BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 8 2012, pages 1114–1121
doi:10.1093/bioinformatics/bts090

Gene expression Advance Access publication February 21, 2012

Causal reasoning on biological networks: interpreting transcriptional changes

Leonid Chindelevitch[1,†], Daniel Ziemek[1,*,†], Ahmed Enayetallah[2], Ranjit Randhawa[1], Ben Sidders[3], Christoph Brockel[4] and Enoch S. Huang[1]

[1]Computational Sciences Center of Emphasis, Pfizer Worldwide Research & Development, Cambridge, MA 02140, [2]Compound Safety Prediction, Pfizer Worldwide Medicinal Chemistry, Groton, CT 06340, [3]Neusentis, Pfizer Worldwide Research & Development, Cambridge CB21 6GS, UK and [4]Translational and Bioinformatics, Pfizer Business Technologies, Cambridge, MA 02140, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: The interpretation of high-throughput datasets has remained one of the central challenges of computational biology over the past decade. Furthermore, as the amount of biological knowledge increases, it becomes more and more difficult to integrate this large body of knowledge in a meaningful manner. In this article, we propose a particular solution to both of these challenges.

Methods: We integrate available biological knowledge by constructing a network of molecular interactions of a specific kind: causal interactions. The resulting causal graph can be queried to suggest molecular hypotheses that explain the variations observed in a high-throughput gene expression experiment. We show that a simple scoring function can discriminate between a large number of competing molecular hypotheses about the upstream cause of the changes observed in a gene expression profile. We then develop an analytical method for computing the statistical significance of each score. This analytical method also helps assess the effects of random or adversarial noise on the predictive power of our model.

Results: Our results show that the causal graph we constructed from known biological literature is extremely robust to random noise and to missing or spurious information. We demonstrate the power of our causal reasoning model on two specific examples, one from a cancer dataset and the other from a cardiac hypertrophy experiment. We conclude that causal reasoning models provide a valuable addition to the biologist's toolkit for the interpretation of gene expression data.

Availability and implementation: R source code for the method is available upon request.

Contact: daniel.ziemek@pfizer.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 18, 2011; revised on February 10, 2012; accepted on February 16, 2012

1 INTRODUCTION

Over the past decade gene expression datasets have been generated at an increasing pace. Other types of large-scale data (e.g. metabolomics or genetics) are also becoming more commonplace. Approaches for inferring correlative or causal structure directly from large datasets, generated by perturbing a biological system and consisting of several data types, e.g. expression data and genetics data, demonstrate great promise in uncovering novel biological insights. One example of a successful inference of causal relationships is the work of (Schadt *et al.*, 2005), in which the authors predict and validate three novel susceptibility genes for obesity based on a rodent model. The inference proceeds from expression data and genetics data from a controlled population of several hundred mice. In practice, most available datasets are much smaller, observational in nature and report a single kind of measurement.

In parallel with an ever-increasing amount of data generated, the biomedical literature is also growing exponentially. The results of new experiments should be evaluated in light of this knowledge for at least two reasons: (i) to discover novel biology, one needs to know what is already known, understand what hypotheses need refinement and what phenomena remain unexplained; and (ii) data interpretation in light of previous experiments can add significant interpretative power, especially given the limitations of small sample size in many current omics experiments.

The goal of our methodology is to predict upstream regulators of observed expression changes based on a set of ~450 000 causal relationships. The resulting putative regulators constitute directly testable hypotheses for follow-up in the laboratory. In this article, we (i) define a scoring scheme to identify putative upstream regulators for any given input dataset based on a set of causal relationships encoded as a causal graph; (ii) analytically compute significance scores for our predictions based on random input gene sets; (iii) analytically quantify the recoverability of embedded signals from regulators for our causal graph under various kinds of noise; and (iv) give concrete examples where our methodology helps elucidate biological phenomena when presented with real data.

1.1 Related work

The use of prior knowledge has a long history in gene expression analysis. (Zien *et al.*, 2000) performed this analysis by harnessing biological networks. (Hartemink *et al.*, 2001) demonstrated how to use the Bayesian network formalism to distinguish between two alternative pre-specified versions of the yeast galactose pathway.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

In both cases, we are not aware of any extensions of the method that scale to the use of prior information encompassing a significant fraction of today's biological knowledge.

The most common approaches utilizing prior knowledge are gene set enrichment analysis methods. All such methods require two inputs: (i) a pre-defined collection of gene sets, e.g. derived from metabolic pathways in the KEGG database (Kanehisa *et al.*, 2010) and (ii) the measured outcome of a biological experiment, e.g. the differentially expressed genes in a microarray study. The output is usually a set of *P*-values quantifying the degree of association of each gene set with the experimentally derived data. The drawback of gene set methods is that they do not take into account any directionality of regulation, and that gene set libraries often capture general biological phenomena (e.g. apoptosis) without any regard to the mechanistic details of the process. This makes it difficult in many cases to define concise hypotheses for wetlab validation.

(Pollard *et al.*, 2005) presented the outline of an approach very similar to ours in spirit. In their paper, an analysis of the most likely regulators of expression changes derived from type 2 diabetes patients recovered known key genes in diabetes and proposed new regulators. As in our method, the reasoning is based on a structured collection of causal relationships. Selventa recently released a white paper on their website describing this approach in detail (http://www.selventa.com). We briefly discuss their significance metric in the Methods section.

2 METHODS

This section describes the methodology used to discover the putative biological causes of observed gene expression profiles based on directed causal relationships. The simple intuition is that we can make predictions of transcriptional effects (up- or down-regulation) starting from any entity using these causal relationships and then compare these predictions to the actual data. We describe the construction of a causal graph from a collection of causal relationships and introduce the reasoning model used on this graph. We then define the scoring function used to rank putative hypotheses and show how to compute the statistical significance of the results analytically.

2.1 Constructing a causal graph

We are concerned with relationships that link exactly two biological entities establishing (1) the direction of causality between them, and (2) the qualitative response (i.e. up- or down-regulation) of the second entity when the first one is up- or down-regulated. Furthermore, each relationship must be clearly linked to a citation in the literature. A causal graph is then a directed graph $G=(V,E)$ whose nodes V are transcript levels, compound concentrations, or states of biological processes, and where a directed edge from node a to node b means that the abundance or activity of b is regulated by the abundance of a. Importantly, the edge $e=(a,b)$ has a $+$ sign if the regulation is positive (an increase in a leads to an increase in b, and a decrease in a, to a decrease in b), and a $-$ sign if the regulation is negative. Additionally, in order to facilitate hypothesis validation for the scientists, each node is annotated with various identifiers and each edge is annotated with the article it is based on and the specific excerpt that gave rise to it.

