Bioinformatics, 32(14), 2016, 2143–2150
doi: 10.1093/bioinformatics/btw154
Advance Access Publication Date: 19 March 2016
Original Paper

Systems biology

Detecting critical state before phase transition of complex biological systems by hidden Markov model

Pei Chen[1], Rui Liu[2], Yongjun Li[1,*] and Luonan Chen[3,4,*]

[1]School of Computer Science and Engineering and [2]School of Mathematics, South China University of Technology, Guangzhou 510640, China, [3]Key Laboratory of Systems Biology, Innovation Center for Cell Signaling Network, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China and [4]Collaborative Research Center for Innovative Mathematical Modelling, Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan

*To whom correspondence should be addressed.

Associate Editor: Cenk Sahinalp

Received on January 18, 2016; revised on February 26, 2016; accepted on March 14, 2016

Abstract

Motivation: Identifying the critical state or pre-transition state just before the occurrence of a phase transition is a challenging task, because the state of the system may show little apparent change before this critical transition during the gradual parameter variations. Such dynamics of phase transition is generally composed of three stages, i.e. before-transition state, pre-transition state and after-transition state, which can be considered as three different Markov processes.

Results: By exploring the rich dynamical information provided by high-throughput data, we present a novel computational method, i.e. hidden Markov model (HMM) based approach, to detect the switching point of the two Markov processes from the before-transition state (a stationary Markov process) to the pre-transition state (a time-varying Markov process), thereby identifying the pre-transition state or early-warning signals of the phase transition. To validate the effectiveness, we apply this method to detect the signals of the imminent phase transitions of complex systems based on the simulated datasets, and further identify the pre-transition states as well as their critical modules for three real datasets, i.e. the acute lung injury triggered by phosgene inhalation, MCF-7 human breast cancer caused by heregulin and HCV-induced dysplasia and hepatocellular carcinoma. Both functional and pathway enrichment analyses validate the computational results.

Availability and implementation: The source code and some supporting files are available at https://github.com/rabbitpei/HMM_based-method.

Contacts: lnchen@sibs.ac.cn or liyj@scut.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

There exist abrupt state changes in many biological processes, in particular, during the progression of complex diseases. For instance, in chronic diseases such as cancer, the malignant deterioration might take place merely within a few months, while the illness might progress gradually and be protracted for years before the onset of such transitions (He *et al.*, 2012; Litt *et al.* 2001; Liu *et al.*, 2001; McSharry *et al.*, 2003; Paek *et al.*, 2005; Roberto *et al.*, 2003; Scheffer *et al.*, 2009; Venegas *et al.*, 2005). In other words, there exists a sudden catastrophic shift during the process of gradual health deterioration that results in a drastic transition from a normal state to a disease state. However, it is a difficult task to detect the critical state just before the phase transition of the system from the observed data due to the lack of apparent state change before the

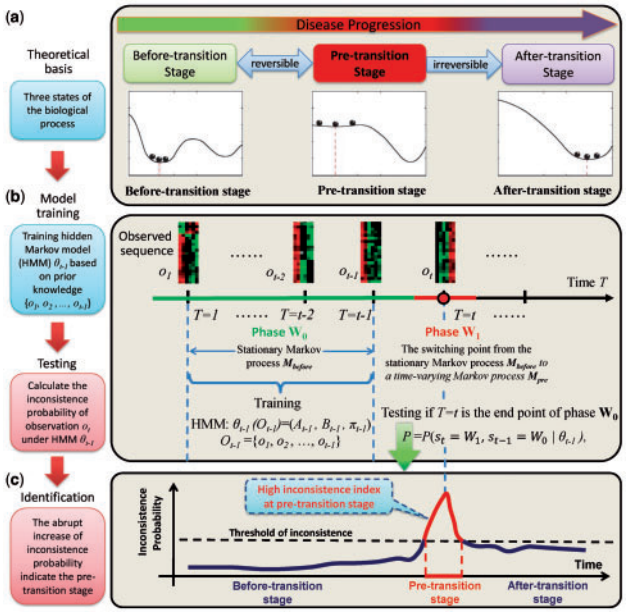

Fig. 1. Outline for identifying the pre-transition state using Markov model. (**a**) A complex biological process can be divided into three states, i.e. the before-transition state, the pre-transition state and the after-transition state. Both the before-transition and after-transition states are relatively stable or stationary with high resilience or low fluctuations, insensitive to the parameter changes. In the two states, external perturbations will not easily drive the system towards an alternative state. The pre-transition state is a time-varying state with low resilience or strong fluctuations, sensitive to the parameter changes, but may be reversible to the before-transition state in contrast to irreversible after-transition state. (**b**) First, based on the observed sequence of samples ($O_{t-1}=\{o_1, o_2, \ldots, o_{t-1}\}$), we train the hidden Markov model (HMM) $\theta_{t-1}(O_{t-1}) = (A_{t-1}, B_{t-1}, \pi_{t-1})$, that is, in view of stable dynamics, the before-transition state is modelled as the stationary Markov process $\theta_{t-1}(O_{t-1})$, while on the other hand, the pre-transition state is defined as a time-varying Markov process, due to the dynamical characteristics that the pre-transition state has strong fluctuations and is also sensitive to parameter changes, i.e. it changes with the time. Thus, based on such dynamical difference of the two states, to detect the pre-transition state during the biological process is equivalent to identify the switching point of the two Markov processes or the end point of the stationary Markov process. Second, under the assumption that time point $T=t$ with observation o_t is the switching point (or end point) of a stationary Markov process described by the trained HMM $\theta_{t-1}(O_{t-1})$, we calculate the probability$P = P(s_t = W_1, s_{t-1} = W_0, \ldots, s_1 = W_0|\theta_{t-1})$, where $s_t = State(o_t)$, and W_0 represents the before-transition state of the system, and W_1 stands for the state that is not consistent to W_0, i.e. the pre-transition state. W_2 is the after-transition state, which may be another stationary Markov process but is not the main focus in this work. (**c**) The large value of P indicates that a candidate point $T=t$ is the switching point of $\theta_{t-1}(O_{t-1})$ with high probability

