ligands. This is a basic strategy of our study toward understanding molecular function of proteins.

EXPLANATION OF OUR ANALYSIS

In this section, we explain the wavelet transformation and the singular value decomposition, the two components of our method. Then, we present an overview of our method to apply time series data of molecular dynamics simulation.

Wavelet Transformation

The wavelet transformation is regarded as a windowed Fourier transformation where the width of the window is adjusted according to the frequency. The transformation is suitable to analyse time series data whose frequency components vary as time goes on. It has been applied in various fields including time series analysis for vibrational motions of small molecules [22, 23, 24, 25, 26, 27]. There exists a variety of wavelet transformations depending on the choice of the window functions [21]. There also exist two types of wavelet transformations, i.e. continuous and discrete ones. Among them, we adopt the Morlet wavelet transformation, one of the continuous transformations. It is the simplest extension of the Fourier transformation, and can be regarded intuitively as a finite time Fourier transformation. This is the reason why we use the Morlet wavelet in our analysis.

For a given time series $f(t)$, the Morlet wavelet transformation $\hat{f}(t,\omega)$ is defined by

$$\hat{f}(t,\omega) \equiv \left(\frac{2\omega^2}{\sigma^2\pi^3}\right)^{\frac{1}{4}} \int_{-\infty}^{\infty} ds f(s) \exp\left(-i\omega(s-t) - \frac{\omega^2}{\sigma^2\pi^2}(s-t)^2\right), \qquad (1)$$

where t is time and ω is frequency. In Eq.(1), the width of the window is $2\pi\sigma/\omega$ and the period of the oscillation is $2\pi/\omega$. Therefore, σ gives the number of oscillations within the window. Thus, the width of the window changes according to the frequency. If the value of σ is too small, we have difficulty of precisely assigning the frequencies. If it is too large, the information concerning the transient features will be lost.

In actual calculation, the data is discrete and their number is finite. Therefore, we approximate the integral over the infinite interval using a sum of finite terms. This approximation introduces an artifact which is caused by the discontinuity between the values of $f(t)$ at the boundary points. This artifact affects the values of the wavelet transformation $\hat{f}(t,\omega)$ for t which lies within the range of $2\pi\sigma/\omega$ from the boundary.

Singular Value Decomposition(SVD)

In the following, the singular value decomposition (SVD) plays the role of reducing the number of degrees of freedom to represent a data. In general, a rectangular complex matrix A of N rows and M columns can be decomposed as follows

$$A = U\Sigma V^\dagger, \tag{2}$$

where U is a $N \times N$ unitary matrix, V is a $M \times M$ unitary matrix, and Σ is a diagonal matrix which has at most $K \equiv \min(N,M)$ non-zero diagonal elements, $s_1 \geq s_2 \geq \cdots \geq s_K \geq 0$. This decomposition is called the singular value decomposition (SVD). Denote the first K column vectors of U and V as $U = (\mathbf{u}_1, \cdots, \mathbf{u}_K, \cdots)$ and $V = (\mathbf{v}_1, \cdots, \mathbf{v}_K, \cdots)$ respectively. Then, the original matrix A is represented by

$$A = \sum_{k=1}^{K} s_k \mathbf{u}_k \mathbf{v}_k^\dagger. \tag{3}$$

Here, the multiplication $\mathbf{u}_k \mathbf{v}_k^\dagger$ indicates the tensor product $\mathbf{u}_k \otimes \mathbf{v}_k^*$ between the vector $\mathbf{u}_k$ and the vector $\mathbf{v}_k^*$ which is the complex conjugate of the vector $\mathbf{v}_k$. We call $\mathbf{u}_1, \cdots, \mathbf{u}_K$ the left singular vectors, $\mathbf{v}_1, \cdots, \mathbf{v}_K$ the right singular vectors, respectively. The non-zero diagonal elements of the matrix Σ, i.e. $s_1 \cdots s_K$, are called the singular values.

Note that the following equalities hold

$$A^\dagger A = V\Sigma^2 V^\dagger \tag{4}$$
$$AA^\dagger = U\Sigma^2 U^\dagger \tag{5}$$

because of Eq.(2). Thus, Eq.(4) is an eigenvalue decomposition of the matrix $A^\dagger A$, and Eq.(5) an eigenvalue decomposition of the matrix $AA^\dagger$. Therefore, the squares of the singular values $s_1^2, \cdots, s_K^2$ are the common eigenvalues of both $A^\dagger A$ and $AA^\dagger$, the column vectors of U are eigenvectors of $AA^\dagger$ and the column vectors of V are eigenvectors of $A^\dagger A$.

