Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 39 (2006) 126–146
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

A control study to evaluate a computer-based microarray experiment design recommendation system for gene-regulation pathways discovery

Changwon Yoo [a,*], Gregory F. Cooper [b], Martin Schmidt [c]

[a] Department of Computer Science, University of Montana, 420 Social Sciences, University of Montana, Missoula, MT 59803, USA
[b] Center for Biomedical Informatics, University of Pittsburgh 8084 Forbes Tower, 200 Lothrop St., Pittsburgh, PA 15213, USA
[c] Department of Molecular Genetics and Biochemistry, University of Pittsburgh W1247 BST, Pittsburgh, PA 15213, USA

Received 12 January 2005
Available online 19 September 2005

Abstract

The main topic of this paper is evaluating a system that uses the expected value of experimentation for discovering causal pathways in gene expression data. By experimentation we mean both interventions (e.g., a gene knock-out experiment) and observations (e.g., passively observing the expression level of a "wild-type" gene). We introduce a system called GEEVE (causal discovery in Gene Expression data using Expected Value of Experimentation), which implements expected value of experimentation in discovering causal pathways using gene expression data. GEEVE provides the following assistance, which is intended to help biologists in their quest to discover gene-regulation pathways:

- Recommending which experiments to perform (with a focus on "knock-out" experiments) using an expected value of experimentation (EVE) method.
- Recommending the number of measurements (observational and experimental) to include in the experimental design, again using an EVE method.
- Providing a Bayesian analysis that combines prior knowledge with the results of recent microarray experimental results to derive posterior probabilities of gene regulation relationships.

In recommending which experiments to perform (and how many times to repeat them) the EVE approach considers the biologist's preferences for which genes to focus the discovery process. Also, since exact EVE calculations are exponential in time, GEEVE incorporates approximation methods. GEEVE is able to combine data from knock-out experiments with data from wild-type experiments to suggest additional experiments to perform and then to analyze the results of those microarray experimental results. It models the possibility that unmeasured (latent) variables may be responsible for some of the statistical associations among the expression levels of the genes under study.

To evaluate the GEEVE system, we used a gene expression simulator to generate data from specified models of gene regulation. Using the simulator, we evaluated the GEEVE system using a randomized control study that involved 10 biologists, some of whom used GEEVE and some of whom did not. The results show that biologists who used GEEVE reached correct causal assessments about gene regulation more often than did those biologists who did not use GEEVE. The GEEVE users also reached their assessments in a more cost-effective manner.

Keywords: Causal discovery; Systems biology; Causal Bayesian networks; Microarray study design

* Corresponding author.
E-mail addresses: cwyoo@cs.umt.edu (C. Yoo), gfc@cbmi.upmc.edu (G.F. Cooper).

1532-0464/$ - see front matter © 2006 Published by Elsevier Inc.
doi:10.1016/j.jbi.2005.05.011

1. Introduction

Most research on causal discovery using causal networks has been based on using passive observational data [1–4]. There are limitations in learning causal relationships from observational data only. For example, if the generating process contains a latent factor (confounder) that influences two variables, it can be difficult, if not impossible, to learn the causal relationships between those two variables from observational data alone.

To uncover such causal relationships, a scientist generally needs to design a study that involves manipulating a variable (or variables) and then observing the changes (if any) in other variables of interest. In such an experimental study, one or more variables are manipulated and the effects on other variables are measured. On the other hand, *observational data* result from passive (i.e., non-interventional) measurement of some system, such as a cell. In general, both observational and experimental data may exist on a set of variables of interest. Limited time and funds restrict the number of variables that can be manipulated and the number of *experimental repeats* that can be collected for the control and experimental groups. For example, a molecular biologist who is interested in discovering the causal pathway of the genes involved in galactose metabolism first has to select the genes he or she is interested in understanding at a causal level. These genes are usually selected based on previously published results or by the molecular biologist's personal interest. Many issues are considered in determining the number of experimental repeats to obtain for each variable in the study design. Having more experimental repeats will typically tighten the statistical confidence intervals in the data analysis. Considering available time, budget, and other constraints, the biologist will make a decision about the number of experimental repeats to obtain.

Developing causal analysis methods is a key focus of several fields. In statistics, jointly with medicine, issues related to randomized clinical trials (RCTs) are studied, including methods for finding an optimal number of cases using stopping rules [5–7]. In molecular biology, developing techniques that generate efficient experimental designs for high throughput methods, such as cDNA microarrays, is gaining interest [8,9]. In artificial intelligence, techniques using graphical models have been used to model experimentation and have been applied to suggest the next experiment for causal discovery [10–12].

All these prior approaches have made contributions to efficient causal study design. They are not, however, sensitive to issues of limited resources and experimenter preferences. The research reported here is concerned with evaluating a decision-analytic system that addresses these issues in helping a biologist design and analyze studies of cellular pathways using high throughput sources of data. In particular, this paper concentrates on the design and analysis of cDNA microarray studies for uncovering gene regulation pathways. The fundamental methodology, however, is applicable to analyzing other high throughput data sources, such as the measurement of protein-levels, which is a rapidly developing area of biology.

Different tools have been developed to assist systems biology research using microarray data [4,10,11]. Unfortunately, there are limited studies that evaluate how useful these tools are to the biologists. In this paper, we provide an evaluation of the GEEVE system with 10 biologists, some of whom used GEEVE and some of whom did not.

In this section, we shortly provide background of gene array chips and give an overview of the GEEVE system.

1.1. Gene array chips

Three major gene-expression measurement technologies are currently available for measuring the expression levels of many genes at once. One is called a cDNA microarray, or simply *DNA microarray* [13]; another is called an *oligonucleotide array*, or GeneChip [14]; and a third technique is called serial analysis of gene expression (SAGE). We concentrate in this paper on the first two techniques, since they are high throughput methods, whereas SAGE is a more time consuming method. The DNA microarray technique uses user-definable probes[1] of DNA microarray, and the oligonucleotide array uses small oligonucleotide (usually 200 or 300 bases) as factory-built probes.

1.2. Problem description

A gene expression study using DNA microarrays usually involves two major steps. The first step typically consists of performing initial experiments to narrow the set of genes to study in more detail. The experimenter can avoid this first step if he or she already knows the specific set of genes of interest. Since the functions of many genes are not known, the first step is usually necessary. A number of microarrays will be assigned to hybridize with a pool of controlled cells and experimental cells. By examining the genes that are differentially expressed in these two groups of cells, the experimenter can decide which genes to study further. After choosing those genes, the experimenter has to produce an experimental design for further study how those genes are functionally related to each other.

2. GEEVE system

This chapter describes the issues related to the implementation of the GEEVE system (causal discovery in Gene Expression data using Expected Value of Experimentation). Tong and Koller [12] used a single-case approach to recommend to the experimenter the best possible pairwise relationship for further investigation. In gene expres-

[1] According to the nomenclature recommended by B. Phimister of *Nature Genetics*, a *probe* is the nucleic acid with known sequence, whereas a *target* is the free nucleic acid sample whose abundance level is being detected.

sion microarray studies, it may not be practical to perform one experiment at a time. Often it is more efficient to repeat a given experiment multiple times in parallel, rather than to repeat the experiment sequentially over time.

Tong and Koller [12] and Ideker et al. [11] used edge entropy loss functions to search for the next best experiment to perform. This approach can be useful when the experimenter is performing a first-phase study to select the genes without any preference toward the relationships among the genes. After the first-phase study, however, the experimenter will usually have some preference for which genes to study in greater detail. As more gene expression experiments (studies) are performed, the experimenter will refine his or her preferences about the relationships to study in more and more detail. Consequently, a recommendation system that incorporates the preferences of the experimenter seems desirable.

GEEVE allows for repeats of an experiment, and as just mentioned it can be sensitive to an experimenter's preferences for which genes to study. These improvements ostensibly make GEEVE more applicable to real-world design of gene expression experiments. GEEVE also incorporates an efficient causal discovery method that is based on an extension of a causal discovery algorithm [15].

The GEEVE system consists of two modules called the causal Bayesian network update (CBNU) module and the decision tree generation and evaluation (DTGE) module (Fig. 1). The CBNU module uses an algorithm called *Implicit Latent Variable Scoring* (ILVS) method [15] to causally analyze the current microarray data in light of the user's prior beliefs about causal relationships among the genes under study. The DTGE module evaluates a decision tree that was generated based on the results of the CBNU module and the experimenter's preferences, which are expressed with GEEVE as a utility function. Finally (under assumptions) the best possible experiments are recommended to the experimenter. The experimenter then chooses the next experiment to perform, which may or may not be the one suggested by GEEVE. When the results are available, they can be submitted to the CBNU module for a new round of analysis.

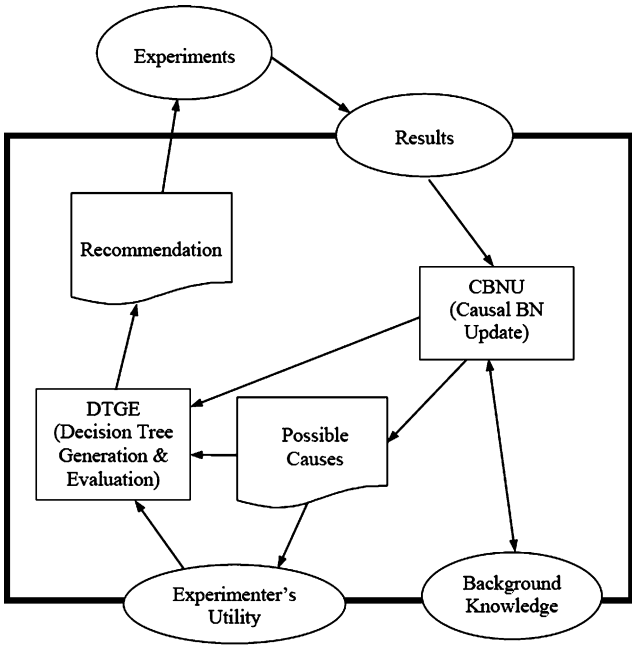

Fig. 1. The GEEVE system. The box with the thick line represents the GEEVE system. Boxes in GEEVE represent system modules. Boxes with wavy lines on the bottom represent outputs from GEEVE. The Experiments oval is an object that is outside of GEEVE. The ovals on the GEEVE border represent objects that communicate with GEEVE from the outside.

2.1. Updating causal bayesian networks

This chapter describes a new method to evaluate causal Bayesian networks using a mixture of observational and experimental data. The algorithm described in the current chapter is then incorporated into the GEEVE system.

A causal Bayesian network (or *causal network* for short) is a Bayesian network in which each arc is interpreted as a direct causal influence between a parent node (variable) and a child node, relative to the other nodes in the network [16]. Fig. 2 illustrates the structure of a hypothetical causal Bayesian network structure containing five nodes that represent genes. The probabilities associated with this causal network structure are not shown.

The causal network structure in Fig. 2 indicates, for example, that the *Gene*1 can regulate (causally influence) the expression level of the *Gene*3, which in turn can regulate the expression level of the *Gene*5. The causal Markov condition gives the conditional independence relationships specified by a causal Bayesian network:

A variable is independent of its non-descendants (i.e., non-effects) given its parents (i.e., its direct causes).

The causal Markov condition permits the joint distribution of the n variables in a causal Bayesian network to be factored as follows [16]:

$$P(x_1, x_2, \ldots, x_n|K) = \prod_{i=1}^{n} P(x_i|\pi_i, K), \quad (1)$$

where x_i denotes a state of variableX_i, π_i denotes a joint state of the parents of X_i, and K denotes background knowledge.

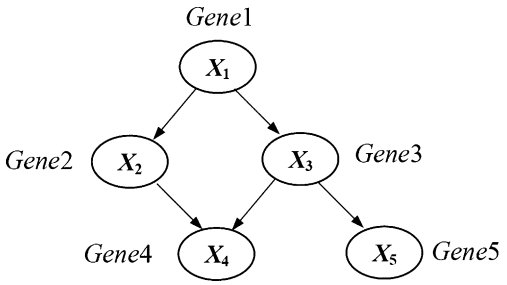

Fig. 2. A causal Bayesian network that represents a portion of a hypothetical gene-regulation pathway.

