Dimensionality Reduction of High Dimensional Time Series based on Artificial Neural Network

Yilin Wang *
School of Mechanical, Electrical, And Information Engineering, Shandong University, Jinan 250100, China

*Corresponding author e-mail: wangyilin@sdu.edu.cn

Abstract. Molecular dynamics is a molecular simulation method which relies on Newtonian mechanics to simulate the motion of molecular system. In this method, some differential equations are integrated, and the results of integration are further processed to obtain the trajectory or momentum evolution process of some particles controlled by dynamic equations, and the technology of extracting the equilibrium state, motion process or related properties of classical particle system can be used. Through molecular dynamics simulation, we can obtain a series of properties of the system, which are widely used in experimental verification, theoretical derivation and other scenarios. Because it can obtain the dynamic state of macromolecules to make up for the limitations of these properties, it is widely used in the study of transmembrane proteins, polypeptide chains and other systems in life sciences. Through the kinetic path reduction of these systems, we can intuitively understand the characteristics of molecular folding, molecular motion and specific binding, which can play a very important role in the study of proteins and peptides. However, due to the characteristics of high-dimensional time series obtained by molecular dynamics simulation, it is difficult for us to pay attention to the collective state or characteristic process of the whole system in a non-equilibrium state or slow process. This is due to the difficulty in data processing and the difficulty in obtaining its characteristic function. This makes it very difficult to study the dynamic process of the whole system, especially the dynamic process at the intermediate non-equilibrium moment. It is difficult to solve this kind of problem by conventional methods, and only a few special simple systems can be solved by experience. Therefore, it is of great significance to find a method to obtain the characteristic function of the system through the trajectory obtained by molecular dynamics, and then reduce the molecular dynamics path. In order to solve this scientific problem, researchers focus on machine learning. In this study, machine learning method will be used to solve the overall non-equilibrium state of the system or the collective state of the slow process in molecular dynamics simulation. Firstly, we use this method to solve a simple one-dimensional four well model. By this method, we obtain a series of characteristic functions describing the motion process of the model. By sorting the eigenvalue contributions, we obtain some main characteristic functions describing the system. It includes the motion description of Markov smooth transition state and the motion description of four potential wells. At the same time, we use the traditional transition probability matrix to calculate. The difference between the characteristic function
obtained by machine learning and the traditional method is very small, but the calculation method is simpler and more universal. After that, we apply the method to the actual scene. By solving the molecular dynamics simulation of alanine dipeptide structure in polymer protein molecule, the characteristic function of dihedral angle folding of alanine dipeptide structure was preliminarily calculated. The results were consistent with the traditional method.

1. Introduction

Dimension reduction refers to mapping the original high-dimensional data such as high-dimensional time series to low latitude data space through some method, usually a mapping method. It is difficult to avoid losing the information of original high-dimensional data space in the process of dimensionality reduction. Therefore, a successful data dimension reduction is to eliminate redundant data and noise in the original high-dimensional data space as much as possible. The principal component analysis method commonly used in statistical and instrumental analysis is the most commonly used dimension reduction method. It represents the data in high-dimensional data space in low-dimensional data space by some linear mapping projection. It is usually used to observe some information between high-dimensional data which is inconvenient to observe in high dimension, such as: it can be used to measure some cancerous biological tissue by mass spectrometry, and compare with the results of normal tissue. This is often difficult to observe in mass spectrometry data, but it can be seen directly whether the test method has the ability to distinguish samples significantly by PCA.

The result of molecular dynamics simulation is also such a high-dimensional time series. Because of the characteristics of its high-dimensional time series, it is unrealistic to analyze the whole slow process directly, so for such data processing, we need to use the means of data dimension reduction. By restoring the trajectory data of high latitude to the characteristic function describing the essence of its motion characteristics, we can intuitively analyze its overall motion.

At present, the most basic method for high-dimensional time series of molecular dynamics is the transfer operator theory. The theory of transition operator is an iterative method which depends on the properties of transition probability matrix and Markov chain, and obtains the eigenfunction and eigenvalue through the diagonalization of matrix. The numerical solutions of such complex differential equations play an important role in many fields, such as molecular dynamics, fluid dynamics, mechanical and electrical engineering and physics. These systems usually exhibit multi-scale behavior, which may be due to the coupling of subsystems with different time scales, such as fast electrical and slow mechanical components. Or because of the inherent characteristics of the system itself, such as rapid vibration and slow conformational changes. [1-3]

