Vol.2016-BIO-48 No.2
2016/12/9

Extraction of Structural Transition Events from Time Seriese Data using Robust Singular Spectrum Transform

Hitoshi AFUSO[1,a)] Ayako OSHIRO[1] Takeo OKAZAKI[2]
Morikazu NAKAMURA[2]

Abstract: To obtaine some knowledge about a protein's function or characteristics, Molecular Dynamics(MD) simulation technique was utilized. Time seriese data representing motions of a protein contains not only essential dynamics but also random motions corresponding to thermal fluctuation of focal system. To solve the problem to extract essential motions we had proposed a method named "GrabRPM". The method allows us to extract and summarize complexed motions contained in a time seriese data from MD simulation. However the part for the extraction of structural transition events depends on two thresholds corresponding the minimal difference of Root Mean Squared Distance(RMSD) between two structures of focal protein and the time period that the effect of structural transition events last. In this study we applied the method to extract change point from time seriese data, named "Robust Singular Spectrum Transform"(RSST)**?** and utilized preliminary experiments using artificial data to show some property of RSST. From the experiments, although RSST could extract the change in the mean value, it has the difficulty for the extraction of the change in the variance.

1. Introduction

Improvements on hardware and software allow us to simulate various biological processes in computer. Simulations about various biological process, such as protein folding and ligand binding were archieved.To execute the simulation of a biological process Molecular Dynamics(MD) method and Monte had been used. As examples of softwares to utilize MD simulation, we can cite CHARMM[1], AMBER[2] and GROMACS[3]. Time seriese data of coordinates of atoms consist of a protein is called "Trajectory". Analysing a trajectory, the information about the focal protein's structure and motions in a system could be obtained. Because structure and its motions are related to the biological function of a protein, obtaining information of them has possibility to archieve the more accurate knowledge about various biological processes.

Instead of the its effectiveness a trajectory contains not only essential motions or dynamics that relate to the functions but also ones that originated from thermal fluctuations in a system. To obtain the information about the structure or motions related to the function it is needed to separate essential dynamics from the random motions. To solve the problem we had proposed a method that consists of density-based clustering[8] and graph representation named "GrabRPM"[7]. Some method to choose the parameters used in the proposal were also developped[10]. In the report[10] a summarization result was shown from the MD trajectory corresponding model protein, HP35. However the extraction of structural transition events depends on two thresholds corresponding the minimal difference of Root Mean Squared Distance(RMSD) between two structures of focal protein and the time period that the effect of structural transition events last. Such thresholds value might be varied depending on focal protein and biological process. Then it is required for a more flexible method to extract the time points structural transitions occured.

In this report we applied the method to extract change point for time seriese data, named "Robust Singular Spectrum Transform" (RSST)**?** and utilized preliminary experiments using artificial data to show some property of RSST.

2. Extraction of Structural Transition

In this section, the explanations about the method for extraction of structural transitions used in previous report[10]. After that we give the introduction of the method, named "Robust Singular Spectrum Transform"(RSST) to extract change points from time seriese data.

2.1 Extraction Method used in the Report[10]

Considering the noise level of given dataset and using

[1] Center for Clinical Research and Quality Management, Ryukyu University Hospital
[2] Department of Engineering, University of the Ryukyus
[a)] h149013@med.u-ryukyu.ac.jp

the optimal Traditional method to extraction of structural transition events were based on a graph-based representation of a protein motions. The representation had been proposed in the report[7] and called "Structural Transition Graph"(STG). Because of that STG contains the information of structural transitions in a trajectory, the analysis of the characteristic sub-structure of STG could allow us the useful information about structural transitions. Considering the fact that a trajectory is a time-series data, Two major cases about about the structure of STG could be assumed. The first one is the structure that is globally linear and contains some locally dense components. And the second is global ring structure and contains some dense local regions. In both cases, structure of STG contains locally dense regions as sub-structure. Global structure of STG corresponds to two major outcomes of biological process. The first outcomes is the case that the end structure was same to the the one at the starting time point in a simulation. And the second is the one that structures at starting time point and end were different such as the simulation of protein folding process. We made a hypothesis that local dense region represents some significant structural transitions such as the resolving a sub-structure that is a energetic barrier on a energy landscape.

According to the method to extract the structural transition events from the time region in a trajectory corresponding to the vertices in local dense region had benn developped. It was based on two assumptions. The first is that in the time region that particular structural transition events occur, the $RMSD$ value of focused protein structure according to a corresponding trajectory. And the second, after the increasing $RMSD$ at certain time point if the variation of $RMSD$ fall into some range then it could be considered that the structure that occured in previous structural transition event might be kept in some time period. The algorithm to detect the time points that structural transition might occure is shonw in Fig.1. Using algorithm described above one can extract the occurence point of that structural transitional events and last time of its effects.

