Parameter Identifiability-based Optimal Observation Remedy for Biological Networks

Yulin Wang[1]

Email: wyl@uestc.edu.cn

Hongyu Miao[2,*]

Email: Hongyu.Miao@uth.tmc.edu

[*]Corresponding author

[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

[2] Department of Biostatistics, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, 77030, USA

Abstract

Background: To systematically understand the interactions between numerous biological components, a variety of biological networks on different levels and scales have been constructed and made available in public databases or knowledge repositories. Graphical models such as structural equation models have long been used to describe biological networks for various quantitative analysis tasks, especially key biological parameter estimation. However, limited by resources or technical capacities, partial observation is a common problem in experimental observations of biological networks, and it thus becomes an important problem how to select unobserved nodes for additional measurements such that all unknown model parameters become identifiable. To the best knowledge of our authors, a solution to this problem does not exist until this study.

Results: The identifiability-based observation problem for biological networks is mathematically formulated for the first time based on linear recursive structural equation models, and then a dynamic programming strategy is developed to obtain the optimal observation strategies. The efficiency of the dynamic programming algorithm is achieved by avoiding both symbolic computation and matrix operations as used in other studies. We also provided necessary theoretical justifications to the proposed method. Finally, we verified the algorithm using synthetic network structures and illustrated the application of the proposed method in practice using a real biological network related to influenza A virus infection.

Conclusions: The proposed approach is the first solution to the structural identifiability-based optimal observation remedy problem. It is applicable to an

arbitrary directed acyclic biological network (recursive SEMs) without bidirectional edges, and it is a computerizable method. Observation remedy is an important issue in experiment design for biological networks, and we believe that this study provides a solid basis for dealing with more challenging design issues (e.g., feedback loops, dynamic or nonlinear networks) in the future. We implemented our method in R, which is freely accessible at https://github.com/Hongyu-Miao/SIOOR.

Keywords: Biological network, Graphical model, Structural identifiability analysis, Structural equation model, Observation strategy

Background

The emergence of young research fields such as systems biology and network medicine [1, 2] reflects some exciting changes in biomedical investigators' view of biology and practice. Particularly, it has been increasingly recognized that thinking in networks may lead to novel scientific insights and findings [3] that the traditional reductionism approaches cannot grant [4]. The recent development of experimental techniques (e.g., a variety of high-throughput omics approaches) also provides unprecedented opportunities for biomedical investigators to construct numerous biological networks at different levels and scales; for instance, protein-protein interaction networks [5, 6], gene regulatory networks [7-10], functional RNA networks [11-13], and metabolic networks [14, 15] can be found in a number of databases or knowledge repositories nowadays [9, 16, 17]. All such previous efforts provide a solid basis for further advancing our understanding of biological systems and the associated outcomes qualitatively or quantitatively.

Graphical models have long been considered as a natural mathematical representation of biological network for various quantitative analysis tasks such as parameter inference [18-21]. Specifically, given a biological network structure and experimental observations of certain variables associated with network nodes, it is often of significant interest to determine the unknown coefficients associated with network edges. For instance, to understand the responses of a biological network (e.g., activation or inhibition) to different environmental signals (e.g., different signaling molecules or different doses of the same signaling molecule), edge coefficients are

likely to vary under different conditions and thus need to be estimated under each condition for the same given network structure [18]. In such a scenario, although the structure of the corresponding graphical model is known and fixed, concerns about the accuracy and reliability of parameter estimates often raise due to, e.g., the existence of unobserved node variables (i.e., latent variables). In practice, latent variables are not uncommon due to various technical limitations, ethic issues, financial affordability, and so on [18, 20]. Therefore, a natural question to ask is: what is the remedy that enables us to obtain reliable parameter estimates for a given graphical model structure with partially observed variables?

To the best knowledge of our authors, the aforementioned important question has rarely been tackled before in the context of quantifying unknown model parameters of biological networks; and in this study, we make the very first attempt to address this question from the structural identifiability point of view. By the definition in Miao et al. [18], an unknown model parameter is structurally identifiable if it can be uniquely determined for a given model structure under the assumptions that sample size is sufficiently large and data quality is not of concern. Of course, one can also take the effects of sample size and data noise into consideration and conduct the so-called practical identifiability analysis [18]; however, this is out of the scope of this study as practical identifiability analysis is not feasible at certain experimental design stage when real data are not available. On the contrary, structural identifiability analysis allows us to detect flaws in model structure and observation scheme before data collection, and thus should be investigated first. Our solution to the question

mentioned at the end of the previous paragraph is thus a strategy that identifies a minimum number of unobserved nodes, for which the associated node variables should be observed in experiments such that all unknown parameters become structurally identifiable. This is a useful and cost-effective remedy if some of the model parameters are not identifiable given the original observation scheme, and we thus name it the structural identifiability-based optimal observation remedy (SIOOR).

Since biological networks can be represented by many different types of mathematical or statistical models, it is impossible to devise the SIOOR strategy for every different model type in one study. Therefore, we consider a linear structural equation model [22] here because it is a representative graphical model type and has been widely applied in various disciplines including systems biology [23-27]. A number of previous studies have investigated the parameter identifiability problem of SEMs, but the majority of these studies only derived theoretical criteria or conditions for identifiability verification, including Pearl's back door and front door criteria [28], Brito and Pearl's generalized instrumental variable criterion [29], Tian's accessory set approach [30]. Only a few studies proposed computerizable identifiability analysis approaches, including Drton's condition [31] and Foygel's half-trek criterion [32] (implemented in R package SEMID), Sullivant's computer algebra method and the more recent Wang's identifiability matrix method [33, 34]. More importantly, all such criterions and methods assume that the observation strategy is given (i.e., it is pre-specified which variables are observed and which are not), and none of them considered the remedy strategy if a given observation strategy does not grant

identifiability to all unknown model parameters. The focus of this study is thus to investigate how to choose a minimum number of nodes that are not observed in the original observation strategy for additional experimental measurements such that all unknown model parameters become identifiable. This study leads to a general and computerizable solution to the SIOOR problem for the first time.

More specifically, in the case that a given observation strategy of a biological network cannot grant identifiability of all unknown parameters in the corresponding SEM due to the existence of unobserved variables, we propose a dynamic programming (DP) approach to search for all possible SIOOR strategies. The proposed approach is a generic and computerizable method that can deal with recursive SEMs. It should be stressed that SIOOR strategy does not involve any power or sample size calculation and thus cannot be compared with the traditional experimental design approaches [35, 36]. Also, it should be stressed that the observability problem in control theory is different from the SIOOR problem because the aim of observability analysis is to determine the internal states of a system from its external outputs [37]. For clarification purpose, we also compare Liu's graphic approach for observability analysis [38] with our SIOOR strategy in this study.

This article is organized as follows. In the Methods Section, the structural identifiability-based optimal observation remedy problem is mathematically formulated. We then propose a dynamic programming approach with theoretical justification to solve the problem for recursive SEMs. In the Results and Discussion Section, we describe our algorithm implementation and validate the proposed method

using selected benchmark networks. Also, a real substructure from the influenza virus A [39] KEEG pathway is chosen as an example to illustrate the application of the proposed method in practice.

Methods

In this section, several key concepts and definitions are introduced for solving the SIOOR problem, including Observation Strategy (OS), Cardinality of Observation Strategy [4], and Identifiability Gain (IG). The design of the dynamical programming algorithm is also described. In addition, we provide the necessary theoretical justification for the proposed method.

Problem formulation

A directed biological network can be denoted by $\mathrm{G}=(\mathbf{V},\mathbf{E})$, where **V** denotes the node set and **E** denotes the edge set. Let V_i $(i=1,2,\ldots,n)$ denote the i-th node, and Y_i denote the variable associated with V_i. If Y_i is a linear function of the remaining node variables, the corresponding SEM can be specified as follows,

$$Y_i=\sum_{j\neq i}c_{ij}Y_j+\varepsilon_i, \qquad i,j=1,\cdots,n,$$

where c_{ij} denotes the coefficient associated with the directed edge $V_j\rightarrow V_i$, and ε_i denotes the disturbance error term that follows a certain distribution (Gaussian or non-Gaussian [40, 41]) with mean zero. For simplicity, all disturbance error terms are assumed to be independent. By definition, **E** specifies the structure of the coefficient matrix $\mathbf{C}=\left[c_{ij}\right]$, i.e., $c_{ij}=0$ if no edge exists in **E** from V_j to V_i for $i\neq j$. When

a network structure contains one or more loops, G is a directed cyclic graph (DCG) and the corresponding SEM is called a non-recursive model; otherwise, G is a directed acyclic graph [42] and the corresponding SEM is called recursive. Although Drton's condition [31] and Foygel's half-trek criterion [32] are applicable to the identifiability analysis of non-recursive SEMs, the identifiability of parameters on a loop may be still inconclusive. Due to the lack of mature structural identifiability analysis techniques for examining every unknown parameter of a non-recursive SEM, this study focuses on recursive SEMs (i.e., DAGs) only.

