Results: focusing only on pairwise interactions to the more general cascading thus needed to fully utilize gene sets and broaden the scope from data do not explicitly consider signal cascading mechanisms that sufficiently exploited. Existing methods accommodating discrete measurements emitted from latent signaling pathways. Their discovery of numerous gene sets, which can be interpreted as discrete measurements hold promises for inferring biological networks. In recent years, gene set combinands and tools for their analysis have become increasingly available due to rapid advancements in high-throughput data acquisition methods (e.g. Subramanian et al. 2005; Medina et al., 2006; Segal et al., 2004; Park et al., 2010). However, challenges remain in exploring signal cascading mechanisms from such data, which can be interpreted as discrete measurements emitted from latent signaling pathway structures.

Many algorithms for biological network inference accommodate discrete inputs (e.g. Altay and Emmert-Streib, 2010) Discretization has especially proved useful in the structural inference of signaling pathways, which are directed networks containing up to a few hundred nodes and several overlapping signal cascades where each cascade represents a directed or ordered chain of molecular interactions. For example, existing non-metabolic pathway structures in the KEGG database (Kanehisa et al. 2010) contain up to 400 nodes. Significant efforts in the inference of signaling pathway structures include Boolean or Probabilistic Boolean networks (e.g. Shmulevich et al. 2002, Kaderali et al. 2009, and Bayesian networks (e.g. Friedman et al. 2000, Segal et al., 2004), which directly benefit from reduced computational complexity by utilizing discrete inputs. Even in the inference of large-scale undirected network topologies using ARACNE (Marogolin et al. 2006), CINNET (Altay and Emmert-Streib, 2010), CLR (Faith et al. 2007), MRNET (Meyer et al. 2004) and Relevance Networks or RNs (Butte and Kohani, 2000), discrete measurements are employed to estimate mutual information (MI) between gene pairs. Therefore, it is increasingly clear that discrete measurements hold promises for inferring biological networks.

Bayesian network methods are commonly used in the inference of signaling pathway structures. However, these methods primarily focus on statistical causal interactions. Thus, the learned networks need not represent signal cascading mechanisms. How to better use discrete measurements available in the form of unordered gene sets, which may be thought of as the observed overlapping and incomplete
Reverse engineering the optimal signaling pathway structures from gene sets

Signal cascading events, remains an open area of research. A few attempts made toward the inference of communication networks from co-occurrence data find applications in biomedical field (e.g. Rabbat et al., 2008), but significant advantages of inferring signaling pathway structures from gene sets are yet to be demonstrated.

We attempt to overcome the issues raised above by presenting a novel computational approach for inferring the optimal signaling pathway structure from partially observed and overlapping gene sets related to a pathway. Identification of pathways from molecular profiling data is a relatively well-studied problem and has been explored in the literature (Xu et al., 2010). However, issues still remain in reconstructing signal cascading mechanisms in the pathways of interest. In our study, we specifically focus on this problem. Our motivation stems from considering a signaling pathway structure as an ensemble of overlapping and linear signaling cascades, which we refer to as information flows (IFs). In other words, the true signaling pathway structure can be constructed by assembling the IFs into a single unit. As a gene may simultaneously participate in multiple IFs, the extent of overlap among IFs is an integral part of the construction. The set of all genes in an IF, with no information about the order in which they appear in the IF, is called an information flow gene set (IFGS) (Acharya et al., 2011).

We observe partial or complete IFGSs but not the order in which their component genes appear in the corresponding IFs. We propose to explore the overlapping information among IFGSs in order to infer underlying IFs, which in turn define the signaling pathway structure.

As there exist $L^l$ different gene orderings for an IFGS with $L$ component genes, a total of $L^{2L}$ signaling pathway structures can be constructed by combining $m$ such IFGSs. An exhaustive search for the true structure among $L^{2L}$ candidate structures may be computationally intractable, even when the values of $m$ and $L$ are controlled. To address this issue, we translate our goal of signaling pathway structure inference from IFGSs into a discrete optimization problem. We then propose a simulated annealing (SA) algorithm to locate the optimal signaling pathway structure. SA (Kirkpatrick et al., 1983) is a well-known search algorithm for solving global optimization problems. SA finds its root in the field of metallurgy, where a metal is heated and then cooled down slowly so that the atoms gradually configure themselves in states of lower internal energy, refining the crystalline structure of the metal. Compared with other global search algorithms such as genetic algorithm (Holland, 1992) and tabu search (Glover, 1989), SA is easier to understand and to implement without sacrificing performance. Since genetic algorithm is a population-based search method and tabu search is a memory-based heuristic, each iteration of SA runs faster than the two approaches. SA also requires a small number of user-specified parameters. In the past, SA has inspired various bioinformatics researches (e.g. Baker et al., 2009; Gonzalez et al., 2006; Chen et al., 2010).

We develop a new gene set-based SA to infer signaling cascades that characterize the optimal signaling pathway structure. Throughout we treat IFGSs as variables and their orders as random.

