Statistical Inference of Discrete Combinatorial Functional Dependency in Biological Systems

Sajal Kumar[1]
[1] Department of Computer Science
New Mexico State University
Las Cruces, NM 88003
sajal49@nmsu.edu

Mingzhou Song[1,2]
[2] Molecular Biology Graduate Program
New Mexico State University
Las Cruces, NM 88003
joemsong@cs.nmsu.edu

Abstract

Inference of a combinatorial function from multiple independent variables (parents) to a dependent variable (child) in a discrete space can be useful in detecting nonlinear relationships in biological systems. Popular conditional independency measures, heavily used in combinatorial inference, are often insensitive to the direction of functional dependency. To address this issue, we define multivariate and conditional functional chi-squared statistics. We also present an algorithm called CFDF for bivariate discrete function inference via an exclusive-effect strategy, in order to identify a best parent set for a given child. It requires each parent to make sufficient contribution beyond any marginal effect. Simulation studies suggest a marked advantage of our framework over alternatives. Applying the method to transcriptome data in genetically perturbed biological systems, we reproduced combinatorial gene interactions known in the literature. Most importantly, we identified combinatorial patterns from joint RNA and protein data to rebut a dispute on the founding principle of molecular biology.

1 Introduction

A combinatorial effect refers to the control of dependent (child) variable by multiple independent (parent) variables. Combinatorial control occurs widely in gene regulation [1, 2]; it drives exponential biodiversity with only linear increase in genetic materials [3]. A fruit fly survives only if it is either male with an active *Sxl* gene or female with an inactive *Sxl* gene; in this example, survival is an exclusive-OR (XOR) function of sex and *Sxl* activity [4].

Many methods have been proposed to infer discrete combinatorial relationships, including switch-like models [5], Boolean networks [6, 7], discrete Bayesian networks [8] and their dynamic version [9], graphical models [10], Petri nets [11], and generalized logical networks [12]. However, these approaches often select one factor at a time [13, 14] and resort to linear [15] or sigmoidal [16] combinatorial functions, to gain computational efficiency at the cost of combinatorial interactions such as exclusive-OR. On the other hand, the more flexible conditional mutual information [17] can tell the difference between $X \not\perp Y|Z$ and $X \perp Y|Z$, and is used in inferring combinatorial transcription regulation [18]; however, it cannot infer $X \to Y|Z$ versus $X \perp Y|Z$, due to its symmetry between X and Y.

Based on the top performing univariate functional chi-squared (FUNCHISQ) method [19] in causal biological network inference [20], we define multivariate FUNCHISQ for joint effects and conditional FUNCHISQ for combinatorial effects. The framework is based on the exclusive-effect premises: 1) the combinatorial effect is captured by conditional statistics exclusive of the marginal effect and 2) when no marginal effect in a given data set can fully explain a child variable, this data set may afford a chance to capture a combinatorial effect. We give a branch-and-bound algorithm called CFDF to

14th Meeting on Machine Learning in Computational Biology (MLCB 2019), Vancouver, Canada.

identify a best bivariate discrete function and demonstrate its effectiveness on both simulated and biological data sets.

2 Methods

Definition 1 (Univariate FUNCHISQ statistic)**.** *Given n samples of discrete random variables X and Y of q and Q levels, respectively, the univariate* FUNCHISQ *is [19, 21]*

$$\chi^2_f(X \to Y) = \left[\sum_{i=1}^{q}\sum_{j=1}^{Q} \frac{[n_{ij} - (n_{i\cdot}/Q)]^2}{n_{i\cdot}/Q}\right] - \left[\sum_{j=1}^{Q} \frac{[n_{\cdot j} - (n/Q)]^2}{n/Q}\right] \tag{1}$$

where n_{ij} is the count of $X = i$ and $Y = j$ in the $q \times Q$ contingency table whose rows represent X and columns Y, $n_{i\cdot}$ is the sum of row i, and $n_{\cdot j}$ is the sum of column j.

It was previously established that $\chi^2_f(X \to Y)$ asymptotically follows a chi-square distribution of $(q-1)(Q-1)$ degrees of freedom under the null hypothesis that Y is independent of X and the assumption that Y is uniformly distributed [19]. A related exact functional test based on the same test statistic is also established [22]. FUNCHISQ is unique in *asymmetric functional optimality* and is maximized if and only if Y is a non-constant function of X given Y's marginal [23]. The statistic is minimized to zero when X and Y are empirically independent. FUNCHISQ previously outperformed mainstream techniques for causal biological network inference [20].

