Non-Markovian properties and multiscale hidden Markovian network buried in single molecule time series

Tahmina Sultana, Hiroaki Takagi, Miki Morimatsu, Hiroshi Teramoto, Chun-Biu Li, Yasushi Sako, and Tamiki Komatsuzaki

Citation: The Journal of Chemical Physics **139**, 245101 (2013); doi: 10.1063/1.4848719
View online: http://dx.doi.org/10.1063/1.4848719
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/139/24?ver=pdfcov
Published by the AIP Publishing

Non-Markovian properties and multiscale hidden Markovian network buried in single molecule time series

Tahmina Sultana,[1,2] Hiroaki Takagi,[3] Miki Morimatsu,[4] Hiroshi Teramoto,[1,2] Chun-Biu Li,[2,5,6] Yasushi Sako,[7] and Tamiki Komatsuzaki[1,2,6,a)]

[1]*Graduate School of Life Science, Transdisciplinary Life Science Course, Hokkaido University, Kita 12, Nishi 6, Kita-ku, Sapporo 060-0812, Japan*
[2]*Molecule and Life Nonlinear Sciences Laboratory, Research Institute for Electronic Science, Hokkaido University, Kita 20, Nishi 10, Kita-ku, Sapporo 001-0020, Japan*
[3]*Department of Physics, Nara Medical University, 840 Shijo-cho, Kashihara Nara 634-8521, Japan and Japan Science and Technology Agency (JST), CREST, Suita, Osaka, Japan*
[4]*WPI Immunology Frontier Research Center (WPI-IFReC), Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan*
[5]*Department of Mathematics, Graduate School of Science, Hokkaido University, Kita 12, Nishi 6, Kita-ku, Sapporo 060-0812, Japan*
[6]*Research Center for Integrative Mathematics, Hokkaido University, Kita 20, Nishi 10, Kita-Ku, Sapporo, Hokkaido 001-0020, Japan*
[7]*Cellular Informatics Laboratory, RIKEN, 2-1 Hirosawa, Wako 351-0198, Japan*

(Received 4 September 2013; accepted 2 December 2013; published online 27 December 2013)

We present a novel scheme to extract a multiscale state space network (SSN) from single-molecule time series. The multiscale SSN is a type of hidden Markov model that takes into account both multiple states buried in the measurement and memory effects in the process of the observable whenever they exist. Most biological systems function in a nonstationary manner across multiple timescales. Combined with a recently established nonlinear time series analysis based on information theory, a simple scheme is proposed to deal with the properties of multiscale and nonstationarity for a discrete time series. We derived an explicit analytical expression of the autocorrelation function in terms of the SSN. To demonstrate the potential of our scheme, we investigated single-molecule time series of dissociation and association kinetics between epidermal growth factor receptor (EGFR) on the plasma membrane and its adaptor protein Ash/Grb2 (Grb2) in an *in vitro* reconstituted system. We found that our formula successfully reproduces their autocorrelation function for a wide range of timescales (up to 3 s), and the underlying SSNs change their topographical structure as a function of the timescale; while the corresponding SSN is simple at the short timescale (0.033–0.1 s), the SSN at the longer timescales (0.1 s to ∼3 s) becomes rather complex in order to capture multiscale nonstationary kinetics emerging at longer timescales. It is also found that visiting the unbound form of the EGFR-Grb2 system approximately resets all information of history or memory of the process.
 [http://dx.doi.org/10.1063/1.4848719]

I. INTRODUCTION

Biological processes such as signal cascade in a cell are comprised of a series of chemical reactions and structural transitions across multiple levels of molecular machinery. Single-molecule (SM) spectroscopy provides a means to uncover the complex mechanism of molecular machinery at single molecule level that one cannot address in bulk measurements.[1–16] For example, it was found in an *in vitro* reconstituted system that the recognition (i.e., dissociation and association) kinetics between epidermal growth factor receptor (EGFR) on the plasma membrane and its adaptor protein Ash/Grb2 (Grb2) show non-exponential kinetics, and association kinetics depends nonlinearly on the Grb2 concentration, suggesting the existence of molecular memory in the signaling process.[16] Such innovative experimental developments of synthetic approaches using intact plasma membrane fractions have offered new challenges in theoretical modeling of the underlying mechanism of molecular machinery.

In the analyses of SM time series, hidden Markov models (HMM) are often used to provide insights on the complex mechanism of molecular machinery.[17–25,27–31] However, derivation of a HMM that incorporates non-exponential kinetics and assignment of the physical correspondence of the constructed states are two of the most contemporary subjects of research yet to be resolved. Moreover, most existing methods (for mathematical modeling) depend on a huge amount of fitting parameters, e.g., algorithms using maximum likelihood estimator to obtain the parameters of the HMM.[19,26] These methods, in general, require the initial assumption of model parameters, such as network structure and number of states.

Recently, a novel, data-driven model was developed to naturally derive the underlying multiscale state space network (SSN) from SM time series based on information theory.[21–23,32–34] The SSN is regarded as a certain class of

a)Email: tamiki@es.hokudai.ac.jp

HMM that represents the complex kinetics such as non-exponential properties and molecular memory of the process, and identifies the physical correspondence. However, conventional methods (e.g., HMM) are state-labeled, i.e., producing a symbol when a state is visited. On the other hand, our SSN is edge-labeled, i.e., producing a symbol when a transition made. The states of SSN depend not only on the present value of the observable but also on the past information along the course of time evolution so that the state-to-state transitions are Markovian even though dynamical correlation may exist in the time series. The original idea of the SSN was developed in 1980s,[32–34] aiming at discovering the pattern of dynamical features buried in a stationary time series. It was recently generalized into nonstationary time series by using wavelet transforms to decompose the time series into components at different scales.[21–23] It was found[21–23] in a single molecule electron transfer experiment of the NADH:flavin oxidoreductase (Fre) complex[35] that the topographical features of the SSN change as a function of timescale to capture the transition from abnormal to normal diffusion observed in the protein fluctuation.

In this article we generalize the wavelet-based multiscale SSN construction scheme to a discrete time series such as an ON/OFF binary time series. We introduce a simple skipping step method (SSM) to decompose the original time series into a set of time series at different timescales. We then derive analytic expressions of kinetic properties, such as the autocorrelation function of a given discrete time series, in terms of the intrinsic properties of SSNs. By using the scheme, we scrutinize the SSN constructed for discrete time series of association and dissociation kinetics between EGFR and Grb2.[16] We show that our analytic formula accurately reproduces the autocorrelation function, and that the underlying SSNs change their properties as a function of timescale. It is also found that one of the states constituting the network serves to reset the memory of the process.

The article is organized as follows: In Sec. II we briefly describe the construction procedure of the SSN and present an analytical expression of autocorrelation by using the general property of the SSN. We introduce a simple scheme called SSM to derive the underlying SSNs at different timescales. In Sec. III, we apply our SSN analysis combined with the SSM to analyze the fluorescence intensity time trace of the binding and unbinding processes between EGFR and Grb2 in an *in vitro* reconstituted system.[16] The conclusion and future perspectives will be given in Sec. IV. In the Appendix we illustrate an example of a three-state toy model whose "observable" is binary as for the demonstration of SSN and SSM.

II. THEORY

A. A brief description of how to construct SSN

Here we briefly explain the procedure for constructing the SSN.[33] If the underlying kinetics have some memories (or information from the past events), the states in the SSN are defined not only by the present value of the observable but also by the past subsequence(s) of the values (with a specific range). In short, the range of the past subsequence, denoted by L_{past} hereinafter, corresponds to the characteristic timescale of correlation of the event. One does not require the system in question to be locally equilibrated or satisfy detailed balance. This method starts from discretizing a continuous time series into a symbolic time series, when the time series of interest is continuous. The symbolization and the number of symbols depend on the nature of time series, experimental setup, signal-to-noise ratio, and so forth.[23,36,38]

The second step is to evaluate the transition probabilities from different subsequences (called past subsequences) to the future symbols, e.g., the transition probability $P(s_i|s_2s_1)$ for any symbol s_i (i=1,..., the total number of symbols) to appear at a time, say t, following a particular subsequence s_2s_1 in which one observes s_1 at time $t-1$ and s_2 at time $t-2$. Similarly, the transition probabilities for all other past subsequences s_js_k with past length two will be evaluated. The suitable value of L_{past} depends on the nature of the underlying dynamics of the time series: L_{past} corresponds to the characteristic timescale of correlation and L_{past} is unity when the process (the time series) is Markovian. The actual value of L_{past} was taken to be the minimal value at which the structural property of the constructed SSN does not change even at increasing the value of L_{past}.[21]

The third step is to derive states in the network by using the transition probabilities for the past subsequences. The transition probability for the past subsequence whose length is optimal in capturing the memory in the process provides all information required to predict the future. The states are defined as follows: for a given L_{past}, if the transition probabilities of two past subsequences are regarded as the same, we group the two past subsequences together into the same set called a "state" (denoted by S_I hereinafter). This is because all composite past subsequences in the "state" generate the same future symbols with the same probabilities, indicating that, although grouping the subsequences simplifies the description by reducing the number of states, there is no loss in the model's predictivity. In the context of SM time series, a state in the constructed SSN is interpreted as a collection of a series of conformational changes that have the same transition properties.

