Inner Composition Alignment for Inferring Directed Networks from Short Time Series

S. Hempel,[1] A. Koseska,[2] J. Kurths,[1,3] and Z. Nikoloski[4]

[1]*Potsdam Institute for Climate Impact Research (PIK), Potsdam, Germany*
[2]*Interdisciplinary Center for Dynamics of Complex Systems, University of Potsdam, Potsdam, Germany*
[3]*Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB243UE, United Kingdom*
[4]*Systems Biology and Mathematical Modeling Group, Max Planck Institute for Molecular Plant Physiology, Potsdam, Germany*

(Received 18 February 2011; published 26 July 2011)

Identifying causal links (couplings) is a fundamental problem that facilitates the understanding of emerging structures in complex networks. We propose and analyze inner composition alignment—a novel, permutation-based asymmetric association measure to detect regulatory links from very short time series, currently applied to gene expression. The measure can be used to infer the direction of couplings, detect indirect (superfluous) links, and account for autoregulation. Applications to the gene regulatory network of *E. coli* are presented.

DOI: 10.1103/PhysRevLett.107.054101 PACS numbers: 05.45.Tp, 87.10.Vg, 87.18.Vf

Many systems can be regarded as complex networks of multiple interacting subsystems [1], e.g., social and economic networks [2], the climate system [3], the brain [4], or gene regulatory networks [5]. Hence, the data-driven reconstruction of these networks is a pressing research problem with valuable applications, where the analysis of multivariate time-resolved data is crucial to infer (causal) relationships. While the current technological advances facilitate the collection of an unprecedented amount of time series data, for several typical systems, particularly in biology, the time series are rather short (≈ 10 time points). Thus, standard association measures, e.g., information-theoretic, correlation, or model-based ones [6,7] may not resolve the couplings, while measures operating on symbolic dynamics appear less sensitive to the length of the time series [7,8]. Moreover, only a few measures address the important problem of directionality of couplings (e.g., Granger causality or transfer entropy [6]).

In this contribution, we develop a permutation-based measure, inner composition alignment (IOTA), denoted by ι, having the following merits: (i) It identifies (unidirectional and bidirectional) coupling and its directionality, (ii) it distinguishes direct from indirect coupling (similarly to partial correlation), (iii) it infers autoregulation (resulting from an internal adaptive mechanism by which a subsystem regulates itself [9]), a problem which, to the best of our knowledge, has not been addressed by any of the available association measures, (iv) it is applicable to very short time series, and (v) it does not depend explicitly on time.

Next, we define IOTA and analyze its properties. Given the time series $y^{(l)}$ and $y^{(k)}$ of the subsystems l and k over the same time domains, let $\pi^{(l)}$ be the permutation which orders $y^{(l)}$ in a nondecreasing order, i.e., $\pi^{(l)}$: $\forall\ i\ [y^{(l)}(\pi^{(l)})]_i \leq [y^{(l)}(\pi^{(l)})]_{i+1}$. The series $g^{(k,l)} = y^{(k)}(\pi^{(l)})$ is the reordering of the time series $y^{(k)}$ with respect to $\pi^{(l)}$. The crucial point of our approach is that, in particular, for gene expression, the reordered time series of 2 interacting subsystems have been observed to be monotonically increasing functions [10]. To quantify the monotonicity, we count the number of intersection (crossing) points of the reordered time series with the horizontal lines which are drawn from each of the time points (Fig. 1). Thus, we can compute ι by

$$\iota^{(l\to k)} = 1 - \frac{\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} w_{ij}\Theta[(g_{j+1}^{(k,l)} - g_i^{(k,l)})(g_i^{(k,l)} - g_j^{(k,l)})]}{\Delta}, \tag{1}$$

where n is the length of the time series, $\Delta = \frac{(n-1)(n-2)}{2}$ is a normalization constant which corresponds to the maximum number of crossings, w_{ij} denotes a weight, and $\Theta[x]$ is the Heaviside step function,

$$\Theta[x] = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

For two coupled subsystems, the number of crossing points tends to zero, rendering a value for ι close to 1. In order to account for noise-induced fluctuations, we compare the