Definition 1 (observation strategy). Given a graph $\mathrm{G}=(\mathbf{V},\mathbf{E})$, its observation strategy can be denoted by a binary vector $O=\left(O_{V_1},\cdots,O_{V_n}\right)^T$, where $O_{V_i}=1$ if node V_i is observed and $O_{V_i}=0$ if V_i is unobserved. □

Observation strategy is important to parameter identifiability. In general, for a given network structure, the more observed nodes an observation strategy contains, the more likely all model parameters are identifiable. However, more observed nodes are usually associated with a higher experiment cost, so it is also desirable to reduce any unnecessary cost. The goal of SIOOR is thus to improve a given observation strategy by observing a minimum number of originally unobserved nodes such that all nonzero parameters in $\mathbf{C}$ become identifiable. For this purpose, let $\mathbf{P}$ denote the vector of all nonzero parameters in $\mathbf{C}$, and let $\mathbf{D}$ denote the vector of identifiability status of every element in $\mathbf{P}$. That is, if P_i is locally or globally identifiable (i.e., P_i has a finite number of possible values or a unique value within the parameter space, see [18]), $D_i=1$; otherwise, $D_i=0$. When all the parameters in a model are locally

or globally identifiable, this model is called identifiable. Consequently, the SIOOR problem can be formulated as follows

$$\min_{\text{observed } V_i} \sum_{i=1}^{n} O_{V_i} \ , \ \text{subject to} \ \ \mathbf{D} = \mathbf{1}, \tag{1}$$

where $\sum_{i=1}^{n} O_{V_i}$ is the total number of observed nodes in an observation strategy O, and $\mathbf{1}$ denotes a vector of ones. For clarification, we stress that the observation measurements are for the random variables associated with network nodes, and we assume $(n-m)$ of them are observed in the original observation strategy, where n denotes the total number of nodes and $0 < m \leq n$.

The objective function above is minimized with respect to the originally unobserved nodes, subject to the constraint $\mathbf{D} = \mathbf{1}$. During the minimization process, it needs to be repeatedly verified whether all parameters have become identifiable (i.e., $\mathbf{D} = \mathbf{1}$). For this purpose, an efficient algorithm for structural identifiability analysis of SEMs is needed. Here we consider the identifiability matrix method proposed by Wang et al. [34]. Briefly, structural identifiability of parameters can be verified by examining the number of solutions to the symbolic polynomial identifiability equations generated by Wright's path coefficient method [43, 44]. To avoid the expensive symbolic computation involved in reducing such identifiability equations, the identifiability matrix method proposes to derive binary matrices from symbolic polynomials and thus enable us to determine the number of solutions via several simple matrix operations. It is noteworthy that Wang's work [34] does not explicitly handle colliders involving bidirectional arcs when generating identifiability equations

with Wright's method, however, the identifiability matrix method is still applicable here as we do not consider bidirectional arcs in DAGs.

Identifiability gain and must-be-observed nodes

The optimization problem in the previous section is combinatorial in nature. Therefore, if the number of the originally unobserved nodes (denoted by m) is not small, enumerating all the 2^m different possible observation strategies over these nodes will be computationally expensive. We thus need an efficient algorithm such as dynamic programming to obtain the solutions. For this purpose, a few more definitions need to be introduced first.

Definition 2 (redundant identifiability equation). Given a set of identifiability equations, an identifiability equation $IE\left(V_i, V_j\right)$ is redundant with respect to that set if it can be expressed as a linear combination of the equations in that set.

Definition 3 (cardinality of observation strategy). Given an observation strategy O for a network G, one symbolic polynomial identifiability equation can be generated for each pair of *d*-connected [28] observed nodes using, e.g., Wright's path coefficient method. Then the total number of non-redundant identifiability equations is called the cardinality of O, denoted by $f\left(O\right)$. □

The Wright's path coefficient method generates identifiability equations for recursive SEMs by calculating the covariance between two node variables, which is equal to the sum of the products of edge coefficients along each *d*-connected path, i.e., $IE\left(V_i, V_j\right)$: $Cov\left(V_i, V_j\right) = \sum_{path_k} \prod_{edge_l} \theta_l$. After removing all redundant identifiability

equations and redundant monomials, the identifiability result of each parameter can be determined by Theorem 1 in [34]. That is, if the number of non-redundant identifiability equations is less than the number of unknown parameters, then the parameters have an infinite number of possible values within the parameter space and are thus unidentifiable; otherwise, the parameters have a limited number of solutions or even a unique solution and are thus at least locally identifiable [45]. Let N_u denote the total number of unknown parameters in $\mathbf{P}$. For every parameter in $\mathbf{P}$ being locally or globally identifiable, the inequality $f(O) \geq N_u$ should hold according to Theorem 1 in [34]. Therefore, the optimization problem can also be formulated as follows

$$\min_{\text{observed } V_i} \sum_{i=1}^{n} O_{V_i} \text{ , subject to } f(O) \geq N_u, \tag{2}$$

where the calculation of $f(O)$ is a key challenge because it depends on specific network structure and observation strategy and thus has no closed-form solution. We thus introduce the following definition.

Definition 4 (identifiability gain). Given a network $\mathrm{G} = (\mathbf{V}, \mathbf{E})$, let $O^{(k)}$ and $f(O^{(k)})$ denote an observation strategy and its cardinality, respectively. Let V_i be an unobserved node in $O^{(k)}$, and only V_i becomes observed in a new observation strategy $O^{(k+1)}$ with the observation statuses of other nodes remaining unchanged. Let $f(O^{(k+1)})$ denote the cardinality of $O^{(k+1)}$. Then the identifiability gain of observing V_i, denoted by $g(V_i, O^{(k)})$, is calculated as $g(V_i, O^{(k)}) = f(O^{(k+1)}) - f(O^{(k)})$. □

By definition, $g\left(V_i, O^{(k)}\right)$ is the difference in cardinality between two consecutive observation strategies $O^{(k)}$ and $O^{(k+1)}$. That is, after V_i becomes observed in $O^{(k+1)}$, we need to find out the number of newly added non-redundant identifiability equations. First, if another node V_j $\left(i \neq j\right)$ is observed in both $O^{(k)}$ and $O^{(k+1)}$ and there exists a Wright's path [46] of length 1 connecting V_i and V_j, it can be shown that the newly added identifiability equation, denoted by $IE\left(V_i, V_j\right)$, is non-redundant (see Lemma 1 and Supplementary Materials for theoretical justification). However, if the length of every Wright's path between V_i and V_j is greater than 1, the identifiability equation $IE\left(V_i, V_j\right)$ is not always redundant, and it depends on both the node's observation status and the structure of the network. Here we introduce the concept of detour-path before we further elucidate the redundancy issue. Consider a DAG $\mathrm{G} = \left(\mathbf{V}, \mathbf{E}\right)$ and two *d*-connected observed nodes V_i and V_j. Assume that there exists a Wright's path P_{ji} between V_i and V_j as well as an observed node $V_k\left(k \neq i, j\right)$ on P_{ji}, and the direction of P_{ji} is from V_i to V_k and then to V_j. Now let P_{ki} and P_{jk} denote the two segments of P_{ji}, then P_{ki} entering node V_k has an arrow pointing into V_k while P_{jk} exiting node V_k has an arrow pointing away from V_k. However, if there exists another Wright's path between V_k and V_j, denoted by $\tilde{P}_{kj}$, which has no any other observed nodes besides V_k and V_j and has an arrow pointing into V_k, then V_k is a collider with respect to P_{ki} and $\tilde{P}_{kj}$. Thus, we call the Wright's path segment P_{jk} the detour-path, and call V_i, V_j and V_k the upstream node, the downstream node, and the collider node of the detour-path P_{jk}, respectively. By definition, a detour-path can have only one downstream node

and one collider node but may have one or more upstream nodes. Moreover, multiple detour-paths can share the same upstream node, the same downstream node or the same collider node. Several examples are given in Fig. 1 to illustrate the concept of detour-path.

