A Novel Information Theoretic Method for Detecting Gene-Gene And Gene-Environment Interactions in Complex Diseases

Pritam Chanda, Aidong Zhang and Murali Ramanthan

Abstract— **Gene-gene and gene-environment interactions play important roles in the etiology of complex multi-factorial diseases. With the advancements in genotyping technology, large genetic association studies based on hundreds of thousands of single-nucleotide polymorphisms are a popular option for the study of complex diseases. In this paper we use information theoretic concepts to develop a novel method for detecting statistical gene-gene and gene-environment interactions in complex disease models. We explore the effectiveness of our method with extensive simulations using different gene-gene interaction models and the rheumatoid arthritis dataset from Genetic Analysis Workshop-15. The performance of the method was compared to the well known multi-factor dimensionality reduction (MDR) and generalized MDR (GMDR) methods. We demonstrate that our method is capable of analyzing a diverse range of epidemiological data sets containing evidences for gene-gene interactions.**

I. INTRODUCTION

The risk of developing many common and complex diseases such as cancer, autoimmune disease and cardiovascular disease involves complex interactions between multiple genes and several endogenous and exogenous environmental factors (or covariates). Because of their abundance on the genome (on average every 100 to 300 bases), the single nucleotide polymorphisms (SNP) have become major source of information for detecting statistical gene-gene and gene-environment interactions underlying the etiology of complex diseases. The successful detection of critical gene-gene and gene-environment statistical interactions can provide the scientific basis for many underlying biological interactions, improves the prospects for uncovering potentially undiscovered genes involved in the disease process and helps to develop preventative and curative measures for particular genetic susceptibilities.

Traditional single-locus based disease-SNP association analysis studies fail to detect all the relevant loci when observable marginal effects at each locus are small [7] [9]. Prominent methods for multi-loci disease association analysis are multi-factor dimensionality reduction (MDR), generalized MDR (GMDR) and regression based methods [15][6]. MDR is a non-parametric method [18] that uses constructive induction wherein the dimensionality of the multi-locus genotype is systematically reduced by pooling into high and low risk groups. The recently proposed GMDR method employs the generalized linear model framework for scoring in conjunction with MDR for dimensionality reduction [14]. GMDR enables inclusion of covariates and handles both discrete and continuous traits in population-based study designs. However, despite availability of a more efficient parallel computing implementation [4], MDR and its variants, including GMDR, are computationally intensive, especially when more than 10 polymorphisms need to be evaluated [19]. The regression based methods are also computationally intensive and model complexity increases rapidly with the increase in number of loci and also with the number of possible allelic states at each locus. Also most regression based methods have been limited to analysis of two-locus interactions.

Information theoretic methods are among the most promising approaches for genetic disease association studies and are versatile and are independent of the underlying genetic models. But only limited research has been done on leveraging these strengths for analysis of multi-locus disease association studies. Several reports have used the Kullback-Leibler divergence and mutual information for genetic analysis. They have been applied for 2-group comparisons such as those used to evaluate ancestry informative markers [2][18][20], as multi-locus linkage disequilibrium (LD) measure to identify tag SNPs [13] and for analytical visualization [3][5]. Information theory based statistics have been proposed for genome-wide data analysis to test for allelic associations [22] and in identifying and visualizing gene-gene and gene-environment interactions [5].

Interaction information between variables was researched upon in diverse areas like physics, information theory, biology, neuroscience, game theory, law and economics. The concept was first introduced by McGill in 1954 [16] as a multivariate generalizations of Shannon's mutual information [21] and Han [8] gave rigorous formal definitions of the concepts of interaction and more recently, Jakulin [10] [11] studied it extensively from a machine learning perspective and provided methods for visualizing interactions between the data attributes. In this paper, we use the interaction information measure to develop a novel method for detecting statistical gene-gene and gene-environment interactions in complex disease models. We critically evaluate the performance of the method using extensive simulation studies and also using the simulated rheumatoid arthritis dataset from Genetic Analysis Workshop 15 (GAW-15) [1].