Definition 2 (Multivariate FUNCHISQ statistic)**.** *Let $X = [X_1, \ldots, X_p]^\top$ be a p-dimension discrete random vector with $q_1, \ldots, q_p$ levels, respectively. Let Y be a discrete random variable of Q levels. We construct a contingency table of $q = q_1 \times \cdots \times q_p$ rows and Q columns. For an observed vector of $x = [x_1, \ldots, x_p]^\top$, we linearize it to calculate a scalar row index i by*

$$i = \left[\sum_{k=1}^{p-1}\left(x_k \prod_{l=k+1}^{p} q_l\right)\right] + x_p \tag{2}$$

Plugging q and i into Eq (1), we define the multivariate FUNCHISQ *statistic $\chi^2_f(X_1, \ldots, X_p \to Y)$, which is the joint effect of X on Y.*

The multivariate FUNCHISQ is also asymptotically chi-squared distributed under the null hypothesis that Y is independent of X and the assumption that Y is uniformly distributed, following a similar argument in [19]. The statistic is subject to spurious combinatorial relationship, as it may pick a parent set involving only one strong parent. To overcome this issue, we introduce conditional functional chi-squared statistic to determine sensible combinations of parents in a multivariate functional relationship.

Definition 3 (Conditional FUNCHISQ statistic)**.** *Let $X = [X_r, X_s]^\top$. We define the conditional functional chi-squared statistic from random variable X_s to Y given X_r by*

$$\underbrace{\chi^2_f(X_s \to Y | X_r)}_{\textit{combinatorial effect}} = \underbrace{\chi^2_f(X \to Y)}_{\textit{joint effect}} - \underbrace{\chi^2_f(X_r \to Y)}_{\textit{marginal effect}} \tag{3}$$

It can be shown that $\chi^2_f(X_s \to Y | X_r)$ follows an asymptotic chi-squared distribution with $q_r(q_s - 1)(Q-1)$ degrees of freedom under the null hypothesis that Y is independent of X_r and X_s and the assumption that Y is uniformly distributed. Conditional FUNCHISQ promotes combinatorial patterns with a large difference between joint and marginal effects. It promotes parents that perform poorly by themselves but outperform best individual parents when combined. For example, in an XOR relationship, looking at X_1 with respect to $Y = X_1 \oplus X_2$ results in a poor individual (marginal) effect; however, looking at X_1, X_2 together maximizes the joint effect.

The combinatorial functional dependency by FUNCHISQ (CFDF) algorithm (Algorithm 1). It selects a best bivariate combinatorial relationship $X_i, X_j \to Y$ among pairs in X of v variables. It uses a branch-and-bound strategy similar to the A^* search. It maintains $\beta \in [0, 1]$, the lowest combinatorial (conditional FUNCHISQ) p-value so far and $\chi^2_{\max}$, the maximum joint FUNCHISQ

Algorithm 1 CFDF: Combinatorial Functional Dependency by FUNCHISQ

Input: Child Y and parent candidates $\mathbf{X} = [X_1, X_2, \ldots, X_v]$
Output: A best pair of parents $P_1, P_2 \in \mathbf{X}$

1. Calculate univariate FUNCHISQ statistics (individual effects) of each X_i in $\mathbf{X}$
2. Sort $\mathbf{X}$ to $\mathbf{X}'$ by individual effect in non-descending order
3. Initialize $\beta \leftarrow 1$ for minimum conditional FUNCHISQ p-value so far; $\chi^2_{\max} \leftarrow 0$ for maximum joint FUNCHISQ achieved so far
4. **for** each $X^{(i)}$ in $\mathbf{X}'$:
5. **for** each $X^{(j)}$ in $\mathbf{X}'$ and $j > i$:
6. Calculate optimistic conditional FUNCHISQ for both $X^{(i)}, X^{(j)}$
7. **if** optimistic conditional FUNCHISQ p-values for both $X^{(i)}, X^{(j)} \leq \beta$
8. $p[i], p[j] \leftarrow$ actual conditional FUNCHISQ p-values of $\chi^2_f(X^{(i)} \rightarrow Y | X^{(j)})$ and $\chi^2_f(X^{(j)} \rightarrow Y | X^{(i)})$
9. $\beta \leftarrow \min(\beta, \max(p[i], p[j]))$
10. **if** $\chi^2_{\max} < \chi^2_f(X^{(i)}, X^{(j)} \rightarrow Y)$ # *collective effect* of $X^{(i)}, X^{(j)}$
11. $\chi^2_{\max} \leftarrow \chi^2_f(X^{(i)}, X^{(j)} \rightarrow Y)$
12. $P_1 \leftarrow X^{(i)}$ and $P_2 \leftarrow X^{(j)}$
13. **else if** optimistic conditional FUNCHISQ of $X^{(j)} > \beta$
14. Remove $[X^{(j)} X^{(j+1)}, \cdots, X^{(v)}]$ from $\mathbf{X}'$
15. Terminate the inner for-loop
16. **else if** optimistic conditional FUNCHISQ of $X^{(i)} > \beta$
17. Terminate the outer for-loop
18. **Return** P_1, P_2