transition. As shown in Figure 1a, generally disease progression can be modeled into three states or stages: (i) a normal state (or the before-transition state) with the high resilience and robustness to perturbations; (ii) a pre-disease state (or the pre-transition state) with the low resilience and sensitive to perturbations, which is the critical state just before the phase transition (Achiron *et al.*, 2010; Chen *et al.*, 2012); and (iii) a disease state (or the after-transition state), representing a seriously deteriorated stage possibly with high resilience and robustness (Liu *et al.*, 2012). It has been shown that even though there are no significant differences between the before-transition state and the pre-transition state in terms of static features (e.g. average values of state variables), there are significant differences between them in terms of dynamic features (e.g. deviations and correlations of state variables) as described in (Liu *et al.*, 2014a). Specifically, a dominant group or dynamical network biomarkers (DNBs) appears among the observed variables when the system state approaches the pre-transition state, satisfying the following three conditions, i.e. high correlations between the variables among this group, low correlations between this group and other variables and high standard deviations of the variables among this group (Chen *et al.*, 2012; Liu *et al.*, 2012, 2013a, b, 2014b, 2015; Li *et al.*, 2013; Tan *et al.*, 2015; Zeng *et al.*, 2014). These theoretical results indicate that the pre-transition state and the before-transition state are dynamically different, but are statically similar. Thus, the dynamics of these two states can be considered as two different Markov processes, which can be explored to detect the critical state or the pre-transition state from the observed data. Actually, the whole phase transition dynamics can be represented as three different Markov processes, i.e. one stationary Markov process for the before-transition state due to its relatively stable dynamics insensitive to parameter changes (stationary feature), one time-varying Markov process for the pre-transition state due to its strong fluctuated dynamics sensitive to the parameter changes (time-varying feature), and another stationary Markov process for the after-transition state due to their relatively stable dynamics insensitive to parameter changes (stationary feature) (Fig. 1b).

In this work, by exploiting the different dynamical features between the before-transition and pre-transition states, we developed a novel computational method based on hidden Markov model (HMM) for identifying the pre-transition state before the critical point is reached during the biological process of complex diseases. Specifically, to identify the pre-transition state is then equivalent to detect the switching point from one stationary Markov process to one time-varying Markov process (Fig. 1b). Utilizing the time-course or stage-course data, we presented the computational method and algorithm on estimating the inconsistence index of the switching or end point of the stationary Markov process at each candidate sampling point. Our study indicates that such novel index, which measures the inconsistence between observed samples and a stationary Markov process described by a trained HMM, is a model-free approach that can be theoretically applied to diseases or biological systems with clear transition events.

To demonstrate the effectiveness of our method, we applied the algorithm to a simulated regulatory network, and three real datasets, i.e. the microarray dataset of acute lung injury with carbonyl chloride inhalation exposure (GSE2565), the microarray dataset of MCF-7 human breast cancer caused by heregulin (HRG) (GSE13009) and the microarray dataset of HCV-induced dysplasia and hepatocellular carcinoma (HCC) (GSE6764). The pre-transition states were successfully identified for both numerical simulated dataset and real datasets, and the functional and pathway analyses also validated our theoretical detections on the early-warning signals of the imminent critical transitions for those diseases.

2 Methods

We first describe the theoretical basis, i.e. the generic properties of a complex system in the vicinity of the critical point, and then provide the procedures used to preprocess input datasets and the detailed algorithm.

2.1 Theoretical basis

Disease progression or its biological process can be generally divided into the following three states or stages, i.e. (i) the before-transition state (or normal state), (ii) the pre-transition state (or pre-disease

state) and (iii) the after-transition state (or disease state) (Fig. 1a). The before-transition state is a stable state with high resilience presenting a relatively 'healthy' stage, during which the state may change gradually and thus is considered as a stationary Markov process. The pre-transition state is a state defined as the limit of the before-transition state just before the critical phase transition. It is sensitive to the parameter changes, reversible to the before-transition state, and thus is considered as a time-varying Markov process. On the other hand, due to low resilience, even a small perturbation may suffice to trigger a drastic state change to the after-transition state, which is another stable state with high resilience and thus is also considered as a stationary Markov process (Fig. 1 and Supplementary Fig. S2). For a complex disease, the after-transition state represents a seriously ill stage, in which a system is difficult to return to the before-transition state even with intensive interventions. Hence, it is crucial to detect the pre-transition state so as to prevent qualitative deterioration by taking timely intervention actions. Although elucidating the critical phase transition at the network level holds the key to understand the fundamental mechanism of disease development and progression, it is notably hard to reliably identify the pre-transition state because there may be little apparent difference between the before-transition and pre-transition states. This is also the reason why diagnosis based on traditional biomarkers may fail to indicate the pre-transition state.

In this work, by exploring the different dynamics of one stationary Markov process and one time-varying Markov process, We design a method with an inconsistence index (*I*-index) based on hidden Markov model (HMM) to signal the impending critical phase transition during a complex biological process. Therefore, detecting the imminent critical phase transition is considered as identifying the switching point or period of the two Markov processes, or equivalently, identifying the period with the drastic increase of the *I*-index resulting from the abrupt change of the state-transition probability (Fig. 1) (Liu *et al.*, 2015). The detail theoretical derivation is given in Supplementary Information B.

2.2 Different dynamic features before and near critical phase transition

In this section, we present the theoretical derivation of our computational method. The dynamics for the progression of a complex disease can be expressed by the following nonlinear discrete-time dynamical system.

$$Z(t) = f(Z(t-1); P), \tag{1}$$

where $Z(t)=(z_1(t),\ldots,z_n(t))$ is an n-dimensional state vector or variable at time instant t with $t=1,2,\ldots$ and $P=(p_1, \ldots, p_s)$ is a parameter vector or driving factors representing slowly changing factors, e.g. genetic factors (SNP, CNV, etc.), epigenetic factors (methylation, acetylation, etc.) or environment factors. $f: \mathbb{R}^n \times \mathbb{R}^s \times \mathbb{R}^n$ is a nonlinear function. For such a nonlinear system, the system will undergo a phase change at $\bar{Z}$ or a bifurcation from a stable equilibrium when the parameter P reaches the threshold P_c (Gilmore, 1993). Detailed descriptions are presented in Supplementary Information A1.