We observe partial or complete IFGSs but not the order in which their component genes appear in the corresponding IFs. We propose to explore the overlapping information among IFGSs in order to infer underlying IFs, which in turn define the signaling pathway structure.

To address this issue, we translate our goal of signaling pathway structure inference from IFGSs into a discrete optimization problem. We then propose a simulated annealing (SA) algorithm to locate the optimal signaling pathway structure. SA (Kirkpatrick et al., 1983) is a well-known search algorithm for solving global optimization problems. SA finds its root in the field of metallurgy, where a metal is heated and then cooled down slowly so that the atoms gradually configure themselves in states of lower internal energy, refining the crystalline structure of the metal. Compared with other global search algorithms such as genetic algorithm (Holland, 1992) and tabu search (Glover, 1989), SA is easier to understand and to implement without sacrificing performance. Since genetic algorithm is a population-based search method and tabu search is a memory-based heuristic, each iteration of SA runs faster than the two approaches. SA also requires a small number of user-specified parameters. In the past, SA has inspired various bioinformatics researches (e.g. Baker et al., 2009; Gonzalez et al., 2006; Chen et al., 2010).

We develop a new gene set-based SA to infer signaling cascades that characterize the optimal signaling pathway structure. Throughout we treat IFGSs as variables and their orders as random.

We also introduce a novel score function to measure the optimality, referred to as energy, of a candidate signaling pathway structure. Annealing refers to taking educated jumps in a feasible set of signaling pathway structures, where the true structure has the lowest energy. In the search process, the algorithm may jump to a neighboring structure with lower energy, resulting in a better move, or may accept to jump to a structure possessing higher energy in order to avoid getting trapped in a local minimum. Initially, when the temperature is high, the algorithm actively explores the feasible set. As cooling takes place, it spends more time around the global minimum. At any time instant, the algorithm only needs to keep track of the best-so-far structure. Figure 1 presents the work flow of the proposed approach.

We evaluated the performance of SA in three different case studies. The first study was conducted on 83 gene set compilations derived from the KEGG database, where SA demonstrated a significantly better performance in recovering the true signaling mechanisms than Bayesian network methods. Since both SA and Bayesian network methods accommodate discrete inputs, use a ‘search and score’ network learning strategy and output a directed network, they can be compared in terms of performance and computational time. Non-search-based approaches, e.g. MI-based gene regulatory network inference methods, are computationally more efficient than search algorithms and can be used to infer large-scale networks with thousands of genes. However, these approaches are suitable for inferring undirected pairwise dependencies. Thus, only the comparison between SA and Bayesian network methods is relevant to the present context. In the second study, we compared the performance of SA and Bayesian network methods using four benchmark *Escherichia coli* datasets available from the DREAM initiative. In the final study, we inferred two context-specific signaling pathway structures activated in breast cancer.

2 METHODS

2.1 Reconstruction of signaling pathway structures as a discrete optimization problem

Throughout we denote an IFGS (unordered gene set) by $X_i$ and an IF (ordered gene set) by $(X_i, \theta_i)$, where $\theta_i$ represents an ordering of genes (nodes) in $X_i$, $i = 1, \ldots, m$. Notations $\Sigma$ and $(\Sigma \Theta)$ are used for an IFGS compendium and a signaling pathway structure, respectively, where $\Sigma = (X_1, \ldots, X_m)$ and $(\Sigma \Theta)$ consists of $\Sigma$ such that $\Sigma \Theta = (\Sigma \Theta)$. Throughout we denote an IFGS (unordered gene set) by $X_i$ and an IF (ordered gene set) by $(X_i, \theta_i)$, where $\theta_i$ represents an ordering of genes (nodes) in $X_i$, $i = 1, \ldots, m$. Notations $\Sigma$ and $(\Sigma \Theta)$ are used for an IFGS compendium and a signaling pathway structure, respectively, where $\Sigma = (X_1, \ldots, X_m)$ and $(\Sigma \Theta)$ consists of $\Sigma$ such that $\Sigma \Theta = (\Sigma \Theta)$.
We propose a novel function to score a candidate signaling pathway structure with the minimum energy. Each likelihood term captures the overlapping information among IFs. The likelihood of an IF, say $\ell(X_i, \theta_i)$, is interpreted as its energy and is defined as

$$\ell(X_i, \theta_i) = \sum_{m=1}^{\infty} \log(\ell(X_i, \theta_i)),$$

where $\ell(X_i, \theta_i)$ stands for the likelihood of IF ($X_i, \theta_i$). Indeed, we compute the likelihood of ($X$, $\theta$) as

