SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:

NONLINEAR PHENOMENA

GENE REGULATORY NETWORKS

Received
17 February 2014

Accepted
25 November 2014

Published
12 December 2014

Correspondence and requests for materials should be addressed to K.A. (aihara@sat.t.u-tokyo.ac.jp) or L.C. (lnchen@sibs.ac.cn)

Detecting Causality from Nonlinear Dynamics with Short-term Time Series

Huanfei Ma[1,2], Kazuyuki Aihara[2] & Luonan Chen[2,3]

[1]School of Mathematical Sciences, Soochow University, China, [2]Collaborative Research Center for Innovative Mathematical Modelling, Institute of Industrial Science, The University of Tokyo, Japan, [3]Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China.

Quantifying causality between variables from observed time series data is of great importance in various disciplines but also a challenging task, especially when the observed data are short. Unlike the conventional methods, we find it possible to detect causality only with very short time series data, based on embedding theory of an attractor for nonlinear dynamics. Specifically, we first show that measuring the smoothness of a cross map between two observed variables can be used to detect a causal relation. Then, we provide a very effective algorithm to computationally evaluate the smoothness of the cross map, or "Cross Map Smoothness" (CMS), and thus to infer the causality, which can achieve high accuracy even with very short time series data. Analysis of both mathematical models from various benchmarks and real data from biological systems validates our method.

Detecting causal relationship between variables, especially from observed time series, has attracted great attentions from multiple-disciplines. Though there still has been no universally accepted definition of causality, various measures of causality have been reported and extensively studied. Among a variety of methods based on linear regression, Granger causality[1] is undoubtedly a widely accepted definition and method to detect causal relationship between different factors. However, the Granger method mainly focuses on linear models and needs a key condition of separability, namely, assumes that the driven information from causative factors can be removed from the effects[2]. In fact, it has been noted that the causality in Granger sense may be not suitable to detect directional coupling between nonlinear systems[3], especially deterministic dynamical systems with weak or moderate couplings. To quantify causal relationship between the intertwined variables of nonlinear systems, methods developed from transfer entropy[4], conditional mutual information[5], recurrence plots[6,7] and nonlinear extension of Granger causality[8,9] have been proposed and extensively studied. Moreover, various kinds of mutual nonlinear cross map methods based on state space reconstruction (SSR) technique have been also studied both theoretically[10–13] and numerically[14] for a long time. In particular, recently the mutual cross map method has been successfully applied to solve complex relationship in ecological systems[2,15].

Though these methods have been demonstrated to correctly identify causal relations for many systems, they all require sufficiently long time series to achieve a reasonable result. This stems from the fact that training regression models, calculating correlations, determining transition probabilities and finding nearest neighbors all need a sufficiently large training set or a large number of samples.

However, in practical situations, the measured time series are always limited rather than sufficiently long, and sometimes are even rather short, e.g., the high throughput microarray or RNA-seq data for gene expressions of a biological process are typically measured less than 20 time points due to both experimental and economical constraints[16]. Though various methods based on Bayesian inference, regression analysis, econometrics models and standard similarity measures have been used to analyze such short time series data[17–19], inferring genetic networks from short data is still regarded as an 'ill-posed' inverse problem and a challenging task[20,21]. On the other hand, in some occasions even though long-term data can be measured, only short (recent) pieces can correctly reflect the causal relation between subsystems due to the nonstationary and fast switching property of the concerned systems. Therefore it is in urgent need of developing new methods to detect causality based on short-term data or a small number of samples.

In contrast to the traditional knowledge that short-term data cannot provide enough information to infer the causal relation, here we show that we can detect causality from very short time series in an accurate manner by exploiting global information of data. Specifically, we propose a measurement "Cross Map Smoothness" (CMS) based on the embedding theory of attractors[22,23] in this paper, which can not only detect causal relationship but also derive a cross map between any two observed variables even with short-term time series data, and then we provide an efficient algorithm to construct such a map for inferring the causal relation. The key idea behind our method is that measuring smoothness of a cross map between two observed variables implies causal relations, which can be computationally achieved even with short time series data, comparing with the traditional methods, e.g., the nearest neighbor method. Analysis of mathematical models from various benchmarks validates our results and real data from biological systems confirms the method can be used to infer genetic networks from short data.

Methods

To begin, we revisit the mutual cross map method based on state space reconstruction. Consider two scalar time series $x(t)$ and $y(t)$ measured from two variables x and y in an unknown nonlinear dynamical system. With appropriately chosen embedding dimension L and proper delay τ[24,25], one can obtain time delayed coordinate vectors $\boldsymbol{x}(t) = [x(t), x(t-\tau), \ldots, x(t-(L-1)\tau)]^T$ and $\boldsymbol{y}(t) = [y(t), y(t-\tau), \ldots, y(t-(L-1)\tau)]^T$ respectively. According to delayed embedding theory[22], the set of vectors $\boldsymbol{x}(t)$ forms the reconstructed attractor M_x, and one can define M_y in an analogous way. For each point $\boldsymbol{y}(t_0)$ on M_y, one can find its k nearest neighbors $\boldsymbol{y}(t_{y1}), \boldsymbol{y}(t_{y2}), \ldots, \boldsymbol{y}(t_{yk})$ with time indices $t_{y1}, t_{y2}, \ldots, t_{yk}$. Moreover, one can define the mutual neighbors for $\boldsymbol{x}(t_0) \in M_x$ as $\boldsymbol{x}(t_{y1}), \boldsymbol{x}(t_{y2}), \ldots, \boldsymbol{x}(t_{yk})$ and the map from nearest neighbors to mutual neighbors is defined as cross map Φ_{yx}: $M_y \rightarrow M_x$. In the case that x is causally influencing y, or x is a driving factor of y (i.e., $x \rightarrow y$), the information of x is included in the dynamics of y, and thus two close states on M_y correspond to two close states on M_x, or explicitly, the mutual neighbors of $\boldsymbol{x}(t_0)$ are also in the neighborhood of $\boldsymbol{x}(t_0)$. Inversely, in the case that y has no influence over x, the dynamics of x is insensitive to the state of y and the mutual neighbors of $\boldsymbol{y}(t_0)$ are not necessarily to be close to $\boldsymbol{y}(t_0)$, as illustrated in Figs. 1(a) and (b). Therefore, the geometry property of mutual neighbors can be used to detect causality[2,10,11]. The details of mutual neighbors and cross map are revisited in Supplementary Information.

Here, it should be stressed that one requirement is essential, that is, the nearest neighbors for both x and y are required to be sufficiently close to the true neighborhood so that the local geometric information can be correctly measured. If this requirement is not fulfilled, contradictory results may be derived, e.g., the studies of refs. 10 and 11 used different assumptions for mutual predictions which both gave the same result. In fact, due to the computational way of state space reconstruction using delayed embedding technique, sufficiently long time series are required to guarantee that the nearest neighbors on the reconstructed attractor converge to the true neighborhood. Figure 2 shows the relationship between nearest neighbors and the time series length, where the nearest neighbors found on the attractor reconstructed from short time series (Fig. 2(b)) are apparently far away from the true neighborhood of the underlying center point (Fig. 2(a)). Detailed discussions and explanations on the necessity of convergence of nearest neighbors can be further referred to ref. 2. Thus, detecting causality based on nearest neighbors and mutual neighbors essentially requires sufficiently long time series data to make reliable causality detection.

Here, we notice that the key idea behind the method of finding mutual neighbors is actually measuring the smoothness of the map. Specifically, if x causally influences y, the nearest neighbors of $\boldsymbol{y}(t_0)$ are mapped to close states of $\boldsymbol{x}(t_0)$, i.e., the cross map Φ_{yx}: $M_y \rightarrow M_x$ maps the neighborhood of $\boldsymbol{y}(t_0)$ to the neighborhood of $\boldsymbol{x}(t_0)$, which actually implies that Φ_{yx} is locally smooth around $\boldsymbol{y}(t_0)$, as shown in Fig. 1(a). On the reverse direction, if y has no influence over x, the image of $\boldsymbol{x}(t_0)$'s neighborhood under the cross map Φ_{xy}: $M_x \rightarrow M_y$ is not necessarily the neighborhood of $\boldsymbol{y}(t_0)$, thus the cross map Φ_{xy} is not necessarily smooth around $\boldsymbol{x}(t_0)$, as shown in Fig. 1(b). If the cross map Φ is locally smooth in the neighborhood of every point on the attractor, then the map is globally smooth on the whole attractor, and vice versa. Thus, the global smoothness of Φ_{yx} and Φ_{xy} can be built from local properties, as illustrated in Figs. 1(c) and (d). Moreover, when the coupling strength increases, information becomes more distinct in the causally influenced variables. As a result, their attractors will contain stronger historical information from the causes. Thus, within one system, the relative smoothness can indicate the relative strength of causative effectiveness.