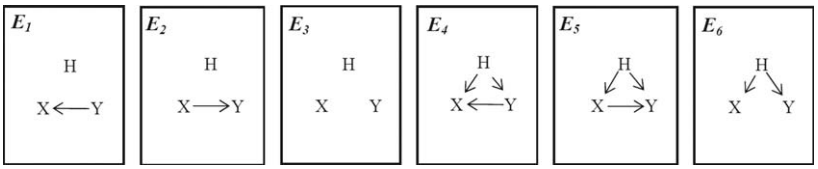

Fig. 3. Six local causal hypotheses.

We introduce six equivalence classes (E_1 through E_6) among the structures (Fig. 3). The causal networks in an equivalence class are statistically indistinguishable for any observational and experimental data on X and Y where H represents a latent variable.

Using the previously published structure scoring method [1,17], we introduced the ILVS method to score the six hypotheses in Fig. 3 [15,18]. *Local ILVS Method* (*LIM*) was introduced to score structures with more than pairwise variables [15]. A high level pseudo code is given in Fig. 4. More detail information of ILVS and LIM could be found at Yoo and Cooper [19] and Yoo [20].

2.2. GEEVE utility model

GEEVE is capable of incorporating an experimenter's utility model [20]. In the research reported in this paper, we did not explore this aspect of GEEVE, because we empirically compare GEEVE's performance to other methods that do not allow modeling utilities flexibly. Instead, we used the following utility assumptions, where E_i^{XY} denotes the node pair X and Y with causal relationship E_i: (1) For all pairs (X, Y), $U(X, Y) = 0.5$, which means that all gene pairs are of equal interest; (2) $U(E_i^{XY}|E_j^{XY}) = 1$ for all $i = j$, which that when the predicted structure E_i^{XY} matches the generating structure E_j^{XY}, the utility is assigned to be the highest possible value (=1.0); (3) $U(E_i^{XY}|E_j^{XY}) = 0.5$ for all E_i^{XY} and E_j^{XY} that have equivalent causal relationships with respect to a latent confounder, that is, E_1^{XY} and E_4^{XY}, E_2^{XY} and E_5^{XY}, and E_3^{XY} and E_6^{XY} are equivalent causal relationships with respect to latent confounder; and otherwise (4) $U(E_i^{XY}|E_j^{XY}) = 0$.

The GEEVE utility for reporting the relationship E_i^{XY} to the user (experimenter) is derived as follows. The weights $w_{ij} = U(E_i^{XY}|E_j^{XY})$ are used as a shorthand notation. The following term is then derived: $q_i = \sum_j w_{ij} \cdot P(E_j^{XY}|D, K)$. Finally, the experimenter's utility for discovering a novel and interesting causal relationship is calculated as $q_i \cdot U(X, Y)$.

2.3. Generating a decision tree

Based on the experimenter's utility specification and the causal Bayesian network output (generated by LIM [15] through a local heuristic search and model selection) the GEEVE system builds a decision tree and evaluates it. GEEVE concentrates on pairwise relationships of genes and generates Fig. 5, where R_j represents a pair of genes, np represents the number of pairs among the genes, m represents a maximum measurements that are obtained for an experimental study, ne_h represents the experimental conditions (explained later in this section) to impose for dataset simulation, tE_i represents the situation where the true structure is E_i, and q_i is defined as in Section 2.2.

For the decision tree shown in Fig. 5, assuming that there are at most l states for each variable and assuming there are k variables modeled in LIM's local structure (see Fig. 4), then the number of possible datasets $nd \leqslant l^{km}$, which is exponential in the number m of microarray experiments (cases). LIM uses a simulation method [21] to make

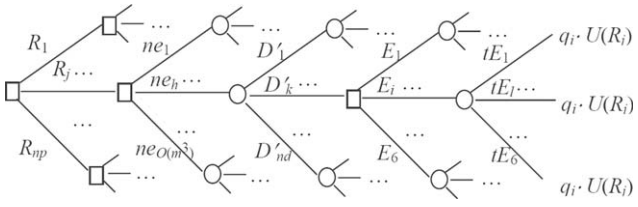

Fig. 5. Specifying the experimental condition decision branch.

```
For each (X, Y), X ≠ Y and X, Y ∈ {All modeled variables}
  Ω ← Select Correlated Variables of (X, Y) /* |Ω| < k */
  For i = 1 to 6
   S ← Greedy Hill Climbing Structure Search with variables in Ω and X and Y
   /* Perturb S and perform Greedy Hill Climbing Structure Search on S
within the user-defined number of iterations */
   /* Score Ei using ILVS and model averaging */
  EndFor
  /* Normalize score of Ei for (X,Y)*/
EndFor
```

Fig. 4. A high level pseudo code of LIM. Note that S is *a local structure* in which it does not include all modeled variables (recall that the set Ω is limited to include only k variables).

the number of possible datasets manageable. LIM keeps track of the highest scoring local structure given an experimental condition and a dataset D. Using the highest scored local structure, LIM generates possible experimental results such as $D'_1, D'_2, \ldots, D'_{nd}$ [21].

The computation of the decision tree evaluation is exponential in the number of microarray experiments (cases). Therefore, we need an approximation method to evaluate the decision tree. Several different approximation methods are available with some assumptions [22,23].

Heckerman et al. [22] introduced a non-myopic approximation method assuming that for a large decision tree, the central limit theorem holds. The method was non-myopic in the number of chance nodes but not in the number of decision nodes.

Chavez and Henrion [23] assumed additive expected utility independence and used linear regression to estimate the expected value of perfect information (EVPI) and expected value of information (EVI). However, Heckerman et al. [22] and Chavez and Henrion [23] approximations are not suitable with large number of decision branches because they assume binary decision nodes. Thus, we use a random heuristic search to approximate the expected value of experimentation. For detail information about the heuristic search, refer to Yoo and Cooper [24].

3. Related work

The GEEVE system incorporates an experimenter's preferences into a decision model to give recommendations about designing a gene-expression experimental study. The decision model it uses is based on decision theory [25,26]. Many different fields concentrate on study design for causal discovery. Traditionally, in statistics and medicine, research on causal discovery is actively pursued in research on controlled trials [5,6,27]. In computer science, causal discovery is also an active research topic, especially in the machine learning community [1–3,20,28,29]. In biology, recent microarray technologies have fueled a field known as *systems biology*, which seeks to discover causal relationships among a large number of genes and other cellular constituents [30,31]. In this section, we will review work related to this paper, concentrating especially on the fields just mentioned.

3.1. Genetic pathway models

Before describing pathway models, we first place them in the context of gene clustering methods, which have been very popular the last few years. Indeed most of the early work on gene expression data analyses used clustering methods. Gene expression levels that were measured by cDNA microarray in the yeast cell-division cycle were analyzed for the first time using a cluster analysis [30]. A cluster analysis typically searches for groups of genes that show similar expression pattern among different experimental conditions. Other analyses followed using similar cluster analyses applied to microarray data [32–35]. Cluster and classification analyses do not necessarily provide causal information, which is at the heart of gene pathway discovery. On the other hand, knowledge of causal pathways can be used to produce a causal clustering of the genes.

Tsang [36] and Dutilh [37] each give a review of genetic networks. Reviews that are focused more on modeling methods are given by de Jong [38] and—especially on Bayesian network—Friedman [39]. A thorough review based on biological context was published by Smolen et al. [40], who suggested that current microarray techniques are limited in delineating intracellular signaling pathways[41]. Smolen et al. [40] argues that since microarray technology is measuring an average expression level of a gene among millions of cells, there is little we can learn about gene-regulation pathways information from the data. We will discuss this issue in Section 5.2 with respect to latent variable detection.

3.2. Experimentation recommendation models

Computational models of scientific discovery were actively studied in artificial intelligence (in conjunction with psychology) in the late 1980s [42]. In molecular biology in particular, Karp [43] created systems in bacterial gene regulation that could describe the initial conditions of an experiment, generate a hypothesis, and refine it.

An extension of supervised learning, *active learning* was applied to learning causal Bayesian networks in scientific discovery [12]. Tong and Koller used edge entropy loss functions and a myopic search to recommend the next best experiment to perform. Their main assumptions are: (1) discrete variables only; (2) no missing data; and (2) no modeling of latent (hidden) variables. They modeled manipulation and selection using the manipulation representation in Cooper and Yoo [44].

Ideker et al. [11] used binary networks to model the perturbation on a gene network and used entropy loss function to recommend the next best perturbation to perform, where perturbation on a gene means forcing the gene to take a fixed value. They implemented two methods to infer a genetic network built from a gene expression dataset. To implement the genetic network, they used a deterministic Boolean model. This model is a simplified version of Bayesian networks (see Section 3.1) where all variables are binary and all conditional distribution tables are simply truth tables.

Similar Boolean networks were used to model the experiments involving the gene networks, and the set-covering method was used to recommend the next best experiment for more than one experimental repeat [10]. Karp et al., used a Boolean circuit model of a biological pathway [45] to model experimentation.

The results of Yoo and Cooper [24] show that the GEEVE system gives better results than systems of Ideker et al. [11] and Tong and Koller [12] (1) in learning the generating models of gene regulation and (2) in recommending experiments to perform.

4. Evaluation

This section describes an evaluation of the GEEVE system. In the evaluation, we used a simulator to generate gene expression data and compared the performance of 10 biologists, some of whom used GEEVE and some of whom did not.

4.1. Simulator for the evaluation

Only a few gene expression simulation systems are currently available [46–48]. Limited functions are available in most of the systems because they are in their early development stages. For example, Tomita et al. [46] simulate a cell by developing a computer program shell that can execute any specified cell model. But the system is limited in its (1) available cell models, (2) exporting the gene expression levels to a file, and (3) modeling of measurement errors.

We used the Scheines and Ramsey [47] simulator system (which we will call the SR Simulator) to generate gene expression data. The SR simulator models genes within a cell and incorporates biological variance, such as that due to signal loss or gene mutation, as well as measurement error. The simulator uses a user-defined number of cells in each probe (we set each probe to contain 100,000 cells in this study). It allows measurement at different time points and uses the following so called *Glass function* [49] to update an expression level of a gene X:

$$eX^t = eX^{t-1} + rate[-eX^{t-1} + F_X(\text{causes\_of}(X^t) \setminus X^{t-1})] + \varepsilon_X, \quad (2)$$

where X^t represents the gene X at time t and eX^t represents the gene expression level of the gene X at time t, $0 < rate \leqslant 1$, causes_of(X^t) are the direct causes of X^t in the model, "\" is the set difference operator, ε_X is an error term drawn from a given probability distribution, and F_X is a binary function specified by the user [49]. Binary functions have been used to model natural phenomena including gene causal pathway [50]. Also note that the model used in this evaluation study contain only a one-stage time-lag, an example of this is shown in Fig. 6, i.e., if a gene has a causal relationship with another gene, it means the relationship is modeled as in Fig. 6.

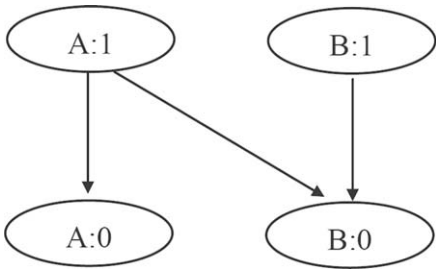

Fig. 6. A one-stage time-lag model. A:0 represents the expression level of gene A at current time and A:1 represents the expression level of gene A at one time-step before the current time.

A burn-in period is desirable in applying the SR Simulator. In particular, for the simulated networks discussed in this section (1) it is often after 80 time lags that the most interesting interactions start among the modeled genes; and (2) the simulated system usually goes into a steady state after 300 time lags. Therefore we used 80 time lags for a burn-in period for evaluation study reported here.