Manuscript submitted June 15, 2008. This work was partly supported by the NSF and the NIH.

P. Chanda is PhD student with the Department of Computer Science and Engineering, State University of New York, Buffalo, NY, USA. pchanda@cse.buffalo.edu

A. Zhang is with Faculty of Computer Science and Engineering, State University of New York, Buffalo, NY, USA. azhang@cse.buffalo.edu

M. Ramanthan is with Faculty of Pharmaceutical Sciences, State University of New York, Buffalo, NY, USA. murali@buffalo.edu

II. RELEVANCE AND REDUNDANCY

In this section, we first define the information theoretic metrics and then describe in details an algorithm that uses the metrics in detecting gene-gene interactions.

A. Information theoretic measures

Interaction Information. Assume $S = \{V_1; V_2; ...; V_n\}$ be the set of genetic or environmental variables in a given data set and C be the disease status (phenotype) variable. The uncertainty of a variable V_i taking values from set U_i is given by Shannon's entropy [21] as,

$$H(V_i) = -\sum_{v \in U_i} p(V_i = v) log_2 p(V_i = v) \quad (1)$$

The interaction information among the k variables (referred to as k-way interaction information ,we shall call it $KWII$) in set $S' = \{V_1; V_2; ...; V_k\}$, $S' \subseteq S$ is the multivariate generalizations of Shannon's mutual information. It is defined as the amount of information (redundancy or synergy) present in the set of variables, which is not present in any subset of these variables [10]. For set S', the $KWII$ can be written succinctly as an alternating sum over all possible subsets τ of S' using the difference operator notation of Han [8]:

$$KWII(V_1; V_2; ...; V_k) = -\sum_{\tau \subseteq S'} (-1)^{|S'|-|\tau|} H(\tau) \quad (2)$$

Relevance. The interaction information given by $KWII(S', C) = KWII(V_1; V_2; ...; V_k; C)$ is a measure of the *relevance* of the set of variables in set $S' = \{V_1; V_2; ...; V_k\}$ towards the disease phenotype variable C (i.e. how well the set explains the disease phenotype). If variables V_1,...,V_{k-1} are already known to be *relevant* for C, then $KWII(S'; C)$ gives the relevance of V_k in combination with the others towards C. The value of $KWII(S'; C)$ can be both positive and negative where larger positive values indicate stronger interaction information (hence higher relevance) among the variables in S' and C. So we shall use only positive $KWII$ values as the measure of relevance in our algorithm.

Redundancy. Let $S_1 = \{V_i; ...; V_j\}$ and $S_2 = \{V_t; ...; V_k\}$ be two sets of variables. Then the mutual information (MUI) between the variables in sets S_1 and S_2 is a measure of the amount of information (i.e. redundancy) shared between the variables of each set and is given by $MUI(V_i...V_j; V_t...V_k) = I(V_i...V_j; V_t...V_k)$ (note the placement of the ; indicating that MUI is actually $KWII$ involving the joints of the variables in S_1 and S_2). The mutual information is maximum when the two sets consist of identical variables so that the redundancy is also maximized.

The *Relevance* and *Redundancy* criteria mentioned above are used together in an iterative search algorithm to select promising genetic and environmental variables along with the interactions in which these variables participate.

B. The Relevance-Redundancy Algorithm

The goal of the algorithm is to determine all the *non-redundant* and *relevant* variables and their associated interactions that involve the disease phenotype. Let $S = \{V_1; V_2; ...; V_n\}$ denote the set of all genetic and environmental variables to be searched for interactions and C is the disease phenotype variable. The algorithm proceeds in an iterative fashion and in each iteration it examines the relevancy of each variable for C and also its redundancy with variables already chosen to be relevant in previous iterations. Let L^i denote the set of variables that participate in one or more interactions with the phenotype variable that are relevant by the above definition till the i^{th} iteration, i.e. L^i is the set of relevant variables at iteration i. Let Q^i denote the set of relevant interactions detected till the i^{th} iteration. Thus $L^0 \subset L^1 \subset ...L^i...$ and $Q^0 \subset Q^1 \subset ...Q^i....$ In the beginning L^0 and Q^0 are empty. At iteration 1, L^1 and Q^1 are updated as,

