# A Novel Reverse Engineering Method for Gene/Protein Network Reconstruction - Divergence Weighted Independence Graphs

Xiang Yang[1], Talikka Marja[1], Belcastro Vincenzo[1], Sperisen Peter[1], Peitsch Manuel[1], Hoeng Julia[1], and Whittaker Joe[2]

[1]Philip Morris International R&D, Philip Morris Products S.A., Neuchâtel, Switzerland
[2]Department of Mathematics and Statistics, Lancaster University, UK

## Introduction

Identification of the interactions between molecular entities within cells is the key to understanding the biological processes involved. Unfortunately, it is difficult to identify these interactions entirely by experiments. Although numerous methods have been developed for inferring gene/protein regulatory networks from expression data, reliable network inference from gene/protein expression data remains an unsolved problem. Recently, a novel method, DWIG (divergence weighted independence graphs), was developed. A simulated data set with 160 virtual animals was generated from the mathematical model (Sedaghat et al., 2002) of insulin signaling pathway to evaluate the performance of DWIG. This simulated data characterized both the within-individual and between-individual variabilities, which other widely used public simulation data, such as the DREAM challenge, did not mimic fully. The performance of three reverse engineering methods, ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) (Margolin et al., 2006), Banjo (Bayesian Network with Java Objects) (Yu et al., 2004), and DWIG, were compared based on these simulated data. The area under the curve (AUC) of receiver operating characteristic (ROC) curve showed that DWIG outperformed ARACNE and Banjo. ARACNE uses the marginal mutual information, while DWIG uses the conditional mutual information, which could be the reason for DWIG's superior performance over ARACNE. After prefiltering some weak links out by ARACNE, DWIG was applied to a protein dataset consisting of cytokines and chemokines that were measured in bronchoalveolar lavage fluid (BALF) of female A/J mice exposed to cigarette mainstream smoke for 3 and 5 months. An association network with 25 cytokines/chemokines was built.

## Materials & Methods

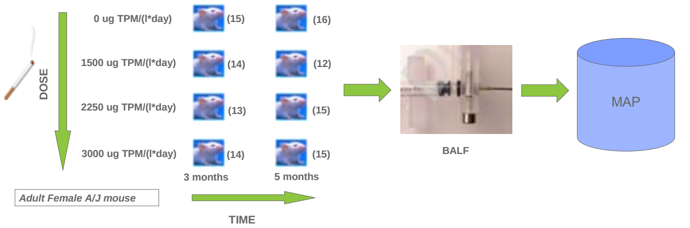
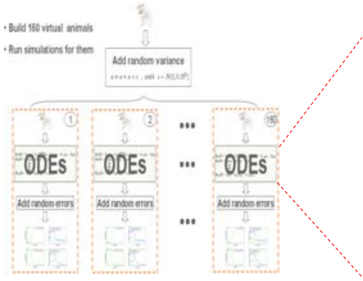


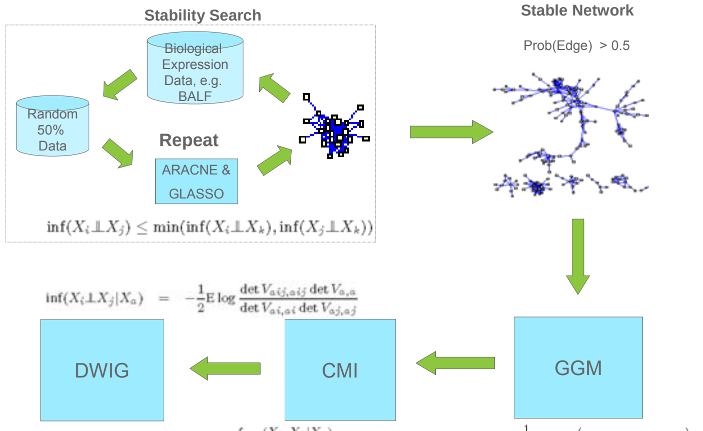
**Fig 1**. Female A/J mice, approximately 6 months of age, were exposed to diluted mainstream smoke from the Reference Cigarette 2R4F for 2, 3, and 4 h per day, 5 days per week for 3 and 5 months. The number of animals per group is in brackets. The mean smoke concentration throughout the study was 735 µg total particulate matter/l. All mice were exsanguinated under deep pentobarbital anesthesia. The tracheae were cannulated and the lungs lavaged with 1 ml of Ca$^{2+}$- and Mg$^{2+}$- free phosphate buffered saline (PBS). After centrifugation, the supernatants from the 1st lavage cycle were aliquoted in adequate volumes, stored below -60 ºC, and sent to RBM (Rules Based Medicine, http://www.rulesbasedmedicine.com/) for MAP analysis.



