BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 11 2014, pages 1579–1586
doi:10.1093/bioinformatics/btu084

Systems biology

Advance Access publication February 10, 2014

Identifying critical transitions of complex diseases based on a single sample

Rui Liu[1], Xiangtian Yu[2,3], Xiaoping Liu[4], Dong Xu[5], Kazuyuki Aihara[4] and Luonan Chen[2,4,*]

[1]School of Mathematics, South China University of Technology, Guangzhou 510640, China, [2]Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for PreDiabetes, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China, [3]School of Mathematics, Shandong University, Jinan 250100, China, [4]Collaborative Research Center for Innovative Mathematical Modelling, Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan and [5]Department of Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Unlike traditional diagnosis of an existing disease state, detecting the pre-disease state just before the serious deterioration of a disease is a challenging task, because the state of the system may show little apparent change or symptoms before this critical transition during disease progression. By exploring the rich interaction information provided by high-throughput data, the dynamical network biomarker (DNB) can identify the pre-disease state, but this requires multiple samples to reach a correct diagnosis for one individual, thereby restricting its clinical application.

Results: In this article, we have developed a novel computational approach based on the DNB theory and differential distributions between the expressions of DNB and non-DNB molecules, which can detect the pre-disease state reliably even from a single sample taken from one individual, by compensating insufficient samples with existing datasets from population studies. Our approach has been validated by the successful identification of pre-disease samples from subjects or individuals before the emergence of disease symptoms for acute lung injury, influenza and breast cancer.

Contact: lnchen@sibs.ac.cn.

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

Received on October 25, 2013; revised on January 18, 2014; accepted on February 4, 2014

1 INTRODUCTION

Based on only one single sample from an individual, traditional biomarkers can indicate whether the sample is in a given disease state or not; however, they cannot generally diagnose the pre-disease state, which can be viewed as a special normal state just before major deterioration or critical transition to a disease (Fig. 1a). In other words, it is a highly challenging problem to distinguish the pre-disease state from the normal state because there may be little apparent difference between these two states, in contrast to the generally significant difference between the normal state and the disease state (or the early disease state). One way to detect early signals of an abrupt change in a system, is to directly construct a mathematical model for the system and predict the progression details (Hirata *et al.*, 2010; May *et al.*, 1977). However, in most realistic cases, such a model for a specific system or a disease is not available due to the complicated and uncertain nature of most such systems. The critical slowing-down theory (Strogatz *et al.*, 1994) provides a useful way to detect the early warning signals of critical transitions for a general system (Scheffer *et al.*, 2009), and has been applied to ecosystems (Carpenter *et al.*, 2005, 2006, 2011; Drake and Griffen, 2010; Scheffer *et al.*, 2001), climate systems (Dakos *et al.*, 2008; Lenton *et al.*, 2008; Held and Kleinen, 2004; Kleinen *et al.*, 2003), economics and global finance (Kambhu *et al.*, 2007; May *et al.*, 2008). However, there are two main limitations to this method: (i) it requires a series of time-course data for each individual, i.e. a large number of samples for each subject; (ii) the measurements are required to cover those variables that show the critical slowing-down dynamics, i.e. extensive knowledge on the system is required. Clearly, for many complex systems, in particular, biomedical systems, it is generally difficult to obtain data (or a model) of an individual satisfying both these conditions. In fact, typical data currently available for molecular biology and medicine are high-throughput OMICS data with a small number of samples (for each individual) but high dimensions. Although these data usually lack dynamic information due to the small number of time-course samples, they are still valuable because of the rich information regarding correlations or interactions among many variables of the high-throughput measurements. With rapid advancements in high-throughput technologies, OMICS data have been measured and made available for complex diseases such as asthma attacks (Venegas *et al.*, 2005), epileptic seizures (Litt *et al.*, 2001) and many others (He *et al.*, 2012; Liu *et al.*, 2001; McSharry *et al.*, 2003; Paek *et al.*, 2005; Roberto *et al.*, 2003). To overcome the two problems of the critical slowing-down method, a new model-free approach, the dynamical network biomarker (DNB), was developed to detect early warning signals of complex diseases even with a small number of samples (Chen *et al.*, 2012; Liu *et al.*, 2012). By exploring fluctuation and correlation information among molecules, the DNB can identify the pre-disease state (Fig. 1a) and thus predict the critical transition before a serious deterioration, in contrast to diagnosing the disease state by traditional

*To whom correspondence should be addressed.