For system (1) near $\bar{Z}$, before P reaches P_c, the system is supposed to stay at a stable equilibrium $\bar{Z}$ and therefore all the eigenvalues are within (0, 1) in modulus. The parameter value P_c at which the state shift of the system occurs is called a bifurcation parameter value or a critical value, and the state just before such a bifurcation is called pre-transition state. Generally, a real system is constantly perturbed by noise, and thus has stochastic dynamics. The following dynamic and statistic features have been proven when the system approaches the pre-transition state from the before-transition state, i.e. a dominant group or dynamical network biomarkers (DNBs) appears among the observed variables when the system state approaches the pre-transition state, satisfying the following three conditions (Chen *et al.*, 2012; Liu *et al.*, 2012, 2013a, 2014b).

- Correlations between the variables $z_i(t)$ among this group increase;
- Correlations between variables $z_i(t)$ of this group and other variables $z_j(t)$ decrease;
- Standard deviations of the variables $z_i(t)$ among this group increase.

Therefore, the dynamics between the before-transition state and the pre-transition state are significantly different. The before-transition state is a stable state with high resilience, insensitive to parameter perturbations, and thus can be modeled as a stationary Markov process. There is no significant change between the distributions of $Z(t)$ and $Z(t-1)$ when the system is in a before-transition state, i.e. the probability distribution almost keeps constant with the time evolution. In contrast, the pre-transition state is sensitive to the parameter changes with low resilience, and its dynamics or probability distribution changes with the time evolution. Thus, the pre-transition state is modeled as a time-varying Markov process. The distribution of $Z(t)$ is significantly distinct to that of $Z(t-1)$ when the system is in a pre-transition state. Based on such dynamic features, we can identify the switching period from the before-transition state to the pre-transition state. The derivation of the Markov process and the corresponding algorithm is presented in the following Section 2.3.2.

2.3 Identifying the switching period from stationary Markov process to time-varying Markov process based on HMM

We regard that the progression of a biological system in the before-transition stage is a stationary Markov process. To detect the onset of the pre-transition stage is equivalent to identify the changing period from this stationary Markov process to another time-varying Markov process.

2.3.1 Inconsistence index (I-index)

Specifically, we propose an inconsistence index (*I*-index) to measure the probability of a candidate time point as the changing or switching point from the stationary Markov process to the time-varying Markov process. On the basis of an observed sequence $O_{t-1}=\{o_1, o_2, \ldots, o_{t-1}\}$, i.e. the preceding $t-1$ sets of samples from time points $1, 2,\ldots, t-1$, we train and obtain a hidden Markov model (HMM) $\theta_{t-1}=(A, B, \pi)$, where the subscript $t-1$ of θ represents that the HMM is derived from the training samples up to time point $t-1$, A is a state transition matrix, B is an emission matrix, and π is a probability vector for the initial state. The training process based on an unsupervised learning procedure, i.e. Baum–Welch algorithm, is provided in Supplementary Information A.3.

Then, *I*-index, or HMM-based probability P_t measuring the inconsistence between the system state s_t at time t and the preceding system states under the HMM θ_{t-1}, is given as follows:

$$I(t) = P_t(s_t = W_1 | s_{t-1} = W_0, s_{t-2} = W_0, \ldots, s_1 = W_0, \theta_{t-1}, O), \tag{2}$$

where θ_{t-1} is the trained HMM, $O_t=\{o_1, o_2, \ldots, o_{t-1}, o_t\}$ is the observed sequence up to time t, $\{s_1, s_2, \ldots, s_{t-1}, s_t\}$ is the state sequence with subscript $i \times \{1, 2,\ldots, t-1, t\}$ stands for the state of the

system at the ith time point, W_0 and W_1 are unobserved (hidden) states while W_0 represents the before-transition state of the system, and W_1 stands for the state that is not consistent to W_0, i.e. the pre-transition state. W_2 is the after-transition state, which is not studied in this work. Obviously, the inconsistence index I satisfies that

$$I(t) = 1 - Q_t(s_t = W_0 | s_{t-1} = W_0, s_{t-2} = W_0, \ldots, s_1 = W_0, \theta_{t-1}, O), \quad (3)$$

where Q_t actually represents the probability of consistence between the system state with observation o_t and the preceding system states under the HMM $\theta_{t\text{-}1}$. According to the Markov chain, it follows

$$\begin{aligned} &Q_t(s_t = W_0 | s_{t-1} = W_0, s_{t-2} = W_0, \ldots, s_1 = W_0, \theta_{t-1}, O) \\ &= Q_t(s_t = W_0 | s_{t-1} = W_0, \theta_{t-1}, O) \\ &= \frac{P(s_{t-1} = W_0, s_t = W_0 | \theta_{t-1}, O)}{P(s_{t-1} = W_0 | \theta_{t-1}, O)} \end{aligned}$$

The calculation of the probability Q_t with $\theta_{t\text{-}1}$ is based on the forward algorithm, shown in Supplementary Information A.3.

Clearly, if the testing time point t is in the stationary Markov process which is described by HMM $\theta_{t\text{-}1}$, or observation o_t is derived in the before-transition stage, then the probability $I(t)$ has no significant change comparing with $I(t$ - 1) (see Fig. 1, or Supplementary Fig. S2), that is, I-index remains stationary when the system is in the before-transition stage. However, if the testing time point t is in the time-varying Markov process, or the observation o_t is derived in the pre-transition stage, then $I(t)$ increases drastically, indicating the high inconsistence between observation o_t and the HMM $\theta_{t\text{-}1}$ based on the preceding t - 1 sample sequences $O_{t\text{-}1} = \{o_1, o_2, \ldots, o_{t\text{-}1}\}$. Clearly the inconsistent data appears at the switching period between the stationary Markov process in the before-transition stage and the time-varying Markov process in the pre-transition stage. Therefore, the abrupt increase of the inconsistence index $I(t)$ signals the impending of the critical transition, or the occurrence of the pre-transition stage.