$$\begin{aligned} V_{max} &= \underset{V_k}{\text{argmax}}\{KWII(V_k; C)\}, k = 1, 2, ..., n \\ L^1 &= L^0 \cup \{V_{max}\} = \{V_{max}\} \\ Q^1 &= Q^0 \cup \{\{V_{max}; C)\}\} \end{aligned} \quad (3)$$

Thus since Q^0 is empty, only the relevance of each variable is determined and is used as the sole criteria to select the a variable and its associated interactions to be added to L^1 and Q^1 respectively. In iteration $i > 1$, a variable V_k not already in L^{i-1} is tested for inclusion in L^i. Since the variable V_k may be interacting with zero or more variables already in L^{i-1} and C, the maximum relevance of V_k in combination with a subset of already selected variables is determined. At the same time, V_k may be redundant with zero or more variables already in L^{i-1}; therefore, the maximum redundancy of V_k in combination with a subset of already selected variables is determined. The following equations are used to choose the variable to be included in L^i and update Q^i in step i,

$$\begin{aligned} V_{max} &= \underset{V_k}{\text{argmax}}\{\max_{\tau \in Q^{i-1}} KWII(\tau; V_k; C) \\ &\quad - \max_{\gamma \in Q^{i-1}} MUI(\gamma; V_k)\}, V_k \notin L^{i-1}, k = 1, 2, ..., n \\ L^i &= L^{i-1} \cup \{V_{max}\} \\ Q^i &= Q^{i-1} \cup \{\underset{\tau \in Q^{i-1}}{\cup}\{\tau; V_k\}\} \end{aligned} \quad (4)$$

Thus the variable with the maximum relevance and minimum redundancy is selected for inclusion. A variable that has high relevance but high redundancy with some variable(s) already selected in pervious iterations will be ignored compared to a variable that has, say, moderate relevance and much lower redundancy with already selected variable(s). The details of the algorithm are given below.

Algorithm *Relevance-Redundancy Search*(S, C, κ)
Input: S(Set of variables), C(phenotype), κ(# of iterations)
Output: Q(Relevant interactions)
1. $L \leftarrow \phi$; $Q \leftarrow \{\phi\}$; $max\_relv \leftarrow -\infty$; $V_{max} \leftarrow \phi$;
2. **for** each $V_k \in S$ **do**

3. $relv \leftarrow KWII(V_k; C)$;
4. **if** $relv \geq max\_relv$
5. $V_{max} \leftarrow V_k$;
6. $max\_relv \leftarrow relv$;
7. **end**
8. **end**
9. $L \leftarrow \{V_{max}\}; Q \leftarrow Q \cup \{\{V_{max}\}\}$;
10. **for** $iter \leftarrow 2$ **to** κ **do** /*each iteration*/
11. /*Q_{temp} retains interacting combinations and $KWII$
12. values for the most relevant variable in this iteration*/
13. $Q_{temp} \leftarrow \phi$; $max\_score \leftarrow -\infty$;
14. **for** each $V_k \in S \backslash L$ **do** /*each new variable*/
15. $max\_relv \leftarrow -\infty$;$R \leftarrow \phi$;
16. **for** each $\tau \in Q$ **do** /*max relevance of V_k*/
17. $relv \leftarrow KWII(V_k; \tau; C)$;
18. **if** $relv \geq max\_relv$
19. $max\_relv \leftarrow relv$;
20. **end**
21. /*retain $\langle$combination,relevance$\rangle$ pair*/
22. $R \leftarrow R \cup \langle \{V_k; \tau\}, relv \rangle$;
23. **end**
24. **if** $max\_relv > 0$
25. $max\_redncy \leftarrow -\infty$;
26. **for** each $\tau \in Q$ **do** /*max redundancy of V_k*/
27. $redncy \leftarrow MUI(V_k; \tau)$;
28. **if** $redncy \geq max\_redncy$
29. $max\_redncy \leftarrow redncy$;
30. **end**
31. **end**
32. $score \leftarrow max\_relv - max\_redncy$;
33. **if** $score \geq max\_score$
34. $max\_score \leftarrow score$;
35. $V_{max} \leftarrow V_k$;
36. $Q_{temp} \leftarrow R$;
37. **end**
38. **end**
39. **end**
40. /*retain only combinations with $KWII > 0$*/
41. **for** each $\langle \tau, relv \rangle \in Q_{temp}$ **do**
42. **if** $relv > 0$
43. $Q \leftarrow Q \cup \{\tau\}$;
44. **end**
45. **end**
46. $L \leftarrow L \cup \{V_{max}\}$;
47. **end**
48. **for** each $\tau \in Q$ **do**
49. **if** $KWII(\tau; C)$ is not significant
50. $Q \leftarrow Q \backslash \{\tau\}$;
51. **end**
52. **end**
53. **return** Q;

