Optimal Parameter Selection for Construction of Motion Representation Graphs and Extraction of Motion Transitional Events

Hitoshi AFUSO[†1,1,a)] Toshinori ENDO[2]

Abstract: Improvements on both hardware and software allows us simulations of various biological process on computer. On the other hand, such simulation results contains not only essential dynamics for protein function, but also random motions originated by thermal fluctuation in the system. The task that distinguish or extract only essential motions from such mixtured results remains as challenging. So far, authors proposed a method, called "GrabRPM", to describe the motions of protein. Using the method, it becomes possible to extract the non-linear motions , such as partial rotations in a protein, that is difficult to be handled with traditional method. However, previous experiments contains the difficulty such as arbitrary property in the step of graph construction. In addition, the extraction of the structure transitional events, that is important to understand the protein-folding phenomena, had not been discussed enough. In this study, a method to choose optimal parameters for constructions of structure transition graphs. In addition, based on the sub-structure contained in structure transition graphs, the extraction of occuring points of sub-event in protein folding was archieved.

1. Background

Improvements on hardware and software allow us to simulate various biological processes in computer. So far, simulations about protein folding, ligand binding, accumlation of proteins and their behavior in solution were archieved. Such computer simulations give us the another way to analyse the biological knowledge that obtained with pure-biological experiments. In addition, computer simulations have another advantage that it can utilize experiments that are difficult to control the experimental parameters in pure-biological experiments. From this point of view, computater simulations and biological experiments are complement, each other. Using both methods effectively, it is expected that more detailed biological knowledge could be obtained. By using detailed biological knowledge, the effective treatments of number of difficult disease such as Alzheimer's disease, might be realized.

For computater simulations of biological process, Molecular Dynamics(MD) method and Monte Carlo-based method had been used. As examples of softwares that used MD, we can cite CHARMM[1], AMBER[2] and GROMACS[3]. And also, to improve the efficiecy of MD simulations, more sophisficated algorithm, such as Replica-Exchanging Molecular Dynamics(REMD) had been developed.

In MD simulations, time seriese data of coordinates of atoms consist of a protein is called a "Trajectory". Analysing the changes of coordinates in a trajectory, the motions of focal protein could be analysed. On the other hand, a trajectory contains not only the motions that relate to the function of focal protein, but also ones that occured by thermal fluctuations in the system. From this fact, it is requred to extract only the motions essential for the function of the protein. As a traditional method to tackle this problem, we can cite Essensial Dynamics Analysis(EDA)[4] based on Principal Component Analysis(PCA). As another example, the method using PCA based on random matrix theory have been proposed. However, because of the fact that they are both linear method, it becomes difficult to handle non-linear motions, such as partial rotation of a protein. Description of such non-linear motions is important for understanding of the function of a protein because it might relate to the formation of sub-structure in a protein, such as α-helix that is essential for the function. To solve that problem, some methods for motion extraction based on manifold learning have been developed[5][6]. In despite of their success, they need some assumptions for dataset such as the sampling is enough to approximate the data manifold. Then, in the case that condition is not satisfied the methods don't work well. Such situation could be led by sub-sampling from huge trajectory or drastic motions of

[1] Meme Media Laboratory, Hokkaido University
[2] Graduate School of Information Science and Technology, Hokkaido Univesrity
[†1] Presently with Graduate School of Medical Science, University of the Ryukyus
[a)] afuso@ibio.jp

simulated protein. For more flexible representation of protein motions, some methods based on graph representation have been proposed[7]. As one extension of that approach, authors developed the method that consists of density-based clustering[11] and graph representation [9]. In the proposal, it was shown that obtained graph structure, called Structure Transition Graph(STG), had consistency with traditional results. In addition, the hypothesis that strongly-connected components in a STG might give some information about protein folding was suggested. However, three points still remain as subjects. First is the relationship between parameters used in the method and its result. Second is the settings for parameters to archieve suitable result. And finally, the extraction of motion transition events was not argued about.

In this report, we utilized three studies corresponding above subjects. First, to show the effect of parameter to the structure of obtained STG, some experiments were executed. Second, the method to select more suitable parameter values was proposed. And third, the method to extract motion transition events was designed based on the information about sub-structure contained STG. In addition to those proposals, the experiments to show the validity and applicability of proposals were conducted.