statistic so far. The optimistic conditional FUNCHISQ is the p-value associated with the theoretical maximum of the statistic, given the individual effect. The returned parents of Y have a large joint effect constrained by sufficient conditional effects. The algorithm has a worst-case runtime of $O(v^2)$. Speedup occurs when a variable fails to secure an optimistic contribution better than or equal to β, in which case we terminate the outer loop if the variable was $X^{(i)}$, since any other $X^{(k)}$, $k > i$ in the ordered list cannot do better; and in case of $X^{(j)}$, we remove all $X^{(k)}$, $k \geq j$ from the ordered list followed by inner loop termination.

3 Results

Simulation studies. To evaluate the performance of CFDF, we randomly generated 10 independent variables and constructed 20 bivariate functions from polynomial, trigonometric, logarithmic and exponential function families. All 10 parents and 20 children were discretized using optimal univariate clustering from the 'Ckmeans.1d.dp' package [24, 25], with the number of clusters determined using the 'mclust' package [26]. We generated 1500 samples with noise added by the R function `jitter`.

We compared CFDF to four other discrete conditional independency tests from the 'bnlearn' package [27] including conditional mutual information, conditional Pearson's chi-square, semi-parametric conditional mutual information and semi-parametric conditional Pearson's chi-square. Each method evaluated all 30 variables as a potential child variable to return a ranked list of the remaining 29 variables, ideally ranking the true parents (when present) higher than others, distinguishing a functional from an independent relationship and also predicting the parent→child direction. Figure 1a,b summarize AUPR and AUROC over multiple runs of the same setup with increasing noise. CFDF performs similarly in AUROC but outperforms other strategies in AUPR, suggesting that CFDF better ranks true directional interactions, as CFDF was the only directional measure out of all competitors.

Pattern discovery on leukemia gene expression data. We applied CFDF on two real data sets for both validation and discovery of combinatorial interactions. All molecular quantities were log transformed and discretized using 'Ckmeans.1d.dp' [24], with quantization levels determined by 'mclust' package [26]. We used the immortalized K562 leukemia cancer cell-line RNA-seq data collected after the knockdown of ~250 RNA-binding proteins (RBPs) from the ENCODE project [28]. Each sample measured the abundance of ~195K transcripts in transcripts per million. For each RBP, we searched for its best combinatorial parents amongst the remaining RBPs. Figure 1c shows a known interaction that we were able to validate from literature [29–31]. More encouragingly, Figure 1d shows a putative combinatorial interaction close to an exclusive-OR function.

Rebutting a dispute on the central dogma of molecular biology. The central dogma of molecular biology states that "DNA makes mRNA and mRNA makes protein" [32]. While a good correlation has been observed between DNA copy number and mRNA expression of the same gene [33–35], recent joint mRNA and protein studies [33–35], reported low correlation between the mRNA and protein levels of a gene, challenging the central dogma. We offer alternative evidence to show that the central dogma is not violated and the observed weak mRNA-protein correlation can be explained by

