Sparse Pathway-Induced Dynamic Network Biomarker Discovery for Early Warning Signal Detection in Complex Diseases

Abolfazl Doostparast Torshizi, Linda Petzold

Abstract—In many complex diseases, the transition process from healthy stage to catastrophic stage does not occur gradually. Recent studies indicate that the initiation and progression of such diseases are comprised of three steps including healthy stage, pre-disease stage, and disease stage. It has been demonstrated that a certain set of trajectories can be observed in the genetic signatures at the molecular level, which might be used to detect the pre-disease stage and to take necessary medical interventions. In this paper, we propose two optimization-based algorithms for extracting the dynamic network biomarkers responsible for catastrophic transition into the disease stage, and to open new horizons to reverse the disease progression at an early stage through pinpointing molecular signatures provided by high-throughput microarray data. The first algorithm relies on meta-heuristic intelligent search to characterize dynamic network biomarkers represented as a complete graph. The second algorithm induces sparsity on the adjacency matrix of the genes by taking into account the biological signaling and metabolic pathways, since not all the genes in the ineractome are biologically linked. Comprehensive numerical and meta-analytical experiments verify the effectiveness of the results of the proposed approaches in terms of network size, biological meaningfulness, and verifiability.

Index Terms—Dynamic Network Biomarker, Disease Progression, Time-Course Gene Expression Data, Optimization.

1 Introduction

HIGH-throughput data-generating technologies have provided scientists with a huge wealth of data corresponding to various levels of genomic interactions such as transcriptome, metabolome, DNA methylation, gene expression etc. These datasets enable researchers to decipher the underlying complex mechanisms of the evolution and progression of such diseases, by employing their molecular interactome structure in a holistic manner. Although the number of features e.g., probes, in such datasets is quite large, only a fraction of them reflect the behavior of the disease of interest. This can be due either to stability of the expression of these probes before and after the disease progression or to their correspondence to non-coding regions of the genome. As a result, a certain set of features called biomarkers are generally used for downstream analysis such as phenotype classification. A common approach for biomarker discovery is to use basic statistical tools such as the t-test to detect differentially expressed genes between diseased patients and controls. This test indicates the difference of the abundance of the given molecules between healthy and disease samples over the course of time. Such biomarkers yield insight into the disease of interest when dealing with relatively simple diseases. Nevertheless, biomarkers playing a crucial role in complex diseases might not necessarily be differentially expressed but they act as a whole in the context of networks. This can be observed in genetically related diseases like cancer [1]. Network-based biomarker discovery approaches have recently received a great deal of attention both in static and time-variant high-throughput studies, due to their capacity to explain emergent properties such as modularity, phenotypic variation and biological heterogeneity [2], [3].

Dynamic Network Biomarker discovery (DNB) approaches conform to the dynamic mutations in molecular signatures of complex diseases and can portray the time-dependent alterations of such biomarkers monitored and evaluated at different stages of the disease [4]. One of the recent advances in the DNB field is in detection of early signals of complex diseases [4]. Similar to many real systems such as ecosystems [5] or climate systems [6] whose current states abruptly shift to another state at a tipping point, recent studies have revealed that a similar phenomenon is observable in clinical medicine. For instance, in chronic diseases such as cancer, deterioration is abrupt rather than smooth [7], [8]. In other words, there is a drastic shift from the healthy stage to the catastrophic stage. As a result, the disease progression evolution can be categorized into three distinct stages: healthy stage, pre-disease stage, and disease stage. During the healthy stage, which is also called the incubation stage or chronic inflammation period [9], the disease is under control and no clinical symptoms are observed. The pre-disease stage reflects the limit of the healthy stage before transition to the disease stage. This stage may be reversible to the healthy stage under appropriate and on time therapeutic interventions [9]. After transition to the disease stage, it would be a daunting task to reverse the disease progression trajectory. Thus, detecting the sub-network biomarkers representing the early warning signals of diseases has been the key objective of our research.

Using bifurcation point theory in dynamical systems theory, based on time-course high-throughput microarray data, Chen [9] *et al.* have shown that during the pre-disease

• *A. Doostparast and L. Petzold are with the Department of Computer Science, Univesity of California, Santa Barbara, CA, 93106.*

stage, a sub-module of gene-regulatory networks represents informative behavior. At this stage, the intra-cluster correlation coefficient among its constituent genes radically increases, while the inter-cluster correlations between the members of the biomarker sub-module and the remaining genes drastically decrease. Additionally, the standard deviation of expression levels of the sub-module biomarkers tends to increase. Given these three signals, they propose a composite index, and they find the biomarkers representing these patterns at each time step, where the sub-module yielding the highest index is treated as the early-warning network biomarker set. Chen *et al.*, Li *et al.* [11] and Liu *et al.* [10] have proposed similar approaches. Li *et al.* [11] employed the DNB discovery approach to identify the tissue-specific biomarker networks associated with type-2 diabetes mellitus (T2DM). They have used gene expression data from multiple tissues of diabetic rat models. In another similar study, Liu *et al.* [10] identified the DNBs of type-1 diabetes mellitus (T1DM) using time-variant gene expression profiles obtained from pancreatic lymph nodes of non-obese diabetic mouse samples.

Based on the core DNB principle, Yu *et al.* [12] proposed the concept of edge-networks. Compared to ordinary networks, each node in an edge network contains a pair of molecules and each edge connects two pairs of molecules. The authors captured the stochastic dynamics of biological systems using the second-order statistical information of a dynamical system by assuming a Gaussian distribution over the expression level of each molecule.

