IEICE Proceeding Series

When Computational Mechanics Meets Single Molecule Time Series

Chun-Biu Li, Tamiki Komatsuzaki

Vol. 2 pp. 445-447
Publication Date: 2014/03/18
Online ISSN: 2188-5079

2013 International Symposium on Nonlinear Theory and its Applications
NOLTA2013, Santa Fe, USA, September 8-11, 2013

When Computational Mechanics Meets Single Molecule Time Series

Chun-Biu Li and Tamiki Komatsuzaki

Research Institute for Electronic Science, Hokkaido University
Kita 20 Nishi 10, Kita-ku, Sapporo 001-0020, Japan
Email: cbli@es.hokudai.ac.jp

Abstract—Developed in the context of information theory, computational mechanics (CM) has been formulated to construct the minimal but the most predictive hidden Markov model, originally termed ϵ-machine, which is able to reproduce the causal structures statistically from time series. Here I will present several generalizations of CM to the study of complex dynamics and kinetics of single molecule (SM) time series. These include the incorporation of wavelet decomposition into CM to construct the multiscale state-space networks for non-stationary SM time series, and the introduction of soft (or lossy) clustering in defining states when noise and measurement error present in the data.

1. Introduction

Single molecule (SM) experiments provide us with unique information on the distribution of molecular properties and their dynamic behaviors, which are inaccessible from ensemble-averaged measurements. In general, the complexity observed in the dynamics and kinetics of a protein originates in the underlying multidimensional energy landscape. The dynamics can be understood as the protein traversing from one state (node) to another along a complex network in the state space. The network properties of biological systems can therefore offer us new perspectives to address the nature of hierarchical organizations in the multidimensional state space and its implications in the SM complex kinetics. Here we address how computational mechanics (CM) [1] extracts the state space network (SSN) of biological systems explicitly from a SM time series, free from a priori assumptions on the underlying physical model. Moreover, we discuss several generalizations of the scheme to handle possible complications one may face in handling real SM time series, such as nonstationarity and the presence of measurement errors.

2. Computational Mechanics: Construction of State-Space Network

We briefly describe how the original CM defines states and constructs connections among them from scalar time series. For a given time series $\mathbf{x} = (x(t_1), x(t_2), \cdots, x(t_N))$ of a physical observable x with continuous values (e.g. the interdye distance reported by fluorescent probes), we first discretize it to obtain the symbolic sequence $\mathbf{s} = (s(t_1), s(t_2), \cdots, s(t_N))$ with $s(t_i)$ denoting the symbolized observable at time t_i. Since CM requires a statistical sampling of subsequences in the symbolic time series $\mathbf{s}$, the choice of discretization scheme depends not only on the experimental resolution but also on the statistical properties of the time series. A reasonable discretization is such that the topological properties of the constructed network are insensitive to the increase in the number of symbols.

The next step in the construction is to trace along $\mathbf{s}$ for each time step t_i to record which subsequence of length L_{future}, $s_A^{future} = \{s(t_{i+1}), \cdots, s(t_{i+L_{future}})\}$, follows consecutively after a subsequence of length L_{past}, $s_B^{past} = \{s(t_{i-L_{past}+1}), \cdots, s(t_i)\}$ ($A, B, \cdots$ represent different symbolic subsequences that appear in $\mathbf{s}$). The transition probability from s_B^{past} to s_A^{future}, denoted by $P(A|B)$, is then obtained for the time series $\mathbf{s}$. In CM, a "state" (denoted by S_i hereinafter) is defined by the set of past subsequences $\{s_{B'}^{past}, s_{B''}^{past}, \cdots\}$ with length L_{past} whose transition to the future subsequence s_A^{future} takes place with the (nearly) same transition probabilities (i.e., $P(A|B') \cong P(A|B'') \cong P(A|B''') \cong \cdots$ for all A). A transition from a state S_i to another S_j is constructed with its weight equal to the transition probability $P(s_A^{future}|S_i)$ if the subsequence s_A^{future} is generated from a transition from S_i to S_j along the time series $\mathbf{s}$. The extraction of all states and transitions among them yields a SSN associated with the time series $\mathbf{s}$.