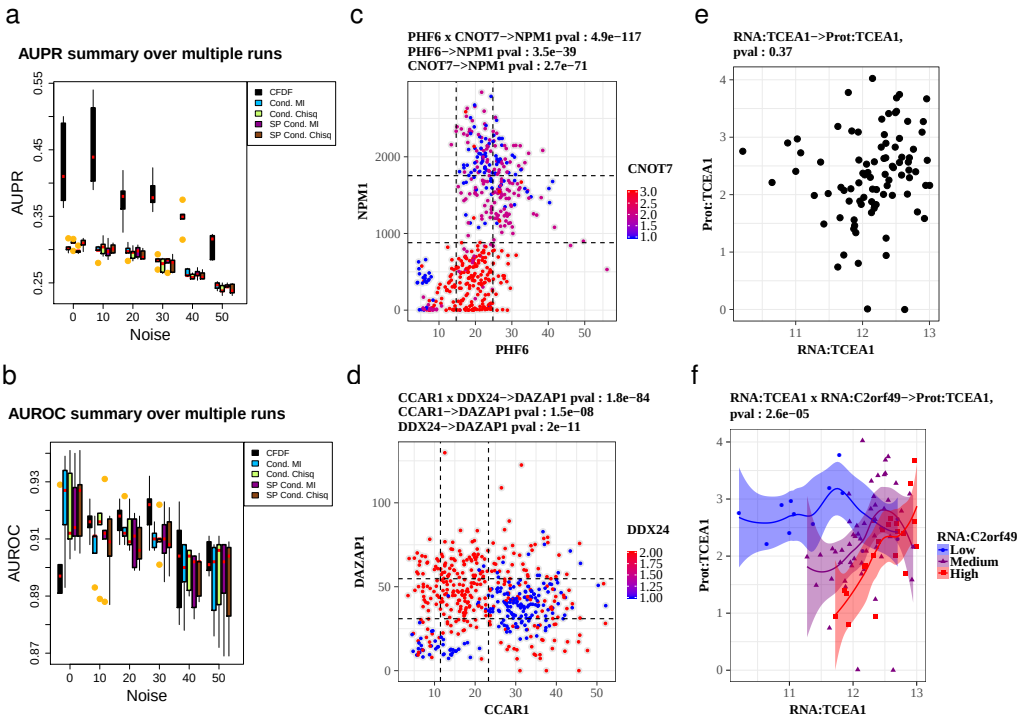

Figure 1: Result summary. (a) AUPR and (b) AUROC over multiple runs with increasing jitter noise. (c) *PHF6* × *CNOT7* → *NPM1* and (d) *CCAR1* × *DDX24* → *DAZAP1*, where the X-axis is the first parent, Y-axis the child, and color represents the second parent. The dotted lines represent the quantization levels. (e) and (f) *C2orf49*, *TCEA1*'s mRNA and protein form a strong combinatorial relationship (P=2.6×10^{-5}), rebutting central dogma violation.

combinatorial gene regulation. The Clinical Proteomics Tumor Analysis Consortium [35] provided 3718 proteins of 30 matched normal and 90 colorectal tumor samples from The Cancer Genome Atlas (TCGA) project. About 57000 mRNA transcripts for matched samples were extracted from TCGA. For each of the 3718 proteins, univariate FUNCHISQ between a protein and its mRNA was computed. If the mRNA was found to be a bad predictor for the protein (p-value>0.05), this protein was taken to the CFDF algorithm with one parent fixed to be the mRNA of the protein. CFDF then searched for a possible second mRNA/protein for a combinatorial effect.

We found a strong combinatorial pattern in translation of gene *TCEA1* mRNA to its protein involving a second gene *C2orf49* (Figure 1). The mRNA and protein of *TCEA1* are only weakly correlated (Figure 1e). However, a strong pattern emerges if we consider *C2orf49*'s mRNA in three levels: at each level, a curve exhibits different translation rates and basal levels for *TCEA1*. Thus, given the mRNA level of *C2orf49*, the mRNA level of *TCEA1* indeed predicts the protein level of TCEA1. *TCEA1* and *C2orf49* are related as belonging to the gene expression SuperPath from PathCards [36]. As translation may involve multiple factors, our finding supports that combinatorial effects reinforce rather than violate the central dogma of molecular biology.

4 Conclusions

The FUNCHISQ based combinatorial functional inference can capture both nonlinear and nonmonotonic functional relationships. CFDF selects sensible combinations of variables by requiring sufficient contributions from each parent variable to the child variable. Our simulation studies demonstrate marked improvement of CFDF over alternative methods on a variety of noisy multivariate functions. We uncovered combinatorial patterns to rebut a dispute on a basic principle of molecular biology. As the method does not assume a predefined functional model, it is applicable to combinatorial pattern discovery in many under-studied biological systems.

Acknowledgements. The reported work is supported by US National Science Foundation grant 1661331 and USDA grant 2016-51181-25408.

References

[1] Buchler NE, Gerland U, Hwa T. On schemes of combinatorial transcription logic. PNAS. 2003;100(9):5136–5141.

[2] Weingarten-Gabbay S, Segal E. The grammar of transcriptional regulation. Human Genetics. 2014;133(6):701–711.

[3] Levine M, Tjian R. Transcription regulation and animal diversity. Nature. 2003;424(6945):147–151.

[4] Page D, Ray S. Skewing: An efficient alternative to lookahead for decision tree induction. In: IJCAI; 2003. p. 601–612.

[5] Ghaffarizadeh A, Flann NS, Podgorski GJ. Multistable switches and their role in cellular differentiation networks. BMC Bioinformatics. 2014;15(Suppl 7):S7.

