xSyn: A Software Tool for Identifying Sophisticated 3-Way Interactions From Cancer Expression Data

Cancer Informatics
Volume 16: 1–3
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1176935117728516

Baishali Bandyopadhyay[1], Veda Chanda[1] and Yupeng Wang[1,2,3]

[1]BDX Research & Consulting LLC, Fairfax, VA, USA. [2]Washon MedData, Inc, McLean, VA, USA.
[3]International Applied Technology Research Institute, Vienna, VA, USA.

ABSTRACT

BACKGROUND: Constructing gene co-expression networks from cancer expression data is important for investigating the genetic mechanisms underlying cancer. However, correlation coefficients or linear regression models are not able to model sophisticated relationships among gene expression profiles. Here, we address the 3-way interaction that 2 genes' expression levels are clustered in different space locations under the control of a third gene's expression levels.

RESULTS: We present xSyn, a software tool for identifying such 3-way interactions from cancer gene expression data based on an optimization procedure involving the usage of UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and synergy. The effectiveness is demonstrated by application to 2 real gene expression data sets.

CONCLUSIONS: xSyn is a useful tool for decoding the complex relationships among gene expression profiles. xSyn is available at http://www.bdxconsult.com/xSyn.html.

KEYWORDS: synergy, mutual information, optimization, 3-way interaction, gene expression, cancer, software

RECEIVED: May 17, 2017. **ACCEPTED:** August 3, 2017.

PEER REVIEW: Two peer reviewers contributed to the peer review report. Reviewers' reports totaled 353 words, excluding any confidential comments to the academic editor.

TYPE: Methodology

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Yupeng Wang, BDX Research & Consulting LLC, 3201 Lothian Rd, Fairfax, VA 22031, USA. Email: ywang@bdxconsult.com

Introduction

Constructing gene co-expression networks from cancer expression data is important for studying the genetic mechanisms underlying cancer.[1,2] Gene co-expression networks are frequently found to be different between normal and cancer samples.[3] Moreover, correlations between gene expression profiles are often dynamic, even depending on the genotypes of single-nucleotide polymorphisms.[4,5] However, correlations or linear regression models describe the general trends among gene expression profiles, thus are not able to capture sophisticated interactions among gene expression profiles.

Synergy measures the joint interaction of 2 genes toward the explanation of the occurrence (or degree of occurrence) of a specific phenotype,[6] which is equal to the difference between the mutual information[7] of the 2 genes' expression profiles with the phenotype and the sum of the mutual information of each gene profile alone with the same phenotype. It is defined formally as $I(G_1, G_2; C) - [I(G_1; C) + I(G_2; C)]$, where I represents mutual information, G_1 and G_2 are 2 interacting genes, and C is the phenotype. A positive value indicates that the combined interaction overwhelms the individual ones, suggesting an interaction between the 2 genes under the phenotype.

Conventionally, the phenotype C is binary sample status (eg, tumor or normal). In this study, we address the scenario that C coincides with the high or low expression level of a control gene x (G_x). Because binary stratification of expression levels may vary among genes, the 3 genes actually comprise a 3-way interaction. High-order genes interactions are widespread in human genomes and influence complex traits.[8] Thus, investigation of 3-way interactions, the basic form of high-order interactions, has practical significance toward fully understanding of high-order gene interactions and genotype-phenotype relationships.

Methods

Sophisticated relationships between 2 interacting genes are modeled using the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) clustering algorithms.[9] Synergy on top of UPGMA indicates how likely points in individual clusters of 2 genes' expression levels tend to show the same phenotypic classes, and in our specific case, the chance that 2 genes' expression levels were clustered in different space locations under the control of a third gene's expression levels. To overcome computational intractability of 3 gene combinations, an optimization procedure was designed. First, an optimal synergy is computed for any gene pair via a greedy algorithm which flips sample statuses to increase synergy. If the optimal synergy passes a threshold (e.g., 0.9), the new sample statuses are searched for a third gene which shows the highest degree of differential expression. We use the new sample statuses to approximate the binary stratification of the control gene's expression levels, which is acceptable considering the high levels of noises in high-throughput expression data. If the identified differential expression is statistically significant, the 3-way interaction is kept as a result. Our algorithm and software are named xSyn.