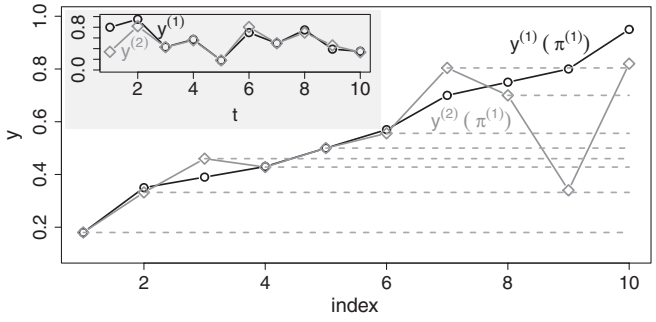

FIG. 1. Time series $y^{(1)}$ and $y^{(2)}$ reordered by the permutation $\pi^{(1)}$. Horizontal lines are drawn at points of $y^{(2)}$ $(\pi^{(1)})$. The plot in the inset shows the time series in their original order.

properties of ι using different weights (summarized in Table I). If the values of the time series are rescaled to the interval $[0, 1]$, ι fulfills the properties of a normalized association measure, independent of the chosen weight. Generally, IOTA is asymmetric, i.e., $y^{(l)}(\pi^{(k)}) \neq y^{(k)}(\pi^{(l)})$, which renders the inference of unidirectional links from short time series possible. As IOTA is permutation based, it is also capable of detecting nonlinear interactions, similarly to mutual information. Note that if the same permutation is applied to all time series, the value of ι does not change.

Additionally, IOTA can be used to address the problem of indirect (superfluous) couplings (links), and possible autoregulation. To identify indirect links, two permutations are applied consecutively. Given the subsystem m regulating the subsystems k and l directly, the pairwise measure predicts an additional link from k to l. To check whether the link is indirect, we determine the permutations $\pi^{(k)}$ and $\pi^{(m)}$ and evaluate whether applying the permutation composition $\pi^{(k)}(\pi^{(m)})$ on $y^{(l)}$ instead of the permutation $\pi^{(k)}$ alone changes the value of the measure. Hence, the partial version of IOTA [Eq. (2)] is formulated by comparing the triplets deduced from the pairwise measure:

$$\iota_p^{((k\to l)|(m\to k),(m\to l))} = |\iota(h^{(l,k,m)}) - \iota(g^{(l,m)})|, \qquad (2)$$

with $g^{(l,k)} = y^{(l)}(\pi^{(k)})$ and $h^{(l,k,m)} = y^{(l)}(\pi^{(k)}(\pi^{(m)}))$. Here, the value of ι_p is expected to tend to zero for $(k \to l)$ being an indirect link. For $k = l = m$, a low value of ι_p is obtained if the time series are almost monotonic, which on the other hand indicates a low probability for autoregulation.

Next, we briefly discuss the similarities and differences between IOTA and Kendall's τ, shown to most reliably infer coupling from very short time series of 10 time points (in comparison to other similarity measures) [7]. To calculate Kendall's rank correlation for each time series, the permutation is determined that arranges the respective series in nondecreasing order, namely, $\pi^{(1)}$ for $y^{(1)}$ and $\pi^{(2)}$ for $y^{(2)}$. If the matching values in $\pi^{(1)}$ and $\pi^{(2)}$ are linked together, then the number of intersections among these links is the number of discordant pairs: $n_d = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\Theta[(r_j^{(1)} - r_i^{(1)}) \times (r_i^{(2)} - r_j^{(2)})] = \frac{n(n-1)}{2} - n_c$, where $r_i^{(l)}$ is the rank of value $v_i^{(l)}$ in the time series $y^{(l)}$, associated with the permutation $\pi^{(l)}$, and $n_c = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\Theta[(r_i^{(1)} - r_j^{(1)})(r_i^{(2)} - r_j^{(2)})]$ is the number of concordant pairs. Finally, Kendall's τ is given by