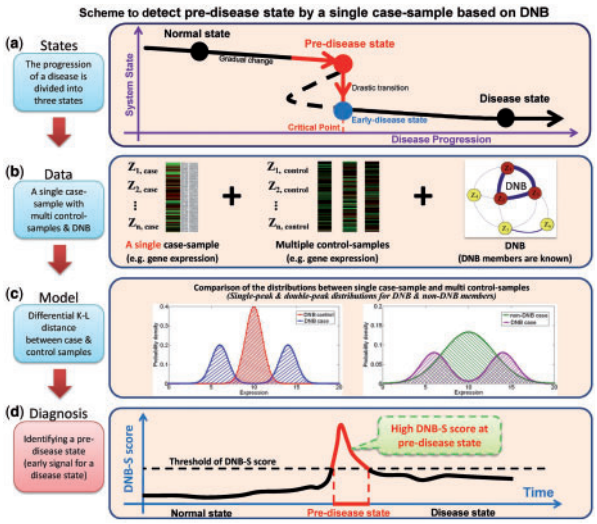

Fig. 1. Outline for identifying the pre-disease state based on a single sample by DNB. This outline shows how the DNB-S score identifies the pre-disease state based on a single sample. **(a)** Disease progression can be divided into three stages, i.e. normal, pre-disease and disease states, where the pre-disease state is a special 'normal' state just before the drastic transition (serious disease deterioration). There are generally significant differences between the normal and disease states, but there may not be apparent differences between the normal and pre-disease states. **(b)** Assume that there is a single case-sample and a number of control–samples (or normal samples). Furthermore, the DNB is assumed to be available, e.g. from previous studies. We aim to diagnose whether a single case sample is in the pre-disease state or not. **(c)** DNB molecules have double-peak distribution in the pre-disease state (the DNB case) but show single-peak distribution in the normal state (the DNB control), while non-DNB molecules show single-peak distribution in the pre-disease state (the non-DNB case). The horizontal axis represents gene expression levels in the log scale as an example. By exploiting such differential distributions in a single sample, i.e. the K–L distribution distance, a composite index (i.e. the DNB-S score) is constructed to diagnose single samples. **(d)** The DNB-S score is relatively high when the case sample is in the pre-disease state and relatively low when the case sample is in the normal state. Thus, we can identify the pre-disease state with a single sample using the DNB-S score, i.e. we can detect the early warning signal of the disease state

biomarkers. Moreover, it has been shown that those molecules in DNB are not necessarily the result of the disease but make the first move from the normal state toward the disease state in a collective manner through a critical transition (Liu *et al.*, 2012). One significant feature of DNB members (or molecules) is that in the pre-disease state, they fluctuate dynamically and are strongly correlated. Therefore, this group of molecules, or the DNB can characterize the dynamic features of the underlying system, and provide early warning signals for detecting the imminent critical transition (Liu *et al.*, 2013a, b, c). The DNB has been successfully applied to many complex diseases, such as liver cancer, to detect sudden deterioration and study the underlying mechanisms (Chen *et al.*, 2012; Liu *et al.*, 2012, 2013a, b, c; Li *et al.*, 2013).

Although the DNB can identify the pre-disease state, it still requires multiple samples for an individual, thereby restricting its clinic application. Clearly, identifying the pre-disease state with only a single sample in a reliable manner is challenging but of great importance, since it provides not only a feasible and cost-effective method for clinic diagnosis but also significantly relieves the burden of sample collections for individuals. In this article, by exploring the differential distributions of DNB members and non-DNB members, we developed a novel computational method for efficiently and accurately identifying the pre-disease state before the critical transition in a complex disease with just a single sample (Fig. 1b).

Specifically, DNB molecules typically have a double-peak distribution in the pre-disease state owing to their strongly fluctuating and correlated nature in the pre-disease stage, but have a single-peak distribution in the normal state due to their dynamically stable behavior in the normal stage (see Supplementary Material A for details). On the other hand, non-DNB molecules are observed to show a single-peak distribution even in the pre-disease state due to their stable behavior (Fig. 1c). Thus, differential distributions between those groups of molecules can be identified from a single sample provided there are multiple control samples, which are generally available in practice. In this article, we use Kullback–Leibler divergence (K–L divergence), which measures the difference between two data distributions to formulate the DNB single-sample score (the DNB-S score) by three factors. Each factor in the DNB-S score reflects an aspect of the dynamic features when the system is near a transition point (Fig. 1). Based on this scheme, a new scoring criterion, i.e. the DNB-S score, was proposed as a quantitative measurement of each sample (Fig. 1d). Note that too few genes in the DNB may give a biased result due to incomplete information on the distribution.

Moreover, we have theoretically and numerically shown that the DNB-S score is relatively high when the system is in a pre-disease state, and relatively low when the system is in a normal or disease state. Hence, the DNB-S score provides a reliable signal to identify the pre-disease state. Both theoretical and computational results show that high-dimensional information of data can be utilized to compensate insufficient samples (Huang *et al.*, 2011; Sciuto *et al.*, 2005; Saeki *et al.*, 2009), which is the main reason why DNB can detect the pre-disease state by a single sample with high-throughput measurements.

In addition, our previous study indicates that the DNB-S score is a model-free approach that can be theoretically applied to any disease or biological system with clear transition events. Furthermore, the DNB-S score in this work is not used for identifying the critical transition point or early disease state but for detecting the state just before the critical transition point (i.e. the pre-disease state), and therefore, in comparison with existing methods, it is of great importance for the early diagnosis of complex diseases. To demonstrate the effectiveness and efficiency of the theoretical work, we applied our method to detect the pre-disease state for three diseases by single samples, i.e. lung injury induced by carbonyl chloride inhalation exposure (GSE2565), MCF-7 human breast cancer resulting from heregulin (HRG) (GSE13009) and human influenza infection caused by *H3N2* virus (GSE30550), all of which successfully validated our predictions.

2 METHODS

We first describe the theoretical basis, i.e. the DNB theory, and the mathematical basis of the DNB-S score, and then provide the procedures used

to preprocess input datasets. The detailed algorithm for the DNB-S score is given in Supplementary Material C.

2.1 Theoretical basis

Disease progression can be generally divided into three stages, i.e. (i) the normal state, (ii) the pre-disease state and (iii) the disease state (Fig. 1a). The normal state is a 'healthy' stage, in which the state change is gradual. The pre-disease state is actually the limit case of the normal state just before the critical transition. Further progression in the pre-disease state will result in a drastic state change to the disease state. However, the pre-disease state is considered reversible to the normal state because appropriate medical treatments or a change in lifestyle can convert this state back to a normal state. On the other hand, the disease state represents a seriously ill stage, from which is difficult to return to the normal state, even with advanced treatment, in the case of major complex diseases such as cancer and diabetes.