The final step is to link the states with each other to form a network. The transition probability from state S_I to state S_J producing symbol s_j, denoted by $P(S_Js_j|S_I)$, yields the weight of the transition from S_I to S_J in the network with the generated s_j. Here, we require that the next state S_J is uniquely determined by the current state S_I and the next symbol s_j. The advantage of this property is that there is a one-to-one correspondence between the symbolic sequences (i.e., the time series) and the state sequences that are generated by the SSN. Since all memory effects are encoded in the definition of states, the transition from S_I to S_J is Markovian even if the transition from symbol to symbol is non-Markovian.

In practice, the convergence of the topographical nature of the SSN is thoroughly examined using the information amount, i.e., the Shannon entropy for the residential probabilities of states in SSN, and the converged SSN is regarded mathematically as the minimal but most predictive model to capture all the statistical information of a given time series. Readers who are interested in the mathematical details can refer to the reviews.[23,33,39]

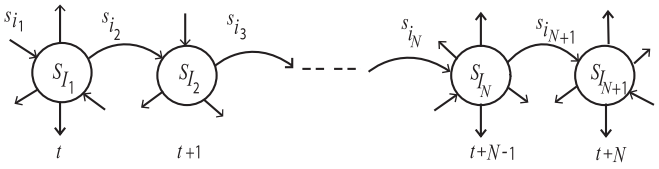

FIG. 1. A possible time series from the state S_{I_1} to $S_{I_{N+1}}$. Here the state S_{I_τ} is represented by a circle with the incoming symbol denoted by s_{i_τ} ($\tau = 1, 2, \ldots$) at time $t = \tau - 1$.

B. Derivation of time autocorrelation function

Using the properties of SSN, we can analytically solve for the autocorrelation function of the time series. Autocorrelation is the expectation value of two values at time t and $t + N$, $s(t)$ and $s(t + N)$. For the sake of brevity we denote these values at time t and $t + N$ as s_{i_1} and $s_{i_{N+1}}$ hereinafter. For this derivation, we first need to derive an analytic expression for a joint probability between s_{i_1} and $s_{i_{N+1}}$. The detailed procedure is as follows: Let us define the notations of symbols to be generated and the associated states denoted as in Fig. 1. By using a chain rule, the joint probability between the two symbols s_{i_1} at time t and s_{i_2} at time $t + 1$ becomes

$$\begin{aligned} P(s_{i_2}, s_{i_1}) &= \sum_{I_1} P(s_{i_2}, S_{I_1}, s_{i_1}) \\ &= \sum_{I_1} P(s_{i_2}|S_{I_1} s_{i_1}) P(S_{I_1}|s_{i_1}) P(s_{i_1}) \\ &= \sum_{I_1} P(s_{i_2}|S_{I_1}) P(S_{I_1}|s_{i_1}) P(s_{i_1}), \end{aligned} \tag{1}$$

where the last equality of Eq. (1) follows from the Markovian property of SSN. The conditional probability of s_{i_2}, given a state S_{I_1} and a symbol s_{i_1}, i.e., $P(s_{i_2}|S_{I_1}, s_{i_1})$, does not depend on the symbol s_{i_1} resulting in $P(s_{i_2}|S_{I_1}, s_{i_1}) = P(s_{i_2}|S_{I_1})$. In the same manner, the joint probability between s_{i_1} at time t and s_{i_3} at time $t + 2$ is given by

$$\begin{aligned} P(s_{i_3}, s_{i_1}) &= \sum_{I_2 i_2 I_1} P(s_{i_3}, S_{I_2}, s_{i_2}, S_{I_1}, s_{i_1}) \\ &= \sum_{I_2 i_2 I_1} P(s_{i_3}|S_{I_2}) P(S_{I_2} s_{i_2}|S_{I_1}) P(S_{I_1}|s_{i_1}) P(s_{i_1}) \\ &= \sum_{I_2 i_2 I_1} P(s_{i_3}|S_{I_2}) T^{(i_2)}_{I_2 I_1} P(S_{I_1}|s_{i_1}) P(s_{i_1}) \\ &= \sum_{I_2 I_1} P(s_{i_3}|S_{I_2}) \Big(\sum_{i_2} T^{(i_2)}_{I_2 I_1} \Big) P(S_{I_1}|s_{i_1}) P(s_{i_1}) \\ &= \sum_{I_2 I_1} P(s_{i_3}|S_{I_2}) \mathbf{T}_{I_2 I_1} P(S_{I_1}|s_{i_1}) P(s_{i_1}). \end{aligned} \tag{2}$$

Here the transition probability $P(S_{I_2} s_{i_2}|S_{I_1})$ is denoted by $T^{(i_2)}_{I_2 I_1}$ and, for the sake of simplicity, we introduce a notation $\mathbf{T}_{I_2 I_1} = \sum_{i_2} T^{(i_2)}_{I_2 I_1}$.

Finally we get the expression for the joint probability between s_{i_1} at time t and $s_{i_{N+1}}$ at time $(t + N)$ as follows:

$$\begin{aligned} P(s_{i_{N+1}}, s_{i_1}) &= \sum_{I_N, I_{N-1}, \ldots, I_1} P(s_{i_{N+1}}|S_{I_N}) \mathbf{T}_{I_N I_{N-1}} \mathbf{T}_{I_{N-1} I_{N-2}} \cdots \\ &\quad \mathbf{T}_{I_3 I_2} \mathbf{T}_{I_2 I_1} P(S_{I_1}|s_{i_1}) P(s_{i_1}) \\ &= \sum_{I_N I_1} P(s_{i_{N+1}}|S_{I_N}) (\mathbf{T}^{N-1})_{I_N I_1} P(S_{I_1}|s_{i_1}) P(s_{i_1}). \end{aligned} \tag{3}$$

We factorize this transition matrix $\mathbf{T}$ into $\mathbf{T} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$ where $\mathbf{Q}$ means the square matrix $(n \times n)$ (n is number of states in SSN) whose column is the eigenvector of the transition matrix, $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues of the transition matrix, and $\mathbf{Q}^{-1}$ is the inverse matrix of $\mathbf{Q}$, respectively. By using the factorization, Eq. (3) can be written as

$$\begin{aligned} &P(s_{i_{N+1}}, s_{i_1}) \\ &= \sum_{I_N I_1} P(s_{i_{N+1}}|S_{I_N}) \left[(\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1})^{N-1} \right]_{I_N I_1} P(S_{I_1}|s_{i_1}) P(s_{i_1}) \\ &= \sum_{I_N I_1} P(s_{i_{N+1}}|S_{I_N}) (\mathbf{Q}\mathbf{\Lambda}^{N-1}\mathbf{Q}^{-1})_{I_N I_1} P(S_{I_1}|s_{i_1}) P(s_{i_1}) \\ &= \sum_{C} \left[\sum_{I_N} P(s_{i_{N+1}}|S_{I_N}) \mathbf{Q}_{I_N C} \right] \lambda_C^{N-1} \\ &\quad \left[\sum_{I_1} \mathbf{Q}^{-1}_{C I_1} P(S_{I_1}|s_{i_1}) P(s_{i_1}) \right] \\ &= \sum_{C} A_C B_C \lambda_C^{N-1}, N \geq 1, \end{aligned} \tag{4}$$

where λ_C is the Cth diagonal element of $\mathbf{\Lambda}$ and $A_C = \sum_{I_N} P(s_{i_{N+1}}|S_{I_N}) \mathbf{Q}_{I_N C}$ and $B_C = \sum_{I_1} \mathbf{Q}^{-1}_{C I_1} P(S_{I_1}|s_{i_1}) P(s_{i_1})$. Since $\mathbf{T}$ is a probability matrix, it always has one eigenvalue equal to unity.