In addition, when an upstream node V_i is shared by two or more detour-paths that have the same downstream node, V_i is called a shared upstream node; otherwise, V_i is called an exclusive upstream node. Note that a detour-path can have both exclusive and shared upstream nodes in the same time, and the collider node of one detour-path can be an upstream node of another detour-path. Consider two detour-paths that have no exclusive upstream nodes, if they share the same downstream node and at least one upstream node, or one upstream node of one detour path is the collider node of the other detour-path, then two detour-paths are intersecting. One can tell that if P_{jk_1} intersects with P_{jk_2} and P_{jk_2} intersects with P_{jk_3}, then P_{jk_1} also intersects with P_{jk_3}. Then we consider a downstream node V_j, let S_IDP denote all the intersecting detour-paths, and let S_SUN denote all the shared upstream nodes of S_IDP. Similar to a single unknown parameter, the coefficient product $WP = \prod_{edge_l} \theta_l$ of a Wright's path P can be deemed as a single parameter and one can tell its structural identifiability based on identifiability equations. If a detour-path P has at least one exclusive upstream node, then the Wright's coefficient WP of P is globally identifiable (see Lemma 2 and Supplementary Materials for theoretical justification). Also, for a group of intersecting detour-paths, if the node number of S_SUN is equal to or greater than the number

of intersecting detour-paths in S_IDP, then the Wright's coefficient of each detour-path in S_IDP is globally identifiable (see Lemma 3 and Supplementary Materials for theoretical justification).

Given a DAG $\mathrm{G}=(\mathbf{V},\mathbf{E})$, consider two observed nodes V_i, V_j and an unobserved node V_u. V_u may not be on any Wright's paths between V_i and V_j. For this case，if only V_u becomes observed in $O^{(k+1)}$, then for each observed node V_i in $O^{(k)}$, one can check whether the identifiability equation $IE(V_i,V_u)$ is redundant according to Lemma 4 (see Supplementary Materials for theoretical justification). That is, when none of the Wright's paths between V_i and V_u contains detour-paths, $IE(V_i,V_u)$ is redundant if and only if each Wright's path between V_i and V_u passes at least one observed node other than V_i and V_u; otherwise, $IE(V_i,V_u)$ is redundant if and only if the Wright's coefficient of each detour-path between V_i and V_u is globally identifiable in $O^{(k)}$ and each Wright's path between V_i and V_u passes at least one observed node other than V_i and V_u. If V_u is on a Wright's path between V_i and V_j, and the sufficient and necessary condition for one of the identifiability equations $IE(V_i,V_u)$ and $IE(V_j,V_u)$ being redundant is similar to Lemma 4 and given in Lemma 5 (see Supplementary Materials for theoretical justification). Note that it can be determined whether the Wright's coefficient of a detour-path is globally identifiable according to Lemma 2 and Lemma 3.

Based on Lemma 4 and Lemma 5, we propose a novel graphic method to calculate the identifiability gain $g(V_i,O^{(k)})$. Let des_i denote the descendant node set of V_i, anc_i denote the ancestor node set of V_i, rel_i denote the set of nodes that

are not included in des_i or anc_i. Moreover, let $bound_i \subset anc_i$ denote the boundary node set, in which every node has at least one outgoing edge to a node in rel_i. Then we can calculate $g\left(V_i, O^{(k)}\right)$ by removing the following edges from the original graph G: i) all the incoming edges to the observed nodes that are not collider nodes of detour-paths in anc_i; ii) all the outgoing edges from some observed nodes in des_i and rel_i and these observed nodes are not the collider nodes of the detour-paths whose Wright's coefficients are unidentifiable in $O^{(k)}$; and iii) all the outgoing edges from the observed nodes in $bound_i$ to nodes in rel_i, and then we get a new graph denoted by G'. Let N_w denote the total number of the observed nodes that are connected with V_i via any Wright's path in graph G'. Furthermore, one can tell from the edge-removal operation that there still exist some redundant identifiability equations in G', because the following two types of redundancy cases have not been considered in the edge-removal operation: V_i is the downstream node of an arbitrary detour-path, and V_i is on a Wright's path between two observed nodes in G'. Let N_r denote the number of redundant identifiability equations in G'. According to the topological structure of G' and the node's observation status, we can obtain N_r based on Lemma 4 and Lemma 5 (see the details in Implementation and Verification Section). It can be shown that the identifiability gain is $g\left(V_i, O^{(k)}\right) = N_w - N_r$ (see Theorem 1 and Supplementary Materials for theoretical justification).

For a given DAG G and an observation strategy $O^{(k)}$, different unobserved nodes may associate with different identifiability gains. Naturally, our strategy is to

choose the unobserved node in $O^{(k)}$ with the maximum identifiability gain if it becomes observed in $O^{(k+1)}$. However, we also recognize that, to assure that all model parameters are at least locally identifiable, certain nodes of a DAG must be observed if they are unobserved in an observation strategy (see Lemma 6 and Supplementary Materials for theoretical justification). For convenience, we call such nodes the must-be-observed [14] nodes, and let $O^{(0)_M}$ denote the observation strategy, in which only the MBO nodes are observed.

Lemma 1. Given a DAG $\mathrm{G}=(\mathbf{V},\mathbf{E})$, an observed node V_i, and an unobserved node V_u in $O^{(k)}$, if only V_u becomes observed in $O^{(k+1)}$, the identifiability equation $IE(V_i,V_u)$ is non-redundant if there exists a Wright's path of length 1 connecting V_i and V_u.

Lemma 2. If a detour-path P has one or more exclusive upstream node, the Wright's coefficient WP of P is globally identifiable.

Lemma 3. For a group of intersecting detour-paths, if the number of the shared upstream nodes in S_SUN is equal to or greater than the number of intersecting detour-paths in S_IDP, then the Wright's coefficient of each detour-path in S_IDP is globally identifiable.

Lemma 4. Given a DAG $\mathrm{G}=(\mathbf{V},\mathbf{E})$, an observed node V_i, and an unobserved node V_u in $O^{(k)}$, if only V_u becomes observed in $O^{(k+1)}$, there exist two cases:

1) each Wright's path between V_i and V_u passes at least one observed node other than V_i and V_u when none of the Wright's paths between V_i and V_u contains detour-paths;

2) each Wright's path between V_i and V_u passes at least one observed node other than V_i and V_u, and the Wright's coefficient of each detour-path between V_i and V_u is globally identifiable in $O^{(k)}$ when certain Wright's paths between V_i and V_u contain detour-paths.

Then the identifiability equation $IE(V_i, V_u)$ is redundant if and only if one of the above conditions holds.

Lemma 5. Given a DAG $G = (\mathbf{V}, \mathbf{E})$, two *d*-connected observed nodes V_i and V_j, and an unobserved node V_u in $O^{(k)}$, if V_u is on a Wright's path between V_i and V_j and only V_u becomes observed in $O^{(k+1)}$, there exist two cases:

1) each Wright's path between V_i and V_j passes at least one observed node other than V_i and V_j when none of the Wright's paths between V_i and V_j contains detour-paths;
2) each Wright's path between V_i and V_j passes at least one observed node other than V_i and V_j, and the Wright's coefficient of each detour-path between V_i and V_j is globally identifiable in $O^{(k)}$ when certain Wright's paths between V_i and V_j contain detour-paths.

Then one of the two identifiability equations $IE(V_i, V_u)$ and $IE(V_j, V_u)$ is redundant if and only if one of the above conditions holds.

Theorem 1. Given a DAG $G = (\mathbf{V}, \mathbf{E})$ and an unobserved node V_i in an observation strategy O, let G' denote the sub-graph after the edge-removal operation. Then the identifiability gain is $g(V_i, O) = N_w - N_r$, where N_w denotes the total number of the observed nodes that are connected with V_i via any Wright's

path in graph G', and N_r denotes the number of redundant identifiability equations in graph G'.