$$\mathcal{L}(X, \theta) = \prod_{i=1}^{m} \ell(X_i, \theta_i).$$

Since log function is monotonically increasing, searching for a structure with the maximum likelihood is equivalent to seeking a structure with the minimum energy. Each likelihood term $\ell(X_i, \theta_i)$ is computed using the estimates of two Markov chain parameters, the initial probability vector $\pi_0$ and the transition probability matrix $\Pi$. If there are $d$ distinct genes among the IFs ($X_i, \theta_i$), $i = 1, ..., m$, we estimate $\pi_0$ as

$$\pi_0 = \left(\frac{c_0}{m}, \cdots, \frac{c_m}{m}\right),$$

where $c_i$ is the total number of times $i$-th gene appears as the first node among $m$ IFs, for $i = 1, ..., m$. If $c_{rs}$ is the total number of occurrences of a directed edge from $r$-th gene to $s$-th gene among $m$ IFs, then

$$\Pi = [p_{rs}]_{r<s}$$

where $p_{rs} = c_{rs}/\sum_{i=1}^{m} c_{rs}$, $r, s = 1, ..., n$. Note that $\Pi$ captures the overlapping information among IFs. The likelihood of an IF, say $x \rightarrow y \rightarrow z$, can now be computed as

$$\ell(x \rightarrow y \rightarrow z) = P(x) \times P(y|x) \times P(z|y),$$

where prior and conditional probability terms in the above equation are known from $\pi_0$ and $\Pi$. The energy of a structure ($X$, $\theta$) can now be computed using Equation 3.

Algorithm 1: Optimal pathway structure by SA

1. Input: IFGS $X_i$, $i = 1, ..., m$, cooling schedule constant $c$, number of jumps $J$.
2. Output: The reconstructed signaling pathway structure.
3. Initialization: At $k=0$, randomly select a feasible structure ($X$, $\theta$). Let BestNetwork = ($X$, $\theta$) and BestEnergy = $\mathcal{E}(X, \theta)$. 
4. for $k = 1, ..., J$ do
5. Randomly choose a network ($X$, $\theta$) from the neighborhood of ($X$, $\theta$), $\mathcal{N} = (\theta, \theta')$. 
6. if $\mathcal{E}(X, \theta) < \mathcal{E}(X, \theta')$ then 
7. $\mathcal{E}(X, \theta) = \mathcal{E}(X, \theta')$. 
8. if $\mathcal{E}(X, \theta) <$ BestEnergy then 
9. BestEnergy = $\mathcal{E}(X, \theta)$. 
10. BestNetwork = ($X$, $\theta$). 
11. end if 
12. else 
13. Draw a Bernoulli sample with probability of TRUE as 
14. if TRUE then 
15. $\mathcal{E}(X, \theta) = \mathcal{E}(X, \theta')$. 
16. end if 
17. end if 
18. end for 

2.3 Feasible signaling pathway structures

Not all $\prod_{i=1}^{m} L_i$ signaling pathway structures, which can be constructed from $X$, exhibit the topological properties of real-world biological networks. To eliminate random structures from the search space, we only consider candidates, which possess certain low-level topological properties such as the degree distribution of underlying structure. The degree distribution of underlying signaling pathway structure, say ($X$, $\theta$), is a weighted asymmetric adjacency matrix $W$ obtained by counting the number of occurrences of directed edges between all gene pairs among $m$ IFs ($X_i, \theta_i$), $i = 1, ..., m$. Note that except for the pair of terminal nodes, the incoming and outgoing degrees of all intermediate nodes in an IF is 1. Since we consider ($X$, $\theta$) as a set of information flows, it can be easily verified that structures obtained by randomly permuting the orders of intermediate nodes in each IF ($X_i, \theta_i$), $i = 1, ..., m$, also have degree distribution $W$. Such structures preserve the marginal degree distributions of genes and form the feasible set $\mathcal{X}_F$ of size $\prod_{i=1}^{m} (L_i - 2)$. In simulation studies, $W$ can be obtained from the true signaling cascades. In real-world studies, it can be approximated by using database knowledge.

2.4 Justification of the energy function

We design and perform an empirical statistical test to show that the true signaling pathway structure has the lowest energy in the feasible set. Given the true signaling pathway structure ($X$, $\theta$), we randomly select $N$ feasible structures and compute the empirical $P$-value $M/N$, where $M$ is the number of structures with energy lower than that of ($X$, $\theta$). The true signaling pathway structure has the lowest energy if the empirical $P$-value is zero. We also perform the above test for a randomly selected feasible structure and expect the empirical $P$-value to vary in the interval $[0, 1]$. 