2. Relation between Parameters and Structure of STG

In this section, the explanations about the relation between parameter variations and resulted STGs would be given. Before the explanation, the outline of previously proposed mehod, GrabRPM, would be shown. After that, the experiments to show the effect of parameter variations for the structure of obtained STG was conducted.

2.1 Outline of GrabRPM

GrabRPM[9], that was proposed by authors is the method to construct the graphs that represent protein motions contained in given trajectory. The method consists of three steps below:

(1) Using OPTICS algorithm[11], assign the order and *ReachabilityDistance* to the structure contained in a trajectory

(2) Applying cluster-tree algorithm[12], construct the clusters of structures in a trajectory.

(3) For obtained clusters, construct a graph considering obtained clusters as vertices and temporal adjacency between clusters as edges.

We give short explanation about the OPTICS algorithm used in Step.(1) above. OPTICS(Ordering Points To Identify Cluster Structure) assigns the order to each data, based on two values, $CoreDistance$ and *ReachabilityDistance*. The definitions of those values are shown in Exp.(1) and Exp.(2).

$$CoreDistance_{\epsilon,M}(p) \qquad (1)$$
$$= \begin{cases} \text{Undefined} & |N_\epsilon(p)| < M \\ d(p, q_M) & \text{Otherwise} \end{cases}$$

$$ReachabilityDistance_{\epsilon,M}(p,o) \qquad (2)$$
$$= \begin{cases} \text{Undefined} & |N_\epsilon(o)| < M \\ \max(CoreDistance(o), d(o,p)) & \text{Otherwise} \end{cases}$$

In Exp.(1) and Exp.(2), $N_\epsilon(p)$ denotes the ϵ-neighbor of point p, $d(p,q)$ is a distance between p and q. q_M denotes the M-th neighbor of p. M represents minimal number of niearest neighbors. If the number of ϵ-neighbor is lesser than M, then such point p is considered as noise point. If p is noise point, order and the value of *ReachabilityDisntace* are not assigned. OPTICS algorithm iterates the assignments of order and *ReachabilityDistance* obtained from minimum number of nearests M and nearest-radius ϵ. Major different point between OPTICS algorithm and other clustering algorithm, such as Ward method or k-means method is that OPTICS algorithm uses the information about the point density around focal sample point directly. In a trajectory from MD simulation, the point density around focal conformation of the protein corresponds to the free-energy surface on the conformation space. In other words, a protein in a simulation tends to be conformations in the region that denotes low free-energy. Then, The point density approximates free-energy surface of corresponding protein. From above consideration, it is expected that OPTICS algorithm could reflect more accurately than other clustering methods. That is the reason why authors used density-based algorithm in the proposal of GrabRPM. To calculate $CoreDistance$ and *ReachabilityDistance*, some method to measure the distance(dissimilarity) between the structure of a protein. As standard of such methods, we can site Root Means Square Deviation($RMSD$) and Distance Matrix Error(DME)[8]. $RMSD$ focuses the distance between corresponding atoms in two different conformations of a protein. On the other hand, DME measures the similarity of the distances intra structure of a protein. Suppose that $RMSD(a,b)$ and $DME(a,b)$ denotes $RMSD$ and DME between structure a and b. The formulas of each measure are shown in Exp.(3) and Exp.(4)

$$RMSD(a,b) = \sqrt{\frac{\sum_{i=1}^{N}(x_i^{(a)} - x_i^{(b)})^2}{N}} \qquad (3)$$

$$DME(a,b) = \frac{2}{N(N-1)}\sqrt{\sum_{i>j}(d_{ij}^{(a)} - d_{ij}^{(b)})^2} \qquad (4)$$

Where N denotes the number of the atoms in a protein and $d_{ij}^{(a)}$ is the distance between i-th and j-th atom contained in structure a. $RMSD$ is varied by rotation and translation of focal structure. Since it does NOT satisfy the triangle inequality that one of the axiom of mathematical distance, $RMSD$ itself is not distance strictly. To make it strict distance, alignment of the structures is required. DME doesn't need such operation. However, it could not distingush two structures if they are mirrored image each other.

OPTICS algorithm outputs the *ReachabilityDistance* and order of each point, not clusters. The plot that its x-axis denotes the order of points and y-axis represents *ReachabilityDistance* of corresponding point is called as "Reachability Plot". In the result of OPTICS algorithm, points that likely form a cluster have small *ReachabilityDistance* values. In opposite, the assignment of order to two points belonging to different clusters results larger *ReachabilityDistance*. Cluster-Tree algorithm[12] extract the tree structure that represents cluster relationships in the dataset. Such tree structure is called as Cluster Tree. Cutting the cluster tree at specified height, one can obtain a clustering of the dataset.

