Performance Research of Clustering Methods for Detecting State Transition Trajectories in Hemoglobin

Kei TAKAMIa, Yukichi KITAMURAa,b*, Masataka NAGAOKAa,c†

a Department of Complex Systems Science, Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
b Department of Applied Chemistry and Biochemical Engineering, Faculty of Engineering, Shizuoka University, 3-5-1 Johoku Naka-ku Hamamatsu 432-8561, Japan
c Future Value Creation Research Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
* kitamura.yukichi@shizuoka.ac.jp †mnagaoka@i.nagoya-u.ac.jp

(Received: January 14, 2021; Accepted for publication: February 22, 2021; Online publication: April 13, 2021)

The time-series clustering method is one of unsupervised machine learning techniques that classify time-series data. In this article, we applied three methods to the clustering analysis for 200 molecular dynamics (MD) trajectories of human adult hemoglobin (HbA), and have reported their clustering performances for detecting the T-R state transition trajectories (TrajT-R). By compared with their silhouette indices, we have discussed the proper clustering conditions.

Keyword: Time-series clustering method, Dynamic time warping distance, Silhouette analysis, Hemoglobin, Allostery regulation

1 Introduction

Human adult hemoglobin (HbA) is an oxygen-carrying protein contained in red blood cells. It is known that the ability of HbA to transport oxygen is regulated by the transition between two states with different oxygen affinities, the T- and the R-state [1]. Using molecular dynamics (MD) simulation, it was proposed that its T-R transition is enhanced by the dissolved oxygen molecules without formation of O2 binding with heme group, called "non-site-specific allosteric regulation" [2, 3]. However, the T-R state transition MD trajectory (TrajT,R) was characterized by the experimental criterion, i.e. their dimer rotation angles \( \varphi \) obtained by both X-ray crystal structures. Therefore, to non-empirically detect the TrajT,R, we have applied the time series clustering methods to the MD trajectories, which is one of the unsupervised machine learning techniques for time-series data [4]. In this letter, we discussed their clustering performances to classify structural changes in the HbA system.

2 Theory and Methods

2.1 Time-series Clustering Methods [4]

For time-series data, unlike the normal clustering analysis for point data, it is necessary to define a distance for comparing characteristics of time-series data. The simplest definition is Euclidean distance (ED), which compares two time series \( x = x_1, x_2, \ldots, x_m \) and \( y = y_1, y_2, \ldots, y_m \) of length \( m \) as follows:

\[
ED(x, y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}
\]

As another definition of distance, the dynamic time warping (DTW) distance [5, 6] is well-known:

\[
DTW(x, y) = \min \left\{ \sum_{k=1}^{K} w_k : (x, y) \right\}
\]

where a pair of \( x \) and \( y \) is warped non-linearly in time to form a warping path \( W = w_1, w_2, \ldots, w_k, \ldots, w_K (m < K < 2m-1) \), i.e., a set of elements \( w_k = (d_{ij})_k \) chosen from an \( m \)-by-\( m \) matrix (\( d_{ij} \)) that defines all of possible distance mappings between \( x \) and
As shown in equation (2), the DTW distance is defined by the path that minimizes the warping cost. Among numerous possible warping paths, the DTW path can be found via dynamic programming for partially summation under several constraints [6]:

\[ y_k = \sum_{i=1}^{n} \sum_{j=1}^{n} d(x_i, y_j) \]  

(3a)

\[ 1 \leq k \leq K, 1 \leq \rho \leq I, 1 \leq \sigma \leq J. (\kappa \Leftrightarrow (\rho, \sigma)) \]  

(3b)

to finally minimize \(y_k\) by using the recurrence formula:

\[ y_k = d_{\rho\sigma} + \min \{\gamma_{(k-1), (\rho-1, \sigma-1)}, \gamma_{(k-1), (\rho, \sigma-1)}, \gamma_{(k-1), (\rho, \sigma)}\} \]  

(4a)

or, equivalently

\[ y(\rho, \sigma) = d(x_\rho, y_\sigma) + \min \{y(\rho-1, \sigma-1), y(\rho-1, \sigma), y(\rho, \sigma-1)\} \]  

(4b)

which is the cumulative distance \(\gamma_k\) (or \(\gamma(\rho, \sigma)\)) from the element \((1,1)\) (i.e., \(k=1\)) up to \((\rho, \sigma)\) (i.e., \(k=\kappa\)) as a sum of the distance \(d(\rho, \sigma)\) and the minimum of the cumulative distances of the adjacent three elements \((\rho-1, \sigma-1), (\rho-1, \sigma)\) and \((\rho, \sigma-1)\) [5, 6].

In the ED, the distance is calculated by summing up distances at the same time point, while in the DTW distance, it is evaluated by a non-linear alignment, allowing similar shapes to match even if they are out of phase along the time axis (see Figure 1). Presently, DBA (DTW Barycenter Averaging) [6] and soft-DTW where the soft-minimum was considered in either equation (4a) or (4b), were comparatively used [7].