$$\tau = 2\frac{n_c - n_d}{n(n-1)} = 1 - \frac{2}{n(n-1)}2n_d. \qquad (3)$$

In contrast, IOTA uses the permutation $\pi^{(1)}$, employed in the reordering of the first time series, to reorder the second one. Thus, ι includes the ordering information from one subsystem and its effect on the second. Applying $\pi^{(1)}$ on the ranks of $y^{(1)}$ and $y^{(2)}$ leads to the series $\rho^{(1)} = r^{(1)}(\pi^{(1)})$ and $\rho^{(2)} = r^{(2)}(\pi^{(1)})$, which are subsequently used to calculate the value of ι:

$$\iota^{(1\to 2)} = 1 - \frac{2}{(n-1)(n-2)}c, \qquad (4)$$

where the number of crossings in Fig. 1 matches $c = \sum_{k=1}^{n-2}\sum_{i=k+1}^{n-1}\sum_{j=i+1}^{n}\Theta[(\rho_k^{(2)} - \rho_j^{(2)})(\rho_i^{(2)} - \rho_k^{(2)})]\Theta[\rho_i^{(1)} - \rho_j^{(1)} + 2]w_{ij}$. It is important to note that the graphical representation of Kendall's τ and ι are comparable only if $y^{(1)}$ is monotonically increasing, i.e., $\rho = r$. However, in the general case, ι includes the ordering information which is neglected in the case of Kendall's τ.

In order to evaluate the capabilities of IOTA to infer directed networks particularly from very short gene expression time series, we apply ι to reconstruct the gene

TABLE I. Different weights w_{ij} for IOTA.

Uniform weighting	1
Arithmetic mean	$\frac{1}{2}(g_{j+1}^{(k,l)} + g_j^{(k,l)})$
Geometric mean	$\sqrt{g_{j+1}^{(k,l)} g_j^{(k,l)}}$
Harmonic mean	$2(\frac{1}{g_{j+1}^{(k,l)}} + \frac{1}{g_j^{(k,l)}})^{-1}$
Maximal excursion	$\max(\lvert g_{j+1}^{(k,l)} - g_i^{(k,l)}\rvert, \lvert g_j^{(k,l)} - g_i^{(k,l)}\rvert)$
Slope	$\lvert g_{j+1}^{(k,l)} - g_j^{(k,l)}\rvert$
Squared slope	$(g_{j+1}^{(k,l)} - g_j^{(k,l)})^2$

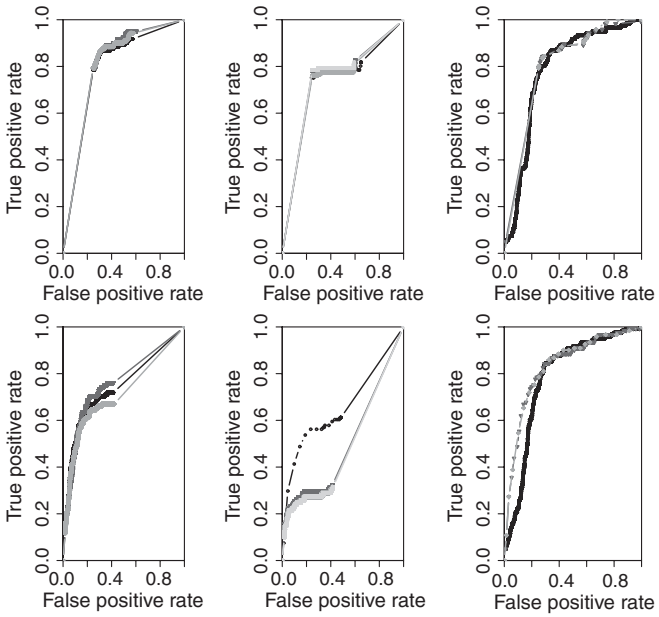

FIG. 2. ROC curves for a network of 100 genes of *E. coli* (upper panels: noise-free case; lower panels: noise level of 0.1). The left and middle panels are obtained for ι (superfluous removed) including a significance test with 10 000 realizations for different weights—left: slope (black), squared slope (dark gray), maximal excursion (gray); middle: uniform (black), arithmetic mean (dark gray), geometric mean (gray), harmonic mean (light gray). The right panels relate to the correlations: Pearson (black), Spearman (dark gray), Kendall (gray). These results do not incorporate a significance test.

