



The power of Boolean implication networks

Debashis Sahoo*

Institute of Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, CA, USA

Edited by:

Hans Westerhoff, University of Manchester, UK

Reviewed by:

Andrzej Michal Kierzek, University of Surrey, UK

Noriko Hiroi, Keio University, Japan
Kristina Gruden, National Institute of Biology, Slovenia

***Correspondence:**

Debashis Sahoo, Institute of Stem Cell Biology and Regenerative Medicine, Stanford University, 265 Campus Drive, Rm G3101B, Stanford, CA, USA.
e-mail: sahoo@stanford.edu

Human diseases have been investigated in the context of single genes as well as complex networks of genes. Though single gene approaches have been extremely successful in the past, most human diseases are complex and better characterized by multiple interacting genes commonly known as networks or pathways. With the advent of high-throughput technologies, a recent trend has been to apply network-based analysis to the huge amount of biological data. Analysis on Boolean implication network is one such technique that distinguishes itself based on its simplicity and robustness. Unlike traditional analyses, Boolean implication networks have the power to break into the mechanistic insights of human diseases. A Boolean implication network is a collection of simple Boolean relationships such as “if A is high then B is low.” So far, Boolean implication networks have been employed not only to discover novel markers of differentiation in both normal and cancer tissues, but also to develop robust treatment decisions for cancer patients. Therefore, analyses based on Boolean implication networks have potential to accelerate discoveries in human diseases, suggest therapeutics, and provide robust risk-adapted clinical strategies.

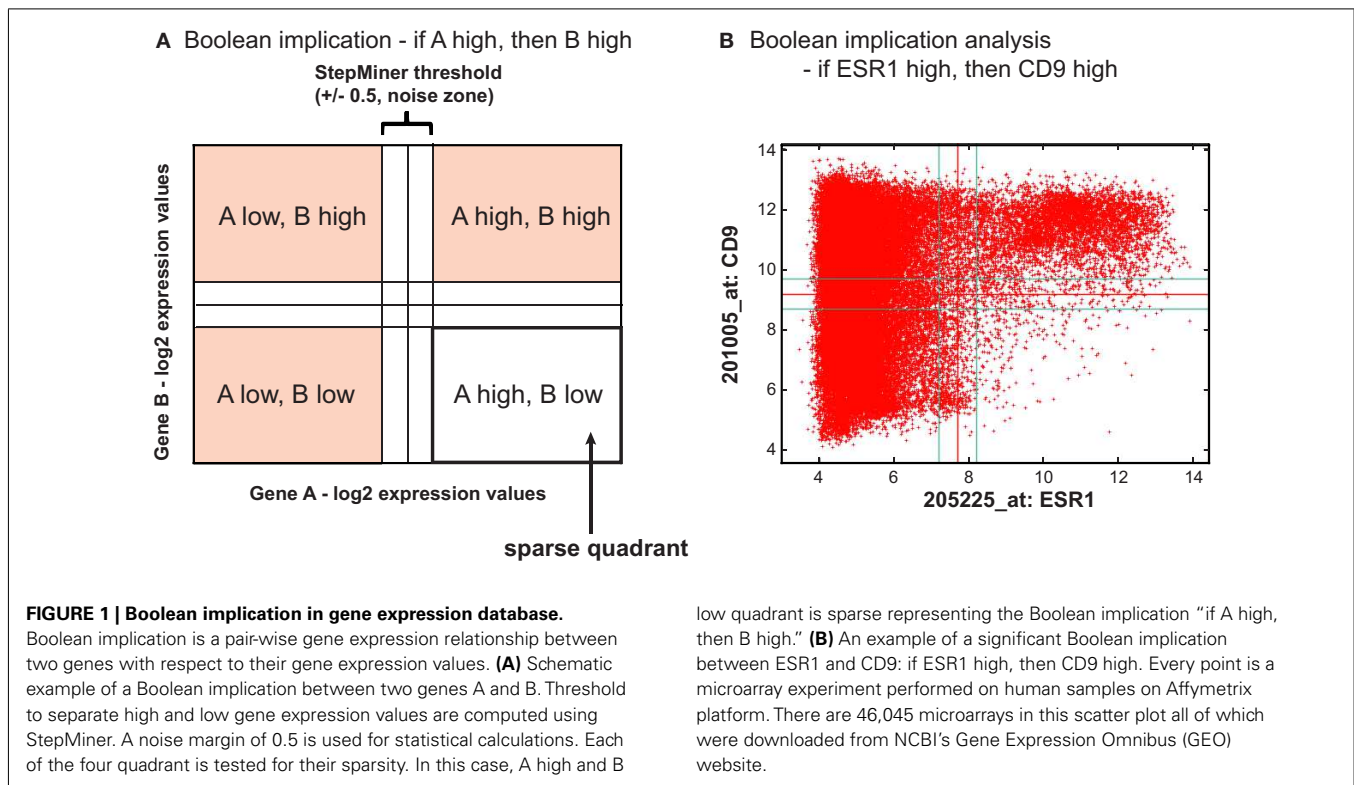
Keywords: bioinformatics, cancer, computational biology, differentiation, microarray analysis, prognostic biomarkers, stem cell, systems biology