2.3.2 Algorithm

There are two steps to detect the switching period in our algorithm, i.e.

- Step-1: In the training step, obtain the HMM $\theta_{t\text{-}1} = (A, B, \pi)$ based on the preceding t-1 observed data (time sequence data) $O_{t\text{-}1} = \{o_1, o_2, \ldots, o_{t\text{-}1}\}$. Here $A = (a_{ij}(t\text{-}1))_{2\times 2}$ with

$$a_{ij}(t-1) = P(s_{t-1} = W_i | s_{t-2} = W_j), \quad i, j \in \{0, 1\} \quad (4)$$

  $B = (b_{jk}(t\text{ - }1))_{2\times(n+1)}$ with $b_{jk}(t$ - 1) representing the probability of the kth possible observation under the assumption that the system state is S_j at time $t-1$, i.e.
  $b_{jk}(t-1) = P(\#1(t-1) = k | s_{t-1} = W_j), \quad j \in \{0,1\}, \quad k \in \{0,1,\ldots,n\}$. (5)
  where case #1(t - 1)=k reflects that there are k molecules differentially expressed in one observation.

- Step-2: In the testing step, calculate I-index of the testing sample o_t at time point t by Eq. (2) based on both the HMM $\theta_{t\text{-}1}$ and the additional observation o_t. If there is a drastic increase of I-index, then t is the switching point, at which system is in the pre-transition stage. Otherwise go to Step-1 for next time point.

The detailed algorithm including the training of HMM and the testing of candidate time point for the I-index is also shown in Supplementary Information A.3.

2.4 Data processing

We applied our method to real datasets, i.e. the microarray data of acute lung injury of mice induced by carbonyl chloride inhalation exposure (GSE2565), the microarray data of MCF-7 human breast cancer caused by heregulin (HRG) (GSE13009) and the microarray data of HCV-induced dysplasia and hepatocellular carcinoma (HCC) (GSE6764). These microarray datasets were downloaded from the NCBI GEO database (www.ncbi.nlm.nih.gov/geo).

In these datasets, redundant probe sets without the corresponding gene symbols were discarded. The expression values of different probe sets mapped to the same gene symbol were averaged. For each species, we downloaded the biomolecular interaction networks from various databases, including STRING (http://string-db.org/), TRED (www.rulai.cshl.edu/cgi-bin/TRED/), KEGG (www.genome.jp/kegg) and HPRD (www.hprd.org), i.e. the available functional linkage information for *Mus musculus* and *Homo sapiens* was downloaded from these databases and combined. After removing any redundancy, we obtained 65 625 linkages in 11 451 human proteins/genes and 37 950 linkages in 6683 mouse proteins/genes. Next, the genes evaluated in these microarray datasets were mapped individually to these integrated functional linkage networks. The network information is employed in the post-processing step for visualizing results and functional analysis. For each disease dataset, the expression profiling information was mapped to the integrated networks individually. Specifically, the gene expression profiling dataset of acute lung injury was obtained from an experiment of a toxic-gas-induced lung injury such as pulmonary edema, in which genomic approach was applied to investigate the molecular mechanism of phosgene-induced lung injury. The dataset for breast cancer was obtained in an experiment on MCF-7 cell line with HRG stimulation. Besides, gene expression profiles of 75 tissue samples were analyzed representing the stepwise carcinogenic process from pre-neoplastic lesions (cirrhosis and dysplasia) to HCC, including four neoplastic stages (very early HCC to metastatic tumors).

Finally, the networks with selected top differential-expression gene sets were visualized using Cytoscape (www.cytoscape.org) and part of the functional analysis was based on the website tool DAVID Bioinformatics Resource (Huang *et al.*, 2008). The detailed data description and processing procedures are provided in Supplementary Information C.

3 Results

3.1 Identifying the pre-transition state for a ten-node network

To demonstrate the effectiveness of the computational method and the inconsistence index, we used a theoretical model of a ten-node gene regulatory network (Fig. 2a) to generate data and show the detection of early-warning signals near a critical point. These types of gene regulatory networks are often used to study transcription, translation, diffusion and translocation processes that affect gene regulatory activities (Chen *et al.*, 2009). The specific set of stochastic differential equations representing the gene regulation among ten genes in the network (Fig. 2a) is given in Supplementary Information B, from which the simulated dataset is generated. The numerical simulation shows that a drastic boost of the inconsistence

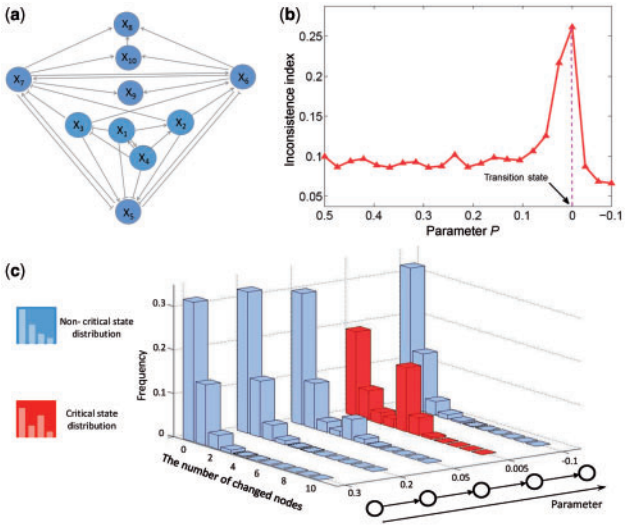

Fig. 2. The validation of HMM-based method on a simulation dataset. To validate the sensitivity and effectiveness, we calculated the inconsistence index using the simulated dataset from a ten-node network. (**a**) Theten-node network, in which the arrow represents positive regulation, whereas the blunt line denotes negative regulation. (**b**) The inconsistence index of the network. It can be seen that an abrupt increase of the probability signals the imminent critical transition at $P=0$. The simulations were performed in MATLAB(R2013a) using the Euler-Maruyama integration method with the Ito calculus. The system undergoes a bifurcation at $P=0$. (**c**) We illustrate that the distribution of the frequency of nodes with large changes, i.e. the ratio of state-transition nodes. It can be seen that when the parameter P is far away from the critical value $P=0$ ($P=0.3$, $P=0.2$), there are few nodes with large-scale changes. However, while the parameter P approaches the critical value $P=0$ ($P=0.005$), the distribution is quite different, i.e. the ratio of 4-changing-nodes increases considerably