The timescale(s) of correlation (called lifetime(s) t_C) of the process can be calculated straightforwardly by

$$t_C = -\frac{1}{\log \lambda_C}. \tag{5}$$

In the signal process, autocorrelation function is often used without normalization, that is, without subtraction of the mean and division by the variance.[40] Since our SSN behaves as a stationary process at least for a timescale for which the SSN is constructed, we can define the autocorrelation function as the expectation value of the product of $s(t + \tau)$ and $s(t)$ without normalization, i.e.,

$$\begin{aligned} C(\tau) &= E[s(t + \tau) s(t)] \\ &= \sum_{i_{\tau+1} i_1} s_{i_{\tau+1}} s_{i_1} P(s_{i_{\tau+1}}, s_{i_1}) \\ &= \sum_{C} \sum_{i_{\tau+1} i_1} s_{i_{\tau+1}} s_{i_1} A_C B_C \lambda_C^{\tau-1} \\ &= \sum_{C} \sum_{i_{\tau+1} i_1} s_{i_{\tau+1}} s_{i_1} A_C B_C \exp\left[-\frac{\tau - 1}{t_C} \right]. \end{aligned} \tag{6}$$

C. Skipping step method

For the application of SSN to SM time series obtained experimentally, there are two related obstacles: one is a problem of insufficient sampling in constructing the transition probabilities in the SSN procedure, and the other is nonstationarity of the time series. The former problem may be inherent to SM measurements such as fluorescence resonance energy transfer measurement especially when the fluorophores (dye) used in the experiment suffer from photobleaching, which shortens the lifetime of the fluorophores and, thus, the length of the time series available.

The latter, nonstationarity problem occurs in two different situations. One is that a given time series is intrinsically nonstationary irrespective of the length of time series. The other is as follows: if all finite characteristic timescales of the system are sufficiently shorter than the length of the time series observed by a measurement, the time series in the experiment should be stationary at equilibrium. However, the time series is regarded as nonstationary even at equilibrium if there exists characteristic timescales of the system comparable or longer than the length of time series monitored. To analyze nonstationary time series, the original algorithm presented in Sec. II A cannot be applied straightforwardly because it is formulated for stationary time series.

To overcome these problems, a generalization of SSN was developed by a multiscale decomposition scheme based on discrete wavelet transforms.[21–23] In the scheme, first, the original nonstationary time series is decomposed into a set of time series at different timescales in a hierarchical manner. Time series that are much longer than their transition timescales are expected to be stationary. It was found for single molecule electron transfer experiment of the NADH:flavin oxidoreductase (Fre) complex that the time series constructed at each timescale shows stationary behavior for a time region shorter than the individual timescales.[21–23] The original SSN construction scheme is applied to the stationary time series components, and the set of SSNs constructed for the stationary time series components are combined to get a single SSN covering a wide range of the original time series.

Possible drawbacks for this scheme developed for continuous time series are: First, the result of the wavelet decomposition depends on the choice of wavelet basis function. For example, most wavelet basis functions result in the so-called Gibbs phenomenon[41] when applied to discrete time series. That is, a finite sum of the Fourier series has artificially large oscillations near a discontinuous jump, and a huge number of Fourier components is required to approximate the discontinuous jumps. (A similar situation should meet for most wavelet basis functions.) Second, in the wavelet decomposition, as timescales of wavelet components become longer, the number of "independent" samples at the long timescales becomes fewer (known as down sampling problem). One may obtain almost the same number of samples at all different timescales by shifting the time origin along the original time series. However, the more the timescale increases, the less the generated time series become independent. This is because the wavelet basis function operates on segments of time series which are overlapping despite the shifting of the time origin.

In turn, the SSM proposed in this article does not need to specify any basis function and is free from the Gibbs phenomenon. Every possible skipping step yields an independent skipped step time series, and, therefore, it is free from the downsampling problem. On the other hand, the properties of SSN (as a HMM) do not disappear even in nonconvergent SSN by using SSM. Regardless of the skipping step used and the convergence of the SSN with respect to the past subsequence length (L_{past}), the SSN remains minimal in a sense of state complexity of network (i.e., Shannon entropy of residential probabilities of states) and maximal in predictivity at least up to L_{past}. First, since different past subsequences with the same transition probabilities to the future are grouped to the same state, the state complexity attains its minimum for a given L_{past}. In other words, all other ways of grouping for the same L_{past} necessarily result in a higher state complexity. Second, this grouping of subsequences to construct the SSN together with its Markovian state-to-state transition probabilities preserves the joint probability of the observable sequences and, therefore, is as predictive as the situation where all details of the original past subsequences are kept up to the given L_{past}.[39] Therefore, SSNs constructed from the skipped step time series make it possible to capture kinetics with timescale corresponding to the skipping step. As the size of the skipping step increases, the resultant SSN is expected to describe slower kinetics. However, the skipped time series do not contain any data in between the skipping steps and this makes it difficult to relate the results of the SSM to some of experimentally detectable quantities such as dwell time distributions. (The mathematical derivation for dwell time distribution will be published elsewhere.)

Here we explain the SSM in the case of skipping step three (SK 3). For SK 3, the original time series is decomposed into three time series, each of which is constructed by sampling from the original time series every three steps. Let the original time series be $\mathbf{s} = \{s(t_1), s(t_2)\ldots, s(t_N)\}$ and let $N = 3m + 1$ (m is an integer), for simplicity. Then, the three time series are $\mathbf{s_1} = \{s(t_1), s(t_4)\ldots, s(t_N)\}$, $\mathbf{s_2} = \{s(t_2), s(t_5)\ldots, s(t_{N-2})\}$, and $\mathbf{s_3} = \{s(t_3), s(t_6)\ldots, s(t_{N-1})\}$, respectively. In Fig. 2 we show the decomposition of the original symbolic time series into three symbolic time series. Combining these three time series $\mathbf{s_1}$, $\mathbf{s_2}$, and $\mathbf{s_3}$, we can obtain the original time series $\mathbf{s}$.

The structure of SSN constructed from the skipped time series reflects dynamics at the timescale that corresponds to the skipping step. Time series of complex process can have various correlation timescales. If the time series involves (a) longer correlation timescale(s) than the skipped step, the transition probabilities along the skipped time series may depend on the subsequences (resampled every skipped step), reflecting their histories or memories. Contrastingly, if the time series involves (a) shorter correlation timescale(s) than the skipped step, the subsequences of the skipped time series are expected to have no memory and their future symbol distributions do not depend on these subsequences. These subsequences are grouped into one state in the SSN within

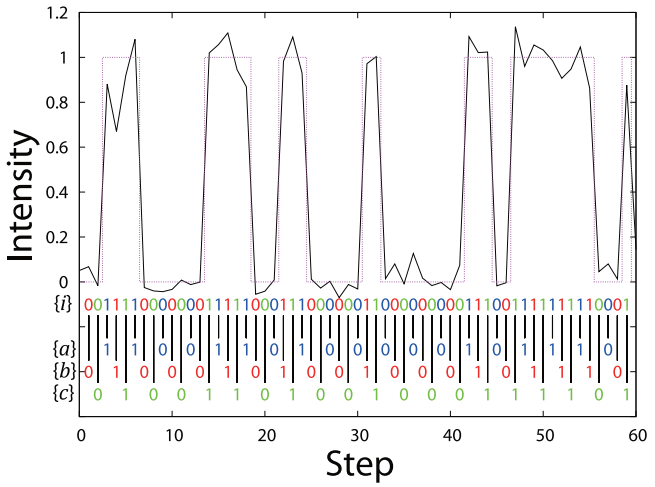

FIG. 2. An original symbolic time series $\{i\}$ colored with red, green, and blue. By skipping every three steps, the time series is decomposed into three distinct time series of a blue time series $\{a\}$, a red time series $\{b\}$, and a green time series $\{c\}$. The black solid and pink dashed lines denote the original time series and the discrete ON/OFF time series, respectively.

the skipped step resolution. Therefore, as the skipping step increases, the SSN tends to have fewer states (unless the system experiences a wider region on the underlying state space than the region scanned by the shorter skipping step). If the timescale of the skipping step becomes longer than the characteristic timescale of correlation, all the subsequences of the skipping step time series lose their memory and are merged into a single state.

The SSM is useful to examine long time memory. Suppose that we are restricted solely on the original time series (i.e., the time series of the skipping step one). If the topographical feature of SSNs constructed for the time series does not converge when L_{past} increases, it indicates that the time series has longer memory than L_{past}, which requires us to examine long time memory existing in the time series. However, when L_{past} increases, the number of the samples (the length of the time series) required to estimate the transition probabilities increases exponentially (the number of possible subsequences grows like $N_s^{L_{\text{past}}}$ where N_s denotes the number of symbols) and roughly L_{past} cannot exceed the logarithm of the length of the time series divided by the logarithm of the number of symbols N_s.[39] Therefore, the examination of long time memory by looking at the change in the topographical property of SSN constructed for the original time series with respect to L_{past} might become erroneous. Contrastingly, with the idea of skipping step, as the skipping step m increases, the number of samples to be required does not increase at all because, as an increase of SK m, the number of the decomposed time series increases as well, which compensates the lack of independent sampling moderately. Therefore, the SSM does not suffer from the lack of sampling and the scrutiny of the topographical properties of SSNs as a function of m makes it possible to examine long time memory buried in the time series.