Lemma 6. For a given DAG $\mathrm{G}=(\mathbf{V},\mathbf{E})$, the following nodes must be observed to assure that all the parameters of the corresponding SEM are at least locally identifiable

1) The nodes with an out-degree 0;
2) The nodes with an out-degree 1;
3) The nodes with an in-degree 0 and an out-degree less than 3.

Dynamic programming strategy

Let $O^{(0)_G}$ denote a given observation strategy. If some of the MBO nodes are not observed in $O^{(0)_G}$, $O^{(0)_M}$ should be incorporated into $O^{(0)_G}$ according to Lemma 6. Therefore, the initial observation strategy, denoted by $O^{(0)}$, should always be $O^{(0)}=\left(O^{(0)_M}\mid O^{(0)_G}\right)$, where the OR operator is an element-wise operation. For example, for a DAG with 6 nodes, if $O^{(0)_M}=[1\ \ 0\ \ 1\ \ 0\ \ 0\ \ 0]^T$ and $O^{(0)_G}=[0\ \ 1\ \ 1\ \ 0\ \ 0\ \ 0]^T$, then $O^{(0)}=[1\ \ 1\ \ 1\ \ 0\ \ 0\ \ 0]^T$.

The dynamic programming strategy starts with the calculation of the cardinality of $O^{(0)}$ (that is, $f\left(O^{(0)}\right)$) based on Theorem 1. Specifically, let R be the number of observed nodes in $O^{(0)}$, $V_{o\_r}\left(r=1,2,\cdots,R\right)$ be the *r*-th observed node in $O^{(0)}$, and $O^{(0)}\left\{V_{o\_1},...,V_{o\_r}\right\}$ be the observation strategy in which only the first *r* observed nodes in $O^{(0)}$ are observed. Then $f\left(O^{(0)}\right)=\sum_{r=1}^{R-1}g\left(V_{o\_(r+1)},O^{(0)}\left\{V_{o\_1},...,V_{o\_r}\right\}\right)$ can be calculated according to Theorem 1. Note that the order at which V_{o_r} is selected into

$O^{(0)}\left\{V_{o\_1},...,V_{o\_R}\right\}$ will not change the observation strategy (e.g., $O^{(0)}\left\{V_{o\_1},V_{o\_2}\right\}=O^{(0)}\left\{V_{o\_2},V_{o\_1}\right\}$) and thus have no effect on the value of $f\left(O^{(0)}\right)$.

The second step of our dynamic programming strategy is to define stages and their associated states. Let S denote the number of unobserved nodes in $O^{(0)}$, and let V_{u_s} $\left(s=1,2,\cdots,S\right)$ denote the *s*-th unobserved node in $O^{(0)}$, then the dynamic programming procedure can be divided into $S+1$ stages. For illustration purpose, we consider a simple example with 5 unobserved nodes, as shown in Fig. 2. The 0-th stage is actually the initialization step as described in the previous paragraph, and it has only one state, i.e., $O^{(0)}$. At the first stage, there are $S=5$ different states; that is, only one of the unobserved nodes $\left\{V_{u\_1},V_{u\_2},\cdots,V_{u\_5}\right\}$ in $O^{(0)}$ will be selected to observe. At the second stage, since one of the five unobserved nodes has been selected at the previous stage, there are only four unobserved nodes for selection and thus four states exist (that is, $\left\{V_{u\_2},V_{u\_3},V_{u\_4},V_{u\_5}\right\}$). Therefore, as shown in Fig. 2, except for stages 0 and 1, each subsequent stage has one less states than its previous stage; also, the upper triangular region (see the area above the labels of stages 1-5 in Fig. 2) is empty because the selection order of unobserved nodes does not affect the eventual observation strategy so the inclusion of such states in the upper triangular region is redundant. One can tell that the proposed stage and state definitions satisfy the optimality principle of dynamic programming [47-49].

The third step is to compute the state transition costs for searching the optimal state transition path(s). According to the definitions of stages and states, there may exist several different states at the *s*-th stage that can transit to the same state at the

(*s*+1)-th stage. For instance, four states V_{u_1}, V_{u_2}, V_{u_3} and V_{u_4} at the first stage can transit to V_{u_5} at the second stage, as shown in Fig. 2. The state transition cost from state V_{u_i} to state V_{u_j} ($i \neq j$) between two consecutive stages is just the identifiability gain $g\left(V_{u\_j}, O^{(k)}\left\{...,V_{u\_i},...\right\}\right)$, where $O^{(k)}\left\{...,V_{u\_i},...\right\}$ means that V_{u_i} is observed in $O^{(k)}$. Then the cardinality of an observation strategy can be computed by adding $f\left(O^{(0)}\right)$ and all the state transition costs along the state transition path. Since the goal of the dynamic programming strategy is to search for the optimal observation strategies, when there exist multiple transition paths from state V_{u_i} in $O^{(k)}$ to state V_{u_j} in $O^{(k+1)}$ ($i \neq j$), the transition path associated with the maximum identifiability gain will be chosen; that is, $f\left(O_{V_{u\_j}}^{(k+1)}\right) = \max\limits_{V_{u\_i}, i\neq j}\left(g\left(V_{u\_j}, O_{V_{u\_i}}^{(k)}\right) + f\left(O_{V_{u\_i}}^{(k)}\right)\right)$, where $O_{V_{u\_i}}^{(k)}$ is a convenient notation for $O^{(k)}\left\{...,V_{u\_i},...\right\}$.

The dynamic programming strategy above can be mathematically described in Eq. (3), and we have implemented this strategy in R (see the "Implementation and verification" Section),

$$\begin{cases} f\left(O_{V_{u\_s}}^{(1)}\right) = f\left(O^{(0)}\right) + g\left(V_{u\_s}, O^{(0)}\right), \quad s = 1,2,...,S, \\ f\left(O_{V_{u\_j}}^{(k+1)}\right) = \max\limits_{V_{u\_i}, i\neq j}\left(g\left(V_{u\_j}, O_{V_{u\_i}}^{(k)}\right) + f\left(O_{V_{u\_i}}^{(k)}\right)\right), \quad k = 1,2,\cdots \ \& \ k \leq S-1. \end{cases} \tag{3}$$

It should be stressed that it is not necessary to finish all the S iterations as shown in Eq. (3). Once the cardinality $f\left(O_{V_{u\_i}}^{(k)}\right)$ at the *k*-th stage becomes equal to or greater than the number of unknown parameters N_u, the dynamic programming process will stop and we get the SIOOR strategies.

Results and Discussion

Overview of the framework

Observation strategy design is an under-investigated problem for biological networks, despite the fact that a variety of biological networks have been actively constructed and used in numerous benchside or bedside studies. However, the existence of latent variables is a common problem due to cost, technical or other limitations, and has significantly hampered our capability to quantitatively investigate and understand such networks via, e.g., key network parameter estimation from experimental data. Identifiability analysis has long been recognized as a powerful tool to assure the accuracy and reliability of parameter estimation techniques; however, identifiability-based observation strategy design for biological networks turns out to be an unexplored field despite its substantial importance to biological network studies like structure identification.

To the best knowledge of our authors, this is the first study that tackles the problem of identifiability-based observation strategy design for biological networks described by linear SEMs. First, we introduce several new concepts such as cardinality of observation strategy and identifiability gain and mathematically formulate the identifiability-based optimal observation problem. Second, for a given network structure, the key idea is to turn a minimum number of unobserved nodes in the original observation strategy into observed such that the number of non-redundant identifiability equations becomes greater than or equal to the number of unknown model parameters (i.e., the whole system is at least locally identifiable). By counting

the number of non-redundant identifiability equations, we avoid performing actual identifiability analysis on SEM and the proposed method is thus computationally efficient. Third, by defining the concepts of stage division and state transition, a dynamic programming strategy is proposed to solve the maximization problem without involving any time-consuming symbolic computation or matrix operations [33, 34]. Fourth, an efficient computing algorithm is proposed to calculate the identifiability gain of each unobserved node in a given observation strategy. More specifically, the computing process is significantly simplified by counting the number of observed nodes that connect with the node of concern via Wright's paths after removing certain edges from the original graph.

