Assessing transfer entropy from biochemical data

Takuya Imaizumi,[1] Nobuhisa Umeki,[2] Ryo Yoshizawa,[2] Tomoyuki Obuchi,[3] Yasushi Sako,[2] and Yoshiyuki Kabashima[4,*]

[1]*Department of Mathematical and Computing Science, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan*

[2]*Cellular Informatics Laboratory, RIKEN Cluster for Pioneering Research, 2-1 Hirosawa, Wako 351-0198, Saitama, Japan*

[3]*Department of Systems Science, Kyoto University, 36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan*

[4]*Institute for Physics of Intelligence, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

(Received 8 November 2021; accepted 16 February 2022; published 8 March 2022)

We address the problem of evaluating the transfer entropy (TE) produced by biochemical reactions from experimentally measured data. Although these reactions are generally nonlinear and nonstationary processes making it challenging to achieve accurate modeling, Gaussian approximation can facilitate the TE assessment only by estimating covariance matrices using multiple data obtained from simultaneously measured time series representing the activation levels of biomolecules such as proteins. Nevertheless, the nonstationary nature of biochemical signals makes it difficult to theoretically assess the sampling distributions of TE, which are necessary for evaluating the statistical confidence and significance of the data-driven estimates. We resolve this difficulty by computationally assessing the sampling distributions using techniques from computational statistics. The computational methods are tested by using them in analyzing data generated from a theoretically tractable time-varying signal model, which leads to the development of a method to screen only statistically significant estimates. The usefulness of the developed method is examined by applying it to real biological data experimentally measured from the ERBB-RAS-MAPK system that superintends diverse cell fate decisions. A comparison between cells containing wild-type and mutant proteins exhibits a distinct difference in the time evolution of TE while any apparent difference is hardly found in average profiles of the raw signals. Such a comparison may help in unveiling important pathways of biochemical reactions.

DOI: 10.1103/PhysRevE.105.034403

I. INTRODUCTION

Cells and their processes are regulated by interactions with biomolecules via chemical reactions. Considerable effort has been made to properly understand the biomolecules and their interactions, which constitutes a tremendous amount of knowledge [1]. Particularly in the last two decades, chains or cascades of reactions have gained increasing attention owing to their potential role in understanding living things as *systems* [2]. Various reactions in cells are mathematically modeled using nonlinear differential equations and Monte Carlo simulations among others. In addition, large reaction pathways composed of tens or hundreds of components are drawn as graphs.

These research efforts have made great progress in understanding the mechanism of cell function control. However, there are still shortcomings. Although fundamental reactions are modeled by differential equations precisely, assessing their significance in the reaction cascade is nontrivial. An increase in the activation levels of enzymes indicates that certain functions are emerging. Nevertheless, one cannot quantify the significance of the reactions by only examining solutions of the differential equations as relevant activation levels can differ from enzyme to enzyme. In addition, biochemical reactions are dynamical processes, which indicates that the temporal alteration of signals is an important medium for transferring relevant information. Therefore, after the reaction pathways are plotted as graphs, there is a need to clarify the timing and how large the transmitted information is to properly control cell function.

For compensating these deficiencies, we focus on the *transfer entropy (TE)* [3,4]. TE is a universal measure of the causal relationship between two time series. It is defined by a functional of the joint distributions of two time series. This makes it possible to quantify causal significance in a unified manner, regardless of the physical mechanism that generates the time series, by representing the generation process in probabilistic models. Some empirical studies support its effectiveness in detecting and characterizing causal relations in complex systems [5,6]. TE has been used for analyzing information processing in various biological organisms from nervous systems [7–15] to gene regulatory networks [16–19] owing to its universality and effectiveness.

Generally, there are two ways to assess TE: *model-based* and *data-driven* approaches. In the model-based approach, the generation processes of the time series are modeled using high-dimensional joint distributions or their equivalent expressions, in which the TE is evaluated analytically or numerically. Although this approach can potentially enable the exact assessment of TE for given models, accurately modeling the generation process of actual systems is difficult, which practically limits its application range to the analysis of theoretical models [20]. Meanwhile, in the data-driven approach,

*Corresponding author: kaba@phys.s.u-tokyo.ac.jp

the TE is assessed directly from experimentally measured data, which avoids the difficulty in accurately modeling actual systems. However, the data-driven approach requires a sufficient number of simultaneously measured data to substitute the joint distribution of the objective time series as samples. Hence, technical constraints have restricted its application mostly to nervous systems, in which large-scale simultaneous measurement is possible and the required sample sizes are relatively small because of the stationary nature of the nervous systems [7–15]. However, the recent advancement and development of measurement techniques has greatly increased the ability of simultaneously measuring the activation levels of biomolecules with high space and time resolutions [21,22]. This may make it possible to assess TE of nonstationary biochemical interactions in cells from measured data.

In such a view of the current situation, we explore the possibilities and limitations for assessing TE from measured data of biochemical reactions. Generally, the assessment of TE is a computationally demanding task involving high-dimensional integrals. For overcoming this difficulty, we employ an approximation using Gaussian models, in which the assessment can be performed efficiently by estimating the covariance matrix of the time series. However, it is difficult to theoretically assess the sampling distributions of TE, which are indispensable for evaluating statistical confidence and significance of the data-driven estimates, due to the nonstationary nature of biochemical signals. We overcome this difficulty by developing a method to screen only statistically significant estimates based on techniques from computational statistics, the utility of which is tested by the application to a theoretically tractable time-varying signal model. In addition, we employ the method for analyzing data experimentally measured from the ERBB-RAS-MAPK system of actual cells. A distinct difference in the time alteration of TE is found in the comparison between cells containing wild-type and mutant proteins, whereas there is no apparent difference in the average profiles of raw time series. Such a comparison may be useful in identifying unknown pathways of biochemical reactions.

This paper is organized as follows: we review the definition of TE and how it can be computed under the Gaussian approximation in Sec. II. In addition, we present methods for assessing the statistical confidence and significance of the inferred results using techniques from computational statistics. In Sec. III the utility of the methods is examined by applying them to data from a theoretically tractable time-varying signal model. Based on the results obtained for the theoretically tractable model, we propose a method for screening only statistically significant estimates of TE. In Sec. IV real-world data from the ERBB-RAS-MAPK system are examined using the method. The final section presents the summary and discussion of the study.

II. ASSESSMENT OF THE TRANSFER ENTROPY FROM MEASURED DATA

A. Transfer entropy

Here we briefly review the definition of TE [3,4]. We use a conventional matrix-vector notation in which the bold type denotes column vectors and the uppercase type represents matrices or random variables. The symbol $\top$ indicates the matrix-vector transpose, such that the (column) vector $\boldsymbol{x} \in \mathbb{R}^n$ is represented as $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$. Given multiple vectors such as $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{z} \in \mathbb{R}^\ell$, etc., their concatenation is expressed as $\boldsymbol{x} \oplus \boldsymbol{y} \oplus \boldsymbol{z} \oplus \cdots = (x_1, \ldots, x_n, y_1, \ldots, y_m, z_1, \ldots, z_\ell, \ldots)^\top$. The notation $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denotes a Gaussian distribution, where $\boldsymbol{\mu}$ is the mean vector and Σ is the covariance matrix.

We suppose two interdependent discrete-time stochastic processes X_t and Y_t $(t = 0, 1, \ldots, T)$, and denote $\boldsymbol{X} \equiv (X_T, X_{T-1}, \ldots, X_0)^\top$ and $\boldsymbol{Y} \equiv (Y_T, Y_{T-1}, \ldots, Y_0)^\top$. We also use the notation $\boldsymbol{X}_t^{(p)} \equiv (X_t, X_{t-1}, \ldots, X_{t-p+1})^\top$ and $\boldsymbol{Y}_t^{(q)} \equiv (Y_t, Y_{t-1}, \ldots, Y_{t-q+1})^\top$. Let us assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ follow a joint distribution $Q(\boldsymbol{x}, \boldsymbol{y})$. Under this setup, the uncertainty of X_t is assessed by *(differential) entropy*

$$H(X_t) = -\int dx_t Q(x_t) \ln Q(x_t), \tag{1}$$

and that given $\boldsymbol{X}_{t-1}^{(p)}$ is quantified by *conditional entropy*

$$H\big(X_t|\boldsymbol{X}_{t-1}^{(p)}\big) = -\int d\boldsymbol{x}_{t-1}^{(p)}\, dx_t Q\big(\boldsymbol{x}_{t-1}^{(p)}\big) Q\big(x_t\big|\boldsymbol{x}_{t-1}^{(p)}\big) \ln Q\big(x_t\big|\boldsymbol{x}_{t-1}^{(p)}\big), \tag{2}$$

where $Q(x_t)$, $Q(\boldsymbol{x}_{t-1}^{(p)})$, and $Q(x_t|\boldsymbol{x}_{t-1}^{(p)})$ represent the marginal and conditional distributions with respect to relevant variables. All of these distributions are reduced from the joint distribution $Q(\boldsymbol{x}, \boldsymbol{y})$, and similar notations are employed hereafter without explanation. $H(X_t|\boldsymbol{X}_{t-1}^{(p)})$ means the remaining uncertainty of X_t on average given $\boldsymbol{X}_{t-1}^{(p)}$. Therefore, the amount of the reduction of the uncertainty

$$I\big(X_t; \boldsymbol{X}_{t-1}^{(p)}\big) = H(X_t) - H\big(X_t\big|\boldsymbol{X}_{t-1}^{(p)}\big), \tag{3}$$

which is referred to as *mutual information* between X_t and $\boldsymbol{X}_{t-1}^{(p)}$ and can be shown to be non-negative, is interpreted as the amount of information conveyed from the past p states $\boldsymbol{X}_{t-1}^{(p)} = (X_{t-1}, \ldots, X_{t-p})^\top$ to the current state X_t.