Mathematical Model of Insulin Pathway

$$
\begin{aligned}
x_1 &= \text{insulin input}\\
dx_2/dt &= k_{-1}x_3 + k_{-3}[\mathrm{PTP}]x_5 - k_1x_1x_2 + k_{-4}x_6 - k_4x_2\\
dx_3/dt &= k_1x_1x_2 - k_{-1}x_3 - k_3x_3\\
dx_4/dt &= k_3x_1x_5 - k_{-2}x_4 + k_{-4'}x_7 - k_{4'}x_4\\
dx_5/dt &= k_3x_3 + k_{-2}x_4 - k_3x_1x_5 - k_{-3}[\mathrm{PTP}]x_5 + k_{-4'}x_8 - k_{4'}x_5\\
dx_6/dt &= k_5 - k_{-5}x_6 + k_6[\mathrm{PTP}](x_7 + x_8) + k_4x_2 - k_{-4}x_6\\
dx_7/dt &= k_{4'}x_4 - k_{-4'}x_7 - k_6[\mathrm{PTP}]x_7\\
dx_8/dt &= k_{4'}x_5 - k_{-4'}x_8 - k_6[\mathrm{PTP}]x_8\\
dx_9/dt &= k_{-7}[\mathrm{PTP}]x_{10} - k_7x_9(x_4 + x_5)/(\mathrm{IR_p})\\
dx_{10}/dt &= k_7x_9(x_4 + x_5)/(\mathrm{IR_p}) + k_{-8}x_{12} - (k_{-7}[\mathrm{PTP}] + k_8x_{11})x_{10}\\
dx_{11}/dt &= k_{-8}x_{12} - k_8x_{10}x_{11}\\
dx_{12}/dt &= k_8x_{10}x_{11} - k_{-8}x_{12}\\
dx_{13}/dt &= k_9x_{14} + k_{10}x_{15} - (k_{-9}[\mathrm{PTEN}] + k_{-10}[\mathrm{SHIP}])x_{13}\\
dx_{14}/dt &= k_{-9}[\mathrm{PTEN}]x_{13} - k_9x_{14}\\
dx_{15}/dt &= k_{-10}[\mathrm{SHIP}]x_{13} - k_{10}x_{15}\\
dx_{16}/dt &= k_{-11}x_{17} - k_{11}x_{16}\\
dx_{17}/dt &= k_{11}x_{16} - k_{-11}x_{17}\\
dx_{18}/dt &= k_{-12}x_{19} - k_{12}x_{18}\\
dx_{19}/dt &= k_{12}x_{18} - k_{-12}x_{19}\\
dx_{20}/dt &= k_{-13}x_{21} - (k_{13} + k_{13'})x_{20} + k_{14} - k_{-14}x_{20}\\
dx_{21}/dt &= (k_{13} + k_{13'})x_{20} - k_{-13}x_{21}
\end{aligned}
$$

**Fig 2**. Generation of a simulated insulin dataset based on a mathematical insulin pathway model (Sedaghat et al., 2002). Protein expression of 160 virtual animals was generated based on 21 ordinary differential equations (ODEs) plus random variance.



$$\inf(X_i \perp X_j) \le \min(\inf(X_i \perp X_k), \inf(X_j \perp X_k))$$

$$\inf(X_i \perp X_j | X_a) = -\frac{1}{2}\mathrm{E}\log\frac{\det V_{aij,aij}\det V_{a,a}}{\det V_{ai,ai}\det V_{aj,aj}}$$

$$\inf(X_i \perp X_j | X_a) = \mathrm{E}\log\frac{f_{ij|a}(X_i, X_j|X_a)}{f_{i|a}(X_i|X_a)\, f_{j|a}(X_j|X_a)}$$

$$f_X(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}}\exp\left(-\tfrac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$

**Fig 3**. Computational workflow for DWIG. Random variable $x_i$ and $x_j$ denote the expression of protein/gene *i* and *j*; V is the variance matrix; inf means the mutual information; The stability search technique together with ARACNE and GLASSO were used for structure estimation and variable selection in high dimensions (Margolin et al., 2006; Meinshausen et al., 2010; Friedman et al., 2008; Whittaker, 1990). The GGM was fitted and the CMI was estimated. Divergence weighted independence graphs were built based on CMI.

## Results

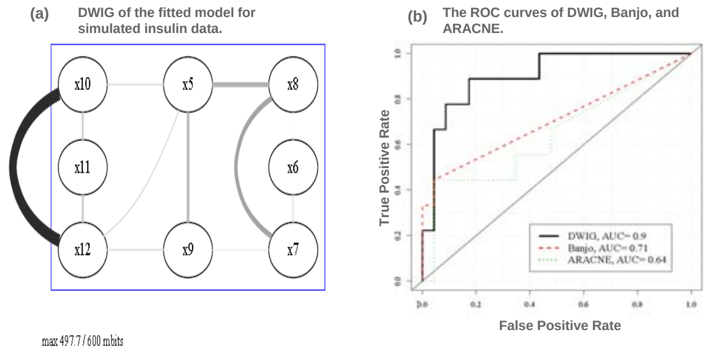


**Fig 4**. (a) The conditional DWIG of the fitted model defined by thresholding; (b) ROC curves of DWIG, Banjo (Yu et al., 2004), and ARACNE (Margolin et al., 2006). The ROC curves were built according to the adjacency matrix constructed from the simulated insulin dataset. DWIG outperformed Banjo and ARACNE in this simulated dataset.