As clustering the structures of the protein, the small difference that originated from thermal fluctuation of the system could be removed. After the operation, considering one cluster as a vertex and time adjacency between two sctructures belonging different clusters as a edge, one directed graph could be constructed. It is expected that constructed graph represents the information about the structural transition of the protein in a trajectory. Authors named that directed graph as "Structure Transition Graph(STG)" and the method to construct STG as "GrabRPM(Graph based Representation of Protein Motion)"[9]. Using GrabRPM, the information about the structural transition in a trajectory could be visualized as a graph diagram. In addition, as visualizing the structural change corresponding certain edge using some software such like UCFS CHIMERA[10], we could see the occurence of structural transitions in a trajectory. GrabRPM could handle the partial rotation of a protein that is difficult to capture with traditional methods. Furthermore, it could be applied in the case that the sampling is not enough to approximate data manifold or trajectory contains drastic structure changes by adjusting two parameters, neighborhood radius ϵ and the minumum number of the neighbors M.

2.2 Parameter Variation Experiment

Using GrabRPM, one can construct a representation of the structural transitions contained a trajectory. However, its result depends on the settings of parameters, ϵ and M in the clustering step. In this report, we conducted the experiments to clarify the effect of variations of parameters to the structure of obtained STG.

In the experiments, the dataset published by Ensign[14] was used. It contains 9 trajectories obtained from MD simulations about a protein, Villin headpiece subdomain of chicken. This protein is known as its high speed of folding and used various studies of protein folding as a model[13]. Each trajectory contained dataset has different initial structure state. They are assigned labels from Run0 to Run8, respectively. In this experiments, the trajectory Run0 was used. For more detailed information about the MD simulations and initial states, refer the original paper[14]. We used grid search like scheme for the variation of the parameters. Concretely says, we varied neighborhood radius ϵ from 2.2 to 3.2 by 0.2 and the minimum number of the nighbors M from 2 to 8 by 2. For each case, GramRPM was applied. The application results are shown in Fig.1. In Fig.1, the number put on each vertex denotes the label of clsuter obtained by OPTICS. As shown in Fig.1, the increasing the neighborhood radius ϵ led to the greater number of the edges in STG. In other words, the variation of the neighborhood radius controlled the structural complexity of resulted STGs. On the other hand, the increasing the minimum number of the neighbors M resulted the less number of the vertices in STGs. This result suggests that the minimum number of neighbors controlls the size of STGs. The greater value of M might lead to the more larger number of the noise points. In other words, by controlling the value of M, we could consider the noise level that contained in given dataset.

In the results the neighborood radius ϵ was set to relatively small, the structure of STGs tended to be nearly linear structure. In opposite, as increasing the radius of neighborhood, the connectivity among the vertices became higher and eventually almost all vertices were reachable each other. In addition, it was observed that the speed of increasing the structural complexity got higher exponentially.

From above results, we concluded that the neighborhood radius ϵ and the minimum number of neighbors M relate to the structural complexity and the size of resulted STGs, respectively.

3. Optimal Parameter Selection Method

From the results of previous experiments, the effects of two parameters were suggested. And also, it was shown that the structural complexity of STGs seems to be controlls by those parameters.

In the case the neighborhood radius is set to small, obtained STGs tends to be nearly linear structure. Such structure might simply represents the time adjacency in a trajectory. From this reason, such STGs might not reflect the characteristic information about the structural transitions. In despite of that, it is difficult to find the where the characteristic transitions occured from too complicated STGs resulted with greater value of ϵ. Although setting the minimum number of the neighbors M to small could result more shrinked STGs, the increasing the neighborhood radius ϵ could lead complicated STGs.

From these results, for the dataset that likely contains more noises, one can adjust the minimum number of the neighbors M. Then essential problem in the selection of parameter settings is how we set the appropriate neighborhood radius ϵ when the minimum number of the neighbors M is given corresponding noise level in dataset. Summarizing above discussion, the method to choose the optimal neighborhood radius ϵ considering the structural complexity of resulted STGs is required.

