A graph theoretic method for assessing topological perturbations in Biological Networks

Florian Martin[1], Alain Sewer[1], Ty Thomson[2], David Drubin[2], Dexter Pratt[2], Andrea Matthews[2], Renee Kenney[2], David de Graaf[2], Julia Hoeng[1], Manuel C. Peitsch[1]
[1]Philip Morris Products SA, Neuchâtel, Switzerland
[2]Selventa, Cambridge, MA, USA

PS 356

Introduction

The description of cellular processes and the quantitative analysis of their perturbations is a crucial step towards understanding disease. It is through the combination of prior knowledge captured in biological network models with high-throughput experimental data that one can explore and understand how cellular processes are impacted by external stressors such as exposure to a broad variety of substances.

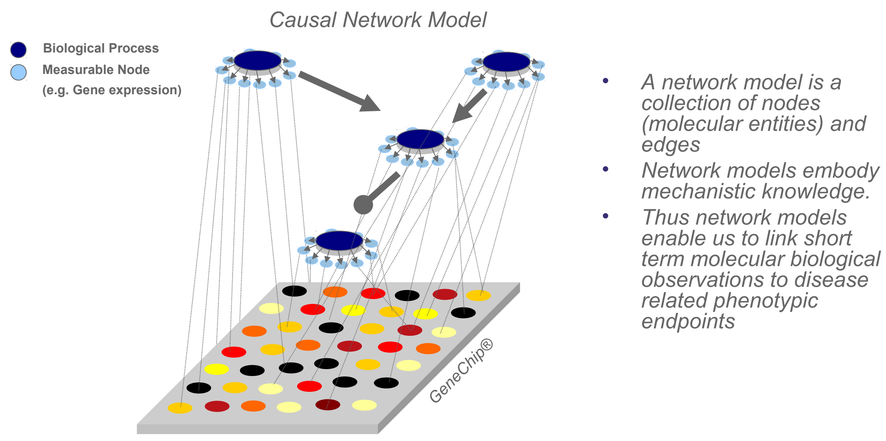
Objective

When using causal network models constructed using Selventa® Knowledge Assemblies [1], not all biological entities in the model can be linked to experimental evidence and, when specific experimental data are gathered, the network will likely be unequally perturbed due to the specific biology represented in the experiment.

This motivated the development of a methodology to identify the most likely perturbed regions of the network by explicitly using the topology.

Materials and methods

Network models We are using network models constructed using Selventa® Knowledge Assemblies [1] to describe non-kinetic causal relationships between biological processes. In such network models, some nodes—termed "hypotheses"—are associated with a set of genes which correspond to the downstream targets of the process described by the node. The agreement between the behavior contained in the model and the behavior observed at the gene expression level in a particular experiment allows us to quantify the activity of the corresponding "hypothesis".



- *A network model is a collection of nodes (molecular entities) and edges*
- *Network models embody mechanistic knowledge.*
- *Thus network models enable us to link short term molecular biological observations to disease related phenotypic endpoints*

Node scoring The first step is to define a perturbation index (PI) of the nodes that are upstream controllers of genes (denoted by $\mathcal{I}$). The PI's give an information about the evidence that the underlying process has been activated (either up or down). The gene expression data are used to estimate the systems response β by mean of a contrast (e.g. Treatment vs. Control). Using the false non-discovery rate (fndr) together with estimated β one computes:

$$PI \propto \| \, fndr \cdot \beta \, \|_1 \quad (1)$$

Random walk and centralities For various reasons (biology, literature model,…), not all "causal" path will be activated uniformly.
Given the graph topology G=(V,E) and the fact that edges represents causality, the idea is to reinforce the simple random walk (SRW, uniform perturbation) toward nodes with more experimental evidence about their activity in the network (high PI values).
The propagation operator of the reinforced random walk is defined by

$$M \in l^2(V) \text{ is defined by } M_{ij} \propto \begin{cases} \frac{1}{n}(1+100\cdot PI_j) & \text{if } i \to j \text{ and } i \in \mathcal{I} \\ \frac{1}{n} & \text{if } i \to j \text{ and } i \notin \mathcal{I} \\ 0 & \text{else} \end{cases} \quad (2)$$

Note: To ensure irreducibility of the chain nodes with zero out-degree randomly jump to nodes with zero in-degree

Then one infers the topological importance of each node in the model with respect to the sequences of causal statements that might occur (i.e path in the graph).