Extending the notion of mutual information, *transfer entropy (TE)* from Y to X at time t with lags q and p is defined as

$$\begin{aligned}\mathcal{T}_{Y\to X}^{(q),(p)}(t) &= I\big(X_t; \boldsymbol{X}_{t-1}^{(p)}, \boldsymbol{Y}_{t-1}^{(q)}\big) - I\big(X_t; \boldsymbol{X}_{t-1}^{(p)}\big) \\ &= H\big(X_t\big|\boldsymbol{X}_{t-1}^{(p)}\big) - H\big(X_t\big|\boldsymbol{X}_{t-1}^{(p)}, \boldsymbol{Y}_{t-1}^{(q)}\big)\end{aligned} \tag{4}$$

and from X to Y as

$$\begin{aligned}\mathcal{T}_{X\to Y}^{(p),(q)}(t) &= I\big(Y_t; \boldsymbol{X}_{t-1}^{(p)}, \boldsymbol{Y}_{t-1}^{(q)}\big) - I\big(Y_t; \boldsymbol{Y}_{t-1}^{(q)}\big) \\ &= H\big(Y_t\big|\boldsymbol{Y}_{t-1}^{(q)}\big) - H\big(Y_t\big|\boldsymbol{X}_{t-1}^{(p)}, \boldsymbol{Y}_{t-1}^{(q)}\big).\end{aligned} \tag{5}$$

Equation (4) indicates how much information is increased regarding X_t or how much uncertainty about X_t is reduced by providing $\boldsymbol{Y}_{t-1}^{(q)}$ on top of $\boldsymbol{X}_{t-1}^{(p)}$, and similarly for (5). These mean that TE stands for the significance of past states of one variable in predicting the current state of the other. In addition, the TE is asymmetric between X and Y unlike mutual information. Therefore, TE is employed as a useful measure of information flow that quantifies the causal relationship between two time series.

B. TE for Gaussian models

TE assessment is computationally demanding even though it is expressed in compact forms, such as (4) and (5). This is because the computational complexity grows exponentially with respect to $p+q+1$ when numerically evaluating (4) and (5), which can be infeasible even for the minimum lags of $p=q=1$ in practical situations. However, when $Q(\boldsymbol{x},\boldsymbol{y})$ is a multivariate Gaussian distribution, the assessment becomes feasible as entropy and conditional entropy can be analytically evaluated using covariance matrices for the Gaussian model.

For explaining this more precisely, we introduce the notation $\Sigma(\boldsymbol{U})$ to generally express the covariance matrix of the random vector $\boldsymbol{U}$. In addition, we use the notation $\Sigma(\boldsymbol{U},\boldsymbol{V})$ to denote the cross-covariance matrix between $\boldsymbol{U}$ and $\boldsymbol{V}$, which is composed of their cross-covariances $\mathrm{cov}(U_i,V_\alpha)$. These provide the *partial covariance* [4] of $\boldsymbol{U}$ given $\boldsymbol{V}\oplus\boldsymbol{W}\oplus\cdots$ as

$$\begin{aligned}&\Sigma(\boldsymbol{U}|\boldsymbol{V}\oplus\boldsymbol{W}\oplus\cdots)\\&\quad=\Sigma(\boldsymbol{U})-\Sigma(\boldsymbol{U},\boldsymbol{V}\oplus\boldsymbol{W}\oplus\cdots)\Sigma(\boldsymbol{V}\oplus\boldsymbol{W}\oplus\cdots)^{-1}.\\&\qquad\times\Sigma(\boldsymbol{U},\boldsymbol{V}\oplus\boldsymbol{W}\oplus\cdots)^{\top}.\end{aligned}\tag{6}$$

When $Q(\boldsymbol{x},\boldsymbol{y})$ is given as a multivariate Gaussian distribution, the properties of the Gaussian random variables yield a formula

$$H\big(X_t\big|\boldsymbol{X}_{t-1}^{(p)}\big)=\tfrac{1}{2}\ln\big(\Sigma\big(X_t\big|\boldsymbol{X}_{t-1}^{(p)}\big)\big)+\tfrac{1}{2}\ln(2\pi e),\tag{7}$$

where $\Sigma(X_t|\boldsymbol{X}_{t-1}^{(p)})=\Sigma(X_t)-\Sigma(X_t,\boldsymbol{X}_{t-1}^{(p)})\Sigma(\boldsymbol{x}_{t-1}^{(p)})^{-1}\Sigma(X_t,\boldsymbol{X}_{t-1}^{(p)})^{\top}$ is the variance of X_t conditioned by $\boldsymbol{X}_{t-1}^{(p)}$. The derivation of this formula is presented in the Appendix. Similarly, we also obtain

$$H\big(X_t\big|\boldsymbol{X}_{t-1}^{(p)},\boldsymbol{Y}_{t-1}^{(q)}\big)=\tfrac{1}{2}\ln\big(\Sigma\big(X_t\big|\boldsymbol{X}_{t-1}^{(p)}\oplus\boldsymbol{Y}_{t-1}^{(q)}\big)\big)+\tfrac{1}{2}\ln(2\pi e).\tag{8}$$

These formulas provide an expression of TE for the Gaussian time series as

$$\mathcal{T}_{Y\to X}^{(q),(p)}(t)=\frac{1}{2}\ln\left(\frac{\Sigma\big(X_t\big|\boldsymbol{X}_{t-1}^{(p)}\big)}{\Sigma\big(X_t\big|\boldsymbol{X}_{t-1}^{(p)}\oplus\boldsymbol{Y}_{t-1}^{(q)}\big)}\right),\tag{9}$$

and similarly for $\mathcal{T}_{X\to Y}^{(p),(q)}(t)$.

Several issues are to be noted with this formula. The first is about the implications of (9). This formula indicates that TE is determined using only covariances between the two time series irrespectively of their averages. This implies that the primary media of information transfer are not the average profiles of the observed signals but their statistical fluctuations, which may be counterintuitive. Nevertheless, in the framework of the Gaussian approximation, the covariances are determined by the Hessian of $-\ln Q(\boldsymbol{x},\boldsymbol{y})$ around the averages for general joint distributions $Q(\boldsymbol{x},\boldsymbol{y})$. This means that the average profile of signals is not a unique but still a major factor for determining TE. The second is about the cost for computation. Given the covariance matrix $\Sigma(X_t\oplus\boldsymbol{X}_{t-1}^{(p)}\oplus\boldsymbol{Y}_{t-1}^{(q)})$, the computational cost for assessing (9) increases as $O((p+q)^3)$ since the most computationally intensive part is the assessment of the matrix inversion $\Sigma(\boldsymbol{X}_{t-1}^{(p)}\oplus\boldsymbol{Y}_{t-1}^{(q)})^{-1}$. In most cases, this is computationally feasible as long as p and q are $O(1)$. The third issue is that, as mentioned in [4], TE is equivalent to the Granger causality [23], which has been extensively studied since the 1970s, for time series generated by multivariate autoregressive (MVAR) models. However, in the framework of the Granger causality, we must introduce many assumptions about how to describe the time series by MVAR models, which becomes nontrivial, particularly when handling nonstationary time series. In contrast, the TE framework is "model agnostic" [4] as (9) can be assessed directly from the covariances, for which we need few assumptions. Therefore, the formula of (9) would be more user-friendly when sufficient knowledge about the data generation process is not available. The final issue is concerning the validity of resorting to Gaussian models. The appropriateness of modeling time series as Gaussians may be criticized for specific physical generation processes. However, in most cases, assessing the TE from the exact formula of (4) and (5) requires significantly heavy computations for non-Gaussian models; therefore, Gaussian models are practically unique choices. Further, fortunately, it is known that even in non-Gaussian cases nonzero values of (9) imply that nonzero values of the true TE given by (4) [24]. For these reasons, we use the formula of (9) for assessing TE.

C. Assessment of TE from data and its statistical significance

Following the above argument, TE can be assessed from samples of time series $\mathcal{D}_M=\{\boldsymbol{x}_\mu,\boldsymbol{y}_\mu\}_{\mu=1}^M$ by substituting $\Sigma(X_t\oplus\boldsymbol{X}_{t-1}^{(p)}\oplus\boldsymbol{Y}_{t-1}^{(q)})$ with its estimate from $\mathcal{D}_M$ in evaluating (9). A natural estimator of $\Sigma(X_t\oplus\boldsymbol{X}_{t-1}^{(p)}\oplus\boldsymbol{Y}_{t-1}^{(q)})$ is

$$\begin{aligned}&\hat{\Sigma}\big(X_t\oplus\boldsymbol{X}_{t-1}^{(p)}\oplus\boldsymbol{Y}_{t-1}^{(q)}\big)\\&\quad=\frac{1}{M-1}\sum_{\mu=1}^{M}\big(\boldsymbol{z}_{t,\mu}^{(p),(q)}-\overline{\boldsymbol{z}_t^{(p),(q)}}\big)\big(\boldsymbol{z}_{t,\mu}^{(p),(q)}-\overline{\boldsymbol{z}_t^{(p),(q)}}\big)^{\top},\end{aligned}\tag{10}$$

where $\boldsymbol{z}_{t,\mu}^{(p),(q)}\equiv(x_{t,\mu},x_{t-1,\mu},\ldots,x_{t-p,\mu},y_{t-1,\mu},\ldots,y_{t-q,\mu})^{\top}$ represents the concatenation of μth samples $x_{t,\mu}$, $\boldsymbol{x}_{t-1,\mu}^{(p)}$, and $\boldsymbol{y}_{t-1,\mu}^{(q)}$, and $\overline{\boldsymbol{z}_t^{(p),(q)}}\equiv M^{-1}\sum_{\mu=1}^{M}\boldsymbol{z}_{t,\mu}^{(p),(q)}$, respectively. This is a consistent and unbiased estimator of $\Sigma(X_t\oplus\boldsymbol{X}_{t-1}^{(p)}\oplus\boldsymbol{Y}_{t-1}^{(q)})$, which guarantees that (10) converges to the true covariance as $M\to\infty$ and the average of (10) with respect to the generation of $\mathcal{D}_M$ accords with it for finite M. The assessment of (9) using (10) also offers a consistent estimator of $\mathcal{T}_{Y\to X}^{(q),(p)}(t)$ under Gaussian assumptions. However, this is generally biased. This is natural because $\mathcal{T}_{Y\to X}^{(q),(p)}(t)$ is a non-negative quantity by nature. Therefore, even if X_t and $\boldsymbol{Y}_{t-1}^{(q)}$ are statistically independent, yielding $\mathcal{T}_{Y\to X}^{(q),(p)}(t)=0$, the statistical fluctuations in (10) always results in positive values for the estimates of $\mathcal{T}_{Y\to X}^{(q),(p)}(t)$. In addition, the convergence rate of the covariance matrix estimator of (10) is rather slow [25]. This makes it challenging to analytically evaluate the sampling distributions of the estimates of TE, which are indispensable for assessing the statistical confidence and significance of the inferred results, although it is known that the maximum likelihood estimator of TE will asymptotically have a χ^2 distribution under the

null hypothesis in which the true TE vanishes, in the case of stationary time series [26,27].

