Impact of parameters in Network-Perturbation Signature identification framework

Qier An

Capstone Project Writing and Presentation Final Paper

MScBMI, the University of Chicago

Abstract

Background:

Irreversible phase transitions exist universally in chronic disease dynamic systems. The network perturbation signature (NPS) framework was designed to detect tipping points and corresponding critical transition signatures (CTS) in disease development. The case study aims to investigate if identified tipping point and CTS remain constant when parameters in the NPS R package are changed.

Methods:

We applied the NPS R package to GSE94016 dataset and adopted disparate parameter combinations in eight independent trials. The dataset assesses whole-genome expression of 20 liver tumor samples. The samples were collected from 5 orthotopic xenograft mice, each with 4 time points. The identified CTS's were analyzed with Fisher's test of significance for potential overlapping contingency.

Results:

A signature and its corresponding tipping point were identified for each of the eight trials. The outputs were validated by robustness simulations. The identified CTS's shared a few common genes while remained generally independent to one another.

Conclusion:

Disparate CTS's were identified by the NPS R package under distinct parameters. Nonetheless the majority of signatures recognized the same tipping point for the dataset, regardless of parameter changes.

Introduction

Phase transitions, including deteriorations and metastasis, exist universally in chronic disease systems. Such transitions are mostly irreversible and feature deregulated gene expressions and dysfunctional interactions[1]. Therefore, successfully predicting the tipping point of these systems is vital to the prevention of further disease developments. Past studies have proved that when a chronic disease system reaches its tipping point, certain properties of gene expression level hold true: there exists a group of molecules whose average Pearson's correlation coefficients (PCCs) of molecules drastically increase in absolute value; the average PCCs of molecules between this group and any other drastically decrease in absolute value; and the average standard deviations (SDs) of molecules in this group drastically increase[2]. Based on these properties, the dynamic network biomarker (DNB) algorithm was introduced to generate a quantitative index which effectively identified the aforementioned group of molecules as critical transition signature (CTS). This index increases as a system approaches its tipping point and peaks at the tipping point.

Yet the existing DNB algorithm was designed for longitudinal data thus could not be efficiently applied to cross-sectional patient profiles. For clinically available data, the method suffers from versatile genomic heterogeneities and transcriptomic noise, as well as the inadequacy of health controls[3]. To solve these adversities, we altered the existing algorithm to a network perturbation signature (NPS) framework. It can be applied to datasets in the form of an R tool package. The framework calculates and compares gene expression fluctuation at each time point, thus eliminating the impact of noises from individual patients. And the index being calculated reflects global maximum in the entire disease process, thus reducing the necessity for a healthy control group. The NPS R package comes with several parameters that can be tuned through the actual calculating process: the number of transcriptome profiles pre-selected for variance, the strategy for the pre-selection, and the false discovery rate (FDR) of PCC calculation.

This case study applied the package to dataset GSE94016 which contained gene expression profiles of spontaneous pulmonary metastasis xenograft mouse model. The model contained orthotopic transplanted human hepatocellular carcinoma (HCC) cell line, HCCLM3-REP of high metastatic potential. The model was labeled with a stable fluorescent protein to effectively resolve and quantify the dynamics of tumorigenesis and metastasis[1]. This dataset was analyzed by a previous study in 2018 using the established DNB algorithm[1]. The third week after the orthotopic implantation (W3) was identified as the tipping point and a biomarker with 334 genes was identified as CTS. Such a result was supported by biological evidences from both the observed metastasis process and the roles of a DNB gene, CALML3, in the process. My case study aims to determine whether the established tipping point can be detected using the NPS framework, whether identical tipping points and CTSs can be identified under disparate parameters for a single dataset and to provide recommendations on parameter processing.

Methods

Gene expression data collection

The study used R as its primary analytics platform. All packages can be obtained through CRAN or Bioconductor. The whole-genome expression profiling dataset is publicly available from the NCBI gene expression omnibus database under accession number GSE94016 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE94016). It includes gene expression data of 20 samples of liver tumor cells, gathered from 5 orthotopic xenograft mice at 4 time points: the 2nd, 3rd, 4th and 5th week after the HCC implantation (marked as W2, W3, W4 and W5 respectively). It was downloaded and processed using the GEOquery package to be an expression matrix consisting of transcriptomes id's (unique loci's) as row names and sample names as columns.

NPS analysis

As a case study for the NPS manuscript, this project ran the NPS tool package on GSE94016 in R. The procedure consisted of five steps as illustrated in figure 1: data preprocessing, pre-selecting transcripts, network partition, identifying dynamic network bio-module and identifying tipping point.