### 2.2 Determining the Number of Clusters

In a time-series clustering, the silhouette analysis [8] is often used for determining the "optimal" number of clusters \(k_{opt}\), based on its tightness and separability. The silhouette width \(\bar{S}_k(\alpha)\) [8] of the \(\alpha\)th cluster (hereafter \(\alpha\)) in \(k\) clusters partitioning \((1 \leq \alpha \leq k)\) can be defined as an average of the silhouette \(S_k(i_\alpha)\) [8] of the \(i_\alpha\)th time-series data (hereafter \(i_\alpha\)) averaged over all \(n_\alpha\) time-series data in the \(\alpha\)th cluster \((1 \leq i_\alpha \leq n_\alpha)\):

\[ \bar{S}_k(\alpha) = \frac{1}{n_\alpha} \sum_{i=1}^{n_\alpha} S_k(i_\alpha) \]  

(5a)

\[ S_k(i_\alpha) = \frac{a_k(i_\alpha) - b_k(i_\alpha)}{\max \{a_k(i_\alpha), b_k(i_\alpha)\}} \]  

(5b)

where \(a_k(i_\alpha)\) is the average distance of \(i_\alpha\) to all other time-series data of the same cluster \(\alpha\) (tightness), while \(b_k(i_\alpha)\) is the minimum out of all the average distances between \(i_\alpha\) to all time-series data in all the other clusters (separability). When \(\alpha\) contains only a single time-series data, \(a_k(i_\alpha)\) is simply defined as zero, and then we easily find \(-1 \leq S_k(i_\alpha) \leq 1\) for each time-series data \(i_\alpha\). If the silhouette index \(\bar{S}_k\) is defined by

\[ \bar{S}_k = \frac{1}{k} \sum_{\alpha=1}^{k} \bar{S}_k(\alpha) \]  

we can evaluate of clustering validity [8, 9]. When \(\bar{S}_k\) is closer to 1, it indicates that the time-series data is classified to an appropriate set of \(k\) clusters, whereas when it is closer to −1, it does that the time-series data should be reclassified into any other clusters partitioning.

### 2.3 Model System and Computational Details

All MD simulations were performed using AMBER 16 [10]. First, the crystal structure in the T state (PDB ID: 2DN2 [11]) was adopted as the initial structure of HbA and used to prepare solution, while another solution containing 120 molecules of oxygen (O2: 0.55 mol/L) was prepared by replacing 120 water molecules in the former one. Second, total 200 initial coordinate files with different positions of oxygen and water molecules were generated via equilibration to adjust the present simulation box under the MD simulation...
Third, after the equilibration with the HbA structure fixed, 200 independent MD trajectories of 8.2 ns were performed, without the HbA structure fixing, by GPU-accelerated pmemd packages. By using python machine learning package tslearn (ver. 0.2.5) [12] for time-series data of $\phi$ defined between a pair of $\alpha\beta$ dimers, we have detected the TrajT-Rs by three time-series clustering methods: ED, DBA [4] and soft-DTW (smoothed formulation of DTW) [7]. In the present study, to determine the optimal number of clusters for 200 MD trajectories, we have investigated $k_s$ defined by equation (5b) using three time-series clustering methods. Finally, $k_s$ was calculated as an average of $s_k(t)$ over all the numbers of TrajT-R.

3 Results and Discussion

Figure 2 shows the silhouette indices $\bar{s}_k$ depending on the number of clusters partitioning $k$ ($k=2$ to 5), obtained by three time-series clustering methods. In all methods, $\bar{s}_k$ was the closest to 1 for $k=2$ in the four. This result indicates that the number of clusters partitioning 2 is optimal (i.e., $k_{opt}=2$) and the clustering result is the denser (or tight) and isolated (or separated).

As summarized in Table 1, soft-DTW brought about the silhouette index $\bar{s}_k$ 0.78 with the number of clusters partitioning $k=2$, which was the closest to 1 among all the clustering methods, indicating the clustering result by soft-DTW is optimal. It is true the ED method is the fastest to estimate $\bar{s}_k$. However, it results in large mutual distances among time-series data with shifts and translations, even when these time series data have a similar time-varying shape.

Moreover, the ratios of TrajT-R in the whole 200 MD trajectories by ED, DBA and soft-DTW were 17%, 27% and 12%, respectively. It was previously reported that those using the empirical criteria are 8% [2] and 16% [3]. The clustering results by ED and DBA included many TrajT-Rs in comparison with those in the previous studies, and these trajectories would show relatively small changes in $\phi$. The ratio of TrajT-R was similar to the previously reported values [2, 3]. Thus, the clustering result by soft-DTW is proper. There was a significant difference in Silhouette indices between DBA and soft-DTW, each of which is classified to hard and soft clustering method, respectively. This indicates that many trajectories are staying in intermediately changed states in the present MD of 8 ns (neither TrajT-R nor T-state remaining trajectories). A soft clustering method, therefore, worked comparatively well in clustering these data. In particular, in present study, the magnitude of $\phi$ was small (almost 10 degrees from T- to R-state) but with the non-negligibly larger fluctuations, thus leading to soft-DTW’s better performance.

