Entropy-Based Epistasy Search in SNP Case-Control Studies

Amir Manzour
Department of Electrical and Computer Engineering
Isfahan University of Technology
a_manzour@ec.iut.ac.ir

Mohammad Saraee
Department of Electrical and Computer Engineering
Isfahan University of Technology
saraee@cc.iut.ac.ir

Abstract

The purpose of gene mapping is to identify the causal genetic regions of a specific phenotype mainly a complex disease. Most complex diseases are believed to have multiple contributing loci often having subtle patterns which make them fairly difficult to find in large datasets. We present and discuss a new criterion called conditional mutual information for association mapping and compare it to the previous criterion which is mutual information from different aspects. Furthermore, algorithms are proposed to find relevance chains. The proposed algorithms are especially in favor of diseases having almost equally contributing regions known as being epistatic. These algorithms are applied to both simulated and real data. The real data represents the genotype-phenotype values for AMD disease. Proposed relevance-chain algorithms have detected some highly associated markers with AMD. C# source files for relevance-chains algorithm are freely available at https://www.sharemation.com/amanzour.

1. Introduction

Gene mapping intends to identify the causal genetic regions of a phenotype. A study in which genotypes are Single Nucleotide Polymorphisms (SNPs) and the phenotype under investigation is a complex disease, is referred to as a SNP case/control study. SNPs are believed to have strong relation with genetic diseases. In case of complex diseases, the contributing loci are unknown and there is no complete knowledge of which genes the causal SNPs are located in or how many causal regions exist.

Various methods are used to identify the causal SNPs most of which are statistical and based on assumption. More recent approaches, however, are based on information theory [1, 2]. Mutual information is one of the basic concepts in information theory which has recently been applied to many problems in molecular biology [4, 5]. Information theoretic measures have been widely used in data mining problems such as the on here [12].

Different approaches are presented to locate the causal regions. For this purpose, algorithms presented in [1] have proven useful. However, one of the shortcomings of this method is that it is futile against epistatic models.

In this paper, a new criterion called conditional mutual information is presented and compared to that of in [1]. It is important to note that the proposed criterion is not peculiar to gene mapping and holds for any type of association mapping in general. Furthermore, recursive algorithms inspired by this new criterion are presented to find the relevance chains in SNP case/control studies. The presented algorithms, which are in favor of the detecting epistatic models, are applied to both simulated and real datasets.

2. Information Theory

In this section a brief review will be presented on basic concepts and definitions of information theory. The most basic and conceptual figure of information theory is entropy. The entropy, or uncertainty of a variable X with realizations x is defined as follows:

$$H(X) = \sum_{x} p(x) \log \frac{1}{p(x)} \tag{1}$$

Similarly, the amount of dependence between two variables or two sets or variables is quantized by mutual information figure:

$$I(Y;X) = \sum_{y} \sum_{x} p(y,x) \log \frac{p(y \mid x)}{p(y)} \tag{2}$$

The discussed measures have many convenient and useful characteristics many of which are explained in [9] extensively. A few of these characteristics are

stated here that prove useful to population-based gene mapping. Two useful theorems are presented that state the convexity conditions of these functions.

Theorem 1:

Entropy is a convex-$\cap$ function on probability set of its arguments. Mutual information is also a convex-$\cap$ on the probability set of either of its arguments, here $p(x)$ or $p(y)$.

Theorem 2:

Mutual information is a convex-$\cup$ function on the transition probability set of either of its arguments, here on $p(x|y)$ or $p(y|x)$ while the probability sets $p(y)$ or $p(x)$ are kept constant, respectively.