III. RESULTS AND DISCUSSION

A. Application to recognition kinetics between EGFR and Grb2

1. A brief description of an in vitro reconstituted receptor-adaptor recognition experiment

We apply our multiscale decomposition scheme based on SSM to the association and dissociation kinetics between EGFR and Grb2. This interaction serves as a crucial step in signal processing in a live cell[16] (see also Fig. 3). Here we briefly describe the experimental setting. The plasma membrane fraction from epithelial carcinoma A431cells was immobilized to the coverslip, and Grb2 labeled with the fluorophore Cy3 were added into the solution. Morimatsu *et al.*[16] observed "intermittent pulses" repeatedly at the same positions on the glass surface by Electron Bombardment (EB) CCD camera equipped with a Micro Channel plate (MCP) image intensifier. The pulse arises from binding and release processes of the Cy3-Grb2 with EGFR localized at the cytoplasmic side of the membrane fragments. For the sake of simplicity, we abbreviate Cy3-Grb2 simply Grb2 hereinafter. The onset of EGFR-Grb2 association results in a fluorescent spot at the corresponding location on the camera, and conversely, the termination of fluorescence corresponds to dissociation. Movies of single molecule interactions between EGFR and Cy3-Grb2 were recorded at the video rate of 1/30 s^{-1} during 18 min within a given observation field. The effects of bleaching and blinking of Cy3 on the ON- and OFF-times were expected to be minimal because the decay time for bleaching was found to be 15 s and blinking occurred only once in 147 s under the same excitation conditions. These time constants were much longer than those of the ON-times.[16] Simultaneous binding or release of multiple spots at the same position on the glass surface was hardly detected, suggesting that EGFR exists in the monomer form. The SM time series are symbolized as a bound state by symbol "1" (fluorescent) and an unbound state by "0" (nonfluorescent) by introducing an appropriate threshold.[16] Single instantaneous transitions in one frame such as $1 \rightarrow 0 \rightarrow 1$ and $0 \rightarrow 1 \rightarrow 0$ were excluded from the analysis, because two frame averaging was applied to the raw images as a pretreatment in order to reduce shot noise.[16]

Single exponential kinetics were not observed between the bound and unbound forms, indicating that multiple states exist within both forms. In addition, the association rate does

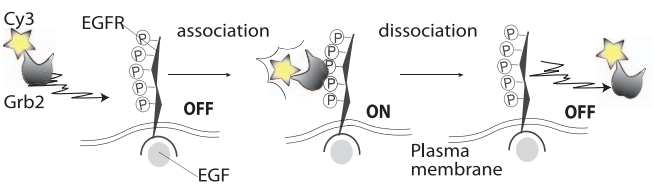

FIG. 3. A schematic picture of the *in vitro* reconstituted system of signal processing. The time duration of the binding between EGFR and Grb2 can be monitored from the duration of high fluorescence intensity from Cy3 attached to Grb2 detected using total internal reflection fluorescence microscope. These correspond to the interaction between Cy3-Grb2 and phosphorylated EGFR in the plasma membrane fragments attached to the glass coverslip.

not increase in proportion to the Grb2 concentration but increases gradually slower than the linear dependence. It was also shown that non-Markovianity exists in the ON/OFF time series of the recognition kinetics between EGFR and Grb2. The non-Markovianity could not be simply interpreted in terms of a multistate Markovian model whose unseen states were solely determined from the individual OFF-time and ON-time distributions.[42] It was thus conjectured that molecular memory exists in the conformational fluctuation of EGFR such that the EGFR conformation may change upon binding with a Grb2, and after the Grb2 dissociates from the EGFR, the EGFR conformation may need to relax to the unbound form that is favored for another (or same) Grb2 binding.

Y1068F mutant. EGFR has five major tyrosine residues that are actively phosphorylated after ligand binding. The Y1068F mutant of EGFR replaces tyrosine (Y) 1068 in EGFR (whose phosphorylation has been reported to construct the primary strong Grb2 binding site[43]) by phenylalanine (F) to prevent phosphorylation.[44] Exponential properties and memory effects in association and dissociation kinetics were analyzed for the Y1068F mutant.[16,42] The Y1068F mutant of EGFR showed non-exponential features in the dissociation kinetics (i.e., in the dwell time distribution of the bound form) at all concentrations, which was the same as the wild type EGFR.[16] However, the association kinetics (i.e., the dwell time distribution of the unbound form) at 1 nM Grb2 concentration showed that it is approximated by a single exponential kinetics for a wide range of timescales with the loss of correlation.[16,42]

B. Correlations in recognition kinetics and the underlying SSNs

To uncover the multiplicity of states and molecular memory in the process and its dependence on the mutation of the Y1068F mutant, we apply the SSN scheme combined with our SSM to the symbolized, binary SM time series of association and dissociation processes of the wide type EGFR and the Y1068F mutant at 1 nM concentration of Grb2. The structural property of the SSNs quantified by Shannon entropy of the states was found to show some locally convergent SSNs at L_{past} equal to 2 or 3 while the entropy continuously increases for a further increase of L_{past}. This local convergence indicates the existence of the timescale separation, i.e., the local convergence of SSN means that the dynamics faster than the timescale of $L_{\text{past}} = 2 - 3$ can be regarded as randomized and stationary such as being trapped in some energy basins. However, the longer timescale dynamics is considered to experience nonstationary basin-hopping processes among different basins. Hence, in this report, we chose $L_{\text{past}} = 3$ for all SSNs to be analyzed[50] with the skipping steps 1, 3, 9, and 27 (corresponding to 0.0333…(=1/30), 0.1, 0.3, and 0.9 s) so that L_{past}-times of a skipping step is equal to one increment of the next skipping step timescale.

Fig. 4 exemplifies the autocorrelation function of the symbolic time series at 1 nM concentration of Grb2 for the wild type EGFR and the Y1068F mutant. The observed autocorrelation function is satisfactorily reproduced by the analytical formula based on the obtained SSNs covering different time domains by different skipping steps. The corresponding time intervals or ranges the SSN can capture the autocorrelation, constructed at different skipping steps, are given at 1 nM concentration for the wild type EGFR and the Y1068F mutant in Table I. For example, at 1 nM concentration of Grb2 with the wild type EGFR, the constructed SSNs can reproduce the autocorrelation function at the timescale from 0 s to 0.133 s with the skipping step 1, abbreviated as SK 1 (in the unit/increment of 1/30 s), from 0.1 s to 0.3 s with SK 3 (0.1 s), 0.3 s to 0.9 s with SK 9 (0.3 s), and 0.9 s to 3.6 s with SK 27 (0.9 s). Remind that the history lengths automatically built in the SSN (i.e., the timescale of each skipping step multiplied by a factor of three ($L_{\text{past}} = 3$)) are 0.1 s, 0.3 s, 0.9 s, and 2.7 s at SK 1, SK 3, SK 9, and SK 27, respectively. This implies that the SSNs at SK 1 and SK 27 capture the autocorrelation beyond the timescale of the history length in the SSN, while those at SK 3 and SK 9 cannot.

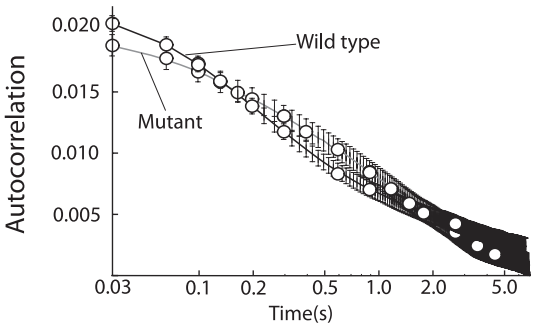

FIG. 4. Autocorrelation for the ON/OFF time series of the association and dissociation processes between the wild type and the Y1068F mutant with 1 nM concentration of Grb2. The autocorrelation function derived numerically from the time series is indicated by black line in the wild type and gray line in the mutant with the error bar denoted by vertical short lines. Those derived analytically in terms of the constructed SSNs are denoted by open circles.

In turn, as seen in Table I, compared to the SSNs of the wild type, the SSN of the Y1068F mutant tends to reproduce longer timescales' autocorrelation at each skipping step: for example, the SSNs with SK 27 can capture the correlation from 0.9 s to 4.5 s compared to those for the wild type (ca. 0.9–3.6 s).

Let us now look into the structure of the SSNs, especially the compositions of the ON and OFF states and their splitting as a function of the skipping step. Fig. 5 presents the corresponding SSN at each skipping step for the wild type EGFR and the Y1068F mutant at 1 nM concentration

TABLE I. The time intervals for which the autocorrelation is reproduced by each SSN constructed for each different skipping step at 1 nM concentration of Grb2 for the wild type and the Y1068F mutant EGFR. Note again that the increment of the time intervals is different with each other dependent on the SK m: 0.033 s, 0.1 s, 0.3 s, and 0.9 s for $m = 1$, 3, 9, and 27, respectively. The unit of time is in seconds.