index indicates the upcoming critical transition at parameter $P = 0$ (Fig. 2b).

The numerical experiment validates that the inconsistence score is reliable and accurate in identifying the pre-transition state and thus provides the early-warning signal of a catastrophic change in the system. Besides, to demonstrate the different dynamics of the system between the before-transition state and the pre-transition state, we illustrate the underlying frequency of nodes with large changes, i.e. the ratio of state-transition nodes (Fig. 2c). The calculation details are presented in Supplementary Information B. Note that there is no clear signal to detect the imminent transition from a single variable (or a few variables) due to the noise (or stochastic fluctuations) of the original biological system, which demonstrates the advantage of exploiting high-dimensional information using our scheme. In other words, if there is no detailed model for a biological system, generally we do not know which variable can reflect the critical change of the system so as to measure it. As shown in this example, given high-throughput data or high-dimensional information, the inconsistence score provides a way to detect the signal for diagnosing the pre-transition state even without a detailed model.

3.2 Predicting critical transitions in real datasets

We applied the HMM-based method in three real experimental datasets, i.e. the microarray data for acute lung injury induced by phosgene gas inhalation (GSE2565), MCF-7 human breast cancer caused by heregulin (HRG) (GSE13009) and HCV-induced dysplasia and hepatocellular carcinoma (HCC) (GSE6764). Here we present the application on the dataset of lung injury as an example. The descriptions of the results on the other two datasets are given in Supplementary Information C.

The dataset GSE2565 was obtained in an experiment on toxic gas-induced lung injury effects, i.e. pulmonary edema (Sciuto *et al.*, 2005). A genomic approach was used to investigate the molecular mechanism of phosgene-induced lung injury. The experiments determined the temporal effects of phosgene exposure on lung tissue antioxidant enzyme concentrations and gene expression levels, and these results were compared with those from air-exposed mice treated in a similar manner to assess the role of the glutathione redox cycle in this oxidative lung injury model. To produce two groups of data, i.e. the control group data and case group data, two groups of CD-1 male mice were exposed to air or phosgene, respectively. Lung tissues were collected from air- or phosgene-exposed mice at eight sampling time point, i.e. 0.5, 1, 4, 8, 12, 24, 48 and 72 h after exposure. The details of the experiment are available in the original paper (Sciuto *et al.*, 2005).

According to the HMM-based method, we regard that each time point is a candidate transition point, i.e. the end point of a Markov process in before-transition state. At each sampling point, the top 500 differential-expression genes are identified on the basis of P-value from the Student t-test. The inconsistence index is then calculated based on these selected genes (Fig. 3a). Clearly, there are eight probability curves respectively corresponding to eight groups of differential-expression genes. Each group of genes is differentially expressed at one time point. Among the probability curves in Figure 3a, the red one presents the inconsistence index calculated based on the top 500 differential genes with the most significant P-values at the 8-h time point. It can be seen that the red curve shows the largest probability at 8 h and thus indicates the imminent critical transition. The selection of 500 genes is from computational consideration. The gene filtering, detailed algorithm and computation progress are shown in Supplementary Information C1. To further elucidate the relation between top differential-expression genes and dysfunctional pathways, we also carried out the clustering analysis based on correlation at the pre-transition stage (4 and 8 h), that is, based on the top 500 differential genes, we selected the most significant gene group (with P-value 1.37E-04 and over 3-fold change comparing with the control group at 8 h, see Figs 3c) for further functional analysis shown in Supplementary Information C1. To illustrate the significance of the results, a comparative figure with control samples and bootstrap is presented in Supplementary Figure S6. Besides, it should be noted that the number of the selected genes is not so sensitive to the inconsistence index (see Supplementary Information C1 for the details).

Besides, to study the top differential-expression genes and their upstream transcriptional factors, we identified the top turnover module or subnetwork, that is, by mapping the top differential-expression 200 genes and their upstream transcriptional factors to the gene regulatory network from STRING, we illustrate a module with over 55% turnover genes in Figure 4a. Each turnover gene is highly (or lowly) expressed before the critical transition and has lowly (or highly) expression after.

Figure 4b presents the dynamical evolution in the whole gene network based on the case data of acute lung injury. It can be seen that those top differential-expression genes (the top right corner in each network) are strongly correlated with wild fluctuation 8 h after the exposure to phosgene gas, which provides a significant signal indicating the pre-transition stage just before the deterioration into differentiation, while other genes show no significant signal. Clearly, when the deterioration is impending, these selected genes