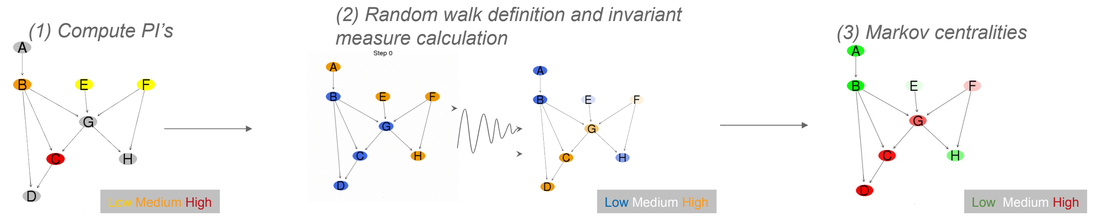
A node (biological process) that is more likely to be perturbed will be more central in the network i.e. the flow of causalities often imply the activation of the biological process. The notion of centrality, called *Markov Centrality* (similar to the notion of [2]), is

$$\mathcal{C}(j) = \sum_{i=1,..,n} (\frac{\pi_j}{\pi_i} G_{ii} - G_{ij}) = \sum_{i=1,..,n} \mathbb{E}_\pi(\text{number of visit to } j \text{ before time } T_i) \quad (3)$$

Where π is the invariant measure of the random walk, $\mathbb{E}_\pi(\text{number of visit to } j \text{ before time } T_i) = \frac{\pi_j}{\pi_i} G_{ii} - G_{ij}$

and $G = \sum_{n \geq 0} (M^n - M^\infty)$

Steps for centrality computation



Topological perturbation Finally, the topological reinforcement of a node due to the perturbation is defined as the log-ratio of the reinforced centrality and the simple random walk (no data) centrality

$$R(j) = \log_{10} \frac{\mathcal{C}(j)}{\mathcal{C}^{SRW}(j)}$$

Additionally one may want to discover which node with a PI is most influencing the centrality ratio of a given node of interest (e.g. a node corresponding to a phenotypic outcome). This can be achieve by computing

$$\frac{\partial \mathcal{R}(j)}{\partial PI_k} = \frac{\partial \mathcal{C}(j)/\partial PI_k}{\mathcal{C}(j)} - \frac{\partial \mathcal{C}^{SRW}(j)/\partial PI_k}{\mathcal{C}^{SRW}(j)}$$

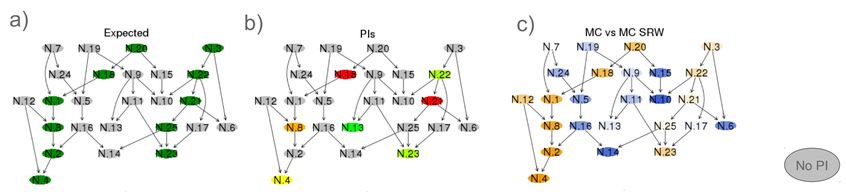
Where $\frac{\partial \mathcal{C}(j)}{\partial PI_k} = \frac{\partial diag(G)}{\partial PI_k}(1/\pi)\pi^T + diag(G)\frac{\partial (1/\pi)}{\partial PI_k}\pi^T + diag(G)(1/\pi)\frac{\partial \pi^T}{\partial PI_k} - \frac{\partial G}{\partial PI_k}$

$\frac{\partial (1/\pi)}{\partial PI_k} = -(1/\pi^2)\cdot\frac{\partial \pi}{\partial PI_k}$ and $\frac{\partial G}{\partial PI_k} = G(\frac{\partial M}{\partial PI_k} - 1\cdot\frac{\partial \pi^T}{\partial PI_k})G$

Results

Illustration

As an illustration of the method, consider the simulated situation where the paths in dark green in panel a) represent the "true" perturbation. The panel b) shows the PI's (from low (**green**) to high (**red**)) and c) the resulting centrality ratios, showing the potential of the approach to unravel perturbed regions of the network (low (**blue**) to high (**orange**) log-ratio).