4.2. Evaluation with a case-control study

We created a simulator that models a gene regulation pathway based on assessments from a molecular biologist, Dr. Martin Schmidt, who has many years of research experience related to gene regulation pathways in yeast, especially in pathways that involve *SNF*1 protein kinase [51–53]. With my technical assistance, Dr. Schmidt developed a model of the *SNF*1 protein kinase pathway for use by the SR simulator described in next section.

4.2.1. Simulated SNF1 protein kinase pathway

The structure of the gene-regulatory simulation model that we used is shown in Fig. 7. This model was generated using only one time-lag point as shown in Fig. 6. Thus, if a gene has a causal relationship with another gene in Fig. 7, it means that the relationship is modeled as in Fig. 6. In Fig. 7 the dotted lines represent the causal relationships that are biologically plausible but not biologically certain. *SSG** represents a group of genes, i.e., *SIP*1, *SIP*2, and *GAL*83. *SSG* was modeled in the simulator but was hidden from the participants in the study that is the expression level of *SSG* was not provided to the participants. Having a latent *SSG* in the model simulates (in limited way) a real microarray experiment in that there can be gene expression levels that are not measured.

ε_X in Eq. (2) was estimated from the cDNA microarray dataset of Gasch et al. [54] and *rate* was estimated as 0.5 by consulting with Dr. Schmidt. F_X (also assessed from

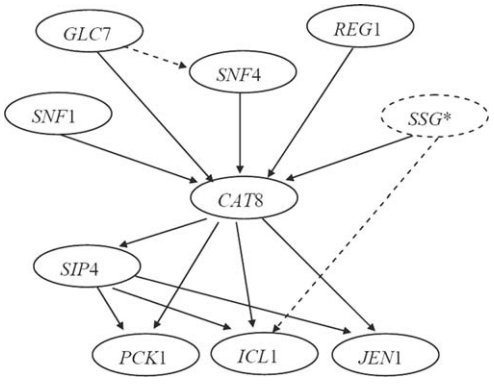

Fig. 7. *SNF*1 simulation pathway model. Dotted lines represent the causal relationships that are biologically plausible, but need further investigation. *SSG** represents a group of genes, i.e., *SIP*1, *SIP*2, and *GAL*83. *SSG* was modeled in the simulator but was hidden to the participants in the control study; i.e., the expression level of *SSG* was not provided to the participants.

Table 1
The values of F_X used in Eq. (2) for simulations involving the *SNF*1 protein kinase pathway (0 = not expressed; 1 = expressed)

(a) *GLC*7 and *SNF*4

	*GLC*7 = 0	*GLC*7 = 1
*SNF*4	0	1

(b) *GLC*7/*REG*1/*SNF*1/*SNF*4/*SSG* and *CAT*8 (R_0: *REG*1 = 0, R_1: *REG*1 = 1, S_0: *SSG*1 = 0, and S_1: *SSG*1 = 1)

	*GLC*7 = 0																*GLC*7 = 1															
	*SNF*4 = 0								*SNF*4 = 1								*SNF*4 = 0								*SNF*4 = 1							
	*SNF*1 = 0				*SNF*1 = 1				*SNF*1 = 0				*SNF*1 = 1				*SNF*1 = 0				*SNF*1 = 1				*SNF*1 = 0				*SNF*1 = 1			
	R_0		R_1		R_0		R_1		R_0		R_1		R_0		R_1		R_0		R_1		R_0		R_1		R_0		R_1		R_0		R_1	
	S_0	S_1	S_0	S_1	S_0	S_1	S_0	S_1	S_0	S_1	S_0	S_1	S_0	S_1	S_0	S_1	S_0	S_1	S_0	S_1	S_0	S_1	S_0	S_1	S_0	S_1	S_0	S_1	S_0	S_1	S_0	S_1
*CAT*8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	1

(c) *CAT*8 and *SIP*4

	*CAT*8 = 0	*CAT*8 = 1
*SIP*4	0	1

(d) *CAT*8/*SIP*4 and *ICL*1/*JEN*1/*PCK*1

	*CAT*8 = 0		*CAT*8 = 1	
	*SIP*4 = 0	*SIP*4 = 1	*SIP*4 = 0	*SIP*4 = 1
*ICL*1	0	1	1	1
*JEN*1	0	1	1	1
*PCK*1	0	1	1	1

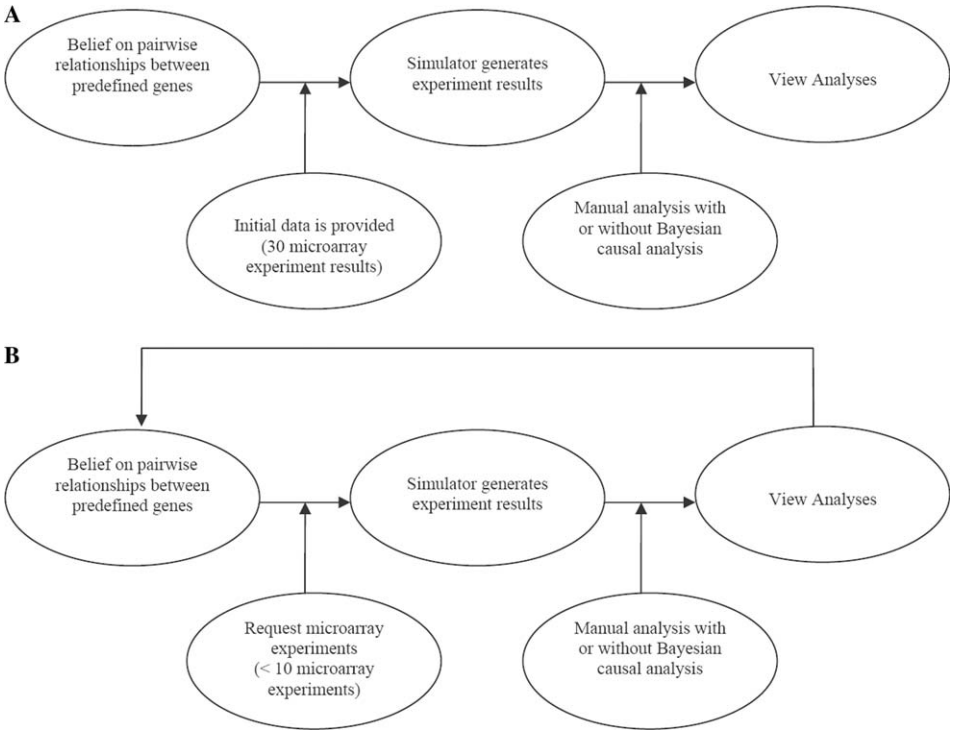

Fig. 8. Phases in the control study. All participants are asked to finish up to Phase 5. (A) Phases 0 and 1. Phase 0 just assesses participants' prior knowledge of predefined genes. Different than phases 2 through 5, phase 1 provides participants with initial microarray experiment results. (B) Phases 2 through 5. Different than in Phase 1, in Phase 2 through Phase 5 participants have to request microarray experiments based on microarray experimental results collected so far.

Dr. Schmidt) from Eq. (2) is defined in Table 1. Table 1 shows the truth table of Glass function that is used in the simulator. For example, in Table 1A, when *GLC*7 is knocked out ($GLC7 = 0$), then the entire system is shut down by regulating all the other genes not to express themselves.

4.2.2. Study design for experiments involving the SNF1 protein kinase pathway

Ten biology faculty members, post-docs, and graduate students were recruited for this study and offered $50 per hour of participation. The biologists expressed at least some knowledge of the SNF1 protein kinase pathway. We randomly divided these study participants into two groups: (1) a control group that did not use GEEVE, and (2) an intervention group that used GEEVE. All participants were able to obtain the gene expressions levels for the nine genes (*SSG* was hidden from the participants) in Fig. 7 under the following experimental conditions:[2]

- a wild-type experiment (i.e., no genes were knocked out) with glucose present;
- a knock-out experiment (with glucose present) for which a single gene (selected from among the nine genes in Fig. 7) was knocked out.

All participants could request up to a total of 50 measured microarray chips over all five phases. The generated results were divided into five experimental phases as follows:

(1) Phase 0 and Phase 1 (Fig. 8A. Participants were initially provided with four wild-type experiments and four experiments in which gene *GLC*7 was knocked out. Thus, a total of eight microarray measurements were initially provided. Four microarray measurements were provided for each experimental condition, as is common in microarray studies. The *GLC*7 knock-out experiment was initially provided because it totally shuts down the cell system. To view the microarray measurements that are provided initially, and as well as the requested microarray measurements after each phase, the control group was provided with a graphical interface that displays gene expression levels in grayscale dots (call this the window GRAYSCALE_DOT window), which is shown in Fig. 9. The intervention group was provided with a causal analysis result from GEEVE (see Fig. 13) along with the GRAYSCALE_DOT window.

[2] There could be other experimental conditions, such as over-expressing a gene, knocking out more than two genes at a time, or setting different environmental conditions, but this initial study is restricted to the experimental conditions listed.

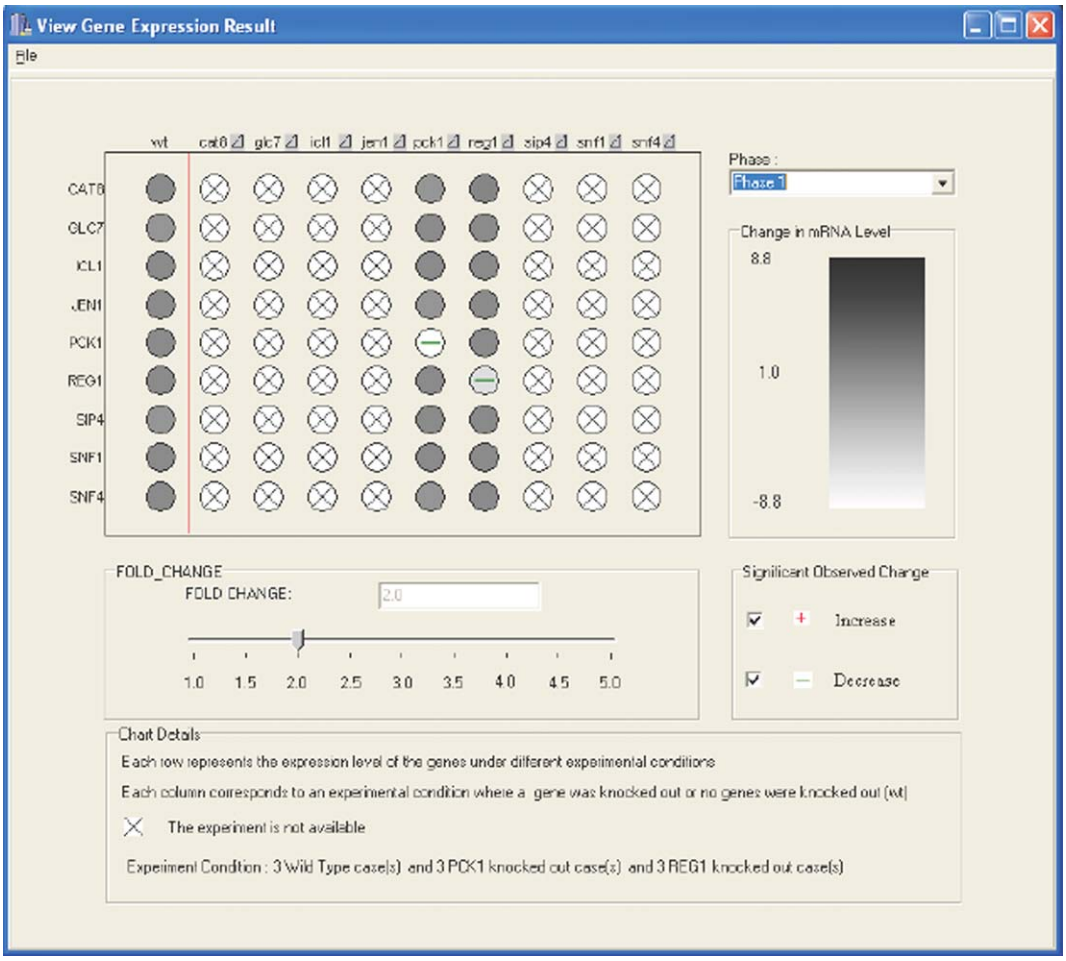

Fig. 9. Experimental result display window, which is called GRAYSCALE_DOT. Each dot represents the gene expression level of a gene that is listed at the left. Each column represents the experimental condition (e.g., reg1Δ represents the experimental condition in which REG1 was knocked out). Participants can choose any phase in the top right corner and view the results of that phase. They can slide the FOLD_CHANGE bar to mark (with "−" and "+" signs) the significant observed changes on the gray scale dot. The "×" marked dots indicate those experimental results that are not avialable at this time.

