Published in final edited form as:

Genet Epidemiol. 2013 April ; 37(3): 283–285. doi:10.1002/gepi.21718.

ViSEN: methodology and software for visualization of statistical epistasis networks

Ting Hu^{1,2,*}, Yuanzhu Chen^{3,*}, Jeff W. Kiralis¹, and Jason H. Moore^{1,2}

¹Institute for Quantitative Biomedical Sciences, Dartmouth College, USA

²Department of Genetics, Geisel School of Medicine, Dartmouth College, USA

³Department of Computer Science, Memorial University, Canada

Abstract

The non-linear interaction effect among multiple genetic factors, i.e. *epistasis*, has been recognized as a key component in understanding the underlying genetic basis of complex human diseases and phenotypic traits. Due to the statistical and computational complexity, most epistasis studies are limited to interactions with an order of two. We developed ViSEN to analyze and visualize epistatic interactions of both two-way and three-way. ViSEN not only identifies strong interactions among pairs or trios of genetic attributes, but also provides a global interaction map that shows neighborhood and clustering structures. This visualized information could be very helpful to infer the underlying genetic architecture of complex diseases and to generate plausible hypotheses for further biological validations. ViSEN is implemented in Java and freely available at <https://sourceforge.net/projects/visen/>.

Keywords

epistasis; gene-gene interaction; high-order interaction; networks; visualization; software; genome-wide association; complex diseases

Introduction

Genome-wide association studies (GWAS) have the great potential identifying single nucleotide polymorphisms (SNPs) associated with complex human diseases and phenotypic traits [Hardy and Singleton, 2009; Hirschhorn and Daly, 2005]. However, current main-effect-centered strategies are only able to find very limited single-locus effects. The non-linear interaction among multiple genetic factors, i.e. *epistasis*, has been recognized as essential in explaining the complex relationship between genetic and phenotypic variations [Cordell, 2009; Moore, 2005]. Powerful computational techniques have been developed to detect and characterize epistasis, but due to the statistical and computational complexity, most analyses are constrained to only pairwise interactions.

Here we present a network-based analysis and visualization tool for epistasis studies. ViSEN reads main effect, two-way interaction, and three-way interaction lists, and is able to generate a single graph visualization of all three orders of effects. In addition, given small-scale candidate-gene population SNP data of a disease or a phenotypic trait, ViSEN can quantify the main effects of all attributes and the pairwise and three-way interactions of all

Corresponding author: Jason H. Moore, jason.h.moore@dartmouth.edu, 1 Medical Center Dr., Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756, Phone: 603-653-9939.

*These authors contributed equally

attribute combinations onsite using information-theoretic quantities. Then the strongest pairwise and three-way epistatic interactions are represented and organized in one network. Networks are very suitable to represent interactions and are able to show not only the neighborhood structure of each attribute but also the global clustering structures of groups of interacting attributes [Newman, 2010]. ViSEN can Visualize such Statistical Epistasis Networks in order to provide a global interaction map showing all three orders of effects and to help identify high susceptibility interacting SNPs.

Methods

Information-theoretic quantities [Chanda et al., 2009; Cover and Thomas, 2006; Jakulin and Bratko, 2003] are well-suited measures for epistasis by considering both genetic attributes, i.e. SNPs, and the phenotypic class as discrete random variables. Specifically, for two SNPs G_1 and G_2 , *mutual information* $I(G_1; C)$ and $I(G_2; C)$, where C is the phenotypic class, quantify the shared information, or dependency, between individual genotypes and the phenotype, i.e. the main effects of G_1 and G_2 . In addition, by joining G_1 and G_2 together, $I(G_1, G_2; C)$ measures how much of the phenotypic class that combining G_1 and G_2 together can explain. The pairwise epistatic interaction between G_1 and G_2 can then be defined using *information gain* $IG(G_1, G_2; C) = I(G_1, G_2; C) - I(G_1; C) - I(G_2; C)$. As such, $IG(G_1, G_2; C)$ is the gained mutual information about C from considering G_1 and G_2 together. A positive $IG(G_1, G_2; C)$ indicates the synergy between G_1 and G_2 on the phenotype C , while a negative information gain indicates the redundancy between them. Such a measure has been found useful as an efficient non-parametric quantification of pairwise epistasis in genetics studies [Fan et al., 2011; Moore et al., 2006].

Furthermore, the information-gain measure was extended to quantify the synergistic effect among three genetic attributes that contribute to disease susceptibility [Hu et al., 2013]. Given three SNPs G_1 , G_2 , and G_3 the mutual information between combining three genotypes together and the phenotypic class C , i.e. $I(G_1, G_2, G_3; C)$, is first obtained. Then the three-way information gain $IG(G_1, G_2, G_3; C)$ is defined by subtracting all three main effects and three pairwise synergies from $I(G_1, G_2, G_3; C)$. Note that when apply to genetic association studies, we usually normalize the pairwise and three-way information gain measures by dividing them by the entropy of the phenotype or the disease status, i.e. $H(C)$. Therefore, the information gain measures describe the percentage that a two-way or a three-way epistatic interaction explains the phenotype (disease). Our proposed $IG(G_1, G_2, G_3; C)$ is a very strict three-way epistasis measure and is designed to detect *pure* three-way epistatic interactions, i.e. excluding all lower-order effects. By applying to a Tuberculosis (TB) data from a West African population, such an approach was able to find a statistically significant three-way epistasis among three important innate immune genes, which could be very helpful for future TB risk prediction and prevention studies [Hu et al., 2013].