However, from the perspective of computation, due to the ‘curse’ of high latitude data, it is usually not feasible or too difficult to use the method based on transfer operator to analyze such problems. This method can only work well in very simple models or special models. One way to avoid this is to project the dynamics of the high-dimensional system onto the low dimensional space, and then analyze the simplified system. [3] In order to solve this problem, researchers have proposed Time delay independent component analysis (TICA), Variational method of conformational dynamics (VAC) and dynamic mode decomposition (DMD). [4-10]

In this study, the whole state of the system in the molecular dynamics simulation will be solved by artificial neural network. Firstly, we use this method to solve a simple one-dimensional four well model. By this method, we obtain a series of characteristic functions describing the motion process of the model. By sorting the eigenvalue contributions, we obtain some main characteristic functions describing the system, it includes the motion description of Markov smooth transition state and the motion description of four potential wells. At the same time, we use the transfer probability matrix to verify the calculation, which depends on the properties of Markov chain. Markov chain is a stochastic model describing a series of possible events, in which the probability of each event depends only on the state reached in the previous event. In a continuous time, this is called a Markov process. It was named after Russian
mathematician Andrey Markov [10] the transfer probability of several steps can be calculated by the properties of Markov chain and one-step transfer matrix. The difference between the feature function obtained by machine learning and the traditional method is very small, but the calculation method is more simple and more universal.

2. Result and Discussion
Firstly, we choose the simple 1d-4 potential well model [9] as the research object. It is discretized into 100 points between -1 and 1, and a transformation matrix is defined so that there is a non-zero probability of moving left, moving right or staying at the current point at each point. The probability can be obtained by the following formula.

\[ p_{ij} = C_i e^{-(V_j - V_i)\alpha}, \text{ if } ||i - j|| \leq 1 \]

Where \( V_j \) and \( V_i \) represent the central potential energy of the discrete point, \( C \) is the normalization factor, and the synthesis of transition probability is 1. If two particles are not adjacent, the transition probability \( P_{ij} = 0 \). In this way, we construct the transition probability matrix. In the treatment of transfer operator matrix method, we use the properties of Markov chain.

Markov chain is a stochastic model describing a series of possible events, in which the probability of each event only depends on the state reached in the previous event. In continuous time, this is called a Markov process. It is named after Russian mathematician Andrey Markov [10] through the properties of Markov chain and one-step transition matrix, the transition probability of several steps can be calculated. We use transfer operator matrix method and machine learning method to simulate trajectories with 100 time steps. The specific calculation code refers to the better encapsulated machine learning method---master library provided by reference [8].

We use the theoretical calculation of the transfer operator method and the artificial neural network with two hidden layers and 100 hidden neurons in each layer for machine learning. By comparison, we find that the state functions of the transition operator method and the machine learning method are almost the same, which shows that the characteristic functions of the machine learning method and the transfer operator method are in good agreement. This proves the accuracy of machine learning method to a certain extent, and it has the advantage that it can be applied to more systems. Compared with other machine learning methods given in the literature, machine learning method has the advantage of good spatial complexity.

2.1. Research Model
The model we used in this study is given by the following formula [9]

\[ V(x) = 2\left(x^8 + 0.8e^{-80x^2} + 0.2e^{-80(x-0.5)^2} + 0.5e^{-40(x+0.5)^2}\right) \]

Where \( V \) is the potential energy of the system and the domain of definition is -1 to 1. We divide the range of -1 to 1 into 100 parts and draw the following figure. The potential energy of each point represents the potential energy of the central point.

![Figure 1. Potential energy location diagram of 1d-4 potential well model](image)
2.2. **Treatment of 1d-4 potential well model by transfer operator matrix method**

Firstly, we calculate the characteristic function of the model motion through the transfer operator theory. The complete calculation code is shown in Appendix 1. For the adjacent parts, the transition probability and potential energy have the following relation:

\[
p_{ij} = \begin{cases} 
  C_i e^{-(\psi_i - \psi_j)}, & \text{if } |i - j| \leq 1 \\
  0, & \text{otherwise,}
\end{cases}
\]

The transition probability matrix \( P(1) \) can be constructed by the above formula. See Appendix 2 for details. The following one-step transition matrix can be obtained.

\[
\begin{bmatrix}
0.39307861 & 0.21156085 & 0 & \ldots & 0 & 0 & 0 \\
0.60692139 & 0.3265427 & 0.23397667 & \ldots & 0 & 0 & 0 \\
0. & 0.46179408 & 0.33076832 & \ldots & 0 & 0 & 0 \\
\ldots \\
0. & 0. & 0. & \ldots & 0.33076843 & 0.4618153 & 0 \\
0. & 0. & 0. & \ldots & 0.23397667 & 0.3265427 & 0.60692139 \\
0. & 0. & 0. & \ldots & 0.21156085 & 0.39307861 & 0
\end{bmatrix}
\]