To tackle this problem, we considered that the measurement for the structural complexity is needed. As such measurement, some methods, called Graph Entropy have been

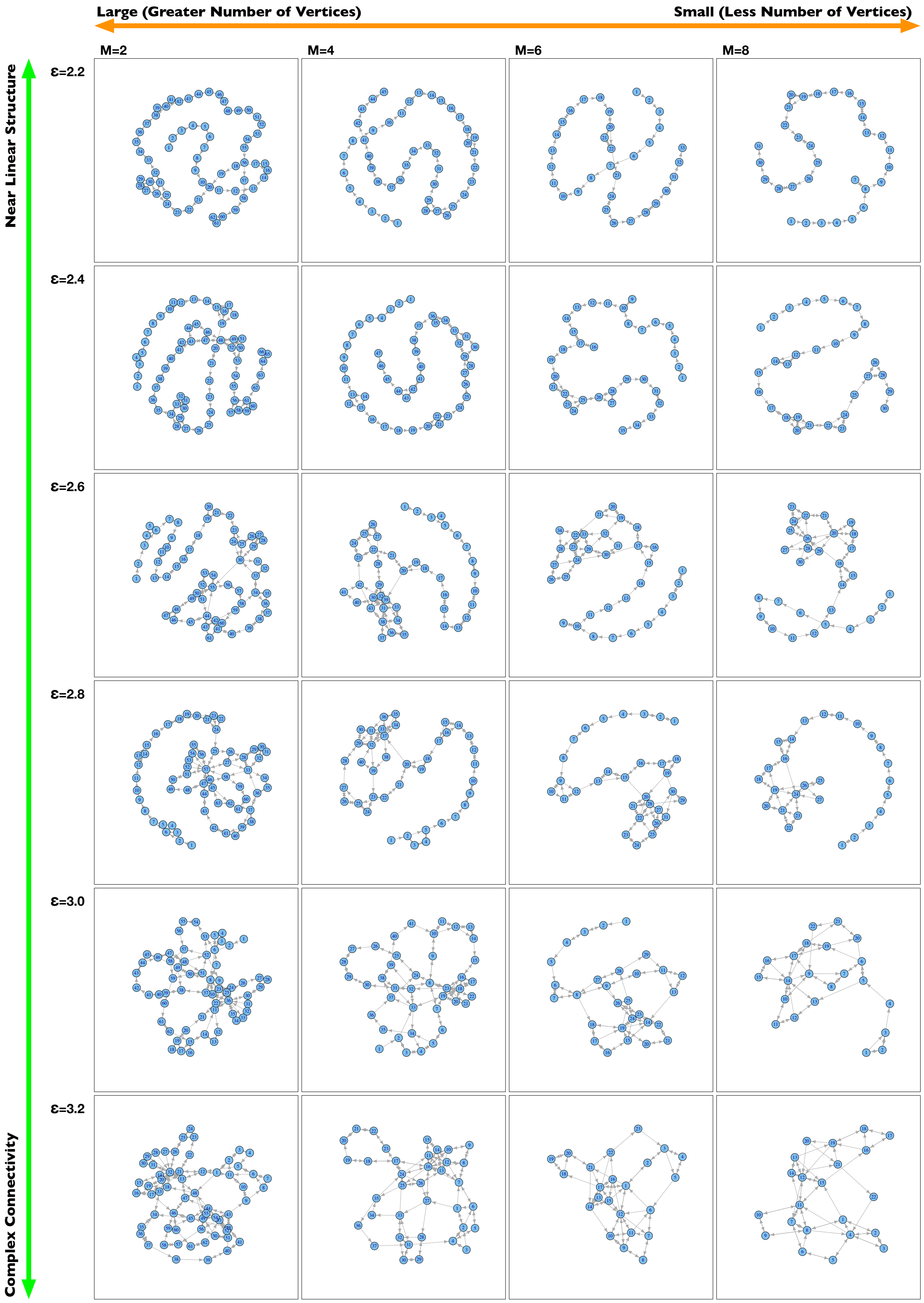

Fig. 1 STGs obtained from Run0 with variations of two parameters: Vertical and horizontal axis denote neighborhood radius ϵ and the minimum number of the neighbors M, respectively. As increasing the value of ϵ, the number of edges contained by STG was increased. As a consequence, the structural complexity of STG became greater. On the other hand, the increasing of the minimum number of neighbors M resulted the less number of the vertices contained in STG.

developed[15]

Next, we give rough explaination about this measurement.