EGFR	SK 1	SK 3	SK 9	SK 27
Wild type	0.0–0.133	0.1–0.3	0.3–0.9	0.9–3.6
Mutant	0.0–0.20	0.1–0.4	0.3–1.2	0.9–4.5

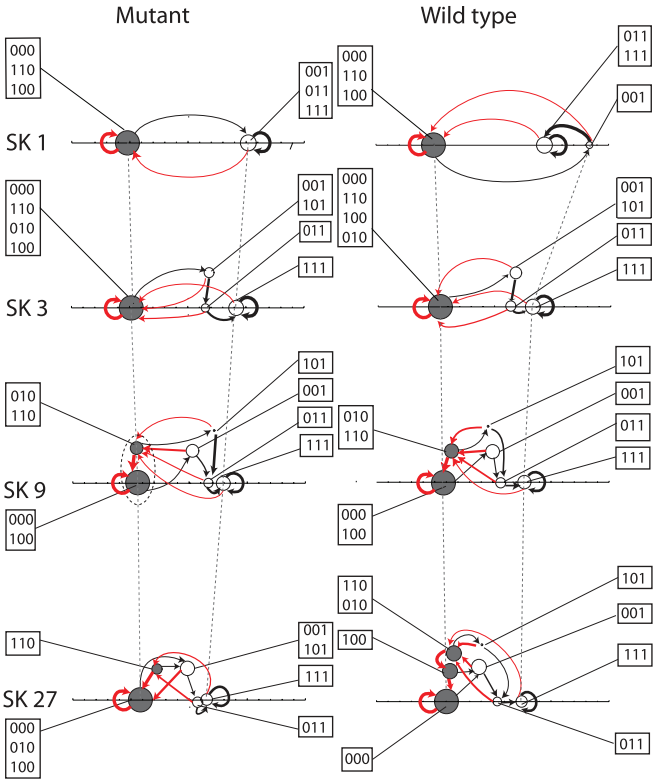

FIG. 5. The SSNs of the wild type and the Y1068F mutant at 1 nM concentration of Grb2 for different skipping steps. The horizontal axis reflects the mutual proximity of the transition probability distributions associated with the individual states in arbitrary unit (a.u.) (see the text in detail). The choice of vertical axis is arbitrary. Open (gray colored) circles denote the ON (OFF) states. The states enclosed by the dashed curve in SK 9 for mutant emphasize that their transition probabilities are almost identical. The size of the circle is proportional to the logarithm of the residential probability of the state: the bigger the circle is, the longer the system resides in that particular state (for visualization of the states whose area is less than 0.005 a.u.^{2}, we introduced the minimum size of 0.005 a.u.^{2}). The red (black) colored links assign as producing next symbol "0" ("1") (whose destination is either of the OFF (ON) states). The weight of the links reflects the extent of the state-to-state transition probabilities.

of Grb2. To visualize which states have mutually similar transition probability distributions, we embed states of *all* SSNs of both the wild type and the Y1068F mutant EGFR[45] into the one-dimensional Euclidean space (the horizontal axis): the more the states have mutually similar transition probability distributions, the closer they are located. To quantify the similarity between two transition probability distributions, we used the Hellinger distance[46] defined by $D[I, J] = \frac{1}{\sqrt{2}}\sqrt{\sum_{s_i}(\sqrt{P(s_i|S_I)} - \sqrt{P(s_i|S_J)})^2}$ where, e.g., $P(s_i|S_I)$ is the transition probability for producing a next symbol s_i from the state S_I. The position of each state along the vertical axis is chosen simply for visual clarity. The size of circle reflects the residential probability of the state; the larger the size of circle, the more the system resides in that state. Hereinafter, we call all states where the EGFR and Grb2 are currently bound (i.e., the (rightmost) last symbol appearing in a subsequence constituting the state is "1") *ON state* (indicated by white circles). Likewise, we call all states where they are currently unbound (i.e., the last symbol appearing in a subsequence constituting the state is "0") *OFF state* (gray circles). A state corresponds to a series of snapshots of conformations to end up either in the bound and unbound forms of the EGFR and Grb2.

The numbers of ON and OFF states, wiring pattern, residential and transition probabilities for the states are dependent on which timescale each SSN represents. It should be noted that this behavior is the same as observed in the analysis of single-molecule electron transfer experiment of the NADH:flavin oxidoreductase (Fre) complex[35] in which the topographical features of the SSN change as a function of timescale in order to recover the hierarchical diffusion property in the protein fluctuation.[21–23]

In both the wild type and the mutant, for SK 1 and SK 3, we have only one OFF state whose composite subsequences of length three are terminated by symbol "0." Note that since time series is binary with L_{past} being three, the maximum numbers of ON states and OFF states are four ($2^2 = 4$), respectively, i.e., eight in total. At SK 1, one can find that one OFF state and two ON states (the wild type) and one ON state (the Y1068F mutant) are sufficient to capture the correlation in this timescale. The existence of only one OFF state implies that irrespective of the paths to reach at the symbol "0," i.e., either $1 \to 1 \to 0$, $1 \to 0 \to 0$, or $0 \to 0 \to 0$, transition probabilities are the same (within a certain significance level). In the other terms, the transition probabilities only depend on the latest symbol "0." On the contrary, while only one ON state exists in the Y1068F mutant, the existence of two ON states in the wild type implies that the next symbol to be generated is dependent on the paths to arrive at the bound form "1" [i.e., either $(1 \to 1 \to 1, 0 \to 1 \to 1)$ or $0 \to 0 \to 1$]. Note that any consecutive paths or links that will end up with each state produce a series of the symbols "0" and "1" along the path (remind that each link has not only the transition probability but also the symbol to be generated), which coincides with the symbolic sequences consisting of the state. Namely, heterogeneous memory effects are encoded in the internal structure of the states, and equivalently in the topology of the SSN.

As an increase of skipping step to three, i.e., for a time series resampled every three steps over the original time series, the OFF state still persists as a single state but the ON state splits further. The number of the ON states reaches at the maximum number of four at the longer skipping timescales SK 9 and SK 27 in the wild type while it changes from four to three as an increase of the skipping timescales from SK 9 to SK 27 in the Y1068F mutant. In turn, the number of the OFF states does not reach the maximum number as the skipping steps increase from 1 to 27. It should be noted in Fig. 5 that, for the Y1068F mutant, the center of the two OFF states at SK 9 is almost identical although the two OFF states at SK 9 for the wild type are located at distinctively different positions (shown with a dashed closed curve in the figure). Namely, the SSN constructed for the Y1068F mutant is interpreted as effectively having a single OFF state up to the timescale of 1/30 s × 9 × 3 (=0.9 s), while it splits to two OFF states when the skipping step increases from 9 to 27.

Compared to the three-state model presented in the Appendix, for which the SSN tends to be simpler as the length of the skipping step increases (i.e., the longer the timescale, the smaller the number of the states), the state splittings in the OFF and ON states at longer timescales (at least, from SK 9

to SK 27) suggest that the multiple states are required to capture the kinetic complexity inherent to those timescales in the EGFR-Grb2 systems. This results from the multiscale, non-stationary nature of the ON/OFF time series of the recognition kinetics. This is interpreted as follows: for short timescale, the SSN (SK 1) structures in both the wild type and the mutant are simple since the conformation fluctuation of the EGFR is more likely to be confined within a single (super) basin on the energy landscape, making the system behave rather stationary and random. However, for the longer timescales (i.e., SK 9 and SK 27) the EGFRs can perform large conformational changes and, therefore, move between larger (super) basins on the energy landscape,[48] implying the emergence of more complex structure of the SSN in order to capture the complex kinetics.

However, it should also be noted worthy that, from the overall tendency of relative positions of the states at 1 nM concentration of Grb2, the locations of the states (i.e., the center of the circles) tend to merge all together as an increase of the skipping step from 1 to 27 in both the wild type and the Y1068F mutant. That is, the longer the skipping step the closer the states' transition probability distributions become, implying that the state transitions become less sensitive on the past (remind the three-state model in the Appendix). Moreover, as for the overall tendency, the network topology is also found to be simpler in the mutant than that in the wild type (i.e., the total number of states in each SSN is equal to or less than that of the wild type at each skipping step and the number of the OFF states does not reach at the maximum of $2^2 = 4$).

We also analyzed the lifetime constants of correlation (Eq. (5)) at different skipping steps SK 1, 3, 9, and 27 for the wild type and the Y1068F mutant of EGFR.[50] As for the overall tendency, as the skipping step increases (with longer timescale), the number of (largely) weighted components of lifetime constants increases more. However, the appearance of lifetime constants with comparable weights is different between the wild type and the Y1068F mutant. For example, at 1 nM concentration there exists only one major component (92%–100%) up to SK 27 in the mutant. On the contrary, while only one major component persists to exist up to SK 9 (86%–98%), it turns out to be diversified at SK 27 in the wild type. The diversification of the lifetime constants as the skipping step increases looks more apparent in the wild type than in the mutant.