3. Mutual Information and Gene Mapping

Most of the previous works done so far have statistical base whereas in recent years, many methods are based on information theoretic concepts including [10, 11]. In a simple case-control study, there is basically an M.L matrix where M is the number of people and L is the number of available SNPs. It is stated in [1] that causal regions $\mathbf{S}_N$ can be found by equation 6.

$$\mathbf{S}_N = [S_1 \quad S_2 \quad \dots \quad S_N]$$
$$\mathbf{S}_N^{old} = \arg\max_{\mathbf{S}_N} I(\mathbf{S}_N; P) \tag{3}$$

The above maximization is a combinatorial problem in nature and can be relaxed to stepwise maximization which is implicitly referred to in [1]. The relaxed method aims to find $\mathbf{S}_N^{old}$ in a stepwise manner. Let's call the resulting vector $\mathbf{S}_N^{old^*}$. Elements of $\mathbf{S}_N^{old^*}$ are computed as follows:

$$\begin{aligned}
&\mathbf{S}_N^{old^*} = [S_1^{old^*} \dots S_N^{old^*}] \\
&where \quad S_N^{old^*} = \arg\max_{1\le i\le L} I(S_i; P \mid S_1^{old^*} S_2^{old^*} \cdots S_{N-1}^{old^*}), \\
&where \quad S_{N-1}^{old^*} = \arg\max_{1\le i\le L} I(S_i; P \mid S_1^{old^*} S_2^{old^*} \cdots S_{N-2}^{old^*}), \\
&\vdots \\
&where \quad S_2^{old^*} = \arg\max_{1\le i\le L} I(S_i; P \mid S_1^{old^*}), \\
&where \quad S_1^{old^*} = \arg\max_{1\le i\le L} I(S_i; P)
\end{aligned} \tag{4}$$

3.1. Mutual information Criterion versus Conditional Mutual Information Criterion

In this section, a new criterion for association mapping is presented and compared to the existing one from different aspects. Let's assume that the number of individuals is in the order of $|SNP|^{genotypes}$. The figure $|SNP|$ is the number of realizations of markers and *genotypes* is the number of SNPs available in the database. The new criterion states:

$$\begin{aligned}
&\mathbf{S}_N^{new} = \arg\max_{\mathbf{S}_N} I(\mathbf{S}_N; P \mid \mathbf{S}_{L-N}), \\
&where \quad \mathbf{S}_N \cap \mathbf{S}_{L-N} = \phi
\end{aligned} \tag{5}$$

The above maximization is equivalent to the minimization in equation 6.

$$\mathbf{S}_{L-N}^{new} = \arg\min_{\mathbf{S}_{L-N}} I(\mathbf{S}_{L-N}; P). \tag{6}$$

Since this minimization also takes a lot of time and complexity, we will resort to the relaxed method in, that is finding $\mathbf{S}_{L-N}^{new^*}$ such that

$$\begin{aligned}
&\mathbf{S}_{L-N}^{new^*} = [S_1^{new^*} \dots S_{L-N}^{new^*}] \\
&where \quad S_{L-N}^{new^*} = \arg\max_{1\le m_N\le L} H(P \mid S_1 S_2 \cdots S_{L-N-1}^{new^*} S_i), \\
&where \quad S_{L-N-1}^{new^*} = \arg\max_{1\le i\le L} H(P \mid S_1^{new^*} S_2^{new^*} \cdots S_{L-N-2}^{new^*} S_i), \\
&\vdots \\
&where \quad S_2^{new^*} = \arg\max_{1\le i\le L} H(P \mid S_1^{new^*} S_i), \\
&where \quad S_1^{new^*} = \arg\max_{1\le i\le L} H(P \mid S_i)
\end{aligned} \tag{7}$$

The interpretation of presented criterion is as follows: In a system where all variable possess unknown dependencies, the causal variant about the system outcome are the ones without which maximum uncertainty occurs about that particular outcome.

The two criteria along with their stepwise relaxed methods will be discussed and compared from different aspects.

3.1.1. Theoretical and Computational Comparisons having looked at the figure of mutual information in the right most equality of equation 3, it is clear that choosing different sets of X only changes the figure through entropy of Y conditioned on X. Since $H(Y)$ is constant, making mutual information a convex-$\cup$ function upon different choices of X.