It takes a non-constant time to compute the node identifiability gain in each iteration, and the algorithm complexity depends on the number of observed nodes. Furthermore, the number of iterations of the dynamic programming algorithm does not depend on the total number of nodes, but the number of unobserved nodes in the original observation strategy. Let S denote the number of unobserved nodes and T denote the number of observed nodes in the original observation strategy, then the computation complexity of the dynamic programing strategy is $O\left(S^2 \cdot T\right)$.

Implementation and Verification

The flowchart of the proposed algorithm for searching the structural identifiability-based optimal observation remedy is shown in Fig. 3. We have implemented the dynamic programming algorithm in R, and all the source codes and

examples are freely accessible at https://github.com/Hongyu-Miao/SIOOR.

Here we describe several important technical details of the implementation. First, at the state transition step, i.e., $f\left(O_{V_{u\_j}}^{(k+1)}\right)=\max_{V_{u\_i},i\neq j}\left(g\left(V_{u\_j},O_{V_{u\_i}}^{(k)}\right)+f\left(O_{V_{u\_i}}^{(k)}\right)\right)$, if there exist multiple transitions that produce the same $f\left(O^{(k+1)}\right)$, our current implementation chooses only one such transition to update the next-stage observation strategy. If needed, the R code can be slightly modified to enumerate all optimal observation strategies. Second, since the boundary node set is just a subset of the ancestor node set for a given node, the processing of the boundary nodes is incorporated into the processing of the ancestor nodes in the current implementation.

In order to verify the implementation, synthetic DAGs can be generated for this purpose, like the two DAG examples in Fig. 4. The first DAG contains 8 nodes and 13 edges, and it has only a single input node and a single output node. Moreover, the first example considers a special initial observation strategy (i.e., all nodes are unobserved) to illustrate the capability of the proposed method to design optimal observation strategy from scratch. The second DAG has multiple input and output nodes, and it considers a more general situation, that is, there exists both observed and unobserved nodes in the initial observation strategy. We analyzed the two examples using the proposed algorithm, and used the identifiability matrix method [34] to verify that the obtained observation remedies do grant (local) identifiability to all model parameters .

Applications to real biological networks

Since it is impossible to cover all the biological networks in various databases and knowledge repositories [16, 17] in one study, we choose the biological network associated with influenza A virus [39] as an application example for illustration purpose. IAV can infect birds as well as mammals including human, and it has been one of the major infectious pathogens that have caused millions of human deaths. It is thus of great scientific significance to systematically understand IAV infection and immune response mechanisms. Therefore, Matsuoka et al. [50] manually curated a comprehensive database, called FluMap, for depicting the influenza virus life cycle at the molecular level from over 500 previous publications. There are mainly five modules in FluMap: virus entry, virus replication and transcription, post-translational processing, transportation of virus proteins, and packaging and budding. Given the critical role of virus replication in influenza virus life cycle, numerous experimental studies (e.g., [42, 51-53]) have made attempts to understand virus replication mechanisms and their clinical implications. Thus, we choose to focus on the IAV replication module and analyze its observation strategy.

Since IAV replication involves many different biomolecules and complex interactions, it is usually infeasible to observe all such components and their interactions in one study. The question of concern here is how to choose a minimal number of nodes in Fig. 5 to observe such that all the model parameters become at least locally identifiable. Note that Fig. 5 is derived from Matsuoka's work [50], and consists of 22 nodes and 26 edges; for simplicity, the catalyzers and inhibitors in this network are treated as reactants.

A relevant concept, called observability, has been previously investigated by Liu et al. [38] for complex dynamic systems. Although observability analysis also deals with observation strategies considering the existence of latent variables, it is very different from identifiability analysis in two aspects: 1) the focus of observability analysis is not model parameters but how to infer the unobserved state variables from experimentally measured outputs of a system; 2) the graphical approach proposed by Liu et al. was developed for the so-called balance equations based on mass-action kinetics, the model structures of which are very different from static linear SEMs. However, it is of interest to compare the identifiability-based observation results with those of the observability-based method. For this purpose, we assume that all the nodes in Fig. 5 are initially unobserved. After applying the proposed dynamic programing method, we get the optimal observation strategy shown in Fig. 6(a) for achieving parameter identifiability. The optimal observation strategy produced by Liu's observability approach is shown in Fig. 6(b). According to Figures 6(a) and 6(b), one can tell that the identifiability-based observation strategy contains 20 observed nodes and 2 unobserved nodes, while the observability-based strategy contains 3 observed nodes and 19 unobserved nodes. That is, for the IAV replication module, the system internal states can be inferred from a few observed output nodes if a balance equation model is used; however, it needs much more observed nodes to achieve parameter identifiability if a linear SEM is used. Such an observation is not only due to the different goals of observability and identifiability analyses, but also the differences in the underlying model structures used in observability or identifiability

analyses.

Moreover, besides the nodes with an out-degree 0 or 1 as mentioned in Lemma 3, the identifiability-based observation strategy is also likely to select the nodes with a high out-degree as unobserved nodes; for instance, the two unobserved nodes viral_RNA and NP(ub) in Fig. 6(a) have the highest out-degrees 2 and 3, respectively. This is because, if an unobserved node has a high out-degree, this node is connected with many out-neighbor nodes; when its out-neighbor nodes are observed, there will exist multiple Wright's paths that connect such out-neighbor nodes and pass this unobserved node, and the corresponding identifiability equations thus contain the parameters associated with the out-edges of this unobserved node such that these parameters can be identifiable. Interestingly, the observability-based strategy tends to select the nodes with a low out-degree as observed nodes, for example, all the nodes with 0 out-degree are observed in Fig. 6(b). It is because the nodes with an out-degree 0 in a DAG are usually the final products of chemical reactions, instead of reactants, and thus the internal states associated with other nodes can be easily inferred based on the balance equations if all the final products of chemical reactions are measured.

Conclusions

In this study, we address an important problem for biological networks: the design of observation strategies for all edge coefficients being identifiable. Linear SEMs are used as the mathematical representation of biological networks, which allows us to formulate the problem as a constrained optimization problem. A dynamic

programming strategy was then developed to solve the constrained optimization problem to obtain the optimal observation strategies at the cost of turning a minimal number of unobserved nodes into observed. The proposed solution is novel and efficient because it avoids both symbolic computation and matrix operations as used in other studies, and we provided necessary theoretical justifications for the proposed algorithm. As verified by multiple examples (synthetic or real networks), the proposed solution is generic and can be applied to an arbitrary DAG (recursive SEMs) without bidirectional edges.

We also recognize that many real biological networks are dynamic, nonlinear, or have feedback loops, which are beyond the capability of the method developed in this study. However, this study provides a basis for determining the identifiability-based optimal observation remedy for more complex biological networks, and we expect to tackle the more challenging problems in the future.

Declarations

List of abbreviations

SEM: structure equation model;

DAG: directed acyclic graph;

SIOOR: structural identifiability-based optimal observation remedy;

IAV: influenza A virus;

DP: dynamic programming;

OS: Observation Strategy;

IG: Identifiability Gain;

DCG: directed cyclic graph.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The network structure data used in this study are all selected from public literature, including the FluMap database [50].

Competing interests

The authors declare no competing interests.

Funding

This work was partially supported by the Fundamental Research Funds for the Central Universities of China ZYGX2014J064 (YW) and NSF grant 1620957 (HM).

Authors' contributions

YW contributed to method development, computational analyses, real network analyses and manuscript writing. HM proposed the idea, oversaw the study, and significantly contributed to manuscript preparation. All authors have read and approved the final version of the manuscript.

Acknowledgements

The authors thank Dr. Yu Luo and Ms. Lijie Wang for useful suggestions and discussions.

Author's information

YW is Assistant Professor at School of Computer Science and Engineering, University of Electronic Science and Technology of China. HM is Associate Professor at the Department of Biostatistics, School of Public Health, University of Texas Health Science Center at Houston, USA.

Endnotes

Not applicable.

Additional files

Additional file 1: Theoretical justifications for identifiability gain computation.