**Figure 2.** One step transition probability matrix example

\( P(100) \) can be obtained by constructing the transfer probability matrix and evolving it 100 times by the properties of Markov chain. The eigenvalues and eigenvectors of \( P(100) \) are solved and sorted. The first four characteristic functions of the system motion can be obtained as follows:

**Figure 3.** The first four characteristic functions of one-dimensional four well model solved by transfer operator method

The first characteristic function corresponds to the smooth transition distribution and is equal to 1 for the transfer operator because of the normalized result in the calculation. The other characteristic functions represent the motion process between the potential wells, that is, when the system crosses a certain potential well, the motion state of the whole system is described. The barrier heights of these wells determine the related eigenvalues \( \lambda \). These four characteristic functions can better describe the overall motion characteristics of the original model. Of course, for the 50 * 50 transfer matrix, we can obtain at most 50 eigenvalues and their corresponding eigenfunctions, but the eigenfunctions ranked by eigenvalues have little contribution to the overall motion of the system, so only the first four state functions with the largest contribution are listed here.
2.3. Treatment of 1d-4 well model by machine learning method

After that, we use machine learning method to deal with the system. In the process of processing, we directly use the potential energy model and one-step transfer matrix provided in 2.1 and 2.2. The time lag parameters are identical to those in 2.2. According to the learning ideas provided by reference [8] and the artificial neural network construction method provided by keras [11], we use the artificial neural network with two hidden layers and 100 hidden neurons in each layer to learn machine learning method. The results show that we choose three characteristic functions corresponding to the unbalanced over balance in 2.2, as shown in the following figure.

![Figure 4. Comparison of the characteristic function calculated by machine learning method and the result of transfer operator method. The purple solid line is the machine learning method, and the blue is the transfer operator method in 2.2.](image)

By comparison, we find that the state functions of the transition operator method and the machine learning method are almost the same, which shows that the characteristic functions of the machine learning method and the transfer operator method are in good agreement. This proves the accuracy of machine learning method to a certain extent, and it has the advantage that it can be applied to more systems.

After that, we compared the calculation results of the model with those of the TICA method reported in the literature [12], and the results are still highly consistent. It is worth mentioning that when I try the TICA code given in the literature on my home computer, there will be memory overflow and incalculability, while machine learning method will not. This is because the spatial complexity of TICA method is exponentially higher than that of machine learning method, which makes many models have to be further optimized and simplified to run successfully, and machine learning method has obvious advantages in this aspect.

2.4. Analysis of alanine dipeptide by machine learning method

After the machine learning method was tested on a simple test model, we selected the alanine dipeptide (N-acetyl-L-alanine-N’-methylamide) as the research object, in which alanine dipeptide was in water phase. Alanine dipeptide is a simple 22 atom peptide, and is often used in a new standard system for biomolecular simulation. Because the molecular structure has four main chain dihedral angles which determine the configuration state, the conformation of the molecule can be determined by calculating the dihedral angle of the main chain. We use the molecular simulation results given by OpenMM software [8, 13]. OpenMM is a software toolkit for performing molecular simulations on a series of high performance computing architectures. It is based on a hierarchical architecture: the lower layer acts as a reusable library that can be called by any application, while the higher layer forms a complete environment for running molecular simulation.

The software is used to simulate the condition [13-14]. The alanine dipeptide in water was simulated for 200 ns under the condition of T = 300, and the force field of amber99sbildn was used to model. In the simulation parameters, K and P = 1 bar, Lennard Jones interaction smoothly switches to zero at the 1.4 nm cutoff, and the particle mesh Ewald is used for electrostatic treatment. The actual space cutoff
value is 1.4 nm, and the space grid spacing between them is 0.12 nm. Save the configuration every 2 PS to generate a track containing 1000000 calculations. The influence of solvent molecules on peptide configuration is implicitly treated. In this case, the theoretical feature function of the transfer operator is not available, so we will directly simulate the machine learning method.

The results of machine learning method are superimposed on the Ramachandran graph. The Laplace diagram shows the conformation of amino residues, and describes whether the dihedral angle of amino acid residues in protein structure is in a reasonable region, which can test the modeling quality after modeling [16] As can be seen from figure 2.3.1, the first slow mode captures the transition along φ torsional motion, the second slow mode captures the transition between α and (β, P / /), while the second slow mode captures the transition between α and (β, P / /) is shown in Fig. 5 [15].