In another study by Zeng *et al.* [13], the dynamical organization of molecular modules at the early disease stage was investigated. They implemented this method to identify early-stage dynamic biomarker networks using gene expression profiles of T1DM on mouse model. To do so, first the tissue-specific and time-specific networks were constructed. Second, each network was divided into several sub-networks using the Markov Clustering Algorithm (MCL) [14]. The pre-disease biomarkers were obtained by a module yielding the largest dynamic index indicated above, and the time when this large index value had occurred was taken as the pre-disease stage.

One of the most recent works in this area was proposed in [15], where DNB discovery was characterized as a model-based framework. The overall pipeline of their method starts by construction of a network through integration of protein interactions with gene expression data followed by an ODE-based dynamic optimization model to infer the dynamic network of interest. The clustering algorithm called ClusterONE [16] was used for identification of network modules, where "High-influence" modules were extracted based on their average degrees across the whole network. One of the recent works in this area is proposed by Vafaee [1]. There have been several newly published papers in this are that can be found in [32] to [37].

The major issue with the mentioned studies lies in the fact that they consider all of the pairwise relationships between the genes. This has several disadvantages including: (1) increased computational cost as a result of dealing with complete graphs as genomic networks; (2) many of the genes which are computationally connected might not have physical interactions, leading to unreasonable biological interpretations. To tackle these issues, we propose two optimization-based algorithms to detect the pre-disease early-warning genetic network sub-modules. In the first algorithm, the complete graph of the genetic interactome is considered, where the dynamic network biomarkers are detected based on a meta-heuristic search algorithm. The second algorithm first constructs the sparse adjacency graph of the ineractome using latent biological knowledge in the context of signaling and metabolic pathways, and then uses meta-heuristic search to detect the underlying dynamic network biomarker modules. Overall, the main contributions of this paper can be summarized as follows:

- Develop an optimized meta-heuristic search algorithm to extract the sub-modules of the genetic interactome at each time step using time-course gene expression data.
- Develop a pathway-based graph construction framework to prepare a sparse and realistic genetic interactome and then extract the early-warning signal sub-networks using the proposed search algorithm.

The remainder of this paper is organized as follows: in Section 2, the mathematical notations and the proposed algorithms are introduced. The time-course data is described in Section 3, and the numerical experiments are demonstrated in Section 4. Concluding remarks are provided in Section 5.

2 METHODOLOGY

According to [9], the DNB discovery task can be accomplished by an index which is composed of three measurements.These measurements include: (1) The pairwise correlation between the genes in the sub-network; (2) The pairwise correlation between the members of the dynamic network biomarker and the other genes; (3) The standard deviation of the expression levels of the member of the sub-network of interest. Consider a time-course high-throughput data set containing M samples. Let $\mathbf{X} = \mathbf{x}_i^t, i \in 1, ..., n$ be the set of the gene samples at $t = 1, ..., T$ time-points. The set of molecular samples at time point t is denoted by $\mathbf{x}_i^t = \{x_{i,1}^t, ..., x_{i,M}^t\}$. Let $\mathbf{X}_c \in \mathbf{X}$ be the set of dynamic network biomarkers of interest. Then the following indices are known as the constituent elements of the DNBs [1]:

1) Members of $\mathbf{X}_c$ show a drastic increase in intra-cluster correlations P_c^t at the pre-disease stage. Such an increase can be modeled as the average of pairwise Pearson correlations between the members of $\mathbf{X}_c$:

$$P_c^t = \frac{1}{|\mathbf{X}_c|.(|\mathbf{X}_c| - 1)} \sum_{\mathbf{x}_i^t, \mathbf{x}_j^t \in \mathbf{X}_c} \rho_{\mathbf{x}_i^t, \mathbf{x}_j^t} \quad (1)$$

2) The inter-cluster correlations $\tilde{P}_c^t$ of the dynamic network biomarker members and the genes outside the network decrease drastically:

$$\tilde{P}_c^t = \frac{1}{|\mathbf{X}_c|.(|\mathbf{X}| - |\mathbf{X}_c|)} \sum_{\mathbf{x}_i^t \in \mathbf{X}_c, \mathbf{x}_j^t \notin \mathbf{X}_c} \rho_{\mathbf{x}_i^t, \mathbf{x}_j^t} \quad (2)$$

3) The standard deviation s_c^t of expression intensities of the dynamic biomarker network at the pre-disease stage increases drastically:

$$s_c^t = \frac{1}{|\mathbf{X}_c|} \sum_{\mathbf{x}_i^t \in \mathbf{X}_c} \sigma_{\mathbf{x}_i^t} \tag{3}$$

Based on these properties, a composite index comprised of the abovementioned indices is used at each time step:

$$I_c^t = \frac{s_c^t . P_c^t}{\tilde{P}_c^t} \tag{4}$$

As can be observed, the composite index considers the entire interactome, implying that the co-expression network is a complete graph. This does not make biological sense because many of the connected genes do not have any biological interaction. Therefore, we will refine this approach by sparsing the co-expression network using the biological pathway knowledge.

Extracting a sub-network of such a large co-expression network can be framed as an optimization problem. The large size of these biological networks necessitates employing meta-heuristic approaches which are able to reach the optimal solution in a timely manner. In the next section we will present the structure of the proposed search algorithm and then couple it with the gene co-expression network construction pipeline.