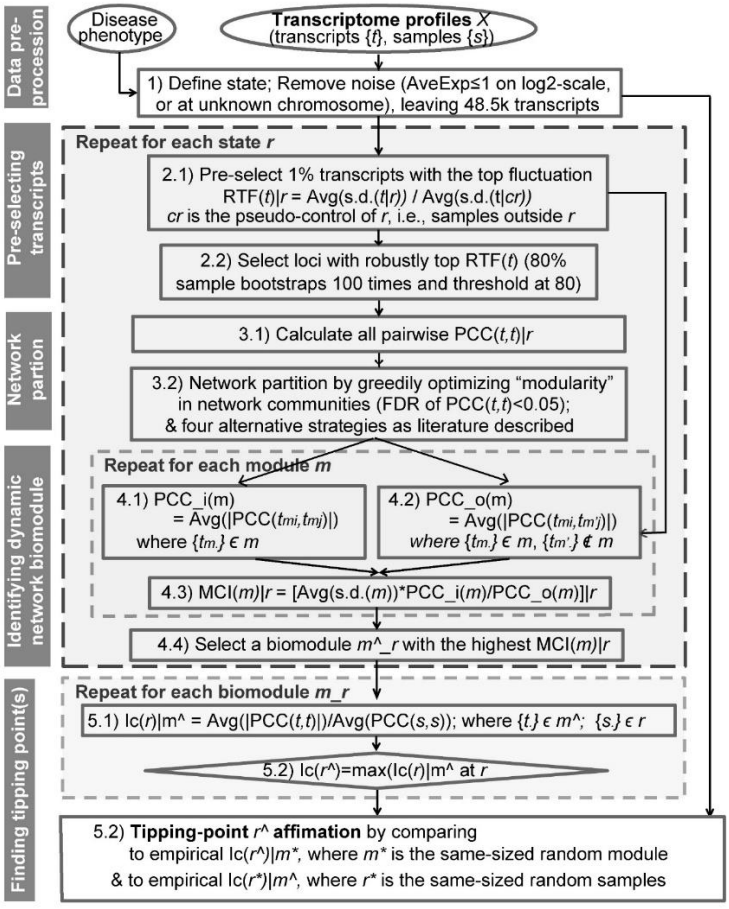

Figure 1. Workflow of the NPS package.

Since increase in standard deviation of the CTS suggests system approaching tipping points, transcriptome profiles with low relative transcript fluctuation (RTF)-score were filtered out, leaving the top 1% fluctuating transcripts. Here the RTF score measures the variance of transcripts t_r of patients in a state r, relative to its complement set t_{cr} (patients outside the state r, serving as a dynamic 'control state')[3],

$$RTF\ (t)|r\ = \frac{sd(t_r)}{sd(t_{cr})}$$

In specific, the complement set t_{cr} was determined with three distinctive methods: single reference group (annotated as 'ref'), all other groups excluding state t (annotated as 'others'), intragroup comparison within group t (annotated as 'itself'). Each method was tested independently and was considered as a part of the parameters. The filtered profiles were clustered by random walk into modules. Afterwards, module-criticality index (MCI) was calculated for each module m by PCC_i (Pearson correlation coefficient of transcriptome profiles within the module), PCC_o (Pearson correlation coefficient of transcriptome profiles between modules) and sd (standard deviation) to achieve the signature group m^. Given,

$$MCI(m)|r = \frac{\overline{sd(m)} * \overline{|PCC\_i(m)|}}{\overline{PCC\_o(m)}}\ |r$$

False discovery rate (FDR), which was used for calculating PCC was the second tunable parameter in this case study. It was set to either 0.05 or 0.1. The module with the highest MCI was identified as the CTS of the system.

Given the CTS m^ in stage r with samples s_r, an index named IC-score was calculated using the PCCs,

$$Ic(r)|m^{\wedge}\ = \frac{|\overline{PCC(t,t)}|}{\overline{PCC(s,s)}};\ where\ \{t.\} \in\ m^{\wedge};\ \{s.\}\ \in\ r$$

Following this equation, the time point corresponding to the highest IC-score was identified as the tipping point r^. Such prediction was validated by two empirical simulations using same-sized random modules and same-sized random samples.

The described work flow was repeated for eight trials. Each trial used a disparate combination of fluctuation pre-selection method and FDR cutoffs of different PCCs.