548
2.5 Search of the optimal signaling pathway structure

For the search procedure, we define the neighborhood of a signaling pathway structure \((X, \Phi)\) as the set of \(X \cup_n (L_i - 2)\) structures obtained by randomly permuting the orders of \(L_i - 2\) intermediate genes in the \(i\)-th IF \((X_i, \theta_i)\), keeping the remaining \(m-1\) IFs in \((X, \Phi)\) fixed, for each \(i = 1, \ldots, m\). This definition justifies the term ‘neighbor’ as only one IF in the given structure is perturbed at a time. Moreover, if we start our search from a feasible structure, the algorithm is guaranteed to take jumps within the feasible set of candidate structures having the same degree distribution as the true signaling pathway. The above definition also satisfies all the properties of a neighborhood presented in (Goldstein and Waterman 1998). We choose the standard cooling schedule, which at the \(k\)-th stage is defined as

\[
T_k = \frac{c}{\log(k+1)}, \quad k = 1, 2, \ldots, \tag{7}
\]

where \(c > 0\) is constant and referred to as cooling schedule constant. The choice of \(c\) is often problem specific. Indeed, a small value of \(c\) may lead SA to get trapped in a local solution, whereas a large value may slow down its speed of convergence. The above cooling schedule has been used to study the convergence properties of a general simulated annealing approach (Hajek, 1988). The probability with which the algorithm accepts a move from a current structure \((X, \Phi)\) to a neighboring structure \((X', \Phi')\) is called the acceptance probability (Chong and Zali 2008) and is defined as

\[
\min \{1, \exp(\varepsilon(X, \Phi) - \varepsilon(X', \Phi')/T)\} \tag{8}
\]

where \(T\) represents the current temperature value, which at the \(k\)-th iteration is given by Equation (7). Note that the algorithm may accept a move to a worse point in order to avoid getting trapped in a local solution. In Algorithm 1, we present the pseudo-code of SA. Algorithm 1 takes an IFGS compendium as input and returns a list of IFs, which are combined to represent the optimal signaling pathway structure.

2.6 Computational complexity

The worst-case running time of SA is \(O(n\text{IntL})\), where \(J\) is the number of jumps, \(m\) is the number of IFGSs and \(L\) is the maximum length of an IFGS in the given compendium. We refer to Section 3 in the Supplementary Material for a detailed discussion on the computational complexity of SA. Overall, SA benefits from a manageable computational load compared with similar search heuristics such as sampling-based Metropolis–Hastings algorithm used in the inference of Bayesian networks. We reemphasize that SA and Bayesian network methods are similar in terms of input, output and network learning strategy. In the inference of Bayesian networks, discrete data are commonly used for a manageable computational complexity. Thus, SA and Bayesian network methods take the same type of input. Both SA and Bayesian network methods share a ‘search and score’ strategy for learning multivariate dependencies. Also, both SA and Bayesian network methods output a directed network. The preceding factors make SA and Bayesian network methods (i) suitable for inferring signaling pathway structures, which are directed networks containing up to a few hundred nodes and (ii) comparable in terms of performance and computational time. Other non-search-based approaches, such as MI-based methods, are computationally more efficient than search methods and can be used for reconstructing gene regulatory networks with thousands of nodes. However, they are suitable for inferring undirected pairwise similarities. Therefore, only the comparison between SA and Bayesian network methods is relevant to the present study.

3 RESULTS

3.1 Case Study I: proof of principle

3.1.1 Description of the datasets

In this study, we evaluate the performance of SA in inferring the true signaling mechanisms, when gene sets are sampled from the true signaling pathway structure. As the input for SA is an IFGS compendium, we first developed a path sampling algorithm (see Section 1 in Supplementary Material) to sample a collection of true IFs from a known pathway structure. The loss of gene ordering information in IFs was simulated by randomly relocating intermediate genes within each IF, keeping the pair of terminal nodes fixed. We used this algorithm on each of the 120 non-metabolic pathways in the KEGG database (Kanehisa et al. 2008) to derive 120 IFGS compendiums. From each compendium, we removed IFGSs of lengths 2 and 3 as they represented true edges and true IFs, respectively. Among the resulting compendiums, we only considered the ones containing a minimum of five IFGSs to allow overlapping among gene sets. The above procedure resulted in 83 non-empty IFGS compendiums composing of under-sampled IFGSs. Since each compendium was derived from a specific KEGG pathway structure, IFGSs in a given compendium shared the same pathway membership. In the derived compendiums, the number and lengths of IFGSs varied in the ranges of 5–723 and 4–13, respectively. We applied SA on each compendium individually to infer the true signaling cascades, i.e. the ones present in the original KEGG pathways. If there are \(m\) gene sets with \(n\) distinct genes in an IFGS compendium, then the input for SA can be given as an \(m \times n\) matrix. If there are \(k\) genes in the \(i\)-th gene set, then the first \(k\) locations in the \(i\)-th row contain non-zero indices representing these genes, and the remaining \(n-k\) locations are set to 0. SA only considers non-zero indices in each row, i.e. genes present in a gene set. The IFs inferred by SA are assembled to reconstruct the signaling pathway structure. We compare the inferred structure with the one constructed from the true IFs.