3.1 Graph Entropy

So far, various definitions of graph entropy have been proposed and they are roughly devided to two categories, deterministics approach and probabilistic approach[15]. As examples of deterministic approach, we can cite the Kolmogolov complexity based on graph encoding[16], the method based on the frequency of certain sub-structure[17], the method based on the number of operations required for transformation from certain initial graph to target graph[18]. Probabilistic approach is further devided into two sub-categories, intrinsic and extrinsic. These sub-categories both consider the probability distribution for certain sub-graphs. And also they are both based on the entropy function[19] proposed by Shannon. The different point between them is the way to define the probability functions. Intrinsic approach obtains the probability function by using the information about characteristics on sub-graphs. On the other hand, in extrinsic approach, the probability function is constructed from the prior information about the sub-graphs. In this report, considering that the prior information about sub-graphs is not given in advance and also particular initial graph structure is not clear, we thought that intrinsic probabilistic approach is suitable for our situation.

For intrinsic probabilistic approach, there are various methods according to the focusing characteristics**?**. In this report, the graph entropy based on degree that is one of the most basic characteristics.

In Exp.(5), the definition of the graph entropy $I(G)$ corresponding graph G is shown.

$$I(G) = \sum \frac{|\delta_i|}{|V|} \log \left(\frac{|\delta_i|}{|V|} \right) \tag{5}$$

Where, $|V|$ denotes the number of vertices in G. $|\delta_i|$ represents the number of vertices that its degree equals i. The sum is over the set of vertices in G. Using Exp.(5), one can measure the information amount of the structural complexity. Next, we design the method to choose optimal values of parameters using graph entropy.

3.2 Optimal Parameter Selection

To choose the optimal parameter value, we need to define the set of candidate values of parameters. In this report, same to previous experiment, we applied the grid search like scheme to generate the candidate value set. As shown in the parameter variation experiment, as increasing the neighborhood radius, the speed of increasing structural complexity get higher exponentially. From this result, the speed of changing of graph entropy is not constant and it is expected that the speed becomes faster at certain point on parameter variations. And if structure of STG changes to more simpler one, such like near linear or near ring, then the value of the graph entropy might decrease. From above discussions, we designed the method to choose optimal neighborhood radius based on the changing speed of graph entropy and variation of its value. In Fig.2, the pseudo code of proposal is shown.

```
procedure ChooseOptimalEps(Data $D$、Minimal candidate value $e$, Maximal candidate value $E$, variation width $\delta e$, Minimum number of neighbors $M$)
    CandidateSet ← $\{e, e + \delta e \cdots, E\}$
    previousInfo ← 0
    previousSpeed ← $\infty$
    for $\epsilon \in$ CandidateSet do
        g ← GrabRPM($D, \epsilon, M$)
        info$_{degree}$ ← CalculateDegInfo($G$)
        if I < previousInfo then
            break
        end if
        diffInfo ← info$_{degree}$ - previousInfo
        if diffInfo > previousSpeed then
            break
        end if
        previousInfo ← I$_{degree}$
        previousSpeed ← diffInfo
    end for
    return $\epsilon$
end procedure
```

Fig. 2 Pseudo code of optimal parameter selection

In Fig.2, function GrabRPM and CalculateDegInfo return constructed STG and calculated graph entropy based on degree. Using algorithm shown in Fig.2 and Exp.5 optimal value for the neighborhood radius ϵ could be obtained.

4. Detection of Structure Transition Events

Considering the noise level of given dataset and using the optimal parameter selection descrived above, the STG that contains information of structural transitions in a trajectory. Since STG reflects the information about the structural transitions contained in a trajectory, by analysing the characteristic sub-structure of STG, it is expected that useful information about structural transitions could be extracted. In such way, we can consider various type of substructure. Then, considering the fact that a trajectory is a time-series data, we can assume two major case about the structure of STGs. First, globally linear structure contains some dense components locally. And second, globally ring structure contains some dense regions locally. In both cases, structure of STG contains locally dense regions as sub-structure. Then we made a hypothesis that such local sub-structure have certain information about the structural transitions and based on this hypothesis, the method to extract the structural transition events from the time region in a trajectory corresponding to the vertices contained in local dense region was designed. This design was based on two major points. 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 second, after the increasing $RMSD$ at certain time point if the variation of $RMSD$ fall into some range

```
procedure ExtractEventTime(Focusing time region $T$, $RMSD$ function in region $T$ $R(t)$, Event occurence threshold $r$, Structure remaining threshold $l$)
    eventTimingSet $\leftarrow \phi$
    for $t \leftarrow T$ do
        if $R(t) > r$ then
            eventTimingSet $Leftarrowt$
        end if
    end for
    eventRegionSet $\leftarrow \phi$
    for $t \leftarrow$eventTimingSet do
        eventRegionSet $\leftarrow t$
        for $u \leftarrow \{t+1, t+2, \cdots,$, NextTiming$(t)\}$ do
            if $R(u) < l$ then
                eventRegionSet $\leftarrow u$
            else
                break
            end if
        end for
    end for
    return eventRegionSet
end procedure
```