Therefore, finding mutual nearest neighbors is equivalent to measuring the smoothness of the cross map Φ, i.e., the smoothness of Φ_{yx} indicates the strength of causative effectiveness from x to y. While mutual neighbors only use the local information around one point, we propose a new framework, i.e., Cross Map Smoothness (CMS), to measure the smoothness of Φ using global information, and consequently we can detect causality even from short time series in an accurate manner. In other words, instead of finding nearest neighbors which requires a large number of samples, we computationally evaluate the smoothness of the cross map by designing an efficient algorithm for the global attractor.

Our fundamental idea is based on the fact that any smooth map can be approximated by a neural network $\mathcal{N}$[26] while training a neural network to approximate an unsmooth map will fail with large training errors, as illustrated in Figs. 3(a)–(d). Furthermore, the training errors reflects the relative smoothness, and thus can be a measure of the relative strength of causative effectiveness. Therefore we can train the neural network $\mathcal{N}$ to approximate the map Φ_{yx}, using the whole set of data $\boldsymbol{y}(t)$ on M_y as input and the whole set of data $\boldsymbol{x}(t)$ on M_x as output. Thus the training error (i.e., the measurement of the relative smoothness)

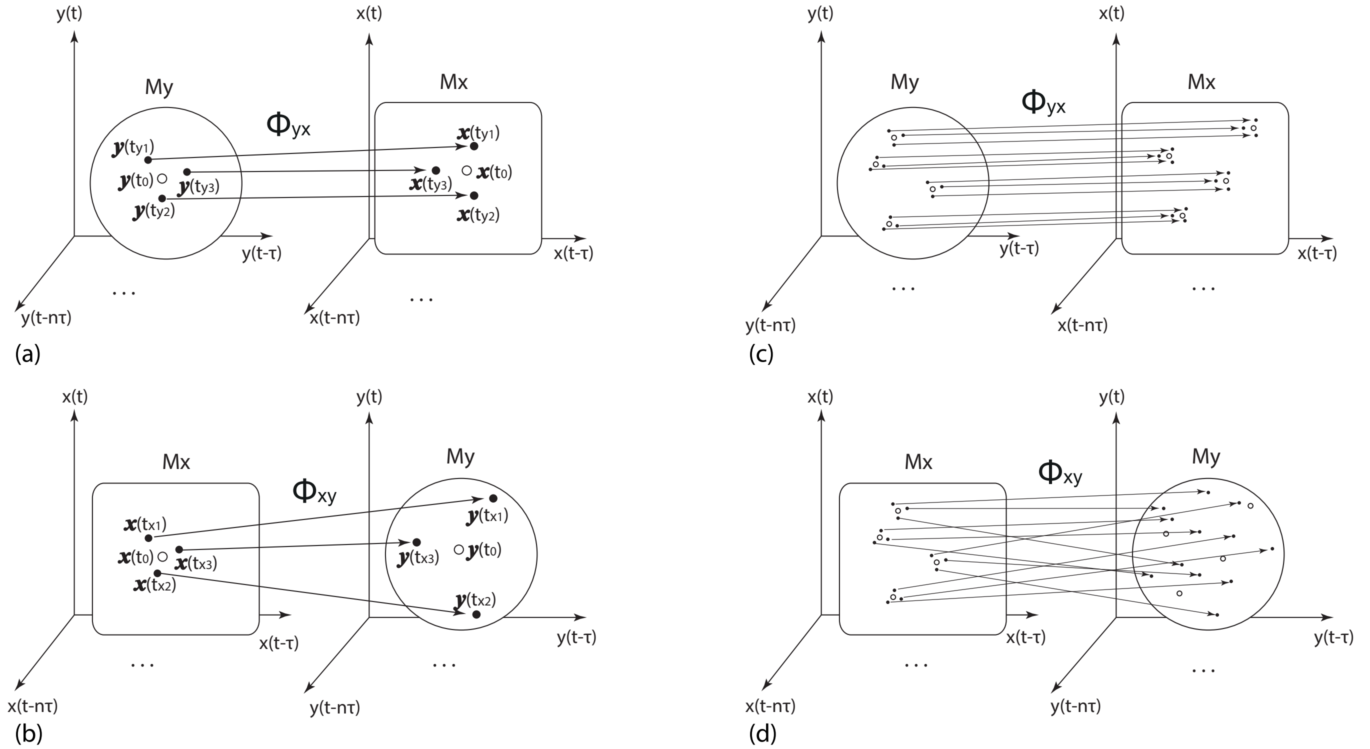

Figure 1 | Illustration of mutual neighbors, cross map and smoothness. (a) For one point $\boldsymbol{y}(t_0) \in M_y$ and its counterpart $\boldsymbol{x}(t_0) \in M_x$, one can find the nearest neighbors $\boldsymbol{y}(t_{y_1}), \boldsymbol{y}(t_{y_2}), \boldsymbol{y}(t_{y_3})$ for $\boldsymbol{y}(t_0)$ and define the mutual neighbors $\boldsymbol{x}(t_{y_1}), \boldsymbol{x}(t_{y_2}), \boldsymbol{x}(t_{y_3})$ for $\boldsymbol{x}(t_0)$. The map between the nearest neighbors and mutual neighbors is defined as cross map Φ_{yx}. In the case x causally influences y, the cross map Φ_{yx} maps a neighborhood to a neighborhood. (b) In the case y does not causally influence x, the cross map Φ_{xy} does not necessarily map a neighborhood to a neighborhood. (c) and (d) The global smoothness of Φ_{yx} and Φ_{xy} built from local smoothness.

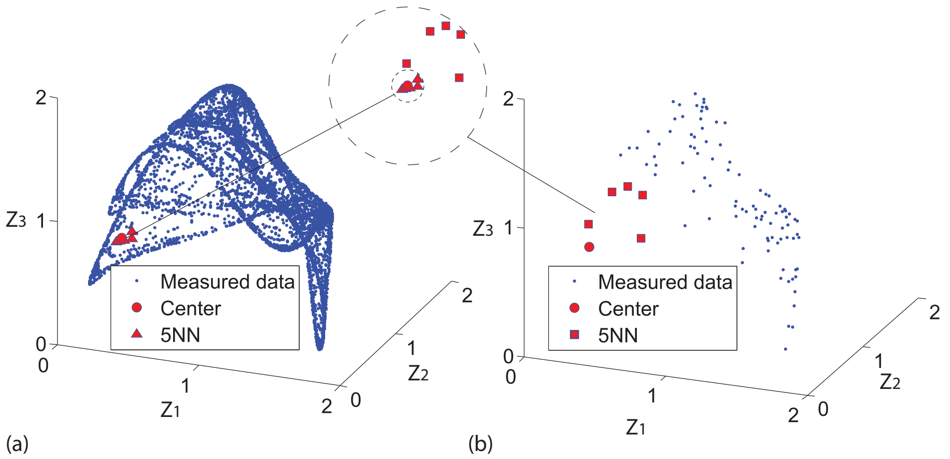

Figure 2 | Illustration for the time series length and convergence of nearest neighbors. Here the time series are generated by one chaotic Lotka-Volterra system. (a) A reconstructed attractor from time series of 7000 samples, and the 5 nearest neighbors (5NN) of one center point. (b) A reconstructed attractor from time series of only 100 sampled points and the 5 nearest neighbors of the same center point. Inset: the comparison of the 5 nearest neighbors for both (a) and (b), where the latter set of points are apparently not close to the center point at all.

between Φ and $\mathcal{N}$ indicates the strength of the causative influence from x to y. The sketch of the Cross Map Smoothness (CMS) with the neural network (NN) method is illustrated in Figs. 3(e) and (f).