2.1 Meta-heuristic Search Algorithm

The primary goal of this research is to pose an optimization problem. To do so, we need to search the pairwise interactions between the genes. Due to the size of the co-expression network in our problem, where the number of nodes is roughly in order of thousands, the complete enumeration methods tend to be computationally troublesome. To avoid such computational burdens, we propose a Simulated Annealing (SA) based search algorithm to find the DNBs yielding the largest composite index value at each time step. In the following, we will discuss each step of the search algorithm in detail.

The search process begins with setting the maximum and minimum annealing temperature, number of iterations at each temperature, and generating a random configuration as the initial solution. During the search process, the current solution is perturbed and new solutions are generated. If the generated solution in the vicinity of the current solution produces a better objective function, then it will be considered as the best solution at its respective search iteration. The solution representation procedure and perturbation mechanisms will be discussed in the following subsections.

2.1.1 Solution Representation

During each iteration of the search process, the current solution of the members of the sub-network is defined as a row vector containing n genes, where n denotes the total number of genes in the gene expression data. This vector contains binary values of 1 and 0, indicating presence or absence of the corresponding gene in the sub-network, respectively.

2.1.2 Perturbation Mechanisms

In order to search the feature space to achieve better solutions, in each iteration of the algorithm the current solution should be perturbed to avoid falling into local optima. We introduce four perturbation mechanisms, three of which apply local search while the fourth implements a major global perturbation. Note that these perturbation mechanisms are randomly selected based on a uniform distribution on the interval [0,1]. These mechanisms are described as follows:

1) Let *S* be the current solution. Randomly select an integer r from 1,..., n-1. Swap the corresponding bits as $NewS[r] = S[r+1]$ and $NewS[r+1] = S[r]$.
2) Let *S* be the current solution. Randomly select an integer R from 1,..., n. Then, $\forall\, S[r] | r \leq R$, if $S[r] = 0$ then $NewS[r] = 1$ and vice versa.
3) Let *S* be the current solution. Randomly generate two unique integers a and b from 1,..., n. Swap the corresponding bits of the generated integers to obtain the new solution as $NewS[a] = S[b]$ and $NewS[b] = S[a]$.
4) Let *S* be the current solution. Compute the number of genes in the current solution by counting the number of 1s in the solution vector. If the number of genes in the module is more than half of the total number of genes, then randomly remove one of the current genes in the solution to keep the network sub-module less complex.

Note that the perturbation mechanisms above are proposed by the authors and other similar perturbation methods can be found in the literature.

2.1.3 Annealing Process

The search procedure starts from an initial Temperature (T_{in}). The maximum temperature is set to T_{max} and the minimum temperature is set to T_{min}. At each temperature the search procedure is repeated $Iter$ times. During each iteration, one of the abovementioned perturbation mechanisms is randomly applied with a uniform distribution. The temperature at iteration t is obtained by

$$T_t = T_{max}(1-\alpha)^t \tag{5}$$

where α is a search parameter between 0 and 1. According to our experiments, $\alpha = 0.2$ yields the most stable objective function values. The overall pipeline of the search method is presented in Algorithm 1. It should be mentioned that at each iteration the newly generated configuration whose objective function value, if worse than the current configuration, can also be accepted as the configuration of choice with probability level which gradually decreases as the search process moves forward. This is called the acceptance probability and is mentioned in Algorithm 1. The algorithm parameters have been set based on recommendations made in [31]. Hence, $T_{max} = 500, T_{min} = 0.1, Iter = 400$.

2.2 Dynamic Biomarker Discovery

We have considered two methods of graph construction. In the first method, all of the pairwise connections between the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2687925, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Algorithm 1 SA-based search pseudo code.

INPUT: T_{max}, T_{min}, T, Iterations, Initial Configuration ($Curr$).
OUTPUT: Best Configuration.
Set $t = 1$
Compute objective function of $Curr$ (E_{Curr})
while $T_t > T_{min}$ **do**
  for $i = 1$ to Iterations **do**
    Create new configuration ($Config_{new}$) by perturbation mechanisms.
    Compute objective function of $Config_{new}$ (E_{New})
    if $E_{New} \geq E_{Curr}$ **then**
      Set: $Curr = Config_{new}$ and $E_{Curr} = E_{New}$
    else
      $Prob_{acceptance} = exp((E_{New} - E_{Curr})/T_t)$, **Set:** $E_{Curr} = E_{New}$
    end if
  end for
  Set: $t = t + 1$, $T_t = T_{max}(1 - \alpha)^t$
end while
Return: $Curr$

entire genes are taken into account. This can be translated to having an adjacency matrix similar to:

$$\mathbf{A} = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}_{n \times n} \quad (6)$$

In this situation, genes having no biological relations might appear in the extracted dynamic network. Such connections might drastically increase the defined objective function, but will limit further biological meta-analysis. To take into account this limitation, we refine the adjacency matrix using the biological knowledge, in the form of biological pathways. In this way, we can ensure that the resulting sub-network biomarkers are biologically meaningful, while preserving the condition of having the largest composite index level. By inducing the pathway knowledge in refining the adjacency matrix of biomarkers, only genes with direct biological relations which appear in at least one pathway will be used during the process of graph construction. The pathway induction process is illustrated in Fig. 1. The overall pipeline of the proposed pathway-induced sub-module dynamic network biomarker discovery is represented in Algorithm 2. According to Algorithm 2, the adjacency matrix (A) of a given set of genes initially represents a complete graph whose elements are entirely equal to 1. At the beginning of this algorithm, the pathways which are stored in the matrix $Path$ are checked. For all of the elements of the adjacency matrix, if two genes a and b cannot be found in at least one of the pathways, then their corresponding value in the matrix A will be changed to 0. The pathway matrix $Path$ is a 1077×933 matrix whose rows represent different pathways. Each row contains the gene Entrez IDs of the genes in its corresponding pathway, hence the lengths of rows are different. The longest pathway contains 933 genes, which is far larger than most of the other pathways. We have put 0 as the elements of the shorter pathways. This is done only for facilitating the computations by making all the pathways having the same length.