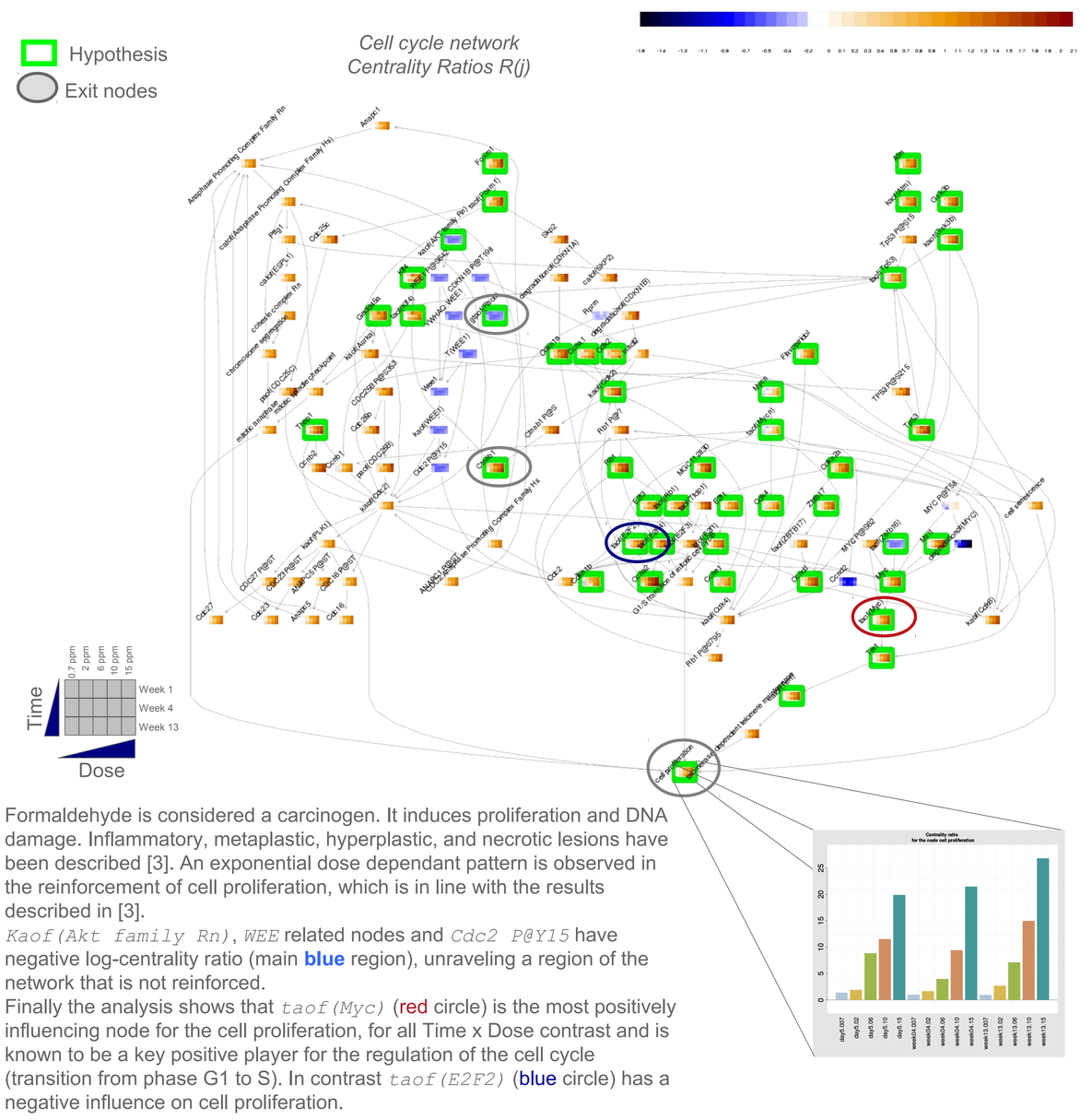


Application to formaldehyde exposure experiment in rat [3]