4 Concluding Remarks

In this article, three time-series clustering methods, ED, DBA and soft-DTW, were applied to determine the T-R state transition trajectory of HbA (TrajT-R), and the proper clustering conditions were investigated by using 200 MD trajectories. We investigated the silhouette indices by the three methods with the number of clusters partitioning $k$ ($k=2$–5), and found that the silhouette index in soft-DTW (0.78) with the number of clusters partitioning of 2 was closest to 1 in all the three. Moreover, in comparison with their ratios of TrajT-R in the whole 200 trajectories, soft-DTW showed a similar ratio to the previously reported empirical ones, leading to conclude that soft-DTW with $k=2$ is optimal in the analysis of HbA system.

Time-series clustering techniques have not been so often used in data-scientific (DS) analysis of the MD simulation data in chemical and biological systems. However, as in the present study, it is expected that DS analysis for time-series MD trajectories will be widely applied to biological system simulations in future. Presently, we have analyzed TrajT-R classified by time-series clustering methods to investigate the effect of oxygen molecules on the structural change of HbA. This is expected to reveal the mechanism of allosteric
regulation due to non-site-specific interactions of oxygen molecules [2].

Acknowledgments

This work was supported partially by the Core Research for Evolutional Science and Technology (CREST) of the Japan Science and Technology Agency (JST); by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan; and also by “Program for Promoting Researches on the Supercomputer Fugaku” (Fugaku battery & Fuel Cell Project). The calculations were partially performed using several computing systems at the Information Technology Center at Nagoya University, and also at the Research Center for Computational Science at the Institute for Molecular Science, Okazaki, Japan.

Table 1. The silhouette indices and the numbers of TrajT-R in three methods, ED, DBA and soft-DTW.

| Methods     | $\bar{s}_k^a$ | # of TrajT-R $^b$ |
|-------------|---------------|------------------|
| This work:  |               |                  |
| ED          | 0.38          | 34 (17%)         |
| DBA         | 0.37          | 54 (27%)         |
| soft-DTW    | 0.78          | 24 (12%)         |
| Previous studies: |           |                  |
| Empirical criterion | - | 8% [2], 16% [3] |

$a$ $\bar{s}_k$ is an average of $\bar{s}_k(i_\alpha)$ over all the number of TrajT-R.

$b$ The percentage in parentheses are the ratio of TrajT-R in all the 200 MD trajectories.

References

[1] J. F. Storz, (2019). "HEMOGLOBIN: Insights into Protein Structure, Function, and Evolution", Oxford, Oxford University Press. ISBN: 978–0-19–881068–1.
[2] M. Tanakayagi, I. Kurisaki, M. Nagaoka, Sci. Rep., 4, 4601 (2014). DOI:10.1038/srep04601 , PMID:24710521
[3] K. Suzuki, (2019) “Theoretical Research of Allosteric T R State Transition of Human Hemoglobin: Application of Principal Component Analysis and Motion Tree Method”, Master thesis, Nagoya University, Japan.
[4] J. Paparrizos, L. Gravano, ACM Trans. Database Syst., 42, 1 (2017). DOI:10.1145/3044711
[5] H. Sakoe, S. Chiba, IEEE Trans. Acoust. Speech Signal Process., 26, 43 (1978). DOI:10.1109/TASSP.1978.1163055
[6] E. Keogh, C. A. Ratanamahatana, Knowl. Inf. Syst., 7, 358 (2005). DOI:10.1007/s10115-004-0154-9
[7] M. Cuturi, M. Blondel, "Soft-DTW: a Differentiable Loss Function for Time-Series," ICML 2017.
[8] P. J. Rousseweu, J. Comput. Appl. Math., 20, 53 (1987). DOI:10.1016/0377-0427(87)90125-7
[9] $\bar{s}_k$ was called the overall average silhouette width in [8].
[10] D. A. Case, R. M. Betz, D. S. Cerutti, T. E. Cheatham, III, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K. M. Merz, G. Monard, H. Nguyen, H. T. Nguyen, I. Omelyan, A. Onufriev, D. R. Roe, A. Roitberg, C. Sagui, C. L. Simmerling, W. M. Botello-Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, L. Xiao, P. A. Kollman, (2016), AMBER 2016, University of California, San Francisco.
[11] S. Y. Park, T. Yokoyama, N. Shibayama, Y. Shiro, J. R. H. Tame, J. Mol. Biol., 360, 690 (2006). DOI:10.1016/j.jmb.2006.05.036 , PMID:16765986
[12] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, E. Woods, J. Mach. Learn. Res., 21, 1 (2020).