INTRODUCTION

In the past detailed single gene investigations in the context of human diseases was extremely successful and produced many useful drugs (Miller et al., 1982; Slamon et al., 2001; Cunningham et al., 2004; Scott et al., 2012). However, the progress was extremely slow and the success was achieved at the cost of a huge number of failed investigations with multiple billions of dollars in investments (Arrowsmith, 2011; Allison, 2012). Unlike in the past years, it is now easy to gather information from tens of thousands of genes simultaneously. Modern approaches can leverage these huge amounts of biological data to understand human diseases. Therefore, a recent trend in analysis has been shifted to multiple genes that are part of a single functional unit commonly known as networks or pathways. The new approaches have been termed network analysis or systems biology. Clearly, these new approaches have the potential to tackle the complexity of human diseases (Mootha et al., 2003; Segal et al., 2003; Basso et al., 2005; Subramanian et al., 2005; Margolin et al., 2006; Bonneau et al., 2007; Lee et al., 2009; Schadt et al., 2010; Bousquet et al., 2011; Gupta et al., 2011; Jornsten et al., 2011). However, the systematic noise in the system has always challenged these approaches. The noise in the system is due to experimental or biological noise and also noise in measuring gene expression values in a microarray hybridization experiment. In addition to noise, other challenge to the network-based approaches is to translate the discoveries to the clinic.

In this mini review, we discuss a systems biology or network-based analysis using Boolean implication network (Sahoo et al., 2008). A Boolean implication network is simply a collection of Boolean implication relationships as described by Sahoo et al. (2008). Boolean typically means a logic calculus of two values,

which are high and low gene expression values in this context. A Boolean implication relationship is a simple “if-then” relationship between the high and low gene expression values between a pair of genes. For example, “if A is high, then B is high” is a Boolean implication relationship between a pair of genes A and B, where A high and B low is ruled out as a possible scenario as shown in **Figure 1**. Therefore, whenever gene expression of A is high, we observe gene expression of B is also high. In other words, A high is a subset of B high. In a two dimensional scatter plot between two genes and their thresholds for high and low values, there are four possible quadrants: “A low B low,” “A low B high,” “A high B low,” and “A high B high.” One or more sparse quadrants in this plot is mathematically represented as a Boolean implication. For example, the Boolean implication “if A high, then B high” represent a sparse “A high B low” quadrant. There are six possible Boolean implication relationships, two of them are symmetric, and other four are asymmetric. The symmetric Boolean implication relationship has two diagonally opposite sparse quadrant and the asymmetric Boolean implication relationship has only one sparse quadrant. As shown in **Figure 1**, the threshold to define “high” and “low” gene expression levels are determined using StepMiner (Sahoo et al., 2007). The expression levels of each probeset are sorted and a step function fitted (using StepMiner) to the sorted expression level that minimizes the square error between the original and the fitted values. We determined the noise margin by using very tightly correlated genes and found that there is still a difference of twofold change (in log scale a value of Miller et al., 1982) among the values that are linearly related. Therefore, we used a noise margin of 1 (threshold -0.5 to threshold $+0.5$) and discarded all the microarrays that fall within these region for Boolean implication analysis. The noise margin was an important consideration



that allowed us to identify many significant Boolean implication relationships.