Without the needs to postulate the number of states and the connectivity of the network, a unique characteristic of the CM is that it extracts the underlying SSN directly from time series with length L_{past} chosen such that the topological feature of the SSN converges as L_{past} increases from zero. It has been proven mathematically [2] that the converged SSN makes all transitions among the states Markovian, i.e., the next state to visit depends only on the current states. Moreover, the converged SSN is a hidden Markov model with minimal complexity and maximum predictive power which can best reproduce the statistics of the time series $\mathbf{s}$. The states in the SSN are defined by not simply the value of $s(t)$ at each instantaneous time but a subsequence of symbols $(s(t_i), s(t_{i-1}), s(t_{i-2}), \cdots)$ when memory exists in the process. Therefore, the CM provides a natural means to lift "degeneracy" (different physical states having the same value in the measured observable) within the limited information of scalar time series.

3. Construction of Multi-timescale SSN using Wavelet Decomposition for SM Nonstationary Time Series

Despite the above attractive features of CM, a difficulty in the original formulation of CM arises when the length of the past sequences L_{past} increases. In particular, the number of possible past sequences s_B^{past} grows exponentially with L_{past} and the statistical accuracy in sampling of s_B^{past} becomes rapidly worse due in real applications. As a result, with the original procedure it may be too hard to properly capture long-time memory effects, especially for the time series obtained from SM measurements in which hierarchies of non-stationarity can exist that spoil the convergence of the SSN with respect to increasing L_{past}.

One possible generalization of the CM to handle nonstationary time series is to perform a wavelet decomposition of the time series into a set of stationary series (the detail components) and a non-stationary series (the approximation component) with different timescales. The combination of wavelet decomposition and CM [3, 4], termed wavelet based CM, allows us to properly quantify the characteristic length of memory for each of the wavelet decomposed time series with definite timescale. This avoids poor statistical accuracy in sampling the past subsequences.

The wavelet based CM can also to some extent resolve the degeneracy problem inherent in observations since the original scalar time series is decomposed into a vector time series with approximation and details as components. More importantly in defining "states" from scale time series, we take into account not only the value itself at each instantaneous time step, but also the time sequence (i.e., history) near the instantaneous time step. The combination of CM with the wavelet decomposition is thus expected to avoid the degeneracy problem more than just either the standard CM or wavelet decomposition alone.

The wavelet based CM has been applied to investigate the dynamics of conformational fluctuations probed by SM electron transfer (ET) experiment detected on a photon-by-photon basis [5]. It has been shown [3, 4] that the topographical features of the SSNs depend on the timescale of observation; the longer the timescale, the simpler the underlying SSN becomes, leading to a transition of the dynamics from anomalous diffusion to normal Brownian diffusion on the multidimensional energy landscape.

4. CM as a Soft Clustering Problem

It is noted that the SSN construction scheme discussed in Section 2 can be viewed as a clustering problem in which past subsequences are grouped to form states according to their transition probability to the future subsequences. In particular, the clustering scheme described in Section 2 corresponds to a "hard" clustering scheme in which each past subsequence can be assigned to only one state. In practice, measurement errors can exist in the time series, implying that uncertainty also presents in the transition probability $P(A|B)$ from the past subsequence s_B^{past} to the future subsequence s_A^{future}. Such uncertainty may not allow us to perform a rigorous hard clustering of the past subsequence into states, and a soft (or lossy) clustering approach is desired. In the soft clustering, each past subsequence can belong to several states with a membership, $P(S_i|s_B^{past})$, which specifies the probability that the past subsequence s_B^{past} belongs to the state S_i. The hard clustering is a special case of the soft clustering in which $P(S_i|s_B^{past})$ equals to either zero or one.