C. Heterogeneity of non-Markovian property buried in SSNs

The most striking consequence of our method is that our method can capture heterogeneity of memory that depends on the states and, equivalently, on the topological nature of the complex network. Fig. 6 presents the corresponding SSNs of Grb2 for the wild type and the Y1068F mutant of EGFR with L_{past} =1, 2, 3, 4, and 5 for SK 1 at 1 nM concentration (the visualization scheme is the same as in Fig. 5). In this article, the memory refers to the degree of non-Markovianity of the time series, i.e., memory of the process, if it exists, is

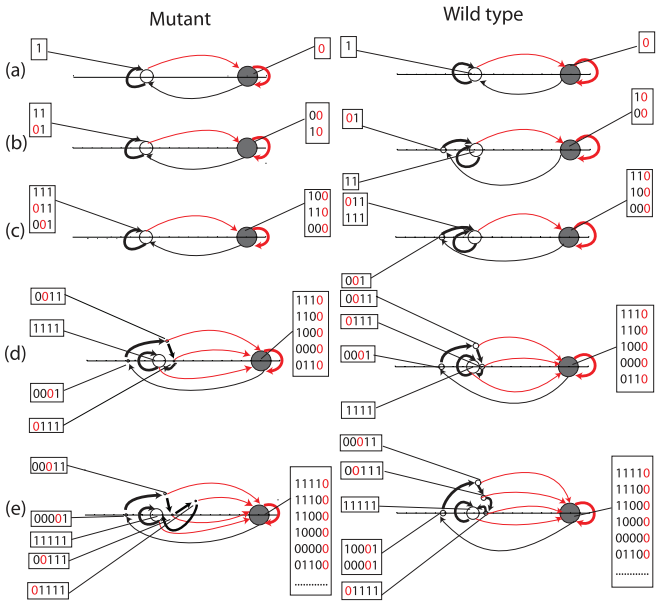

FIG. 6. The state splitting as increasing L_{past} at 1 nM for the recognition reaction between the wild type and the Y1068F mutant EGFR, and Grb2. (a) $L_{\text{past}} = 1$, (b) $L_{\text{past}} = 2$, (c) $L_{\text{past}} = 3$, (d) $L_{\text{past}} = 4$, and (e) $L_{\text{past}} = 5$. The meaning of state, links and their colors are the same as in Fig. 5.

manifest in the length of the optimal past sequences L_{past} used in defining the states of SSN.

Let us look into the change pattern of SSNs as an increase of L_{past}. At $L_{\text{past}} = 2$, the state composed of "10" and "00" subsequences is regarded as the OFF state (denoted by gray color) and the system currently visits the unbound form "0" (i.e., the rightmost symbol is 0). "10" implies that the system visited the bound form "1" at the previous step and makes a transition to the unbound form, while "00" implies that the system maintains to reside in the unbound form. Likewise, both the state composed of "11" and that composed of "01" are regarded as the ON state (i.e., the rightmost symbol is 1) while they are rather differentiated as distinct states in the wild type EGFR. However, in the Y1068F mutant of EGFR, "11" and "01" are regarded as the same ON state up to $L_{\text{past}} = 3$.

The topographical nature of the SSN with $L_{\text{past}} = 3$ is preserved with the SSN constructed with $L_{\text{past}} = 2$ (the SSN converges approximately in such a short timescale from 0.067 s to 0.1 s). However, as the value of L_{past} gets larger, i.e., 4 and 5, the SSN of the wild type and the mutant EGFR changes gradually in order to capture different kinetics emerging at longer timescales: while the OFF state persists as a single state for a timescale from 0 s to 0.167 s (=1/30 s × 5), some ON states split with increasing L_{past} while the others do not. For example, in the wild type the ON state (111,011) observed at $L_{\text{past}} = 3$ splits into three distinct states of (1111), (0111), (0011) at $L_{\text{past}} = 4$ but the ON state (001) at $L_{\text{past}} = 3$ does not split with longer L_{past}, corresponding to (0001) and (10001,00001) at $L_{\text{past}} = 4$ and 5, respectively. This splitting arises from the non-Markovian nature of the process. Hence, scrutiny of the splitting pattern can provide comprehensive understanding of molecular memory in the time domain where the SSNs are constructed. A simple inspection

tells us that, whenever the system once visits the unbound form "0" (the latest timings in the subsequences belonging to the individual states are indicated by red in Fig. 6), all the path information of how the system reaches at the unbound form is "reset" (i.e., the transition probability distributions become independent of the history once the system arrives at the unbound form "0"). When the EGFR-Grb2 system visits the unbound form, the history, or memory, is approximately reset in the original time series for a length of 0.167 s.

It should be noted, however, that the number of possible subsequences grows very rapidly ($\sim 2^{L_{\rm past}}$), which make the sampling statistics worse at larger $L_{\rm past}$. For example, at $L_{\rm past} = 27$, $2^{27} = 134, 217, 728$, the length of the the binary time series observed experimentally is 18 min, recorded every 1/30 s, yielding 32 727 data points. It is apparent that SSN analysis cannot be straightforwardly extended to large values of $L_{\rm past}$. This is the main reason why we introduce the skipping step algorithm combined with the original SSN scheme. In principle, if we could increase $L_{\rm past}$ large enough, we should end up with a Markovian network when the timescale of correlation of the process is finite. However the state of SSN may not be the underlying actual conformational state on the $(3n - 6)$ dimensional coordinate space (n is the number of all atoms). It is because (1) conformational dynamics on the $(3n - 6)$ dimensional coordinate space can, in general, associate memory or correlation, and (2) even if the conformational dynamics has no memory effect in high dimension and a Markovian network is yielded, states constructed from the projected one-dimensional time series cannot be reflected in full by all hidden $(3n - 6)$ dimensional coordinate space, but are expected to correspond to some effective conformational states that can be deduced from the projected time series.

Finally let us articulate the summary of the results of SSNs combined with SSM in relation to the previous experimental observation.[16]

1. It was found in Ref. 16 that the analyses of dwell time distributions (i.e., distributions of the time period for which the system dwells continuously at either 1 or 0) clarified that the ON and OFF states must contain multiple states, though the non-Markovian properties of the EGFR-Grb2 recognition kinetics of ON/OFF time series could not be simply described by the multiple states whose number was evaluated from the non-single exponential features of the dwell time distributions. In our current SSN scheme combined with the SSM, the extraction of the states is based not on the dwell time distribution but on the pattern buried in the time series, multiple states are naturally detected for the EGFR-Grb2 system, with encoding memory effects.
2. Our formula reproduces their autocorrelation function for a wide range of timescales up to about 3 s, and the topographical structure of the SSNs changes as an increase of the timescale: while the corresponding SSN is rather simple at the short timescale (0.033–0.1 s), the SSN at the longer timescales (0.1 s to $\sim$ 3 s) becomes complex for the elucidation of hierarchically organized kinetics appearing at the longer timescale in the wild type and the Y1068F mutant of EGFR. This is interpreted as being captured in some super basin to attain stationary behavior in the short timescale. One possible scenario to make the SSN more complex as an increase of the timescale is that partial interactions between the wild type EGFR and the Grb2 occurring faster than the experimental time resolution.[16] Such partial interactions possibly change the conformation of proteins, and conformational memory produced after dissociation can affect the reduction of the ON-rates, which leads to the structural change of SSNs.
3. It is found that, when the system once visits the unbound form of EGFR$\cdots$Grb2, the system approximately loses history or memory. This manifests the existence of heterogeneity of molecular memory, i.e., dependent on paths or states, the degree to what extent the system possesses the memory can be different. The interpretation is that the conformational dynamics of the unbound EGFR relax fast enough before the binding of the next Grb2. On the contrary, the bound states of the EGFR-Grb2 system keep splitting as $L_{\rm past}$ increases, suggesting slower relaxation dynamics for the bound EGFR-Grb2 complex. In the Y1068F mutant, the bound state does not split up to 0.133 s (i.e., $L_{\rm past} = 3$) in Fig. 6, indicating the dynamics of the bound state relax faster than those of the wild type.
4. In the experiments, the Y1068F mutant showed non-exponential features in the dissociation kinetics at all concentration as similarly as the wild type.[16] However, the association kinetics (=the dwell time distribution at the unbound form) at 1 nM concentration showed that it is approximated by a single exponential kinetics for a wide range of timescales.[16,42] The SSNs for the Y1068F mutant are interpreted to reflect such complexity of kinetics: the SSNs for the Y1068F mutant have effectively a single OFF state up to the timescale of $\sim$1.2 s at 1 nM (see Fig. 5).