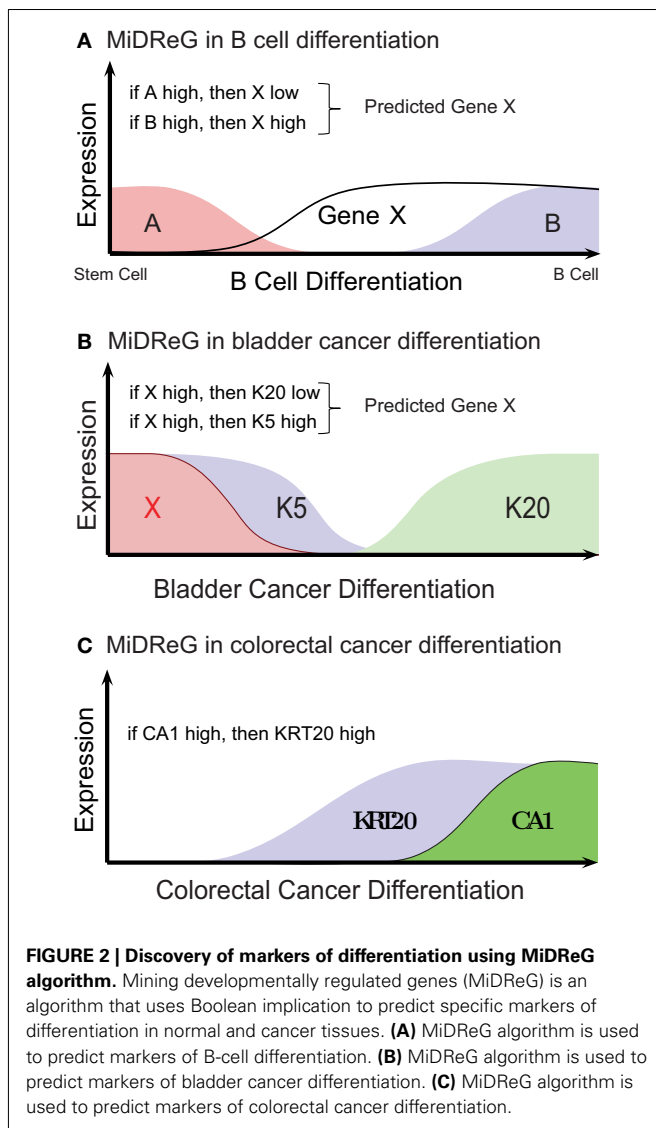
SYSTEMS BIOLOGY USING BOOLEAN IMPLICATION

It is possible to discover Boolean implication relationships in the largest possible dataset that include all publicly available microarrays from Gene Expression Omnibus (GEO) or ArrayExpress. These relationships represent natural invariants in a particular species. For example, a Boolean implication relationship in a particular dataset that contains all human samples on Affymetrix platform represents a natural invariant gene expression relationship in human. Many of these invariants are due to tissue specific gene expression. For example, a brain specific gene and a prostate specific gene can never be expressed together. Therefore, they will have a Boolean relationship of the form “if A high, then B low.” Similarly, many of these relationships can be due to developmental gene expression pattern or related to the biological process of differentiation. Mining developmentally regulated genes (MiDReG) is a simple algorithm that uses Boolean implication to identify genes expressed at different stages of differentiation (Sahoo et al., 2010). The key concept behind this algorithm is to use invariants to predict state of the gene expression pattern. We describe here how MiDReG and Boolean implication are used in B cell, bladder cancer, and colon cancer differentiation.