Traditional biomarkers, e.g. molecular biomarkers and network biomarkers, are designed to distinguish disease samples from normal samples (Liu *et al.*, 2013a, b, c). For example, a molecule is taken as a biomarker if its abundance (or expression) is significantly higher or lower in the disease state than in the normal state, thereby reflecting the severity or presence of the illness in the disease state (Supplementary Fig. S1a). On the other hand, the DNB scheme aims to distinguish pre-disease samples from normal samples according to the correlations and fluctuations of DNB molecules (Supplementary Fig. S1a, b, d and e) (Liu *et al.*, 2013a, b, c). In other words, the DNB method aims to screen out a group of strongly correlated and wildly fluctuating molecules, which are also called 'the leading network' (Liu *et al.*, 2012) because those molecules may together make the first move to lead the whole system from the normal state to the disease state.

Although elucidating the critical transition at the network level holds the key to understanding the fundamental mechanism of disease development and progression, it is notably hard to reliably identify the pre-disease state because there are few apparent differences between the normal and pre-disease states (Fig. 1a); note that a pre-disease state is still a normal state, i.e. it is the limit of the normal state just before the disease state. This is also the reason why diagnosis based on traditional biomarkers may fail to indicate the pre-disease state (Supplementary Fig. S1c). The theoretical basis for the DNB is summarized by the following conditions, which have been proved to hold simultaneously when the system approaches the pre-disease state (see Supplementary Material A for details):

(1) The deviation of a group of molecules, called the DNB from the property of the whole population, drastically increases, i.e. the fluctuation condition.

(2) The average correlation between any two molecules in the DNB increases, i.e. the correlation or internal connection condition.

(3) The average correlation between any molecule in the DNB and another in the non-DNB decreases, i.e. the correlation or external connection condition.

(4) There are no significant changes in the deviations and correlations of molecules among the remaining molecules of the system, i.e. the non-DNB.

The above four conditions together define the DNB. The DNB genes can be detected by the algorithms (Chen *et al.*, 2012; Liu *et al.*, 2013a, b, c) based on the above conditions, but in this work we assume that DNB genes are available for each disease (see Supplementary Materials D and E for cross-validation tests), by which we further identify the pre-disease state with a single sample. Dynamics satisfying the preceding conditions can be viewed as local herding behavior, i.e. members in a group or subnetwork act together without planned direction (or show strongly correlated fluctuation in the whole group). These conditions imply an imminent regime shift or phase transition, and therefore, are used in DNB theory to signal the emergence of the critical transition. Such a phenomenon can also be described as a condition where all the DNB molecules become dynamically correlated or connected so that the system can be reorganized or reconnected in a different way or regime (state).

However, directly applying the above conditions requires multiple samples from an individual who may not even have clinic disease symptoms, which restricts the clinic application. Next, we derive a new criterion, i.e. the DNB-S score from a single sample by exploring interaction information from high-throughput data based on the above conditions, which can compensate for the insufficient samples.

2.2 Differential distributions by K–L divergence

From the theoretical result of DNB, DNB molecules are typically expected to have a double-peak distribution in the pre-disease state due to their significant differential expression and strongly correlated fluctuation nature (conditions 1–3), but non-DNB molecules have a single-peak distribution even in the pre-disease state due to their stable behavior (condition 4) (see Supplementary Material A for details). On the other hand, in the normal state, both the DNB and non-DNB molecules have single-peak distributions due to their dynamically stable behavior (Fig. 1). These differential distributions between DNB and non-DNB molecules and between DNB molecules of normal and pre-disease states can be identified from a single sample provided that there are multiple control samples, which are generally available in practice. In this article, we use the K–L divergence, which measures the difference between two data-distribution patterns to formulate the DNB-S score based on three factors (Fig. 1).

For two discrete probability distributions P and Q, i.e. those of DNB molecules or non-DNB molecules with normalized values, the K–L divergence of Q from P is defined as

$$D_{KL}(P,Q) = \sum_k \ln\left(\frac{P(k)}{Q(k)}\right)P(k), \tag{1}$$

where $P(k) = \mathrm{Prob}_P(x = x_k)$ and $Q(k) = \mathrm{Prob}_Q(y = y_k)$ with

$$\sum_k P(k) = 1 \text{ and } \sum_k Q(k) = 1.$$

In an information theory context (Cover *et al.*, 2005), K–L divergence is actually the relative entropy, i.e.

$$D_{KL}(P,Q) = H(P,Q) - H(P), \tag{2}$$

where $H(P,Q)$ is the cross entropy of P and Q, and is related to the information lost in P if only Q is known. $D_{KL}(P, Q)$ is zero only when the distribution P is identical with Q, and is positive otherwise.

The K–L divergence was originally proposed for measuring the difference between two data distributions (Kullback *et al.*, 1968, 1987) and further extended to serve as a theoretical basis for data differencing (Shamilov and Giriftinoglu, 2010), outlier detection (Oh *et al.*, 2008) and evaluating sample similarity (Lindorff-Larsen and Ferkinghoff-Borg, 2009; Zhou and Chellappa, 2006). Assuming that P_A and P_B are two datasets, respectively, for variables or measurements in two samples (A, B), if the score $D_{KL}(P_A, P_B)$ is zero, then the two samples represent the same amount of information, thereby possessing the most similarity. During the progression of a complex disease, if the K–L divergence of two samples is very small, then it is natural to regard that the stages from which the samples are derived are very similar. Thus, the K–L divergence is a natural choice for comparing two samples with a number of measurements.