IV. CONCLUSION AND PERSPECTIVES

In this article we have presented a novel scheme to extract the multiscale state space network that takes into account multiple nature of the states unseen in measurements and non-Markovianity of the process solely from SM time series. The crux is the combination of a nonlinear time series analysis recently developed on the basis of information theory[21–23,32,34,47] with the skipping step algorithm. We also derived the exact formula for the autocorrelation function of the symbolic time series. In this article, we demonstrated the potential of our theory by applying to the ON/OFF SM time series of the recognition kinetics between the wild type and the Y1068F mutant of EGFR and Grb2 for an *in vitro* reconstituted system at 1 nM concentration of Grb2.[16,42] Mathematically there is no difficulty to generalize our theory into non-binary time series having more than two symbols. Appropriate symbolization of time series depends on the experimental setting. In general, the observed time traces are contaminated by several sources of extrinsic and intrinsic noise.

It is of crucial importance to extract the actual time trace of a physical quantity (e.g., interdye distance) desired for the further analysis from a raw data (e.g., photon arrival times at the donor and acceptor channels).[37] Recently, it was shown that the existence of dynamic disorder of single enzymatic turnover reactions depends on how one assigns the ON/OFF levels and the most widely used binning and thresholding approach may yield a misleading interpretation.[38] Because some subsequences involving one step transitions in the ON/OFF time series were removed from the analyses,[16] it is desired to confirm how such removals would affect in the analyses based on change point detection method.[36,38]

It should also be noted that the SSN scheme groups a set of past subsequences (into a single state) whose transition probability distributions are regarded as identical within an error tolerance determined by the significance level. The grouping and resultant states depend on this significance level and even if two past subsequences are grouped into a certain state for a specific significance level, it might not be the case for another significance level. Therefore, a SSN constructed for a certain significance level might mislead the precise interpretation concerning the underlying SSN. Therefore, the visualization of the SSN projected onto the metric space that reflects the proximity of transition probability distribution is very crucial to capture the essence of the network structure.

In the literature, a state is often regarded as a conformation or, more generally, a set of conformations. However, direct connections between the high-dimensional protein conformations and states obtained from the SSN analysis may not be established solely based on the information of a one-dimensional time series. This limitation is inherent to most of all measurements, but we expect that a systematic survey of the dependence on several amino acid residue mutations provides more concrete identification of the state in relation to the identity of the role of amino acid residues.

The nonstationary features and the timescale dependence of the constructed networks are the natural consequence arising from the property of the system in question, whenever states are defined along the time series. One of the alternative data-driven approaches to construct a network is to utilize dwell-time time series, which is the series of the consecutive dwell times at the individual levels such as ON and OFF.[49] It is interesting to compare the results of these complimentary approaches as one of the forthcoming subjects. The analysis of higher concentration of Grb2 and the mathematical derivation of exact dwell time distribution for binary (or non-binary) time series will be published elsewhere.

ACKNOWLEDGMENTS

We thank Dr. J. Nick Taylor for his critical reading of this manuscript and his valuable comments. This work has been partially supported by Grant-in-Aid for Scientific Research(B), JSPS (to T.K.), Grant-in-Aid for Exploratory Research, JSPS (to T.K.), a Grant-in-Aid for Scientific Research on Innovative Areas "Spying minority in biological phenomena (No. 3306)," MEXT (to T.K.). The computations were partially performed using the Research Center for Computational Science, Okazaki, Japan.

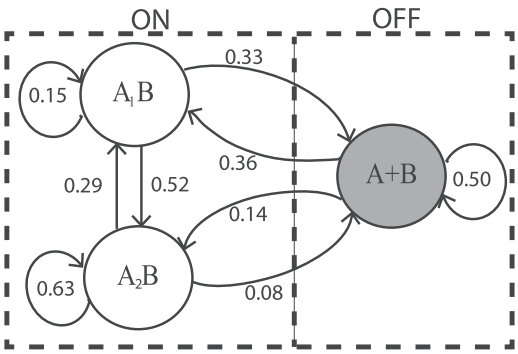

FIG. 7. A Markovian network model composed of three states in which two states belonging to "ON" are not visually distinguishable in the time series. The transition probabilities per unit time among these three states are denoted by the numbers associated with links.

APPENDIX: AN ILLUSTRATION OF OUR CONSTRUCTION SCHEME OF MULTISCALE SSNS FOR A THREE-STATE MARKOVIAN NETWORK

In this subsection we illustrate our method by using a simple model system that consists of a receptor protein A and a substrate B (see Fig. 7). Suppose that the protein A has two conformations A_1 and A_2, and when the protein A binds to the substrate B there exists two distinct bound forms denoted by A_1 B and A_2 B. The other state corresponds to the unbound form denoted by A+B. The former two bound states are assigned as ON states whereas the latter as an OFF state (indicated by gray circle in Fig. 7). For this model, with generating a binary time series of length 100 000 using Monte Carlo method (ON and OFF levels are represented by "1" and "0"), let us construct the underlying SSNs by using the procedure presented in Sec. II A. The SSNs converged at each skipping step with $L_{\text{past}} \leq 2$ are shown in Fig. 8, and the residential probabilities of the states of the SSN of skipping step one are given, with those of the three-state model, in Table II. One can see that the SSN at the skipping step one has the same number of states as the three-state model does: one corresponds to the OFF state whereas the other two correspond

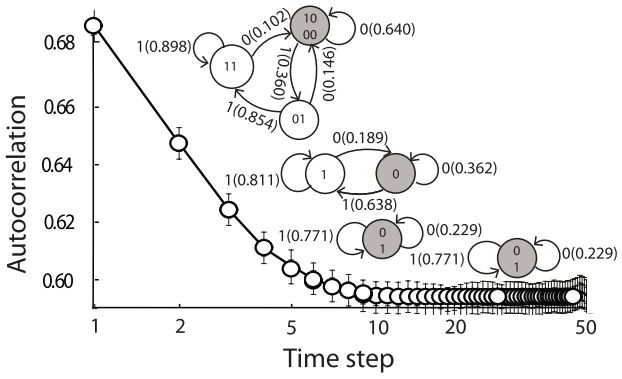

FIG. 8. The SSNs constructed at SK 1, SK 3, SK 9, and SK 27 shown from the left top to the right bottom for the three-state model. The autocorrelation function derived numerically from the time series is indicated by black line with the error bar denoted by vertical short lines. Those derived analytically in terms of the constructed SSNs are denoted by empty circles (the time range each SSN can reproduce is summarized in Table III). The shaded and empty circles denote the OFF and ON states, respectively.

TABLE II. The residential probabilities of the three-state model and the corresponding SSN. The SSN is constructed with the significance level $\alpha = 0.01$, $L_{\text{past}} = 2$, and the skipping step one. The state S_0 contains subsequences 11, the state S_1 01, and the state S_2 10 and 00, respectively.

	Three-state model	SSN
ON	$P(A_1B) = 0.20$	$P(S_0) = 0.69$
	$P(A_2B) = 0.57$	$P(S_1) = 0.08$
OFF	$P(A + B) = 0.23$	$P(S_2) = 0.23$

to the ON state although the residential probabilities in the ON states S_0 and S_1 are different from those of A_1B and A_2B. The reason for this discrepancy between the network structure of the model and that of the SSN is as follows: In general, depending on the topological features of the underlying network, e.g., some networks may have some redundancy, the underlying network may not be the simplest model to generate the time series, and therefore the converged SSN constructed is not necessarily the same to the underlying network. This is because the SSN scheme deals with only a time series, and is designed to construct the simplest but most predictive network by capturing all statistical and kinetic information buried in the one-dimensional time series. It is noted that the statistical complexity (which quantifies how complex the network model is) of the underlying network is larger than that of the constructed SSN for the three-state model, implying that the constructed SSN is indeed a simpler representation of the process (i.e., statistical complexity is 0.796 for the constructed SSN but 0.980 for the original three-state model).

Fig. 8 shows the SSNs constructed at skipping steps 1 (SK 1), 3 (SK 3), 9 (SK 9), and 27 (SK 27) with the autocorrelation computed numerically by the formula $C(\tau) = \frac{1}{N-\tau}\sum_{i=1}^{N-\tau} s_i s_{i+\tau}$ (where s_i and N are the value at time i and the total length of time series), and the comparison to the autocorrelation computed analytically in terms of the obtained SSNs. These different SSNs can reproduce autocorrelation function at different timescales corresponding to their skipping step. At SK 1, we have three states whereas at SK 3 there are two states. At nine steps, the autocorrelation almost converges to the asymptotic value $(\frac{1}{N}\sum_{i=1}^{N} s_i)^2$, implying that, with the increment of more than nine steps, the next outcome does not "remember" or "refer to" the current outcome. Hence, the SSN becomes the same as that from a simple coin toss at SK 9. The network constructed at SK 27 turns out to be the same as at SK 9, which means that the SSNs up to SK 9 are enough to capture all autocorrelation of the system.