Suppose that the first $\bar{K}$ of the singular values are much larger than the rest of them. Then, Eq.(3) can be approximately written as

$$A \approx \sum_{k=1}^{\bar{K}} s_k \mathbf{u}_k \mathbf{v}_k^\dagger. \tag{6}$$

This means that the matrix A can be well represented by the reduced number of vectors $\mathbf{u}_k, \mathbf{v}_k \quad (k = 1, \cdots, \bar{K})$. Thus, SVD provides us with a method of reducing a given data to smaller degrees of freedom. In the following, we will use SVD for this purpose.

Overview of our method

Our analysis combine the wavelet transformation with the low-pass filter and SVD. In the following, we explain our method when we apply it to time series data of the

alpha carbons of the protein. First, we apply the wavelet transformation to each of the time series of the Cartesian coordinates of the alpha carbons, and retain lower frequency components of the wavelet transformation, i.e. we utilize the wavelet transformation with the low-pass filter. Then, for each of the times, singular value decomposition is applied to the matrix thus obtained. In the following, we explain our method for each of the steps.

- Wavelet Transformation with low-pass filter
  For a given times series $q_n(t)$ of the n-th degree of freedom with $n = 0, \cdots, N-1$, we apply the wavelet transformation to obtain $\hat{q}_n(t, \omega)$. In actual calculation, we apply the wavelet transformation to discrete time series $\{q_n(t_i)\}_i \quad (n = 0, \cdots, N-1)$ where i ranges from 0 to $M-1$, $t_i = i\delta t$ with δt the time step of the data. Then, we obtain the transformed data $\{\hat{q}_n(t_i, \omega_l)\}_{i,l}$ where both i and l range from 0 to $M-1$ and $\omega_l = \frac{2\pi l}{M\delta t}$. Note that, for real time series $\{q_n(t_i)\}_i$, $\hat{q}_n(t_i, \omega_{M-l})$ is the complex conjugate of $\hat{q}_n(t_i, \omega_l)$ for $l = 1, \cdots, M/2-1$.
  We expect that oscillations with lower frequencies exhibit collective behavior involving larger number of alpha carbons. Thus, we focus our attention to lower frequency components of the wavelet transformation, that is, $\hat{q}_n(t_i, \omega_l)_{i,l}$ and their complex conjugates $\hat{q}_n(t_i, \omega_{N-l})_{i,l}$ ranging from $l = M_1$ to $l = M_2$ with $0 << M_1 < M_2 << M/2-1$. Here, M_1 is chosen to avoid the artifact caused by the finiteness of the time series.
  For each of the time t_i, we construct the matrix $A(t_i) = \left\{A_{n,l}(t_i)\right\}_{n,l}$ where $A_{n,l}(t_i)$ equals to $\hat{q}_n(t_i, \omega_l)$ for $l = M_1, \cdots, M_2$ or $l = M-M_2, \cdots, M-M_1$ with $n = 0, \cdots, N-1$. Otherwise, $A_{n,l}(t_i)$ is set to be zero.
- Singular Value Decomposition
  Applying SVD to the matrix $A(t_i)$, we obtain the k-th singular value $s_k(t_i)$, the corresponding left singular vector $\mathbf{u}_k(t_i)$, and the right singular vector $\mathbf{v}_k(t_i)$, respectively. Note that the left singular vectors can be chosen to be real and that the $M-l$-th elements of the right singular vectors are the complex conjugates of their l-th elements. While the left singular vectors describe oscillations in space, the right singular vectors capture information concerning frequencies. The singular values indicate the amplitudes of these components. In our study, the number of singular values K is equal to $\min(N, 2(M_2 - M_1 + 1))$ since we apply the low pass filter to construct the matrix $A(t_i)$.

INDEXES CHARACTERIZING COLLECTIVE MOTION

In general, only a few singular vectors are sufficient for describing collective degrees of freedom for proteins [20][29]. After extracting those degrees of freedom describing collective motion of the protein, we characterize how the protein changes its conformation. In order to do it, we define indexes which quantify collectivity of the motion for those cases when the largest singular value is dominant [29].