B-CELL DIFFERENTIATION

B cells are special types of blood cell that are created from a blood stem cell by the process of differentiation. As the stem cell undergoes the process of differentiation, many genes changes their expression pattern. There are genes that are specific to the stem

cell only and also there are genes that are specific to the differentiated B cell. MiDReG algorithm takes advantage of these gene pairs that have a significant Boolean implication “if A high, then B low,” and predict other genes that are expressed in the progenitors or precursors of B cells (Inlay et al., 2009; Sahoo et al., 2010). Let’s assume that gene A is expressed at the blood stem cells and it turns off as the stem cells differentiate to B cell. Similarly, let’s assume that gene B is off at the stem cell and it turns on as the stem cell differentiates to B cells (**Figure 2A**). Therefore, in this narrow view of differentiation gene A and gene B are mutually exclusively expressed. Let’s assume that there is a significant Boolean implication “if A high, then B low.” The significant Boolean implication represents a global invariant in all microarray datasets. In this case, if we want to identify a gene X that turns on after gene A turns off and before gene B turns on, we could simply use Boolean implication “if A high, X low,” and “if B high, X high” (**Figure 2A**). Since the Boolean implication is an invariant, we could hypothesize a state of differentiation where gene A is off, gene X is on, and gene B is off. In addition, this state of differentiation is between stem cell and the mature B cell. Therefore, gene X could potentially mark precursors of the mature B cell. We validated the gene expression patterns of the newly discovered genes using this approach by qPCR on the sorted B-cell progenitors from mouse blood and bone marrow. Review of the published literature of knockout mice revealed that many of our discovered genes were directly involved in B-cell differentiation. Out of 62 MiDReG genes, 41 genes were found to be knocked out in mice. Out of these 41 mice knockouts, 26 (63.4%) genes show defects in B-cell function and differentiation, 9 (22.0%) genes are associated with known B-cell function according to other experiments, and 6



(14.6%) genes could have a B-cell function based on their expression in the B cell and reported other hematopoietic functions. A detailed analysis on mouse lineages using MiDReG revealed a new earliest marker of B-cell differentiation Ly6D. This gene was investigated in detail by Inlay et al. (2009). Overall, our results on the B-cell differentiation suggested that MiDReG is a simple but extremely powerful approach to discover novel markers of progenitor cells.

BLADDER CANCER DIFFERENTIATION

Differentiation within cancer is a very controversial topic (Reya et al., 2001). However, in bladder cancer it is established that there are two different cell types identified by Keratin 5 and Keratin 20 (Chan et al., 2009). Keratin 5 marks immature cell types that can differentiate to Keratin 20 positive cells (Chan et al., 2009). MiDReG algorithm was used to identify an upstream marker Keratin 14 (Volkmer et al., 2012). There is a significant Boolean implication relationship between Keratin 5 and Keratin 20 “if Keratin 5 high, then Keratin 20 low” that enabled the MiDReG

algorithm to predict upstream markers. In this case, we are interested in a marker X that goes down early compared to Keratin 5. Thus, it is expressed at the most immature state of the cancer cell. The candidate markers were chosen based on Boolean implication “if X high, then Keratin 5 high” and “if X high, then Keratin 20 low” (Figure 2B). Keratin 14 was one of the markers that satisfied these two Boolean implication strongly. In addition, Keratin 14 was a single prognostic marker in both gene and protein expression datasets. The prognostic power of Keratin 14 was independent of currently established stage and grade. Therefore, a simple immunohistochemical analysis can identify high risk bladder cancer patients. Since, clinicians decide whether to perform cystectomy which is complete bladder removal based on stage and grade, it is possible to incorporate Keratin 14 based risk stratification into this important clinical decision endpoint. Clinicians are currently developing risk-adapted clinical strategies based on Keratin 14 for bladder cancer patients.