2.3 Data processing and algorithm

Three gene-expression-profiling datasets were downloaded from the NCBI GEO database (ID: GSE2565, GSE13009, GSE30550) (www.ncbi.nlm.nih.gov/geo). In these datasets, probe sets without corresponding gene

symbols were not considered in our analysis. The expression values of probe sets mapped to the same gene were averaged. Genes in the DNBs for the three diseases were linked and correlated based on combined functional coupling information from various databases of protein–protein interactions such as STRING, FunCoup and BioGrid.

In each disease dataset, expression profiling information was individually mapped to the integrated networks to identify the corresponding DNB. For each species, we downloaded biomolecular interaction networks from various databases, including BioGrid (www.thebiogrid.org), TRED (www.rulai.cshl.edu/cgi-bin/TRED/), KEGG (www.genome.jp/kegg) and HPRD (www.hprd.org). First, the available functional linkage information for *Mus musculus* and *Homo sapiens* was downloaded from these databases and combined. For instance, after removing any redundancy, we obtained 37 950 linkages in 6683 mouse proteins/genes for acute lung injury. Next, the genes evaluated in these microarray datasets were mapped individually to their integrated functional linkage networks. For the influenza dataset, gene-expression profiles were obtained and measured on whole peripheral blood drawn from all subjects at an interval of 8 h post inoculation (hpi) through 108 hpi. A total of 267 gene microarrays were obtained for all subjects at 16 time points, including the baseline (224 hpi). Networks were visualized using Cytoscape (www.cytoscape.org). Besides, the dataset for breast cancer was obtained in an experiment on the MCF-7 cell line with HRG stimulation. The algorithm of the DNB-S score and its application to the three real datasets are described in Supplementary Materials C and D.

2.4 Functional analysis

As described in Supplementary Material D, we performed pathway-enrichment analysis and -functional analysis for the identified DNBs of the three diseases. The full list of DNBs is provided in the Supplementary Table 'Identified DNBs'.

3 RESULTS

We use a single sample with high-throughput data, e.g. genomics or proteomics data, to identify the pre-disease state or early warning signal of a disease based on the DNB-S score. Achieving such a reliable diagnosis is of great importance in clinic application since one sample can be obtained much more easily than multiple samples for each individual who does not yet exhibit any disease symptoms during the short period before the critical transition.

3.1 Identifying the pre-disease state from a single sample

By exploring the dynamical properties of the underlying system near a critical point, we are in a position to design a computational method for identifying the pre-disease state based on a single case sample. First, remember that there are two groups of variables in the high-throughput data in a single sample, i.e. a group of DNB members and a group of non-DNB members (or the remaining molecules except DNB members in the system). We define the DNB-S score to identify the pre-disease state when only a single case sample is available (see Supplementary Material A). Specifically, given a single case sample, a number of control samples (or normal samples) and the identified DNB, we can construct a composite index I for the pre-disease state based on the differential distributions between the case sample and the control samples:

$$I = \frac{D_{KL}(\text{case}_{\text{DNB}}, \text{control}_{\text{DNB}}) \times D_{KL}(\text{case}_{\text{DNB}}, \text{case}_{\text{non-DNB}})}{\varepsilon + D_{KL}(\text{case}_{\text{non-DNB}}, \text{control}_{\text{non-DNB}})} \quad (3)$$

which is called the DNB single-sample score (DNB-S score). Here, ε is a small positive number to avoid zero division. We require a number of control samples to obtain a stable background distribution. Owing to the nature of the DNB (conditions 1–4), when the system approaches the pre-disease state from the normal state, the terms of the DNB-S score (3) have the following features:

- The K–L divergence of the case sample and the control samples of the DNB, i.e. $D_{KL}(\text{case}_{\text{DNB}}, \text{control}_{\text{DNB}})$ increases due to high differential distributions of the DNB members between the case sample and the control samples, i.e. DNB molecules are typically expected to have a double-peak distribution in the pre-disease state, which is completely different from the single-peak distribution of DNB molecules in the normal state (as demonstrated in the numerical simulation in Fig. 2c).
- The K–L divergence of the case sample between DNB and non-DNB molecules, i.e. $D_{KL}(\text{case}_{\text{DNB}}, \text{case}_{\text{non-DNB}})$ increases due to high differential distributions between DNB members and non-DNB, i.e. DNB molecules are typically expected to have a double-peak distribution in the pre-disease state, which is completely different from the single-peak distribution of non-DNB molecules in the pre-disease state (as shown in Fig. 2d).

Clearly, the above two terms in the criterion are mainly used to detect the pre-disease state. On the other hand, when the system moves to the disease state after passing the pre-disease state, the third term of the DNB-S score (3) has the following feature:

- The K–L divergence of the case sample and the control samples of non-DNB, i.e. $D_{KL}(\text{case}_{\text{non-DNB}}, \text{control}_{\text{non-DNB}})$ has no significant change in both the normal and pre-disease states, but usually increases in the disease state due to differential distributions of non-DNB members between case and control samples in the disease state, i.e. non-DNB molecules are typically expected to exhibit similar single-peak distributions in both normal and pre-disease states, but have different average values or distributions for normal and disease samples.

Different from the pre-disease and normal states, significant differences between the disease and normal states are expected (from the third feature). Thus, the K–L divergence between case and control samples of non-DNB usually increases in the disease state, resulting in a low level of the DNB-S score. Therefore, the third term in the criterion also contributes to distinguish between the disease state and the pre-disease state. However, the pre-disease state can be detected mainly by the first two terms in the criterion.

Integrating the above properties, we can detect the pre-disease state based on a single case sample by the DNB-S score (3); that is, if the criterion (3) is much higher than a threshold, then the case sample is diagnosed as being in the pre-disease state. Actually, the combined criterion from the three terms also reduces the effects of noise and data errors, and thus improves the sensitivity for detecting the critical information in a pre-disease sample.