(2) Phase i for ($2 \leqslant i \leqslant 5$, Fig. 8B):
- In each phase, participants in the *intervention group* were required to follow GEEVE's recommendations for which experiments to perform (detail in Fig. 10). This requirement was relaxed slightly in Phases 1 and 2, where participants could select among two experimental design choices generated by GEEVE. As described in Chapter 4, the experimental design recommendations by GEEVE were based on an expected value of information calculation and it used all of the microarray data that had been gathered up to the current point of analysis. An experimental design consisted of a specification of (1) a selected pair (X, Y) of genes to focus on, and (2) how many microarray experiments (conditions) to perform with those genes, where an individual experiment could involve either knocking out X and measuring Y, knocking out Y and measuring X, or just measuring X and Y without knocking out either one of them (the wild-type experiment). The total number of microarray experiments that could be performed in a single phase was 10. As stated previously, the total number of microarray experiments over all the phases was allowed to be up to 50. The window that the participants used in the intervention group is shown in Fig. 11. The left table of the experiment request window (in Fig. 11) is the list of causal predictions of LIM and upper right corner table shows the experimental recommendations from GEEVE. Note that because of limited biological variation that is presented in current microarray technology [55], by default, we do not display the analysis results of the latent confounded relationships (Fig. 11). We display only, for example, the best recommendation of GEEVE is to carry out one wild-type experiment, three *REG*1 knock-out experiments, and three *SNF*4 knock-out experiments. Recall from Section 2.3 that GEEVE models three different types of experiments: (1) a wild-type; (2) a knock-out of *Gene*1; and (3) a knock-out of *Gene*2. Participants could indicate an experimental preference that differed from the recommendation

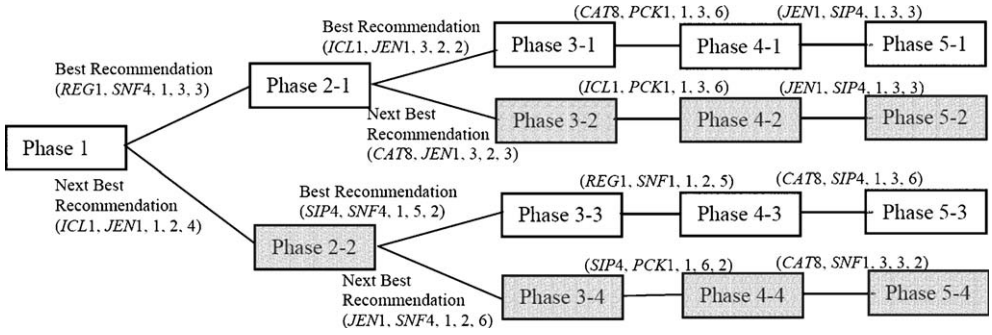

Fig. 10. Generated scenarios of GEEVE recommendations for the intervention group. In Phases 1 and 2, participants are given the option to choose between the best recommendation and the next best recommendation. After Phase 3, participants are asked to follow only the best recommendation of GEEVE. Contents within the parentheses represent recommendations of GEEVE in the form (GeneX, GeneY, NWT, NXK, and NYK) where GeneX and GeneY represent a gene pair, NWT represents the number of wild-type experiments, NXK represents number of GeneX knock-out experiments, and NYK represents number of GeneY knock-out experiments.

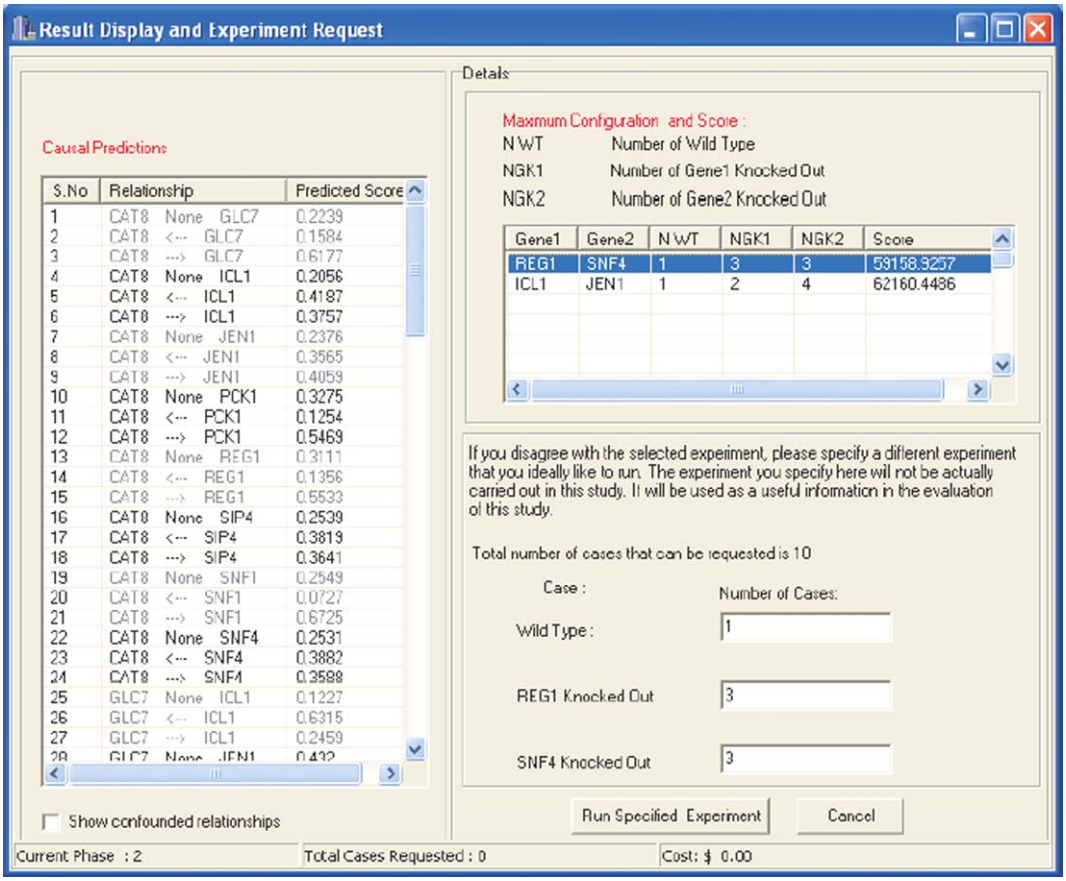

Fig. 11. Experiment request window (intervention group). Participants in the intervention group could select from GEEVE recommendations, which are displayed in the top right table. See main text for detail descriptions of the contents in the window.

by GEEVE (by changing the number of cases (microarray experiments) in the text box on the lower right corner), but this information was logged only and did not influence the experimental simulations actually performed (as shown by the path of phases in Fig. 10).

- Participants in the *control group* could have GEEVE perform (via a simulation) any experiment they desired. As with the intervention group, the total number of microarray experiments that could be performed in a single phase was 10, and the total number of microarray experiments over all the phases was allowed to be up to 50. In designing their experiments, the control group participants had electronic access to all of the previous microarray data that had been gathered up to the current point. The window that the participants used in the control group is shown in Fig. 12.

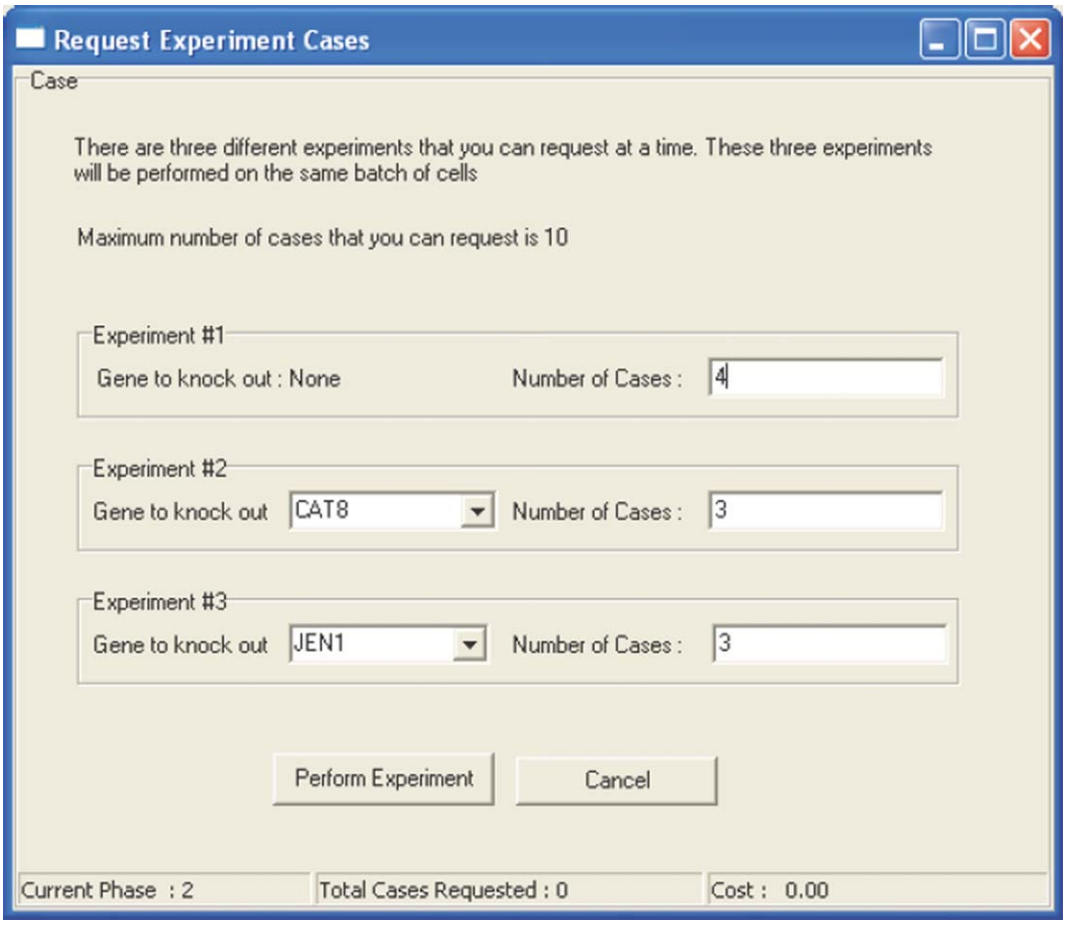

Fig. 12. Experiment request window (control group). Participants in the control group could request up to 10 measurements in three different experiments.

There were three introductory sessions for the participants in the study that described (1) a general overview of the study; (2) the system used in the control group; and (3) the GEEVE system used in the intervention group. After the introductory session, participants were randomly placed either in the control or intervention groups. All participants were asked to download an appropriate program via a designated study website and to use an email attachment to send their results back to me within a week. Three participants needed more than a week to respond. All participants responded within two weeks.

Since a typical GEEVE analysis of a single microarray result takes 3–4 h, it was not practical to ask the participants in the intervention group to wait that long for a recommendation. Thus, we precomputed and cached different experimental scenarios that the user could follow in this study. As shown in Fig. 10 only a limited number of experimental options were available to the user, because of limited time to pre-compute them.