C. Algorithm Description and Computational Complexity

The algorithm takes as input the set of genetic and environmental variables S, the phenotype variable C, and the number of iterations κ. The set of informative variable combinations and their $KWII$ values is the output (Q). Equation 3 is used to update L (the set of relevant variables) and Q (the set of associated interactions) with the first informative variable and its associated interaction in lines 1-9. In each succeeding iteration, equation 4 is used to select a variable according to the information theoretic criteria; for each new variable V_k, lines 15-23 calculates the maximum relevance of V_k in combination with variables already selected in previous iterations and lines 25-31 calculates the maximum redundancy of V_k in combination with variables already selected in previous iterations. The difference of the two quantities is used as a measure to select the most informative variable in the current iteration and L and Q are updated with it. Only the interactions with $KWII > 0$ are retained in Q in each iteration. Finally, only the interactions that are statistically significant are returned as output. Significance of $KWII$ value of each combination can be easily ascertained using permutation or bootstrap based methods. We next present rigorous simulation studies to analyze the effectiveness of the *Relevance-Redundancy* algorithm.

III. SIMULATIONS FOR CASE STUDY

A complex case study involving simulated data sets was used to critically assess the effectiveness of the above algorithm in correctly identifying the interacting variables causing the disease. The simulated data consisted of 42 biallelic SNP variables numbered 0-41 and the case-control phenotype variable C. The SNPs were arranged into six groups ($G1 - G6$) with seven SNPs in each group (see Figure 1). Various levels of linkage disequilibrium (LD) was simulated between the SNPs in each group. Four SNPs (denoted $S1$, $S2$, $S2$ and $S4$), each randomly selected from SNPs in $G1$, $G2$, $G5$ and $G6$ respectively were assumed to be involved in the disease process though models of complex interaction. The remaining six SNPs in each group were simulated to be in various levels of LD (r^2 values 0.9,0.8 and 0.7) with the the causative SNP of that group. The disease was modeled to occur using two-locus gene-gene interaction models between SNPs $S1$ and $S2$ (i.e interaction between $G1$ and $G2$), and $S3$ and $S4$ (i.e interaction between $G5$ and $G6$) that attempt to mimic biological interactions. SNPs in groups $G3$ and $G4$ were not associated with the disease. In an effort to classify the types of interaction in the case of two biallelic loci, Li and Reich [12] have enumerated 512 possible two-locus models and identified a fewer number of non-redundant two-locus models. We choose two widely used models for our simulation. Each model specifies the penetrance of the disease given the genotypes of the two interacting loci. Let the two loci be denoted by L_1 and L_2. Let the two alleles at loci L_1 be A and a (genotypes are aa, Aa and AA), and at loci L_2 be B and b (genotypes are bb, Bb and BB). Let λ_{aa} , λ_{Aa} , λ_{AA} be the marginal penetrances at L_1 and λ_{bb} , λ_{Bb} , λ_{BB} be the marginal penetrances at L_2. Denote the joint penetrances for each genotype g of the two loci by μ_g, i.e. $P(Disease|g) = \mu_g$. Then the marginal penetrances at