Venn diagram and Fisher's test of significance

Venn diagrams of the CTSs detected in the trials were drawn to illustrate the independence of the bio-modules. Fisher's test was conducted for each diagram to affirm significance of the observed pattern.

Results

The functions in the NPS R package were applied with different parameter combinations to the dataset GSE94016 in eight trials. For each trial, a critical transition signature (CTS) and its corresponding tipping point were identified. One of those trials, whose outputs are shown in Figure 2, had fluctuation pre-selection method as 'others' and FDR set to 0.1.

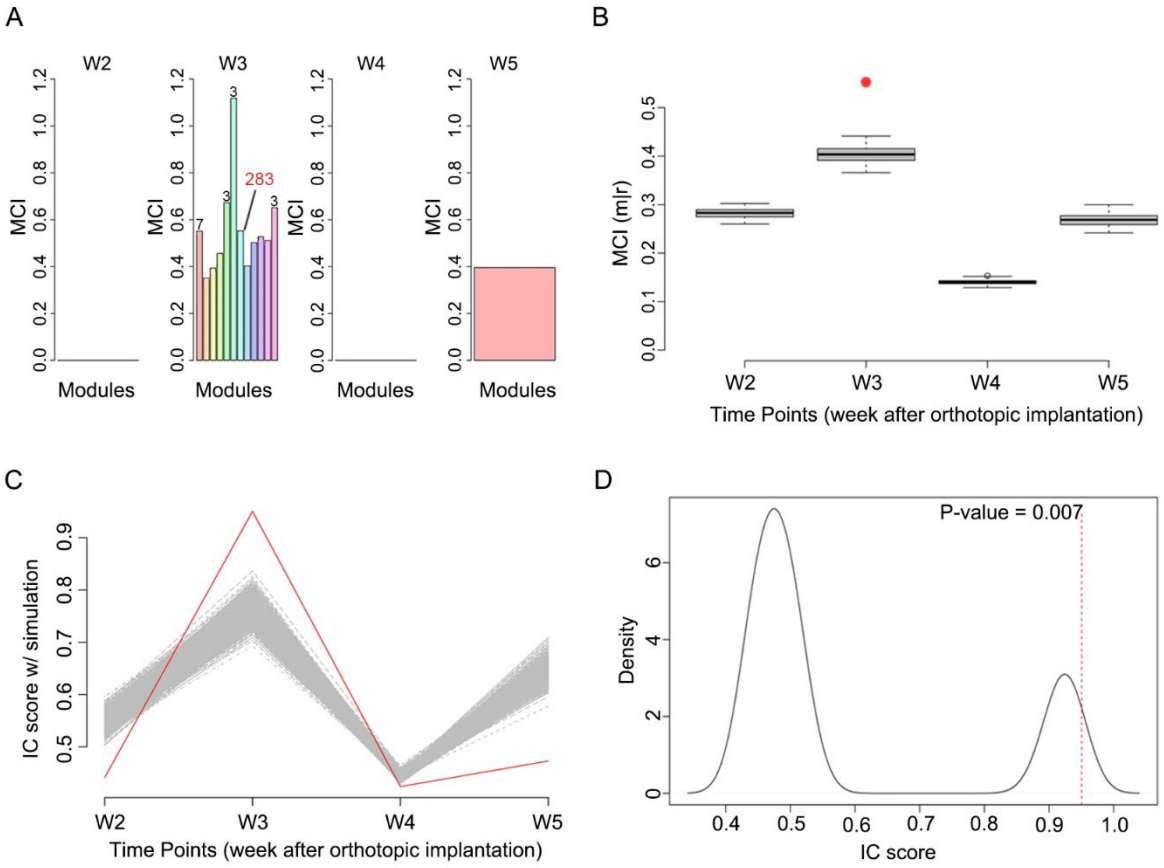

Figure 2. Sample visualizations of outputs for trial "method = 'others', fdr = 0.1" suggests a tipping point at W3. **A**. Bar-plot shows all clustering modules at each time point. **B**. The MCI of CTS (the red point) is higher than the MCI distribution of 100 simulated same-sized modules (as boxplot), validating the significance of CTS. **C**. The IC-score of CTS (in red line) is higher than IC-score of 1000 simulated same-sized modules (in grey lines) at W3, validating the W3 as the tipping point. **D**. Density histogram confirms a higher IC-score of CTS (red dashed line) with p-value < 0.05 comparing to 1000 modules randomly generalized by bootstrapping the same number of samples.