Thus, when a user selected one of these experiments, GEEVE could almost instantly retrieve the data from that simulated experiment as well as its causal analysis of that data. We selected the experimental scenario choices as those that, according to GEEVE, had the lowest expected cost/benefit ratio, where benefit is the term calculated by evaluating the decision tree in Fig. 5. As stated in Section 2.3, GEEVE incorporates experimenter's preferences among the gene pairs and calculates the cost of the experiments. Note that since GEEVE recommendations were precompiled in this study (Fig. 10), the preferences were assessed from Dr. Schmidt and the same preferences were provided to all participants in the study; each participant was asked to adopt those preferences as his or her own. Using those features, GEEVE recommends which experimental condition to consider, i.e., which gene to knock-out, and how many microarray experiments to perform. For example, let us assume that after analyzing the initial data, GEEVE calculates that knocking out *REG*1 and observing expression levels of all other genes may be the best experiment to perform. Additionally, GEEVE recommends doing three repetitions of this experiment. If the user wants to follow the best recommendation, GEEVE would then fetch the analysis result from the appropriate precalculated configuration, e.g., Phases 2–1 in Fig. 10.

The knock-out experiments in both the intervention and the control groups were limited to a single knock-out of one of the nine genes (excluding *SSG*) shown in Fig. 13. After each phase, each participant was asked to give his or her probability assessment of the causal relationships among the genes that the participant had initially selected as being

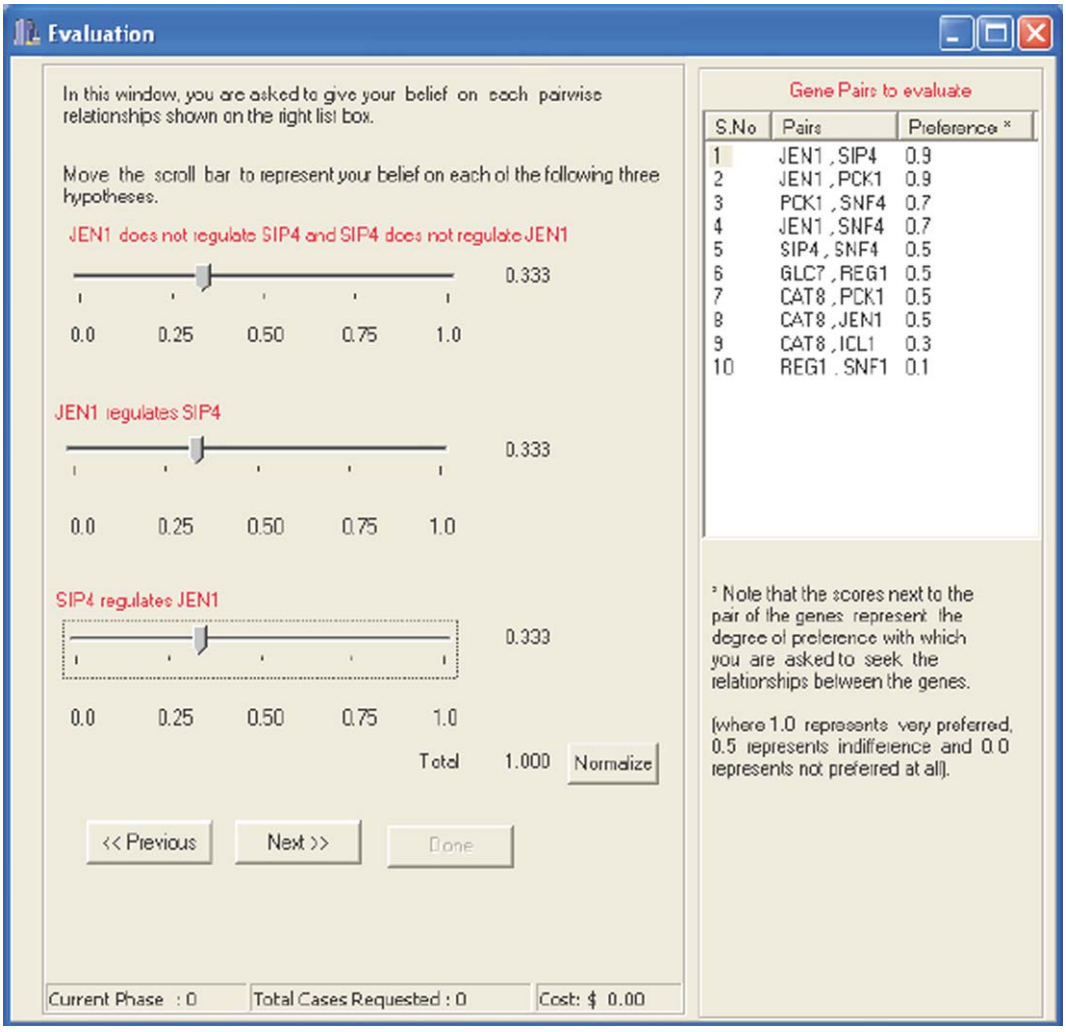

Fig. 13. Evaluation window. Participants in control and intervention groups were asked to assess their beliefs on 10 prespecified gene pairs after each phase. All participants were given the same gene pairs and asked to adopt the preferences that are shown in the upper right corner. Preferences on gene pairs were assessed from Dr. Schmidt.

of most interest to him or her. Fig. 13 shows the window that was used to obtain these assessments. The resources (i.e., participant's time to finish the study and the number of experimental and observational cases) used by each individual to predict the causal relationships were also recorded.

There was a baseline (before the study) and a follow-up (after the study) computer-based questionnaire given to each participant to assess his or her belief about the causal relationships among the 10 genes in the domain of study (see Fig. 13). The participant's beliefs at baseline are tagged as Phase 0 beliefs. The information from these questionnaires is summarized in the next section.

The following results use ROC curves to characterize the discovery performance of each participant on Dr. Schmidt's preferences among 10 pairwise genes shown in Fig. 13 (Note the preferences are shown in the upper right table). To assess Dr. Schmidt's preferences, we randomly selected 20 gene pairs and asked Dr. Schmidt to express his preference among the gene pairs (Dr. Schmidt was also asked to add or drop any gene pairs of the 20 gene pairs we have selected). Before conducting the actual study we have asked my colleagues to test the program that was used in the actual study. Colleagues who participated in the preliminary testing of the program stated that evaluating 20 gene pairs was simply too much. So, out of the 20 gene pairs, we selected 10 gene pairs to get a mixture of 4 gene pairs for which Dr. Schmidt had relatively high preferences (>0.6) and 6 gene pairs for which he had relatively low preferences (<0.6). These preferences were weighted in calculating the AUROC for each participant in each phase, i.e., for all $R \in \{10 \text{ pairwise relationships}\}$, we calculate:

$$\frac{1}{10} \cdot \sum_{R} U(R) \cdot \mathrm{AUROC}(R), \tag{3}$$

where $U(R)$ represents the preference of R and AUROC(R) denote the area under the ROC curve for a prediction that involves R. We also compare the resources that the participants in each group used.

Table 2
Information about the participants in the intervention and control groups

	Professor	Post doc	Ph.D. student (>3rd year)	Ph.D. student (⩽ 3rd year)	Others*	Total
*(a) Positions. *Others include a technician with a Master's degree in a field other than biology*						
Control group	1	0	2	1	1	5
Intervention group	1	1	1	2	0	5
	Understand well	Understand somewhat	Know only the genes	Totally ignorant	Total	
(b) Knowledge in SNF1 pathway						
Control group	0	2	3	0	5	
Intervention group	0	1	4	0	5	
	Understand well	Understand somewhat	Totally ignorant	Total		
(c) Knowledge in cDNA microarray technology						
Control group	0	5	0	5		
Intervention group	1	4	0	5		
	Average	Standard deviation	*p* value			
(d) Subjective self-evaluation of computer expertise using the following values: 0 = Novice, 0.5 = Intermediate, 1.0 = Expert.p value to reject hypothesis H_0: $\mu_1 = \mu_2$ where μ_1 represents the mean of subjective self-evaluation of computer expertise in the intervention group and μ_2 represents the mean of the same value in the control group						
Control group	0.56	0.13	0.34			
Intervention group	0.60	0.16				

4.2.3. Control study results

Table 2 shows more information about the participants in the control and intervention groups. The 10 participants were selected based on their knowledge of the *SNF*1 pathway and cDNA microarray technology. Table 2 shows that participants were equally distributed based on their positions, knowledge of the *SNF*1 pathway, knowledge of cDNA microarray technology, and their expertise in computers (see Appendix A for the pre-study questionnaire). This is because we stratified the participants into the intervention and control groups to balance the dimensions in Table 2 as much as possible.

Fig. 14 plots the AUROC curves to characterize the discovery performance of each group. The intervention group starts at Phase 0 with lower prediction performance in independence and causal predictions than the control group. At the end of Phase 5, the intervention group performs better than the control group in independence and causal predictions. No p values to reject hypothesis $\mu_1 = \mu_2$[3] were lower than 0.05.[4] In other words, we cannot reject the hypothesis $\mu_1 = \mu_2$ with high confidence ($p \leqslant 0.05$) in any phase. Although the differences were not strongly different statistically, Fig. 14 shows a trend in which the intervention group generally performs better (except in Phase 3) than the control group.

Fig. 14 plots the comparison of the two groups in each phase without considering the number of microarray experiments that the participants in the two groups performed (via the SR Simulator). Fig. 15 incorporates that number. In particular, it displays the AUROC per microarray experiment (this unit represents an increased fraction of an AUROC that an experimenter is willing to gain per microarray experiment) for each phase showing plots for the intervention and the control groups.

There are different suggested protocols to analyze a microarray chip [57]. Consulting a technician at the Virginia Bioinformatics Institute, we were told that it usually takes 16 h (two days) of a technician's time to produce and analyze one microarray chip. We were also told that it will usually take 20 and 24 h for him or her to analyze two and three cDNA microarray chips at once respectively (4 h for each additional microarray chip). This is because it usually takes 4 h to finish the first step, extracting DNA. If the technician earns $20 per hour, the costs involved in analyzing two chips in the two different scenarios are: (1) $640 to analyze one chip at a time ($(16\text{ h} \times 2) \times \$20$); and (2) $400 to analyze two chips at once ($20\text{ h} \times \$20$). Similarly, the costs involved in analyzing three chips are: (1) $960 to analyze one chip at a time ($(16\text{ h} \times 3) \times \$20$); and (2) $480 to analyze three chips at once ($24\text{ h} \times \$20$). We use a function that models the expect time to complete microarray experiments, i.e., $f(x) = 4x + 12$, for $x > 1$, where x is the number of microarray experiments requested. Then the cost function becomes $C(x) = \$20 \cdot f(x)$. Results that incorporate the estimated costs involved in performing an experiment are shown in Fig. 16. The resulting trends are similar to Fig. 15, but with even more of a difference between the intervention group and the control group.

The p value to reject the null hypothesis $\mu_1 = \mu_2$, where μ_1 represents the mean of AUROC/case in the intervention group and μ_2 represent the mean of AUROC/case in the control group, is shown in Table 3. Table 3 shows that intervention group has a modestly statistically significant difference from the control group in Phases 2, 4, and 5.

[3] where (1) μ_1 represents the mean of AUROC in the intervention group for causal or independence prediction; and (2) μ_2 represents the mean of AUROC in the control group for causal or independence prediction.

[4] The hypotheses testing analyses in this chapter uses the two-sample t test [56].