When constructing a new graph configuration, only biologically related genes are allowed to be connected. As a result, the constructed graphs will no longer be complete. We have tested the search algorithm in both situations where the co-expression graph is complete or sparse (based on pathway knowledge).

Algorithm 2 Pathway-induced dynamic biomarker discovery

INPUT: T_{max}, T_{min}, T, Iterations, Initial Configuration ($Curr$), Pathway matrix ($Path$).
OUTPUT: Best Configuration.
Set adjacency matrix A to the $n \times n$ matrix with all the elements equal to 1.
for each pair of genes (x, y) in $Curr$ **do**
  if $(x, y) \notin Path$ **then**
    Set: $A(x, y) = 0$
  end if
end for
Set $t = 1$
Compute the objective function of $Curr$ (E_{Curr}) considering pathway-induced A.
while $T_t > T_{min}$ **do**
  for $i = 1$ to Iterations **do**
    Create new configuration ($Config_{new}$) by perturbation mechanisms.
    Compute the objective function of $Config_{new}$ (E_{New}) considering pathway-induced A.
    if $E_{New} \geq E_{Curr}$ **then**
      Set: $Curr = Config_{new}$ and $E_{Curr} = E_{New}$
    else
      $Prob_{acceptance} = exp((E_{New} - E_{Curr})/T_t)$, **Set:** $E_{Curr} = E_{New}$
    end if
  end for
  Set: $t = t + 1$, $T_t = T_{max}(1 - \alpha)^t$
end while
Return: $Curr$

2.3 Computational Complexity

Suppose I to be the number of iterations at each temperature, T to be total number of temperatures during the annealing process, and K to be the total number of nodes in the DNB. Since the number of temperature levels decreases in a logarithmic fashion, then the worst case scenario for the largest number of temperatures will be T. Objective function computation at each iteration can be done in $O(K)$ time. Therefore, whole search process without considering biological knowledge is done in $O(ITK)$ time. If the are P pathways, then each pair of genes should be checked for biological relations. Hence, it will take $O(PT^2)$ to check for biological relevance of the initial adjacency matrix before starting the search process. As a result, the entire search process considering biological knowledge will be performed in $O(ITK + PT^2)$ time.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2687925, IEEE/ACM Transactions on Computational Biology and Bioinformatics

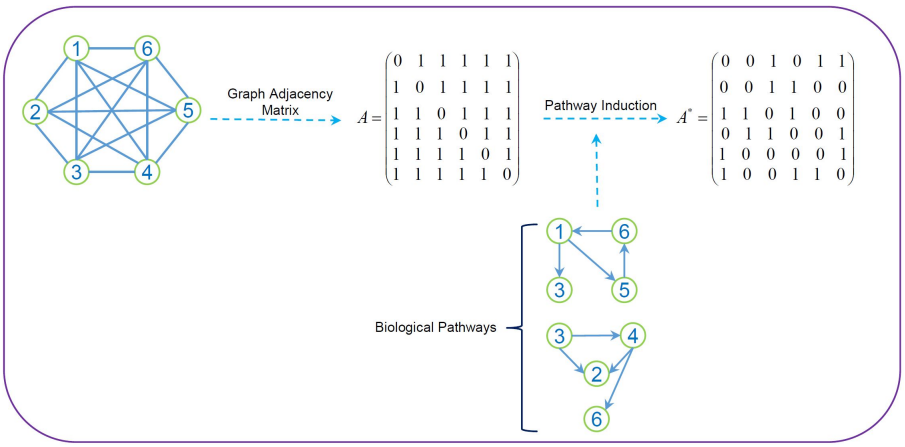

Fig. 1. The pathway induction process on a complete gene co-expression graph.

3 Data

Due to the high temporal resolution of a microarray gene expression experiment by Sciuto *et al.* [17], we have chosen the time-course gene expression data of lung injury with carbonyl chloride inhalation exposure. This data belongs to an experiment for investigating the molecular mechanisms of phosgene-induced lung injury. In order to obtain the control and case cohorts, two groups of male mice were exposed to air and phosgene, respectively. Then, long tissues were collected from both cohorts at the following time points after the exposure: 0.5, 1, 4, 8, 12, 24, 48, and 72h. This dataset is publicly available at the NCBI GEO repository [18] with the accession number GSE2565.

This data contains 22690 probe sets corresponding to 13792 unique genes. Several pre-processing stages were taken. Since many of these genes are not differentially expressed in the case and control cohorts, a two-sided statistical t-test along with false discovery rate (FDR) estimation was carried out at each time step, and genes with q-values larger than 0.01 were removed. Finally, case samples were row-wise z-normalized.

The biological pathways were downloaded from the Molecular Signature Data Base (MSigDB v5.1) [19] which includes 13311 genes in total. Three different sources of biological pathways were used including: KEGG [20], Reactome [21], and BioCarta [22], each containing 186, 674, and 217 biological pathways, respectively.

