

# Causal Reasoning on Biological Networks: Interpreting Transcriptional Changes (Extended Abstract)

Leonid Chindelevitch<sup>1</sup>, Daniel Ziemek<sup>1,\*</sup>, Ahmed Enayetallah<sup>2</sup>,  
Ranjit Randhawa<sup>1</sup>, Ben Sidders<sup>3</sup>, Christoph Brockel<sup>4</sup>, and Enoch Huang<sup>1</sup>

<sup>1</sup> Computational Sciences Center of Emphasis,  
Pfizer Worldwide Research and Development, Cambridge, MA, USA

<sup>2</sup> Compound Safety Prediction,  
Pfizer Worldwide Medicinal Chemistry, Groton, CT, USA

<sup>3</sup> eBiology, Pfizer Worldwide Research and Development, Sandwich, Kent, UK

<sup>4</sup> Translational and Bioinformatics,  
Pfizer Business Technologies, Cambridge, MA, USA

## 1 Introduction

Over the past decade gene expression data sets have been generated at an increasing pace. In addition to ever increasing data generation, the biomedical literature is growing exponentially. The PubMed database (Sayers et al., 2010) comprises more than 20 million citations as of October 2010. The goal of our method is the prediction of putative upstream regulators of observed expression changes based on a set of over 400,000 causal relationships. The resulting putative regulators constitute directly testable hypotheses for follow-up.

## 2 Methods

In order to find those regulators, we first construct a *causal graph*  $G_C$  whose nodes are transcript levels, compound concentrations, and states of biological processes. To represent causality, each node appears twice, once with a + sign (upregulation) and once with a - sign (downregulation). A directed edge from node  $a$  to node  $b$  means that the abundance or activity of  $b$  is regulated by the abundance or activity of  $a$ . 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, to facilitate hypothesis validation for the scientists. We licensed the substrate for our method from two vendors: Ingenuity Inc. and Genstruct Inc. This yields 250,000 unique relationships covering 65,000 full-text articles indexed by PubMed. Pollard et al., 2005 presented a similar approach, but did not provide any details on implementation.

The gene expression data determines the subset  $G^+$  of all genes that are significantly overexpressed and the subset  $G^-$  of all genes that are significantly

---

\* Joint first author, corresponding author: daniel.ziemek@pfizer.com

underexpressed. We define  $G^\pm := G^+ \cup G^-$ . We choose a distance threshold  $\Delta$  which determines the maximum length of the paths we consider. Given a hypothesis  $h \in V(G_C)$ , we classify each node of  $G_C$  into one of three possible sets:  $S_h^+ := \{v \in V(G_C) | d(h, v) \leq \Delta, d(h, v) < d(h, -v)\}$ ,  $S_h^- := \{v \in V(G_C) | d(h, -v) \leq \Delta, d(h, -v) < d(h, v)\}$ ,  $S_h^0 := \{v \in V(G_C) | d(h, v) > \Delta \text{ or } d(h, v) = d(h, -v)\}$ , where  $d(\cdot, \cdot)$  is the distance between two nodes in the graph  $G_C$ . 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  $G^\pm$ . We define  $n_{\sigma, \tau} := |S_h^\sigma \cap G^\tau|$  for  $\sigma, \tau \in \{+, -\}$ . That is, the score of hypothesis  $h$  is  $s(h, G^\pm) = n_{++} + n_{--} - n_{+-} - n_{-+}$ .

However, a good score does not necessarily mean good explanatory power, because of possible connectivity differences between the nodes of  $G_C$ . Therefore we also look at statistical significance. For a given hypothesis  $h$  and a given score  $s_0 := s(h, G^\pm)$ , we would like to compute the probability of  $h$  scoring  $s_0$  or better with a *random* set of genes  $G_R^\pm := G_R^+ \cup G_R^-$ , chosen with  $|G_R^+| = |G^+|$  and  $|G_R^-| = |G^-|$ . We have developed a method for computing this probability in time cubic in  $|G^\pm|$ .

When processing a particular data set, our algorithm begins by computing the scores for each hypothesis and ranks the set of all hypotheses by their score. The correctness  $p$ -value  $p$  of a hypothesis is typically required to be below a certain threshold. The enrichment  $p$ -value  $p_E$  of a hypothesis is also required to pass 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 and represents a standard measure in gene set overrepresentation methods (e.g. Draghici et al., 2003). Finally, we may also filter out those hypotheses whose number of correct predictions,  $C := n_{++} + n_{--}$ , is below a certain user-defined threshold.

**Table 1.** The top five causal hypotheses from the three oncogene expression signatures (Bild et al., 2006) are shown in the table, where C is the number of transcript changes correctly explained by the hypothesis; I incorrectly & A ambiguously. A +/- indicates the inferred directionality of the hypothesis.