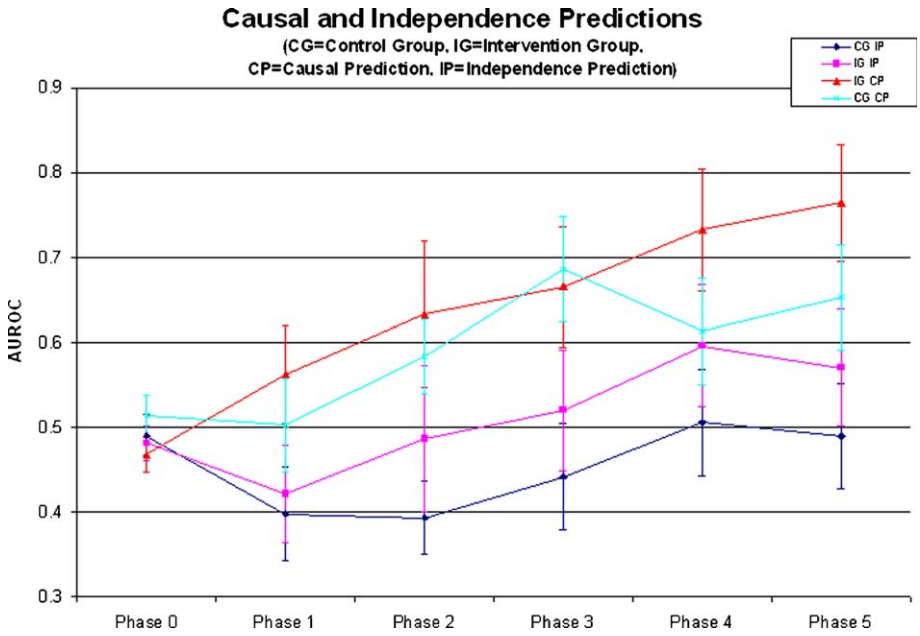

Fig. 14. Area under ROC (AUROC) of the control and intervention groups. Each bar represents a 95% confidence interval.

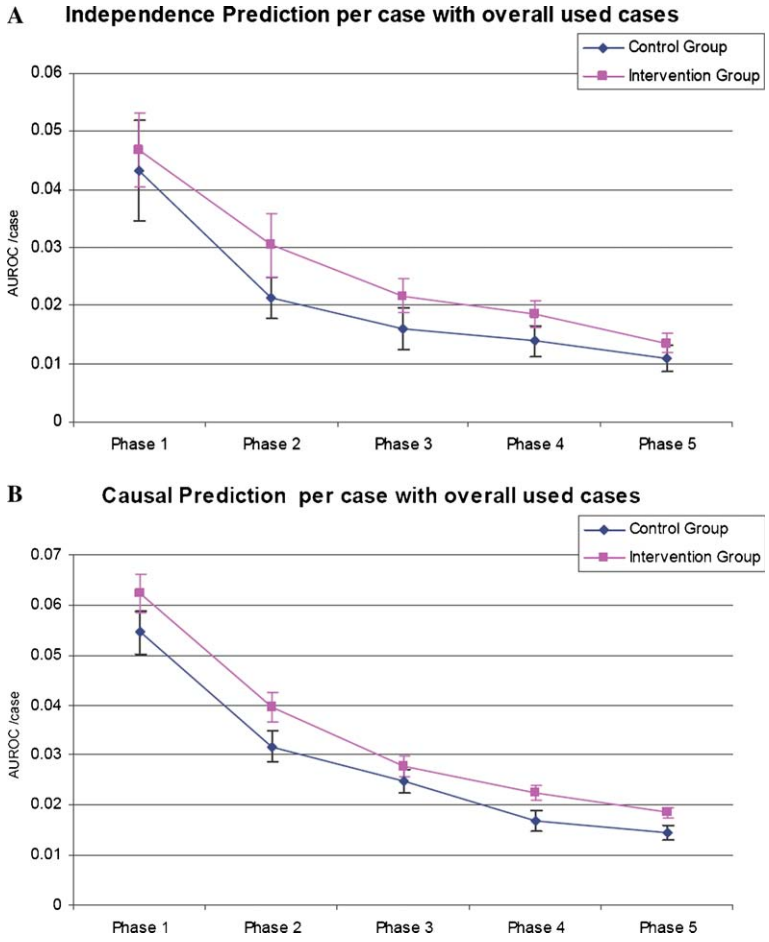

Fig. 15. Area under ROC (AUROC) per experiment for the control and intervention groups. Each bar represents a 95% confidence interval. (A) AUROC per experimental case (microarray experiments) for independence relationship predictions (B) AUROC per experimental case (microarray experiments) for causal relationship predictions.

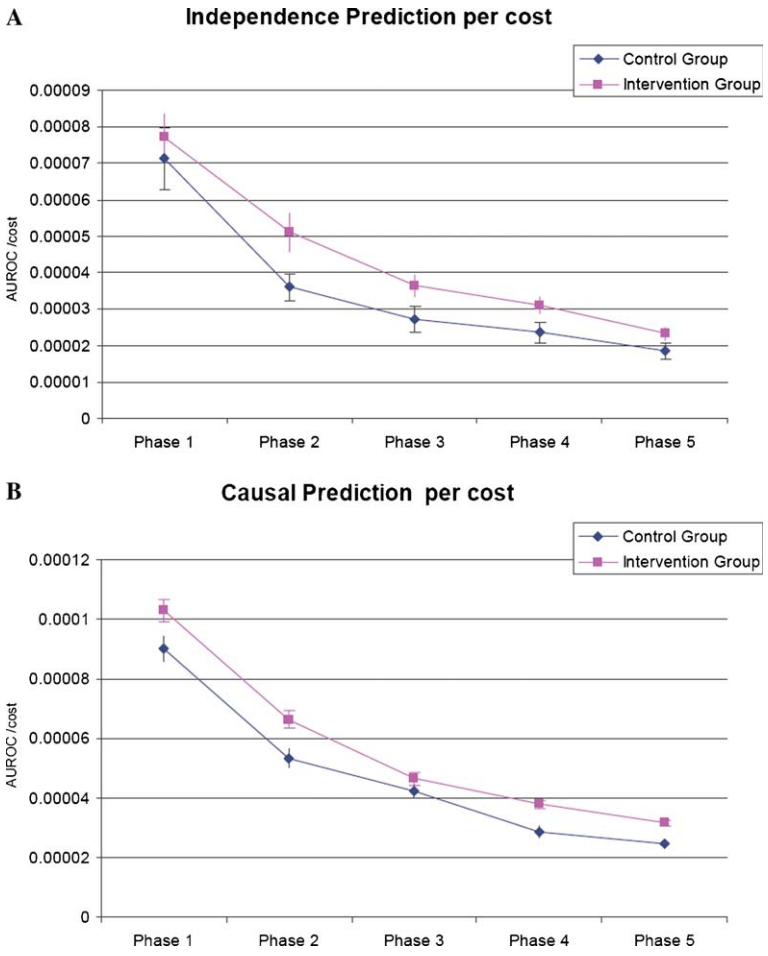

Fig. 16. Area under ROC (AUROC) per cost of the control and intervention groups. Each bar a represents a 95% confidence interval. (A) Independence relationship predictions: AUROC divided by cost. (B) Causal relationship predictions: AUROC divided by cost.

Table 3
p value to reject the hypothesis H_0: $\mu_1 = \mu_2$ where μ_1 represents the mean of AUROC in the intervention group and μ_2 represents the mean of AUROC in the control group

	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
Independence prediction	0.39	0.08	0.14	0.11	0.20
Causal prediction	0.07	0.04[a]	0.14	0.02[a]	0.04[a]

[a] Statistically significant results to support $\mu_1 > \mu_2$.

A similar analysis was performed with only four gene pairs that had higher than 0.7 user preference levels (see the upper right table of Fig. 13 for the four gene pairs and recall that GEEVE incorporates an experimenter's preferences for which genes to study). We are concentrating on these four gene pairs because they represent a group of gene pairs that each experimenter is relatively more interested in. The intervention group performed statistically significantly better than the control group ($p \leqslant 0.05$) in Phases 2, 4, and 5. Fig. 17 shows that using GEEVE with such preferences performed better in causal relationship prediction than an experimenter without using GEEVE's preferences, especially when there were a limited number of microarray measurements.

Fig. 18 shows Fig. 14 in a different way. It calculates the increase of AUROC in each phase. For example, the value in Phase 2 in Fig. 18A represents the AUROC increase from Phases 1 to 2 in Fig. 14. It shows that the increase of causal predictions (Fig. 18B) of the intervention group was statistically significantly better than that of the control group in Phases 1 and 4.

Table 4A shows the difference of the control group and intervention group in terms of average number of microarray experiments obtained among the participants to analyze the data. Table 4B shows the average time the participants spent during the simulated experiments. Although no statistically significant difference was noticed between the two groups, intervention group used the

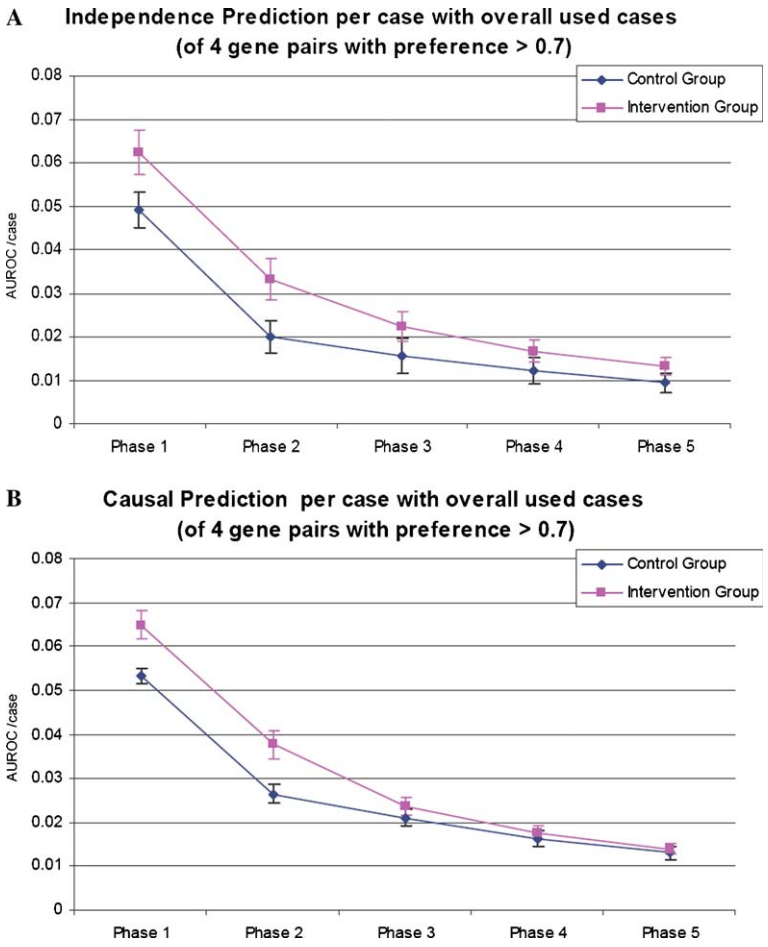

Fig. 17. Area under ROC (AUROC) per case (microarray experiments) of the control and intervention groups considering only gene pairs with a user preference higher than 0.7. Each bar represents a 95% confidence interval. (A) AUROC per case (microarray experiments) of independence relationship predictions. (B) AUROC per case (microarray experiments) of causal relationship predictions.

resources (in terms of the number of microarray experiments obtained and the total amount of time spent) more efficiently. Only five participants responded to the follow-up questionnaire which is shown in full in Appendix B. The results are shown in Table 4C; no statistically significant preference difference was found between the two different result display methods which is not surprising given the small sample size.

Although it was not statistically significant, the intervention group showed a trend to perform better than the control group (Fig. 14) while requesting a smaller number of experiments Table 4A. It may be difficult to clearly specify precisely why such trends were observed in this limited study. However, there are some differences that are observed between the two study groups.

First, as shown in Fig. 10, whatever path the participations in the intervention group take, they eventually perform eight different knock-out experiments, and thus, they do not perform all possible knock-out experiments. For example, if a participant follows a branch that includes Phases 3–3 in Fig. 10, then he or she did not perform a *SNF*1 knock-out experiment (note that *GLC*7 knock-out experiments were initially provided). In the control group 3 (out of 5) participants performed all 9 possible knock-out experiments. This indicates that in the control group many participants preferred a complete experimental design, i.e., an experimental design with all possible knock-out experiments.