For practically overcoming this difficulty, we employ computational methods known as *bootstrapping* [28]. We construct the sampling distributions in the following ways depending on the purpose.

a. For the confidence interval: Suppose that an estimate $\hat{\mathcal{T}}^{(q),(p)}_{Y\to X}(t)$ of (9) is evaluated for a given data set $\mathcal{D}_M$ by substituting $\Sigma(X_t \oplus \boldsymbol{X}^{(p)}_{t-1} \oplus \boldsymbol{Y}^{(q)}_{t-1})$ with $\hat{\Sigma}(X_t \oplus \boldsymbol{X}^{(p)}_{t-1} \oplus \boldsymbol{Y}^{(q)}_{t-1})$ of (10). We need to evaluate the degree of statistical fluctuations in $\hat{\mathcal{T}}^{(q),(p)}_{Y\to X}(t)$, which is inevitable owing to the finiteness of the sample size M. For this purpose, we handle $\hat{\Sigma}(X_t \oplus \boldsymbol{X}^{(p)}_{t-1} \oplus \boldsymbol{Y}^{(q)}_{t-1})$ as if it is the true covariance matrix and generate a new sample set of size M by independently drawing samples from an identical distribution $\mathcal{N}(0, \hat{\Sigma}(X_t \oplus \boldsymbol{X}^{(p)}_{t-1} \oplus \boldsymbol{Y}^{(q)}_{t-1}))$. This provides a surrogate estimate of TE $\hat{\mathcal{T}}^{*(q),(p)}_{Y\to X}(t)$ using $\hat{\Sigma}^*(X_t \oplus \boldsymbol{X}^{(p)}_{t-1} \oplus \boldsymbol{Y}^{(q)}_{t-1})$, which is the covariance matrix estimated from the new sample set. We repeat these procedures for $B(\gg 1)$ times, which results in an empirical distribution of $\hat{\mathcal{T}}^{*(q),(p)}_{Y\to X}(t)$. We construct the $100(1-\alpha)\%$ $(0<\alpha<1)$ confidence interval (CI) by specifying 50α and $100-50\alpha$ percentile points of the empirical distribution to evaluate the degree of statistical fluctuation of the estimate. The same procedure is performed for $\hat{\mathcal{T}}^{*(p),(q)}_{X\to Y}(t)$ as well.

b. For significance threshold: When estimating $\hat{\mathcal{T}}^{(q),(p)}_{Y\to X}(t)$, it always takes a non-negative value even when the true TE $\mathcal{T}^{(q),(p)}_{Y\to X}(t)$ vanishes. We need to evaluate the distribution of $\hat{\mathcal{T}}^{(q),(p)}_{Y\to X}(t)$ for the null hypothesis $H_0 : \mathcal{T}^{(q),(p)}_{Y\to X}(t) = 0$ to distinguish the obtained value from those of chance levels. To construct the distribution, we define the covariance matrix $\Sigma_{H_0}(X_t \oplus \boldsymbol{X}^{(p)}_{t-1} \oplus \boldsymbol{Y}^{(q)}_{t-1})$ for H_0 by making all elements of the cross-covariance between $X_t \oplus \boldsymbol{X}^{(p)}_{t-1}$ and $\boldsymbol{Y}^{(q)}_{t-1}$ vanish in $\hat{\Sigma}(X_t \oplus \boldsymbol{X}^{(p)}_{t-1} \oplus \boldsymbol{Y}^{(q)}_{t-1})$, keeping the other elements fixed. Based on this, we generate a new sample set of size M by independently drawing samples from an identical distribution $\mathcal{N}(0, \Sigma_{H_0}(X_t \oplus \boldsymbol{X}^{(p)}_{t-1} \oplus \boldsymbol{Y}^{(q)}_{t-1}))$. This provides an estimate of TE $\hat{\mathcal{T}}^{\#(q),(p)}_{Y\to X}(t)$ using $\hat{\Sigma}^{\#}(X_t \oplus \boldsymbol{X}^{(p)}_{t-1} \oplus \boldsymbol{Y}^{(q)}_{t-1})$, which is the covariance matrix estimated from the new sample set. We repeat these procedures for $B(\gg 1)$ times, which provides an empirical distribution of $\hat{\mathcal{T}}^{\#(q),(p)}_{Y\to X}(t)$. The $100(1-\alpha)$ $(0<\alpha<1)$ percentile point of the empirical distribution is used as the significance threshold (ST) to distinguish the estimated value from those of chance levels with the significance level of $100\alpha\%$. The same procedure is performed for $\hat{\mathcal{T}}^{\#(p),(q)}_{X\to Y}(t)$ as well.

These methods are very simple in terms of technical aspects. The necessary procedure is just repeating the evaluation of the TE from resampled data many times. The computational burden for repetition can prevent the execution of these methods in cases where a considerable amount of computation is required in obtaining a single estimate. However, in the current case, the computational cost for assessing TE is only $O((p+q)^3)$ under the Gaussian approximation, which would not be an obstacle for the execution. Nevertheless, the validity of using not the true distribution but the empirical distribution provided by the estimated covariance matrices for assessing the sampling distributions, following the *plug-in principle* [28], is debatable. In the next section, we examine this issue by applying the methods to a time-varying signal model, in which the theoretical evaluation of TE is tractable.

III. TESTING THE DEVELOPED METHOD USING THE THEORETICALLY TRACTABLE MODEL

As a simple but nontrivial example for which the theoretical assessment of TE is possible, we consider a d-dimensional time-varying state space model [29], which is expressed as

$$\boldsymbol{u}_t = F_t \boldsymbol{u}_{t-1} + \boldsymbol{\mu}_t + \boldsymbol{\xi}^{\mathrm{U}}_t, \tag{11}$$

$$\boldsymbol{v}_t = G\boldsymbol{u}_t + \boldsymbol{\xi}^{\mathrm{V}}_t, \tag{12}$$

where $t = 1, \ldots, T$, and $\boldsymbol{u}_t \in \mathbb{R}^{d(\geqslant 2)}$ and $\boldsymbol{v}_t \in \mathbb{R}^2$, respectively. $\boldsymbol{\xi}^{\mathrm{U}}_t \in \mathbb{R}^d$ and $\boldsymbol{\xi}^{\mathrm{V}}_t \in \mathbb{R}^2$ are the Gaussian noise vectors, both of which are independent in time. Equation (11) describes how the state vector $\boldsymbol{U}_t$ evolves in time, being subject to time-varying parameters $F_t \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\mu}_t \in \mathbb{R}^d$. Meanwhile, (12) defines the measurement process of $\boldsymbol{U}_t$. We consider that the first and last components of $\boldsymbol{U}_t$ are measured, such that the components of the resulting vector $\boldsymbol{V}_t$ are regarded as X_t and Y_t in Sec. II under the setup of the measurement matrix $G = \left(\begin{smallmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{smallmatrix}\right)$.

Two features are noted for (11) and (12):

(1) Given a set of the time-varying parameters F_t and $\boldsymbol{\mu}_t$, the resulting time series $\boldsymbol{u}_T, \ldots, \boldsymbol{u}_1$ and $\boldsymbol{v}_T, \ldots, \boldsymbol{v}_1$ are provided as linear combinations of $\boldsymbol{\xi}^{\mathrm{U}}_T, \ldots, \boldsymbol{\xi}^{\mathrm{U}}_1$ and $\boldsymbol{\xi}^{\mathrm{V}}_T, \ldots, \boldsymbol{\xi}^{\mathrm{V}}_1$ added to deterministic time series for a given initial state $\boldsymbol{u}_0$. This guarantees that the set of time series $\boldsymbol{U} = \{\boldsymbol{U}_T, \ldots, \boldsymbol{U}_1\}$ and $\boldsymbol{V} = \{\boldsymbol{V}_T, \ldots, \boldsymbol{V}_1\}$ follows a joint Gaussian distribution that varies over time.

(2) The mean parameters $\boldsymbol{\mu}_T, \ldots, \boldsymbol{\mu}_1$ do not contribute to statistical fluctuations of $\boldsymbol{U}$. This indicates that covariances of any components of $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent of the mean parameters.

These features enable us to evaluate $\Sigma(X_t \oplus \boldsymbol{X}^{(p)}_{t-1} \oplus \boldsymbol{Y}^{(q)}_{t-1})$ and $\Sigma(Y_t \oplus \boldsymbol{X}^{(p)}_{t-1} \oplus \boldsymbol{Y}^{(q)}_{t-1})$ efficiently using recursive equations, which results in the theoretical assessment of TE. More specifically, the statistical independence of noise vector $\boldsymbol{\xi}^{\mathrm{U}}_t$ regarding time provides a recursive equation for computing covariance matrix $C^{\mathrm{U}}_t = \mathbb{E}[\boldsymbol{U}_t\boldsymbol{U}_t^\top] - \mathbb{E}[\boldsymbol{U}_t]\mathbb{E}[\boldsymbol{U}_t]^\top$, which is expressed as

$$C^{\mathrm{U}}_{t+1} = F_t C^{\mathrm{U}}_t F_t^\top + \Delta^{\mathrm{U}}_t, \tag{13}$$

where $\mathbb{E}[\cdots]$ is the average with respect to noise vectors and Δ^{U}_t is the covariance matrix of $\boldsymbol{\xi}^{\mathrm{U}}_t$. In addition, the covariance matrix of $\boldsymbol{U}_t$ between two different times is given as

$$D^{\mathrm{U}}_{t+m,t} = \left(\prod_{\tau=t+1}^{t+m} F_\tau\right) C^{\mathrm{U}}_t, \tag{14}$$

where $m = 1, \ldots, T-t$ and $D^{\mathrm{U}}_{t+m,t} = \mathbb{E}[\boldsymbol{U}_{t+m}\boldsymbol{U}_t^\top] - \mathbb{E}[\boldsymbol{U}_{t+m}]\mathbb{E}[\boldsymbol{U}_t]^\top$. Subsequently, the covariance matrices of $\boldsymbol{V}_t$ are computed as

$$C^{\mathrm{V}}_t = GC^{\mathrm{U}}_t G^\top + \Delta^{\mathrm{V}}_t, \tag{15}$$