4 Numerical Experiments

Using the data mentioned in the previous section, we have applied both of the proposed algorithms on this data to investigate the pre-disease stage, along with extracting their driving genes. According to Table 1, the pre-disease stage for Lung Acute Injury (ALI) occurs at 8hr. This is in exact accordance with the other benchmark methods mentioned here.

Compared to Algorithm 1, whose constructed graph is complete, the sparsity of the constructed graph by Algorithm 2 is high due to the proposed pathway induction method. According to Table 1, seven different methods have interrogated the pre-disease stage detection using dynamic network biomarkers. Six methods introduce 8 hr as the pre-disease stage containing diverse number of biomarkers. Previous studies have asserted that large sets of biomarkers usually contain a large fraction of redundant genes having no significant biological relation with the disease of interest [23]. Conversely, small biomarker sets do not provide deep insights into the underpinnings of complex diseases. This is a strength point of the proposed algorithms, which extract biomarkers whose size are reasonably large enough while being in compliance with the other methods in terms of the detected pre-disease stage. The heatmap of the derived DNB is depicted in Figure 2. Here we have separated this sub network by two yellow rectangles. The intersection of these boxes represents the extracted DNBs. Figure 2(a) denotes the constituent genes of the extracted DNB at the pre-disease stage (8hr). The intersection of both yellow boxes in Figure 2(a) demontrates only the DNB of interest, which consists of six samples at 8hr. Figure 2(b) illustrates how correlated the genes in this module are, which goes back to the basic definition of DNBs regarding extreme correlation among the driver biomarker genes in the extracted sub-graph. Additionally, we obtained the Pearson correlation coefficient between the expression levels of the identified dynamic network biomarkers and the phenotype of interest. This is done for all of the methods at their respective identified pre-disease stage. It can be observed that the Algorithm 2 represents the largest association value while Method 4 ranks second. Only these two methods represent association coefficient over 0.7 while the rest of the algorithms lie below 0.7. Nevertheless, Algorithm 2 outperforms Method 4 significantly in terms of the DNB size (55 vs 85) and algorithm running time (210 seconds vs 450 seconds), and

TABLE 1
Comparison of DNBs of ALI predicted by different methodologies

Method	Pre-disease Time Point	DNB size	Running Time (S)	Association Coefficient
Method 1 [9]	8hr	220	400	0.662
Method 2 [12]	8hr	84	550	0.626
Method 3 [13]	8hr	73	390	0.699
Method 4 [11]	8hr	85	450	0.722
Method 5 [1]	4hr	16	650	0.685
Algorithm 1	8hr	144	180	0.697
Algorithm 2	8hr	55	210	0.741

TABLE 2
Enriched pathways by the extracted DNBs

Pathway Titles	P-Value
Protein processing in endoplasmic reticulum	1.2×10^{-8}
MAPK signaling pathway	2.83×10^{-8}
Pathways in cancer	1.46×10^{-7}
Osteoclast differentiation	5.61×10^{-7}
Glutathione metabolism	1.06×10^{-5}
Metabolic pathways	9.19×10^{-6}
Cytokine-cytokine receptor interaction	3.71×10^{-5}
p53 signaling pathway	2.97×10^{-5}
ErbB signaling pathway	6.98×10^{-5}
Antigen processing and presentation	10^{-4}

biological relevance.

Since DNBs are the leading networks representing critical transition of diseases into the catastrophic stage, they are linked to genes involved in pathogenesis. Here we have conducted two sets of meta-analysis experiments in order to investigate functional behavior of DNBs through pathway enrichment analysis and Gene Ontology (GO) analysis.

After performing the pathway enrichment analysis, we selected the top 10 enriched pathways using the DNBs obtained by Algorithm 2. These pathways are listed in Table 2.

According to Laeson *et al.* [24], the Protein Processing in the Endoplasmic Reticulum pathway can contribute to the pathogenesis of pulmonary fibrosis. Not only our results but also previous studies have demonstrated the expression of MAPK pathways in lung tissue of patients with idiopathic pulmonary fibrosis (IPF). For example Antoniou *et al.* [25] have verified the hypothesis of involvement of MAPK signaling pathways in pathogenesis of IPF. Finally, the regulation of DNBs provide suitable conditions for genetic mutations as a main cause of cancer occurrence according to the cancer-related pathways. The enriched pathways verify the meaningfulness of the detected DNBs.

In order to gain a better insight into the objective function trajectories across different time steps, we have provided the objective function level in each separate time steps across the entire algorithm iterations (Fig. 3). In total, based on the set parameters, 15600 iterations are carried out at each simulation on each time step.

During each iteration, the perturbation mechanisms have been chosen randomly by a uniform probability distribution. According to Fig. 4, and as previously mentioned, 8 hr is detected as the critical time step of interest. The objective function of the algorithm reaches its highest level in iteration 12000 and then stabilizes and does not improve anymore. Similar behavior for the other time steps at different iterations can be observed where after a certain iteration level their objective function stabilizes. According to Fig. 4, we have represented the frequency of the chosen perturbation mechanisms for each time step during the simulations run. As mentioned before, each perturbation mechanism is chosen randomly based on a uniform distribution so the chance of getting each perturbation mechanism at each iteration is equal to 0.25. As the critical time step captured by the algorithm, hr 8, mechanism 3 shows the highest frequency (almost 3975) and mechanism 4 ranks second. Mechanism 2 has the least frequency level though he difference between the highest and lowest frequencies is almost 140. By this observation, we can conclude that in hr 8, mechanism 3 is more impactful than the other mechanisms. Nevertheless, in the other time steps a different behavior can be observed. For example, in 48 hr, frequency of mechanism 4 is by far larger than the other three while its respective optimal objective function level is among the lowest. Hr 8 and hr 24 yield the first two highest final objective function. By looking at their perturbation mechanism frequency, we will notice that mechanism 2 and 3 show a higher frequency than the others. Generally speaking, none of these perturbation mechanisms can be specified as the main factor in gaining a higher objective function value since there is no obvious relationship between the frequencies and the optimal objective function at each time step.