However, it is obvious that $I(\mathbf{S}; P)$ has as many as $|P|.|\mathbf{S}|$ distinct theoretical maxima. This makes it unlikely for $\mathbf{S}_N^{old^*}$ to be same as $\mathbf{S}_N^{old}$, since theoretical maximization can not be achieved in a stepwise

manner. The amount of computation of the stepwise approach to satisfy the former criterion is proportional to N while that of the new criterion, it is L-N.

3.1.2. Epistasy. As previously stated, epistatic models are known as models in which the causal regions do not have high correlation with the disease individually. In such cases, the probability for $\mathbf{S}_N^{old}$ and $\mathbf{S}_N^{old^*}$ to be equal is:

$$p(\mathbf{S}_N^{old^*} = \mathbf{S}_N^{old}) = p(no\ epistasy).\frac{1}{|P|.|\mathbf{S}|} + p(epistasy).0$$
$$= \frac{1}{2}|P|.|\mathbf{S}| \tag{8}$$

In the maximization method presented for the second criterion, however, it is virtually impossible to eliminate all elements of an epistatic model. $\mathbf{S}_{L-N}^{new^*}$ can never contain all elements of an epistatic model. Therefore, at least one causal variant is never eliminated through stepwise maximization presented in 11, no matter what kind of model the disease follows.

Therefore, one of the big advantages of the stepwise method presented to satisfy the second criterion is that it can trap one of the disease causal regions through stepwise elimination no matter how many causal regions the disease has.

4. Relevance-Chains Algorithms

Complex diseases are believed to have multiple contributing genetic regions often having subtle patterns. Besides statistical concepts, mutual information has also been used to in relevance-chains algorithms [1, 6]. In a case/control study, the number of available samples are usually much less that the genotypes.

We present new algorithms for finding relevance chains that follow epistatic models. For each chain of SNPs $\{S^r, \mathbf{S}_J^*)\}$ is found such that:

$$S^r = \arg\max_{1\le i\le L} I(S_i; P \mid \mathbf{S}_J^*)$$
$$where \quad \mathbf{S}_J^* = \arg\max_{\mathbf{S}_J} H(\mathbf{S}_J \mid P) \tag{9}$$

Three different approaches are explored in order to find more effective sets of $\mathbf{S}_J^*$ all of which can be done through stepwise maximization.

$$Mode\ A: \quad \mathbf{S}_J^* = \arg\max_{\mathbf{S}_J} H(\mathbf{S}_J)$$
$$Mode\ B: \quad \mathbf{S}_J^* = \arg\max_{\mathbf{S}_J} H(\mathbf{S}_J \mid P) \tag{10}$$
$$Mode\ C: \quad \mathbf{S}_J^* = \arg\max_{\mathbf{S}_J} H(P \mid \mathbf{S}_J)$$

Recursive algorithms are best suited for finding the above relevance-chains because of the recursion in Shannon Entropy. For instance, if J=2, we have:

$$H(\mathbf{S}_2) = H(S_1) + H(S_2 \mid S_1)$$
$$H(\mathbf{S}_2 \mid P) = H(S_1 \mid P) + H(S_2 \mid PS_1)$$
$$H(P \mid \mathbf{S}_2) = H(P \mid S_1) + H(S_2 \mid PS_1) - H(S_2 \mid S_1) \tag{11}$$

Length of each chain is limited to a maximum arbitrary number maxDepth. The selection of SNPs at each step can be done in two ways. In the first approach, all SNPs having entropies higher than a threshold are chosen. $T = H_{\max}(Z) - \varepsilon$, where $H_{\max}(Z)$ is the maximum achievable entropy in the chosen mode. In the second approach, M SNPs are picked that correspond to the M maximum conditional entropy figures in each step. Statistical significance of values is verified using equation 12.

$$I(S; P \mid \mathbf{S}) \ge \Gamma_{1-\alpha}\left(\frac{|\mathbf{S}|}{2}(|S|-1)(|P|-1), \frac{1}{N \ln 2}\right) \tag{12}$$

The relevance-chains algorithm is presented in figure 1. Based on the chosen mode, M SNPs are selected in each depth.

```
Input=Data, maxDepth, M, Mode
Ent=entropyfigure
Repeat (Ent, Depth)
S^M=M corresponding markers with maximum value in Ent
If (Depth =0) then return and save all elements of S^M
Else {
For i=1 to M
Ent=Search(S_i, Depth)
Repeat(Ent, Depth-1)
}
end
```

Figure 1: Relevance-chains algorithm, any of the 3 modes can be selected in line 3.