$$D^{\mathrm{V}}_{t+m,t} = GD^{\mathrm{U}}_{t+m,t} G^\top, \tag{16}$$

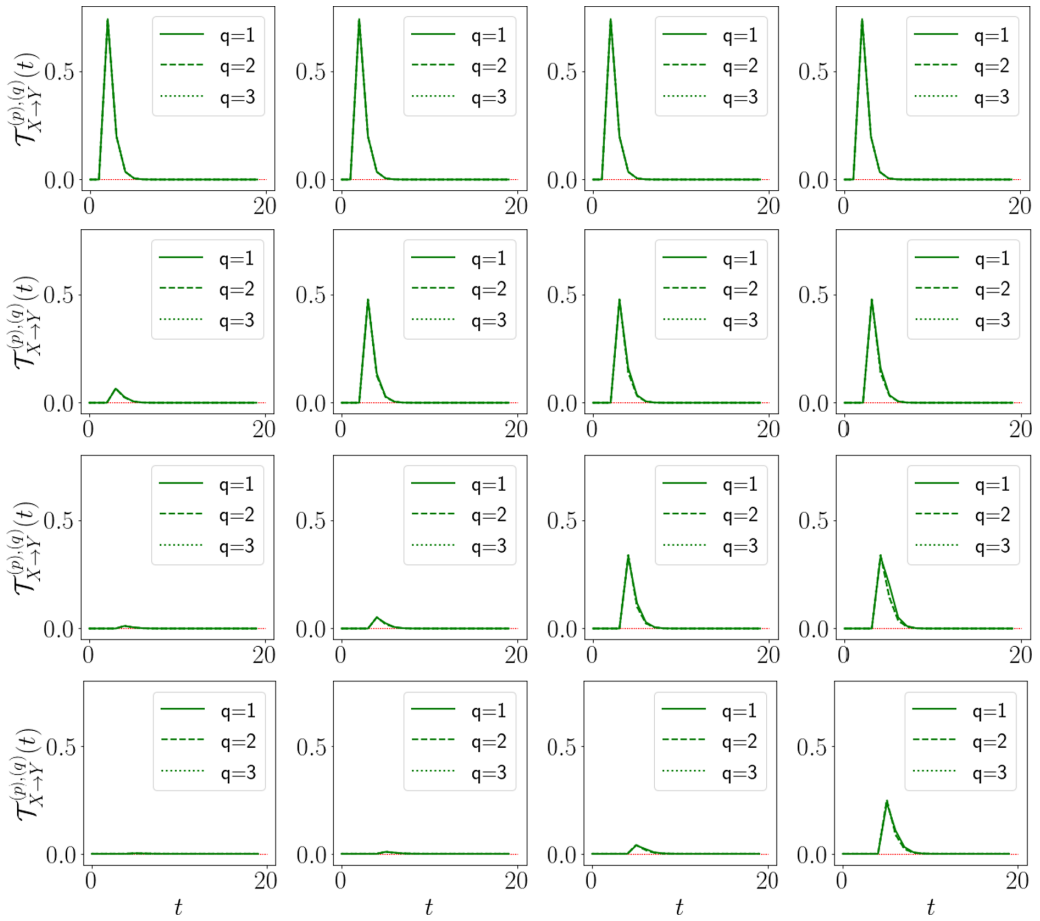

FIG. 1. Theoretically computed $\mathcal{T}^{(p),(q)}_{X\to Y}(t)$ for the system characterized by (17). From top to bottom: The cascade length d varies from 2 to 5. From left to right: The lag p of the cause time series X_t changes from 1 to 4.

where $C^{\mathrm{V}}_t = \mathbb{E}[\boldsymbol{V}_t \boldsymbol{V}_t^\top] - \mathbb{E}[\boldsymbol{V}_t]\mathbb{E}[\boldsymbol{V}_t]^\top$, $D^{\mathrm{V}}_{t+m,t} = \mathbb{E}[\boldsymbol{V}_{t+m}\boldsymbol{V}_t^\top] - \mathbb{E}[\boldsymbol{V}_{t+m}]\mathbb{E}[\boldsymbol{V}_t]^\top$, and Δ^{V}_t is the covariance matrix of $\boldsymbol{\xi}^{\mathrm{V}}_t$. The recursive Eqs. (13) and (14), in conjunction with (15) and (16), offer all components that constitute $\Sigma(\boldsymbol{X} \oplus \boldsymbol{Y})$ with $O(T^2)$ computational cost. Performing this is feasible as long as p and q are $O(1)$. Picking up all components of $\Sigma(X_t \oplus \boldsymbol{X}^{(p)}_{t-1} \oplus \boldsymbol{Y}^{(q)}_{t-1})$ from $\Sigma(\boldsymbol{X} \oplus \boldsymbol{Y})$ and substituting them into (9) provide $\mathcal{T}^{(q),(p)}_{Y\to X}(t)$, and similarly for $\mathcal{T}^{(p),(q)}_{X\to Y}(t)$.

For simulating time series that are simultaneously measured from a cascade of reactions, we consider a very simple model, which is given by

$$F_t = \begin{pmatrix} \alpha & 0 & \dots & \dots & 0 \\ F_{21}(t) & \alpha & 0 & \dots & 0 \\ 0 & \beta & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \alpha & 0 \\ 0 & 0 & \ddots & \beta & \alpha \end{pmatrix}, \tag{17}$$

where $0 < \alpha < 1$ is the decay rate of each component and β is the activation caused by the reaction with the component of the previous subscript. We assume that the first component of $\boldsymbol{U}$ triggers the reaction cascade in a short duration $\gamma > 0$, which is taken into account by the time-varying matrix element

$$F_{21}(t) = \begin{cases} \delta \exp(-t/\gamma), & t \geqslant 0 \\ 0, & t < 0 \end{cases}. \tag{18}$$

Indeed, we need to employ nonlinear equations with respect to $\boldsymbol{u}_t$ to precisely describe realistic cascades of chemical reactions. However, the media for information transfer in the framework of the Gaussian treatment is the statistical fluctuations of the state vectors around their averages, which justifies the current linearized description at least as a first approximation. In addition, the uniform setting of the parameters is not crucial when examining how the length of the reaction cascade d is reflected in the dependence of TE on the lag parameters p and q, which we will focus on in the following discussions.

We set $\alpha = 0.5$ and $\beta = 1$ to ensure that the influence of the trigger of the first component propagates forward without decay. Meanwhile, we set $\delta = 5$ and $\gamma = 1$, $\Delta^{\mathrm{U}}_t = 0.1^2 \times I_{d\times d}$, and $\Delta^{\mathrm{V}}_t = 0$ for other parameters, where $I_{n\times n}$ generally denotes $n \times n$ identity matrix. Figure 1 presents the plot of $\mathcal{T}^{(p),(q)}_{X\to Y}(t)$ computed from (13)–(15) for various pairs of lag parameters p and q by varying the length of the reaction cascade d from 2 to 5. The TE of the reverse direction can also be computed, but it is not presented because the statistical independence between the first component of $\boldsymbol{U}_t$ and the sub-

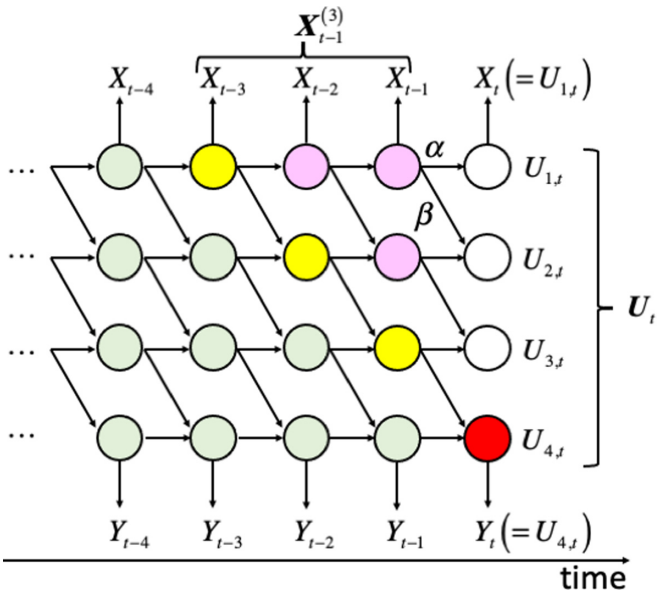

FIG. 2. Computational graph of (17) in the case of $d = 4$.

sequent components in the past state $\boldsymbol{U}_{t-m}$ $(m \geqslant 1)$ guarantees that $\mathcal{T}_{Y\to X}^{(q),(p)}(t)$ vanishes trivially. All the plots indicate that $\mathcal{T}_{X\to Y}^{(p),(q)}(t)$ does not significantly depend on the lag q of the *effect* time series Y_t. This is presumably because in the current setup, $\Delta_t^{\mathrm{V}} = 0$ is set to zero for a relatively small α. When Δ_t^{V} is set finite, the dependence of the TE on the lag q of the effect time series varies in a nontrivial manner depending on the value of α.

On the other hand, $\mathcal{T}_{X\to Y}^{(p),(q)}(t)$ monotonically increases as the lag p of the *cause* time series X_t increases, and almost saturates at $p = d - 1$. This is because the influence of the first component $U_{1,t}(=X_t)$ of $\boldsymbol{U}_t$ reaches the last component $U_{d,t}(=Y_t)$ late by the lag of $d - 1$ due to to the one-dimensional nature of the reaction cascade, as shown in the computational graph in Fig. 2. This graph indicates that the prediction accuracy of Y_t (red node) is maximized when $U_{d-1,t-1}$ is given in addition to Y_{t-1}. Unfortunately, $U_{d-1,t-1}$ is a hidden variable and cannot be directly observed. However, it is correlated with past states $X_{t-1}(=U_{1,t-1}), X_{t-2}(=U_{1,t-2}), \ldots$ as they share the same ancestors, and the strength of the correlation is maximized by X_{t-d+1} as it is connected to $U_{d-1,t-1}$ without decaying through the path of the sequence of yellow nodes in the graph. Therefore, $\mathcal{T}_{X\to Y}^{(p),(q)}(t)$ increases as p increases from 1 to $d - 1$. However, for $p > d - 1$, the information from $X_{t-d}, \ldots, X_{t-p}$ has already been considered in $Y_{t-1}(=U_{d,t-1}), \ldots, Y_{t-p+d}(=U_{d,t-p+d})$, which are denoted as light green nodes in the graph, and yields little innovative gain on top of $\boldsymbol{Y}_{t-1}^{(q)}$ in predicting Y_t. Therefore, $\mathcal{T}_{X\to Y}^{(p),(q)}(t)$ hardly increases for $p > d - 1$, which is the reason for the saturation at $p = d - 1$. Such a dependence of TE on the lag parameter of the cause time series does not change qualitatively unless decay rates α and β vary depending on sites significantly around the above-mentioned values. This, in conjunction with rough knowledge about the time scale of elemental reactions, may be useful for characterizing the length of the reaction cascade.

Figure 3 shows $\mathcal{T}_{X\to Y}^{(p),(q)}(t)$ assessed from $N_{\mathrm{samp}} = 200$ samples generated by (11) and (12) for the $d = 4$ case presented in Fig. 1 using the methods discussed in Sec. II. Estimates of $\mathcal{T}_{Y\to X}^{(q),(p)}(t)$, the true values of which are constantly zero, are also plotted for reference. For all plots, the upper limit of 99% CIs surpassed the true value (green) at every point. However, its lower limit is also larger than the true value at points where the true values are small (marked by red crosses). In such cases, the CIs do not cover the true values, leading to inaccurate estimations.