Myc						E2F3						H-Ras								
Gene	Rank	Score	C	I	A	$p$	Gene	Rank	Score	C	I	A	$p$	Gene	Rank	Score	C	I	A	$p$
MYC+	1	22	23	1	1	$2 \cdot 10^{-14}$	CDKN2A -	1	12	13	1	1	$3 \cdot 10^{-9}$	TNF +	1	36	47	11	6	$1 \cdot 10^{-15}$
ZBTB16 -	2	10	10	0	0	$4 \cdot 10^{-11}$	E2F1 +	2	10	11	1	0	$8 \cdot 10^{-6}$	IL1B +	2	28	32	4	1	$5 \cdot 10^{-15}$
ALK +	3	9	9	0	0	$3 \cdot 10^{-12}$	E2F family +	3	5	5	0	0	$7 \cdot 10^{-5}$	F2 +	3	23	27	4	0	$4 \cdot 10^{-16}$
TP53 -	4	8	12	4	0	$2 \cdot 10^{-3}$	PROX1 +	4	4	4	0	0	$4 \cdot 10^{-6}$	EGF +	4	21	26	5	0	$1 \cdot 10^{-12}$
HDAC6 -	5	3	3	0	0	$6 \cdot 10^{-5}$	ITGB1 -	5	3	3	0	0	$6 \cdot 10^{-5}$	TGFBI +	5	21	31	10	2	$5 \cdot 10^{-8}$
...							...							HRAS +	10	15	19	4	0	$5 \cdot 10^{-9}$

### 3 Validation and Results

Using simulations we established that our method is able to recover embedded regulators given our causal graphs with high-accuracy in the presence of noise. In order to test the performance of the causal reasoning algorithm on a biological data set 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  $\beta$ -catenin. The data from this paper was not present in the causal interaction knowledge base when we applied our causal reasoning algorithm to these published signatures. For three signatures (c-Myc, H-Ras, E2F3) either the overexpressed protein or a protein immediately downstream from it, is correctly identified by our algorithm as the top-ranked predicted hypothesis (Table 1). c-SRC and  $\beta$ -catenin had very few matching genes. Our method did not return highly significant results in those cases, meaning that no confident predictions were possible.

We also used our algorithm to compare myocardial gene expression changes associated with isoprenaline-induced (pathological) hypertrophy with exercise-induced (adaptive) hypertrophy in mice, obtained from the public domain (Galindo et al., 2009). In the isoprenaline group, the analysis supports biological networks of several hallmarks of cardiac disease and cardiomyocyte stress (e.g. Aragno et al., 2008). These include hypotheses indicative of increased hypoxia, increased NOS production, oxidative stress, inflammatory response and endoplasmic reticulum stress. In contrast, the exercise-induced hypertrophy demonstrates perturbation of the same biological networks as in the isoprenaline group but with reversed direction of regulation, e.g. decreased hypoxia.

The outlined results provide evidence that method based on the outlined score and statistical measures can accurately detect the underlying cause of a biological gene expression signature and identify regulatory modules from within a larger, more complex data set. In our experience the output of our method was easy to interpret for biologists, and several hypotheses have already been selected for follow-up. It is our hope that the interplay between experimental work based on our method, the discovery of novel biology and the subsequent enrichment of the causal graph will lead to a virtuous cycle allowing for the continued expansion of the boundaries of biological knowledge.

## References

- [Aragno et al.,2008] Aragno, M., Mastrocola, R., Alloatti, G., Vercellinatto, I., Bardini, P., Geuna, S., Catalano, M.G., Danni, O., Boccuzzi, G.: Oxidative stress triggers cardiac brosis in the heart of diabetic rats. *Endocrinology* 149(1), 380–388 (2008)
- [Bild et al., 2006] Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M.-B., Harpole, D., Lancaster, J.M., Berchuck, A., Olson, J.A., Marks, J.R., Dressman, H.K., West, M., Nevins, J.R.: Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439(7074), 353–357 (2006)
- [Draghici et al., 2003] Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C., Krawetz, S.A.: Global functional profiling of gene expression. *Genomics* 81(2), 98–104 (2003)
- [Galindo et al., 2009] Galindo, C.L., Skinner, M.A., Errami, M., Olson, L.D., Watson, D.A., Li, J., McCormick, J.F., McIver, L.J., Kumar, N.M., Pham, T.Q., Garner, H.R.: Transcriptional profile of isoproterenol-induced car- diomyopathy and comparison to exercise-induced cardiac hypertrophy and human cardiac failure. *BMC Physiol.* 9, 23 (2009)

- [Pollard et al., 2005] Pollard, J., Butte, A.J., Hoberman, S., Joshi, M., Levy, J., Pappo, J.: A computational model to define the molecular causes of type 2 diabetes mellitus. *Diabetes Technol. Ther.* 7(2), 323–336 (2005)
- [Sayers et al., 2010] Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Lu, Z., Madden, T.L., Madej, T., Maglott, D.R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Wang, Y., Wilbur, W.J., Yaschenko, E., Ye, J.: Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 38(Database issue), D5–D16 (2010)