Figure 1a gives an illustrative example. In general, our causal graph can contain conflicting and redundant information. Redundant information provides additional evidence for a certain causal relationship. The amount of evidence will not influence our reasoning as it is difficult to assess whether findings reported in two articles constitute independent evidence. Conflicting information can either be due to errors in the extraction of information from literature or to the complexity of the underlying processes. For instance, relationships might have a different directionality of effect in different contexts. In principle, it should be possible to distinguish contextual effects by explicitly including the context in the computation. In practice, however, we found that only limited information is available for each context, so we decided to construct a comprehensive collection of causal relationships without taking the context into account for the computation. To obtain maximum coverage for each expression dataset, we transformed all molecular entities to their human homologs using the Homologene database.

Though text-mining technology is maturing at a rapid pace, we found the only large-scale sources of reliable causal relationships to be databases of manually extracted facts. Unfortunately, no public-domain or academic databases exist that contain a sufficient number of causal relationships. Therefore, we licensed the substrate for our methodology from two vendors: Ingenuity Inc. (http://www.ingenuity.com) and Selventa Inc. (http://www.selventa.com). Both provide manually curated high-quality content. They yield a causal graph containing ∼450 000 causal relationships, of which over 250 000 are unique, between nearly 37 000 entities, representing ∼65 000 full-text articles indexed by PubMed.

2.2 Reasoning on the causal graph

Algorithmically, causal reasoning models the effects of a change in the abundance of a on the abundance of z by tracing the shortest path from a to z in G and then evaluating its sign, determined by the product of the signs of the edges along the path. It is expected that a upregulates z if this overall sign turns out to be positive, and that a downregulates z if it is negative.

It is advantageous to transform $G=(V,E)$ into the *computational causal graph* $G_C=(V_C,E_C)$ in order to be able obtain information about the shortest positive and negative paths by applying standard algorithms to it: breadth-first search if the edges are unweighted, which is the case we consider here, or Dijkstra's algorithm if the edges are weighted. G_C is obtained by creating two copies of each node of G, one with a $+$ sign, one with a $-$ sign, and letting the corresponding edges be duplicated as well, so as to separate positive paths (those between nodes of the same sign) from negative paths (those between nodes of the opposite sign). See Figure 1b for an example. This transformation also allows us to remove redundancies in the set of edges since the same interaction may be described in multiple articles. The resulting graph is *simple* as it has no loops or parallel edges.

2.3 Scoring hypotheses

We assume that we deal with expression data, and that the only observable changes are levels of *gene transcripts*. Let $T(G)$ and $T(G_C)$ denote those nodes of G and G_C, respectively, that correspond to gene transcripts.

The gene expression data allows us to determine the subset G^+ of all gene transcripts that are significantly overexpressed and the subset G^- of all gene transcripts that are significantly underexpressed. We define $G^\pm := G^+ \cup G^-$ to be the set of all significant transcripts in the gene expression profile, and G^0 to be the set of all gene transcripts that are not differentially expressed. We also denote by $-v$ the node v with the opposite sign. We choose a distance threshold Δ which determines the maximum length of the paths we consider. Given a hypothesis $h \in V(G_C)$, we partition the set $T(G_C)$ into three subsets:

$$S_h^+ := \{v \in T(G_C) | d(h,v) \leq \Delta, d(h,v) < d(h,-v)\}$$
$$S_h^- := \{v \in T(G_C) | d(h,-v) \leq \Delta, d(h,-v) < d(h,v)\}$$
$$S_h^0 := \{v \in T(G_C) | d(h,v) > \Delta \text{ or } d(h,v) = d(h,-v)\}$$

The elements of S_h^+ are predicted to be upregulated, those of S_h^-, to be downregulated, and those of S_h^0 are predicted ambiguously.

In order to evaluate the goodness-of-fit of a hypothesis h to the observed expression data, we score 1 for each *correct* prediction, -1 for each *incorrect* prediction and 0 for each *ambiguous* prediction made by h about $G^\pm$. That is,

$$s(h, G^\pm) = (|S_h^+ \cap G^+| + |S_h^- \cap G^-|) - (|S_h^+ \cap G^-| + |S_h^- \cap G^+|).$$

Note that we identify the node $v \in T(G)$ with the node $v^+ \in T(G_C)$ for the purposes of this computation, since $G^\pm \subseteq T(G)$ while $S_h^+, S_h^- \subseteq T(G_C)$. The scores computed for each putative hypothesis provide us with an overall

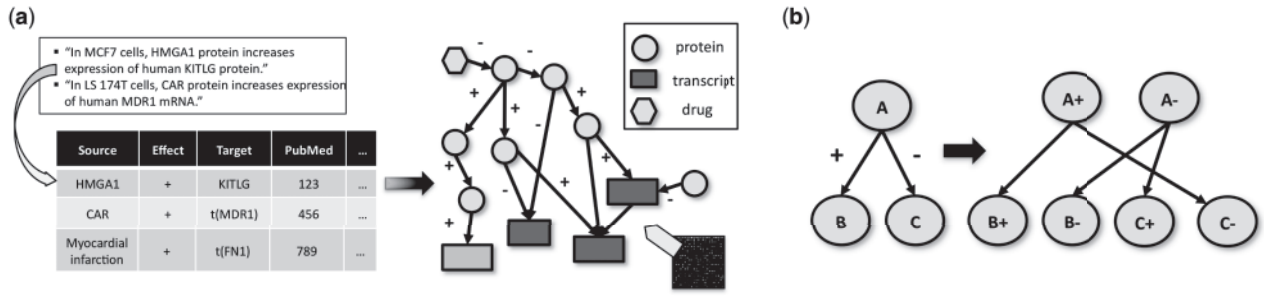

Fig. 1. (**a**) Schematic depiction of a set of relationships curated from the literature and transformed into a causal graph. (**b**) For computational purposes, this graph is transformed into a *computational causal graph*.

ranking of all hypotheses (ties are broken arbitrarily). However, a good score does not necessarily mean good explanatory power, because of possible connectivity differences between the transcript nodes of G_C. In particular, 'hubs' are likely to have higher scores no matter what the expression data is like. Therefore we also need to look at the statistical significance of each score.

2.4 Computing statistical significance

For a given hypothesis h and a given score $s_0 := s(h, G^{\pm})$, we would like to know how likely h is to score s_0 or better with a *random* set of genes $G_R^{\pm} := G_R^{+} \cup G_R^{-}$, chosen with $|G_R^{+}| = |G^{+}|$ and $|G_R^{-}| = |G^{-}|$. This question can clearly be answered if we can compute the distribution of scores that h can have over all such sets.