Measuring pairwise and three-way epistasis helps identify important pairs and trios of susceptibility SNPs. Furthermore, an emerging systematic approach to representing and analyzing the interacting genetic factors is to use networks. Statistical epistasis networks (SEN) [Hu et al., 2011] are built by including strong interacting pairs and thus provide a global interaction map that shows not only the neighborhood structure of each attribute but also the topology of attributes clustered together. When applied to a population-based bladder cancer dataset, such a network approach was able to characterize a large connected structure of SNPs associated with bladder cancer that infers the complex genetic architecture of the disease.

We extend our previous work and present the methodology and visualization software ViSEN that is able to show both pairwise and three-way statistical epistatic interactions, in

addition to individual main effects, in a single graph. The user uploads three files listing the significant main effects, pairwise, and three-way epistatic interactions. Then ViSEN is able to show a single graph that visualizes all three orders of effects, where nodes are SNPs, edges represent pairwise epistatic interactions, and triangle-shaped hyper-edges represent three-way epistatic interactions. For smaller-scale pre-selected population SNP data, ViSEN can also calculate mutual information and information gain measures onsite, in addition to visualization. By organizing and representing all three orders of effects, ViSEN provides a global genetic interaction map of a disease or a phenotypic trait.

Implementation

In order to reach the broadest audience, we chose Java as the programming language for ViSEN using a well-received graph computation library called JUNG (Java Universal Network/Graph Framework). ViSEN has a graphical user interface (GUI) to layout the epistasis networks in a two-dimensional space using a force-based model (Fig. 1). The circular nodes are SNPs, solid-line edges represent pairwise interactions, and triangles represent three-way interactions. We use the area of the geometric shapes and width of the edges to indicate their strength. ViSEN has a set of controls to read user data and to save the graph layout. The three lists of one to three orders of effects follow the standard tab-delimited network file format with each line of texts consisting of an attribute name (two and three attribute names for two-way and three-way files, respectively) followed by the effect strength. The format of the user population SNP data is also a tab-delimited plaintext. The first line contains a header row of labels assigned to each column of the data, and each following line contains a data row. The last column of the file is the class.

After the initial layout is displayed, the user can reposition the nodes and triangles for fine-tuning. ViSEN also provides the user with a set of controls to turn on and off the labels for the strength of epistatic effects as needed. In addition, the user can control the number of pairwise and three-way interactions being visualized in the GUI. While edges and triangles are inserted to or removed from the layout, ViSEN animates these changes for the user to observe how the network evolves. When a satisfactory layout is achieved, ViSEN can export the visualization to a PNG file.

Discussion

ViSEN shows both pairwise and three-way epistasis, in addition to main effects, in one network. To the best of our knowledge, it is the first visualization software that shows three orders of effects simultaneously. Such an idea embraces the complexity of genetic architecture underlying complex diseases and phenotypic traits, and can serve as a very useful map to identify groups of risk-associated SNPs and to depict their unique interacting patterns.

In the future development, we plan to integrate the computation of network and hyper-graph statistics in ViSEN, such that the user can be provided with network property analysis in addition to visualization. Moreover, current statistical epistasis quantifications in ViSEN only consider discrete traits, and it is important to extend the quantification measures to continuous traits. Last, it will be very interesting to enable the biology annotation feature in ViSEN, such that we can interpret the interaction topology using the functional information of SNPs, including biological pathways and gene ontology categories. This analysis will be very useful for translating statistical and computational findings to biological mechanisms, and can help us gain a better understanding of complex human diseases and phenotypic traits.

Acknowledgments

This work was supported by the National Institute of Health R01-LM009012, R01-LM010098, R01-AI59694, R01-EY022300 to JHM, and the Natural Sciences and Engineering Research Council of Canada Discovery Grant 327667-2010 to YC.

References

- Chanda P, Sucheston L, Zhang A, Ramanathan M. The interaction index, a novel information-theoretic metric for prioritizing interacting genetic variations and environmental factors. *European Journal of Human Genetics*. 2009; 17:1274–1286. [PubMed: 19293841]
- Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nature Review Genetics*. 2009; 10:392–404.
- Cover, TM.; Thomas, JA. *Elements of Information Theory*. 2. Wiley; 2009.
- Fan R, Zhong M, Wang S, Zhang Y, Andrew A, Karagas M, Chen H, Amos CI, Xiong M, Moore JH. Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment interactions/correlations of complex diseases. *Genetic Epidemiology*. 2011; 35:706–721. [PubMed: 22009792]
- Hardy J, Singleton A. Genome-wide association studies and human disease. *New England Journal of Medicine*. 2009; 360:1759–1768. [PubMed: 19369657]
- Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nature Review Genetics*. 2009; 6:95–108.
- Hu T, Chen Y, Kiralis JW, Collins RL, Wejse C, Sirugo G, Williams SM, Moore JH. An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *Journal of the American Medical Informatics Association*. 2013 in press.
- Hu T, Sinnott-Armstrong NA, Kiralis JW, Andrew AS, Karagas MR, Moore JH. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics*. 2011; 12:364. [PubMed: 21910885]
- Jakulin A, Bratko I. Analyzing attribute dependencies. *Lecture Notes in Artificial Intelligence*. 2003; 2838:229–240.
- Moore JH. A global view of epistasis. *Nature Genetics*. 2005; 37:13–14. [PubMed: 15624016]
- Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*. 2006; 241:252–261. [PubMed: 16457852]
- Newman, MEJ. *Networks: An Introduction*. Oxford University Press; 2010.

Fig. 1. Graphical visualization of two-way (solid edges) and three-way (triangles) epistatic interactions using ViSEN. Nodes are genetic attributes. Labels in red show the strengths of main and interaction effects.