Fig. 3 Pseudo code of the algorithm for structural transition event detection

then it could be considered that the structure that occured in previous structural transition event might be kept in some time range.

From discussion above, we designed the algorithm to detect the timing that structural transition might occure. The pseudo code is shonw in Fig.3. Using algorithm described above, the occurence point of that structural transitional events and last time of its effects. In addition, using VMD[20] the characteristics structural transition events in extracted time region could be visualized.

5. Application Experiment

Using algorithms shown in Fig.2 and Fig. 3, we can construct the directed graph that represents structural transition contained in given trajectory and also detect the occurence of characteristic structural transition events. From the information about the time region, the analysis of the structural transition while protein folding could be archieved.

To show the validity and applicability of proposal method, we applied our method to the dataset obtained by Ensign it et.al[14].

5.1 Validity of the Parameter Selection

At first, application of the algorithm to choose optimal radius was conducted. In the experiment, the trajectory Run0 contained in the dataset of Ensign[14]. As candidate set of parameter ϵ, we set its minimum value to 2.0 and maximum value to the maximum $RMSD$ of the trajectory. We set the number of elements in candidate set to 10. And the minimum number of neighbors was also varied from 2 to 8 by 2. STG is directed graph. Then corresponding graph entropy has some arbitrary property for selecting the characteristics, in-degree or out-degee. In this experiment, we used graph entropy based on in-degree. The application results are shown in Fig.4. In Fig.4, the graph entropy corresponding the minimum number of neighbors M and candidate neighborhood radius set $\{\epsilon_i\}$ was plotted on the left. In the figure, the dashed rectangle shows the chosen value of neighborhood radius. Four STGs shown in the right side in Fig.4 are the result under the condition that the minimum number of neighbors $M = 4$ and neighborhood radius was set to optimal value. Comparing STG (a) and (b) in Fig.4, STG (a) has near linear structure. From the fact that a trajectory is time-seriese data, this structure might be trivial. On the other hand, in STG (b), local dense components shown in gree circles. This suggests that occurence of similar structures in different time points. Then STG (b) might represent certain not-trivial information about the structural transitions. STG (d) contains large strongly-connected component shown in beige circle. It is difficult to find characteristic sub-structure in the STG.

From above results, it was shown that proposed method could choose optimal value from candidate value set so that resulted STG is easy to find the sub-structure that some characteristic structural transition might occur.

5.2 Extraction of Structural Transition Events

Next, to show the applicability of the method for detection of structural transition events and the time region that events occured, some experiments were conducted. In this experiment, the trajectory Run2 was used in the dataset to construct STGs. In this experiment, we assumed that dense regions in a STG have certain information about the characteristic structral transitions. Based on the assumption, the time regions corresponding to local dense regions was analysed to detect the structural transition events. After that analysis, the results were aminated with VMD[20]. Using the animation, we analysed which kind of structural transition had occured. Some time steps in the animation were shown in Fig.5.2.

As shown in Fig. 5.2, at the first of focused time region, the incomplete helix shown in white circle was formed. As time goes along, exchanging of left and right region of the protein shown in white rectangle was occured. In addition, by rotation of the region shown with red dashed rectangle, in last state (h) the incomplete helix at time (a) was solved without distruction of other complete helices. From this results, focused sub-graph of STG represents the solution of the incomplete helix that is inconvenient for the progression of the folding.

From discussion above, by anlysing the characteristic substructure of STG, the structural transition events that is important for folding event, could be extracted.