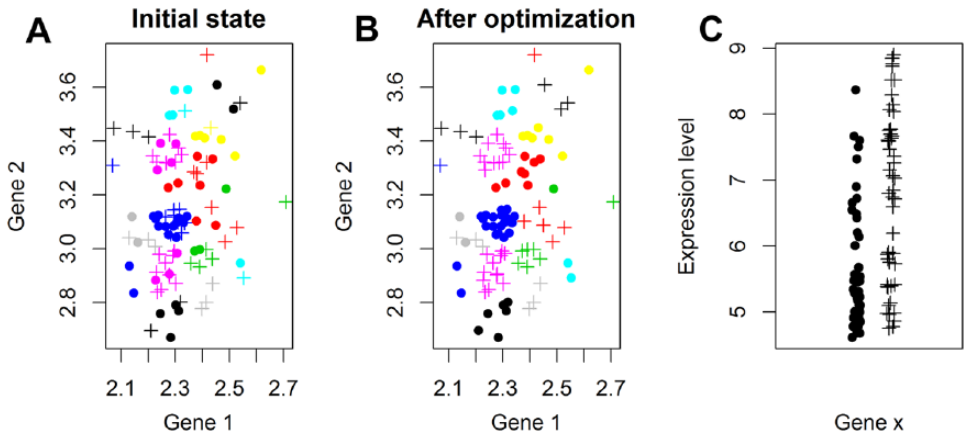

Figure 1. Visualization of a 3-way interaction identified by the xSyn software. "." represents sample status "0," whereas "+" represents sample status "1." (A) Clusters are mixed with different sample statuses before optimization. (B) Clusters tend to be filled with the same sample status after optimization, and thus, an optimal synergy is achieved. (C) The gene x shows differential expression under the new sample statuses.

The synergy between 2 gene expression profiles was computed via mutual information and conditional entropy, as previously noted.[10] The detailed algorithm is described as follows:

1. Select a number of genes showing the lowest mutual information (i.e., highest conditional entropy) with the phenotype C.
2. For each pair G_i and G_j among the selected genes:
   a) Compute the synergy under the phenotype C (sample statuses).
   b) An optimization procedure is applied. The procedure starts from the phenotype C, makes iterations to flip the status of one sample at each iteration if the synergy can be best increased, and stops when the synergy reaches convergence (i.e., cannot be increased).
   c) If the synergy reaches a threshold (e.g., 0.9), this gene pair is retained for putative 3-way analysis.
   d) Retrieve the new sample statuses generated by the optimization procedure.
   e) Loop through each gene to assess whether that gene could be G_x. Compute the *P* value of the *t* test for assessing differential expression under the new sample statuses. If the *P* value (after Bonferroni correction) is smaller than the cutoff (e.g., .05), then G_i, G_j, and G_x are output as a 3-way interaction.

The xSyn software package was written in C++ and should be run on Linux operating systems. Before compiling, c++11 and the GNU Scientific Library (GSL) need to be installed. The "readme.txt" file contains detailed instructions for executing the programs. The input gene expression data file should be tab-delimited, with the first 2 rows specifying the sample names and phenotypes (0 or 1). All subsequent rows should contain expression data for each gene/transcript/probe set. There are 2 types of output files. The first type records intermediate results, which contain gene pairs, optimal synergy, optimal sample labels, and clusters. A program for assessing the statistical significance of synergy based on permutations is provided. The second type contains the generated 3-way interactions. Script examples for multiple threading are provided. We note that the default thresholds for steps 1 and 2c were chosen heuristically based on a single workstation with 20 cores. Users may relax the thresholds to increase solution coverage.

Results

We applied the xSyn software to a prostate cancer microarray expression data set.[11] The data set contained 50 normal samples, 52 cancer samples, and 12 625 probe sets. Top 1000 probe sets showing the lowest mutual information with the phenotype were selected to assess 3-way interactions (however, all probe sets were considered for assigning gene x). The synergy cutoff was 0.9, and *P* value cutoff of *t* test (after Bonferroni correction) was .05. Twenty computer cores were used to parallelize computation, and the computation completed within 5 days. A total of 2415 3-way interactions were generated.