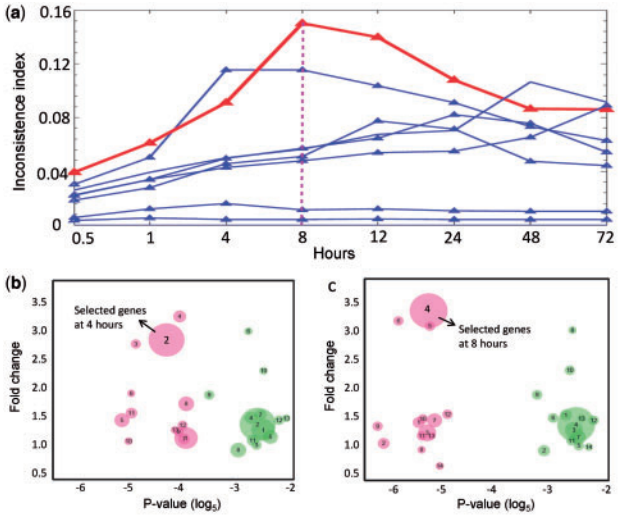

Fig. 3. The application of HMM-based method to the dataset of acute lung injury induced by phosgene inhalation. (**a**) The inconsistence indices based on the top 500 differential-expression genes from each candidate transition time points. The red curve represents the inconsistence index calculated from the top 500 differential-expression genes which are selected at 8 h, while the seven blue curves are those from other time points. It can be seen that the most significant signal appears at 8 h the exposure to phosgene gas, which is in coincidence with the experimental observation, i.e. the main physiological effects occurred within the first 8 h after exposure. (**b**) We illustrate the clustering result respectively for genes at 4 h (red bubbles) and 1 h (green ones). (**c**) We illustrate the clustering result respectively for genes at 8 h (red bubbles) and 1 h (green ones). Clearly, comparing with the control group, the *P*-value of each gene group is more significant at 8 h than that at 1 h. The top differential-expression genes were selected to proceed with functional analysis (Color version of this figure is available at *Bioinformatics* online.)

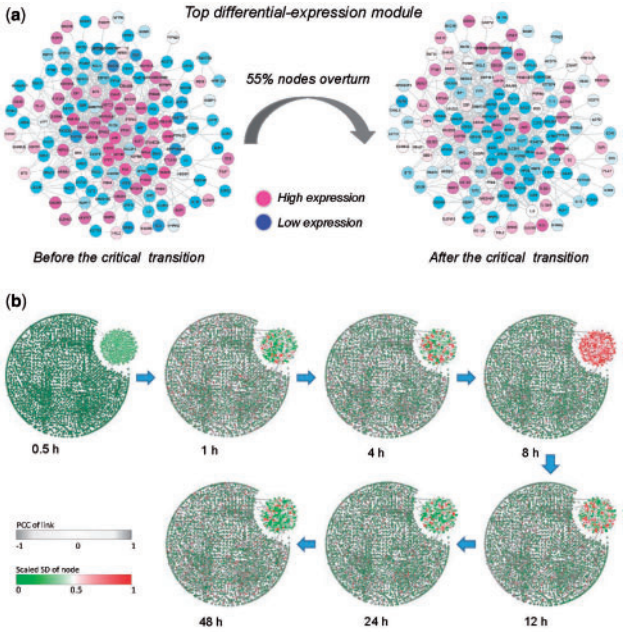

Fig. 4. Dynamical changes in the network during the progression of phosgene-induced acute lung injury. To validate the results from HMMbased method, we show the dynamical evolution of the network structure. (**a**) A subnetwork with top 200 differential-expression genes is identified from the 4th clustering group in Figure 3c. In this network, over 55% genes have reversal (or turnover) expressions when the system progresses from the before-transition state to the after-transition state, comparing with only 18% turnover ratio for all genes. (**b**) The figures show the dynamical changes of the molecular network at 0.5, 1, 4, 8, 12, 24 and 72 h, respectively. Each network was constructed from the overall mapped mouse molecular interaction network based on the expression data. The color of nodes represents the fluctuation of expression, and the thickness of links stands for the correlation between each pair of nodes. The group of the 200 genes of most significant differential expressions is located in the top right corner in each network. It can be seen that the selected genes show wild fluctuation in terms of their expressions during 4–8 h. These critical phenomena do not appear before or after this period, i.e. the before- or aftertransition state. Thus, the pre-disease stage is around 4–8 h, where the network structure exhibits the most significant change, just before the critical transition triggered by phosgene inhalation (Color version of this figure is available at *Bioinformatics* online.)

form a special subnetwork. It can be seen that, oppositely, neither the whole gene network nor the selected differential-expression genes present a signal before or after the transition, which shows the sensitivity of the HMM-based method at the pre-transition state. In fact, the inconsistence index reveals the existence of the pre-transition state, which, however, cannot be shown by any single bio-molecule. Therefore, the benefits brought by the HMM-based method in signaling the pre-transition state make the identification and management of high-risk cases more effective.

Briefly, those studies suggested that the main physiological effects occurred within the first 8 h after exposure, resulting in common observations of enhanced bronchioalveolar lavage fluid (BALF) protein levels, increased pulmonary edema and ultimately decreased survival rates (Sciuto *et al.*, 2005). At the concentration delivered, 50–60% mortality was routinely observed after 12 h while 60–70% mortality was observed after 24 h (Sciuto *et al.*, 2005). Early-warning signals of lung injury based on our method are shown in Figures 3a and 4b, which showed that the pre-transition state may start around 8 h, whereas the system may enter the after-transition state (or disease state) after 12 h. Our prediction based on the inconsistence index agreed with the actual development of the disease.

Figure 5 presents the second application of HMM-based method for HRG-caused breast cancer. Based on Figure 5a, each probability curve is calculated based on a group of differential-expressed genes at the corresponding sampling time point. The red curve in Figure 5a shows the most significant increase of inconsistence index at the 90-min time point, which indicates that the most possible critical transition point is around 90-min time point. Figure 5b shows the dynamical evolution in the whole feature network based on the case data of breast cancer. It can be seen that the network structure also changes significantly at the 90-min time point. In fact, the original assay showed that the AP-1 complex in HRG-treated MCF-7 cells contains c-JUN, c-FOS and FRA-1, although the association of c-JUN in the complex is transient (Saeki *et al.*, 2009). Besides, the stimulation of MCF-7 breast cancer cells with EGF and HRG resulted in very similar early transcription profiles up to 90 min; however, subsequent cellular phenotypes differed after 3 h (Saeki *et al.*, 2009), which suggests that the differentiation is around 3 h (the 9th sampling time point). Therefore, our application results are in coincidence with the experimental observation and successfully detect the early-warning signal of the impending critical transition.