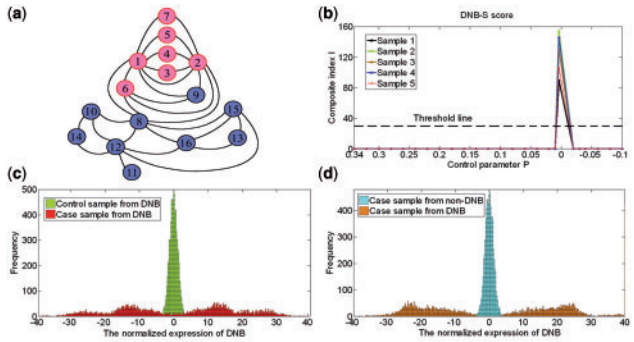

Fig. 2. Numerical validation of the theoretical model. **(a)** The graph is a 16-node network. $z_1, z_2, \ldots, z_7$ are DNB members derived from the dynamic data. The critical transition is at parameter $P = 0$ in the theoretical model, where the system undergoes a critical transition driven by DNB or the leading network composed of z_i ($i = 1, 2, \ldots, 7$). **(b)** shows five samples of the composite index (DNB-S score). Clearly, the DNB-S score for each sample is very high (far beyond the threshold line) when the system is in the pre-disease state as the parameter P approaches the critical value 0. **(c)** The distributions, respectively, for the normalized DNB control data and normalized DNB case data, which are clearly distinct. The distribution of DNB molecules in a case data (or sample) shows a double peak whereas the distribution in a control data has a single peak. **(d)** presents the distributions for the normalized DNB case data (double peak) and normalized non-DNB case data (single peak), which show significant differences. **(c)** and **(d)** present the two important terms of the DNB-S score

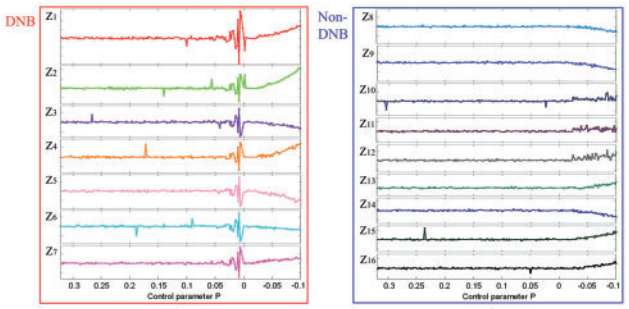

Fig. 3. States of the 16-node network with changing the value of the parameter P. The expression values of the 16 nodes (or state variables) under different values of the parameter P are shown in the graphs. Clearly (i) there are random fluctuations for each node far before the critical transition due to noise; (ii) there are no significant changes for any non-DNB node, e.g. $z_8, z_9, \ldots, z_{16}$, before the critical transition. Thus, a traditional scheme (e.g. critical slowing-down) based on only one or a few variables cannot correctly signal the pre-disease state. In other words, based on the integration of the collective dynamics of a subnetwork, the DNB, i.e., $z_1, z_2, \ldots, z_7$ (the graphs in left column), is expected to provide reliable and correct early warning signals in the pre-disease state, i.e. the expressions of DNB members increasingly fluctuate in a collective manner as the system approaches the critical point

3.2 Numerical experiments

To demonstrate the effectiveness of the DNB-S score, we used a theoretical model of a 16-node network (Fig. 2a) to generate data for each subject or sample. Data for five single-sample subjects were generated and used for validation. Detailed descriptions of the network represented by a set of stochastic differential equations are provided in Supplementary Material B, and the results of numerical experiments are provided in Figure 2. In particular, we obtained the following results.

- When the system is near the critical point (i.e. the parameter P approaches the critical value 0), the DNB-S score (or the two terms in the nominator of the criterion) is in a high level or well above the threshold line (Fig. 2b).
- When the system passes the critical point and is in the disease state (i.e. the parameter P is negative), the term in the denominator of the composite index is at a high level, which results in a lower value of the DNB-S score.

The numerical experiment validates that the DNB-S score is reliable and accurate in identifying the pre-disease state and thus provides the early warning signal of a catastrophic change in the system. Besides, the state changes of the system for all nodes are also presented in Figure 3, from which it can be seen that the DNB group, i.e. a dominant group composed by nodes z_1, $z_2, \ldots, z_7$, shows a clear early warning signal by their coordinated dynamic behavior in a collective manner. 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 (Fig. 3), which demonstrates the advantage of exploiting high-dimensional information using the DNB 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 DNB-S score provides a way to detect the signal for diagnosing the pre-disease state even without a detailed model.

3.3 Application to three diseases

We further applied the DNB-S score to three diseases using high-throughput real data, i.e. microarray data for live influenza infection (humans) caused by *H3N2* virus (GSE30550), acute lung injury (rats) induced by carbonyl chloride inhalation exposure (GSE2565) and MCF-7 human breast cancer caused by heregulin (HRG) (GSE13009). The detailed algorithm and data descriptions are presented in Supplementary Materials C and D, respectively. It is worth mentioning that although time series data are available for each of the three diseases, we identified the pre-disease states in one data point at a time independently, i.e. using a single sample in each identification. Figures 4–6 show the identified pre-disease states just before the critical deteriorations based on the DNB-S score, which agrees well with the observed biological phenotypes described in the original datasets (Huang *et al.*, 2011; Saeki *et al.*, 2009; Sciuto *et al.*, 2005).