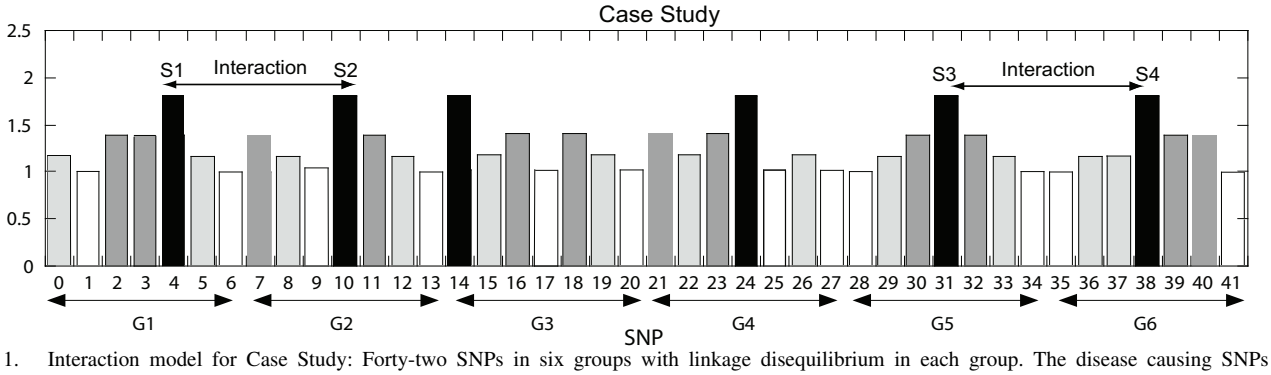

Fig. 1. Interaction model for Case Study: Forty-two SNPs in six groups with linkage disequilibrium in each group. The disease causing SNPs are $S1$,$S2$,$S3$,$S4$ (black). In each group, two SNPs each are in LD of 0.9 (dark grey),0.8(light grey) and 0.7(white) with the black colored SNP of that group. The y-axis gives the pairwise mutual information between each SNP and the SNP colored black in that group which increases with increasing LD between SNP pairs indicating redundancy. We would like to detect only the diseased SNPs (black) from groups $G1$, $G2$, $G5$ and $G6$.

each locus is given by,

$$\lambda_{g_1} = \sum_{g_2 \in \{bb,Bb,BB\}} \mu_{g_1 g_2} P(g_2), \quad g_1 \in \{aa, Aa, AA\}$$
$$\lambda_{g_2} = \sum_{g_1 \in \{aa,Aa,AA\}} \mu_{g_1 g_2} P(g_1), \quad g_2 \in \{bb, Bb, BB\} \quad (5)$$

And the overall population prevalence of the disease is given by,

$$P(Disease) = \sum_{g_1 \in \{aa,Aa,AA\}, g_2 \in \{bb,Bb,BB\}} \mu_{g_1 g_2} P(g_1 g_2) \quad (6)$$

The genotype frequencies can be calculated using the allele frequencies at each loci under Hardy Weinberg equilibrium assumptions. The two models (Figure 2) are summarized below. Model 1 is an additive model that has a baseline penetrance for genotype $aabb$ and it increases in an additive fashion with each copy of the disease causing allele in the genotype. Model 2 incorporates a multiplicative interaction with a baseline value that increases the chance of disease multiplicatively when at least one disease causing allele from each locus is present [12]. Given fixed values of the disease prevalence and the allele frequencies at each locus, for each model, the marginal effects at each locus are bounded by some maximum value η and the interaction effects (θ and α) are solved for by working backwards using the above equations. The bounds on the marginal effect sizes at each locus are specified as,

$$\lambda_{AA}/\lambda_{Aa} \leq \eta, \lambda_{Aa}/\lambda_{aa} \leq \eta,$$
$$\lambda_{BB}/\lambda_{Bb} \leq \eta, \lambda_{Bb}/\lambda_{bb} \leq \eta \quad (7)$$