References

1. Butcher EC, Berg EL, Kunkel EJ. Systems biology in drug discovery. Nat Biotech. 2004;22(10):1253-9.
2. Barabasi A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56-68.
3. Seiple IB, Zhang Z, Jakubec P, Langlois-Mercier A, Wright PM, Hog DT, et al. A platform for the discovery of new macrolide antibiotics. Nature. 2016;533(7603):338-45.
4. Bansal M, Yang J, Karan C, Menden MP, Costello JC, Tang H, et al. A community computational challenge to predict the activity of pairs of compounds. Nat Biotech. 2014;32(12):1213-22. doi: 10.1038/nbt.3052
5. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005;437. doi: 10.1038/nature04209.
6. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, et al. A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. Cell. 2005;122(6):957-68. doi: 10.1016/j.cell.2005.08.029.
7. Carninci P, Kasukawa T, Katayama S, Gough J, Frith M, Maeda N. The transcriptional landscape of the mammalian genome. Science. 2005;309. doi: 10.1126/science.1112014.
8. Minguez P, Parca L, Diella F, Mende DR, Kumar RD, Helmercitterich M, et al. Deciphering a global network of functionally associated post-translational modifications. Molecular Systems Biology. 2012;8(1):599-.
9. Minguez P, Letunic I, Parca L, Bork P. PTMcode: a database of known and predicted functional associations between post-translational modifications in proteins. Nucleic Acids Res. 2013;41. doi: 10.1093/nar/gks1230.
10. Liu Z, Wu H, Zhu J, Miao H. Systematic identification of transcriptional and post-transcriptional regulations in human respiratory epithelial cells during influenza A virus infection. BMC Bioinformatics. 2014;15(1):336-.
11. Reynolds A, Leake D, Boese Q, Scaringe S, Marshall W, Khvorova A. Rational siRNA design for RNA interference. Nature Biotechnology. 2004;22(3):326-30.
12. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120(1):15-20.
13. Ponting CP, Oliver PL, Reik W. Evolution and Functions of Long Noncoding RNAs. Cell. 2009;136(4):629-41.
14. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. Nature. 2000;407. doi: 10.1038/35036627.
15. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(6):1777-82.
16. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, et al.

Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489(7414):91-100.
17. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic acids research. 2014;42(D1):D199-D205.
18. Miao H, Xia X, Perelson AS, Wu H. On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM review. 2011;53(1):3-39.
19. Giraud C, Tsybakov A. Discussion: Latent variable graphical model selection via convex optimization. The Annals of Statistics. 2012;40(4):1984-8.
20. Shamaiah M, Lee SH, Vikalo H. Graphical Models and Inference on Graphs in Genomics: Challenges of high-throughput data analysis. IEEE Signal Processing Magazine. 2012;29(1):51-65. doi: 10.1109/MSP.2011.943012.
21. Domke J. Learning graphical model parameters with approximate marginal inference. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2013;35(10):2454-67.
22. Mazman SG, Usluel YK. Modeling educational usage of Facebook. Computers in Education. 2010;55(2):444-53.
23. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120. doi: 10.1016/j.cell.2004.12.035.
24. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci. 2007;104. doi: 10.1073/pnas.0610772104.
25. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell. 2009;136. doi: 10.1016/j.cell.2009.02.006.
26. Minguez P, Parca L, Diella F, Mende DR, Kumar R, Helmer‐Citterich M. Deciphering a global network of functionally associated post‐translational modifications. Mol Syst Biol. 2012;8.
27. Cai XBJ, Giannakis GB. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. PLoS computational biology. 2013;9. doi: 10.1371/journal.pcbi.1003068.
28. Pearl J. Causality: models, reasoning, and inference (2nd Edition). Cambridge: Cambridge University Press; 2009.
29. Brito C, Pearl J. Generalized instrumental variables. Uncertainty in Artificial Intelligence2002. p. 85-93.
30. Tian J. A criterion for parameter identification in structural equation models. arXiv preprint arXiv:12065289. 2012.
31. Drton M, Foygel R, Sullivant S. Global identifiability of linear structural equation models. The Annals of Statistics. 2011:865-86.
32. Foygel R, Draisma J, Drton M. Half-trek criterion for generic identifiability of linear structural equation models. The Annals of Statistics. 2012;40(3):1682-713.
33. Sullivant S, Garcia-Puente LD, Spielvogel S, editors. Identifying causal effects with computer algebra. Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI) AUAI Press; 2010.

34. Wang Y, Lu N, Miao H. Structural identifiability of cyclic graphical models of biological networks with latent variables. BMC Systems Biology. 2016;10(1):1-15. doi: 10.1186/s12918-016-0287-y.
35. Kreutz C, Timmer J. Systems biology: experimental design. FEBS Journal. 2009;276(4):923-42.
36. Marvel S, Williams CM. Set membership experimental design for biological systems. BMC Systems Biology. 2012;6(1):21-.
37. Liu AR, Bitmead RR. Stochastic observability in network state estimation and control. Automatica. 2011;47(1):65-78.
38. Liu Y, Slotine JE, Barabasi A. Observability of complex systems. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(7):2460-5.
39. Pirsiavash H, Ramanan D, Fowlkes CC, editors. Globally-optimal greedy algorithms for tracking a variable number of objects. computer vision and pattern recognition; 2011.
40. Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. A linear non-Gaussian acyclic model for causal discovery. The Journal of Machine Learning Research. 2006;7:2003-30.
41. Hoyer PO, Hyvarinen A, Scheines R, Spirtes PL, Ramsey J, Lacerda G, et al. Causal discovery of linear acyclic models with arbitrary distributions. arXiv preprint arXiv:12063260. 2012.
42. Watanabe T, Kiso M, Fukuyama S, Nakajima N, Imai M, Yamada S, et al. Characterization of H7N9 influenza A viruses isolated from humans. Nature. 2013;501(7468):551-5. doi: 10.1038/nature12392
43. Wright S. The method of path coefficients. The Annals of Mathematical Statistics. 1934;5(3):161-215.
44. Wright S. Path coefficients and path regressions: alternative or complementary concepts? Biometrics. 1960;16. doi: 10.2307/2527551.
45. Garcia C, Li T. On the Number of Solutions to Polynomial Systems of Equations. SIAM Journal on Numerical Analysis. 1979.
46. Sullivant S, Talaska K, Draisma J. Trek separation for Gaussian graphical models. Annals of Statistics. 2010;38(3):1665-85.
47. Felzenszwalb PF, Zabih R. Dynamic Programming and Graph Algorithms in Computer Vision. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2011;33(4):721-40. doi: 10.1109/TPAMI.2010.135.
48. Tran D, Yuan J, Forsyth D. Video Event Detection: From Subvolume Localization to Spatiotemporal Path Search. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014;36(2):404-16. doi: 10.1109/TPAMI.2013.137.
49. Jiang H, Tian T, Sclaroff S. Scale and Rotation Invariant Matching Using Linearly Augmented Trees. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015;37(12):2558-72.
50. Matsuoka Y, Matsumae H, Katoh M, Eisfeld AJ, Neumann G, Hase T, et al. A comprehensive map of the influenza A virus replication cycle. BMC systems biology. 2013;7(1):97.

51. Honda A, Mizumoto K, Ishihama A. Minimum molecular architectures for transcription and replication of the influenza virus. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(20):13166-71.
52. Konig R, Stertz S, Zhou Y, Inoue A, Hoffmann HH, Bhattacharyya S, et al. Human host factors required for influenza virus replication. Nature. 2010;463(7282):813-7.
53. York A, Hutchinson E, Fodor E. Interactome analysis of the influenza A virus transcription/replication machinery identifies protein phosphatase 6 as a cellular factor required for efficient virus replication. Journal of Virology. 2014;88(22):13284-99.

Figure List

Figure 1. Four examples for illustrating the detour-path concept, where observed nodes are colored green and unobserved nodes are colored grey. (a) A simple detour-path; (b) Two detour-paths P_{jk}, P_{nk} share the same collider node V_k and upstream nodes V_i, V_o; (c) Two detour-paths P_{jk}, P_{nk} share the same collider node V_k, upstream nodes V_i, V_m and edge e_{jk}; (d) Two detour-paths P_{jk}, P_{nk} share the same upstream node V_i and downstream node V_j.