Let $G_R^{\pm}$ be fixed, and let $G_R^{0} := T(G) - G_R^{\pm}$. Define $q_\sigma := |S_h^\sigma|$ and $n_\sigma := |G_R^\sigma|$, where $\sigma \in \{+,-,0\}$. Note that although $|S_h^{+}| = |S_h^{-}|$ in $T(G_C)$, this is not the case in $T(G)$ since the two sets may have different numbers of positive entities. Therefore $q_+ \neq q_-$ in general. Let us further define $n_{\sigma\tau} := |S_h^\sigma \cap G_R^\tau|$ for $\sigma, \tau \in \{+,-,0\}$. This gives us the 3×3 contingency table of values:

n_{++}	n_{+-}	n_{+0}	q_+
n_{-+}	n_{--}	n_{-0}	q_-
n_{0+}	n_{0-}	n_{00}	q_0
n_+	n_-	n_0	$\|T(G)\|$

This table will be identical for a large number of sets G_R, and this number, $D[n_{++}, n_{+-}, n_{-+}, n_{--}]$, depends only on the top left 2×2 corner of the table, since the other entries are determined by the constraints on row and column sums. Using multinomial coefficients, we can write $D[n_{++}, n_{+-}, n_{-+}, n_{--}]$ as

$$\binom{q_+}{n_{++}, n_{+-}, n_{+0}} \binom{q_-}{n_{-+}, n_{--}, n_{-0}} \binom{q_0}{n_{0+}, n_{0-}, n_{00}}.$$

The score for a set G_R yielding this table will simply be

$$S[n_{++}, n_{+-}, n_{-+}, n_{--}] := (n_{++} + n_{--}) - (n_{+-} + n_{-+}).$$

It also follows from a generalization of Vandermonde's identity that the total number of possible such sets is

$$T := \sum_{n_{++}, n_{+-}, n_{-+}, n_{--}} D[n_{++}, n_{+-}, n_{-+}, n_{--}] = \binom{|T(G)|}{n_+, n_-, n_0}.$$

Thus, the distribution we are seeking will be a sum of the $D[n_{++}, n_{+-}, n_{-+}, n_{--}]$, grouped by $S[n_{++}, n_{+-}, n_{-+}, n_{--}]$ and normalized by T. This distribution can be computed with a quartic dynamic programming algorithm that produces all the D values. However, special summation techniques can be used to obtain a cubic algorithm (Petkovšek *et al.*, 1996).

Generally, when processing a particular dataset, our algorithm begins by computing the scores for each hypothesis and ranking the hypotheses by their score. Then, several filters are applied to constrain the output. First, the P-value P of a hypothesis, computed according to the methodology above, is typically required to be below a certain threshold. Second, an enrichment P-value P_E of a hypothesis is also required to be below a certain threshold. P_E is the probability of finding $n_{++} + n_{--} + n_{+-} + n_{-+}$ differentially expressed transcripts for a putative hypothesis h under the null model of assigning random transcripts to be differentially expressed, and is computed by a Fisher Exact Test. This is a standard approach in gene set enrichment methods [e.g. Draghici *et al.* (2003)] and treats $S_h^{\pm}$ as a gene set without regard to directionality. Finally, we also filter out those hypotheses whose number of correct predictions, $C := n_{++} + n_{--}$, is below a certain user-defined threshold. Each resulting hypothesis is reported to the user in the form *Entity* $+/-$ (e.g. $MYC+$, $AKT-$) and can then be inspected in detail. These predictions explicitly include hypotheses about non-transcriptional molecular events. As tested hypotheses exhibit an intricate dependency structure, it is not straightforward to apply false discovery rate estimation procedures. In this work, we generally consider only highly significant hypotheses as the starting point of any analysis ($P, P_E < 10^{-3}$) and let the user decide to expand to less significant ones based on biological expertise.

2.5 Hypothesis recoverability

In order to understand how well the hypothesis that explains a dataset can be recovered with our methodology we examine the effect of a family of perturbations to the 'perfect' dataset. Suppose that a hypothesis h and a distance threshold Δ are fixed and a set of significant genes $G_R^{\pm}$ of size N is to be generated. The hypothesis h partitions $T(G_C)$ into $S_h^{+}, S_h^{-}, S_h^{0}$. We can decide to pick a certain fraction of $G_R^{\pm}$ from S_h^{+} (signal), a certain fraction from S_h^{-} (adversarial noise) and the rest from S_h^{0} (random noise). We call these fractions γ, δ and ϵ, respectively. However, the random data sets we obtain this way may contain the same entity with opposite signs, something that never occurs in real data. Fortunately, this problem can be easily overcome, as we explain below.

Surprisingly, we can in fact compute the expected effect of such a perturbation (with fixed N, γ, δ and ϵ but averaged over all possible $G_R^{\pm}$) analytically. The key observation is that the expected rank of h depends only on the probabilities that a hypothesis h' beats h. Indeed, the expected rank of h is

$$E[\text{rank}(h)] = 1 + \sum_{h' \neq h} \left(\Pr[s(h) < s(h')] + \frac{1}{2} \Pr[s(h) = s(h')] \right)$$

because h will be placed below all the hypotheses that beat it and half the hypotheses that tie with it (since ties are broken arbitrarily).

Let us now fix a competing hypothesis h'. Let us define $m_{\sigma\tau} := |S_h^\sigma \cap S_{h'}^\tau|$ for $\sigma, \tau \in \{+,-,0\}$. Let $k_{\sigma\tau}$ denote the number of elements chosen for G_R in $S_h^\sigma \cap S_{h'}^\tau$. In order to obtain a dataset with the specified properties we need to choose $|G_R^{+}| = \gamma N$ elements from S_h^{+}, $|G_R^{-}| = \delta N$ from S_h^{-} and $|G_R^{0}| = \epsilon N$ from S_h^{0}. This gives us the following conditions on the possible choices of k values:

$$0 \leq k_{\sigma\tau} \leq m_{\sigma\tau} \; \forall \sigma, \tau, k_{\sigma+} + k_{\sigma-} + k_{\sigma 0} = |G_R^\sigma| \; \forall \sigma.$$

However, only five of the nine parameters $m_{\sigma\tau}$ are effectively independent, because of the relationships

$$S_h^- \cap S_{h'}^- = -(S_h^+ \cap S_{h'}^+); \qquad S_h^- \cap S_{h'}^+ = -(S_h^+ \cap S_{h'}^-);$$
$$S_h^- \cap S_{h'}^0 = -(S_h^+ \cap S_{h'}^0); \qquad S_h^0 \cap S_{h'}^- = -(S_h^0 \cap S_{h'}^+);$$
$$S_h^0 \cap S_{h'}^0 = -(S_h^0 \cap S_{h'}^0).$$