regulatory network of the bacterium *E. coli*, as described previously in [11]. More precisely, we analyze a subnetwork of 100 genes (representing the nodes) that approximate significantly well the statistical properties of the whole network [12]. The investigated subnetwork is sparse, having 121 unidirectional links, 6 of which are autoregulatory. The dynamics of each node (gene) is governed by Michaelis-Menten and Hill kinetics, rendering the simulated gene expression time series very similar to real microarray mRNA measurements. Moreover, the considered gene expression time series consist of 10 time points each, corresponding to real experimental measurements.

Our statistical analysis is based on the following permutation test: We select 10 000 permutations uniformly at random, shuffle the data according to these permutations, recalculate ι, and estimate the empirical p values at the significance level 0.01. Additionally, we determine the significance of the direction $(l \to k)$ relying on an analogous permutation test, where we check whether the distance $\iota^{(l\to k)} - \iota^{(k\to l)}$ of the original time series is larger than that of the randomized ones. To test the reconstruction efficiency of IOTA and particularly the influence of different weights (Table I), we consider the resulting receiver operating characteristics (ROC) curves, which illustrate the change of the relative trade-offs between benefits [true positive rate (TPR)—correctly inferred links] and drawbacks [false positive rate (FPR)—incorrectly inferred links], while continuously tuning the threshold that is used to identify a link [13].

We investigate both deterministic and stochastic time series. Figure 2 illustrates that the range of values of ι depends on the choice of the weight, where all mean-based weights are less robust against the influence of noise than the slope-based ones. IOTA has the lowest noise sensitivity using the squared slope weight—an important feature, especially when dealing with biological data. Hence, we employ the squared slope weight in our further study.

By comparing the reconstruction efficiency of ι to those of the rank correlations, we observe that the lower boundary of the FPR is similar in both cases, which poses a direct control on the false positives. However, the ROC curve for ι is more continuous than those of the rank correlations. Hence, the network topology as inferred with ι is less sensitive to the threshold chosen to decide which nodes are to be linked. This is of particular value when dealing with experimental data.

Next, we compare various reconstruction scenarios obtained with IOTA and Kendall's τ, when different thresholds (as they were previously employed to obtain the ROC curves) are used to identify the links (Fig. 3). While in the deterministic case the proper choice of the threshold is evident, the situation is more complicated when stochasticity is present, since it is difficult to quantify the influence of noise. For instance, for noise intensity 0.1 and threshold 0.95, Kendall's rank correlation gives a TPR of less than 10% (FPR $\approx$ 1%), whereas a threshold of 0.5 renders a TPR $\approx$ 70% (FPR $\approx$ 25%). On the other hand, when IOTA is applied under the same conditions, the TPR at