We randomly selected a 3-way interaction for validation. Figure 1 visualizes the effectiveness of xSyn in generating the optimal synergy for a pair of gene expression profiles and identifying the corresponding differential expression from a third gene. In the initial state (Figure 1A), 2 genes' expression profiles are clustered, but all clusters are mixed with different sample statuses. After optimization (Figure 1B), most clusters are filled with the same sample statuses, and thus, an optimal synergy is achieved. Next, a 3-way interaction is identified. The gene x shows differential expression with the new sample statuses generated by the optimization procedure (Figure 1C). We then applied xSyn to another expression data set. The data set had 80 samples, of which 40 were treated as cancer samples (E2F-null samples).[12] xSyn was executed under single-thread mode and default parameters, and 80 three-way interactions were obtained. A randomly chosen 3-way interaction was visualized and similar results to Figure 1 were obtained (Figure S1). Collectively, these validations indicate that the xSyn software is effective in identifying the addressed 3-way interactions.

Conclusions

xSyn is a software tool for identifying sophisticated 3-way interactions from cancer gene expression data based on optimization algorithms and information theory. The effectiveness of xSyn has been demonstrated by application to 2 gene expression data sets. xSyn is a useful tool for investigating the complex relationships among gene expression profiles in all types of gene expression data such as microarray, RNA-Seq, and array comparative genomic hybridization (CGH).

Acknowledgements

The authors thank the IBM Global Entrepreneur Program for offering credits for the usage of IBM SoftLayer cloud servers.

Author Contributions

YW conceived the methodology. BB and YW developed the software. BB and VC tested the software. BB and YW wrote the paper. All authors read and approved the final manuscript.

Availability and Requirements

- Project name: xSyn;
- Project home page: http://www.bdxconsult.com/xSyn.html;
- Operating system: Platform independent;
- Programming language: C++;
- Dependent libraries: c++11 and GSL;
- License: GPLv3.

Availability of Data

The microarray gene expression data used as examples for xSyn are available in GEO under accession numbers GSE68907 and GSE24594.

REFERENCES

1. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat Commun*. 2014;5:3231.
2. van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP. Gene co-expression analysis for functional classification and gene-disease predictions [published online ahead of print January 10, 2017]. *Brief Bioinform*. doi:10.1093/bib/bbw139.
3. Choi JK, Yu U, Yoo OJ, Kim S. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*. 2005;21:4348–4355.
4. Wang Y, Joseph SJ, Liu X, Kelley M, Rekaya R. SNPxGE(2): a database for human SNP-coexpression associations. *Bioinformatics*. 2012;28:403–410.
5. Kayano M, Takigawa I, Shiga M, Tsuda K, Mamitsuka H. Efficiently finding genome-wide three-way gene interactions from transcript- and genotype-data. *Bioinformatics*. 2009;25:2735–2743.
6. Anastassiou D. Computational analysis of the synergy among multiple interacting genes. *Mol Syst Biol*. 2007;3:83.
7. Cover TM, Thomas JA. *Elements of Information Theory*. New York, NY: Wiley-Interscience; 2006.
8. Taylor MB, Ehrenreich IM. Higher-order genetic interactions and their contribution to complex traits. *Trends Genet*. 2015;31:34–40.
9. Sokal R, Michener C. A statistical method for evaluating systematic relationships. *Univ Kansas Sci Bullet*. 1958;38:1409–1438.
10. Watkinson J, Wang X, Zheng T, Anastassiou D. Identification of gene interactions associated with disease from gene expression data using synergy networks. *BMC Syst Biol*. 2008;2:10.
11. Singh D, Febbo PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*. 2002;1:203–209.
12. Fujiwara K, Yuwanita I, Hollern DP, Andrechek ER. Prediction and genetic demonstration of a role for activator E2Fs in Myc-induced tumors. *Cancer Res*. 2011;71:1924–1932.