Specifically, Figure 4 shows the DNB-S scores for live influenza infection of 17 subjects (17 humans), in which nine subjects were diagnosed as having influenza infection or clinic symptoms (symptomatic subjects) 45 h later and eight subjects were classified as non-symptoms (asymptomatic subjects) during the whole study period (Huang *et al.*, 2011). Figure 4a shows the clinic symptoms (S) and non-symptoms (N) among the 17 subjects with live influenza infection based on real clinic tests. In Figure 4b, we identified

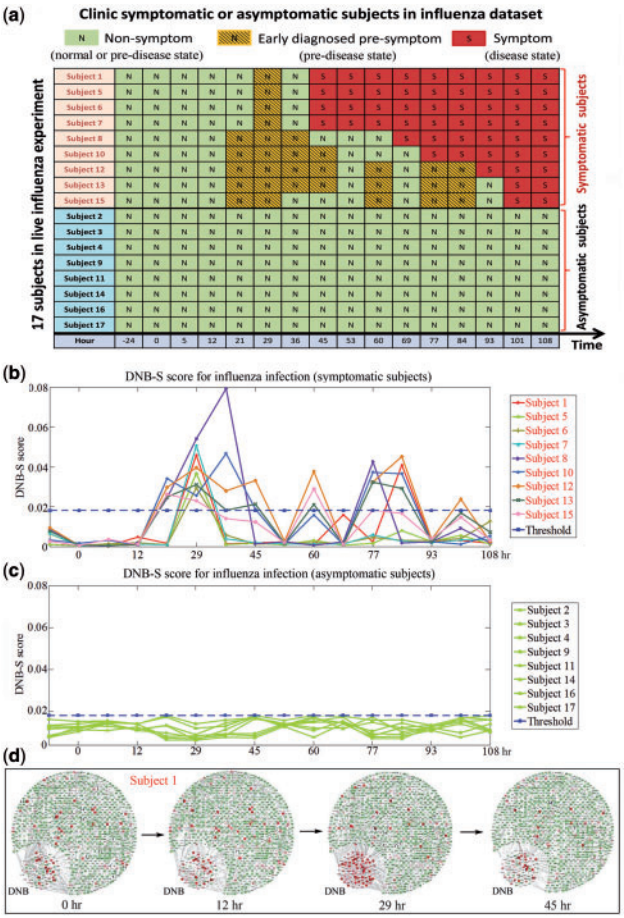

Fig. 4. Identifying the pre-disease state for live influenza infection of 17 humans based on a single sample. We demonstrate early diagnosis of live influenza infection in 17 humans using a real biomedical dataset. **(a)** The clinic symptoms (S) and non-symptoms (N) at different time points among the 17 subjects with live influenza infection based on real clinic tests. **(b)-(c)** show the DNB-S scores of nine single-sample symptomatic subjects (humans) and eight asymptomatic subjects (humans) for influenza infection resulting from *H3N2* virus. The pre-disease states or pre-symptom for influenza infection occurred around 29 h (i.e. 29, 36 and 45 h), whereupon the DNB-S scores became respectively higher than the threshold shown in **(b)**. All symptomatic subjects were correctly identified before the clinical diagnosis of the disease state **(b)**, whereas all asymptomatic subjects showed no signals of the pre-disease states and were also correctly classified **(c)**. **(d)** The dynamic changes in the molecular network of a single-sample subject (Subject 1) at 0, 12, 29 and 45 h (sliding window) with the corresponding DNB, where the color of the nodes represents the fluctuation strength of molecular expressions, and each edge represents the correlation between two nodes. It can be seen that at 29 h, there is a strong signal to indicate the pre-disease state or pre-symptom

the pre-disease states (early diagnosed pre-symptom) of nine symptomatic subjects before the appearance of clinic symptoms, while our DNB-S scores indicated the absence of pre-disease states for the other eight asymptomatic subjects shown in Figure 4c. Clearly, the identified pre-disease states are respectively from 29 to 45 h, which are well before the earliest clinic symptom that appeared at 45 h for the nine symptomatic subjects

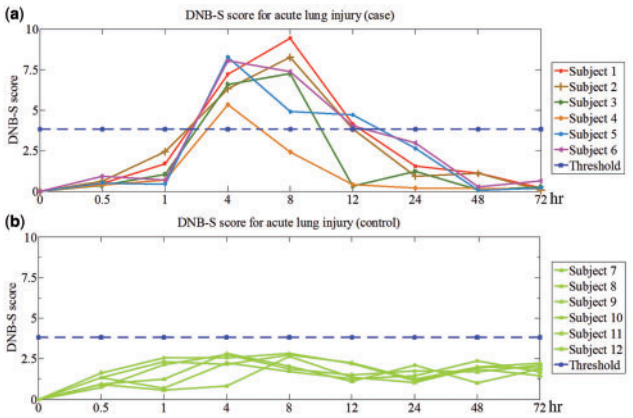

Fig. 5. Identifying the pre-disease state for acute lung injury induced by carbonyl chloride inhalation exposure based on a single sample. **(a** and **b)** Six single case subjects (rats) and six control subjects (rats) for acute lung injury induced by carbonyl chloride inhalation exposure. In **(a)**, it can be seen that the DNB-S scores are well above the threshold line (blue line) at 4 and 8 h, much before the 24 h time point in the disease state. All case samples or case subjects were correctly identified before the onset of serious deterioration of the disease state **(a)**, whereas all normal samples show no signal pertaining to the pre-disease state and were also correctly classified **(b)**