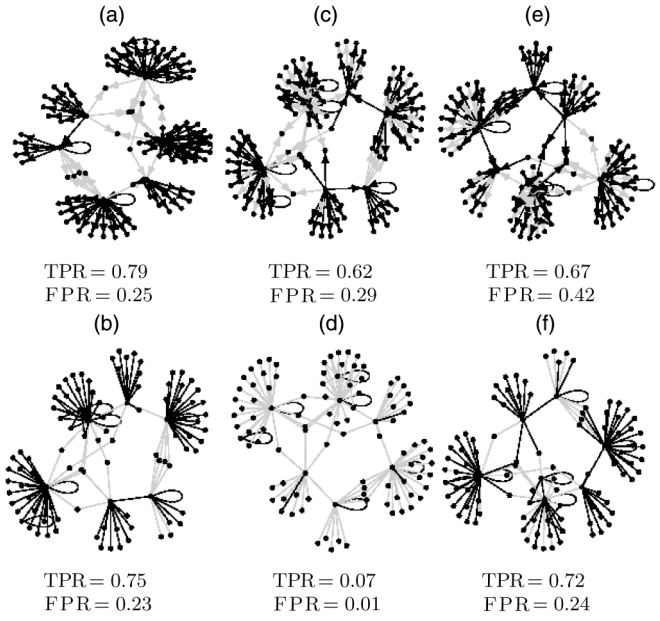

FIG. 3. Reconstruction of a regulatory network of 100 genes of *E. coli* from (a),(b) noise-free time series using threshold 1, (c), (d) time series simulated with noise level 0.1 using threshold 0.95, and (e),(f) with noise level 0.1 using threshold 0.5. Panels (a),(c),(e) show the networks obtained with IOTA, whereas (b),(d),(f) are obtained with Kendall's τ. The original network (in the lower panels the undirected version) is shown in light gray; correctly identified links are marked in black. False positive links are not shown.

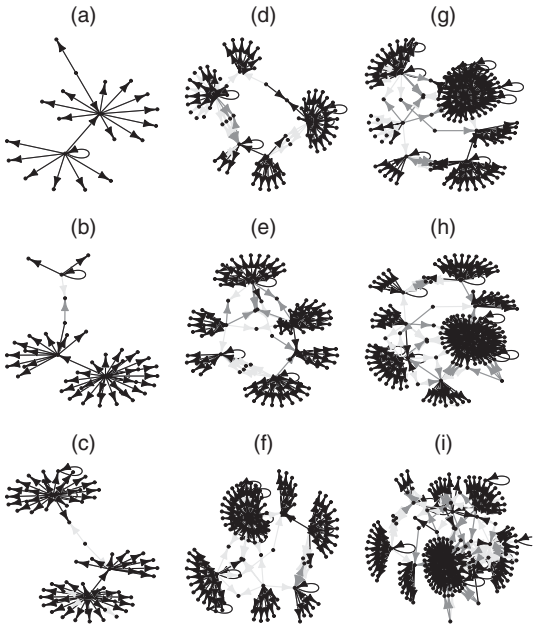

FIG. 4. Network reconstruction with IOTA for networks of different sizes [(a)–(i) refer to Table II] and stochastic time series (noise intensity 0.01) of 10 time points. Black links are obtained with threshold 0.75 (dark gray additionally with threshold 0.99); light gray indicates not identified links.

TABLE II. Reconstruction efficiency with IOTA (networks shown in Fig. 4, threshold 0.99).

	No. of nodes	Autoregulated		Unidirectional	Bidirectional	IOTA (Kendall)	
		Yes	No			TPR	FPR
(a)	20	1/1	0/19	20/20	0/0	1.00 (0.80)	0.73 (0.43)
(b)	40	2/2	0/38	39/41	0/0	0.95 (0.90)	0.81 (0.79)
(c)	60	3/3	0/57	60/62	0/0	0.92 (0.73)	0.38 (0.24)
(d)	80	4/4	1/76	73/89	0/3	0.73 (0.52)	0.23 (0.14)
(e)	100	6/6	0/94	98/121	3/5	0.79 (0.63)	0.27 (0.13)
(f)	120	10/10	0/110	124/147	3/5	0.81 (0.63)	0.28 (0.16)
(g)	140	14/14	0/126	160/179	3/5	0.84 (0.64)	0.27 (0.20)
(h)	160	15/16	14/144	170/210	5/12	0.79 (0.65)	0.21 (0.17)
(i)	180	19/19	0/161	179/255	2/13	0.65 (0.48)	0.19 (0.10)

both threshold levels is $\approx 65\%$ [FPR $\approx 30\%$ (40%) for threshold 0.95 (0.5)]. Hence, in contrast to Kendall's τ, where the number of correctly and falsely identified links is strongly dependent on the threshold, the values obtained with IOTA are almost constant. Thus, IOTA results in robust predictions with respect to varying thresholds, as demanded in practical applications. Furthermore, in contrast to Kendall's τ, which assumes all genes to be autoregulated per definition, IOTA infers correctly all of the included autoregulatory links, but also partially identifies genes which are not autoregulated (1% in the noise-free case and even 5% from the noisy time series).