COLON CANCER DIFFERENTIATION

Many important markers in the differentiation of colon cancer cells follow Boolean implication (Dalerba et al., 2011). For example, there is a significant Boolean implication between Keratin 20 and CA1 “if CA1 high, then Keratin 20 high” (Figure 2C). This relationship is particularly strong with no exception. There are no tumors with CA1 high and KRT20 low. Even in a tumor when CA1 positive cells are present they have to go through a KRT20 positive precursor cell during differentiation. Accordingly, CA1 positive cells are a subset of Keratin 20 positive cells in both normal colon and colorectal cancer tissues. In addition, Keratin 20 negative patients have worse outcome compared to CA1 positive and Keratin 20 positive cancer patients. Other markers such as MS4A12, CD177, and SLC26A3 follow similar Boolean implication relationships.

STRENGTHS AND LIMITATIONS

In this review we show that Boolean implication can be used to identify markers of differentiation in both normal and cancer tissues. The strength of Boolean implication is its ability to identify asymmetric gene expression relationships. In contrast, most other approaches focus on using symmetric gene expression relationship to build gene expression network. We have shown that some of the gene expression patterns in differentiation can be modeled using asymmetric Boolean implication. Therefore, it would be useful for predicting important genes involved in the process of differentiation. In addition, markers of differentiation are most likely robust prognostic biomarkers in cancer patients. Using these markers, clinicians may be able to develop better risk-adapted treatment decisions for cancer patients. The limitation of Boolean implication is that it requires large number of samples. Also, it might miss many other important genes that are involved in differentiation but do not have significant Boolean implication. Accordingly, Boolean implication is a very stringent criterion. Therefore, it pulls out many important genes and appears to be less noisy compared to traditional approaches.

An important distinction between Boolean implication analyses compared to other traditional network-based analyses is that most of these other analyses are focused on identifying gene regulatory networks or signal transduction pathways. Boolean

implication has not been utilized to identify gene regulatory networks or signaling networks which contains simple feed-back and feed-forward structure. Instead, it was used to identify cell type or tissue specific gene expression patterns and they are interpreted in terms of development and differentiation. This is very different from Bayesian or mutual information based networks that primarily identify transcription factors and their targets (Segal et al., 2003; Basso et al., 2005; Margolin et al., 2006; Lee et al., 2009). Similarly, Boolean implication analyses are also different from traditional Boolean networks that are used to build a functional executable model or a circuit model (Glass and Kauffman, 1973; Shmulevich and Kauffman, 2004). There are also networks based on ODE models which describes mechanistic biochemical interactions (Ferrell et al., 2011). Both the Boolean and ODE based approaches described above models non-linear dynamical systems (Glass and Kauffman, 1973; Shmulevich and Kauffman, 2004; Ferrell et al., 2011). In contrast, Boolean implication analyses models static invariant relationships in a large biological dataset.

In summary, Boolean implication is an empirically observed relationship in the data, which may not hold for data gathered for different tissue types or under different conditions. Like correlation networks, Boolean implication networks do not capture

causality. Boolean implication captures both symmetric as well as asymmetric relationships. It provides a powerful platform for discovery of novel markers of differentiation in both normal and cancer tissues.

ACKNOWLEDGMENTS

Boolean implication and MiDReG tools were developed as part of Dr. Sahoo's Ph.D. at Stanford University with significant contribution from Prof. David Dill as Ph.D. advisor, Prof. Sylvia Plevritis as co-advisor, Prof. Rob Tibshirani and Andrew Gentles. The application of these tools were developed in collaboration with the Weissman lab and the Clarke lab at Stanford University. The author thank I. L. Weissman, M. F. Clarke, J. Lipsick, M. van de Rijn, L. D. Shortliffe, J. D. Brooks, J. Pollack, R. Levy, J. Seita, M. Inlay, D. Bhattacharya, R. K. Chin, J. Volkmer, P. Dalerba, K. S. Chan for critical discussions, helpful suggestions, and technical advice. Dr. Sahoo is supported by National Institutes of Health (NIH) Grant K99CA151673-01A1, Department of Defense Grant W81XWH-10-1-0500, Ludwig Institute Grant (PI: Irv Weissman), and a grant from the Siebel Stem Cell Institute and the Thomas and Stacey Siebel Foundation. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of the NIH and other grant agencies.