Thus, we propose the Cross Map Smoothness(CMS) algorithm using a Radial Basis Function (RBF) network to detect causality between two variables x and y. The details of the algorithm is listed in Supplementary Information Section 2. Here we adopt a leave-one-out strategy to fully use the short time series, i.e., we train one RBF network based on each leave-one-out data set and make prediction on the one test point. Finally, we compute the causality index R_{xy} based on the the normalized training error, which measures the causative effective strength from x to y.

Results

To validate the method, several representative examples are considered as benchmarks.

Theoretical model validation. Let us begin with several representative causality patterns which can be used as motifs in many complex situations. We first consider two coupled variables with both unidirectional and bidirectional couplings in the following form,

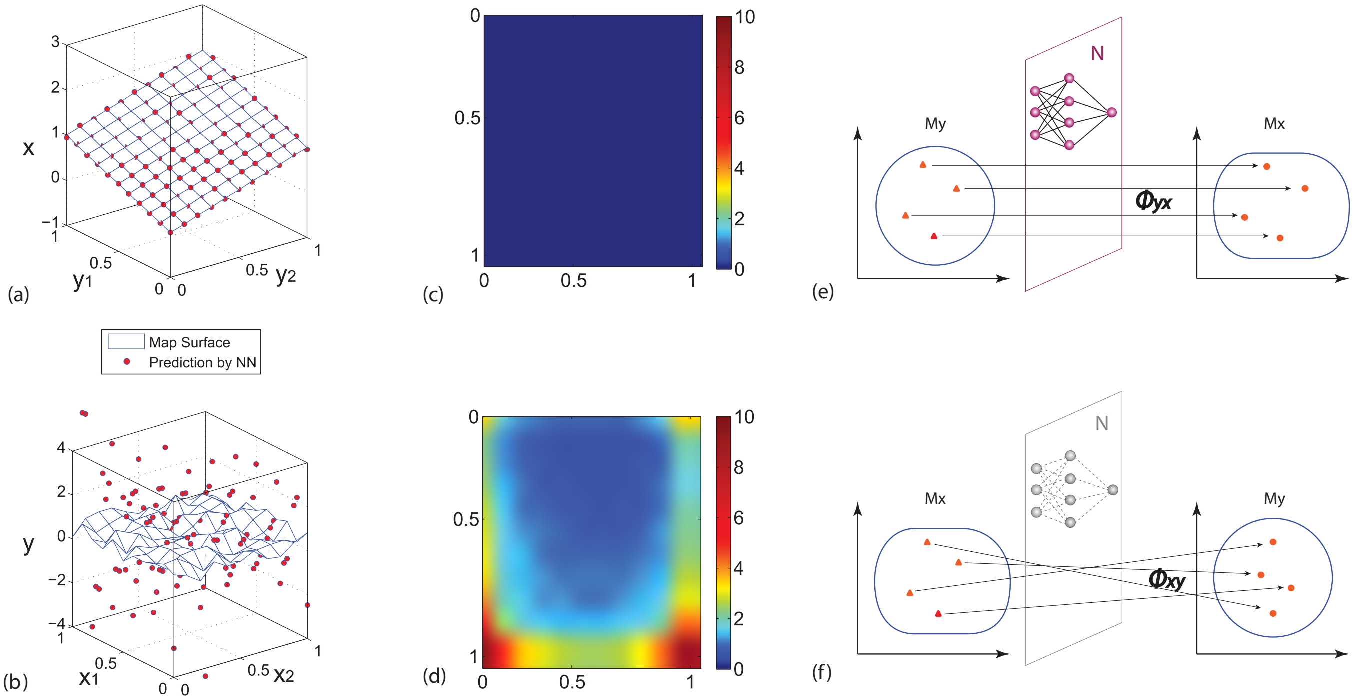

Figure 3 | Sketch of the cross map smoothness learned by a neural network (NN). (a) and (b) Illustrations for the neural network's approximation ability for smooth map and unsmooth map. Here the map surface in (a) is assumed to be $x = y_1 + y_2$ and the surface in (b) is simply generated by random points. (c) and (d) The prediction error (or the smoothness of Φ) for cases in (a) and (b) respectively, where the leave-one-out scheme is used to calculate errors. (e) Assume that x causally influences y, the information of x has been encoded in M_y and consequently Φ: $M_y \to M_x$ maps a neighborhood of y to a neighborhood of x, implying Φ_{yx} is smooth. Thus a neural network $\mathcal{N}$ can be trained to approximate the map based on the measured data on M_x and M_y. (f) Assume that y has no impact on x, then M_x has no information from y. Training a neural network to approximate the unsmooth map Φ: $M_x \to M_y$ will fail.

$$\begin{aligned}X(t+1)&=X(t)\left[r_x-r_xX(t)-\gamma_{xy}Y(t)\right],\\ Y(t+1)&=Y(t)\left[r_y-r_yY(t)-\gamma_{yx}X(t)\right],\end{aligned}\tag{1}$$

where $r_x = 3.7$ and $r_y = 3.8$ are two coefficients. Here γ_{xy} and γ_{yx} are two coupling parameters and indicate the strength of causative effectiveness. In the first case, we set $\gamma_{xy} = 0$ and $\gamma_{yx} = 0.32$, which implies that X and Y have a driving-response relation, namely there is unidirectional causality from X to Y, while the inverse is not true. We use time series with length of 20 time points and apply the CMS method to the data set. The causality detection result obtained by CMS Algorithm is shown in Fig. 4(a), where nonzero R_{xy} and zero R_{yx} clearly fit with the unidirectional causality pattern. Then we set $\gamma_{yx} = 0.1$ and $\gamma_{xy} = 0.02$ which makes it a mutually coupled system, and thus there is mutual causative effectiveness between X and Y. With the same setting in the first case, we detect the causality between X and Y using CMS, as shown in Fig. 4(b). The detected result with $R_{xy} = 0.69$ and $R_{yx} = 0.32$ not only shows the bidirectional causality between X and Y but also indicates that the relative strength of the causative effectiveness from X to Y is stronger than the inverse direction.

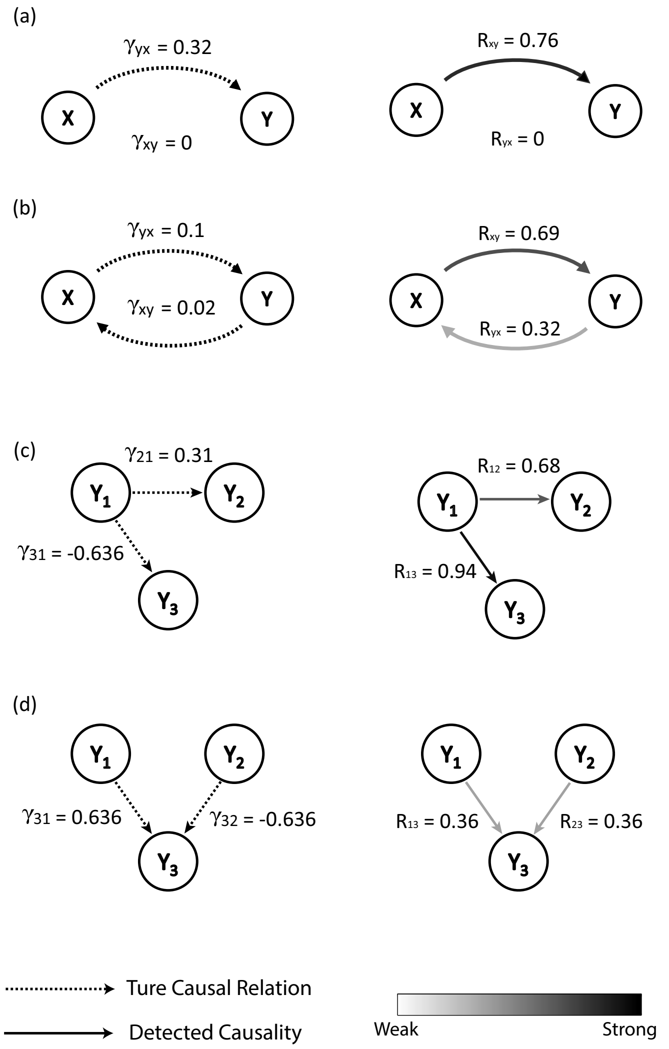

Figure 4 | Coupling relationship patterns (coupling strength γ in the left column) and the corresponding causality patterns (detected index R in the right column), where only the significantly detected causal relations above threshold are shown. (a) Unidirectional causality pattern in the 2 species model. (b) Bidirectional causality pattern in the 2 species model. (c) Fan-out causality pattern. (d) Fan-in causality pattern.