The results for all eight trials were summarized in table 1. Small-sized modules were identified as CTSs in the trials whose RTF scores were calculated using intragroup comparison. Since the significance of the tipping points were not supported by simulations, we modified the method by giving weight adjustment to the modules based on module sizes when calculating MCIs. This new method was conveniently named "itself_weighted".

	Method = Ref						Method = Other				
	Tipping Point	Sig MCI	Sig IC	Module size	Fisher odds		Tipping Point	Sig MCI	Sig IC	Module size	Fisher odds
FDR=0.05	W3	True	True	183	4.94	FDR=0.05	W3	True	True	84	2.29
FDR=0.1	W5	True	True	123	1.05	FDR=0.1	W3	True	True	283	2.21
	Method = Itself						Method = Itself_weighted				
	Tipping Point	Sig MCI	Sig IC	Module size	Fisher odds		Tipping Point	Sig MCI	Sig IC	Module size	Fisher odds
FDR=0.05	W2	True	False	10	0	FDR=0.05	W3	True	True	162	0
FDR=0.1	W2	True	False	20	0	FDR=0.1	W3	True	True	292	0.65

Table 1. Summary table for the parameter sets tested in this case study. "Method" stands for fluctuation pre-selection methods. The "itself_weighted" method was under the "itself" criteria and had the "adjust.size" option turned on in the R package. "True" for "Sig MCI" (or "Sig IC") stands for consonance of the CTS (or tipping point) identified by the NPS framework and the robustness simulation result, and vice versa.

Furthermore, a Venn diagram was drawn for each CTS identified by the NPS framework and the established biomarker identified by the DNB method (Figure 3). Fisher's test of significance was used to demonstrate the extent of overlapping between the bio-modules. As a result, the CTS which was identified from top fluctuating transcripts involving a reference group (W2 in this case) and FDR as 0.05 (ref0.05) overlapped the most with the DNB biomarker, giving an odds ratio of 4.94 with p-value < 0.01. On the other hand, the CTS identified by the weighted-adjusted intragroup comparison method with a 0.05 FDR (itselfw0.05) was mostly independent from DNB biomarker with odd ratio down to 0.

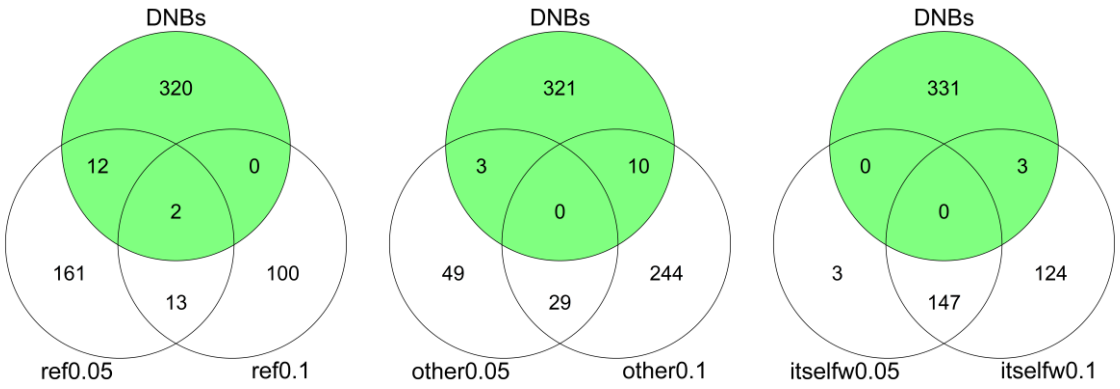

Figure 3. Venn Diagrams illustrating overlapping between critical transition signatures (CTS) identified by NPS package and DNB algorithm.

Discussion

Conclusion

Judging from the study results, the NPS framework successfully detected the established tipping point for the GSE94016 dataset on xenograft mouse model. Changing the parameters led disparate biomarkers to be identified as signatures. Yet the IC-score simulations performed with disparate module sizes reached the global maximum at W3 congruously, indicating it as an incontrovertible tipping point (Figure 4).