Noting that $|T(G)| = m_{++} + m_{+-} + m_{+0} + m_{0+} + (1/2)m_{00}$, we can partition $T(G)$ into five pairwise disjoint sets, and then further specify the sign of each gene depending on which of the two mutually opposite overlap sets it should belong to. This ensures that we never consider random sets where the same transcript enters two different subsets with opposite signs. It follows that each choice of k values now represents a number of sets given by $N_1 \cdot N_2$, where:

$$N_1 := \binom{m_{++}}{k_{++}, k_{--}, r_{++}} \binom{m_{+-}}{k_{+-}, k_{-+}, r_{+-}} \binom{m_{+0}}{k_{+0}, k_{-0}, r_{+0}};$$
$$N_2 := \binom{m_{0+}}{k_{0+}, k_{0-}, r_{0+}} \binom{m_{00}/2}{k_{00}} 2^{k_{00}}.$$

Here, $r_{\sigma\tau} := m_{\sigma\tau} - (k_{\sigma\tau} + k_{-\sigma-\tau})$ for each appropriate σ, τ, and the last term corresponds to the freedom of choosing the sign of each entity in the self-opposite set $S_h^0 \cap S_{h'}^0$. It remains to compute the distribution of the score obtained by h', $s(h')$, when the k values are subject to the constraints above. For a given choice of k values the score is

$$s(h') := \left(\sum_{\sigma \in \{+,-,0\}} k_{\sigma+} \right) - \left(\sum_{\sigma \in \{+,-,0\}} k_{\sigma-} \right) =$$
$$= (k_{++} - k_{+-} + k_{-+} - k_{--}) + (k_{0+} - k_{0-}) := s_\pm + s_0.$$

Here, we call $s_\pm$ the part of the scores corresponding to $S_h^\pm$ and s_0, the part corresponding to S_h^0. To compute the distribution of $s_\pm$, we note that it is precisely the same as the distribution of scores computed in the previous section, when the parameters q_+, q_-, q_0 and n_+, n_-, n_0 are replaced by m_{++}, m_{+-}, m_{+0} and $\gamma N, \delta N, (m_{++} + m_{+-} + m_{+0}) - (\gamma N + \delta N)$.

To compute the distribution of s_0, we note that the probability P_{s_0} of having a fixed score s_0 is just

$$p_{s_0} := \sum_{\substack{k_{0+} + k_{0-} + k_{00} = \epsilon N \\ k_{0+} - k_{0-} = s_0}} \binom{m_{0+}}{k_{0+}, k_{0-}, r_{0+}} \binom{m_{00}/2}{k_{00}} 2^{k_{00}} =$$
$$= \sum_k \binom{m_{0+}}{k} \binom{m_{0+} - k}{k - s_0} \binom{m_{00}/2}{\epsilon N - 2k + s_0} 2^{\epsilon N - 2k + s_0},$$

which, though it looks complicated, is a one-parameter summation that can either be computed directly in linear time or in average constant time per score s_σ by using special summation techniques (Petkovšek *et al.*, 1996). By the Vandermonde identity we have

$$\sum_{s_0} p_{s_0} = \binom{m_{0+} + m_{00}/2}{\epsilon N} 2^{\epsilon N}.$$

Finally, the overall distribution of scores is obtained by a convolution of the distributions of $s_\pm$ and s_0, and the probability of h' beating h is then the sum of the probabilities of scores exceeding the score of h, $(\gamma - \delta)N$.

In order to compute the sizes of the overlaps, $m_{\sigma\tau}$, we proceed as follows. First, we take A to be the adjacency matrix of G_C, and then get a reduced adjacency matrix A_r by keeping only those columns that correspond to transcripts. We then compute $L := A_r A_r^T$ to get the values of m_{++} and m_{+-} as entries corresponding to $L_{h,h'}$ and $L_{h,-h'}$, respectively. Finally, we compute the other m values by using $\sum_{\tau \in \{+,-,0\}} m_{\sigma\tau} = |S_h^\sigma|$.

An important observation greatly reduces the computational effort required for computing the recoverability of a graph. If two hypotheses, say h and h', are *disjoint*, that is, if $m_{++} = m_{+-} = 0$, then $s_\pm = 0$ for all choices of k values, and the score of h' equals the random noise term s_0. Furthermore, its distribution only depends on the sizes of $S_{h'}^\tau$ but not on h. Therefore, it can be computed once for each h' and aggregated over all h' that are disjoint from h before computing the probability of a score equal to or exceeding that of h. In particular, if $\gamma > 1/2$, the maximum score of any h' disjoint from h, ϵN, cannot equal or exceed the score of h, $(\gamma - \delta)N$, and so the disjoint hypotheses do not make any contribution to h's recoverability.

2.6 Graph perturbations

In order to examine the sensitivity of the model's predictions to graph changes we investigated three ways of perturbing a causal graph. First, we looked at deleting a fraction μ of the edges at random while preserving the sign distribution. Second, we looked at inserting an extra fraction ν of edges uniformly at random. Finally, we compared with a graph generated uniformly at random from the family of graphs with the same in- and out-degree distribution as our causal graph. This graph was generated by a method adapted from (Rao *et al.*, 1996) for causal graphs. The results are presented below.

Rather than looking at the recoverability of the perturbed graph (which will tell us about the likelihood of recovering the correct hypothesis from that graph), we are more interested in computing the expected rank of the correct hypothesis for a dataset generated with the original graph, but analyzed by our methodology on the perturbed graph, which we term the two graphs' *cross-recoverability*. It turns out that computing the cross-recoverability is only slightly more difficult than computing the recoverability.

Indeed, let G' be the perturbed graph [importantly, $V(G) = V(G')$]. The key insight is that it suffices to understand the distribution of scores $s_{h,h'}$ of each hypothesis h' in G' when a random dataset coming from hypothesis h in G is given as input. This in turn only depends on the sizes of the overlaps between the sets S_h^σ in G and $S_{h'}^\tau$ in G', which can be read off from the matrix $L := A_r B_r^T$, where B_r is the reduced adjacency matrix of G'. One final change is that we also get a distribution of scores $s_{h,h}$ for h in G' (rather than a single score as before), and in order to compute the contribution h' makes to the cross-recoverability of h we need to perform one more convolution to get $Pr[s_{h,h'} > s_{h,h}] + (1/2)Pr[s_{h,h'} = s_{h,h}]$.