2.2 Outline of Robust Singular Spectrum Transform

The extraction method shown in Fig.1 uses two thresholds that represents some property of focal protein and biological process. Then these values might be varied depends on a protein and biological process. From this fact the more flexible method to extract the time points that structural transition events is required.

As such method we can site Robust Singular Spectrum Transform(RSST)**?**. RSST calculates some score for each time points that represents the likelihood that the time point is a change point. The steps of RSST are shown in Fig.**??** and its pseudo code is shown in Fig.**??**. the main concepts of RSST is the comparison the current state and the past changes in the past. Firstly Hankel matrix is calculated at time point t before or after by sliding window manner using

```
1: procedure ExtractEventTime(Focusing time region T, RMSD function in time region T R(t), Event occurence threshold r, Structure remaining threshold l)
2:     eventTimingSet ← φ
3:     for t ← T do
4:         if R(t) > r then
5:             eventTimingSet Leftarrowt
6:         end if
7:     end for
8:     eventRegionSet ← φ
9:     for t ←eventTimingSet do
10:        eventRegionSet ← t
11:        for u ← {t + 1, t + 2, ···,, NextTiming(t)} do
12:            if R(u) < l then
13:                eventRegionSet ← u
14:            else
15:                break
16:            end if
17:        end for
18:    end for
19:    return eventRegionSet
20: end procedure
```

Fig. 1 Pseudo code of the algorithm for structural transition event detection in the report[10]

the window that its width w. The number of windows(subsequence) contained in Hankel matrix is denoted as n. Hankel matrix at time point t corresponding to the past changes is denoted as $H(t)$ and the one corresponding the current state is denoted as $G(t)$.

$$H(t) = (seq(t-n), \ldots, seq(t-1))^T \quad (1)$$

$$G(t) = (seq(t), \ldots, seq(t+n-1))^T \quad (2)$$

In Eq.2 $seq(t)$ represents a sub-sequence starting at time point t that its length is w. After the calculation of the matrices current change state and the changes in the past are captured using the singular value decomposition and eigen value decomposition of them.

$$H(t) = U(t)S(t)V(t)^T \quad (3)$$

$$G(t)G(t)^T v_i = \lambda_i v_i \quad (4)$$

where $\lambda_{i+1} \leq \lambda_i \leq \lambda_{i-1}$. The singular values in $S(t)$ and corresponding vectors in $U(t)$ are sorted with increasing order same as λ_i. After the decomposition to remove the components that originated by random fluctuation the vectors corresponding small singular values or eigen values are filtered out. The number of vectors after the filtering is denoted as l for $H(t)$ and m for $G(t)$. To measure the significance of the vectors the singular values are accumulated. The points that tangent to the curve of accumulation is greater than $\frac{\pi}{4}$ is determined as the threshold to filter the vectors originated by random fluctuations out. For parameter m same steps are used to determine its value. For comparison between the current state and past changes the projection of the vector v_i onto the hyper plane U_l and normalization are utilized.

$$\alpha(t) = \frac{U_l v_i}{||U_l v_i||} \quad (5)$$

where $i \in \{1, \ldots, m\}$. If there is no change of the dynamics in the time seriese the vector v_i representing current state will be near on the hyper plane represented by U_l. In other words the vectors $\alpha(t)$ and v_i are very similar in such case. To measure the difference the cosine of the angle between them is calculated.

$$cs_i(t) = 1 - \alpha(t)v_i \quad (6)$$

After the calculatin of $cs(t)_i$ for each significant vector v_i the weighted sum of them is calculated as change score using λ_i as weight values.

$$cs(t) = \frac{\sum_i^m \lambda_i \times cs_i(t)}{\sum_i^m \lambda_i} \quad (7)$$

In addition RSST executes one more filtering using the information about the static state of time seriese around the time point t.

$$\hat{cs(t)} = cs(t) \times |\mu_f(t) - \mu_p(t)| \times |\sigma_f(t) - \sigma_p(t)| \quad (8)$$

where $\mu_f(t)$ and $\mu_p(t)$ represent the mean values in the past and the future at time point t. $\sigma_f(t)$ and $\sigma_p(t)$ are standard deviation in the past and the future.

Follwing the above steps we can calculate the change score that represents the likelihood that the time point is change point.

3. Performance of RSST in Parameter Variation

In RSST we have to set two parameters, the width of window w and the number of the windows n for the construction of Hankel matrix. To check the performance of RSST according to the variation of these parameters, we utilized experiments using artificial time seriese data.