We consider collective motion of the protein as a kind of motion when neighboring Cα atoms oscillate along similar directions. Then, we characterize collective motion

around the p-th Cα as follows. Note that the three-dimensional vector $\bar{\mathbf{u}}_p(t)$ is defined using the first left singular vector by $\mathbf{u}_{k=1}(t) = (\bar{\mathbf{u}}_1(t), \cdots, \bar{\mathbf{u}}_p(t), \cdots, \bar{\mathbf{u}}_N(t))$. We call $\bar{\mathbf{u}}_p(t)$ the oscillation vector of the p-th Cα 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 Cα atom can be taken either along the sequence of the protein or within its three-dimensional conformation. Thus, we can introduce four indexes $x_p^{(i)}(t)$ $(i = 1, \cdots, 4)$,

$$x_p^{(1)}(t) \equiv \frac{1}{2n-1} \left| \sum_{|p-q|<n} \bar{\mathbf{u}}_p(t) \cdot \bar{\mathbf{u}}_q(t) \right|, \tag{7}$$

$$x_p^{(2)}(t) \equiv \frac{1}{b_{p,r}(t)} \left| \sum_{|\mathbf{r}_p(t)-\mathbf{r}_q(t)|<r} \bar{\mathbf{u}}_p(t) \cdot \bar{\mathbf{u}}_q(t) \right|, \tag{8}$$

$$x_p^{(3)}(t) \equiv \frac{1}{2n-1} \left| \sum_{|p-q|<n} \frac{\bar{\mathbf{u}}_p(t) \cdot \bar{\mathbf{u}}_q(t)}{|\bar{\mathbf{u}}_p(t)|\,|\bar{\mathbf{u}}_q(t)|} \right|, \tag{9}$$

$$x_p^{(4)}(t) \equiv \frac{1}{b_{p,r}(t)} \left| \sum_{|\mathbf{r}_p(t)-\mathbf{r}_q(t)|<r} \frac{\bar{\mathbf{u}}_p(t) \cdot \bar{\mathbf{u}}_q(t)}{|\bar{\mathbf{u}}_p(t)|\,|\bar{\mathbf{u}}_q(t)|} \right|, \tag{10}$$

where n is the difference of Cα atoms from the p-th Cα atom along the primary structure, $\mathbf{r}_p(t)$ is the position of the p-th Cα atom in the three-dimensional space at time t, r is the distance from the p-th Cα atom in the three-dimensional space, and $b_{p,r}(t)$ is the number of Cα atoms within the distance r from p-th Cα atom at time t.

Our application of these indexes to AKE [20] and TLL [29] reveals that time dependence of the indexes characterize nonstationary features of conformational change for these proteins. For TLL, our analysis shows the following; First, time evolution of the collective motion involves not only the dynamics within a single potential well but takes place wandering around multiple conformations. Second, correlation of the collective motion between secondary structures shows that collective motion exists involving multiple secondary structures. These results indicate that time series analysis of molecular dynamics simulation is a fruitful approach for understanding dynamical behavior of proteins.

INTRINSICALLY DISORDERED PROTEINS (IDPS)

As a future target of our method, "intrinsically disordered proteins (IDPs)" are important subjects [31]. These proteins exhibit large conformational motion to the extent that their secondary structures are not fully determined. Moreover, such motion is supposed to play an important role in their functions. For example, "fly-catching" mechanism is proposed meaning that unfolding of a part of these proteins is efficient for searching for ligands which they bind [32]. Then, new methodology is necessary which is applicable

to large conformational motion which is exhibited by IDPs. We expect that our method becomes an important tool to understand dynamical behavior of IDPs since our method is applicable to such transient movement.

CONCLUSIONS

Here, we have explained our method to extract nonstationary features of slow collective motion toward time series data of molecular dynamics simulation for proteins. The method consists of the following two steps: (1) the wavelet transformation and (2) the singular value decomposition (SVD). The wavelet transformation enables us to characterize time varying features of oscillatory motions and SVD enables us to reduce the degrees of freedom of the movement. Moreover, we have introduced indexes to characterize collective motion of proteins. These indexes provide us with information of nonstationary deformation of protein structures. As a future prospect, we have discussed application of our method toward "intrinsically disordered proteins (IDPs)".

ACKNOWLEDGMENTS

This study has been done under collaboration with the following people, Ms. N. Sakurai, Ms. M. Kamada, Ms. S. Kimura, Dr. S. Fuchigami, Prof. A. Kidera, Prof. M. Sekijima, Dr. M. Takada and Prof. K. Joe. I would like thank them for fruitful outcome of our collaboration. This work has been supported by Priority Area "Molecular Theory for Real Systems", Grant-in-Aid for challenging Exploratory Research and Grant-in-Aid for Scientific Research (C) from the Ministry of Education, Culture, Sports, Science and Technology, the Cooperative Research Program of "Network Joint Research Center for Materials and Devices", Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures, and Nara Women's University Intramural Grant for Project Research.