Cross covariances in the estimated matrices $\hat{\Sigma}(X_t \oplus \boldsymbol{X}_{t-1}^{(p)} \oplus \boldsymbol{Y}_{t-1}^{(q)})$ and $\hat{\Sigma}(Y_t \oplus \boldsymbol{X}_{t-1}^{(p)} \oplus \boldsymbol{Y}_{t-1}^{(q)})$ are always nonzero even if two time series $\boldsymbol{X}$ and $\boldsymbol{Y}$ are statistically independent. This means that TE assessed from samples is always biased positively even for statistically independent two time series, which could lead to a risk of giving false positives in judging the finiteness of TE. One approach to reduce the risk is to estimate the covariance matrices extremely accurately by collecting a huge number of samples. However, this is difficult to carry out in practice. Another approach is to take into account the positive biases in judging the statistical significance. In Fig. 3 the lower limit of the 99% CI is smaller than the 1% ST at all points that are marked by the red crosses, indicating that the estimates are not statistically significant in the worst case. In other words, statistically significant estimates can be screened under given statistical confidence and significance levels by accepting only cases where the lower limit of the CI is larger than ST (marked by red arrows). This is the method for assessing TE that we propose in this study.

The degrees of freedom of covariance matrices $\hat{\Sigma}(X_t \oplus \boldsymbol{X}_{t-1}^{(p)} \oplus \boldsymbol{Y}_{t-1}^{(q)})$ and $\hat{\Sigma}(Y_t \oplus \boldsymbol{X}_{t-1}^{(p)} \oplus \boldsymbol{Y}_{t-1}^{(q)})$ that are necessary for assessing TE with lag parameters p and q are $(p + q + 2)(p + q + 1)/2$. This is as high as 6 even for the smallest case of $p = q = 1$, and grows up to 28 for $p = q = 3$. As N_{samp} must be sufficiently larger than the degrees of freedom for accurate estimation of the matrices, the sample size should be at least hundreds even if the proposed method is employed.

IV. APPLICATION TO DATA FROM ERBB-RAS-MAPK SYSTEM

The proposed method for assessing TE was applied to real biological data obtained from single living cells.

A. ERBB-RAS-MAPK system and simultaneous signal measurement by the fluorescence microscope

The ERBB-RAS-MAPK system is an intracellular signal transduction network responsible for cell fate decisions [30]. ERBB is a cell surface receptor protein activated by small proteins, including epidermal growth factor (EGF), applied to the cell culture medium. The ERBB activation is recognized by the GRB2/SOS protein complex in the cytoplasm and is associated with the ERBB at the cytoplasmic side of the cell membrane (Fig. 4). Subsequently, SOS activates RAS protein on the membrane to induce an association of RAF protein in the cytoplasm with the active RAS. The recruitment of SOS and RAF to the cell surface can be detected under a fluorescence microscope, reflecting the activation of ERBB and RAS, respectively [21]. SOS and RAF can be observed

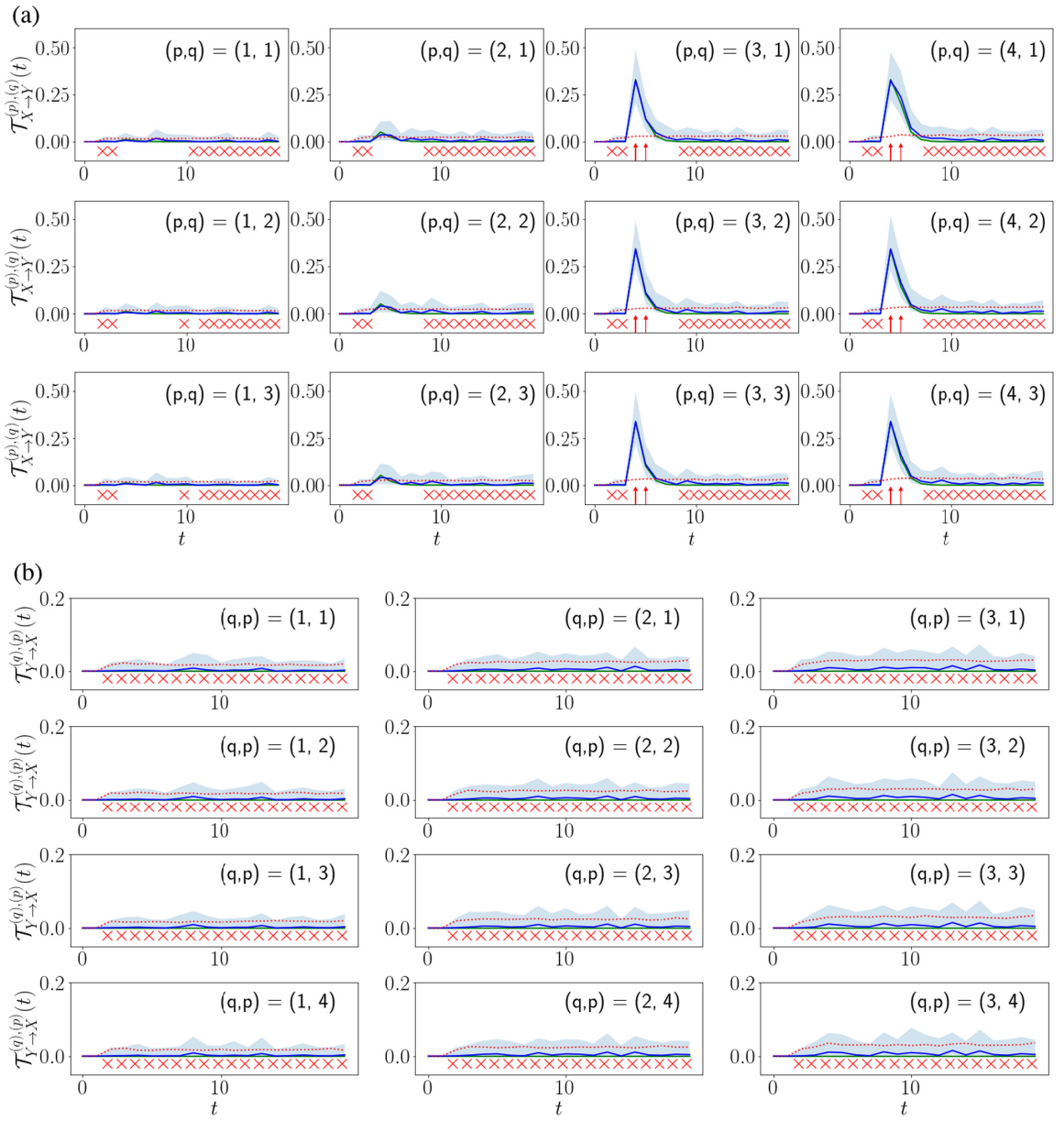

FIG. 3. (a) $\mathcal{T}^{(p),(q)}_{X\to Y}(t)$ assessed from 200 samples for the $d = 4$ case presented in Fig. 1. (b) That for $\mathcal{T}^{(q),(p)}_{Y\to X}(t)$. In both panels, blue and green full lines stand for the estimated and true values, respectively. The true values of $\mathcal{T}^{(q),(p)}_{Y\to X}(t)$ are constantly zero in this setting. Light blue areas and red dotted lines represent 99% CIs and 1% STs. The lower limit of the 99% CI exceeds the true value of the TE at points marked by red crosses, which may overestimate TE. However, one can screen statistically significant estimates (marked by red arrows) by accepting only cases in which the lower limit of the CI is larger than the 1% ST.

simultaneously in the same single cells using two different colors of fluorescent tags [22].

B. Experiment and measured signals

We stimulated ERBB in HeLa cells expressing the wild-type or mutant (i.e., R1131K) SOS with EGF under a microscope at time $t = 0$ and measured the changes in the fluorescence signals from SOS and RAF on the cell membrane (Fig. 5). R1131K is a mutant of SOS, in which the arginine (R) at position 1131 in the amino acid sequence is replaced by lysine (K). This mutation is observed in Noonan syndrome, a human genetic disease [31]. The hyperactivation of RAS has been reported under this mutation, but its molecular mechanism is not entirely known. One possibility is the defect of a negative feedback regulation caused by serine phosphorylation around R1131, which is in the GRB2 association site of SOS [21,32].

Figures 6 and 7 show 248 and 282 samples of the signals measured from cells with wild-type (wt) or mutant SOS, respectively. The unit time is 1 minute, and the signals represent the increase of the fluorescence intensity from the average levels of the first three points corresponding to $t = -3, -2$, and -1 [min]. As shown in these figures, transient increases of the fluorescence signals of SOS and RAF are observed after EGF application. In addition, the RAF activation dynamics

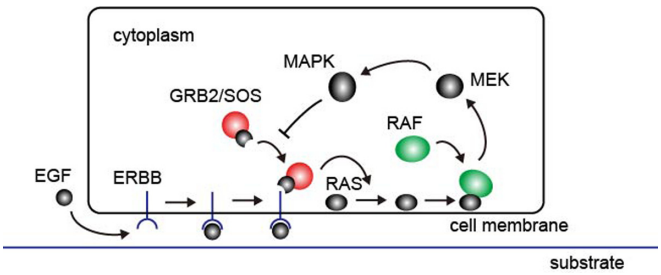

FIG. 4. Signal transduction pathway of the ERBB-RAS-MAPK system. In cells stimulated with EGF, successive translocations of SOS and RAF from the cytoplasm to the cell membrane occur, which recognizes ERBB and RAS activation, respectively. The membrane associations of SOS and RAF and RAF-induced MAPK activation create a negative feedback loop, which is disrupted by the R1131K mutation of SOS.

are delayed and sustained after the initial peak. The difference in the activation dynamics between cells with wt and mutant SOS is barely noticed in the average and single-cell trajectories.

C. Detecting the difference between cells with wild-type and mutant biomolecule by TE

Figures 8 and 9 show the assessment of TE values for the samples presented in Figs. 6 and 7 by changing the lags p and q systematically. A code and raw data for reproducing

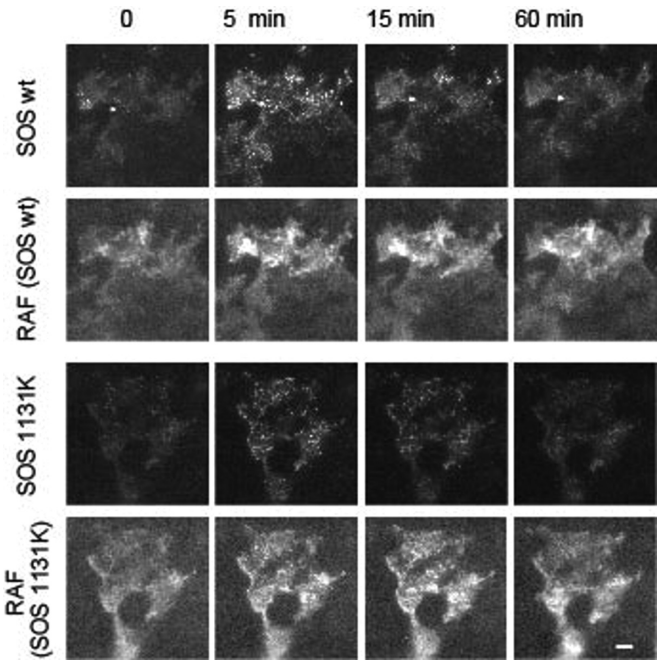

FIG. 5. Recruitments of SOS and RAF to the cell membrane in cells stimulated with EGF. Fluorescence signals from the basal surface of the cells were selectively observed using a total internal reflection fluorescence microscope. At time 0, cells were stimulated with EGF. The upper (SOS) and lower (RAF) images were acquired at the same field of view using a dual color microscopy. Scale bar: 10 μm.