3.1 Experimental Setup

In the experiments we considered two major cases of the changes in time seriese, mean shift and the change in variance. For mean shift a time seriese that its length = 1,000 was generated by concatenating 5 sub-sequences. The length of each sub-sequence is 200 and the mean value was varied in 0.0, -1.0, 2.0, 0.5 and 0.0. Each sub-sequence has same variance value $\sigma^2 = 0.2$. For variance change a time seriese data that its length = 1,000 was generated in same manner with mean shift. Each sub-sequence has length 200 and the variance values were 0.1, 0.2, 0.4, 0.8, and 1.0. The mean value was set to 0.0. RSST was utilized to the aritificial data by setting parameter w to 5, 10, 20, 50 and 100 and parameter n to 10, 20, 50 and 100.

3.2 Result and Discussion

The results for the data with mean shift are shown in Fig.2. As shown in Fig.2, in the cases which relatively small value of n RSST could capture the change points corresponding to the shifts of mean more accurately. On the other hand the results in the cases which smaller window size was tend to have much more false positive change points in the region where mean value = 0.0. Such region contained only random variation (noise) the result of RSST could be noisy in the region where essential signal (or dynamics) is small. In addition larger window size had tend to be relatively smoothen curve of change scores. The window size reflect the sub-sequence space in which the essential dynamics of time seriese has to be contained. From this fact it is implied that larger window size correspond to so large sub-sequence space and this lead to the situation that sub-sequences in the space are distributed sparsely. Then the difference between two hyper plane spanned by sub-sequences adjacent in the time seriese data could be relatively large.

As shown in Fig.3, RSST could capture only the change point correctly between $\sigma = 0.1$ to 0.2 and $\sigma = 0.2$ to 0.4. In addition there were much false positive change points in the region that the variance value is higher than 0.4. And in the case $w = 5$ much false positive points were detected in whole region. These results correspond to previous ones in the case of mean shift. Because of the settings parameter $\mu = 0$ in variance change data RSST results became noisy. And it is implied from these results that the results of RSST degradate in the region that time seriese data has high variance.

From these results it is implied that RSST could capture mean shift events more effectively but its results are easily degradated by increasing the variance contained in focal data.

4. Conclusion

In this report we utilized RSST to artificial data to check its performance according to the change of the static states μ and σ of time seriese data. Although RSST is more flexible than the method that we had proposed before, its results are easily degradated by the increasing the variance. As next task the development of the method that could handle the variance change more robustly and parameter decision rules.

References

[1] B. R. Brooks, C. L. Brooks III, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, (2009), "CHARMM: The Biomolecular simulation Program", *J. Comp. Chem*, Vol.30

[2] D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossvry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu and P.A. Kollman (2014), AMBER 14, University of California, San Francisco.

[3] Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005). "GROMACS: fast, flexible, and free". *J Comput Chem*, Vol.26(16), pp.1701-1718

[4] Balsera, M.A., Wriggers, W., Oono, Y. and Shulten, K., (1996) Principal component analysis and long time protein dynamics, J. Phys. Chem., Vol.100, pp.2567-2572

[5] Lei, H., Su, Y., Jin, L. and Duan, Y., (2010) Folding network of Villin headpiece subdomain, Biophysical Journal, Vol.99, pp.3374-3384

[6] Pilar Cossio, Alessandro Laio and Fabio Pietrucci, (2011), "Which Similarity Measure is Better for Analyzing Pro-