Here it is stressed that within one system when all the other conditions are the same, as the coupling strength increases, information becomes more distinct in the causally influenced variables, and consequently larger causality indices will be detected. However, this strength of causative effectiveness is relative but not absolute, i.e., the relation is not always monotonous between the coupling parameter values and the coupling strength. Therefore, the detected index reflects the relative strength of the causative effectiveness between different pairs of variables within one system. To this end, we consider varying coupling strength values. In the unidirectional case, we fix $\gamma_{xy} = 0$ and vary the values of γ_{yx} in the range [0, 0.32]. For each value of γ_{yx}, we generate data and use the CMS algorithm to detect causal relations with the same setting. The result is shown in Fig. 5(a) where the detected R_{yx} is always zero while R_{xy} firstly jumps from zero to nonzero, then shows an ascending trend as the coupling strength increases. In the bidirectional case, we consider the following varying form: $\gamma_{yx} = \alpha, \gamma_{xy} = 0.2 - \alpha$ where α is a varying factor in the range [0, 0.2] measuring the coupling strength in both directions. With the same setting as in the unidirectional case, the detected result is shown in Fig. 5(b) where zero causality indices reflect zero couplings and the detected causality indices R_{xy} and R_{yx} show ascending or descending trend as the coupling strength varies in the same way. Moreover, for small α, it clearly shows $R_{yx} > R_{xy}$ which coincides with the fact of relative stronger causative effectiveness from X to Y, and vice versa.

Then we consider a more complicated system involving three variables as follows:

$$Y_j(t+1)=Y_j(t)\left(\gamma_{jj}-\sum_{i=1,2,3}\gamma_{ji}Y_i(t)\right),\ j=1,2,3,\tag{2}$$

where γ_{ij} are coupling parameters. With particular settings of the coupling parameters, shown in Supplementary Information, the causal relations between the three variables can show fan-out or fan-in patterns, as shown in Figs. 4(c) and (d). For the fan-out case in Fig. 4(c), there are two unidirectional couplings from Y_1 to Y_2 and Y_3, while Y_2 and Y_3 have no direct relationship with each other. Since Y_2 and Y_3 are both driven by the common source from Y_1, the dynamics of Y_2 and Y_3 both contain the information from Y_1. Thus the time series $Y_2(t)$ and $Y_3(t)$ are correlated but have no causality between them, which is a difficult situation for causality detec-

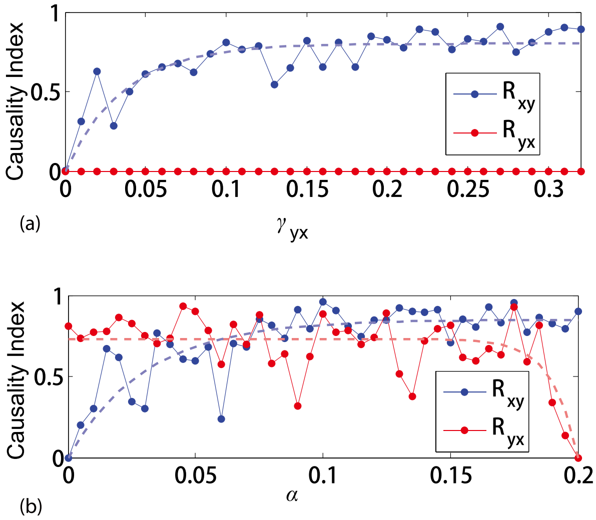

Figure 5 | Causality index detected for varying the coupling strength values. Dotted lines are the fitted trend curves. (a) Unidirectional case. (b) Bidirectional case.

tion[7]. Here we apply the CMS method to the time series with length of 20 points, and get the mutual relationship between Y_1, Y_2 and Y_3, which is shown in Fig. 4(c), where only the detected causality over significance threshold is shown. The result in Fig. 4(c) indicates that we can detect causality from Y_1 to Y_2 and Y_3 but not vice versa. Furthermore, there is no causal relationship detected between Y_2 and Y_3, which confirms that our method is effective for common source causality pattern even with short-term data. As for the fan-in case in Fig. 4(d), there are two unidirectional couplings to Y_3, i.e., Y_3 is driven by both Y_1 and Y_2 simultaneously. With the same setting as the fan-out case, we use CMS to detect the mutual relationship between the three variables, as shown in Fig. 4(d). The result illustrates that we can correctly detect the fan-in causality pattern. Here we stress that the strength of the detected nonzero causality in Fig. 4(d) is much weaker than the previous cases. Actually, a fan-in motif is generally considered as a difficult pattern to infer[17], this is mainly due to the fact that the dynamics of Y_3 are affected by both Y_1 and Y_2 at the same time, which weakens the effect of each single driving force.

The above cases validate that our method can be effective for discrete-time dynamical systems. To test our method with continuous-time systems, we consider the Lorenz system driven by a chaotic signal from the Rössler system, which was used as a benchmark in ref. 11. We use the standard parameter values with which the coupled system has chaotic dynamics. We assume that 50 time points with an even measurement interval are observed from the systems and use CMS to detect the causality between the two systems. The detected causality indices are $R_{LR} = 0$ and $R_{RL} = 0.27$, which clearly shows the unidirectional causal relation from the Rössler system to the Lorenz system.

Here we note that though we use the normalized error in CMS Algorithm, the final causality index R_{YX} for non-causal situation does not reach exactly zero. Therefore in order to decide whether the causality relation exists, we set a threshold value ξ based on the significance test[18,27–29]. Our statistical analysis is based on the permutation test: we run 1000 independent permutations uniformly at random, shuffle the time points according to the permutations, and run CMS on the shuffled data. With the empirical distribution, we estimate the threshold as $\xi = 0.001$ at a significance level $p < 0.05$, i.e., we treat a causality index below 0.001 as zero. The details of the significant test is shown in Supplementary Information.

The above illustrations show that our method is effective even for the situations where only short-term data can be obtained. On the other hand, in many occasions, owing to strong nonstationary and irregular behavior of many real-world systems, the causal relationship between system variables may switch quickly and thus even if a long time series is available the causality detected based on the long-term time series is meaningless. To consider such problems, we assume that the coupling parameters in system (1) are no longer constant but switching at random intervals between two sets of values so that the causal relationship between the two variables change from time to time, as shown in Fig. 6. We assume that a time series of 1000 time points are measured for the system, and it is obvious that using the whole time series to calculate one constant causality between X and Y will yield a false result. Therefore we use a time window of 20 time points moving along the whole time series and use the CMS method to detect causality from every short piece of data during one time window. Figure 6 shows the detected result, where the dashed square waves represent the random switching of the coupling parameters between zero and nonzero values and the solid lines represent the detected strengths of causative effectiveness over each time window. The switching value of γ_{xy} decides whether or not there is causative effectiveness from Y to X and it is clear that the detected causality from Y to X coincides with the square wave of γ_{xy} quite well in Fig. 6(a). The similar result can be observed for the causality from X to Y and the switching values of γ_{yx} in Fig. 6(b), which confirms the effectiveness of our method for such causality-varying situations.