REFERENCES

- Allison, M. (2012). Reinventing clinical trials. *Nat. Biotechnol.* 30, 41–49.
- Arrowsmith, J. (2011). Trial watch: phase III and submission failures: 2007–2010. *Nat. Rev. Drug Discov.* 10, 87.
- Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* 37, 382–390.
- Bonneau, R., Facciotti, M. T., Reiss, D. J., Schmid, A. K., Pan, M., Kaur, A., Thorsson, V., Shannon, P., Johnson, M. H., Bare, J. C., Longabaugh, W., Vuthoori, M., Whitehead, K., Madar, A., Suzuki, L., Mori, T., Chang, D. E., Diruggiero, J., Johnson, C. H., Hood, L., and Baliga, N. S. (2007). A predictive model for transcriptional control of physiology in a free living cell. *Cell* 131, 1354–1365.
- Bousquet, J., Anto, J. M., Sterk, P. J., Adcock, I. M., Chung, K. F., Roca, J., Agusti, A., Brightling, C., Cambron-Thomsen, A., Cesario, A., Abdelhak, S., Antonarakis, S. E., Avignon, A., Ballabio, A., Baraldi, E., Baranov, A., Bieber, T., Bockaert, J., Brahmachari, S., Brambilla, C., Bringer, J., Dauzat, M., Ernberg, I., Fabbri, L., Froguel, P., Galas, D., Gojobori, T., Hunter, P., Jorgensen, C., Kauffmann, E., Kourilsky, P., Kowalski, M. L., Lancet, D., Pen, C. L., Mallet, J., Mayosi, B., Mercier, J., Metspalu, A., Nadeau, J. H., Ninot, G., Noble, D., Oztürk, M., Palkonen, S., Préfaut, C., Rabe, K., Renard, E., Roberts, R. G., Samolinski, B., Schünemann, H. J., Simon, H. U., Soares, M. B., Superti-Furga, G., Tegner, J., Verjovski-Almeida, S., Wellstead, P., Wolkenhauer, O., Wouters, E., Balling, R., Brookes, A. J., Charron, D., Pison, C., Chen, Z., Hood, L., and Auffray, C. (2011). Systems medicine and integrated care to combat chronic noncommunicable diseases. *Genome Med.* 3, 43.
- Chan, K. S., Espinosa, I., Chao, M., Wong, D., Ailles, L., Diehn, M., Gill, H., Presti, J. Jr., Chang, H. Y., van de Rijn, M., Shortliffe, L., and Weissman, I. L. (2009). Identification, molecular characterization, clinical prognosis, and therapeutic targeting of human bladder tumor-initiating cells. *Proc. Natl. Acad. Sci. U.S.A.* 106, 14016–14021.
- Cunningham, D., Humblet, Y., Siena, S., Khayat, D., Bleiberg, H., Santoro, A., Bets, D., Mueser, M., Harstrick, A., Verslype, C., Chau, I., and Van Cutsem, E. (2004). Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. *N. Engl. J. Med.* 35, 337–345.
- Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P. S., Rothenberg, M. E., Leyrat, A. A., Sim, S., Okamoto, J., Johnston, D. M., Qian, D., Zabala, M., Bueno, J., Neff, N. E., Wang, J., Shelton, A. A., Visser, B., Hisamori, S., Shimono, Y., van de Wetering, M., Clevers, H., Clarke, M. F., and Quake, S. R. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.* 29, 1120–1127.
- Ferrell, J. E., Tsai, T. Y., and Yang, Q. (2011). Modeling the cell cycle: why do certain circuits oscillate? *Cell* 144, 874–885.
- Glass, L., and Kauffman, S. A. (1973). The logical analysis of continuous, non-linear biochemical control networks. *J. Theor. Biol.* 39, 103–129.
- Gupta, P. B., Fillmore, C. M., Jiang, G., Shapira, S. D., Tao, K., Kuperwasser, C., and Lander, E. S. (2011). Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* 146, 633–644.
- Inlay, M. A., Bhattacharya, D., Sahoo, D., Serwold, T., Seita, J., Karsunky, H., Plevritis, S. K., Dill, D. L., and Weissman, I. L. (2009). Ly6d marks the earliest stage of B-cell specification and identifies the branchpoint between B-cell and T-cell development. *Genes Dev.* 23, 2376–2381.
- Jornsten, R., Abenius, T., Kling, T., Schmidt, L., Johansson, E., Nordling, T. E., Nordlander, B., Sander, C., Gennemark, P., Funai, K., Nilsson, B., Lindahl, L., and Nelander, S. (2011). Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Mol. Syst. Biol.* 7, 486. doi:10.1038/msb.2011.17
- Lee, S. I., Dudley, A. M., Drubin, D., Silver, P. A., Krogan, N. J., Pe'er, D., and Koller, D. (2009). Learning a prior on regulatory potential from eQTL data. *PLoS Genet.* 5, e1000358. doi:10.1371/journal.pgen.1000358
- Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl. 1), S7.
- Miller, R. A., Maloney, D. G., Warnke, R., and Levy, R. (1982). Treatment of B-cell lymphoma with monoclonal anti-idiotype antibody. *N. Engl. J. Med.* 306, 517–522.
- Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34, 267–273.
- Reya, T., Morrison, S. J., Clarke, M. F., and Weissman, I. L. (2001). Stem cells, cancer, and cancer stem cells. *Nature* 414, 105–111.
- Sahoo, D., Dill, D. L., Gentles, A. J., Tibshirani, R., and Plevritis, S. K. (2008). Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biol.* 9, R157.