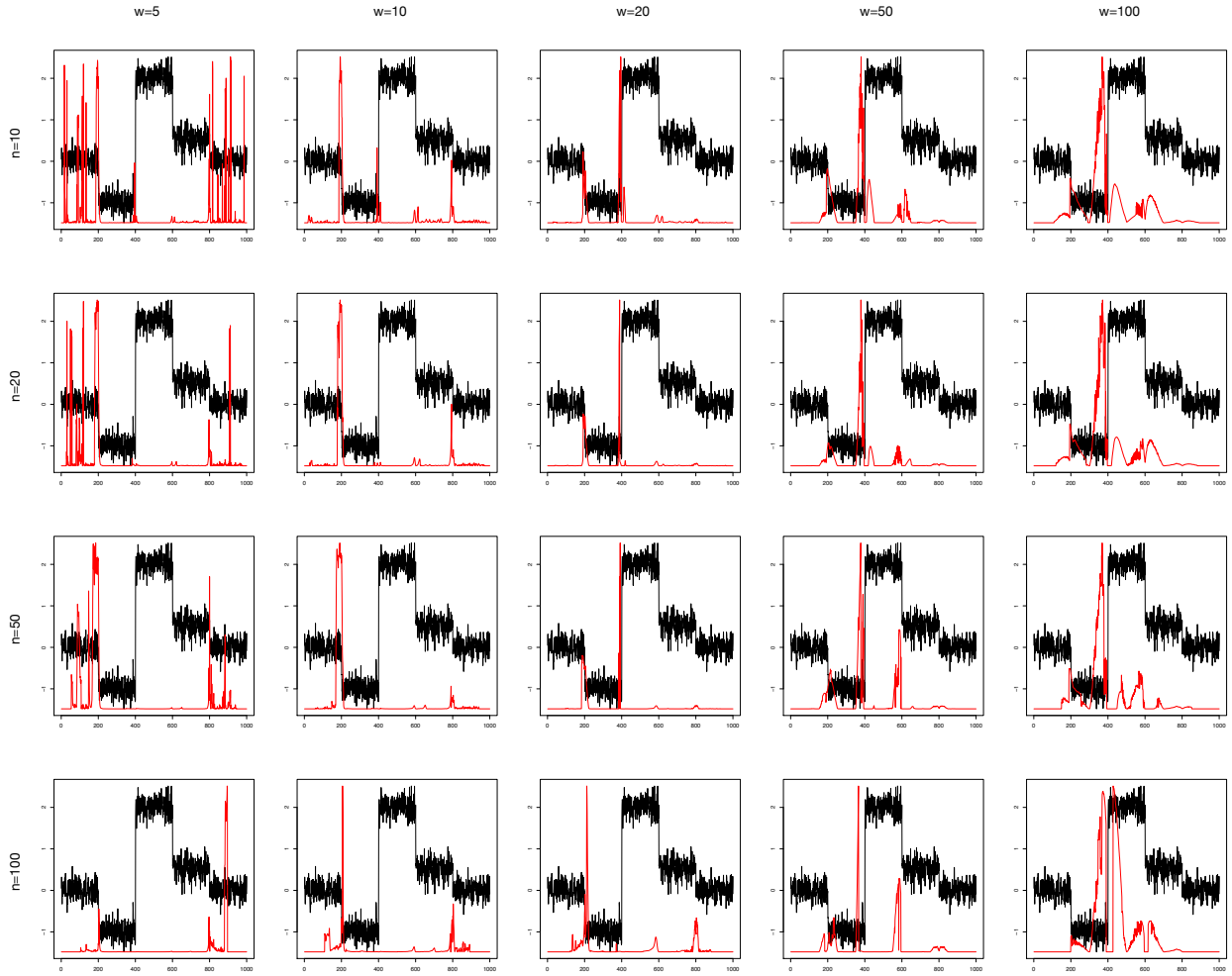

Fig. 2 The results obtained from mean shift data. The black line represents the mean shift data and red one denotes the corresponding change scores at each time point.