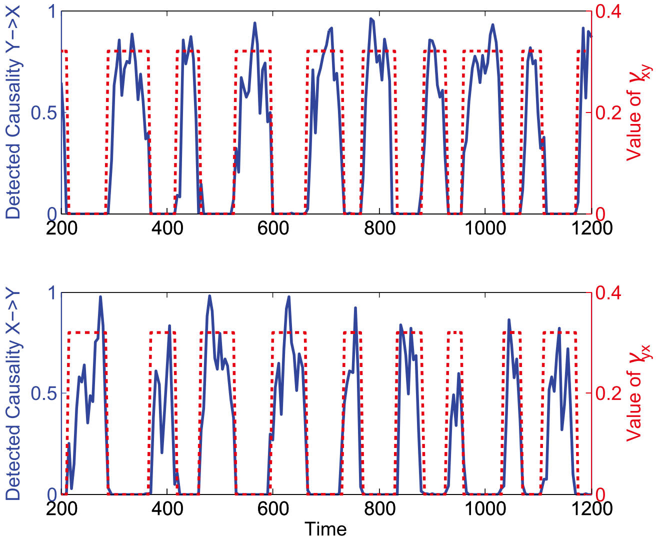

Figure 6 | Causality detection for a parameter-varying system in a piecewise manner. The dashed square waves represent the random switching of the coupling parameters between zero and nonzero values, and the solid lines represent the detected strengths of causative effectiveness over each time window.

Unraveling gene regulatory networks. How to infer gene regulatory interactions from transcriptomics time-resolved data, and further unravel the gene regulatory network (GRN) is of paramount importance to gain a deeper insight into the complexity and functions of the underlying biological systems. Due to the limit of experiment technique and other constraints, usually only very short-term and often noisy timeresolved measurements can be available in gene expressions. Though various methods based on Bayesian inference, regression analysis, econometrics models and standard similarity measures have been used to analyze such short time series data[17–19], inferring genetic networks from short data is still regarded as an 'ill-posed' inverse problem and a challenging task[20,21].

Here, we note that in a gene regulatory network, the regulation mechanism obeys some biochemistry rules and thus the regulatory dynamics can be described by standard kinetics models, such as Michaelis-Menten and Hill kinetics[30,31]. Therefore, the regulatory interactions can be measured by causal relationship in nonlinear dynamical systems and the proposed CMS method can be particularly suitable for such a task, i.e., reverse engineering GRN from short-term data.

Since in the real time-resolved expression data, e.g., microarray chip data, not every regulatory subnetwork contains information of all the participating genes, particularly over a specific time period and a specific condition of interest. These facts render it challenging to give a comprehensive evaluation of network inference with real time-resolved expression data. On the other hand, it is widely accepted these years to evaluate inference methods using standard synthetically generated data sets[17–19]. Therefore, before applying our method with real data, we give a comprehensive validation with synthetically generated data sets of the bacterium *E. coli*[30], as described in ref. 32.

Specifically, we consider a subsystem consisting of 50 genes picked out randomly from the whole network, whose regulatory relation is shown in Fig. 7(a). The regulation network of the selected subsystem consists of several clusters, as illustrated in Fig. 7(a), which can well approximate the statistical properties of the whole network[30]. Here each node's dynamics is governed by Michaelis-Menten or Hill kinetics, so that the simulated gene expression time series are similar to real microarray measurements. We assume that only 10 time points are available for each gene expression time series to simulate the real experimental measurements and use CMS to detect causality between each pair of genes. In order to evaluate the inference efficiency of CMS to detect GRN structure, we use the resulting receiver operating characteristics (ROC) curves, which plot the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Actually, ROC curves show the relative trade-offs between benefits of correctly inferred links (TPR) and the drawbacks of incorrectly inferred links (FPR) with different values of threshold to identify a link. The ROC curve for noise-free data is shown in Fig. 7(b), from which it can be concluded that the ROC curve is very close to the perfect classification; i.e., when FPR is controlled to be less than 2.5%, TPR can reach more than 90%, and the area under curve (AUC) is 0.98 which is very close to be one. Therefore, the CMS method can achieve a very good result for noise-free data. Meanwhile, generally noise-free expression data cannot be obtained in real experiment, and thus we need to further test the robustness of the method to include various kinds of noise in the time series. Here we consider three kinds of noise simultaneously, namely, the biological noise, the experimental noise and noise on correlated inputs with three different noise intensities, namely 0.1, 0.2 and 0.3. Figure 7(b) shows the three ROC curves for data with noisy perturbations, where we can conclude that though the ROC curves for noisy data are lower than the ROC curve for noise-free data, the accuracy is still very high; especially all the three ROC curves have AUC values around 0.9.

Moreover, to consider the influences of network properties such as size and degree, we further test two additional data sets for *S.cerevisiae* with 100 and 150 genes respectively. The selected subsystems are shown in Figs. 8(a) and (b), and based on 10 output of the gene expressions, we apply CMS to infer the network interactions respectively. The ROC curves for both non-noise and noise disturbed cases are given in Figs. 8(c) and (d). Here, it is noticed that there are two central genes which regulate other downstream genes at the same time for the selected subnetworks of *S.cerevisiae*, which makes fan-in and fan-out motifs abundant. As discussed in Figs. 4(c) and (d), the fan-in motif may lead the causality less significant to be detected, making it more difficult to detect the true network.

Next, we test our method with real gene expression data. Here we consider the data of the laboratory rat (Rattus norvegicus) cultured cells sampled from suprachiasmatic nucleus (SCN) for studying circadian rhythm, where the gene expression profiles are measured with Affimetrix microarray (Genechip Rat Genome 230 2.0)[33–35]. To elucidate the gene regulation network architecture, we select the data set consisting of 16 measured time points after the drug perturbation in the 19th hour. For the mammalian circadian clocks, it has been identified that there are around 17 genes involved in the core regulation network, where the transcriptional circuits are formed by regulation of *E*/*E*′ boxes, *DBP*/*E4BP4* binding elements and *RevErbA*/*ROR* binding elements respectively[36,37]. Moreover, besides the gene-level interactions, there are also regulation interactions at the protein level; e.g., the transcription factor *Clock* is phosphorylated by *PFK* family genes and the crytochrome genes *Cry*1 and *Cry*2 are phosphorylated by *MAPK* family genes[33]. Therefore, we consider the 17 core circadian genes as well as 18 kinase genes, whose relations are depicted in Fig. 9(a). With 16 time points measured for each gene's expression, we apply the CMS method to detect the regulation relation between all the selected 35 genes. As a comparison, we also apply IOTA, partial IOTA[18] and CCM[2] to the same data set, where IOTA is a newly proposed permutation-based asymmetric association measure to detect regulatory links from very short time series and CCM is a mutual cross map-based method. Based on the core regulation network in ref. 36, we carry out the ROC analysis for the regulation detection, and the results are shown in Fig. 9(b).

Here, we stress that inference of GRN based on only one single short-term data set is a challenging task due to the extremely short measurements. The existing methods for GRN inference can usually reach an AUC around 0.7 for synthetical data but only around 0.5 for real experimental data[21]. The CCM method, which relies on finding nearest neighbors and thus requires long-term data for the convergence, has AUC around 0.5 in Fig. 9(b). Particularly, the core transcriptional circuits of mammalian circadian clocks consist of complexly integrated regulatory loops involving three kinds of middle elements: *E*/*E*′ boxes, *DBP*/*E4BP4* binding elements and *RevErbA*/*ROR* binding elements. Therefore in the circadian data set we used here, there are many fan-in motifs and the relation between two interacting subsystems may no longer be monotonic, and thus the IOTA method and the partial IOTA method may lead to a false result, as shown in Fig. 9(b) where the AUC of the IOTA method is less than 0.5.