4.1 ALI associated cellular mechanisms

In this section, we will investigate possible roles of the obtained DNBs in the pathogenesis of ALI. To do so, the DNBs were functionally profiled using Gene Ontology (GO) [26]. The goal of the GO project is to hierarchically attribute genes to terms organized as graphs. Basically, these terms fall into three groups including: biological processes, cellular components and molecular functions. We have employed the BiNGO tool [27] to determine the GO terms that are statistically overrepresented by the DNBs. BiNGO maps the functional specification of the queried gene sets on the GO hierarchy and outputs this mapping as a graph. In our use of BiNGO, the parameters were set as follows: the hypergeometric test is used for enrichment where the p-values were FDR corrected; the significance level was set to 0.01 and the organism of interest was set to Mus Musculus. The biological processes being enriched by DNBs are shown in Figure 5.

It can be observed that GO yields a holistic view of the biological processes being enriched by the obtained DNBs. In this figure, different regions pertaining to certain functionalities are annotated by a unique color, where nodes are the GO terms and edges represent the relationships between them. Note that the size of each node indicates the number of DNBs annotated to its corresponding GO category. Many of the GO terms represented in Figure 5 have direct relationships with the initiation and progression of ALI. Some of these terms are as follows: response to toxic substances, response to oxygen levels, positive regulation of cell death, response to temperature stimulus, cellular response to external stimulus, regulation of inflammatory response, response to ketone, and positive regulation of extrinsic apoptotic signaling pathway. To support these extracted terms, we have looked at the literature to find clinical evidence verifying these results. According to [28], [29], dysregulation of apoptosis plays a vital role in the initiation of ALI and other related disorders. Another evidence verifies that the gained DNBs are in direct relationship with ALI regarding the regulation of the inflammatory response, since ALI is the widespread manifestation of inflammatory responses of the lung to direct insults [30].

Overall, the medical evidence represented here support the obtained GO terms enriched by the output DNB of the proposed algorithms. The list of the genes detected in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2687925, IEEE/ACM Transactions on Computational Biology and Bioinformatics

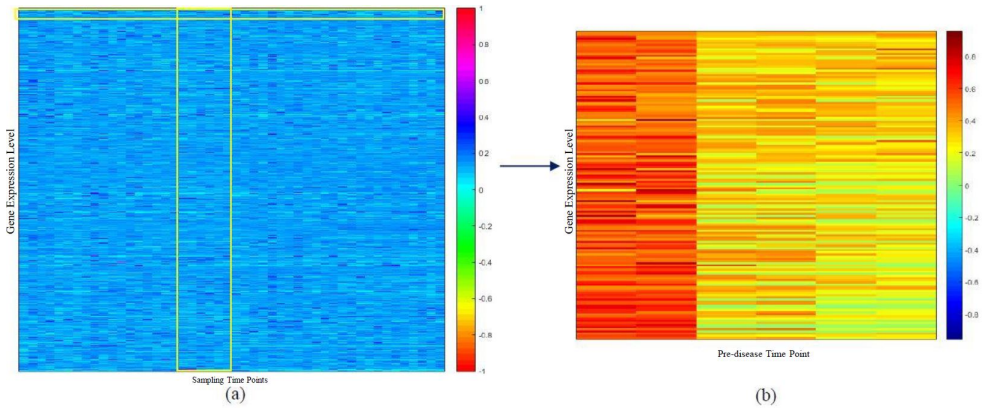

Fig. 2. Heatmap representation of the extracted DNBs.

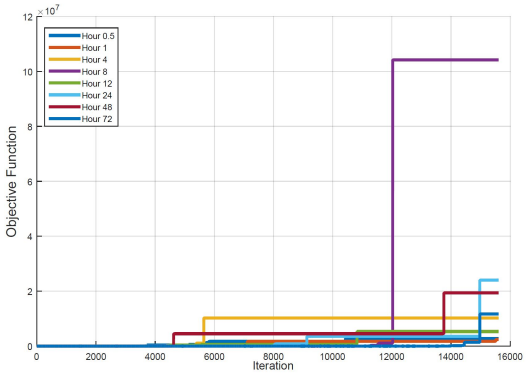

Fig. 3. Objective function trajectory on different time steps across the entire iterations of the algorithm.

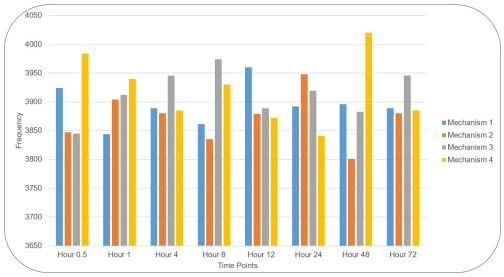

Fig. 4. Frequency of the perturbation mechanisms in entire iterations.

DNB network by our algorithm and other algorithms are represented in the supplementary file.