In order to test the capabilities of IOTA, we modify the original source network to include additional bidirectional links. Subnetworks of various sizes and simulated time series of 10 time points each are generated from the modified network, as shown in Fig. 4 and Table II. Applied to networks of intermediate size (100–140 nodes), ι inferred approximately 60% of the bidirectional links present (Table II bidirectional). However, we observe that when applied to very short time series, IOTA identifies in most of the cases only one significant direction. Furthermore, similarly to the previous example, nearly all autoregulated genes are correctly identified, while slightly reducing the FPR for several of the investigated networks (Table II autoregulated). We also observe that the capability of IOTA to correctly infer couplings depends on the density of the network, performing particularly well on sparse networks (e.g., gene regulatory networks).

In summary, we introduce IOTA, a permutation-based measure, as an efficient tool to identify relations between subsystems, together with the associated directionality, currently possible only for a small number of measures operating exclusively on long time series. We find that IOTA is robust to noise and can be used to infer statistically significant (nonlinear) couplings from very short time-resolved data of gene expression. Moreover, even from short time series, IOTA can infer autoregulation and the direction of coupling under certain conditions (in particular, if the dynamics of the coupled systems involves small time delays). The dependence of IOTA's reliability on the length of the time series and the ability of the measure to analyze chaotic time series remains to be investigated in a future study.

We acknowledge support from the GoFORSYS project (Grant No. 0313924) funded by the German BMBF.

[1] R. Albert and A.-L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002); R. V. Donner, Y. Z. Zou, J. F. Donges, N. Marwan, and J. Kurths, New J. Phys. **12**, 033025 (2010).
[2] M. O. Jackson and A. Watts, J. Econ. Theory **106**, 265 (2002).
[3] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, Europhys. Lett. **87**, 48 007 (2009).
[4] M. Ding, Y. Chen, and S. L. Bressler, in *Handbook of Time Series Analysis* edited by B. Schelter, M. Winterhalder, and J. Timmer (Wiley-VCH Verlag, Weinheim, 2007), pp. 437–460.
[5] H. de Jong, J. Comput. Biol. **9**, 67 (2002).
[6] P. Capitani and P. Ciaccia, Data Know. Eng. **62**, 438 (2007); F. D. Gibbons and F. P. Roth, Genome Res. **12**, 1574 (2002); T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000); K. Hlavackova-Schindler, M. Palus, M. Vejmelka, and J. Bhattacharya, Phys. Rep. **441**, 1 (2007); W. Li, J. Stat. Phys. **60**, 823 (1990); S. Guo, A. K. Seth, K. M. Kendrick, C. Zhou, and J. Feng, J. Neurosci. Methods **172**, 79 (2008); J. Nawrath, M. C. Romano, M. Thiel, I. Z. Kiss, M. Wickramasinghe, J. Timmer, J. Kurths, and B. Schelter, Phys. Rev. Lett. **104**, 038701 (2010).
[7] S. Hempel, A. Koseska, Z. Nikoloski, and J. Kurths, BMC Bioinf. **12**, 292 (2011).
[8] N. Wessel, A. Suhrbier, M. Riedl, N. Marwan, H. Malberg, G. Bretthauer, T. Penzel, and J. Kurths, Europhys. Lett. **87**, 10 004 (2009).
[9] U. Alon, Nat. Rev. Genet. **8**, 450 (2007).
[10] T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and K. Marchal, BMC Bioinf. **7**, 43 (2006).
[11] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, Nat. Genet. **31**, 64 (2002).
[12] T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and K. Marchal, SynTRen generator, version 1.1.3, 2006, http://homes.esat.kuleuven.be/~kmarchal/SynTReN/software.html.
[13] T. Fawcett, Pattern Recogn. Lett. **27**, 861 (2006).