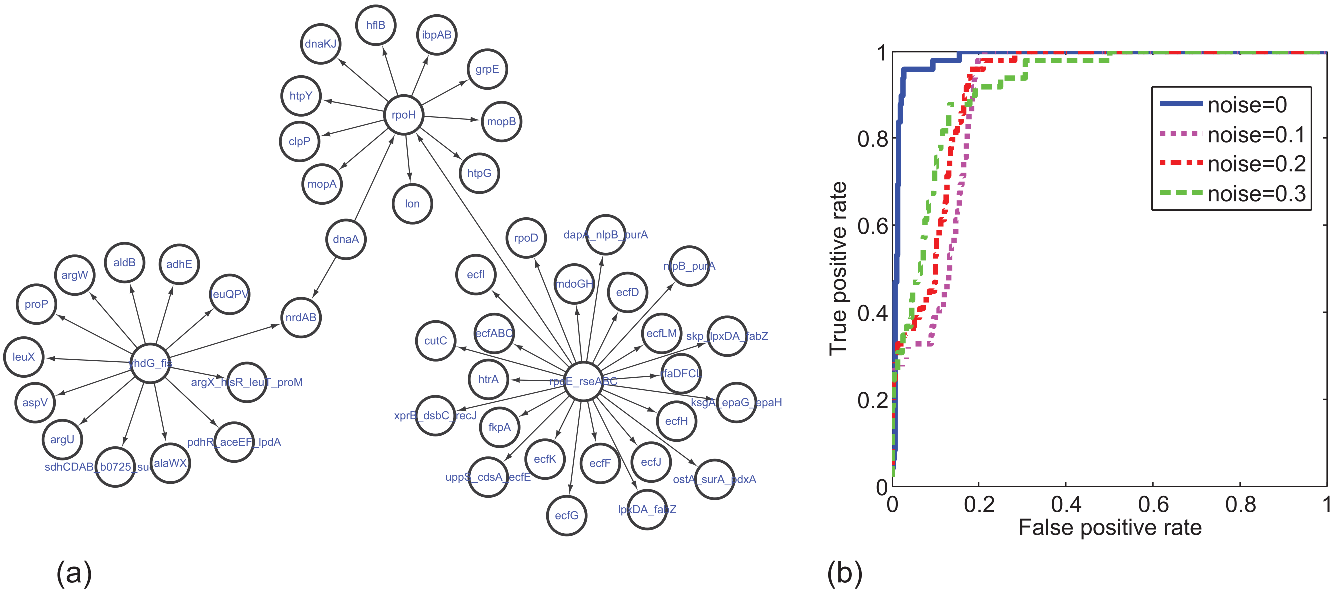

Figure 7 | (a) Regulatory network with the selected 50 genes of *E. coli.* (b) The ROC curves of the detection results by our method (CMS), with different levels of the noise condition.

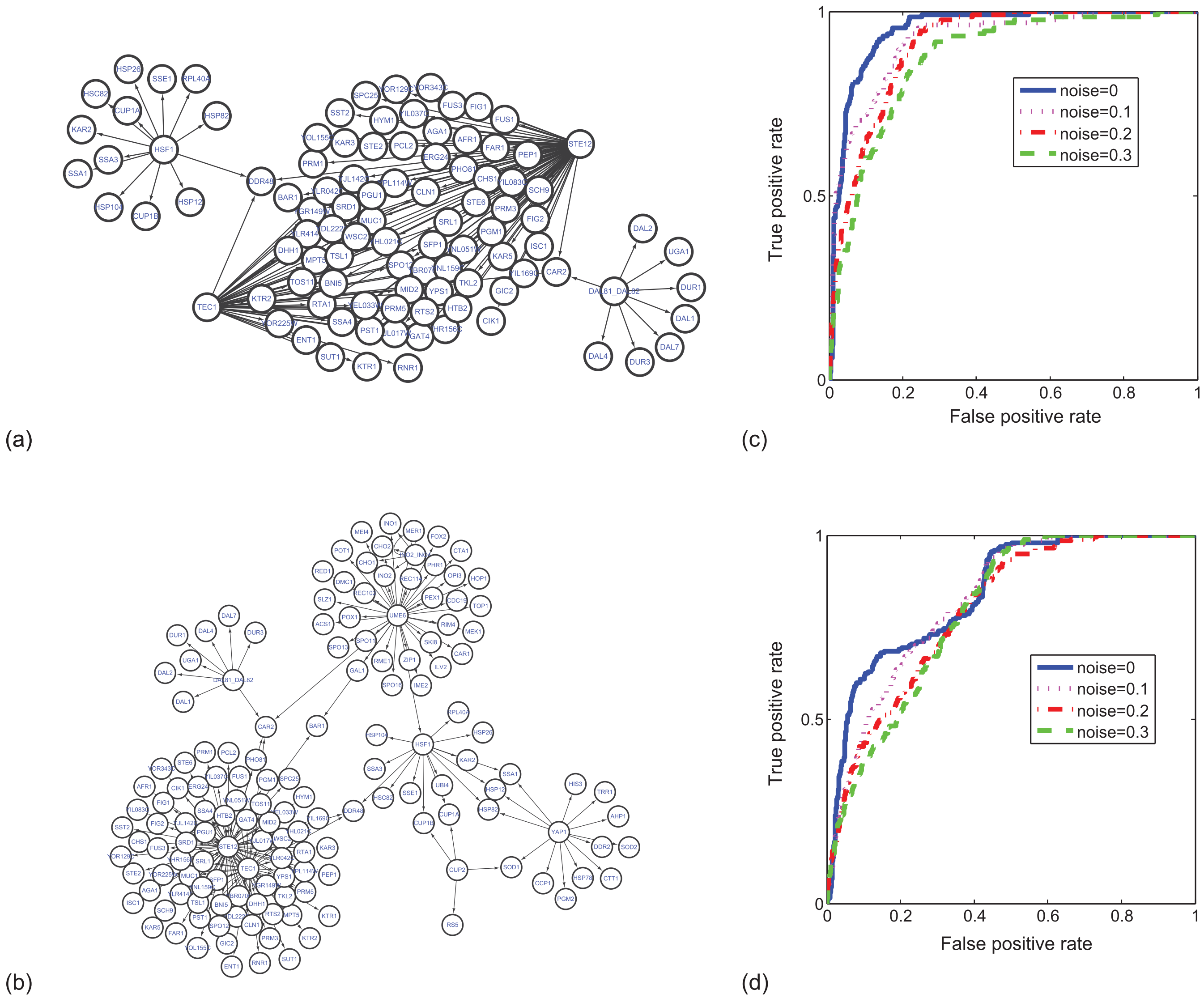

Figure 8 | (a) Regulatory network with the selected 100 genes of *S.cerevisiae.* (b) Regulatory network with the selected 150 genes of *S.cerevisiae.* (c) The ROC curves of the results for the network in (a) by our method, with different levels of the noise condition. (d) The ROC curves of the results for the network in (b) by our method, with different levels of the noise condition.

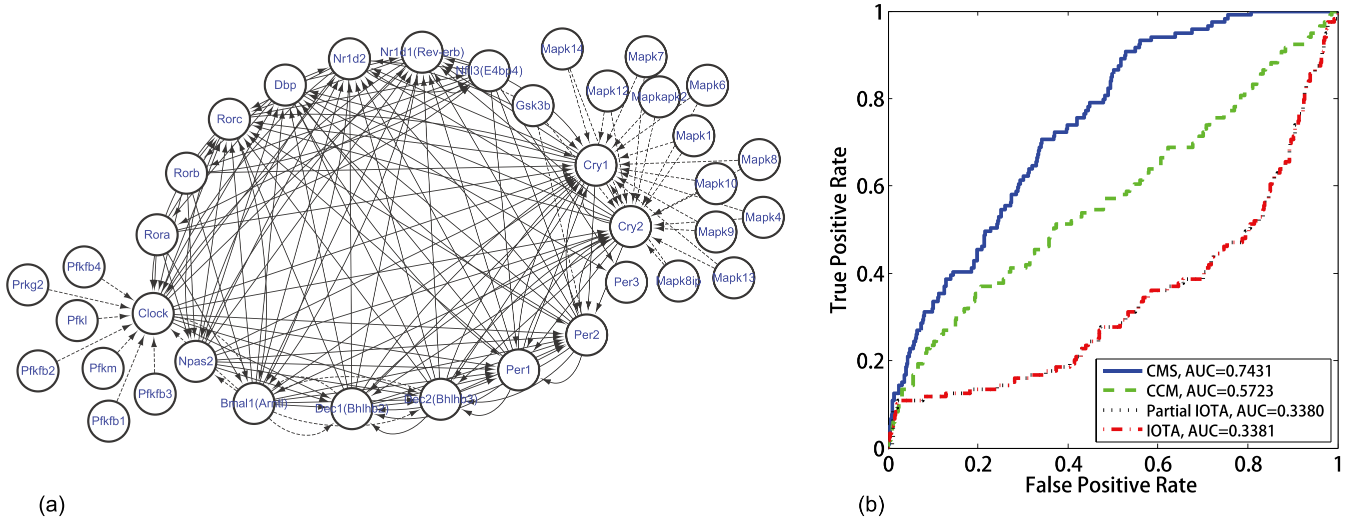

Figure 9 | (a) Regulatory network with the selected circadian genes, where the solid lines indicate gene-level regulations and the dashed lines imply protein-level interactions. (b) The ROC curves of the results, with four methods tested on the same data set.