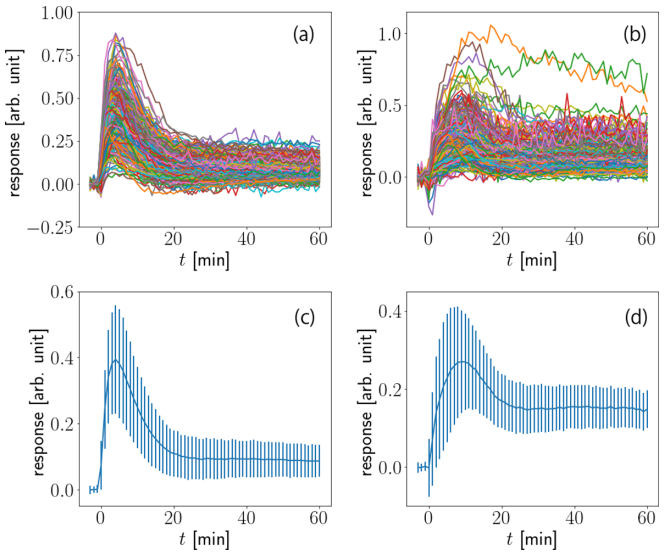

FIG. 6. Top panels: 248 samples of raw signals measured from cells with wt SOS. Bottom panels: Their averages together with one standard deviation. Left (a, c) and right (b, d) panels correspond to SOS and RAF, respectively.

these figures are available from [33]. The sample sizes of 248 and 282 for cells with wt and mutant SOS, respectively, are considered not to be too small compared with the degrees of freedom $(p+q+2)(p+q+1)/2$ of the covariance matrices to be estimated within the range of $1 \leqslant p \leqslant 3$ and $1 \leqslant q \leqslant 3$. For all cases, the number of bootstrapping repetitions B for assessing CI and ST was set to 1000.

Significantly large values of TE from SOS to RAF are observed at the early stage of cell signaling in all (p, q) combinations for both wt and mutant SOS, which is expected from the signal transduction cascade (Fig. 4). However, from RAF

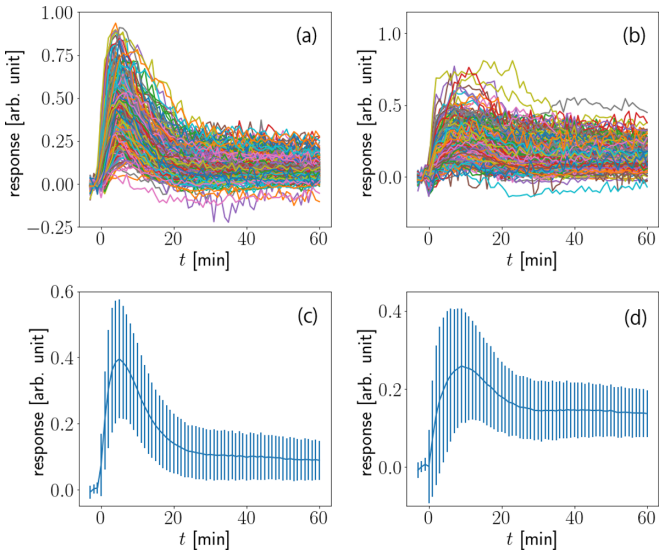

FIG. 7. Top panels: 282 samples of raw signals measured from cells with mutant (R1131K) SOS. Bottom panels: Their averages together with one standard deviation. Left (a, c) and right (b, d) panels correspond to SOS and RAF, respectively.

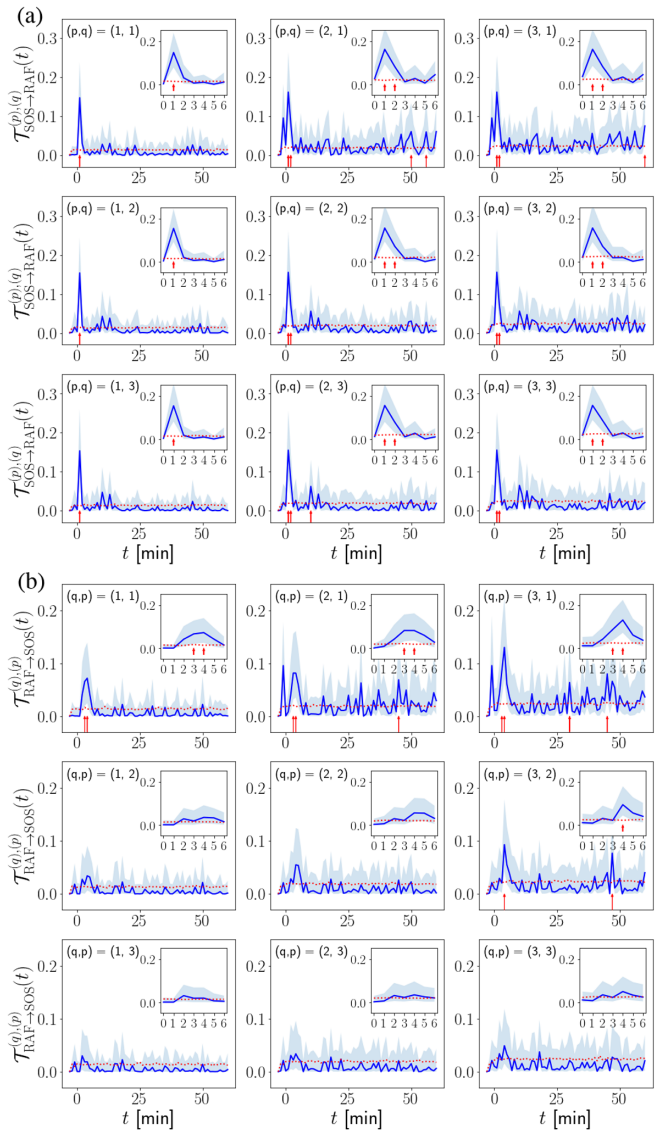

FIG. 8. Assessed TE (blue full lines) for cells with wt SOS from samples shown in Fig. 6. (a) From SOS to RAF. (b) From RAF to SOS. Light blue areas and red dotted lines represent 99% CIs and 1% STs, respectively. The red arrows indicate statistically significant estimates, which are screened by the statistical confidence and the significance levels. In response to the initial peak of TE from SOS to RAF, the significant TE in the reverse direction was found after 2–3 min (insets). Statistically significant TE is also observed even in the later stage.

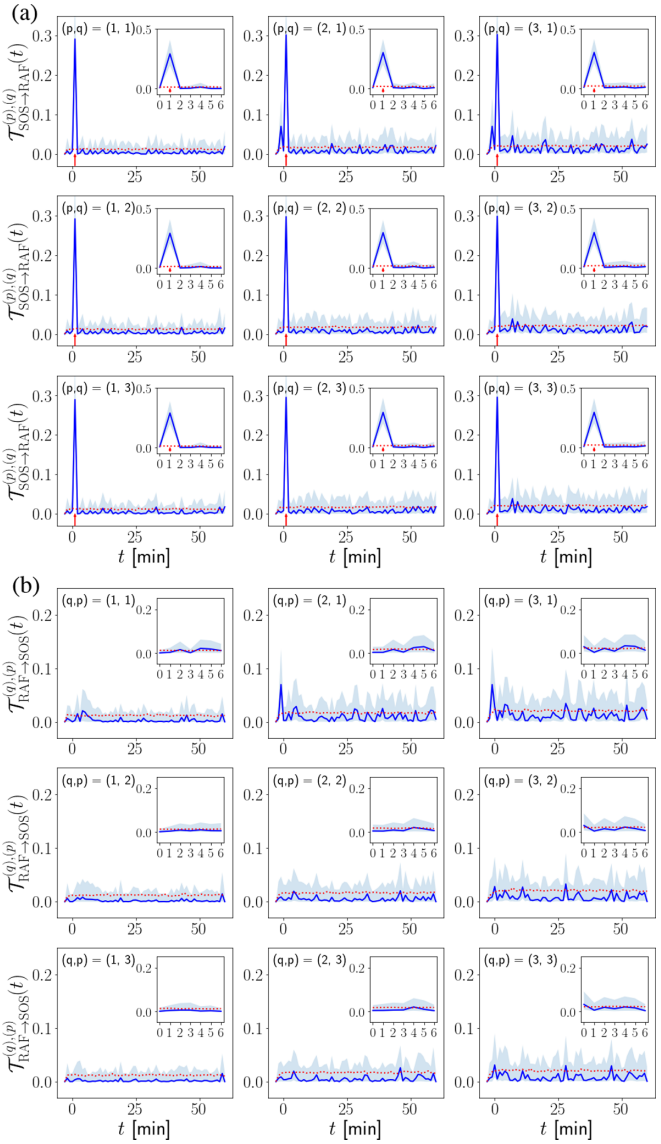

FIG. 9. Assessed TE (blue full lines) for cells with mutant SOS from samples shown in Fig. 7. (a) From SOS to RAF. (b) From RAF to SOS. Light blue areas and red dotted lines represent 99% CIs and 1% STs, respectively. The red arrows indicate statistically significant estimates, which are screened by the statistical confidence and the significance levels. No statistically significant TE was found in the direction from RAF to SOS, even in the early stages (insets).

to SOS, significant TE values are obtained only in cells with wt SOS (Fig. 8). The initial peaks of TE from RAF to SOS delay by approximately 2–3 minutes from those from SOS to RAF (insets), suggesting reversed information flow caused by the negative feedback loop from RAF to SOS via MAPK and/or other proteins downstream of RAF. In addition, the initial peak of TE from RAF to SOS increases as the lag q of the RAF is set larger, while that of the reverse direction does not exhibit such a tendency. The result of the theoretical model in Sec. II implies that this may be because there is a longer cascade of reactions in the signal transduction pathway from RAF to SOS than in that from SOS to RAF. The first peak values of TE from RAF to SOS decrease as p increases, while those from SOS to RAF hardly depend on q. This may be due to the difference of noise levels in the measurement between the SOS and RAF. The noise levels in our measurement are affected by the nature of the fluorescence tag bound to the proteins. We used tetramethylrhodamine and GFP for SOS and RAF, respectively. The former is a chemical probe that is brighter and more stable than fluorescent proteins like GFP.