The results of the third application for the hepatocellular carcinoma (HCC) (GSE6764) are illustrated in Supplementary Information C.

3.3 Functional analysis

Phosgene gas is one of the most important and common chemical industry gases (Schneider and Diller, 2000). Some pathogenic mechanisms of the acute lung injury induced by phosgene have been identified (Sciuto *et al.*, 2005). According to the results above, a

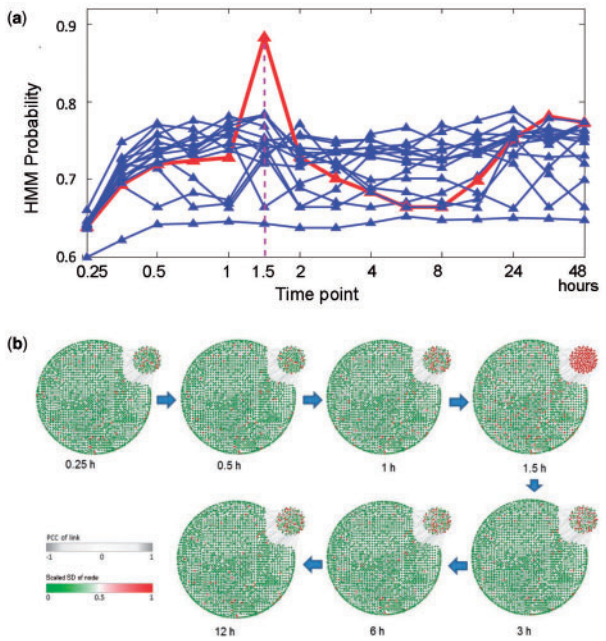

Fig. 5. The application of HMM-based method on the dataset of HRG-induced breast cancer. (**a**) By applying HMM-based method to the microarray data of HRG-induced breast cancer, we show the inconsistence indices based on the top 500 differential-expression genes from each candidate transition time points. The red curve represents the inconsistence index calculated from the top 500 differential-expression genes which are selected at the time point 1.5 h, while the blue curves are those from other time points. It can be seen that the most significant signal appear at 1.5 h, which agrees with the experimental observation, i.e. the main physiological effects occurred within the first 1.5 h after exposure. (**b**) We illustrate the dynamical changes in the network during the progression of breast cancer. The 200 top expressed genes are placed in the top right corner (Color version of this figure is available at *Bioinformatics* online.)

major change in the inconsistence index of the top differential genes occurs from 4 to 8 h. Pathway enrichment analysis and GO functional analysis showed that the genes in the top significant group (the 4th red group in Fig. 3c) were closely related to the mechanism of disease progression (Sciuto *et al.*, 2005; Wang *et al.*, 2013). Dysfunction in glutathione metabolism and the chemokine signaling pathway related to the inflammatory immune response was caused in vivo, which also reflected protection against the oxidant-like activity of phosgene. Pathways affected by the oxidant reaction became disordered, especially signal transduction via protein-modified activation, such as the mitogen-activated protein kinase (MAPK) signaling pathway and the Wnt signaling pathway. The decrease in PH induced by the HCl-release reaction affected some pathways that are sensitive to intracellular conditions and related to communication or transport channels, e.g. gap junctions. Some signaling pathways may also be relevant to repair, survival, apoptosis and reproduction, such as the gonadotropin-releasing hormone (GnRH) signaling pathway, MAPK signaling pathway and TGF-beta signaling pathway (Sciuto *et al.*, 2005; Wang *et al.*, 2013). At the gene ontology (GO) function level, some biological processes were also highly related to acute lung injury. For example, the expression profiles of some genes were related to abnormal changes in primary metabolic processes. This indicates the denaturation of lipids, proteins and nucleic acids that may have been oxidized by phosgene (Sciuto *et al.*, 2005; Wang *et al.*, 2013).

The functional analysis for HRG-induced breast cancer and HCC liver cancer are shown in the Supplementary Information C.

4 Discussion

Complex diseases significantly damage the health of many people all over the world. Detecting the early-warning signal of the sudden deterioration provides an opportunity to interrupt and prevent the continuing costly cycle of managing these diseases and their complications. Although it is crucial to detect the pre-transition state so as to prevent the qualitative deterioration by taking appropriate intervention actions, it is a challenging task to reliably identify the pre-transition state because the state of the system may show neither apparent change nor clear phenomenon before this critical transition during the disease progression. This is also the reason why diagnosis based on traditional biomarkers may fail to indicate a pre-transition state.

In this work, we presented a computational method with an inconsistence index based on HMM to identify the imminent critical transition, which has been shown to be effective by real datasets. It is worth noting that this method aims to detect the early-warning signal generating from the pre-transition state (or pre-disease state), rather than to find the indication of the after-transition state (or disease state) in which the qualitative deterioration has taken place. As shown in Figure 1 and in METHODS, generally there are significant differences between the before-transition state and the after-transition state, which is why we can find many molecular biomarkers to accurately diagnose the disease based on the differential expressions of those biomarkers. But there may be no significant differences between the before-transition state and the pre-transition state in terms of expressions, which requires the different types of biomarkers based on different signals. DNB theory provides such conditions of the pre-disease state, and we developed HMM-based method to quantitatively detect the signals of the pre-disease state based on those conditions. We applied our method to the identification of the pre-transition state based on both the simulated dataset and the microarray data from acute lung injury experiment. Given that the lung damage is not generally detectable until approximately 8 h after exposure to phosgene gas, we tested the hypothesis that the disease progression can be modeled into three states: (i) a before-transition state with high resilience and robustness to perturbations; (ii) a pre-transition state, defined as the prelude to catastrophic shift into the transition state, occurring just before the phase transition point is reached, therefore, with low resilience and robustness to perturbation due to its dynamical structure; (iii) an after-transition state, representing the seriously deteriorated stage possibly with high resilience and robustness to perturbation. Then, based on the microarray data, we indicated the existence of the pre-transition stage right before the critical transition induced by phosgene inhalation. Actually, an indicative early-warning signal is presented at 8 h after the expose to phosgene gas. The validation based on simulated dataset (Fig. 2) and the microarray data of acute lung injury (Figs 3 and 4) demonstrates the sensitivity and effectiveness of our method. With the genomics or proteomics survey of the CD-1 model rats, we constructed bio-molecular networks (Fig. 4) to gauge the dynamical regulations among genes at different hours after exposure to phosgene. Besides, we showed that some metabolic pathways respond to the phosgene-induced interruptions and become increasingly disordered. Therefore, the HMM-based method provides a computational way of prying into the underlying mechanism of cell differentiation and thus helping to achieve the timely intervention. Our dynamic network analysis suggests, in regard to lung injury, by focusing at the pre-transition state of the rat model, we are able to probe the in situ environment changes preceding the development of lung tissue abnormalities. This may not only lead to deep insights of