The number of relevance chains formed is equal to $M^{\max Depth}$. By increasing M more diverse chains are formed but at the cost of a higher computational time.

Table 1 shows a comparison between the proposed relevance-chains algorithm and that of Zaher Dawy et al presented in [1].

Table 1: Comparison between the proposed relevance-chains algorithm and and that of Zaher Dawy et al [1].

	Proposed algorithm	Zaher Dawy et al. algorithm
Generality	Acceptable	Acceptable
Computational Complexity	Acceptable	Acceptable
Inter-chain and intra-chain diversity	Capable	Not capable
Discovering epistatic models	Capable	Not capable
Discovering other models	Capable	Capable

5. Application

5.1. Data sets

The proposed methods were tested on simulated and real datasets. The simulated data assumed 10000 SNPs which were assumed ternary variables. Hundred samples, were simulated half of which were considered as controls and the other half as cases. All variables assumed to be iid and therefore were generated randomly.

The real data set that we implemented the relevance-chains algorithm on related to AMD disease which was previously used in [7]. The database is available as an open source in [8]. 116212 SNPs were extracted and recorded for 146 individuals. 50 individuals were controls and the other 96 were cases. The genotypes all had three realizations as the simulated dataset; one for heterozygous and two for homozygous combinations.

5.2. Results

By applying the relevance-chains algorithm in [1] on the random dataset, maximum two-locus mutual information achieved was 0.4007. This figure was achieved while M=10 and did not improve even at M=40. However, all three modes of the relevance-chains algorithms presented in equation 10 were applied to the same simulated data and the results are shown in figure 2. The highest mutual information value achieved by relevance chains is illustrated for different M while depth is one. It is clear that by increasing M more diversified chains are formed and hence, higher mutual information is achieved.

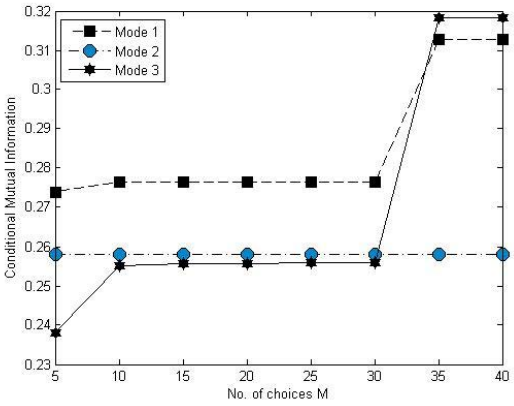

Figure 2: Results for all 3 modes of relevance-chin algorithm applied to random datasets for different M and depth=1. The maximum figure is illustrated at each step.

Applying different modes of the proposed algorithm on AMD disease dataset also revealed chains having epistatic models.

The relevance-chain algorithm was applied with depth=1 and M=5 and some 2-tuple chains were observed. Markers rs2187210 (SNP_A-1702347) and rs10498077 (SNP_A-1667319) together caused mutual information of 0.9216. However, marker rs2187210 and the disease vector alone make only 0.0034 bits of mutual information while marker 43884 makes 0.2056 bits. Another relevance chain of equal value which was 0.9216 is made between rs2187210 and rs9304554 (SNP_A-1689432). Rs9304554 alone has mutual information of 0.0631 bits. Both modes B and C detected these chains. Mode B also detected a somewhat epistatic-model chain between markers rs2187210 and rs6845733 (SNP_A-1702798) that lead to a joint mutual information of 0.81076. Rs6845733 by it self makes as small as 0.0238 bits of information about the disease. Rs542359(SNP_A-1673543) which only has 0.0034 bits of mutual information, jointly with rs10498077 made up to about 0.9218 bit of information about the disease which was detected by mode B.