TABLE III. The time region for which each SSN constructed for different skipping steps 1, 3, 9, and 27 reproduces the autocorrelation. Note that the increment of the intervals is different with each other dependent on the SK m, and the total step length is 100 000.

SK 1	SK 3	SK 9	SK 27
0–100, 000	3–99, 999	9–99, 999	27–99, 981

For comparison, we superimpose an autocorrelation function calculated directly from the ON/OFF time series on the autocorrelation function evaluated analytically by the SSNs at different skipping steps in Fig. 8. These two autocorrelation functions coincide within the error bar of the autocorrelation function. Table III presents the time intervals for which the autocorrelation is reproduced by each SSN constructed for each different skipping step. Note that for this three-state model the generated time series is perfectly stationary and the SSN at SK 1 can reproduce correlations at all time ranges. As for the further demonstration of the reproducibility of the SSNs on, e.g., higher order correlation, we showed the third-order correlation and the mutual information of the time series evaluated by the three-state toy model and the SSN.[50]

[1] A. A. Deniz, S. Mukhopadhyay, and A. E. Lemke, J. R. Soc., Interface **5**, 15–45 (2008).
[2] X. Michalet, S. Weiss, and M. Jager, Chem. Rev. **106**, 1785–1813 (2006).
[3] Y. Jai, D. S. Talaga, W. L. Lau, H. S. M. Lu, W. F. DeGrado, and R. M. Hochtrasser, Chem. Phys. **247**, 69–83 (1999).
[4] D. S. Talaga, W. L. Lau, H. Roder, J. Y. Tang, Y. W. Jia, W. F. DeGrado, and R. M. Hochstrasser, Proc. Natl. Acad. Sci. U.S.A. **97**, 13021–13026 (2000).
[5] D. S. Talaga, Y. Jia, M. A. Bopp, A. Sytnik, W. A. DeGrado, R. J. Cogdell, and R. M. Hochstrasser, Springer Ser. Chem. Phys. **67**, 313–325 (2001).
[6] E. Rhoades, E. Gussakovsky, and G. Haran, Proc. Natl. Acad. Sci. U.S.A. **100**, 3197–3202 (2003).
[7] E. Mei, J. Tang, J. M. Vanderkooi, and R. M. Hochstrasser, J. Am. Chem. Soc. **125**, 2730–2735 (2003).
[8] B. Schuler, ChemPhyChem **6**, 1206–1220 (2005).
[9] X. S. Xie and J. K. Trautman, Annu. Rev. Phys. Chem. **49**, 441–480 (1998).
[10] T. Ha, A. Y. Ting, J. Liang, W. B. Caldwell, A. A. Deniz, D. S. Chemla, P. G. Schultz, and S. Weiss, Proc. Natl. Acad. Sci. U.S.A. **96**, 893–898 (1999).
[11] L. Edman and R. Rigler, Proc. Natl. Acad. Sci. U.S.A. **97**, 8266–8271 (2000).
[12] N. J. Dovichi, R. Polakowski, A. Skelley, D. B. Craig, and J. Wong, Springer Ser. Chem. Phys. **67**, 241–256 (2001).
[13] X. Zhuang, L. E. Bartley, H. P. Babcock, R. Russell, T. Ha, D. Herschlag, and S. Chu, Science **288**, 2048–2051 (2000).
[14] L. Ying and X. S. Xie, J. Phys. Chem. B **102**, 10399–10409 (1998).
[15] P. J. Rothwell, S. Berger, O. Kensch, S. Felekyan, M. Antonik, B. M. Wohrl, T. Resle, R. S. Goody, and C. A. M. Seidel, Proc. Natl. Acad. Sci. U.S.A. **100**, 1655–1660 (2003).
[16] M. Morimatsu, H. Takagi, K. G. Ota, R. Iwamoto, T. Yanagida, and Y. Sako, Proc. Natl. Acad. Sci. U.S.A. **104**, 18013–18018 (2007).
[17] M. Andrec, R. M. Levy, and D. S. Talaga, J. Phys. Chem. A **107**, 7454–7464 (2003).
[18] T. C. Messina, H. Kim, J. T. Giurleo, and D. S. Talaga, J. Phys. Chem. B **110**, 16366–16376 (2006).
[19] S. A. McKinney, C. Joo, and T. Ha, Biophys. J. **91**, 1941–1951 (2006).
[20] L. R. Rabiner, Proc. IEEE **77**, 257–289 (1989).
[21] C. B. Li, H. Yang, and T. Komatsuzaki, Proc. Natl. Acad. Sci. U.S.A. **105**, 536–541 (2008).
[22] C. B. Li, H. Yang, and T. Komatsuzaki, J. Phys. Chem. B **113**, 14732–14741 (2009).
[23] C. B. Li and T. Komatsuzaki, *Cell signaling Reactions: Single-Molecular Kinetic Analysis* (Springer, London, UK, 2011), Chap. 11, pp. 221–263.
[24] I. V. Gopich and A. Szabo, J. Chem. Phys. **124**, 154712–154727 (2006).
[25] K. A. Merchant, R. B. Best, J. M. Louis, I. V. Gopich, and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **104**, 1528–1533 (2007).
[26] D. Colquhoun and A. G. Hawkes, Proc. R. Soc. London, Ser. B **211**, 205–235 (1981).
[27] I. V. Gopich and A. Szabo, J. Phys. Chem. B **107**, 5058–5063 (2003).
[28] A. Baba and T. Komatsuzaki, Proc. Natl. Acad. Sci. U.S.A. **104**, 19297–19302 (2007).
[29] O. Flomenbom and R. J. Silbey, Phys. Rev. E **76**, 041101-1–041101-5 (2007).
[30] B. English, W. Min, A. M. van Oijen, K. T. Lee, G. Luo, H. Sun, B. J. Cherayil, S. C. Kou, and X. S. Xie, Nat. Chem. Biol. **2**, 87–94 (2006).

[31]K. Kamagata, T. Kawaguchi, Y. Iwahashi, A. Baba, K. Fujimoto, T. Komatsuzaki, Y. Sambongi, Y. Goto, and S. Takahashi, J. Am. Chem. Soc. **134**, 11525–11532 (2012).
[32]J. P. Crutchfield, Nat. Phys. **8**, 17–24 (2012).
[33]C. R. Shalizi and J. P. Crutchfield, J. Stat. Phys. **104**, 817–881 (2001).
[34]J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105–108 (1989).
[35]H. Yang, G. Luo, P. Karnchanaphanurach, T. M. Louie, I. Rech, S. Cova, L. Xun, and X. S. Xie, Science **302**, 262–266 (2003).
[36]L. P. Watkins and H. Yang, Biophys. J. **86**, 4015–4029 (2004).
[37]L. P. Watkins, H. Chang, and H. Yang, J. Phys. Chem. A **110**, 5191–5203 (2006).
[38]T. Terentyeva, H. Engelkamp, A. Rowan, T. Komatsuzaki, J. Hofkens, C. B. Li, and K. Blank, ACS Nano **6**, 346–354 (2012).
[39]C. R. Shalizi, An algorithm for building Markov models from time series, 2003, see http://www.cscs.umich.edu/~crshalizi/CSSR/.
[40]P. F. Dunn, *Measurement and Data Analysis for Engineering and Science* (McGraw-Hill, New York, USA, 2005).
[41]S. E. Kelly, Appl. Comput. Harmon. Anal. **3**, 72–81 (1996).
[42]H. Takagi, M. Morimatsu, and Y. Sako, Adv. Chem. Phys. **146**, 195–215 (2012).
[43]J. Downward, P. Parker, and M. D. Waterfield, Nature (London) **311**, 483–485 (1984).
[44]R. Iwamoto, K. Hanada, and E. Mekada, J. Bio. Chem. **274**, 25906–25912 (1999).
[45]T. F. Cox and M. A. A. Cox, *Multidimensional Scaling* (Chapman and Hall, London, UK, 2001).
[46]H. H. Kuo, *Gaussian Measures on Banach Spaces* (Springer, Berlin, Germany, 1975).
[47]C. R. Shalizi, K. L. Klinkner, and J. P. Crutchfield, Technical Report, Santa Fe Institite, 2002; e-print arXiv:cs/0210025v3[cs: LG].
[48]Y. Matsunaga, K. S. Kostov, and T. Komatsuzaki, J. Phys. Chem. A **106**, 10898–10907 (2002).
[49]C. B. Li and T. Komatsuzaki, Phys. Rev. Lett. **111**, 058301 (2013).
[50]See supplementary material at http://dx.doi.org/10.1063/1.4848719 for the analysis of local convergence in constructing the SSN and that of the lifetime constants of the wild type and the Y1068 mutant EGFR at 1 nM concentration of Grb2 for each SSN and their constants (in Table I) which are calculated by Eq. (5). In addition, the third-order correlation function and the mutual information of the three-state toy model and the SSN are also given as for the comparison.