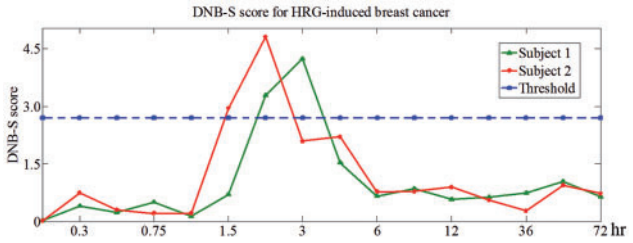

Fig. 6. Identifying the pre-disease state for human breast cancer caused by heregulin (HRG) based on a single sample. In this figure, the pre-disease state is identified as being ~2 h, at which both of the DNB-S scores are higher than the threshold (blue line). Both the case samples and case subjects were correctly identified before the serious deteriorated disease state

based on the original clinic tests, thereby validating our method and the pre-disease states (see Fig. 4a). Hence, our early diagnoses agree with the original results (Huang *et al.*, 2011). In addition, a figure illustrating dynamic changes of the whole molecular network of a single-sample subject (i.e. Subject 1) from 0 to 45 h is shown in Figure 4d, where a strong signal for the pre-disease state can be observed at 29 h, and a complete representation of all the three terms in the DNB-S score is provided as Supplementary Figure S10. From Figure 4a, it can be seen that for all the nine symptomatic subjects, the DNB-S score is higher in the pre-disease states and thus detects the early warning signal of the deterioration before clinic symptoms could be observed, whereas the DNB-S score correctly shows no signals for eight clinic asymptotic subjects. It can be seen that the pre-disease stages for Subjects 8, 10, 12, 13 and 15 are longer than those for Subjects 1, 5, 6 and 7. Furthermore, there are multiple critical stages (pre-disease states) for Subjects 12, 13 and 15, which may

represent multiple deterioration processes for these three subjects. Actually, there is an approximate 24 h interval for these three critical stages, i.e. $36 \rightarrow 60 \rightarrow 84$ h, which may be related to the immune responses of the three subjects before the eventual defeat to the disease state; the analyses of their mechanisms will be considered as a future research topic.

Figure 5a shows the DNB-S scores for six single case subjects of acute lung injury, where between 1 and 12 h, the DNB-S scores were all above the threshold line. Therefore, we identified the pre-disease state at ~4 h. In the original experiment, a 50–60% mortality was routinely observed after 12 h and a 60–70% mortality was observed after 24 h (Sciuto *et al.*, 2005). The major deterioration on average thus emerges ~24 h (the 7th sampling time point). It can be seen from Figure 5a that the DNB-S scores are well above the threshold line at 4 and 8 h, which indicates the pre-disease state before the critical point at 24 h. Moreover, all control samples correctly show no signal based on the DNB-S score (Fig. 5b). Therefore, using a single case sample, the DNB-S score is able to identify the pre-disease state, which is consistent with our previous results (Chen *et al.*, 2012; Liu *et al.*, 2012) and the observed experimental results (Sciuto *et al.*, 2005). The curves for the three terms in the criterion are provided in Supplementary Figure S8.

Figure 6 shows the DNB-S scores for two single-sample subjects of HRG-induced breast cancer, where the DNB-S scores are all above the threshold line at 2 h. Therefore, we identified the pre-disease state in each case before 4 h. In the original experiment, the stimulation of MCF-7 breast cancer cells with epidermal growth factor 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 deterioration is ~3 h. These results are in agreement with the original experimental results (Saeki *et al.*, 2009). The curves for the three terms in the composite index are provided in Supplementary Figure S9.

The successful application of the DNB-S score in the three real datasets demonstrate the effectiveness of our method in identifying the pre-disease state and thus provide early diagnosis of critical transitions on the basis of just a single sample. The detailed algorithm and data descriptions for the three diseases are provided in Supplementary Materials C and D, respectively. To validate the effectiveness and accuracy of the identified DNBs as well as the DNB-S score, leave-one-out cross-validation was carried out and is provided in Supplementary Material E. The identified DNBs for the diseases are given in the Supplementary Table 'Identified DNBs'.

4 DISCUSSION

In this article, we developed a novel computational method, i.e. the DNB-S scoring method, to identify the pre-disease state of a disease on the basis of a single sample, which facilitates early diagnosis before the disease state or its serious deterioration. From the viewpoints of both theoretical analysis and numerical computation, we have demonstrated that the DNB-S score is sensitive to any sample near the pre-disease state. All the results show that we identified the pre-disease state, i.e. the state just before a critical transition to the disease state, rather than the disease state targeted by traditional biomarkers. By developing this method, we also found that a single case-sample with high-throughput measurements actually has enriched information sufficient for early diagnosis even if there is no reliable disease model (i.e. the underlying mechanism of disease deterioration is unclear). In other words, the DNB-S score is a model-free approach, which is capable of exploiting high-dimensional information (or interaction information on high-dimensional data) and thus distinguishing pre-disease samples from normal ones.

Our study makes two main contributions. First, the DNB-S score can identify the pre-disease state before moving into the deteriorated disease state by a critical transition, rather than diagnosing the disease state. Therefore, it has profound potential to achieve 'real' early diagnosis for complex diseases. Second, the DNB-S score can detect the pre-disease state with only a single sample for each individual, which is of great importance for clinic applications in realistic cases such as clinic testing and personalized healthcare. All the results show that high-dimensional information of data can be used to compensate for insufficient samples, which is the major reason why the DNB-S score can detect the pre-disease state from a single high-throughput sample.

It is also worth noting that this method is based on the DNB theory and thus the molecules in the DNB can be related to the leading network of the disease. Those molecules may make the first move from the normal state toward the disease state through a critical transition, and thus, be causally related to disease-driving genes or networks. Therefore, the DNB-S score can be a reflection of the leading factors (the driving network or disease-driving genes) to the serious deterioration of complex diseases. In addition to the diagnosis of diseases, our method can be applied to the analysis of other complex processes or phenomena in biology, physics, ecology or even economics in a similar manner provided that drastic state transitions are involved in these processes, e.g. cell-differentiation and cell-cycle processes.