Eight week old male F344/CrlBR rats were exposed to formaldehyde through whole body inhalation. Whole body exposures were performed at doses of 0, 0.7, 2, 6, 10, and 15 ppm (6 hours per day, 5 days per week). Animals were sacrificed at 1, 4, and 13 weeks following initiation of exposure. Following sacrifice, tissue from the Level II region of the nose was dissected and digested with a mixture of proteases to remove the epithelial cells. The epithelial cells acquired from this section of the nose consisted primarily of transitional epithelium with some respiratory epithelium. Gene expression microarray analysis was performed on the epithelial cells [3].

Cell Proliferation model: Cell cycle subnetwork

To further a systems-level assessment of the biological impact of perturbations on nondiseased mammalian lung cells, a lung-focused causal network for cell proliferation was constructed by Westra et al. [4] which encompasses diverse biological areas that lead to the regulation of normal lung cell proliferation (Cell Cycle, Growth Factors, Cell Interaction, Intra- and Extracellular Signaling, and Epigenetics), and contains a total of 848 nodes (biological entities) and 1597 edges (relationships between biological entities). The network was verified using four published gene expression profiling data sets associated with measured cell proliferation endpoints in lung and lung-related cell types. Predicted changes in the activity of core machinery involved in cell cycle regulation (RB1, CDKN1A, and MYC/MYCN) are statistically supported across multiple data sets, underscoring the general applicability of this approach for a network-wide biological impact assessment using systems biology data. The Cell-cycle subnetwork is used here.



Formaldehyde is considered a carcinogen. It induces proliferation and DNA damage. Inflammatory, metaplastic, hyperplastic, and necrotic lesions have been described [3]. An exponential dose dependant pattern is observed in the reinforcement of cell proliferation, which is in line with the results described in [3].
`Kaof(Akt family Rn), WEE` related nodes and `Cdc2 P@Y15` have negative log-centrality ratio (main **blue** region), unraveling a region of the network that is not reinforced.
Finally the analysis shows that `taof(Myc)` (red circle) is the most positively influencing node for the cell proliferation, for all Time x Dose contrast and is known to be a key positive player for the regulation of the cell cycle (transition from phase G1 to S). In contrast `taof(E2F2)` (blue circle) has a negative influence on cell proliferation.

Summary and Conclusion

The proposed method is applied to the cell cycle sub network model for Rat Nasal Epithelial cells exposed to varying concentrations of formaldehyde for various time periods. The perturbed regions are identified and reveal a time- and dose-dependent reinforcement, but also reveals regions with opposite signs. Thus, the structure of the overall systems response hidden in the noisy behavior of thousands of downstream-controlled genes is elegantly captured by our approach.

The method provides a insightful way to describe global effects of external perturbations on a biological network by combining the knowledge contained in a causal model and the systems response measured by gene expression technology.

References

[1] White paper, *Reverse Causal Reasoning,* www.selventa.com
[2] S. White and P. Smyth., *Algorithms for estimating relative importance in networks. International Conference on Knowledge Discovery and Data Mining,* Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining(2):266-275,2003
[3] Andersen ME, Clewell HJ 3rd, Bermudez E, Dodd DE et al. *Formaldehyde: integrating dosimetry, cytotoxicity, and genomics to understand dose-dependent transitions for an endogenous compound*. *Toxicol Sci* 2010 Dec;118(2):716-31. PMID: 20884683
[4] Westra et al., *Construction of a Computable Cell Proliferation Network Focused on Non-Diseased Lung Cells*, BMC Systems Biology 2011, **5**:105

PMI RESEARCH & DEVELOPMENT

ICSB 2011
Heidelberg, Germany
August 28 – September 1

Philip Morris International Research & Development, Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland
T: +41 58 242 21 11, **F**: +41 58 242 28 11, **W**: www.philipmorrisinternational.com

SELVENTA™ Analyze. Accelerate. Clarify.