From equations 5-7, once the allele frequencies, $P(Disease)$, and η are known, we can solve for the interaction effects (θ and α) using iterative numerical methods such that the bounded marginal effects are maximized at each locus.

IV. Experimental Results

For the case study, a population of 500,000 individuals with genotypes in Hardy-Weinberg equilibrium and given

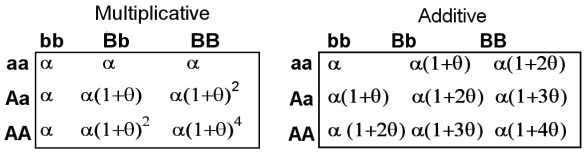

Fig. 2. The disease models used in the case study. Each entry indicates the disease penetrance given a genotype , e.g. $P(Disease|AaBb) = \mu_{Aabb} = \alpha(1+\theta)$ for the Multiplicative Model.

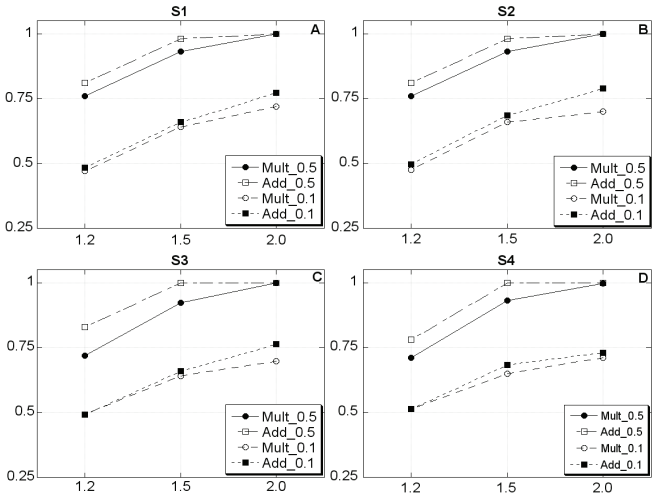

Fig. 3. The power of detecting the diseased SNPs $S1$, $S2$, $S3$ and $S4$ for disease allele frequencies of 0.1 and 0.5 (y-axis) against η (x-axis). The disease model is either Additive (between diseased SNPs $S1$, $S2$ and $S3$, $S4$) or Multiplicative (between diseased SNPs $S1$, $S2$ and $S3$, $S4$).

allele frequencies was generated and 2000 cases and 2000 controls were randomly selected from the population. The same disease model (Multiplicative or Additive) was assumed in both the interactions. The disease prevalence was fixed at 0.1 and the maximum marginal effect size (η) was varied as 1.2, 1.5 and 2.0. The magnitudes of the maximum marginal effect sizes were chosen based on known results about complex diseases and previous works [15]. The frequencies of the disease alleles were assigned 0.1 and 0.5 in two separate experiments, while the uninformative SNPs in $G3$ and $G4$ had allele frequencies of 0.5. We conducted 1000 independent simulations for each of the

maximum marginal effect sizes and the two allele frequencies at the two loci. For each experiment with a given allele frequency, the significance of the observed $KWII$ values of each interacting combination output by the algorithm was determined using strategies similar to that described in [5]: the null distribution of $KWII$ for each combination was obtained by calculating it on genotypes simulated with a marginal effect size of unity at each locus (θ=0) and each observed $KWII$ was deemed significant if it exceeded the 95^{th} percentile value of the corresponding null distribution. A one-sided analysis was assumed since we are interested in variables involved in interactions with $KWII > 0$ (since positive values indicate the presence of an interaction).