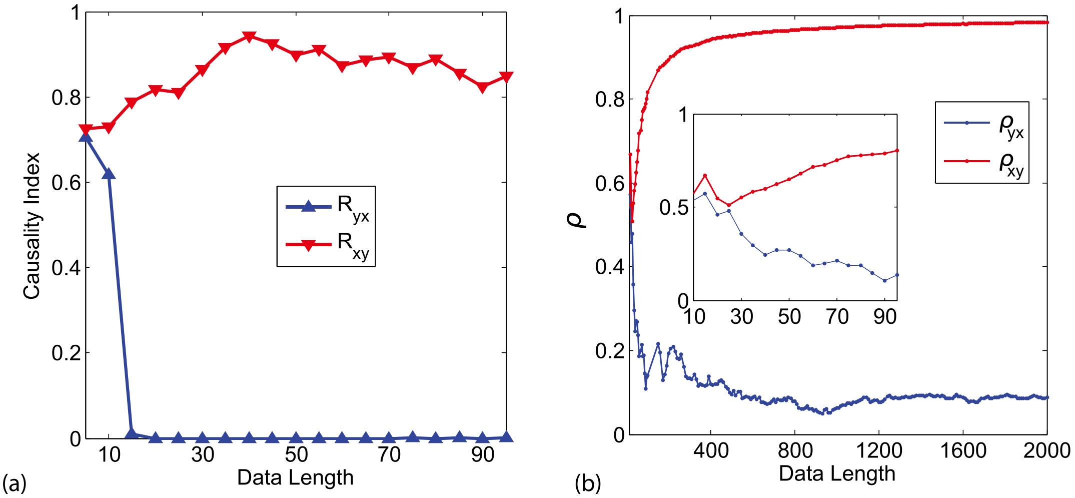

Figure 10 | (a) The causality detected by CMS based on different lengths of time series. (b) The causality detected by CCM based on different lengths of time series, where the inset is the enlarged part for the same data length as in (a).

As shown for the both synthetical GRN data and the real GRN data, our CMS method which is particularly effective for short-term data can achieve a very good result in a much accurate way in ROC analysis.

Discussion

In this section, we discuss several issues related to CMS' application in real situations. First a question arises naturally: how short can be the time series so that CMS can be effective? Or explicitly, what is the lower bound of the length of time series that CMS requires to guarantee a reliable result. Intuitively, the longer the time series can be observed, the more information the data can provide. Therefore just as pointed out in ref. 2, one should consider the convergence of the causality index over different lengths of time series and use the limit value as the truly detected causality index. Here we use system (1) with unidirectional causality setting as a benchmark to test the lower bound for the CMS method. The result is shown in Fig. 10(a), which indicates that with around 20 time points, it is enough for CMS to distinguish zero and nonzero causality. Meanwhile, as a comparison, we also test the CCM[2] method for the same data set. Since the CCM uses the convergence of nearest neighbors to detect causality, it needs much longer data length to converge. As shown in Fig. 10(b) as well as the inset there, CCM cannot distinguish zero and nonzero causality with short length of data, and with as long as 2000 time points, CCM can give a trend of convergence for nonzero causality though still no convergence for zero causality is achieved. Therefore, we conclude that CMS method can be effective for short-term data with length $n \sim O(10)$ while the existing neighborhood-based method requires data with length $n \sim O(10^3)$ to reach a reasonable result.

Then we compare our method with some representative existing methods for causality detection. Here we choose two kinds of methods for comparison, namely, the mutual cross map based on nearest neighbors and the composition alignment method. For the former, we use the newly proposed CCM[2] and for the latter, we use IOTA which is purposely designed for inferring gene networks from short time series[18]. We test all the numerical results in our paper with the same condition for the two methods, and the comparison results for the theoretical models are shown in Fig. 11. Moreover, for the gene networks, we consider several criteria to compare the methods, i.e., we consider the area under the ROC curve (AUC(ROC)), the Youden index (YOUDEN = max(the true positive rate - the false positive rate)), and the area under the Precision/Recall curve (AUC(PvsR)) which is based on the comparison between the true edges and the inferred ones. The results for these ROC analysis are shown in Table 1. Generally, it is suggested that a method has an excellent performance if conditions $AUC(ROC) > 0.8$; $YOUDEN > 0.5$ and $AUC(PvsR) > 0.05$ are satisfied simultaneously[19]. We highlighted the scores with the excellent performances in Table 1. Clearly, we see that CMS performs well in all the three cases. Since CCM needs long term data for convergence, the accuracy of results by CCM based on short term data here is poor. As for IOTA, it is specifically designed for gene network inference, and one crucial point of the IOTA approach lies on the assumption that two interacting genes have monotonic relationship. Therefore, for a general nonlinear dynamical system or gene expression which does not obey the monotonic assumption, the IOTA method may fail.

The above comparison results also imply that R_{xy} designed in CMS Algorithm can indicate the relative probability or the strength of

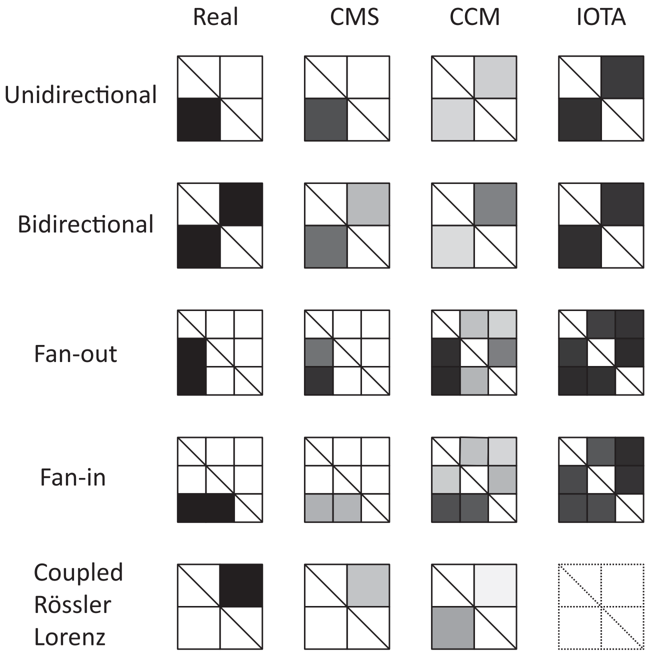

Figure 11 | **Comparison results for three methods on theoretical models.** The left column shows the causality patterns in 5 models, where black blocks shows causality from vertical variables to horizonal variables. The gray scale represents the strength of the detected causality between 0 (white) and 1 (black).

Table 1 | Comparison results for three method on four GRN cases, where three criteria are listed and the scores with the excellent performances are highlighted

	AUC(ROC)			YOUDEN			AUC(PvsR)		
	CMS	CCM	IOTA	CMS	CCM	IOTA	CMS	CCM	IOTA
Ecoli50	0.98	0.43	0.88	0.93	0.15	0.67	0.45	0.02	0.16
Yeast100	0.95	0.68	0.75	0.79	0.51	0.5	0.18	0.02	0.17
Yeast150	0.84	0.59	0.74	0.53	0.28	0.51	0.05	0.01	0.12
SCN	0.74	0.53	0.37	0.38	0.13	0.03	0.25	0.14	0.09

causality taking values between 0 and 1. Therefore, we can use the CMS index R to detect whether there is causality and how strong the causal relation is between two variables, as shown in the gray scale used in Fig. 4 and Fig. 11. Note that though the CMS method uses the prediction error to measure causality, it is different from the measurement of the Granger method which uses a series $y(t)$ to predict $y(t)$ by constructing $x(t) \rightarrow y(t)$ correlation; our method uses a series $y(t)$ to predict $x(t)$ by constructing a $y(t) \rightarrow x(t)$ map, for inferring the causality from $x(t)$ to $y(t)$. It is also noted that the prediction error based methods cannot detect autoregulation, i.e., the causal effect from one variable to itself.