**(a)** **Symbols and names for the 25 cytokines/chemokines analyzed in the BALF of A/J mice dataset.**

| Abbreviation used | Protein name |
|---|---|
| CD40 | CD40 |
| CD40L | CD40 ligand |
| Eotaxin | Eotaxin |
| FGF9 | Fibroblast growth factor 9 |
| GCP2 | Granulocyte chemotactic peptide-2 |
| Haptogl | Haptoglobin |
| IgA | Immunoglobulin A |
| IL10 | Interleukin-10 |
| IL1b | Interleukin-1 beta |
| Insulin | Insulin |
| IP10 | Inducible Protein-10 |
| Leptin | Leptin |
| LIF | Leukemia inhibitory factor |
| MCP1 | Monocyte chemotactic protein-1 |
| MCP3 | Monocyte chemotactic protein-3 |
| MCSF | Macrophage colony stimulating factor |
| MDC | Macrophage-derived chemoattractant |
| MIP1b | Macrophage inflammatory protein -1 beta |
| MIP1g | Macrophage inflammatory protein -1 gamma |
| MIP2 | Macrophage inflammatory protein -2 |
| MMP9 | Matrix metalloproteinase-9 |
| Myogl | Myoglobin |
| TF | Tissue factor |
| VCAM1 | Vascular cell adhesion molecule-1 |
| VEGF | Vascular endothelial growth factor |

**(b)** **DWIG of smoke exposure group for A/J mice dataset.**

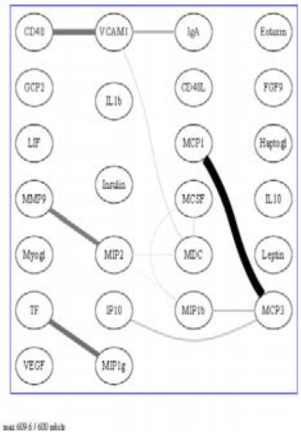


**Fig 5**. (a) Symbols and names for the 25 cytokines/chemokines analyzed in the BALF of A/J mice dataset; (b) DWIG of the fitted graphical models obtained by using the common edges from the stabilized versions of an ARACNE and a GLASSO search for smoke exposure group. The links CD40 ~ VCAM1 and MMP9 ~ MIP2 were supported by several publications (Yoon et al., 2007; Lanone et al., 2002; Gonzalo et al., 1996; Lei et al., 1998; Propst et al., 2000).

## Conclusions

**In conclusion, a novel computational method—DWIG—for exploratory analysis of biological expression data was established. DWIG outperformed other widely used methods such as Banjo and ARACNE using the simulated dataset. DWIG was applied to a MAP data set obtained from BALF samples of female A/J mice exposed to cigarette mainstream smoke for 3 and 5 months. An association network with 25 cytokines/chemokines was built. Some links identified by DWIG were supported by several publications. DWIG is a powerful method for reconstruction of a biologically meaningful network from biological expression data set.**

## Abbreviations

**BALF:** Bronchoalveolar lavage fluid
**MAP**: Multi-analyte profile
**DWIG:** Divergence weighted independence graphs
**ARACNE**: Algorithm for the reconstruction of accurate cellular networks
**Banjo**: Bayesian network with Java objects
**CMI**: Conditional mutual information
**GGM**: Graphical Gaussian model

## References

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Biostatistics, 9(3), 432-41.
Gonzalo, et al. (1996). Journal of Clinical Investigation, 98(10), 2332-2345.
Lanone, S., et al. (2002). Journal of Clinical Investigation, 110(4), 463–474.
Lei, X., et al. (1998). Journal of Clinical Investigation, 101(6), 1342-1353.
Margolin, A., et al. (2006). BMC Bioinformatics, 20;7, Suppl 1:S7.
Meinshausen, N. and Buhlmann, P. (2010). Stability selection. J. Royal Statist. Society B, 72(4), 117-132.
Propst, S., et al. (2000). The Journal of Immunology, 165(4), 2214-2221.
Sedaghat, A. R., et al. (2002). American Journal of Physiology – Endocrinology and Metabolism, 283, E1084-E1101.
Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics. Wiley, Chichester.
Yoon, H., Cho, H., and Kleeberger, S. (2007). Environmental health perspectives, 115(11), 1557-1563.
Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED. Bioinformatics. 2004 Dec 12;20(18):3594-603.

PMI RESEARCH & DEVELOPMENT

19th Annual International Conference on Intelligent Systems for Molecular Biology and 10th European Conference on Computational Biology
- AUSTRIA CENTER VIENNA, July 17 -19, 2011

**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