Markers 43884 and 93244 might be considered as major contributing loci since they always appeared at the end of the relevance chains. Most of the achieved figures have low statistical significance due to small number of available samples for certain SNPs. This is because the values of many SNPs were not successfully extracted from the samples and hence were marked as “No Call” which greatly affected the statistical significance of the results.

6. Discussion and Conclusions

Interaction between genetic regions can be tricky and subtle. The presented algorithms mentioned in this work tend have the advantage of not being based on any apriori assumption while most of the statistical approaches are based on such assumptions.

As stated in the literature, one of the major disadvantages of a case/control study is that high association does not necessarily establish causality. In this work, a different criterion was proposes for assessing cause and effect relationship amongst data. The relaxed version of this criterion which is based on stepwise maximization proves useful in detecting causal regions of epistatic models while having a reasonable computational complexity. This opens way for dealing with complex diseases having unknown number of susceptibility regions.

Algorithms were also presented in order to find relevance chains having epistatic models. Epistasy-model search is what makes this approach distinct from other SNP case-control studies. The idea here is to avoid the full search by only checking those pairs that have a higher probability of containing the subtle patterns.

Parameters M and T can be used to control computational time or amount of epistasy achieved in resulting chains, respectively. In the first case, by setting a threshold T, the amount of diversity of each chain is controlled whereas in the second case, by setting M, computational complexity can be handled. However, even if the disease possesses models other than epistatic, the proposed algorithms can also detect the major contributing loci at the last step of the algorithm.

The available dataset for AMD disease contains many epistatic models amongst its markers. This can imply that the above disease might be controlled by the epistatic model. However, the statistical significance of the results is fairly low which calls for further investigation on the available genotypes.

Both stepwise maximization of the proposed criterion and the presented algorithms have reasonable computational complexity, meaning that their complexity only grows linearly with increase in the number of markers or samples. This reasonable computational time makes them convenient for applications on a *genome-wide* level.

References

[1] Z. Dawy, B. Goebel, "Gene mapping and marker clustering using Shannon's mutual information," *IEEE Trans. On computational biology and bioinformatics*. Vol. 3, NO. 1, January-March 2006.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech J.* vol 27, pp 623-656, July-October 1948.

[3] Cover T and Thomas J. Elements of Information Theory. Jon Wiley Sons
Inc (1993).

[4] A. Butte and I. Kohane, "Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements," Proc. Pacific Symp. Biocomputing, pp. 418-429, Jan. 2000.

[5] J. Kasturi, R. Acharya, and M. Ramanathan, "An Information Theoretic Approach for Analyzing Temporal Patterns of Gene Expression," Bioinformatics, vol. 19, no. 4, pp. 449-458, 2003.

[6] A. Butte and I. Kohane, "Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements," Proc. Pacific Symp. Biocomputing, pp. 418-429, Jan. 2000.

[7] Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, Sangiovanni JP, Mane SM, Mayne ST, et al.: Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science 2005, 308:385-389.

[8] http://variation.yale.edu/dataDownload.html.

[9] Cover, T. and Thomas, J., *Elements of Information Theory*. New York: John Wiley & Sons, 1991.

[10] J. Hampe, S. Schreiber, and M. Krawczak, "Entropy-Based SNP Selection for Genetic Association Studies," Human Genetics, vol. 114, no. 1, pp. 36-43, Dec. 2003.

[11] L. Cardon and J. Bell, "Association Study Designs for Complex Diseases," Nature Rev. Genetics, vol. 2, no. 2, pp. 91-99, Feb. 2001.

[12] Toivonen, H., Onkamo, P., Hintsanen, P., Terzi, E., and Sevon, P., *Data Mining for Gene Mapping,* Department of Computer Science and Helsinki Institute of Information Technology, University of Helsinki, Finland, Ph. D. Thesis, 2004.