2.7 Validation on simulated data: robustness

Figure 2 illustrates the recoverability of hypotheses in the causal graph, as well as the cross-recoverability with the causal graph perturbed by random edge deletion. The plots show the fraction of the hypotheses correctly recovered at each rank cutoff, and can be thought of as analogs of ROC curves. While we ran this computation with a larger number of parameters, we just show some representative results for illustration.

Figure 2a shows the recoverability of hypotheses in our causal graph for datasets of size $N = 10$ when $\gamma = 0.4$, i.e. only 40% of the dataset is composed of signal (the correctly predicted consequences S_h^+). When $\epsilon = 0.6$, i.e. the other six entities in the dataset are random noise, there is almost perfect recoverability (circles). When $\delta = 0.2$, i.e. two of the entities are chosen in an adversarial manner (i.e. from S_h^-), the recoverability significantly deteriorates (triangles). Finally, performance is comparable to that of a uniformly random ranking (red dots) when the adversarial noise level almost reaches signal level (crosses).

Figure 2b shows the cross-recoverability of hypotheses when the prediction occurs on the causal graph perturbed by the deletion of a fraction of its edges. Here, $\gamma = 0.8$ and $\delta = 0.2$. Once again, we consider datasets of size $N = 10$. Here, we see a clear deterioration of the performance as progressively more edges are deleted.

We also analyzed the case of additional edges inserted uniformly at random, with a dataset of size $N = 10$. While the perturbations do affect performance in a negative way compared to the baseline, the impact is minor even for a large number of edge insertions. We believe that this is due to the sparsity of our causal graph, so that the randomly added edges are unlikely to appreciably affect the connections between hypotheses and transcripts.

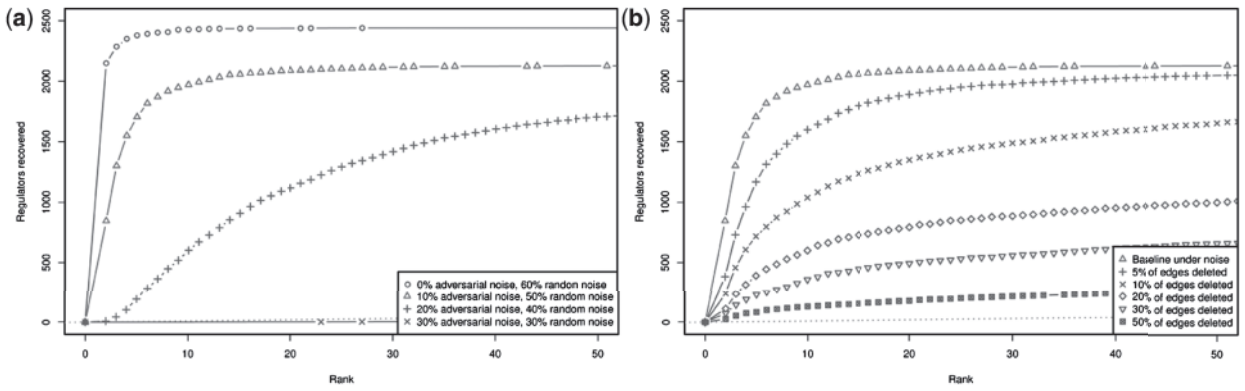

Fig. 2. Estimation of robustness of each hypothesis against noise. Conceptually, we generate all perfect expression data sets of a given size for each hypothesis and compute the average rank of each hypothesis when ordered by score. Each curve represents the number of correctly detected regulators when thresholding at a given average rank. Figure (**a**) compares different noise levels applied to the perfect expression data set for each regulator. Figure (**b**) assesses recoverability of the regulator when edges have been deleted from the network. The red curve is a common reference point in both plots.

Finally, the cross-recoverability for a randomly generated graph with the same in- and out-degree distribution as the original causal graph was even worse than that of a uniformly random ranking.

The conclusion we can draw from Figure 2a is that our causal graph is robust to random noise, but not to adversarial noise. The robustness to random noise is a critical feature for any computational model, since the gene expression data we deal with in practice is likely to be quite noisy. The much lower robustness to adversarial noise is not problematic, because we do not expect gene expression data to 'try' to confuse the method by steering it toward the opposite inference of the one it should make. However, we note that as long as the ratio of signal to adversarial noise remains >1 we can still make correct predictions with high probability.

Figure 2b shows that small perturbations to the causal graph do not significantly affect its ability to make correct inferences. Thus, even though some of the edges in our causal graph may be incorrect (or incorrect for the particular context we are investigating), so that the 'correct' graph looks like a version with a number of edges deleted, we can still make meaningful predictions. The same holds true if other edges are missing, which is likely to be the case due to the incomplete nature of our current biological knowledge. On the other hand, we also establish that the important information does not lie in the connectivity of the causal graph alone; if that had been the case, we would have been able to make equally good predictions using the randomly generated graph with the same in- and out-degree distribution as our causal graph.

The robustness of our method is a desirable feature, and as new knowledge is added to or removed from the causal graph, we will be able to reassess the effects that this has on its robustness.

2.8 Alternative assessment of significance

Very recently, the company Selventa released a white paper on their website providing more details of the approach outlined in (Pollard *et al.*, 2005). We discuss similarities and differences between the approaches in this section.

Fundamentally, the approaches utilize the same computational model of a causal graph and attempt to detect upstream regulators. However, no assessments of robustness of the model in the presence of noise are given in the white paper. Furthermore, Selventa's approach does not provide an aggregate score for each hypothesis, instead ranking them by their statistical properties. These statistical properties are discussed below in greater detail. Since the score of a hypothesis, defined as the number of correct predictions minus the number of incorrect predictions, is a useful indicator for the goodness-of-fit of the hypothesis to the observed data, we believe that it is valuable to provide this easily understandable information to the user,

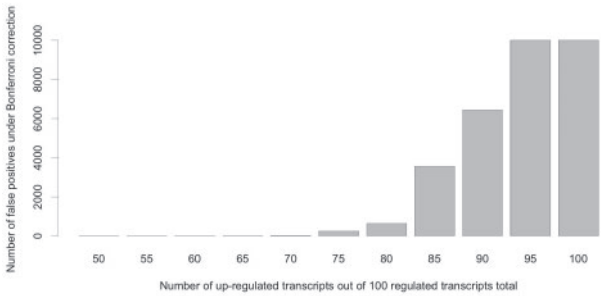

Fig. 3. Overestimation of significance for highly unbalanced hypotheses based on the concordance *P*-value.

in addition to the less directly interpretable *P*-values. In our experience, although a high score does not always result in a meaningful hypothesis, meaningful hypotheses always score highly. This score enabled us to derive the properties of our method in the presence of noise and demonstrate that a significant amount of noise can be tolerated given the current causal graph.