A. Simulation Results

Figure 3 shows the power of detecting the disease causing SNPs $S1$, $S2$, $S3$ and $S4$. A SNP is deemed detected when it participates in one or more interactions with significant $KWII$ values output by the method. The number of iterations, κ was set to 6. We observe that for disease allele frequency of 0.5, our method achieves high power of 85-88% (Multiplicative model) and 96-100% (Additive model) at maximum marginal effect size (η) of 1.5 and 75-80% power at very low $\eta = 1.2$. Power decreases with decrease in allele frequency to 0.1, but is still about 50% for $\eta = 1.2$ and increases to about 70% for $\eta = 2.0$.

B. Analysis of GAW-15 Data

We further evaluated the performance of the *Relevance-Redundancy* method using the data corresponding to problem 3 of the Genetic Analysis Workshop 15 (GAW-15) which consisted of 100 replicates simulated after the epidemiology and familial pattern of Rheumatoid Arthritis (RA), a complex genetic disease in which it is hypothesized that several loci contribute to disease susceptibility. The data contains: i) 730 microsatellite markers with an average spacing of 5 cM; ii) 9,187 SNPs distributed on the genome to mimic a 10K SNP chip set, and iii) 17,820 SNPs on chromosome 6. In addition RA affectation status (case-control variable), sex, age, smoking status, AntiCCP (anti-cyclic citrullinated peptide antibody) measure, IgM (immunoglobulin M) measure, severity, DR allele from father, DR allele from mother, age at onset, age at death are included as covariates (i.e. environmental variables). The data had 8000 samples (3468 cases of RA and 4532 controls). The AntiCCP and IgM measures were defined for the RA cases only. We have used the 9187 SNPs distributed on all the chromosomes from the first of the replicates to evaluate the our method and the remaining replicates were used to obtain the 95% confidence intervals for $KWII$ of each combination of variables found by the algorithm. Three separate analyses were done with the 9187 SNPs and Sex, Age, Smoking status as covariates and (i) RA status (ii) IgM measure and (iii) AntiCCP measure as the phenotype variable.Although phase information was provided, we chose to not include it and treated the data as unphased genotype data. Age, AntiCCP and IgM being continuous measures were each discretized by binning into five intervals of equal width. The number of iterations, κ was set to 10.

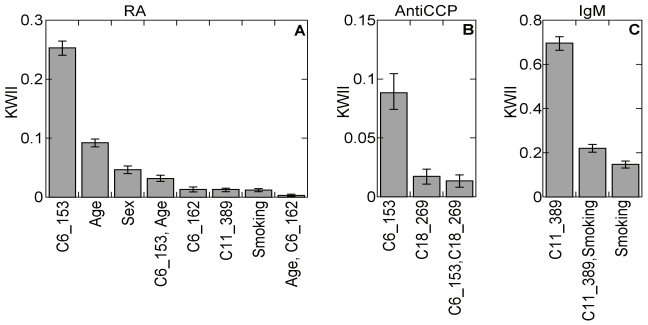

Fig. 4. The interacting variables detected using the algorithm on GAW-15 RA data using the three phenotypes.The x-axis shows the interaction combinations obtained and the phenotypes are implicit in each combination. The confidence intervals are shown on the $KWII$ values for each combination.

Figure 4 present the results for three analyses using the algorithm. The interactions in the figures were deemed significant since their confidence intervals did not span zero (zero indicates absence of an interaction). We find that our method detects the SNPs and covariates that were simulated to have associations with the RA disease. In the figures, C{chromosome no.}_{SNP no.} is used as the naming convention for the markers. In figure 4A, the combinations consist of Locus C or DR (both SNPs C6_153), Locus D (C6_162), Locus F (C11_389) and the environmental variables Age, Sex and Smoking that had associations with the RA affection status in the simulated data set [17]. The simulated data contained pronounced effects of DR and locus F on RA affection status and IgM levels, respectively, and this was confirmed by the high interaction values corresponding to the DR locus. Locus D also had a direct effect on RA risk. Although it had a very low disease allele frequency (only 0.0083, making minor allele homozygotes very rare), our information theoretic method detected it successfully. Figures 4B and 4C show the combinations obtained with AntiCCP and IgM as phenotype variables, respectively. We successfully detect Locus C or DR (SNP C6_153) and Locus E (C18_269) with AntiCCP in figure 4B and the effects of Locus F (C11_389) and Smoking on IgM in figure 4C using our method.