[6] Akutsu T, Miyano S, Kuhara S, et al. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: Pacific Symposium on Biocomputing. vol. 4; 1999. p. 17–28.

[7] Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics. 2002;18(2):261–274.

[8] Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning. 1995;20(3):197–243.

[9] Murphy KP. Dynamic Bayesian Networks: Representation, Inference and Learning [PhD Thesis]. University of California. Berkeley, CA; 2002.

[10] Madigan D, York J, Allard D. Bayesian graphical models for discrete data. International Statistical Review/Revue Internationale de Statistique. 1995;p. 215–232.

[11] Peterson JL. Petri Net Theory and the Modeling of Systems. Prentice Hall PTR; 1981.

[12] Song M, Lewis CK, Lance ER, Chesler EJ, Yordanova RK, Langston MA, et al. Reconstructing generalized logical networks of transcriptional regulation in mouse brain from temporal gene expression data. EURASIP Journal on Bioinformatics and Systems Biology. 2009;2009:545176.

[13] Storey JD, Akey JM, Kruglyak L. Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biology. 2005 Aug;3(8):e267.

[14] Jiang X, Neapolitan RE, Barmada MM, Visweswaran S, Cooper GF. A fast algorithm for learning epistatic genomic relationships. In: Proceedings of AMIA Annual Symposium; 2010. p. 341–345.

[15] Knijnenburg TA, Wessels LF, Reinders MJ. Combinatorial influence of environmental parameters on transcription factor activity. Bioinformatics. 2008;24(13):i172–i181.

[16] Titsias MK, Honkela A, Lawrence ND, Rattray M. Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison. BMC Systems Biology. 2012;6(1):53.

[17] Cover TM, Thomas JA. Elements of Information Theory. John Wiley & Sons; 2012.

[18] Wang K, Saito M, Bisikirska BC, Alvarez MJ, Lim WK, Rajbhandari P, et al. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nature Biotechnology. 2009;27(9):829–837.

[19] Zhang Y, Song M. Deciphering interactions in causal networks without parametric assumptions. arXiv Molecular Networks. 2013 Nov;p. 1311.2707. Available from: `http://arxiv.org/abs/1311.2707`.

[20] Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, Carlin DE, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. Nat Methods. 2016 Apr;13(4):310–318.

[21] Zhang Y, Zhong H, Nguyen H, Sharma R, Kumar S, Song J. FunChisq: Model-Free Functional Chi-Squared and Exact Tests; 2019. R package version 2.4.9.1. https://CRAN.R-project.org/package=FunChisq.

[22] Zhong H, Song M. A Fast Exact Functional Test for Directional Association and Cancer Biology Applications. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019 May;16(3):818–826.

[23] Nguyen HH. Inference of Functional Dependency via Asymmetric, Optimal, and Model-free Statistics [PhD Thesis]. Department of Computer Science, New Mexico State University. Las Cruces, USA; 2018.

[24] Wang H, Song M. Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming. The R Journal. 2011;3(2):29–33.

[25] Song J, Zhong H, Wang H. Ckmeans.1d.dp: Optimal, Fast, and Reproducible Univariate Clustering; 2019. R package version 4.3.0. https://cran.r-project.org/package=Ckmeans.1d.dp.

[26] Fraley C, Raftery A, Scrucca L. Normal mixture modeling for model-based clustering, classification, and density estimation. Department of Statistics, University of Washington. 2012;23:2012.

[27] Scutari M. Learning Bayesian Networks with the bnlearn R Package. J Stat Softw. 2010;35(3):1–22.

[28] Consortium EP, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.

[29] Todd MA, Picketts DJ. PHF6 interacts with the nucleosome remodeling and deacetylation (NuRD) complex. Journal of Proteome Research. 2012;11(8):4326–4337.

[30] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Research. 2006;34(suppl_1):D535–D539.

[31] Navarro SL, White E, Kantor ED, Zhang Y, Rho J, Song X, et al. Randomized trial of glucosamine and chondroitin supplementation on inflammation and oxidative stress biomarkers and plasma proteomics profiles in healthy humans. PLoS One. 2015;10(2):e0117534.

[32] Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561.

[33] Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. Cell. 2016;166(3):755–765.

[34] Mertins P, Mani D, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016;534(7605):55.

[35] Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014;513(7518):382–387.

[36] Belinky F, Nativ N, Stelzer G, Zimmerman S, Iny Stein T, Safran M, et al. PathCards: multi-source consolidation of human biological pathways. Database (Oxford). 2015;2015.