5 Conclusions

This paper deals with the dynamic network biomarker discovery problem using high-throughput gene expression data collected across different time steps. We have proposed two algorithms using Simulated Annealing meta-heuristic search. In the first algorithm, it is assumed that the dynamical network module is a complete graph and all genes are in mutual interaction. Although this has been the main pillar of similar studies, it might not appropriately handle real world problems where all genes are not necessarily related in terms of biology. Accordingly, the second proposed algorithm refines the adjacency graph between the genes using the biological pathway knowledge, such that only genes having direct biological links can be connected in the interactome graph. Both algorithms were tested on a real microarray study on ALI, and the resulting biomarkers were tested using pathway enrichment analysis and GO terms. Computational results were verified using the clinical evidence from the literature. Compared to the existing methods, our methods provide well-sized dynamcial networks representing the pre-disease stage. We expect that our methodology can be extended to other complex diseases.

6 Acknowledgments

The authors gratefully acknowledge the financial support by the Institute for Collaborative Biotechnologies through Coagulopathy grant W911NF-1O-2-0114 from the U.S. Army Research Office. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] F. Vafaee, *Using multi-objective optimization to identify dynamical network biomarkers as early-warning signals of complex diseases*, Sci. Rep., vol. 6, no. 22023, pp. 1-11, 2016.
[2] R.L. Barter, S.J. Schramm, G.J. Mann, Y.H. Yang, *Network-based biomarkers enhance classical approaches to prognostic gene signatures*, BMC Syst. Bio., vol. 8, no. 55, pp. 1-16, 2014.
[3] T. Ideker and N. J. Krogan, *Differential network biology*, Mol. Syst. Bio., vol. 8, no. 565, pp. 1-9, 2012.
[4] X. Wu, L. Chen, X. Wang, *Network biomarkers, interaction networks and dynamical network biomarkers in respiratory diseases*, Clin. Trans. Med., vol. 3, no. 16, pp. 1-7, 2014.
[5] J. M. Drake, B. D. Griffen, *Early warning signals of extinction in deteriorating environments*, Nature, vol. 467, pp. 456-459, 2010.
[6] J. M. Drake, B. D. Griffen, *Tipping elements in the earth's climate change*, Proc. Natl. Acad. Sci., vol. 105, pp. 14308-14312, 2008.
[7] C. Vogel, E.M. Marcotte, *Insights into the regulation of protein abundance from proteomic and transcriptomic analyses*, Nat. Rev. Genet., vol. 13, pp. 227-232, 2012.
[8] C. Trefois, P.M. Antony, J. Goncalves, A. Skupin, R. Balling, *Critical transitions in chronic disease: transferring concepts from ecology to systems medicine*, Curr. Opin. Biotech., vol. 34, pp. 48-55, 2015.
[9] L. Chen, R. Liu, Z.P. Liu, M. Li, K. Aihara, *Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers*, Sci. Rep., vol. 2, no. 342, pp. 1-8, 2012.
[10] X. Liu, R. Liu, X. M. Zhao, L. Chen, *Detecting early-warning signals of type-1 diabetes and its leading biomolecular networks by dynamical network biomarkers*, BMC Med. Gen., vol. 6, no. s8, pp. 1-8, 2013.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2687925, IEEE/ACM Transactions on Computational Biology and Bioinformatics

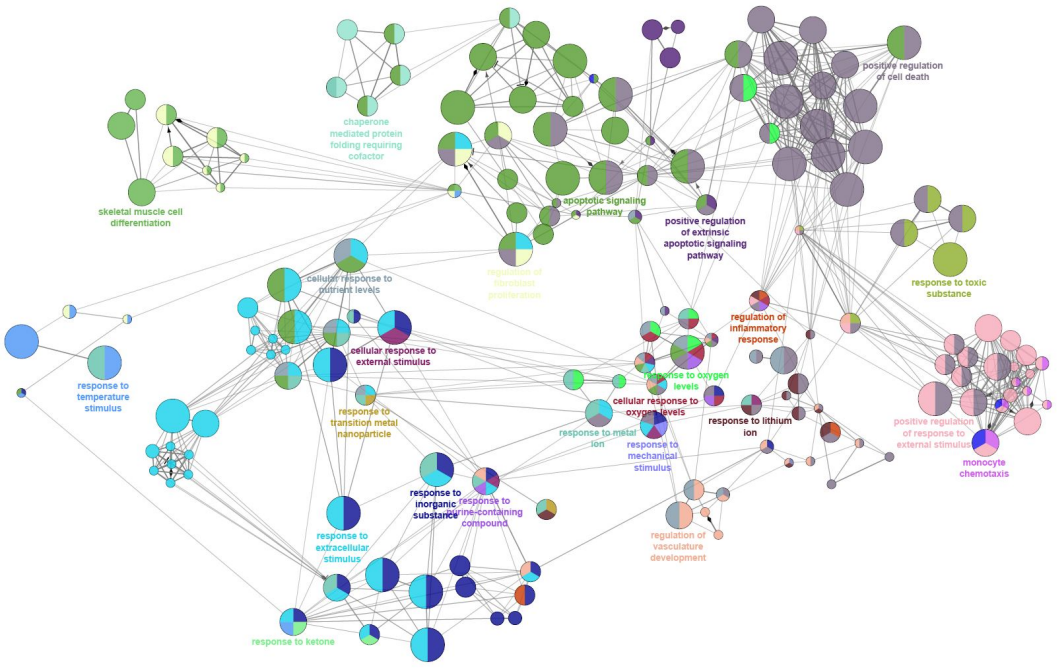

Fig. 5. The biological processes hierarchy enriched by the DNBs.