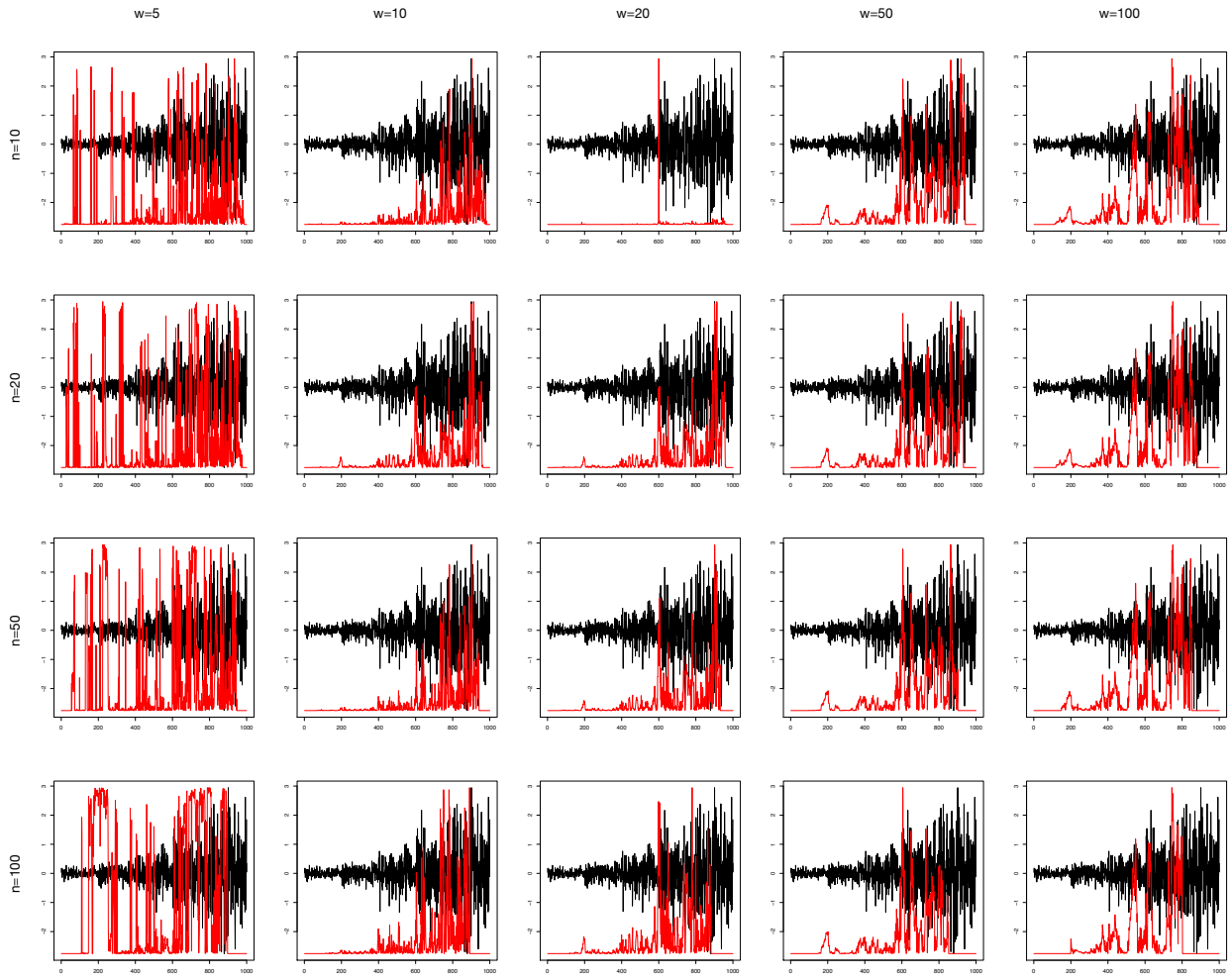

Fig. 3 The results obtained from variance change data. The black line represents the original data and red one denotes the corresponding change scores at each time point.

tein Structures in a Molecular Dynamics Trajectory?", *Phys. Chem. Chem. Phys*, Vol.13, pp.10421-10425
[7] Afuso, H, Mineta, K, and Endo, T, (2013), "Graph based Representation of Nonlinear Motions of Proteins", SIG-BIO Technical Report, 2013-BIO-36, pp.1-7
[8] Ankerst, M., Breuning, M.M., Kriegel, H-P. and Sander, J. (1999) OPTICS: Ordering Points To Identify the Clustering Structure, ACM SIGMOD International Conference on Management of Data, ACM Press, pp.4960
[9] Sander, J., Zin, X., Lu, Z., Niu, N. and Kovorsky, A., (2003) Automatic extraction of clusters from hierarchical clustering representations, Pro-ceeding PAKDD '03, Proceedings of the 7th Pacific-Asia conference on Advances in knowledge discovery and data mining, pp.75-87
[10] Afuso, H. and Endo, T., "Optimal Parameter Selection for Construction of Motion Representation Graphs and Extraction of Motion Transitional Events", SIG-BIO Technical Report, 2015-BIO-48, pp.1-8
[11] Ensign, DL., Kasson, P.M. and Pande, V.S., (2007) Heterogeneity egen at the speed limit of folding: large-scale moleculer dynamics study of a fast-folding variant of the Villin headpiece, J. Mol. Biol, Vol.384, pp.806-816
[12] Mowshowitz, A., and Dehmer, M., (2012) Entropy and Compresity of Graph Revisited, *Entropy*, Jan, Vol.17, pp.1-11
[13] Shetty, J. and Adibi, J. (2005) Discovering Important Nodes through Graph Entorpy The Case of Enron Email Database, *KDD'2005*, Illinois
[14] Constantine, G. (1990) Graph Complexity and the Laplacian Matrix in Blocked Experiments. *Linear and Multilinear Algebra*, Vol.28, pp.49-56
[15] Bonchev, D. (1995) Kolmogorov's Information, Shannon's Entropy and Topologcal Complexity of Molecules, *Bulg. Chem. Commun*, Vol.28, pp.567-582
[16] Shannon, C.E. and Weaver, W., (1949) 'The Mathematical Theory of Communication, Univesity of Illinois Press
[17] Humphrey, W., Dalke, A. and Schulten, K. (1996(VMD - Visual Molecular Dynamics", *J.Molec.Graphics*, Vol.14, pp.33-38.