external environment interactions, but also identify the effective time window for novel therapeutic strategies in phosgene-triggered lung injury.

There are limitations of this work. First, the validity of the identified pre-transition state and the accurate result need further supports from animal experiments or clinical studies. Second, the method is insensitive when the genes are not differentially expressed. Although this work is merely a step forward towards detecting the early-warning signals of a critical transition during disease progression and the algorithm is expected to be improved in both sensitive and accurate ways, it opens a new way for identifying the early-warning signals of a critical transition during the progression of complex diseases.

Funding

The work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (No. XDB13040700); National Natural Science Foundation of China (Grant numbers 91530320, 91529303, 91439103, 61134013, 61370228, and 11401222); Pearl River Science and Technology Nova Program of Guangzhou.

Conflict of Interest: none declared.

References

Achiron,A. *et al*. (2010) Microarray analysis identifies altered regulation of nuclear receptor family members in the pre-disease state of multiple sclerosis. *Neurobiol. Dis*., **38**, 201–209.

Arnol'd,V.I. (1994) *Dynamical Systems V: Bifurcation Theory and Catastrophe Theory*. World Scientific, Springer, Berlin.

Chen,L. *et al*. (2009) *Biomolecular Networks: Methods and Applications in Systems Biology*. John Wiley & Sons, Hoboken, NJ.

Chen,L. *et al*. (2012) Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci. Rep*., **2**, 1–8.

Gilmore,R. (1993) *Catastrophe Theory for Scientists and Engineers*. Dover Publications, New York.

He,D. *et al*. (2012) Coexpression network analysis in chronic hepatitis B and C hepatic lesion reveals distinct patterns of disease progression to hepatocellular carcinoma. *J. Mol.Cell Biol*., **4**, 140–152.

Huang,D.W. *et al*. (2008) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat. Protoc*., **4**, 44–57.

Li,M. *et al*. (2013) Detecting tissue-specific early-warning signals for complex diseases based on dynamical network biomarkers: study of type-2 diabetes by cross-tissue analysis. *Brief. Bioinf*., **15**, 229–243.

Litt,B. *et al*. (2001) Epileptic seizures may begin hours in advance of clinical onset: a report of five patients. *Neuron*, **30**, 51–64.

Liu,J.K. *et al*. (2001) Pituitary apoplexy. *Semin. Neurosurg*., **12**, 315–320.

Liu,R. *et al*. (2012) Identifying critical transitions and their leading networks for complex diseases. *Sci. Rep*., **2**, 1–9.

Liu,R. *et al*. (2013a) Dynamical network biomarkers for identifying critical transitions and their driving networks of biologic processes. *Quant. Biol*., **1**, 105–114.

Liu,X.P. *et al*. (2013b) Detecting early-warning signals of type 1 diabetes and its leading biomolecular networks by dynamical network biomarkers. *BMC Med. Genomics*, **6**, S8.

Liu,R. *et al*. (2014a) Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. *Med. Res. Rev*., **34**, 455–478.

Liu,R. *et al*. (2014b) Identifying critical transitions of complex diseases based on a single sample. *Bioinformatics*, **11**, 1579–1586.

Liu,R. *et al*. (2015) Identifying early-warning signals of critical transitions with strong noise by dynamical network markers. *Sci. Rep*., **5**, 17501.

McSharry,P.E. *et al*. (2003) Prediction of epileptic seizures: are nonlinear methods relevant? *Nat. Med*., **9**, 241–242.

Paek,S. *et al*. (2005) Hearing preservation after gamma knife stereotactic radiosurgery of vestibular schwannoma. *Cancer*, **104**, 580–590.

Roberto,P.B. *et al*. (2003) Transition models for change-point estimation in logistic regression. *Stat. Med*., **22**, 1141–1162.

Saeki,Y. *et al*. (2009) Ligand-specific sequential regulation of transcription factors for differentiation of MCF-7 cells. *BMC Genomics*, **20**, 545–552.

Scheffer,M. *et al*. (2009) Early-warning signals for critical transitions. *Nature*, **461**, 53–59.

Sciuto,A.M. *et al*. (2005) Genomic analysis of murine pulmonary tissue following carbonyl chloride inhalation. *Chem. Res. Toxicol*., **18**, 1654–1660.

Schneider,W. and Diller,W. (2000) *Phosgene, in Ullmann's Encyclopedia of Industrial Chemistry*. Wiley-VCH, Weinheim, pp. 411–414.

Tan,Z. *et al*. (2015) Cerebrospinal fluid protein dynamic driver network: at the crossroads of brain tumorigenesis. *Methods*, **15**, 36–43

Venegas,J.G. *et al*. (2005) Self-organized patchiness in asthma as a prelude to catastrophic shifts. *Nature*, **434**, 777–782.

Wang,P. *et al*. (2013) Mechanism of acute lung injury due to phosgene exposition and its protection by cafeic acid phenethyl ester in the rat. *Exp. Toxicol. Pathol*., **65**, 311–318.

Zeng,T. *et al*. (2014) Deciphering early development of complex diseases by progressive module network. *Methods*, **3**, 334–343.