- Sahoo, D., Dill, D. L., Tibshirani, R., and Plevritis, S. K. (2007). Extracting binary signals from microarray time-course data. *Nucleic Acids Res.* 35, 3705–3712.
- Sahoo, D., Seita, J., Bhattacharya, D., Inlay, M. A., Weissman, I. L., Plevritis, S. K., and Dill, D. L. (2010). MiDReG: a method of mining developmentally regulated genes using Boolean implications. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5732–5737.
- Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., and Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.* 11, 647–657.
- Scott, A. M., Wolchok, J. D., and Old, L. J. (2012). Antibody therapy of cancer. *Nat. Rev. Cancer* 12, 278–287.
- Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34, 166–176.
- Shmulevich, I., and Kauffman, S. A. (2004). Activities and sensitivities in Boolean network models. *Phys. Rev. Lett.* 93, 048701.
- Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J., and Norton, L. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* 344, 783–792.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550.
- Volkmer, J. P., Sahoo, D., Chin, R. K., Ho, P. L., Tang, C., Kurtova, A. V., Willingham, S. B., Pazhanisamy, S. K., Contreras-Trujillo, H., Storm, T. A., Lotan, Y., Beck, A. H., Chung, B. I., Alizadeh, A. A., Godoy, G., Lerner, S. P., van de Rijn, M., Shortliffe, L. D., Weissman, I. L., and Chan, K. S. (2012). Three differentiation states risk-stratify bladder cancer into distinct subtypes. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2078–2083.
- commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 March 2012; paper pending published: 04 April 2012; accepted: 27 June 2012; published online: 23 July 2012.

Citation: Sahoo D (2012) The power of Boolean implication networks. *Front. Physiol.* 3:276. doi: 10.3389/fphys.2012.00276

This article was submitted to *Frontiers in Systems Physiology*, a specialty of *Frontiers in Physiology*.

Copyright © 2012 Sahoo. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any