REFERENCES

1. H. Frauenfelder, P. G. Wolynes, and R. H. Austin, *Rev. Mod. Phys.* **71**, S419 (1999).
2. D. M. Leitner, and J. E. Straub, editors, *Proteins : Energy, Heat and Signal Flow*, CRC Press, 2010.
3. S. Fuchigami, Y. Matsunaga, H. Fujisaki, and A. Kidera, *Adv. Chem. Phys.* **145**, 35–82 (2011).
4. M. Ikegami, J. Ueno, M. Sato, and A. Kidera, *Phys. Rev. Lett.* **94**, 178102 (2005).
5. M. Kubitzki, and B. De Groot, *Structure* **16**, 1175–1182 (2008).
6. T. Ichiye, and M. Karplus, *Proteins* **11**, 205 (1991).
7. A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, *Proteins* **17**, 412 (1993).
8. A. Kitao, S. Hauward, and N. Go, *Proteins* **33**, 496 (1998).
9. H. Lou, and R. L. Cukier, *J. Phys. Chem. B* **110**, 24121–24137 (2006).
10. K. Moritsugu, O. Miyashita, and A. Kidera, *J. Phys. Chem. B* **107**, 3309 (2003).
11. Y. Matsunaga, S. Fuchigami, and A. Kidera, *J. Chem. Phys.* **130**, 124104 (2009).
12. Y. Naritomi, and S. Fuchigami, *J. Chem. Phys.* **134**, 065101 (2011).
13. J. A. Hanson, K. Duderstadt, L. P. Watkins, S. Bhattacharyya, J. Brokaw, J. W. Chu, and H. Yang, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 18055 (2007).
14. E. Barkai, F. Brown, M. Orrit, and H. Yang, editors, *Theory and Evaluation of Single-Molecule Signals*, World Scientific, 2008.

15. O. Miyashita, J. Onuchic, and P. Wolynes, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12570 (2003).
16. N. A. Temiz, E. Meriovitch, and I. Bahar, *Proteins* **57**, 468 (2004).
17. P. Maragakis, and M. Karplus, *J. Mol. Biol.* **352**, 807–822 (2005).
18. C. Snow, G. Qi, and S. Hayward, *Proteins* **67**, 325 (2007).
19. A. Ansari, J. Berendzen, S. F. Bowne, H. Frauenfelder, T. B. Sauke, E. Shyamsunder, and R. D. Young, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 5000 (1985).
20. N. Sakurai, M. Toda, S. Fuchigami, and A. Kidera (in preparation).
21. I. Daubechies, *Ten Lectures on Wavelets*, Springer-Verlag, 1992.
22. L. V. Vela-Arevalo, and S. Wiggins, *Int. J. Bifurcation and Chaos* **11**, 1359 (2001).
23. C. Chandre, S. Wiggins, and T. Uzer, *Physica D* **181**, 171 (2003).
24. A. Shojiguchi, A. Baba, C.-B. Li, T. Komatsuzaki, and M. Toda, *Laser Physics* **17**, 1097 (2006).
25. A. Shojiguchi, C.-B. Li, T. Komatsuzaki, and M. Toda, *Phys. Rev. E* **75**, 035204(R) (2007).
26. A. Shojiguchi, C. B. Li, T. Komatsuzaki, and M. Toda, *Phys. Rev. E* **76**, 056205 (2007).
27. A. Shojiguchi, C. B. Li, T. Komatsuzaki, and M. Toda, *Phys. Rev. E* **77**, 019902(E) (2007).
28. G. H. Golub, and C. F. V. Loan, *Matrix computations, 3rd edition*, John Hopkins, 1996.
29. M. Kamada, M. Toda, M. Sekijima, M. Tamada, and K. Joe, *Chem. Phys. Lett.* **502**, 241 (2011).
30. K. Arora, and C. L. Brooks, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 18496 (2007).
31. A. Dunker, J. Lawson, C. Brown, R. Williams, P. Romero, J. Oh, C. Oldfield, A. Campen, C. Ratliff, K. Hipps, et al., *Journal of Molecular Graphics and Modelling* **19**, 26–59 (2001).
32. E. Trizac, Y. Levy, and P. G. Wolynes, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 2746–2750 (2010).