Second, as shown in Fig. 10, the number of experiments requested in the intervention group varies within different phases. For example, if a participant follows a branch that includes Phases 3–2 in Fig. 10, then he or she ends up requesting 7, 8, 10, and 7 experiments for Phases 2, 3, 4, and 5, respectively, [we denote these sequence of requests as (7, 8, 10, and 7)]. In the control group, 2 participants requested the same number of experiments for all the phases [(9, 9, 9, and 9) and (10, 10, 10, and 10)] and others showed the following sequence of requests: (9, 10, 10, and 10), (9, 10, 10, and 10), and (9, 8, 6, and 6). This indicates that participants in the control group requested the number of experiments mostly based on the maximum number of experiments that they could request.

5. Conclusions

Systems biology emphasizes large scale discovery of the *interactions* of genes, proteins, and other cell elements. Systems biology is confronted with a huge number of interac-

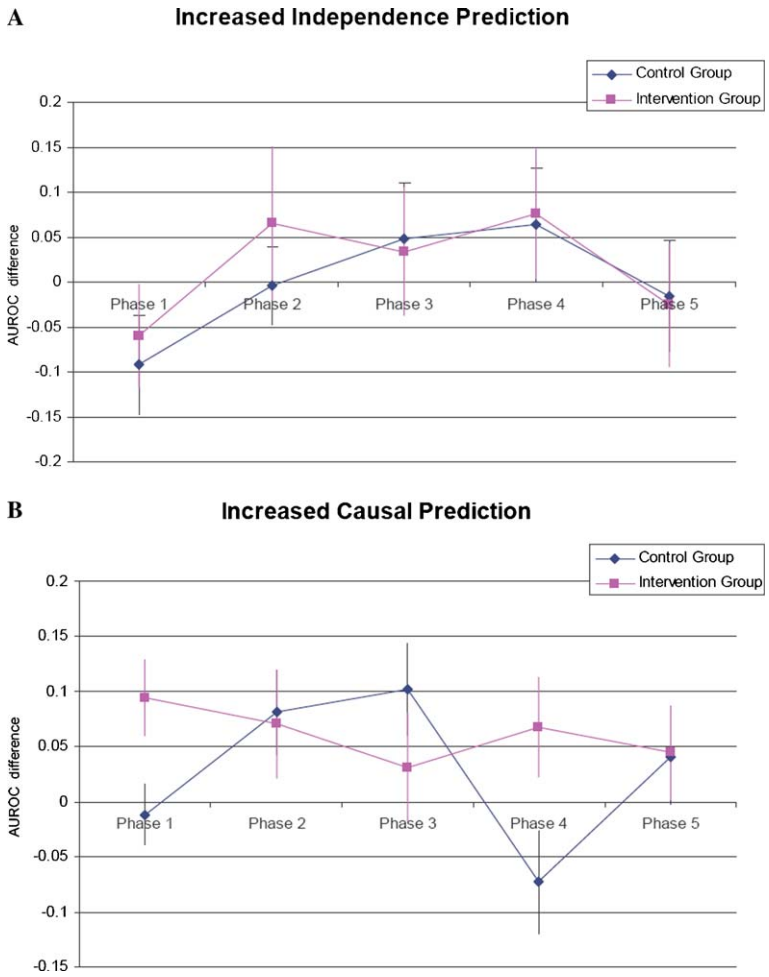

Fig. 18. Difference in AUROC between phases. Each bar represents a 95% confidence interval. (A) AUROC difference between phases of independence relationship predictions. (B) AUROC difference between phases of causal relationship predictions.

Table 4
Resource used in both groups and feedback of the causal analysis result window from GEEVE (Fig. 13) and the GRAYSCALE_DOT window (Fig. 9)

		Average	Standard deviation	*p* value
(a) Number of experiments requested. p value to reject hypothesis H_0: $\mu_1 = \mu_2$ where μ_1 represents the mean of the number of experiments requested in the intervention group and μ_2 represents the mean of the number of experiments requested in the control group.				
Control group		45.2	5.31	0.11
Intervention group		41.8	0.45	
(b) Total time spent in the study (seconds). p value to reject hypothesis H_0: $\mu_1 = \mu_2$ where μ_1 represents the mean of time spent for overall study in the intervention group and μ_2 represents the mean of time spent for overall study in the control group. Note that this is a rough estimate because we cannot guarantee whether a user was working on the control study all the time that the system window was open.				
Control group		4519.00	2509.65	0.27
Intervention group		3388.00	422.96	
(c) Subjective evaluation of GRAYSCALE_DOT and GEEVE causal output. Subjective values used are: 0.0 = not helpful at all, 0.25 = somewhat not helpful, 0.5 = indifference, 0.75 = somewhat helpful, 1.0 = very helpful				
Control group	GRAYSCALE_DOT	0.83[a]	0.14	
Intervention group	GRAYSCALE_DOT	0.62[b]	0.53	
	GEEVE result	0.75[b]	0	

[a] Based on three responses.
[b] Based on two responses (no free text comments were provided in all responses).

tions, not the least of which is the interaction of genes. There are challenges in designing high throughput experiments, such as cDNA microarrays, and for analyzing the high volume of data generated by those experiments to discover gene regulation networks. Intrinsically, these issues are causal in nature. We have introduced a new causal analysis method along with a computer system that uses that method to recommend the gene-regulation experiments to perform.

Unlike clinical randomized controlled trials, where an experimenter is interested in the causal relationship of a handful variables (e.g., an experimenter is interested in a new drug and its treatment effect) in systems biology an experimenter is usually interested in the causal relationships among thousands of entities, such as genes. Different approaches are needed in systems biology for causal discovery and experimental design recommendation. This paper has explored one such approach. In the remainder of this section, we summarize the contributions made by this paper and then discuss open problems.

5.1. Local causal search with experimentation recommendations

We developed a system called GEEVE that incorporates an experimenter's preferences regarding which genes to study to discover causal relationships among those genes. Among the genes of interest, GEEVE models their likely causal relationships, based on prior biological knowledge and experimental data.

Experiments provide benefit in terms of information, but they also have costs in terms of human labor and the laboratory costs. Considering preferences, costs, and a current model of causal relationships, GEEVE recommends the most cost-effective experiment it can find in its search of the space of experiments.

We conducted a randomized controlled study that involved 10 biologists who are familiar with the SNF1 pathway in yeast. This study showed that most of the time the intervention group performed better—although not always statistically significantly so—than the control group in predicting whether pairs of genes (of interest to the biologist study participant) act independently or have a causal relationship. It also showed that the intervention group designed experiments more efficiently than the control group by not selecting gene pairs and number of experiments using a blocked design with a maximum allowed number of experiments, which appears to be the design most commonly used by participants in the control group.

Calculating the predictive performance (area under ROC curve) per (simulated) experiment performed, the intervention group showed better results, some of which were statistically significant, than did the control group. These results suggest that GEEVE did improve the efficiency of the intervention group in discovering causal relationships among the genes of interest.

5.2. Future work and open issues

Regarding external experimental conditions, such as nutrient conditions, they could be modeled as exogenous variables in a causal Bayesian network. Currently, GEEVE models only experiments that involve wild-type gene levels and single gene knock-outs. In the future, more general experiments, such as over-expression experiments, more than one gene knock-out and so forth, should be modeled.

Regarding modeling the time course of gene expression, and determining precisely when to sample cells during experimentation, temporal Bayesian networks appear a natural choice [58,59]. It will be interesting to explore models that use both continuous and discrete variables within temporal Bayesian networks. Temporal Bayesian networks also provide one approach to modeling gene regulation feedback. The six pairwise causal hypotheses used in this research could be extended to model such feedback. This is an important issue for future research because feedback is widely observed in many cellular pathways.

Currently GEEVE only generates decision trees based on the discovery of pairwise gene relationships. More generally, R_j in Fig. 5 (Section 2.3) should include more than pairwise relationships. Doing so will allow GEEVE to (1) model beyond a single gene perturbation experiments, such as a knock-out of two or more genes at a time; and (2) incorporate (in the decision tree) the effects on other genes besides genes (X, Y) when gene X (or Y) is perturbed.

We have also introduced a causal discovery system that can score latent structures. Since the most closely related prior methods assume no latent variables, there is no straightforward way to evaluate GEEVE's prediction of latent structures with these other methods. Also since cDNA microarray is measuring the average expression level of millions of cells, the variance that we observe in the levels (when an experiment is repeated several times) is due almost entirely to measurement error and not to biological variation [55]. Biological variation is needed to discover latent structure, certainly with LIM, and we believe with any method. Measuring the expression level of genes under various experimental conditions (e.g., measuring at different time points or in different temperatures) can provide biological variation among groups of cells; it is an open question how helpful biological variation of this particular variety will be in discovery of latent structure.

Another way to obtain biological variation in gene expression would be to measure gene expression at the level of a single cell. Such measurements will require new technology. We anticipate that such methods will be developed within the next decade. If so, the methods in this paper will be applicable to suggesting when latent factors (such as unknown proteins) may be influencing two or more specific genes.

Due to time and budget constraints, the evaluation reported in this paper only addresses an initial, limited set of issues. For example, the study does not answer the following question: How would the participants in the intervention group perform if they were placed in the control group? Such a cross-over design could be conducted by extending the present study to include a different simulation pathway network. In this additional study, we would ask participants in the intervention group of the current study (reported herein) to participate in the control group in the additional study and vice versa.

Many other ways to improve the controlled study can be considered, such as:

(1) Recruiting more participants, which will facilitate finer matching of the control and intervention groups in terms of their knowledge of the *SNF*1 pathway. We expect that having additional participants also will tighten the confidence intervals in the AUROC analyses.
(2) Providing more personal attention to each participant. We could have visited each participant personally and spent the entire time with him/her during his/her participation in the study. In this case, we could have asked all the post-study questions immediately after the participant completed using the study.
(3) Having an independent person conducting the introductory session. This would remove the chance that I, as the developer of GEEVE, might unintentionally have biased training in a way that favors the participants view of GEEVE.

This paper presents only an initial study of the causal analysis and experimental design modules of GEEVE. Clearly, much more extensive testing can be done using simulated data and real data. When GEEVE is sufficiently refined, it would be interesting to make it available as a web-based system and let biologists submit their microarray data along with their preferences to GEEVE and then receive the analysis results online or via e-mail. To accomplish this, it might be helpful to use already developed hypothesis ontology [60,61].

Ideker et al. [31] describe four steps in discovering causal pathways among the genes: (1) gather and formulate the current knowledge about the genes and their pathways; (2) design and perform experiments; (3) analyze the data from the experiments; and (4) formulate new hypotheses to explain the analysis results not predicted by Step 1 and then repeat steps 2, 3, and 4. There are many open issues in how to complete this loop. The soundness of microarray measurements needs to be studied further, e.g., studying the relationship between mRNA levels and protein expression levels, and studying and quantifying the various sources of measurement error related to detecting gene expression levels. Other open issues include detecting genes and their promoter regions from sequence information, compiling known gene regulatory knowledge (and other cell-network knowledge) from the literature, and standardizing causal pathway representations to interact with interaction with existing databases of biological pathways, e.g., KEGG [62], and WIT [63].

Acknowledgments

This research has been supported in part by grants from the National Science Foundation (IIS-9812021), the National Cancer Institute, the Mellon Fellowship of University of Pittsburgh, the National Institute of Health, and the National Aeronautics and Space Administration (NRA2-37143).

Appendix A. Pre-study questionnaire

Thank you for your interest in participating in a study that involves designing and analyzing simulated cDNA microarray experiments in yeast. I am writing to ask the following six brief questions, which should take a total of less than a minute of your time. Your responses will be very helpful in guiding the selection of the best balance of participants for the study. I hope to be back in contact with you soon.