6. Conclusion

In this report, the method to choose optimal parameter for construction of STG was developed. First, to clarify the effect of two parameters to resulted STGs, parameter variation experiment was conducted. Based on the results of the experiment, we designed the method to choose optimal parameter value. And also, using the informaion of charac-

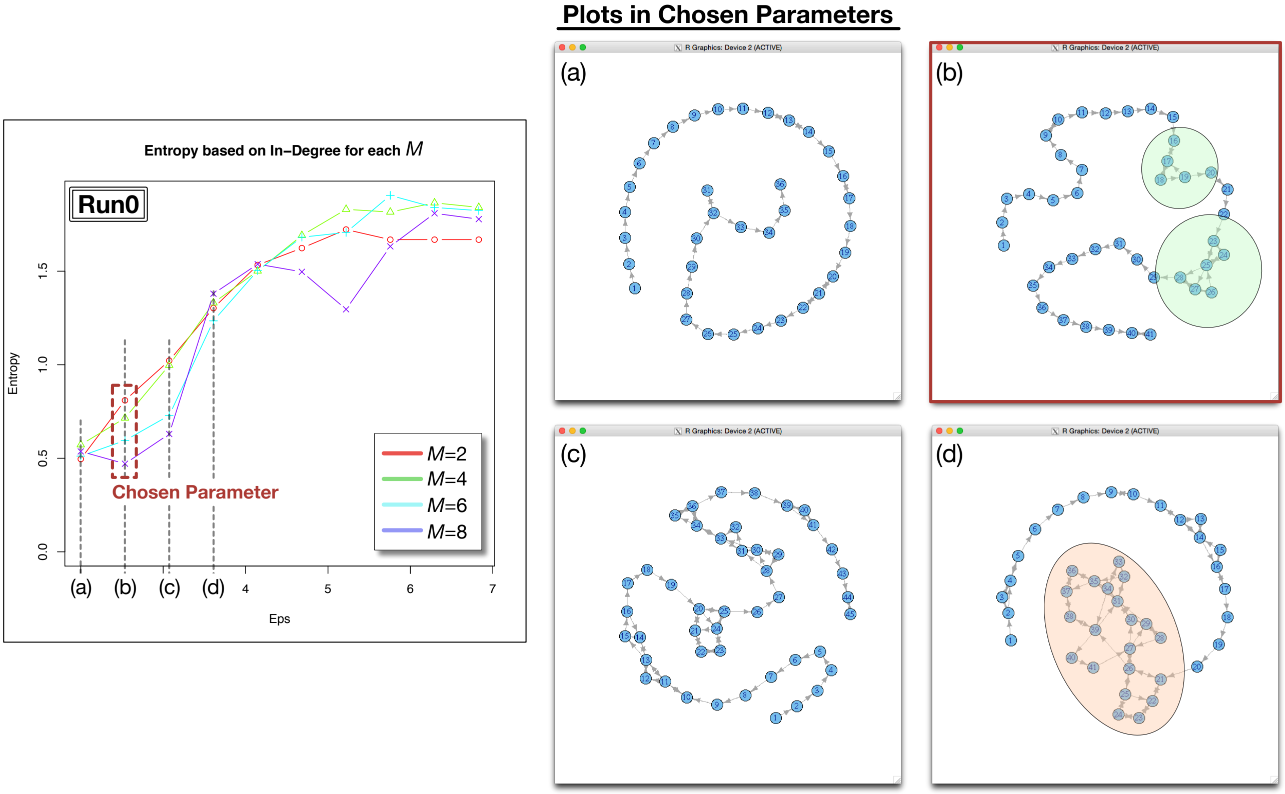

Fig. 4 Results of Parameter Selection with Run0: Selected parameter value is surrounded by dashed rectangle. The graph entropy values were shown in the plot on the left. Appling the proposal method for parmeter selection, the value corresponding point (b) was chosen.