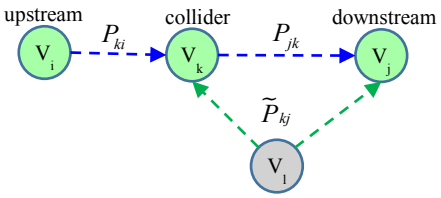
Figure 2. Schematic illustration of the stages, states and state transition costs in the proposed dynamic programming strategy.

Figure 3. Flowchart of the proposed dynamic programming algorithm.

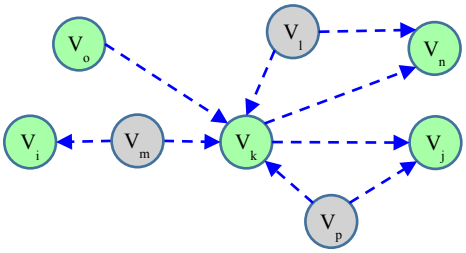
Figure 4. Two DAG examples for algorithm implementation validation, where the green nodes are unobserved and the blue ones are observed in the initial observation strategy. (a) A DAG with a single input and a single output; (b) A DAG with multiple inputs and multiple outputs.

Figure 5. An application example based on the influenza A virus replication module, where all nodes are initially unobserved and in green color.

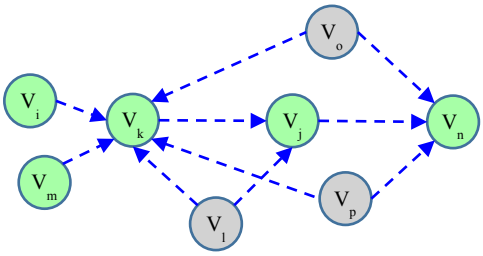
Figure 6. The optimal observation strategies for the influenza A virus replication module based on (a) identifiability and (b) observability, where the yellow nodes are observed and the green nodes are unobserved.



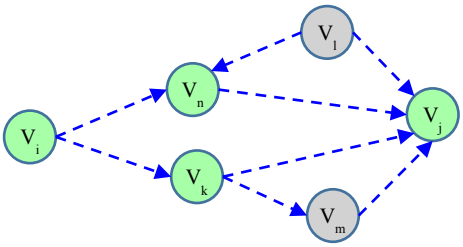
(a) A simple detour-path.



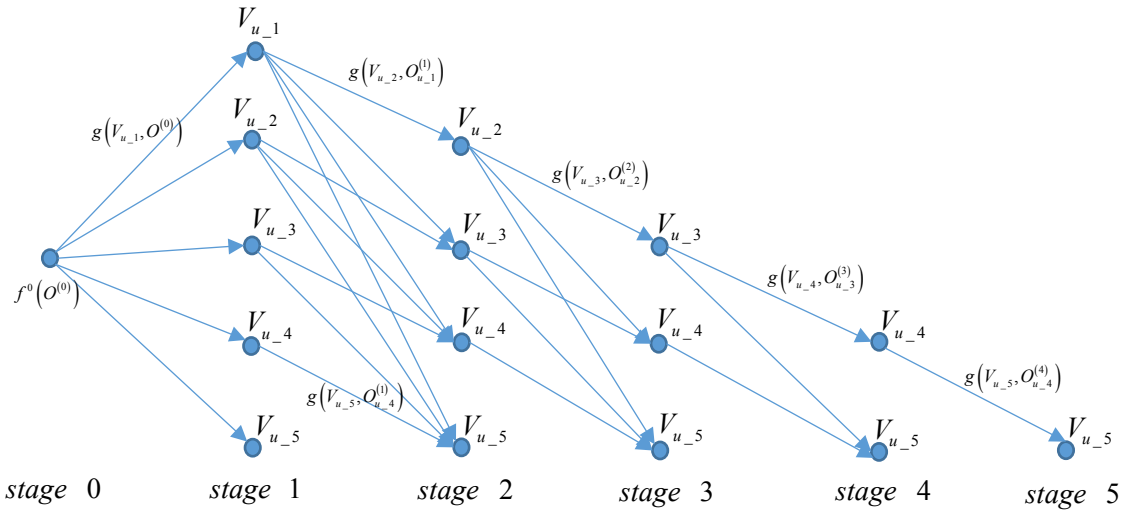
(b) Two detour-paths P_{jk}, P_{nk} share the same collider node V_k and upstream nodes V_i, V_o .

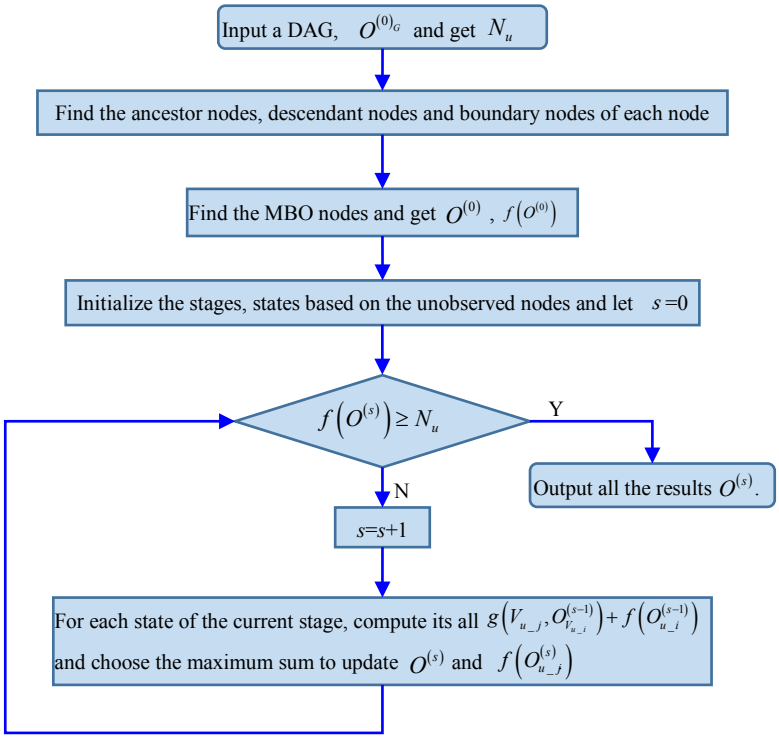


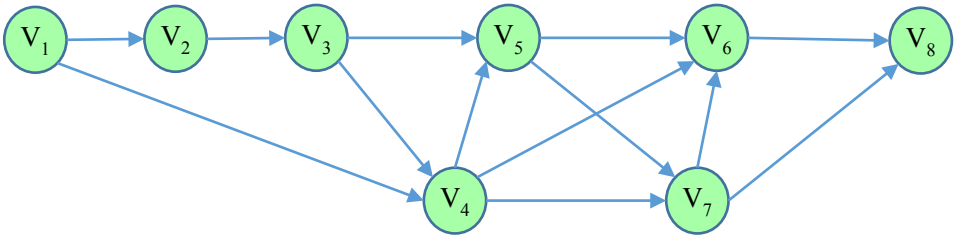
(c) Two detour-paths P_{jk}, P_{nk} share the same collider node V_k, upstream nodes V_i, V_m and edge e_{jk}



(d) Two detour-paths P_{jk}, P_{nk} share the same upstream node V_i and downstream node V_j.