Thank you.
Changwon Yoo

1. To what degree are you familiar with cDNA microarray technology? _____
   A = Not familiar with cDNA microarray technology.
   B = Understand the basic idea of cDNA microarray technology but not the details.
   C = Understand the concept of cDNA microarray technology well.
2. Have you ever analyzed a dataset of a cDNA microarray study (yes/no)? _____
3. Have you ever designed a cDNA microarray study (yes/no)? _____
4. Have you ever carried out a cDNA microarray study in the lab (yes/no)? _____
5. To what degree are you familiar with the gene regulation pathway involving SNF1 protein kinase in yeast? _____
   A = Know nothing about the pathway.
   B = Know some genes involved in the pathway but do not know about the pathway.
   C = Know some genes involved in the pathway and some knowledge about the pathway.
   D = Know most of the genes involved in the pathway and some knowledge about the pathway.
6. How would you rate your level of computer expertise? (0 = Novice, 0.5 = Intermediate, 1.0 = Expert)
7. Please choose one of the following that best describes you. _____
   A = Master's degree student.
   B = Master's degree holder but not a doctoral student.
   C = Doctoral student (3 or less years in the program).
   D = Doctoral student (more than 3 years in the program).
   E = Post doc.
   F = Faculty member (assistance professor or higher).
   G = Other (please specify).

Appendix B. Follow-up questionnaire

Thanks for your participation in the control study. I am writing to ask the following four brief questions. Your responses will be very helpful in analyzing the study.

Thank you.
Changwon Yoo

- Did you experience any difficulties interpreting the gene expression level (analysis results displayed in grayscale dots in a separate window)? If yes, please describe them briefly (free text)?
- Were the program's display of the gene expression levels helpful in learning the causal relationships between the genes (values between 0.0 and 1.0, 0.0 = not helpful at all, 0.25 = somewhat not helpful 0.5 = indifference, 0.75 = somewhat helpful, 1.0 = very helpful)?
- Did you experience any difficulties interpreting the program's causal predictions (analysis results displayed in "CAT8->JEN1 0.823" format in the Evaluation Window) (yes/no)? If yes, please describe them briefly (free text)?*
- Were the program's causal predictions helpful in learning the causal relationships between the genes (values between 0.0 and 1.0, 0.0 = not helpful at all, 0.25 = somewhat not helpful 0.5 = indifference, 0.75 = somewhat helpful, 1.0 = very helpful)?*
- Any other comments (free text)?

*Only asked to the intervention group.

References

[1] Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn 1992;9:309–47.

[2] Heckerman D. A Bayesian approach to learning causal networks. In: Proceedings of the conference on uncertainty in artificial intelligence. Morgan Kaufmann; 1995. p. 285–95.

[3] Spirtes P, Glymour C, Scheines R. Causation, prediction, and search. 2nd ed. Cambridge, MA: MIT Press; 2000.

[4] Bay SD, Shrager J, Pohorille A, Langley P. Revising regulatory networks: from expression data to linear causal models. J Biomed Inform 2002;35:289–97.

[5] Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approach to randomized trials. J R Stat Soc 1994;157(Part 3):357–416.

[6] Berry DA, Stangl DK. Bayesian methods in health-related research. In: Berry DA, Stangl DK, editors. Bayesian biostatistics. New York: Marcel Dekker; 1996. p. 3–66.

[7] Friedman LM, Furberg CD, DeMets DL, Chapter 7. Sample size, In: Fundamentals of clinical trials. 3rd ed. Mosby-Year book: St. Louis; 1996. p. 94–129.

[8] Karp PD, Krummenacker M, Paley S, Wagg J. Integrated pathway/genome database and their role in drug discovery. Trends Biotechnol 1999;17(7):275–81.

[9] Kerr MK, Churchill GA. Experimental design for gene expression microarrays. Biostatistics 2001;2:183–201.

[10] Karp RM, Stoughton R, Yeung KY. Algorithms for choosing differential gene expression experiments. Res Comput Biol 1999:208–17.

[11] Ideker T, Thorsson V. Karp RM. Discovery of regulatory interactions through perturbation: inference and experimental design. In: Pacific symposium biocompution. 2000. p. 305–16.

[12] Tong S, Koller D. Active learning for structure in Bayesian networks. In: International joint conference on artificial intelligence. Seattle; WA: 2001.

[13] Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nat Genet 1999;21(Suppl):33–7.

[14] Lipshutz RJ, Fodor SPA, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. Nat Genet 1999;21(Suppl):20–4.

[15] Yoo C, Cooper G. Discovery of gene-regulation pathways using local causal search. In: AMIA. San Antonio, Texas; 2002. p. 914–8.

[16] Pearl J. Probabilistic Reasoning in Intelligent Systems. In: Brachman RJ, editor. Representation and reasoning. San Mateo, CA: Morgan Kaufmann; 1988.

[17] Heckerman D, Geiger D, Chickering D. Learning Bayesian networks: the combination of knowledge and statistical data. Mach Learn 1995;20:197–243.

[18] Yoo C, Thorsson V, Cooper GF. Discovery of a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. In: Pacific symposium on biocomputing. Maui, Hawaii: World Scientific: 2002. p. 498–509.

[19] Yoo C, Cooper G. Causal discovery of latent-variable models from a mixture of experimental and observational data. In: Center for Biomedical Informatics Research Report CBMI-173. Center for Biomedical Informatics: Pittsburgh, PA: 2001.

[20] Yoo C, Expected value of experimentation in causal discovery from gene expression studies. Ph.D. dissertation, 2002.

[21] Henrion M. Propagating uncertainty in Bayesian networks by probabilistic logic sampling. In: Lemmer JF, Kanal LN, editor. Uncertainty in artificial intelligence 2. North-Holland, Amsterdam. 1988. p. 149–63.

[22] Heckerman D, Horvitz E, Middleton B. An approximate non-myopic computation for value of information. In: Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence. 1991.

[23] Chavez T, Henrion M. Efficient estimation of the value of information in Monte Carlo models. In: Uncertainty in artificial intelligence. 1994. p. 119-27.

[24] Yoo C, Cooper G. An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. J Artif Intell Med 2004;31:169–82.

[25] von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton, NJ: Princeton University Press; 1944.

[26] Keeney RL, Raiffa H. Decisions, with multiple objectives: preference and value tradeoffs. New York: John Wiley; 1976.

[27] Achcar JA. Use of Bayesian analysis to design of clinical trials with one treatment. Commun Stat Theory Methods 1984;13:1693–707.

[28] Pearl J. Causality: models, reasoning, and inference. Cambridge, UK: Cambridge University Press; 2000.

[29] Heckerman D, Meek C, Cooper GF. A Bayesian approach to causal discovery. In: Glymour C, Cooper GF, editors. Computation causation and discovery. Menlo Park, CA: AAAI Press; 1999. p. 141–65.

[30] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol Biol Cell 1998;9:3273–97.

[31] Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Bumgarner R, et al. Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. Science 2001;292:929–34.

[32] Michaels GS, Carr DB, Askenazi M, Fuhrman S, Wen X, Somogyi R. Cluster analysis and data visualization of large-scale gene expression data. In: Pacific symposium on biocomputing. 1998. p. 42–53.

[33] Herwig R, Poustka AJ, Muller C, Bull C, Lehrach H, O'Brien J. Large-scale clustering of cDNA-fingerprinting data. Genome Res 1999;9:1093–105.

[34] Golub TR, Slonim DK, Tamayo P, Huard C, Caasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.

[35] Bushel PR, Hamadeh HK, Bennett L, Green J, Ableson A, Misener S, et al. Computational selection of distinct class- and subclass-specific gene expression signatures. J Biomed Inform 2003;35:160–70.

[36] Tsang J. Gene expression, DNA arrays, and genetic network. In: Unpublished manuscript Bioinformatics Laboratory at University of Waterloo. 1999.

[37] Dutilh, B. Gene networks from microarray data. In: Unpublished manuscript. Literature thesis at Utrecht University. 1999.

[38] de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol 2002;9(1):67–103.
[39] Friedman N. Inferring cellular networks using probabilistic graphical models. Science 2004;303(5659):799–805.
[40] Smolen P, Baxter DA, Byrne JH. Modeling transcriptional control in gene networks—methods, recent results and future directions. Bull Math Biol 2000;62:247–92.
[41] Chu T, Glymour C, Scheines R, Spirtes P. A statistical problem for inference to regulatory structure from associations of gene expression measurement with microarrays. Carnegie Mellon University; 2003.
[42] Shrager J, Langley P. Computational models of discovery and theory formation, In: Shrager J, Langley P, editor. San Mateo, CA: Morgan Kaufman; 1990.
[43] Karp PD. Hypothesis formation as design. In: Shrager J, Langley P, editors. Computational models of discovery and theory formation. San Mateo, CA: Morgan Kaufman; 1990. p. 276–317.
[44] Cooper GF, Yoo C. Causal discovery from a mixture of experimental and observational data. In: Proceedings of the conference on uncertainty in artificial intelligence. Morgan Kaufmann; 1999. p. 116–25.
[45] Akutsu T, Miyano S, Kuhara S, Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: Pacific symposium on biocomputing. Hawaii: 1999. p. 17–28.
[46] Tomita M, Hashimoto K, Takahashi K, Shimizu T, Matsuzaki Y, Miyoshi F, et al. E-CELL: software environment for whole cell simulation. Bioinformatics 1999;15(1):72–84.
[47] Scheines R. Ramsey J. Gene simulator. 2001: Available from: http://www.phil.cmu.edu/projects/tetrad/tetrad4.html.
[48] Saavedra R, Glymour C. A regulatory network simulator. In: Simulator based on (Yuh et al. 1998) under development. 2001.
[49] Edwards R, Glass L. Combinatorial explosion in model gene networks. Chaos 2000;10:691–704.
[50] Kauffman S. Origins of order—self-organization and selection in evolution. Londan: Oxford University Press; 1993.
[51] McCartney R, Schmidt M. Regulation of Snf1 kinase. Activation requires phosphorylation of threonine 210 by an upstream kinase as well as a distinct step mediated by the Snf4 subunit. J. Biol. Chem. 2001;276(39):36460–6.
[52] Schmidt M, McCartney R, Zhang X, Tillman T, Solimeo H, Wolfl S. Std1 and Mth1 proteins interact with the glucose sensors to control glucose-regulated gene expression in *Saccharomyces cerevisiae*. Mol Cell Biol 1999;19:4561–71.
[53] Schmidt M, McCartney R. Beta-subunits of Snf1 kinase are required for kinase function and substrate definition. EMBO J 2000;19(18):4936–43.
[54] Gasch A, Spellman P, Kao C, Carmel-Harel O, Eisen M, Storz G, et al. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 2000;11(12):4241–57.
[55] Spirtes P, Glymour C, Scheines R, Constructing Bayesian network models of gene expression networks from microarray data. In: to appear in the Proceedings of the Atlantic symposium on computational biology, genome information systems and technology. 2001.
[56] deGroot. Probability and Statistics. Reading, Massachusetts: Addison-Wesley; 1986.
[57] Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, et al. A concise guide to cDNA microarray analysis. Biotechniques 2000;29(3):548–62.
[58] Murphy K, Mian S. Modelling gene expression data using dynamic Bayesian networks. In: Technical report, U.B. Department of Computer Science, Editor. 1999.
[59] Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol 2000;7:601–20.
[60] Racunas SA, Shah NH, Albert I, Fedoroff NV. HyBrow: a prototype system for computer-aided hypothesis evaluation. Bioinformatics 2004;20(Suppl. 1):257–64.
[61] Sim I, Olasov B, Carini S. An ontology of randomized controlled trials for evidence-based practice: content specification and evaluation using the competency decomposition method. J Biomed Inform 2004;37(2):108–19.
[62] KEGG, *KEGG (Kyoto encyclopedia of genes and genomes)*, in Available from: http://www.genome.ad.jp/kegg/.
[63] WIT, *WIT (What is there?) is a www-based system to support the creation of function assignments made to genes and the development of metabolic models*, in Available from: http://wit.mcs.anl.gov/WIT2/.