Finally, the statistical measures we compute take more information into account than the statistical measures computed by Selventa's approach. It is easy to see that the richness of a hypothesis is identical to the enrichment *P*-value, P_E, as defined in our methods. Therefore, we will focus on comparing the concordance value of a hypothesis and the *P*-value, *P*, as defined in our methods. We begin by noting that, using our notation, the concordance can be rewritten as

$$C := \frac{1}{2^{n_{++}+n_{--}}} \sum_{j \geq n_{++}+n_{--}} \binom{n_+ + n_-}{j}.$$

In other words, the concordance of a hypothesis is the probability of getting at least the same number of correct predictions out of the number of predictions made by the hypothesis. The concordance is a fundamentally different metric as it does not take into account how many transcripts are regulated in the experimental data. This information is crucial in order to assess whether the overlap of the experimental data with the hypothesis is significant. Consequently, the concordance must be used in conjunction with enrichment to be useful. In contrast, our correctness *P*-value is a direct generalization of enrichment that takes the directionality of regulation into account, and it reduces to the enrichment in the special case when all regulation is of the same sign.

Perhaps even more importantly, the concordance defined in Selventa's white paper does not take into account any imbalance in the number of

up- and down-regulated transcripts. Specifically, we consider a setting with 100 transcripts, and a hypothesis H which makes only unambiguous predictions (i.e. $|S_H^{\pm}|=100, |S_H^0|=0$). We generate K expression data sets $G_R^{\pm}$ uniformly at random, with the constraints $n_+ := |G_R^+| = 80$ and $n_- := |G_R^-| = 20$. In each of the K trials, we compute the resulting concordance and correctness P-values for H. In this scenario, we expect no significant results at a P-value threshold of $0.05/K$, according to the Bonferroni correction method for multiple testing. Figure 3 shows the number of erroneously significant findings (false positives) for hypotheses with q_+ (the number of transcripts predicted to be upregulated) ranging from 50 to 100, q_- (the number of transcripts predicted to be downregulated) set to $100-q_+$ and K set to 10 000. It demonstrates how the concordance P-value can severely overestimate significance in an unbalanced case. In contrast, our correctness P-value correctly controls the false positive rate under the null hypothesis, giving no false positives as expected. Conversely, the concordance P-value will severely underestimate significance if the experimental data shows the opposite imbalance for a given hypothesis.

3 IMPLEMENTATION

The analysis results are communicated using the Causal Reasoning Browser, a Java application based on the open-source biological network viewer Cytoscape (Shannon *et al.*, 2003). We have developed a plugin that enables browsing, clustering, merging and filtering of all statistically significant predicted upstream hypotheses in conjunction with the relevant subgraphs of our causal graph. Each subgraph (Fig. 4) can be evaluated by the user to assess the validity of a specific hypothesis. An overview graph allows users to quickly visualize hypotheses and shows how they are related to each other. Users are able to edit and manipulate the overview graph to group hypotheses together in the context of existing literature. The relatedness between hypotheses is displayed in the form of similarity statistics in a hypothesis table. The browser enables users to merge two or more hypotheses as a first step toward building larger explanatory subnetworks. Merging maintains the edge and node information and thus allows users to hover over edges to show their sources, providing additional investigative information.

4 BIOLOGICAL RESULTS

4.1 Validation: recovering known perturbations

In order to test the performance of the causal reasoning algorithm on a biological dataset we sought out experimental data which had a single, well-defined perturbation that should be identified by the algorithm. (Bild *et al.*, 2006) used recombinant adenoviruses to infect non-cancerous human mammary epithelial cells with a construct to overexpress one of five oncogenes; c-Myc, H-Ras, c-Src, E2F3 and β-catenin. DNA microarray analysis was performed to identify gene expression signatures for each of the oncogenes which were further refined by a supervised classification method to select genes that differentiated each oncogene from the others, see methods in (Bild *et al.*, 2006). Importantly, the data from this paper was not present in our collection of causal interactions, making it a suitable test set. The gene expression signatures contained 62 (c-SRC), 72 (β-catenin), 196 (c-Myc), 223 (H-Ras) and 238 (E2F3) transcripts, respectively. Of these, 31 (50%, c-SRC), 28 (38%, β-catenin), 153 (78%, c-Myc), 190 (85%, H-Ras) and 186 (78%, E2F3) could be matched with transcripts in our causal graph.

For three signatures (c-Myc, H-Ras and E2F3), either the overexpressed protein or a protein immediately downstream from it is correctly identified by the causal algorithm when looking at the top-ranked predicted hypothesis (Table 1). c-SRC and β-catenin had few matching genes. Our method did not return significant results in those cases, meaning that no confident predictions were possible.

For both the Myc and Ras expression signatures, these proteins are returned as significant regulators of the expression data by our causal analysis. MYC+ is the top causal result from the Myc signature at one interaction away from the expression data ($P=2\cdot10^{-14}$). The causal graph correctly links the up-regulation of Myc to $C=23$ of the expression changes.

For the E2F3 expression data, the E2F family is returned as a significant result, as is E2F1, but not E2F3; see Table 1. E2F1 and E2F3 have a very similar role as transcription factors that function to control the cell cycle and are similarly implicated in cancer (Chen *et al.*, 2009). However, the top hypothesis is CDKN2A− (also known as ARF). The ARF tumor suppressor is a key component of the p53 tumor surveillance pathway and is repressed by E2F3 in normal cells (Aslanian *et al.*, 2004), which supports ARF as our causal hypothesis and provides a mechanistic link back to E2F3 even if it has not been reported as significant.

HRAS+ is returned as a significant hypothesis ($P=4\cdot10^{-9}$) in the H-Ras dataset and is ranked 10th. The top hypothesis is tumor necrosis factor alpha (TNF+), which was also up regulated in the original H-Ras expression signature. HRAS induces the expression of TNF, and taken together they are able to explain a large portion (38%) of the gene expression signature. TNF is multi-functional, having a role in inflammatory responses, cell proliferation, differentiation, apoptosis, lipid metabolism and coagulation. (Suganuma *et al.*, 1999) demonstrated that TNF is essential for tumorigenesis. Causal reasoning correctly found that activating H-Ras increases TNF levels, promoting tumor formation.

These data provide evidence that causal reasoning can accurately detect the underlying cause of a biological gene expression signature and identify regulatory modules from within a larger, more complex dataset. We are also able to differentiate the immediate transcriptional events resulting from pathway activation from the subsequent downstream activity of secondary responses.