3.1.2 Description of Bayesian network methods

First, we note that a gene set compendium can be written as a matrix of binary discrete values (Fig 2). A gene set can be naturally interpreted as a set of genes expressed in an experiment and thus corresponds to a vector (sample) of binary values obtained by considering the
Fig. 3. (A) empirical P-values computed for true signaling pathway structures (Left) and randomly selected feasible pathway structures (Right) corresponding to 83 IFGS compendiums derived from the KEGG pathways. (B) Energy values computed by varying the cooling schedule constant for a total of $2 \times 10^5$ jumps. The IFGS compendium was derived from the generic vascular smooth muscle contraction pathway in KEGG.

Presence (1) or absence (0) of genes in the gene set. We only consider genes belonging to the given IFGS compendium. For example, if there are m gene sets with n distinct genes in a compendium, then the binary discrete data is an $m \times n$ matrix. If there are k genes in the i-th gene set, then the corresponding k locations in the i-th row of data are set to 1 and the remaining $n-k$ locations are set to 0. Such matrices serve as input to Bayesian network methods.

We considered two Bayesian network approaches: K2 (Cooper and Herskovits, 1992) and Metropolis–Hastings or MH (Murphy, 2001a) implemented in the Bayes Net Tool Box (BNT) (Murphy, 2001b). Given an initial ordering of nodes, the K2 approach is based on incrementally assigning a parent to a node whose addition increases the score of the resulting structure the most. MH algorithm starts from an initial directed acyclic network and sequentially samples networks from the neighborhood of the most recent network. Neighborhood in the context of MH is the collection of all directed acyclic networks that differ from the given network by addition, deletion or reversal of a single edge. For scoring a structure, BNT provides the Bayesian Information Criterion (BIC) and Bayesian score function. We used both BIC and Bayesian scoring (with Dirichlet prior) functions to infer Bayesian networks. In the case of K2, the maximum number of parents allowed for a node was set at three for a manageable computational complexity.

3.1.3 The proof-of-principle study. We began by examining that the true signaling pathway structure has the lowest energy in the feasible set. We considered two collections of feasible structures. The first collection composed of all 83 signaling pathway structures constructed from the true IFs. The second collection contained 83 randomly selected structures, one from each of the 83 feasible sets. Figure 3A presents the empirical P-values calculated for each structure in the two collections, where we fixed $N=1000$ (see Section 2). We observed that the empirical P-value for each of the 83 true structures was always zero while it fluctuated in the interval (0 1) in the case of randomly selected feasible structures. This justified the choice of the energy function used in our algorithm.

For choosing the cooling schedule constant and number of jumps, we considered 10 IFGS compendiums with the number of IFGSs in the range 30–723. Note that the signaling pathway structures in public databases are often generic in nature. So, only a part of a signaling pathway structure will be activated under a specific context, as opposed to the entire structure. Therefore, the above gene set compendiums are a reasonable representation of underlying context-specific signaling mechanisms. As a result, the choice of parameters based on our evaluation is also applicable to other similar scenarios.

We evaluated the performance of SA by setting the cooling schedule constant at five different levels $c=1$, 10, 20, 30 and 40 and the number of jumps at four different levels $J=1 \times 10^4$, $5 \times 10^4$, $1 \times 10^5$, $2 \times 10^5$. In general, a small value of $c$ may result in a local solution, whereas a large value of $c$ may require large computational time. This fact is also evident from Figure 3B, where we present energy values from four different runs of SA with cooling schedule constant set at $c=1$, 10, 20 and 30. Thus, a value of $c$ should be chosen to comprise between inference accuracy and computational time. We summarize the performance of SA in terms...
Reverse engineering the optimal signaling pathway structures from gene sets

Table 1. Comparison of SA and Bayesian network methods MH and K2 (using Bayesian score) in terms of computational time (in minutes) and F-score