Another striking feature is that statistically significant TE values are observed in cells with wt SOS at the later stage, while no such behavior is found in cells with mutant SOS. As RAF activation is sustained in cells with SOS of both types

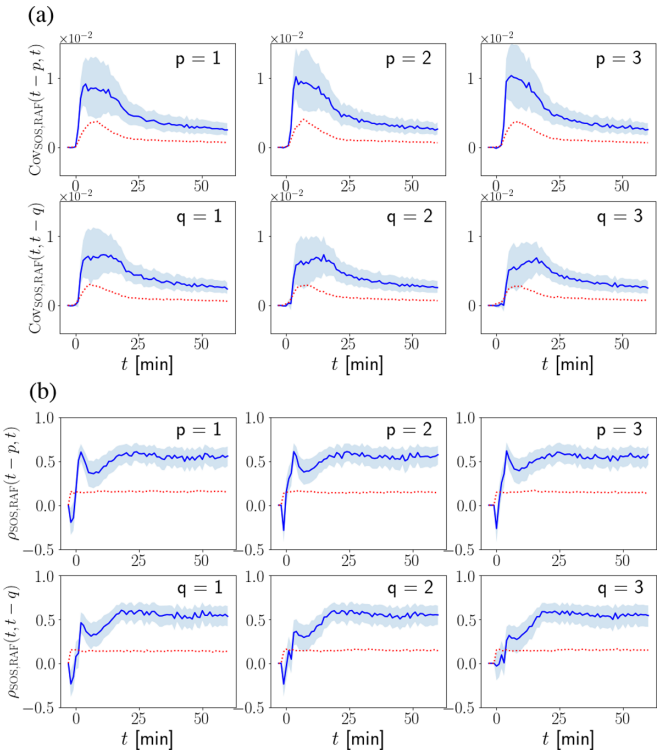

FIG. 10. (a) Covariance and (b) Pearson's correlation coefficient between SOS and RAF activities with time lags for cells with wild-type SOS (blue full lines). 99% CIs (light blue areas) and CTs (red dotted lines), which represent one percentile point from the top for the null hypothesis, are computed with the bootstrapping method.

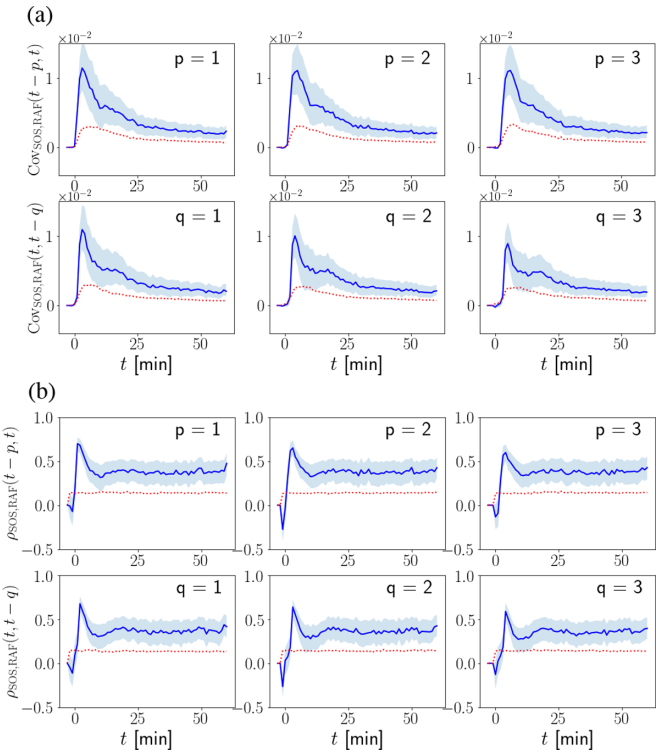

FIG. 11. (a) Covariance and (b) Pearson's correlation coefficient between SOS and RAF activities with time lags for cells with mutant SOS (blue full lines). 99% CIs (light blue areas) and CTs (red dotted lines), which represent one percentile point from the top for the null hypothesis, are computed with the bootstrapping method.

(Figs. 6 and 7), this difference may be a collateral evidence of the defect of a negative feedback regulation from RAF to SOS in cells with mutant SOS.

Under the Gaussian approximation, TE can be computed from covariances of two time series with given time lags. This may invoke a naive question whether similar results can also be obtained by more conventional covariance based analyses. For answering such a question, we plot covariance and Pearson's correlation coefficient, which is defined by normalizing the covariance with product of standard deviations, between SOS and RAF changing time lags in Figs. 10 and 11 for cells with wt and mutant SOS, respectively. The plots show that the two time series with the time lags are correlated with the statistically significant level in the almost entire range of observation time regardless of whether the SOS type is wild or mutant. Although the profiles of peaks and the strength of correlations differ slightly, it is difficult to find distinct qualitative difference between the two cases from the plots. This implies that TE is a more suitable measure for characterizing the information flow conveyed between the two time series with a high time resolution.

V. SUMMARY AND DISCUSSION

In summary, we examined the possibilities and limitations of assessing the transfer entropy (TE) from the measured data of biochemical reactions. We employed the Gaussian approximation, which enables us to efficiently assess TE based on covariance matrices estimated from samples of objective time series. In general, it is necessary to evaluate the sampling distributions to guarantee the accuracy of the estimated results. However, an analytical evaluation of the sampling distributions of TE is difficult for nonstationary time series. We resolved this difficulty by computationally assessing the sampling distributions using bootstrapping techniques from computational statistics. The computational methods were tested by the application to a theoretically tractable model of a stochastic process, which led to the development of a method for screening only statistically significant estimates under given levels of statistical confidence and significance. In addition, this method was applied to assess the dynamics of the information flow in a real biological reaction network inside living cells. Although the raw signals measured from cells with the wild-type and a mutant molecule are hardly distinguished, the method successfully detected the difference between them in the time course of TE. This implies that the developed method may serve as a useful tool in studying intracellular reaction networks in which large-scale simultaneous measurement of activities of biomolecules has been made possible owing to the recent advancement of fluorescence microscope technologies.

It should be noted that the finiteness of TE is not a sufficient sign of the causal relationship between two time series [34]; nonzero TE can be a spurious causality detector when observations are performed incompletely due to the presence of unobserved states [35]. Nevertheless, the TE based analysis

shown in the current paper would still be useful at least for the following two purposes. One is to quantify the efficiency of information transmission for *known pathways* for which unobserved states are absent or their influence is negligible. As mentioned in Introduction, many pathways of biochemical reactions have been identified with a high accuracy for these decades. However, little is known about when and how large information is transmitted through the pathways. Our methodology could fill the "missing piece" with a high time resolution, although hot debates are continuing on the appropriateness of TE as a causality quantifier [36–41]. The other is to offer clues for finding *unknown pathways*. Although the finiteness of TE is not a sufficient condition of the causal relationship, it still serves as a necessary condition. Therefore, the assessment of TE would provide useful guidelines for screening possible candidates of relevant pathways. In addition, even if pathways are not identified completely, comparison of TE between healthy and disordered systems might lead to more efficient diagnosis and treatment of various disorders.

Although we restricted the application domain to biochemical reactions in this study, the proposed methodology can be utilized to analyze information flow in general systems of a wider class. For instance, it may be useful in examining time-varying effective interactions in multiagent or nonlinear dynamical systems from simulation data [42–44] and those in nervous or active matter systems from video imaging data [45,46], as phenomena observed in such systems are fairly reproducible and collecting many data on them is relatively easy. Nevertheless, the necessity of collecting many samples of simultaneously measured data (i.e., at least hundreds of samples) for accurately estimating covariances may be a bottleneck in applying the data to nonstationary time series in many other domains. Meanwhile, in general, appropriate prior knowledge about objective systems can improve the estimation accuracy significantly. The reduction of the necessary sample size incorporating Bayesian inference and/or other machine learning techniques is an important research direction for further research.

ACKNOWLEDGMENTS

We acknowledge the technical assistance provided by Mutsumi Nakanishi. This work was partially supported by the MEXT KAKENHI Grant No. 19H05647 (Y.S.), JSPS KAKENHI Grants No. 17H00764 (T.O., Y.K.), 18K11463, 19H0182 (T.O.), and 20H00620 (Y.K.), and JST CREST Grant No. JPMJCR1912 (Y.S., Y.K.).

APPENDIX: DERIVATION OF EQ. (7)

For two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ that follow joint distribution $Q(\boldsymbol{x},\boldsymbol{y})$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ may be either scalar or any dimensional vector random variables, conditional entropy of $\boldsymbol{X}$ given $\boldsymbol{Y}$ is defined generally as

$$H(\boldsymbol{X}|\boldsymbol{Y}) = -\int d\boldsymbol{y} Q(\boldsymbol{y})\, d\boldsymbol{x} Q(\boldsymbol{x}|\boldsymbol{y}) \ln Q(\boldsymbol{x}|\boldsymbol{y}). \tag{A1}$$

Let us suppose that $\boldsymbol{X}$ and $\boldsymbol{Y}$ follows a multivariate Gaussian,

$$Q(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{(2\pi)^{\frac{n+m}{2}}[\det\Sigma(\boldsymbol{X}\oplus\boldsymbol{Y})]^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}\oplus\boldsymbol{y}-\boldsymbol{\mu}_X\oplus\boldsymbol{\mu}_Y)^\top \Sigma(\boldsymbol{X}\oplus\boldsymbol{Y})^{-1}(\boldsymbol{x}\oplus\boldsymbol{y}-\boldsymbol{\mu}_X\oplus\boldsymbol{\mu}_Y)\right), \tag{A2}$$

where n and m are the dimensions of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$ are means of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Covariance matrix $\Sigma(\boldsymbol{X}\oplus\boldsymbol{Y})$ is expressed as

$$\Sigma(\boldsymbol{X}\oplus\boldsymbol{Y}) = \begin{pmatrix} \Sigma(\boldsymbol{X}) & \Sigma(\boldsymbol{X},\boldsymbol{Y}) \\ \Sigma(\boldsymbol{Y},\boldsymbol{X}) & \Sigma(\boldsymbol{Y}) \end{pmatrix} \tag{A3}$$

using the notation defined in the main text.