Funding: National Natural Science Foundation of China (Grant numbers 91029301, 61134013, 61072149, 11326035 and 11241002); Fundamental Research Funds for the Central Universities (Grant number 2014ZZ0064); the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant Number KSCX2-EW-R-01) and 863 project (Grant number 2012AA020406); FIRST program from Japan Society for the Promotion of Science initiated by the Council for Science and Technology Policy.

Conflict of Interest: none declared.

REFERENCES

Dakos,V. *et al.* (2008) Slowing down as an early warning signal for abrupt climate change. *Proc. Natl Acad. Sci. USA*, **105**, 14308–14312.

Carpenter,S.R. *et al.* (2011) Early warnings of regime shifts: a whole-ecosystem experiment. *Science*, **332**, 1079–1082.

Carpenter,S.R. and Brock,W.A. (2006) Rising variance: a leading indicator of ecological transition. *Ecol. Lett.*, **9**, 311–318.

Carpenter,S.R. (2005) Eutrophication of aquatic ecosystems: bistability and soil phosphorus. *Proc. Natl Acad. Sci. USA*, **102**, 10002–10005.

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

Cover,T. and Thomas,J. (2005) *Elements of Information Theory*. Wiley, New Jersey.

Drake,M.J. and Griffen,D.B. (2010) Early warning signals of extinction in deteriorating environments. *Nature*, **467**, 456–459.
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.
Held,H. and Kleinen,T. (2004) Detection of climate system bifurcations by degenerate fingerprinting. *Geophys. Res. Lett.*, **31**, L23207.
Hirata,Y. *et al.* (2010) Development of a mathematical model that predicts the outcome of hormone therapy for prostate cancer. *J. Theor. Biol.*, **264**, 517–527.
Huang,Y. *et al.* (2011) Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza infection. *PLoS Genet.*, **7**, e1002234.
Kambhu,J. *et al.* (2007) *New Directions for Understanding Systemic Risk: A Report on a Conference Cosponsored by the Federal Reserve Bank of New York and the National Academy of Sciences*. The National Academies Press, Washington D.C.
Kleinen,T. *et al.* (2003) The potential role of spectral properties in detecting thresholds in the earth system: application to the thermohaline circulation. *Ocean Dynam.*, **53**, 53–63.
Kullback,S. (1968) *Information Theory and Statistics*. Dover Press, New York.
Kullback,S. *et al.* (1987) Letter to the Editor: The Kullback Leibler distance. *Am. Stat.*, **41**, 340–341.
Lenton,T.M. *et al.* (2008) Tipping elements in the earth's climate system. *Proc. Natl Acad. Sci. USA*, **105**, 1786–1793.
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. Briefings in Bioinformatics [Epub ahead of print, DOI: 10.1093/bib/bbt027, February 9, 2014].
Lindorff-Larsen,K. and Ferkinghoff-Borg,J. (2009) Similarity measures for protein ensembles. *PLoS one*, **4**, 1–13.
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. *Sem. 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) Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. Medicinal Research Reviews [Epub ahead of print, DOI 10.1002/med.212932013, February 9, 2014].
Liu,R. *et al.* (2013b) Dynamical network biomarkers for identifying critical transitions and their driving networks of biologic processes. *Quant. Biol.*, **1**, 105–114.
Liu,X.P. *et al.* (2013c) Detecting early warning signals of type 1 diabetes and its leading biomolecular networks by dynamical network biomarkers. *BMC Med. Genom.*, **6**, S8.
May,R.M. (1977) Thresholds and breakpoints in ecosystems with a multiplicity of stable states. *Nature*, **269**, 471–477.
May,R.M. *et al.* (2008) Ecology for bankers. *Nature*, **451**, 893–895.
McSharry,P.E. *et al.* (2003) Prediction of epileptic seizures: are nonlinear methods relevant? *Nat. Med.*, **9**, 241–242.
Oh,J.H. *et al.* (2008) Biological data outlier detection based on Kullback–Leibler divergence. In: Hamid,R.A. and Xiaohua,T.H. (eds.) *IEEE International Conference on Bioinformatics and Biomedicine, BIBM'08*. Philadelphia, pp.249–254.
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 Genom.*, **20**, 545–552.
Scheffer,M. *et al.* (2009) Early warning signals for critical transitions. *Nature*, **461**, 53–59.
Scheffer,M. *et al.* (2001) Catastrophic shifts in ecosystems. *Nature*, **413**, 591–596.
Sciuto,A.M. *et al.* (2005) Genomic analysis of murine pulmonary tissue following carbonyl chloride inhalation. *Chem. Res. Toxicol.*, **18**, 1654–1660.
Shamilov,A. and Giriftinoglu,C. (2010) Generalized entropy optimization distributions dependent on parameter in time series. *WSEAS Transact. Informat.*, **7**, 102–111.
Strogatz,S.H. (1994) *Nonlinear Dynamics and Chaos: with Applications to Physics, Biology, Chemistry and Engineering*. Addison-Wesley, Reading, Massachussets.
Venegas,J.G. *et al.* (2005) Self-organized patchiness in asthma as a prelude to catastrophic shifts. *Nature*, **434**, 777–782.
Zhou,S.K. and Chellappa,R. (2006) From sample similarity to ensemble similarity: probabilistic distance measures in reproducing kernel Hilbert space. *IEEE Transact. Pattern Anal. Mach. Intel.*, **28**, 917–929.