[11] M. Li, T. Zeng, R. Liu, L. Chen, *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., vol. 15, pp. 229-243, 2014.

[12] X. Yu, G. Li, L. Chen, *Prediction and early diagnosis of complex diseases by edge-network*, Bioinformatics, vol. 30, pp. 852-859, 2014.

[13] T. Zeng et al., *Deciphering early development of complex diseases by progressive module network*, Methods, vol. 67, pp. 334-343, 2014.

[14] A. J. Enright, S. Van Dongen, C. A. Ouzounis, *An efficient algorithm for large scale detection of protein families*, Nucleic Acid Res., vol. 30, pp. 1575-1584, 2002.

[15] Y. Li, S. Jin, L. Lei, Z. Pan, X. Zou, *Deciphering deterioration mechanisms of complex diseases based on the construction of dynamic networks and systems analysis*, Sci. Rep., vol. 5, 2015.

[16] T. Nepusz, H. Yu, A. Paccanaro, *Detecting protein protein complexes in protein-protein interaction networks*, Nature Methods, vol. 9, pp. 471-472, 2012.

[17] A. M. Sciuto et al., *Genomic analysis of murine pulmonary tissue following carbonyl chloride inhalation*, Chem. Res. Toxicol., vol. 18, pp. 1654-1660, 2005.

[18] R. Edgar, M. Domrachev, A. E. Lash, *Gene expression omnibus: NCBI gene expression and hybridization array data repository*, Nucleic Acid Res., vol. 30, pp. 207-210, 2002.

[19] A. Subramanian et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*, Proc. Nat. Acad. Sci., vol. 102, no. 43, pp. 15545-15550, 2005.

[20] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, *KEGG as a reference resource for gene and protein annotation*, Nucleic Acid Res., vol. 44, pp. 457-462, 2016.

[21] A. Fabreqat et al., *The Reactome pathway Knowledgebase*, Nucleic Acid Res., vol. 44, pp. 481-487, 2016.

[22] A. Nishimura, *A view from the web BioCarta*, Biotech. Soft. Int., vol. 2, no. 3, pp. 117-120, 2001.

[23] J. West, S. Beck, X. Wang, A. E. Teschendorf, *An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem cell differentiation pathways*, Nature Sci. Reps., vol. 3, no. 9283, 2013.

[24] W.E. Lawson et al., *Endoplasmic reticulum stress in alveolar epithelial cells is prominent in IPF: association with altered surfactant protein processing and herpesvirus infection*, Am. J. Physiol. Lung Cell Mol. Physiol., vol. 294, no. 6, pp. 1119-1126, 2008.

[25] K.M. Antoniou et al., *Expression analysis of Akt and MAPK signaling pathways in lung tissue of patients with idiopathic pulmonary fibrosis (IPF)*, J. Recept. Signal. Transduct. Res., vol. 30, no. 4, pp. 262-269, 2010.

[26] M. Ashburneret al., *Gene ontology: tool for the unification of biology*, Nat. Gene., vol. 25, pp. 25-29, 2000.

[27] S. Maere, K. Heymans, M. Kuiper., *BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks*, Bioinformatics, vol. 21, pp. 3448-3449, 2005.

[28] T.R. Martin, M. Nakamura, G. Matute-Bello, *The role of apoptosis in acute lung injury*, Crit. Med. Care., vol. 31, pp. s184-s188, 2003.

[29] M. Chopra, J.S. Reuben, A.C. Sharma, *Acute lung injury: apoptosis and signaling mechanisms*, Exp. Biol. Med., vol. 234, pp. 361-371, 2009.

[30] M. Ragaller, T. Richter, *Acute lung injury and acute respiratory distress syndrome*, J. Emerg. Trauma Shock, vol. 3, no. 1, pp. 43-51, 2010.

[31] A. Doostparast Torshizi, M.H. Fazel Zarandi, *Alpha-plane based automatic general type-2 fuzzy clustering based on simulated annealing meta-heuristic algorithm for analyzing gene expression data*, Com. Bio. Med. vol. 64, pp. 347-359, 2015.

[32] R. Liu, M. Li, Z.P. Liu, J. Wu, L. Chen, K. Aihara, *Identifying critical transitions and their leading biomolecular networks in complex diseases*, Sci. Rep., vol. 2, no. 813, 2012.

[33] R. Liu, K. Aihara, L. Chen, *Dynamical network biomarkers for identifying critical transitions and their driving networks of biologic processes*, Quant. Bio., vol. 1, no. 2, pp. 105-114, 2013.

[34] R. Liu, X. Wang, K. Aihara, L. Chen, *Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers*, Med. Res. Revs., vol. 34, no. 3, pp. 455-478, 2014.

[35] R. Liu, X. Yu, X. Liu, D. Xu, K. Aihara, L. Chen, *Identifying critical transitions of complex diseases based on a single sample*, Bioinf., vol. 30, no. 11, pp. 1579-1586, 2014.

[36] R. Liu, P. Chen, K. Aihara, L. Chen, *Identifying early-warning signals of critical transitions with strong noise by dynamical network markers*, Sci. Rep., vol. 5, no. 17501, 2015.

[37] P. Chen, R. Liu, Y. Li, L. Chen, *Detecting critical state before phase transition of complex biological systems by hidden Markov model*, Bioinf., vol. 32, no. 14, pp. 2143-2150, 2016.