Here we argue that bootstrapping method [6] provides a simple and nonparametric scheme to determine the soft clustering membership $P(S_i|s_B^{past})$ in the SSN construction that incorporates the measurement errors. Suppose that one obtains a symbolic time series, $\mathbf{s} = (s(t_1), s(t_2), \cdots, s(t_N))$, from symbolizing the experimental time series $\mathbf{x}$, we further assume that there are some uncertainties in assigning the symbol due to the measurement errors in $\mathbf{x}$ such that at each time instant t_i, the symbol visited by the system is specified by a symbolization probability, $P(s(t_i)|x(t_i))$. The explicit form of $P(s(t_i)|x(t_i))$ depends on the particular symbolization scheme we use and on the statistics of the measurement errors in $x(t_i)$. Next, we generate an ensemble of bootstrapped symbolic time series $\{\mathbf{s}_1^{boot}, \mathbf{s}_2^{boot}, \cdots\}$, each with the same length of the original time series $\mathbf{x}$, by random sampling from $P(s(t_i)|x(t_i))$ for all t_i. Each bootstrapped symbolic time series $\mathbf{s}_j^{boot}$ can be viewed as a possible realization of the symbolic dynamics associated with the measurement errors. The SSN construction scheme described in Section 2, which is a hard clustering scheme, is then applied to each of the bootstrapped symbolic time series to obtain the corresponding bootstrapped SSN by grouping past subsequence into states. Due to the variations among the bootstrapped symbolic time series, the assignment of past subsequences to states can also vary among different bootstrapped SSNs, Finally, the membership $P(S_i|s_B^{past})$ can be constructed by counting how many times a given s_B^{past} is assigned to the state S_i in the set of bootstrapped SSNs. It can be easily seen that if no measurement error presents in original time series $\mathbf{x}$, the symbolization is unique, i.e., $P(s(t_i)|x(t_i))$ equals to either zero or one, and so all bootstrapped symbolic time series are the same. This implies that there is no variation in assigning a given s_B^{past} to the states, and therefore, the clustering simply reduces to a hard clustering one.

5. Concluding Remarks

In this article, we have discussed two possible generalizations of the CM to handle possible complications one may face when constructing hidden Markov model from SM time series. These include the wavelet based CM that extracts the multiscale SSNs from nonstationary time series, and the extension to soft clustering in the SSN construction that incorporates the effects of measurement error into the analysis. We also note here that the SSN construc-

tion with soft clustering can also be formulated in terms of the information bottleneck method [7] in which the memberships $P(S_i|s_B^{past})$ are obtained by finding the best tradeoff between the predictivity and complexity of the model.

Acknowledgments

The authors would like to thank NOLTA2013 organizing committee members for their fruitful suggestions and comments.

References

[1] J. P. Crutchfield and K. Young, "Inferring Statistical Complexity," *Phys. Rev. Lett.*, vol.63, pp.105, 1989.

[2] C. R. Shalizi and J. P. Crutchfield, "Computational Mechanics: Pattern and Prediction, Structure and Simplicity," *J. Stat. Phys.*, vol.104, pp.816, 2001.

[3] C-B. Li, H. Yang and T. Komatsuzaki, "Multiscale Complex Network of Protein Conformational Fluctuations in Single-Molecule Time Series," *Proc. Natl. Acad. Sci. USA*, vol.105, pp.536, 2008.

[4] C-B. Li, H. Yang and T. Komatsuzaki, "New Quantification of Local Transition Heterogeneity of Multiscale Complex Networks Constructed from Single-molecule Time Series," *J. Phys. Chem. B*, vol.113, pp.14732, 2009.

[5] H. Yang *et. al.*, "Protein Conformational Dynamics Probed by Single-Molecule Electron Transfer," *Science*, vol.302, pp.264, 2003.

[6] B. Efron, "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods," *Biometrika*, vol.68, pp.589, 1981.

[7] N. Tishby, F. C. Pereira and W. Bialek, "The Information Bottleneck Method," *The 37th annual Allerton Conference on Communication, Control, and Computing*, pp368, 1999.