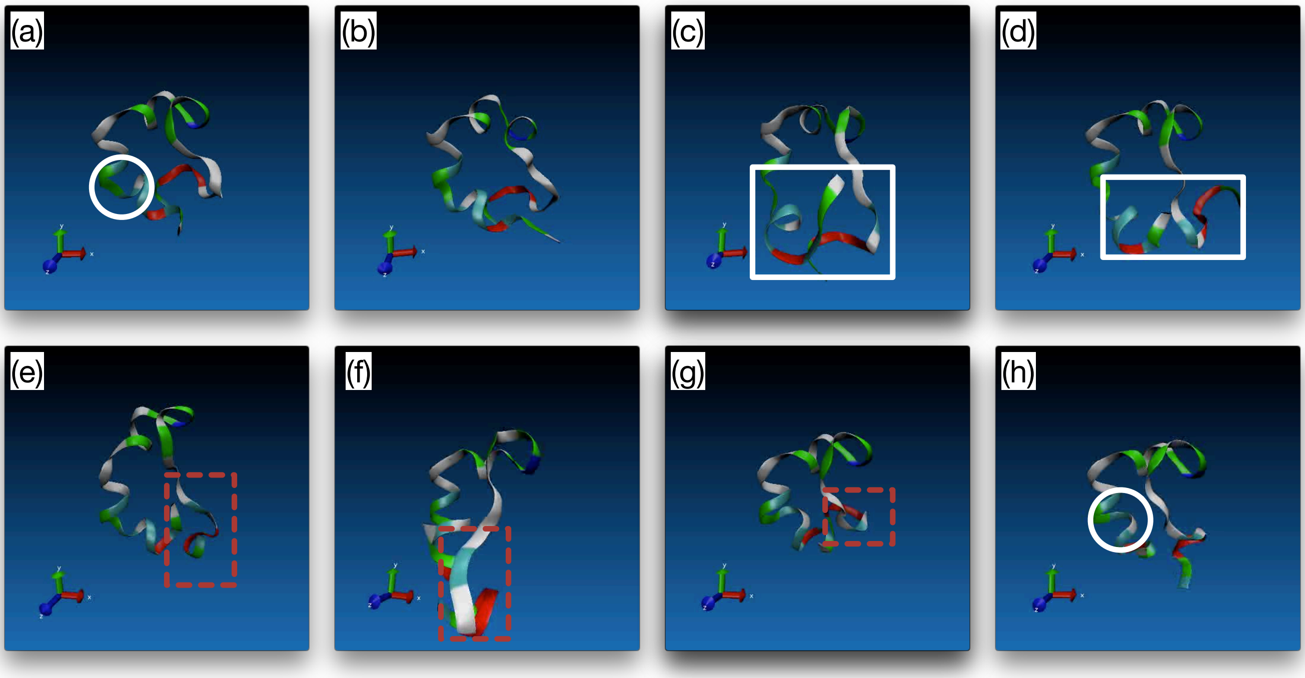

Fig. 5 Animation results of the time region with VMD: In the region shown with white circle, helix was formed incompletely. And in the last part of the time region (h) incomplete formation of helix was solved.

teristics sub-structures, the method for detection of structural transition events was proposed. To show the validity and applicability of two proposals, proposed methods were applied to the dataset consist of trajectories generated MD simulations about the protein, Villin headpiece subdomain of chicken. As a result, their validity and applicability were shown. As future tasks, we have two subjects. First is the extraction of the similar structural transition contained in multiple trajectories. And second is the automation of the extraction of structural transiton events using VMD currently done manually.

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]

[6]

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

[8] Pilar Cossio, Alessandro Laio and Fabio Pietrucci, (2011), "Which Similarity Measure is Better for Analyzing Protein Structures in a Molecular Dynamics Trajectory?", *Phys. Chem. Chem. Phys*, Vol.13, pp.10421-10425

[9] 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

[10]

[11] 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

[12] 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

[13]

[14] 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

[15] Mowshowitz, A., and Dehmer, M., (2012) Entropy and Compresity of Graph Revisited, *Entropy*, Jan, Vol.17, pp.1-11

[16] Shetty, J. and Adibi, J. (2005) Discovering Important Nodes through Graph Entorpy The Case of Enron Email Database, *KDD'2005*, Illinois

[17] Constantine, G. (1990) Graph Complexity and the Laplacian Matrix in Blocked Experiments. *Linear and Multilinear Algebra*, Vol.28, pp.49-56

[18] Bonchev, D. (1995) Kolmogorov's Information, Shannon's Entropy and Topologcal Complexity of Molecules, *Bulg. Chem. Commun*, Vol.28, pp.567-582

[19] Shannon, C.E. and Weaver, W., (1949) 'The Mathematical Theory of Communication, Univesity of Illinois Press

[20] Humphrey, W., Dalke, A. and Schulten, K. (1996(VMD - Visual Molecular Dynamics", *J.Molec.Graphics*, Vol.14, pp.33-38.