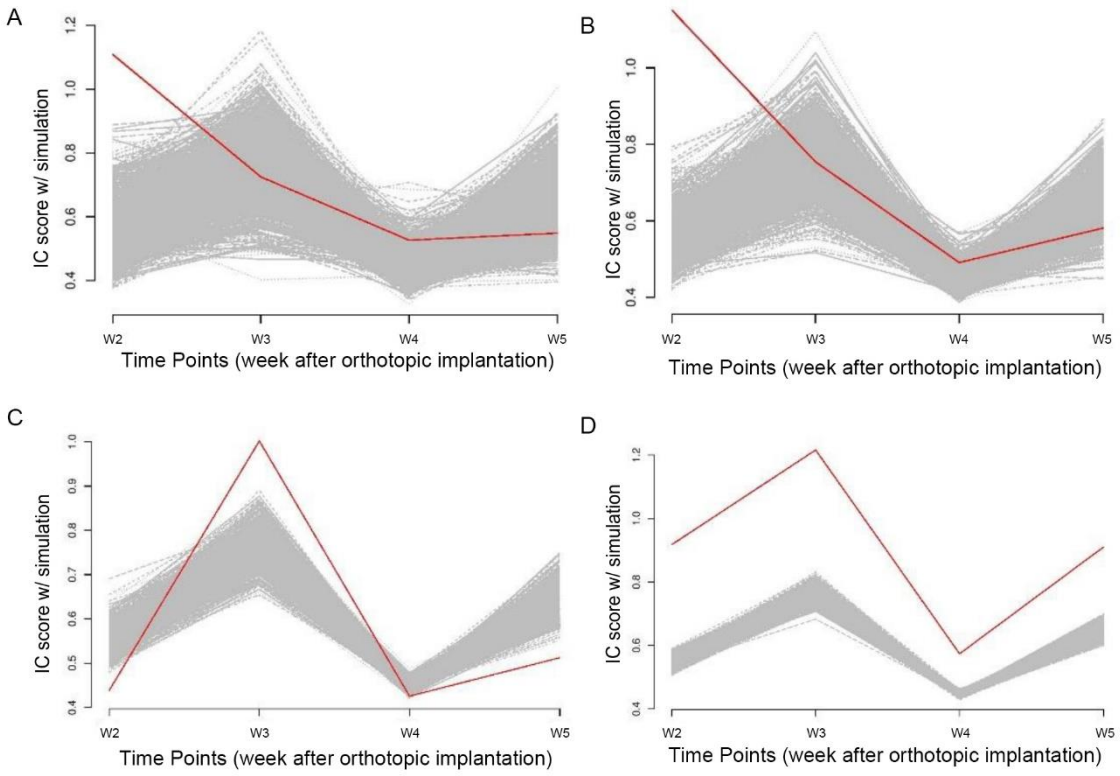

Figure 4 IC scores (in red) with simulation under disparate module size (in grey). **A**. module size = 10 (itself - fdr0.05). **B**. module size = 20 (itself - fdr0.1). **C**. module size = 84 (other - fdr0.05). **D**. module size = 292 (itself_weighted - fdr0.1).

For CTS with small module sizes, IC-score robustness cannot be validated by simulation. In such scenario, weight adjustment is recommended for MCI calculation. This operation gives heavier weight to modules with larger sample sizes and can thus reduce noises caused by intercorrelation.

Next Steps

While the case study focused on exploring outputs of the NPS framework under disparate parameters, it saved some adjustments that might be necessary for other datasets. The dataset GSE94016 has a large sample size with sufficient normally distributed transcriptome expression profiles for analysis. These

features may not hold true for other datasets. In those cases, feature engineering including transformation and normalization is needed.

For this case study, we arbitrarily set the cutoff for minimum module size at 10 when identifying critical transition signatures. The cutoff allowed us to capture significant modules meanwhile eliminated the randomness from smaller nodes. The algorithm for perfect cutoff value, however, was left to be explored in future studies.

Moreover, though the majority of outputs recognized the established W3 as the tipping point for this study, there was indeed a trial which identified W5 instead. In that trial a reference time point W2 was used as the complement set for fluctuation pre-selection and FDR was set to 0.1. The result was validated by IC-score simulation and density simulation of samples under the same parameters. Nonetheless following the majority results, we concluded W3 as the NPS-identified tipping point. Yet a need to investigate reasons for the spotted inconsistency is necessary, both biologically and statistically.

References

1. Yang, B., Li, M., Tang, W., Liu, W., Zhang, S., Chen, L., & Xia, J. (2018). Dynamic network biomarker indicates pulmonary metastasis at the tipping point of hepatocellular carcinoma. Nature Communications, 9(1). doi:10.1038/s41467-018-03024-2
2. Chen, L., Liu, R., Liu, Z., Li, M., & Aihara, K. (2012). Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Scientific Reports,2*(1). doi:10.1038/srep00342
3. Yang, X. H., Wang, Z., Griggs, D., An, Q., Tang, F., & Cunningham, J. M. (2019). Adopting tipping-point theory to transcriptome profiles unravels disease regulatory trajectory. doi:10.1101/668442