<table>
<thead>
<tr>
<th></th>
<th>10³</th>
<th>10⁴</th>
<th>10⁵</th>
<th>2 × 10⁵/Final</th>
<th></th>
<th>10³</th>
<th>10⁴</th>
<th>10⁵</th>
<th>2 × 10⁵/Final</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time</td>
<td>F-score</td>
<td></td>
<td></td>
<td></td>
<td>Time</td>
<td>F-score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA</td>
<td>0.02</td>
<td>0.18</td>
<td>1.9</td>
<td>3.7</td>
<td>SA</td>
<td>0.57</td>
<td>0.89</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MH</td>
<td>0.49</td>
<td>5.14</td>
<td>53.37</td>
<td>118.1</td>
<td>MH</td>
<td>0.11</td>
<td>0.16</td>
<td>0.17</td>
<td>0.21</td>
</tr>
<tr>
<td>K2</td>
<td>0.10</td>
<td>0.01</td>
<td></td>
<td></td>
<td>K2</td>
<td>0.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MH</td>
<td>0.03</td>
<td>0.32</td>
<td>3.2</td>
<td>6.5</td>
<td>MH</td>
<td>0.69</td>
<td>0.91</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>K2</td>
<td>0.12</td>
<td>27.02</td>
<td>0.10</td>
<td></td>
<td>K2</td>
<td>0.08</td>
<td>0.11</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>SA</td>
<td>0.04</td>
<td>0.39</td>
<td>3.9</td>
<td>7.9</td>
<td>SA</td>
<td>0.11</td>
<td>0.16</td>
<td>0.17</td>
<td>0.21</td>
</tr>
<tr>
<td>MH</td>
<td>2.22</td>
<td>21.11</td>
<td>0.10</td>
<td></td>
<td>MH</td>
<td>0.09</td>
<td>0.145</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>K2</td>
<td>0.32</td>
<td></td>
<td></td>
<td></td>
<td>K2</td>
<td>0.27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MH</td>
<td>0.20</td>
<td>2.00</td>
<td>19.91</td>
<td>39.92</td>
<td>MH</td>
<td>0.022</td>
<td>0.48</td>
<td>0.644</td>
<td>0.71</td>
</tr>
<tr>
<td>K2</td>
<td>367.5</td>
<td></td>
<td></td>
<td></td>
<td>K2</td>
<td>1.49</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Performance of SA and MH is evaluated at jump/sample index 10³, 10⁴, 10⁵ and 2 × 10⁵. In the case of K2, total time and F-score is presented. We considered 4 IFGS compendiums with 54, 108, 195 and 723 IFGSs (in the same order). In the case of MH, a structure with the highest F-score among the sampled structures was used for comparison.

aProgram terminated due to memory crash.

Fig. 5. Performance of SA in reconstructing the true signaling cascades and signaling pathway structures corresponding to 83 IFGS compendiums derived from the KEGG database.

Fig. 6. Comparison of SA with Bayesian network methods K2 and MH using BIC and Bayesian score functions. (A and B) Shows F-score and precision, respectively.

of F-score and precision averaged over 10 independent runs. F-score is defined as \(2pr/(p+r)\), where \(p\) and \(r\) stand for precision and recall, respectively. Precision is the proportion of true positives among the inferred edges.

In Figure 4, we observe an increase in the performance of SA with increasing number of jumps (Row 1 to Row 4), for each fixed value of \(c\). Moreover, the F-scores and precision values are overall better in the case of \(c=10\), compared with other values of \(c\). In Table 1, we compare SA, MH and K2 in terms of computational time and F-score, where we use four IFGS compendiums with 54, 108, 195 and 723 IFGSs. We also present the performance of SA and MH at different jump/sample indices. It is clear from Table 1 that the total time required by SA to take 2 × 10⁵ jumps may be smaller than the time required by MH to sample 10³ – 10⁴ structures. MH also suffers from large memory requirements. Moreover, performance of SA is significantly better than MH at different jump indices. Since K2 does not depend on the number of jumps, we list total time required in a single run of the algorithm. At the end of 2 × 10⁵ jumps, total time required by SA is higher than K2 by a manageable difference. On the other hand, F-scores from SA are significantly higher than the ones from K2. By considering 2 × 10⁵ jumps, the F-score could be increased up to 70% in the case of a large compendium with 723 IFGSs. Thus, the parameters \(c=10\) and \(J=2 \times 10⁵\) provide a good compromise between computational time and method performance.

By fixing \(c=10\) and \(J=2 \times 10⁵\), we applied SA on all 83 IFGSs compendiums. Figure 6 demonstrates the performance of SA in reconstructing the true signaling mechanisms. On the left and middle panels of Figure 6, we have plotted the number of structures among 83 reconstructed structures with a certain minimum precision and F-score, respectively. On the right panel, we have considered the.
Fig. 7. An example showcasing the performance of SA in recovering the true structure using the IFGS compendium derived from GnRH signaling pathway in KEGG database. Structures represent true (A) and inferred signaling pathways (B), respectively. The black (solid) and blue (dashed) edges represent true positives and false positives, respectively. Figures were generated using Cytoscape (Shannon et al., 2003).

proportion of signaling cascades accurately inferred by SA in each compendium. The feasibility and validity of SA is evident from the high precisions, $F$-scores and high proportions of accurately inferred signaling cascades.

In Figure 6, we present the results from a comparative study performed using each of the 83 IFGS compendiums. We observe a significantly better performance of SA in recovering the true structure compared with the Bayesian network methods. In each run of MH, the first 1000 samples were collected for a manageable computational complexity and the structure giving the highest $F$-score was selected for comparison. Figure 6 demonstrates the strength of SA in inferring signal cascading mechanisms. As described in Section 3.1.1, each IFGS compendium considered in Figure 6 contained gene sets that represented true signaling events in the corresponding KEGG structure. However, we did not know the ordering of genes in the events. As a result, binary discrete data used for Bayesian network methods is also a true representation of underlying signaling events. Note that in each sample (gene set) of binary data matrix, genes that participate in underlying IF always fall in the same bin. Due to the use of this true data representation, we expect all algorithms to perform well. Nonetheless, the strength of Bayesian network methods lies in inferring causal interactions (column–column association), whereas SA explicitly considers signal cascading mechanism in each row. Therefore, we observe a superior performance of SA.