4.2 Example use case: causal drivers of cardiac hypertrophy

In this section, we use causal reasoning to compare myocardial gene expression changes associated with isoprenaline-induced (pathological) hypertrophy to those associated with exercise-induced (adaptive) hypertrophy in mice. The raw data was obtained from the public domain (Galindo *et al.*, 2009). With the filters set at $P\leq0.05, p_E\leq0.05$ and $C\geq3$, causal reasoning analysis generates 117 and 207 significant hypotheses for the two datasets, respectively. We arrive at a set of testable hypotheses by starting with the highly significant hypotheses and building out subnetworks based on current biological knowledge.

In the isoprenaline group the highest-ranking hypothesis is *response to hypoxia*+, with $P=2.07\cdot10^{-13}, p_E=7.22\cdot10^{-65}$ and $C=122$, whereas the highest-ranking hypothesis in the exercise group is *response to hypoxia*−, with $P=5.3\cdot10^{-3}, p_E=2.97\cdot10^{-32}$ and $C=140$. The cascade of downstream biological events in the subnetworks generated by our approach provides evidence that hypoxia is a major event. Note that an enrichment captured by P_E for *response to hypoxia* seems present in both cases, but

Table 1. The top five causal hypotheses from the three oncogene expression signatures (Bild *et al.*, 2006) are shown in the table

	Myc					E2F3					H-Ras				
No.	Gene	C	I	A	*P*	Gene	C	I	A	*P*	Gene	C	I	A	*P*
1	MYC +	23	1	1	2e−14	CDKN2A −	13	1	1	3e−9	TNF +	47	11	6	1e−15
2	ZBTB16 −	10	0	0	4e−11	E2F1 +	11	1	0	8e−6	IL1B +	32	4	1	5e−15
3	ALK +	9	0	0	3e−12	E2F family +	5	0	0	7e−5	F2 +	27	4	0	4e−16
4	TP53 −	12	4	0	2e−3	PROX1 +	4	0	0	4e−6	EGF +	26	5	0	1e−12
5	HDAC6 −	3	0	0	6e−5	ITGB1 −	3	0	0	6e−5	TGFB1 +	31	10	2	5e−8
10	…					…					HRAS +	19	4	0	5e−9

C is the number of transcript changes correctly explained by the hypothesis, I incorrectly and A ambiguously. A +/− indicates the inferred directionality of the hypothesis.

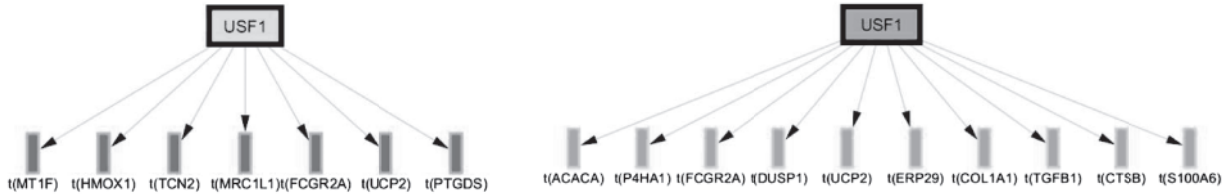

Fig. 4. Example demonstrating opposite predicted effect of USF1 protein abundance or activity based on dissimilar transcript changes (15% overlap). Unaided, such an inference is very difficult from the gene expression data alone. Data taken from the cardiac hypertrophy use case.

a clear directional effect (*response to hypoxia*+) is only asserted in the isoprenaline group. In the exercise group the direction is captured as opposite (*response to hypoxia*-) but the confidence is lower. Biologically, this is a plausible prediction as the effects of isoprenaline are expected to be severe and pathologic with little or no compensation. The effects of exercise, however, should be weaker and adaptive in nature through physiological feedback loops. The analysis shows a biological network that captures several hallmarks of cardiac diseases and cardiomyocyte stress signaling (Aragno *et al.*, 2008; der Heiden *et al.*, 2010; Rona, 1985; Teekakirikul *et al.*, 2010; Wölkart *et al.*, 2006). The subnetwork demonstrates a cluster of increased hypoxia ($C = 146$, or ∼19% of the total). Tissue hypoxia triggers increased activity of the aldosterone/angiotensin axis evidenced by *AGT*+, *angiotensin II*+ and *aldosterone*+, and decreased angiotensin antagonists *losartan*− and *captopril*− ($C = 39$). Subsequently, increased nitric oxide synthase (NOS) production leads to oxidative stress (five hypotheses, $C = 71$). Oxidative stress together with hypoxia can result in a cell stress positive feedback loop that additionally includes an inflammatory response (10 hypotheses, $C = 102$) and endoplasmic reticulum stress (2 hypotheses, $C = 29$). Eventually, the self-reinforcing cell stress signals culminate in (i) DNA damage and apoptosis, and (ii) myogenic disruption and structural remodeling supported by an extensive network that includes *TGF*β+ signaling, *wounding*+ and *pulmonary fibrosis*+, *dystrophin*− and *myogenic differentiation-1 transcription factor*−.

In contrast, the causal subnetwork for exercise-induced hypertrophy demonstrates a number of mechanisms that are generally the reverse of those inferred from the isoprenaline group. *Hypoxia*− and *ischemia*− hypotheses reflect the ability of the cells to compensate for increased cardiac load and enhanced oxygenation ($C = 163$, or ∼11% of the total). *Endothelin*− (EDN1) and *angiotensin II*− are inferred downstream of hypoxia ($C = 53$). Contrary to isoprenaline's extensive pro-inflammatory subnetwork, the exercise group exhibits a strong anti-inflammatory subnetwork (10 hypotheses, $C = 173$). The endoplasmic reticulum (ER) stress subnetwork in the exercise group contains *XBP1*− and *POR*+, while the opposite is true of the isoprenaline group. The DNA damage subnetwork shows hypotheses in the opposite direction to that of isoprenaline. However, *FAS*+ and *mitochondrial DNA damage*+ may reflect some degree of cell stress as well.

Several similarities and differences can be noted between our causal reasoning analysis and the analysis from the original study of this public dataset (Galindo *et al.*, 2009). The overall analysis in the original study was similar in that it referred to pathways captured by causal reasoning in a higher level of detail, such as acute phase response, fibrosis, oxidative stress and cell morphology. A major difference is that causal reasoning generates hypotheses on similar pathways, but with opposite inferred directionality. Note that the inferred direction of the same upstream entity rely on disparate gene sets (Fig. 4). Also, the level of detail provided immediately suggests hypotheses for experimental follow-up.

5 DISCUSSION

In this article, we have presented a comprehensive methodology to derive concise molecular hypotheses that can explain the transcriptional changes observed in genome-wide microarrays. We have compiled a large causal graph of curated interactions, to which we have successfully applied a simple but powerful reasoning method. Our method is very similar in spirit to the method of (Pollard *et al.*, 2005). However, we have shown here for the first time how to analytically compute the statistical significance of detected hypothetical drivers of the expression data.