Computing the matrix inversion of (A3) yields an expression

$$\Sigma(\boldsymbol{X}\oplus\boldsymbol{Y})^{-1} = \begin{pmatrix} \Sigma(\boldsymbol{X}|\boldsymbol{Y})^{-1} & -\Sigma(\boldsymbol{X}|\boldsymbol{Y})^{-1}\Sigma(\boldsymbol{X},\boldsymbol{Y})\Sigma(\boldsymbol{Y})^{-1} \\ -\Sigma(\boldsymbol{Y})^{-1}\Sigma(\boldsymbol{Y},\boldsymbol{X})\Sigma(\boldsymbol{X}|\boldsymbol{Y})^{-1} & \Sigma(\boldsymbol{Y}|\boldsymbol{X})^{-1} \end{pmatrix}. \tag{A4}$$

This means that conditional distribution $Q(\boldsymbol{x}|\boldsymbol{y})$ is expressed as

$$Q(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}\det[\Sigma(\boldsymbol{X}|\boldsymbol{Y})]^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{X|Y})^\top \Sigma(\boldsymbol{X}|\boldsymbol{Y})^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{X|Y})\right), \tag{A5}$$

where $\boldsymbol{\mu}_{X|Y} = \boldsymbol{\mu}_X + \Sigma(\boldsymbol{X},\boldsymbol{Y})\Sigma(\boldsymbol{Y})^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_Y)$. Inserting this expression into (A1) offers

$$H(\boldsymbol{X}|\boldsymbol{Y}) = \frac{1}{2}\ln\det(\Sigma(\boldsymbol{X}|\boldsymbol{Y})) + \frac{n}{2}\ln(2\pi e). \tag{A6}$$

Finally, substituting X_t and $\boldsymbol{X}_{t-1}^{(p)}$ with $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, in (A6) provides (10).

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. Watson, *Molecular Biology of the Cell*, 4th ed. (Garland, New York, 2002).

[2] I. Tavassoly, J. Goldfarb, and R. Iyengar, Systems biology primer: The basic methods and approaches, Essays Biochem. **62**, 487 (2018).

[3] T. Schreiber, Measuring Information Transfer, Phys. Rev. Lett. **85**, 461 (2000).

[4] L. Barnett, A. B. Barrett, and A. K. Seth, Granger Causality and Transfer Entropy are Equivalent for Gaussian Variables, Phys. Rev. Lett. **103**, 238701 (2009).

[5] T. Bossomaier, L. Barnett, M. Harr, and J. T. Lizier, *An Introduction to Transfer Entropy: Information Flow in Complex Systems* (Springer, Cham, 2016).

[6] M. Lungarella, K. Ishiguro, Y. Kuniyoshi, and N. Otsu, Methods for quantifying the causal structure of bivariate time series, Int. J. Bifurcation Chaos **17**, 903 (2007).

[7] M. Wibral, B. Rahm, M. Rieder, M. Lindner, R. Vicente, and J. Kaiser, Transfer entropy in magnetoencephalographic data: Quantifying information flow in cortical and cerebellar networks, Prog. Biophys. Mol. Biol. **105**, 80 (2011).

[8] O. Stetter, D. Battaglia, J. Soriano, and T. Geisel, Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals, PLoS Comput. Biol. **8**, 1 (2012).

[9] M. Ursino, G. Ricci, and E. Magosso, Transfer entropy as a measure of brain connectivity: A critical analysis with the help of neural mass models, Front. Comput. Neurosci. **14**, 45 (2020).

[10] C. J. Honey, R. Kötter, M. Breakspear, and O. Sporns, Network structure of cerebral cortex shapes functional connectivity on multiple time scales, Proc. Natl. Acad. Sci. U. S. A. **104**, 10240 (2007).

[11] L. Faes, G. Nollo, and A. Porta, Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique, Phys. Rev. E **83**, 051112 (2011).

[12] S. Stramaglia, G.-R. Wu, M. Pellicoro, and D. Marinazzo, Expanding the transfer entropy to identify information circuits in complex systems, Phys. Rev. E **86**, 066211 (2012).

[13] J. Sun and E. M. Bollt, Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings, Physica D **267**, 49 (2014).

[14] L. Novelli, P. Wollstadt, P. Mediano, M. Wibral, and J. T. Lizier, Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing, Netw. Neurosci. **3**, 827 (2019).

[15] D. P. Shorten, R. E. Spinney, and J. T. Lizier, Estimating transfer entropy in continuous time between neural spike trains or other event-based data, PLoS Comput. Biol. **17**, 1 (2021).

[16] T. Q. Tung, T. Ryu, K. H. Lee, and D. Lee, Inferring gene regulatory networks from microarray time series data using transfer entropy, in *Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)* (IEEE, New York, 2007), pp. 383–388.

[17] S. Roy, D. Das, D. Choudhury, G. G. Gohain, R. Sharma, and D. K. Bhattacharyya, Causality inference techniques for in-silico gene regulatory network, in *Mining Intelligence and Knowledge Exploration*, edited by R. Prasath and T. Kathirvalavakumar (Springer International Publishing, Cham, 2013), pp. 432–443.

[18] J. C. Castro, I. Valdés, L. N. Gonzalez-García, G. Danies, S. Cañas, F. V. Winck, C. E. Ñústez, S. Restrepo, and D. M. Riaño-Pachón, Gene regulatory networks on transfer entropy (GRNTE): A novel approach to reconstruct gene regulatory interactions applied to a case study for the plant pathogen *Phytophthora Infestans*, Theor. Biol. Med. Modell. **16**, 1 (2019).

[19] J. Kim, S. T. Jakobsen, K. N. Natarajan, and K.-J. Won, TENET: Gene network reconstruction using transfer entropy reveals key regulatory factors from single cell transcriptomic data, Nucleic Acids Res. **49**, e1 (2020).

[20] L. R. Garcia Michel, C. D. Keirns, B. C. Ahlbrecht, and D. A. Barr, Calculating transfer entropy from variance-covariance matrices provides insight into allosteric communication in ERK2, J. Chem. Theory Comput. **17**, 3168 (2021).

[21] Y. Nakamura, N. Umeki, M. Abe, and Y. Sako, Mutation-specific mechanisms of hyperactivation of Noonan syndrome SOS molecules detected with single-molecule imaging in living cells, Sci. Rep. **7**, 14153 (2017).

[22] R. Yoshizawa, N. Umeki, A. Yamamoto, M. Okada, M. Murata, and Y. Sako, p52Shc regulates the sustainability of ERK activation in a RAF-independent manner, Mol. Biol. Cell **32**, 1838 (2021).

[23] C. W. J. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica **37**, 424 (1969).

[24] D. Marinazzo, M. Pellicoro, and S. Stramaglia, Kernel Method for Nonlinear Granger Causality, Phys. Rev. Lett. **100**, 144103 (2008).

[25] O. Ledoit and M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivariate Anal. **88**, 365 (2004).

[26] C. Granger, Economic processes involving feedback, Inf. Control. **6**, 28 (1963).

[27] P. Whittle, The analysis of multiple stationary time series, J. R. Stat. Soc. B Stat. Methodol. **15**, 125 (1953).

[28] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability No. 57 (Chapman & Hall/CRC, Boca Raton, FL, 1993).

[29] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods* (Oxford University Press, Oxford, 2001).

[30] M. A. Lemmon and J. Schlessinger, Cell signaling by receptor tyrosine kinases, Cell **141**, 1117 (2010).

[31] F. Lepri, A. De Luca, L. Stella, C. Rossi, G. Baldassarre, F. Pantaleoni, V. Cordeddu, B. J. Williams, M. L. Dentici, V. Caputo, S. Venanzi, M. Bonaguro, I. Kavamura, M. F. Faienza, A. Pilotta, F. Stanzial, F. Faravelli, O. Gabrielli, B. Marino, G. Neri *et al.*, SOS1 mutations in Noonan syndrome: Molecular spectrum, structural insights on pathogenic effects, and genotype-phenotype correlations, Human Mutation **32**, 760 (2011).

[32] S. Corbalan-Garcia, S. S. Yang, K. R. Degenhardt, and D. Bar-Sagi, Identification of the mitogen-activated protein kinase phosphorylation sites on human Sos1 that regulate interaction with Grb2, Mol. Cell. Biol. **16**, 5674 (1996).

[33] https://github.com/kabashiy/transfer_entropy/.

[34] M. Baldovin, F. Cecconi, and A. Vulpiani, Understanding causation via correlations and linear response theory, Phys. Rev. Research **2**, 043436 (2020).

[35] D. A. Smirnov, Spurious causalities with transfer entropy, Phys. Rev. E **87**, 042917 (2013).

[36] D. A. Smirnov, Quantification of causal couplings via dynamical effects: A unifying perspective, Phys. Rev. E **90**, 062921 (2014).

[37] P. A. Stokes and P. L. Purdon, A study of problems encountered in Granger causality analysis from a neuroscience perspective, Proc. Natl. Acad. Sci. U. S. A. **114**, E7063 (2017).

[38] M. Dhamala, H. Liang, S. L. Bressler, and M. Ding, Granger-Geweke causality: Estimation and interpretation, NeuroImage **175**, 460 (2018).

[39] L. Barnett, A. B. Barrett, and A. K. Seth, Solved problems for Granger causality in neuroscience: A response to Stokes and Purdon, NeuroImage **178**, 744 (2018).

[40] D. A. Smirnov, Transfer entropies within dynamical effects framework, Phys. Rev. E **102**, 062139 (2020).

[41] D. A. Smirnov, Spectral causalities within dynamical effects framework, Europhys. Lett. **128**, 20006 (2019).

[42] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, Novel type of Phase Transition in a System of Self-Driven Particles, Phys. Rev. Lett. **75**, 1226 (1995).

[43] R. G. Andrzejak, A. Ledberg, and G. Deco, Detecting event-related time-dependent directional couplings, New J. Phys. **8**, 6 (2006).

[44] A. Cavagna and I. Giardina, Bird flocks as condensed matter, Annu. Rev. Condens. Matter Phys. **5**, 183 (2014).

[45] K. Ota, Y. Oisi, T. Suzuki, M. Ikeda, Y. Ito, T. Ito, H. Uwamori, K. Kobayashi, M. Kobayashi, M. Odagawa, C. Matsubara, Y. Kuroiwa, M. Horikoshi, J. Matsushita, H. Hioki, M. Ohkura, J. Nakai, M. Oizumi, A. Miyawaki, T. Aonishi *et al.*, Fast, cell-resolution, contiguous-wide two-photon imaging to reveal functional network architectures across multi-modal cortical areas, Neuron **109**, 1810 (2021).

[46] J. Iwasawa, D. Nishiguchi, and M. Sano, Algebraic correlations and anomalous fluctuations in ordered flocks of Janus particles fueled by an ac electric field, Phys. Rev. Research **3**, 043104 (2021).