We also evaluated the performance of SA when IFGSs in a compendium shared multiple pathway memberships (see Section 4 in Supplementary Material). Results from this evaluation were similar to the ones in Figure 7.

In Figure 8, we present a signaling pathway structure inferred by our approach. Structures on the left and right correspond to the true and inferred signaling pathway structures, respectively. The black (solid) and blue (dashed) edges represent true positives and false positives, respectively. Figure 8 demonstrates high precision and recall in the structure reconstructed by SA, resulting in a high $F$-score.

3.2 Case Study II: evaluation using E. coli datasets

3.2.1 Description of the datasets In this study, we compared the performance of SA and Bayesian network methods using four benchmark E. coli datasets available from DREAM3 network challenges in the DREAM initiative (Marbach et al., 2009, 2010; Prill et al., 2010). The first two datasets comprise of 50 genes and 51 samples, whereas the remaining two datasets contain 100 genes and 101 samples. The corresponding gold standard networks comprise 62, 82, 125 and 119 edges, respectively. We compared the inferred structures with the corresponding gold standards. We first derived four IFGS compendiums from the above datasets by declaring the top 10% of the measurements in each dataset as 1 and the remaining measurements as 0. This discretization produced IFGSs of diverse lengths across different samples. In each compendium, we considered IFGSs with lengths in the range 3–9. This resulted in four IFGS compendiums with 47, 45, 45 and 49 IFGSs, respectively.

3.2.2 Performance evaluation We used SA to explore the search spaces formed by considering all possible gene orderings of IFGSs present in each compendium. We applied K2 and MH on the binary equivalent data corresponding to each compendium. Since we could not discover any structure in several runs of K2 on some of the compendiums, we present the performance of SA and MH. In Figure 8A, we show the performance of SA and MH in terms of $F$-score ratio, which is the ratio of $F$-score from SA and the one from MH. In Figure 8B, we present the performances in terms of precision ratio. A ratio $>$1 indicates a better performance of SA.

In the case of SA, a structure was inferred by fixing the cooling schedule constant at 10 and the number of jumps $2 \times 10^5$. In the case
Reverse engineering the optimal signaling pathway structures from gene sets

Table 2. Comparison of SA and MH in terms of computational time (in minutes) using four E.coli datasets from the DREAM initiative

<table>
<thead>
<tr>
<th>Method</th>
<th>E. coli 1</th>
<th>E. coli 2</th>
<th>E. coli 3</th>
<th>E. coli 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>3.41</td>
<td>3.25</td>
<td>4.47</td>
<td>4.50</td>
</tr>
<tr>
<td>MH-BIC</td>
<td>24.95</td>
<td>22.41</td>
<td>62.65</td>
<td>47.98</td>
</tr>
<tr>
<td>MH-BAYES</td>
<td>25.19</td>
<td>22.61</td>
<td>174.61</td>
<td>72.62</td>
</tr>
</tbody>
</table>

While the figures do not attempt to portray a comprehensive view of signaling pathways, SA algorithm has the potential to uncover biologically relevant mechanisms that have not been previously considered or understood.

3.3 Case Study III: ERBB and PMOM pathways activation in breast cancer

3.3.1 Description of the datasets In this study, we showcase two context-specific signaling pathways, ERBB and PMOM (progesterone-mediated oocyte maturation), activated in breast cancer. We considered 87 genes participating in the ERBB signaling pathway and 35 genes in the giant connected component (GCC) of the PMOM pathway from the KEGG database. We analyzed 299 clinical breast cancer tissue gene expression profiles from the DREAM initiative.

While the figures do not attempt to portray a comprehensive view of signaling pathways, SA algorithm has the potential to uncover biologically relevant mechanisms that have not been previously considered or understood.

ERBB/HER family receptors play important roles in many types of cancer including breast cancer. Disregulation/mutation in the epidermal growth factor receptor (EGFR) and ERBB2 (HER2) have been known to promote angiogenesis and metastasis in breast cancer [Furue and Tendler, 2008; Navone et al., 2009]. Some known signaling cascades that contribute to breast cancer progression include RAS/MEK/ERK and PI3K/PDK1/AKT signaling pathways that regulate apoptosis and cell cycle. These signaling events are reflected in the edges depicted in Figure 9B. For instance, in breast cancer ERBB2/HER2 receptor can constitutively activate the PI3K/PDK1/AKT cascade and the downstream effector, the mammalian target of rapamycin (MTOR). This known signaling cascade is conformed as a direct action between ERBB2 and MTOR in Figure 9B.