![Conformation of alanine dipeptide α, β and P //](image)

**Figure 5.** Conformation of alanine dipeptide α, β and P //. Among them, yellow is the central carbon atom, gray is the carbon atom, blue is the nitrogen atom, and red is the oxygen atom.

The conformations of the molecule of αL and γ are shown in Figure 6. Through these structural diagrams, we can clearly understand the path of molecular motion among several metastable conformations.

![Conformation of alanine dipeptide α, β and P //](image)

**Figure 6.** Conformation of alanine dipeptide α, β and P //. Among them, yellow is the central carbon atom, gray is the carbon atom, blue is the nitrogen atom, and red is the oxygen atom.

Compared with the results of KTICA method reported in the literature, its basic characteristics are completely consistent, but its time scale is generally about 20% [17], which needs to be further explored.

### 3. Conclusion

In this study, machine learning method will be used to solve the overall non-equilibrium state of the system or the collective state of the slow process in molecular dynamics simulation. Firstly, we use this method to solve a simple one-dimensional four well model. By this method, we obtain a series of characteristic functions describing the motion process of the model. By sorting the eigenvalue contributions, we obtain some main characteristic functions describing the system, it includes the motion description of Markov smooth transition state and the motion description of four potential wells. At the same time, we use the traditional transition probability matrix to calculate. The difference between the
characteristic function obtained by machine learning and the traditional method is very small, but the calculation method is simpler and more universal.

After that, we apply the method to the actual scene. By solving the molecular dynamics simulation of alanine dipeptide structure in polymer protein molecule, the characteristic function of dihedral angle folding of alanine dipeptide structure was preliminarily calculated. We use the method reported in the literature to carry out the machine learning method, and process the results of molecular simulation of alanine dipeptide. The results of machine learning method are superimposed on the Ramachandran diagram, the first slow mode captures the transition along $\phi$ torsional motion, the second slow mode captures the transition between $\alpha$ and $(\beta, P/\gamma)$, while the second slow mode captures the transition between $\alpha L$ and $\gamma$. Results compared with the results of KTICA method reported in the literature, its basic characteristics are completely consistent, but its time scale is generally about 20%. This needs to be further explored.

References

[1] Wang Bing. Analysis of Molecular Dynamics Simulation Trajectory of BamA Protein Based on Machine Learning Algorithms [D]. Jilin University, 2019. (in Chinese)

[2] Chen Haihong. Principle and application of machine learning [M]. University of Electronic Science and Technology Press, 2017 (in Chinese)

[3] Stefan Klus, et al. Data-Driven Model Reduction and Transfer Operator Approximation [J]. Journal of Nonlinear Science. 985–1010, 2018.

[4] Christian R, et al. Modeling Molecular Kinetics with tICA and the Kernel Trick [J]. Journal of Chemical Theory and Computation 2015, 11, 2, 600–608.

[5] Matthew P, et al. Landmark Kernel tICA for Conformational Dynamics [J], bioRxiv, 2017.

[6] T. Chen and H. Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems [J], IEEE Trans. Neural Networks 6, 911, 1995.

[7] Jürgen Schmidhuber, Deep learning in neural networks: An overview [J], Neural Networks, 85-117, 2015.

[8] Wei Chen, et al. Nonlinear discovery of slow molecular modes using state-free reversible VAMPnets [J]. The Journal of Chemical Physics. 150, 214114 2019.

[9] Prinz, et al. Markov models of molecular kinetics: Generation and validation [J], Journal of Chemical Physics, 134 (17), 2011.

[10] Gagniuc, Paul A. Markov Chains: From Theory to Implementation and Experimentation [M]. USA, NJ: John Wiley & Sons, 1–235 2017.

[11] F. Chollet, Keras([CP/DK], https://keras.io.

[12] Christian R, et al. Modeling Molecular Kinetics with tICA and the Kernel Trick [J]. Journal of Chemical Theory and Computation, 2015, 11, 2, 600–608.

[13] P. Eastman, et al. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation [J], Journal of Chemical Theory and Computation, 2013, 9, 1, 461–469.

[14] P. Eastman, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics[J], Plos Computational Biology, 2017, 13, e1005659.

[15] Alexander N, et al. Role of Solvent in Determining Conformational Preferences of Alanine Dipeptide in Water[J], Journal of the American Chemical Society, 2004, 126, 8, 2574–2581.

[16] Morris AL, et al. Stereochemical quality of protein structure coordinates [J]. Proteins, 1994 12, 345-364.

[17] Benjamin Trendelkamp-Schroer, et al. Estimation and uncertainty of reversible Markov models [J]. The Journal of Chemical Physics, 143, 174101 (2015).