In conclusion, based on the state space reconstruction theory for nonlinear dynamics, we have developed a new method of CMS to detect causality between variables, even with short observed time series. The key idea of our method is to detect causative effectiveness by measuring the smoothness of the cross map between two observed variables rather than finding the nearest neighbors, thereby avoiding the requirement of long-term time series data. The method is validated with both theoretical benchmark models and real-world data from gene networks. Our method is particularly effective in situations where only short-term data are available, such as high throughput biological data. In this paper we adopted a neural network model to train a smooth map, and other methods constructing a smooth map can be also used in a similar way. As a future topic, we will consider to extend this method further to detect the causal relations of the measured variables just before the critical transitions[38–40] and high dimensional measured variables of nonlinear dynamics[41].

1. Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438 (1969).
2. Sugihara, G. *et al.* Detecting causality in complex ecosystems. *Science* **338**, 496–500 (2012).
3. Haufe, S., Nikulin, V. V., Mller, K.-R. & Nolte, G. A critical assessment of connectivity measures for EEG data: A simulation study. *NeuroImage* **64**, 120–133 (2013).
4. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **85**, 461 (2000).
5. Paluš, M., Komárek, V., Hrnčíř, Z. & Štěrbová, K. Synchronization as adjustment of information rates: detection from bivariate time series. *Phys. Rev. E* **63**, 046211 (2001).
6. Feldhoff, J. H., Donner, R. V., Donges, J. F., Marwan, N. & Kurths, J. Geometric detection of coupling directions by means of inter-system recurrence networks. *Phys. Lett. A* **376**, 3504–3513 (2012).
7. Hirata, Y. & Aihara, K. Identifying hidden common causes from bivariate time series: a method using recurrence plots. *Phys. Rev. E* **81**, 016203 (2010).
8. Chen, Y., Rangarajan, G., Feng, J. & Ding, M. Analyzing multiple nonlinear time series with extended Granger causality. *Phys. Lett. A* **324**, 26–35 (2004).
9. Ancona, N., Marinazzo, D. & Stramaglia, S. Radial basis function approach to nonlinear Granger causality of time series. *Phys. Rev. E* **70**, 056221 (2004).
10. Schiff, S. J., So, P., Chang, T., Burke, R. E. & Sauer, T. Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble. *Phys. Rev. E* **54**, 6708 (1996).
11. Quyen, M. L. V., Martinerie, J., Adam, C. & Varela, F. J. Nonlinear analyses of interictal EEG map the brain interdependences in human focal epilepsy. *Physica D* **127**, 250–266 (1999).
12. Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M. & Bhattacharya, J. Causality detection based on information-theoretic approaches in time series analysis. *Phys. Rep.* **441**, 1–46 (2007).
13. Chicharro, D. & Andrzejak, R. G. Reliable detection of directional couplings using rank statistics. *Phys. Rev. E* **80**, 026217 (2009).
14. Arnhold, J., Grassberger, P., Lehnertz, K. & Elger, C. A robust method for detecting interdependences: application to intracranially recorded EEG. *Physica D* **134**, 419–430 (1999).
15. Deyle, E. R. *et al.* Predicting climate effects on Pacific sardine. *Proc. Nat. Acad. Sci. USA* **110**, 6430–6435 (2013).
16. Wang, Y., Joshi, T., Zhang, X. & Chen, L. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics* **22**, 2413–2420 (2006).
17. Marbach, D. *et al.* Revealing strengths and weaknesses of methods for gene network inference. *Proc. Nat. Acad. Sci. USA* **107**, 6286–6291 (2010).
18. Hempel, S., Koseska, A., Kurths, J. & Nikoloski, Z. Inner composition alignment for inferring directed networks from short time series. *Phys. Rev. Lett.* **107**, 054101 (2011).
19. Hempel, S., Koseska, A., Nikoloski, Z. & Kurths, J. Unraveling gene regulatory networks from time-resolved gene expression data–a measures comparison study. *BMC bioinformatics* **12**, 292 (2011).
20. Wang, X., Wu, M., Li, Z. & Chan, C. Short time-series microarray analysis: Methods and challenges. *BMC Systems Biology* **2**, 58 (2008).
21. Wang, M. *et al.* LegumeGRN: A Gene Regulatory Network Prediction Server for Functional and Comparative Studies. *PloS one* **8**, e67434 (2013).
22. Takens, F. *Dynamical Systems and Turbulence* Rand, D. A. & Young, L. S. (ed.) 366–381 (Springer-Verlag, New York, 1981).
23. Sauer, T., Yorke, J. A. & Casdagli, M. Embedology. *Journal of statistical Physics* **65**, 579–616 (1991).
24. Kantz, H. & Schreiber, T. *Nonlinear time series analysis* (Cambridge University press, Cambridge, 2004).
25. Kennel, M. B. *et al.* Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A* **45**, 3403–3411 (1992).
26. Park, J. & Sandberg, I. W. Universal approximation using radial-basis-function networks. *Neural Comput.* **3**, 246–257 (1991).
27. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).
28. Luo, Q., Ge, T. & Feng, J. Granger causality with signal-dependent noise. *Neuroimage* **57**, 1422–1429 (2011).
29. Guo, S. *et al.* Uncovering interactions in the frequency domain. *Plos. Comput. Bio.* **4**, e1000087 (2008).
30. Van den Bulcke, T. *et al.* SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics* **7**, 43 (2006).
31. Hofmeyr, J.-H. S. & Cornish-Bowden, H. The reversible Hill equation: how to incorporate cooperative enzymes into metabolic models. *Comput. Appl. Biosci.* **13**, 377–385 (1997).
32. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.* **31**, 64–68 (2002).
33. Wang, Y., Zhang, X.-S. & Chen, L. A network biology study on circadian rhythm by integrating various omics data. *OMICS* **13**, 313–324 (2009).
34. Kawaguchi, S. *et al.* Establishment of cell lines derived from the rat suprachiasmatic nucleus. *Biochem. Bioph. Res. Co.* **355**, 555–561 (2007).
35. Morioka, R. *et al.* Phase Shifts of Circadian Transcripts in Rat Suprachiasmatic Nucleus. *In The Second International Symposium on Optimization and Systems Biology*. 109C114 (World Publishing Corporation, Lijiang, China, 2008).
36. Ueda, H. R. *et al.* System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nat. Genet.* **37**, 187–192 (2005).
37. Ko, C. H. & Takahashi, J. S. Molecular components of the mammalian circadian clock. *Hum. Mol. Genet.* **15**, R271–R277 (2006).
38. Chen, L. *et al.* Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci. Rep.* **2**, 342 (2012).
39. Liu, R. *et al.* Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. *Medicinal Research Reviews* **34**, 455–478 (2013).
40. Liu, R. *et al.* Identifying critical transitions of complex diseases based on a single sample. *Bioinformatics* **30**, 1579–1586 (2014).

41. Ma, H. *et al.* Predicting Time-Series from Short-Term High-Dimensional Data. *Int. J. Bifurcat. Chaos* **24**, 1430033 (2014).

Acknowledgments

We thank Prof. Hajimi Tei on the gene expression data and prior information collection. This research was supported by the Aihara Project, the FIRST program from JSPS, initiated by CSTP. This work was also supported by the NSFC (Nos. 61134013, 91029301, 61072149, 11301366, and 11326035), the Knowledge Innovation Program of CAS (No. KSCX2-EW-R-01), and 863 project (No. 2012AA020406).

Author contributions

H.M., L.C. and K.A. conceived the research; H.M. and L.C. performed the experiments, analyzed the data, and prepared the figures; All the authors wrote the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Ma, H., Aihara, K. & Chen, L. Detecting Causality from Nonlinear Dynamics with Short-term Time Series. *Sci. Rep.* **4**, 7464; DOI:10.1038/srep07464 (2014).