Given the importance of current knowledge for the creation of the causal network, we investigated how its characteristics such as redundancy and coverage impact our ability to retrieve clear signals. In order to understand this based on our current causal graph of over 250 000 unique causal relationships, we defined the concept of recoverability and derived an analytical method to assess it for

any causal graph. Applying it to our causal graph clearly showed that upstream drivers can be recovered within a wide range of noise parameters. Furthermore, the signal is not due to the connectivity structure of the graph alone, as evidenced by a comparison to a random graph with the same in- and out-degree distribution.

To assess the biological validity of the approach, we have also applied causal reasoning to two biological datasets. The first had over-expressed various oncogenes in a normal mammary epithelial cell line on which microarray analysis was performed. Causal reasoning was able to correctly identify the overexpressed oncogene as the cause of the expression changes, and delimit downstream pathways. This echoes the results of the tests on simulated data and gives us confidence that causal reasoning is an effective methodology to interrogate gene expression data. The second biological use case compared transcriptional changes in isoprenaline- and exercise-induced hypertrophy and demonstrates the power of causal reasoning on an experimental dataset. We were able to identify known hallmarks of disease as well as novel differences between the two forms of hypertrophy. In comparison to traditional methods, such as gene set enrichment, our method provides considerable additional insights.

There are several limitations to the method described here. We are certainly limited by the amount of biological knowledge currently available in the form of scientific publications, so our method is unlikely to elucidate completely unexplored areas of biology. We are also limited by the incomplete translation of scientific information from the literature into computationally usable causal relationships, whether due to the labor-intensiveness of the translation process or to the more complex nature of many biological relationships that goes beyond simple causality. The impact of this limitation is therefore expected to decrease as the coverage of our causal graph increases, and targeted curation efforts can be made to improve the breadth and the depth of causal information for specific high value datasets. The output produced by our method is a ranked list of succinct molecular hypotheses with all of the associated evidence to allow deep biological review and interpretation. The ability to critically review the model is important, as the results presented here demonstrate. For example, the causal analysis of the E2F3 oncogene expression data did not report E2F3 itself as a significant hypothesis, but did return E2F1, a closely related member of the same family of transcription factors. E2F1 and E2F3 have a similar role and are similarly implicated in cancer (Chen *et al.*, 2009).

We are currently pursuing several extensions of our methods. First, our scoring function naturally extends to combinations of upstream nodes, and we are exploring optimization procedures to suggest such combinations to the computational biologists. Combinations of hypotheses will present new mathematical and algorithmic challenges. However, the biggest challenge will be to combine hypotheses in a way meaningful to our end users in the biology field. Furthermore, we currently rely on external sources for the interaction data (Ingenuity and Selventa). In the future, we would like to allow biologists to add to and edit this data as they review our models, so that we could slowly build up a much more comprehensive and accurate set of interactions in a decentralized way. Finally, the approach we described provides a natural framework to enable the integrated analysis of multiple data types such as proteomics, metabolomics and genetics. We would like to extend the causal reasoning approach to work with heterogeneous data as more and more of it becomes available.

We believe that the appropriate encoding of the current state of the biomedical literature and the routine application of methods to interpret results from new experiments in light of what is already known will guide the way to novel insights by reevaluating known facts in different contexts or identifying unexplored areas of biology.

In our experience, the output of our method was easy to interpret for biologists, and several hypotheses have already been selected for follow-up. We hope that the discovery of novel biology and the enrichment of the causal graph will lead to a virtuous cycle and a continued expansion of the boundaries of biological knowledge.

ACKNOWLEDGEMENT

The authors would like to thank Ketan Patel for help with the selection of the validation data sets.

Funding: All authors were funded (as employees or contractors) by Pfizer, Inc.

REFERENCES

Aragno,M. *et al*. (2008) Oxidative stress triggers cardiac fibrosis in the heart of diabetic rats. *Endocrinology*, **149**, 380–388.

Aslanian,A. *et al*. (2004) Repression of the arf tumor suppressor by e2f3 is required for normal cell cycle kinetics. *Gene. Dev.*, **18**, 1413–1422.

Bild,A.H. *et al*. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.

Chen,H.-Z. *et al*. (2009) Emerging roles of e2fs in cancer: an exit from cell cycle control. *Nat. Rev. Cancer*, **9**, 785–797.

der Heiden,K.V. *et al*. (2010) Role of nuclear factor kappab in cardiovascular health and disease.*Clin. Sci.*, **118**, 593–605.

Draghici,S. *et al*. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Galindo,C.L. *et al*. (2009) Transcriptional profile of isoproterenol-induced cardiomyopathy and comparison to exercise-induced cardiac hypertrophy and human cardiac failure. *BMC Physiol.*, **9**, 23.

Hartemink,A.J. *et al*. (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks.*Pac. Symp. Biocomput.*, 422–433.

Kanehisa,M. *et al*. (2010) Kegg for representation and analysis of molecular networks involving diseases and drugs.*Nucleic Acids Res.*, **38**, D355–D360.

Petkovšek,M. *et al*. (1996) *A=B*. Wellesley, MA: A K Peters.

Pollard,J. *et al*. (2005) A computational model to define the molecular causes of type 2 diabetes mellitus. *Diabetes Technol. Ther.*, **7**, 323–336.

Rao,A.R. *et al*. (1996) A Markov chain Monte Carlo method for generating random (0, 1)-matrices with given marginals. *Indian J. Stat.*, **58**, 225–242.

Rona,G. (1985) Catecholamine cardiotoxicity. *J. Mol. Cell. Cardiol.*, **17**, 291–306.

Schadt,E.E. *et al*. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.*, **37**, 710–717.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Suganuma,M. *et al*. (1999) Essential role of tumor necrosis factor alpha (TNF-alpha) in tumor promotion as revealed by TNF-alpha-deficient mice. *Cancer Res.*, **59**, 4516–4518.

Teekakirikul,P. *et al*. (2010) Cardiac fibrosis in mice with hypertrophic cardiomyopathy is mediated by non-myocyte proliferation and requires tgf-beta. *J. Clin. Invest.*, **120**, 3520–3529.

Wölkart,G. *et al*. (2006) Role of endogenous hydrogen peroxide in cardiovascular ischaemia/reperfusion function: studies in mouse hearts with catalase-overexpression in the vascular endothelium. *Pharmacol. Res.*, **54**, 50–56.

Zien,A. *et al*. (2000) Analysis of gene expression data with pathway scores. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 407–417.