In Figure 9C, the reconstructed ERBB signaling pathway revealed a previously unknown direct link from ERBB3 to ARAF. ARAF (A-Raf proto-oncogene serine/threonine-protein kinase) is known to phosphorylate and activate MEK1 (MAP2K1) and MEK2 (MAP2K2), leading to the suppression of apoptosis in cancer cells [Huang et al., 2004]. However, the possible role of ERBB3 as an upstream regulator is a novel implication that clearly warrants further investigation. In addition, PI3K family members are known to phosphorylate and activate MEK1 (MAP2K1) and MEK2 (MAP2K2), leading to the suppression of apoptosis in cancer cells [Huang et al., 2004]. However, the possible role of ERBB3 as an upstream regulator is a novel implication that clearly warrants further investigation. In addition, PI3K family members are known to...
to be the downstream targets of FGFR and ERBB2/HER2, but not ERBB3. Thus, the direct link between ERBB2 and PI3K inferred by SA is in accordance with the previously established results. The direct link between ERBB3 and PI3K, on the other hand, suggests a potential role of ERBB3 receptor tyrosine kinase in breast cancer. A major clinical challenge of breast cancer treatment is acquired resistance to hormone therapy as the tumor develops alternative survival signaling such as enhanced cross-talk between the estrogen receptor (ER) and ERBB1/ERBB2. Thus combinatorial therapeutic intervention targeting both ER and ERBB2 (HER2) is currently under intensive clinical studies. Revelation of the novel link between ERBB3 and PI3K family proteins is significant because it represents yet another adaptive pathway in breast cancer that needs to be fully understood in order to develop a more effective regimen blocking this survival signaling.

In the case of PMOM pathway (Fig. D), we show a highlighted role of the Fizzy protein (FZR1/CDC20) in breast cancer. It is an indication that the ubiquitin ligase activity of the anaphase promoting complex (APC) plays an important role in breast cancer progression. Previous studies have established an association between APC and FZR1 implicating FZR1 regulation of ANAPC isoforms 1, 2, 4, 5, 7 and 10. We observe additional regulation mechanisms involving ANAPC 11 and 13, apparently in a way specific to breast tumor tissues. The reconstructed PMOM signaling pathway also reveals a novel direct action of mitogen-activated protein kinase 1 (MAPK1) upon FZR1. The MAP kinase cascade is associated with the control of cell cycle progression, but in a manner that is far upstream of FZR1-mediated APC. It is possible that this direct action may be a result of the non-genomic signaling of progesterone that rapidly and constitutively activates the MAP kinase signaling cascade in breast cancers that are ER positive but progesterone receptor (PGR) negative.

If experimentally validated and mechanistically elucidated, the novel activation of FZR1 by MAPK1 will have important outcomes in breast cancer research. For example, studies can be designed to investigate if inhibiting the kinase can block FZR1-mediated APC, and if any effector proteins are involved in this signaling cascade. Such studies can be driven by hypotheses generated from SA-based reconstruction of signaling pathways, and can lead to the discovery of new biomarkers as potential diagnostic, prognostic, or therapeutic targets for breast cancer.

4 CONCLUSION

In this article, we presented a novel SA approach to learn the optimal signaling pathway structures from gene sets. We hypothesized a true signaling pathway structure as an ensemble of overlapping signal cascades. We then translated its reconstruction from unordered gene sets corresponding to signaling cascades into a discrete optimization problem. Throughout we treated gene sets as random variables and their orders as random. We also introduced a novel energy
function to measure the optimality of a signaling pathway structure. Overall, our approach benefits from the following: (i) treatment of unordered gene sets as random variables and building blocks of a signaling pathway allows us to explicitly consider signal cascading mechanisms in the underlying structure. (ii) The problem easily fits into the framework of discrete optimization, where the feasible space is finite but is difficult to explore. (iii) The computational complexity of SA is manageable. In Case Study I, performance evaluation using 83 gene set comprehends derived from the KEGG pathways demonstrated that SA could recover the underlying structures more efficiently than Bayesian network methods. In Case Study II, we compared the performance of SA and Bayesian network methods using four E.coli datasets available from the DREAM initiative. In Case Study III, breast cancer-specific reconstruction of two signaling pathway structures from the KEGG database further proved the advantages of using SA in real-world scenarios.

The proposed study is useful since the prior known pathway structures may not represent a complete picture of underlying signal cascading mechanisms. There might exist additional mechanisms among genes related to the pathways. Also the pathway structures in databases are often generic, whereas scientists may be interested in learning context-specific networks of genes in the pathways. SA can be used in such scenarios. As gene set-based structural inference of signaling pathways is new to the biomedical field, refinement and extension of our algorithm is an important future research direction for us. For example, the current setting can be combined with the identification of pathway components from high-throughput transcriptomics data. SA will also benefit by penalizing random structures in the search space and improving the current jump strategy to locate the optimal solution. We believe that gene set-based approach is an important step toward the reconstruction of signaling pathway structures from molecular profiling data available in diverse forms.

Funding: This work was supported by NIH (grants R21LM010137 and 5G12RR026265-2 to D. Z. and G. W.), respectively.

Conflict of interest: none declared.

REFERENCES