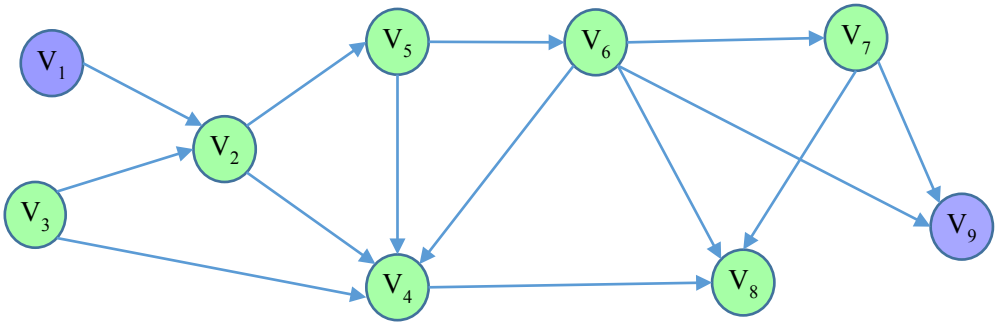

V_{u_1}
V_{u_2}
V_{u_3}
V_{u_4}
V_{u_5}
$g\left(V_{u\_1}, O^{(0)}\right)$
$f^0\left(O^{(0)}\right)$
$g\left(V_{u\_2}, O_{u\_1}^{(1)}\right)$
$g\left(V_{u\_5}, O_{u\_4}^{(1)}\right)$
$g\left(V_{u\_3}, O_{u\_2}^{(2)}\right)$
$g\left(V_{u\_4}, O_{u\_3}^{(3)}\right)$
$g\left(V_{u\_5}, O_{u\_4}^{(4)}\right)$
stage 0
stage 1
stage 2
stage 3
stage 4
stage 5


Input a DAG, $O^{(0)_G}$ and get N_u
Find the ancestor nodes, descendant nodes and boundary nodes of each node
Find the MBO nodes and get $O^{(0)}$, $f(O^{(0)})$
Initialize the stages, states based on the unobserved nodes and let $s=0$
$f(O^{(s)}) \geq N_u$
Y
Output all the results $O^{(s)}$.
N
$s=s+1$
For each state of the current stage, compute its all $g(V_{u\_j}, O^{(s-1)}_{V_{u\_i}}) + f(O^{(s-1)}_{u\_i})$
and choose the maximum sum to update $O^{(s)}$ and $f(O^{(s)}_{u\_j})$



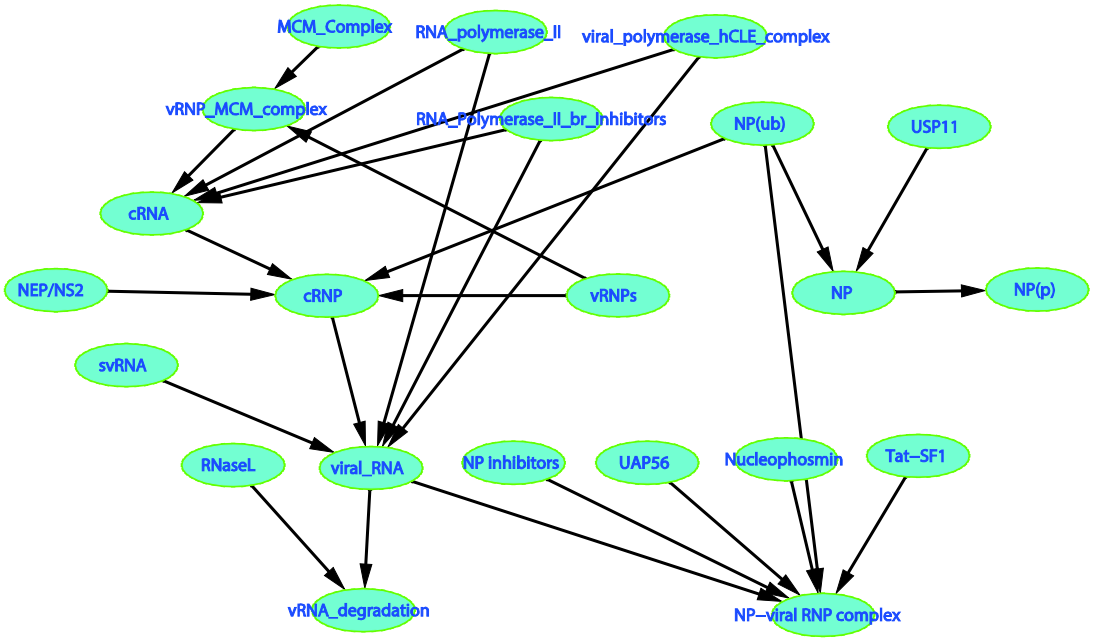
The optimal observation remedy is: V_1,V_2,V_4,V_6,V_8 and any two nodes of V_3, V_5,V_7

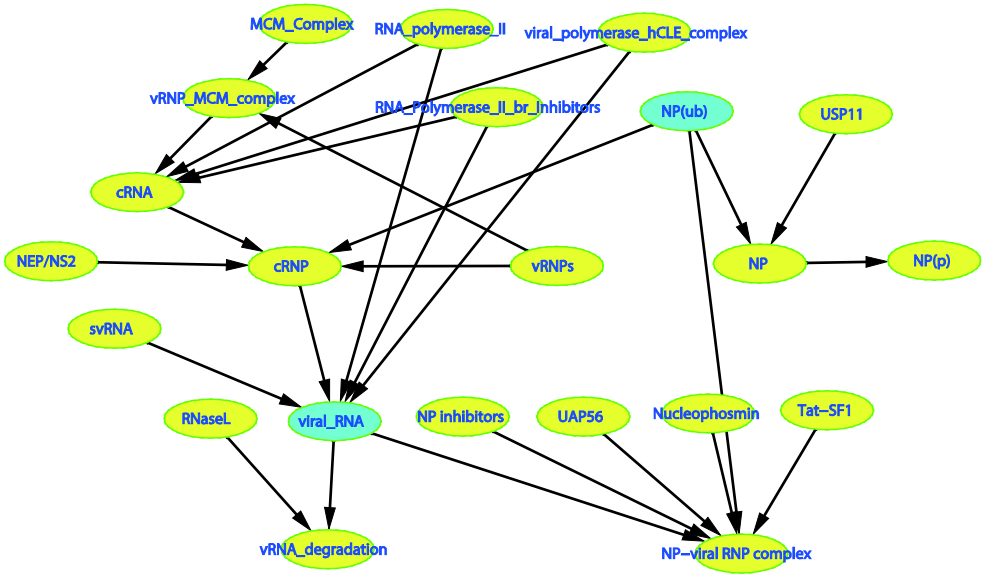
(a) A DAG with a single input and a single output



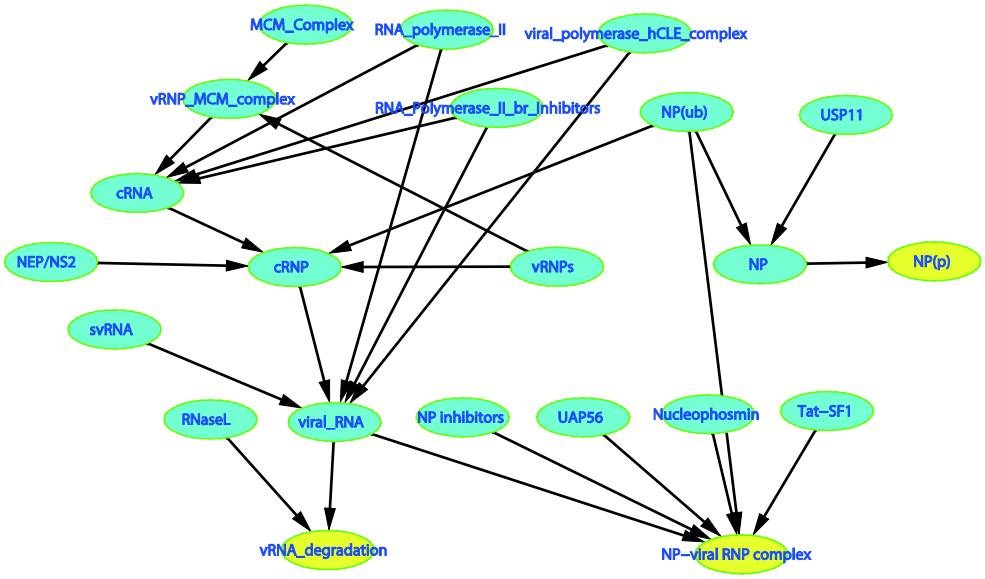
The optimal observation remedy is: V_1,V_2,V_3,V_4,V_8, V_9 and any one node of V_5, V_6,V_7

(b) A DAG with multiple inputs and multiple outputs


MCM_Complex
RNA_polymerase_II
viral_polymerase_hCLE_complex
vRNP_MCM_complex
RNA_Polymerase_II_br_Inhibitors
NP(ub)
USP11
cRNA
NEP/NS2
cRNP
vRNPs
NP
NP(p)
svRNA
RNaseL
viral_RNA
NP Inhibitors
UAP56
Nucleophosmin
Tat–SF1
vRNA_degradation
NP–viral RNP complex



(a) Identifiability-based optimal observation strategy.



(b) Observability-based optimal observation strategy.