C. Comparison with other methods

We compare our method with two well-known methods for analysis of gene-gene interactions, MDR [18] and GMDR [14]. Both these methods attempt to explore the interaction space in a combinatorial fashion and have exponential time complexity and exploring all possible subsets containing more than one variable was not possible with 9187 SNP variables in a reasonable timeframe. So we selected 100 SNPs from among the 9187 SNPs in replication 1 of GAW-15 data to create a smaller data set that could be analyzed by all three competing methods and explored upto three variable combinations. This data set included the covariates Smoking, Age, and Sex and contained the key informative loci and the remaining SNPs were selected randomly from the rest of the 9,187 SNPs. The RA affection status was used

as the phenotype since MDR can handle only case-control phenotypes.

The MDR analysis detected {C6_153}, {C6_154, Age}, {C6_153, Age, Sex} as associated with RA. Both the SNPs C6_153 and C6_154 denote the chromosome locus C or DR. The MDR analysis did not detect Locus D (C6_162) and Smoking. GMDR requires a priori calculation of covariate effects, which are then incorporated into the analysis. Covariates cannot be analyzed alone. So the GMDR analysis was performed with Sex, Age and Smoking as the covariates and RA as trait. The method identified the following SNP combinations: {C6_153}, {C6_153, C6_162} and {C6_153, C6_154, C11_389} with Age,Sex and Smoking where SNPs C6_153, C6_154 both denote the chromosome 6 locus C or DR; C6_162 denotes locus D on chromosome 6 and C11_389 denotes locus F on chromosome 11. The Relevancy-Redundancy method detected all the relevant combinations: {C6_153}, {Age}, {Sex}, {C6_153,Age}, {C6_162}, {C11_389} and {Smoking}. Note that both MDR and GMDR detected redundant loci C6_153 and C6_154 both of which represent locus C or DR (owing to high LD between them), however, the Relevancy-Redundancy method detected only one of them (C6_153) and is thus more parsimonious (i.e avoid detecting redundant variables).

These results demonstrate that our method performs reasonably well compared to existing prominent methods and are capable of analyzing a diverse range of epidemiological data sets containing evidences for gene-gene as well as gene-environment interactions.

V. Discussion

We have presented an information theoretic method and evaluated its performances using complex simulation strategies that uses two different models of gene-gene statistical interaction. Detecting genes and environmental factors interacting to increase the susceptibility to disease risk is a very challenging task due to many reasons, particularly due to the large size of the data and presence of confounding factors like linkage disequilibrium, phenocopies and locus heterogeneity. We have shown that our information theoretic method has high power in detecting gene-gene interactions and the method is appealing not only because it is simple and performed well in the experiments and the GAW-15 data, but also because it is flexible and can be used when the genetic and environmental variables have different numbers of classes or when the phenotype has more than two classes. This means that SNP and microsatellite markers can be analyzed together if necessary. Also they are naturally extensible to study models with more than two loci and environmental variables.

We have used simulated data modeled after real disease data. The GAW-15 data set was sufficiently rich and complex because it was modeled based on a real rheumatoid arthritis data set and the ground truth is established during the simulation. For future work, we would like to test our method on several publicly available SNP data sets and also using more interaction models, particularly with models containing complex gene-gene and gene-environment interactions involving 3 or more loci in a manner similar to our simulations in [5]. We also intend to incorporate additional biological knowledge e.g. gene expression and biological pathway information along with the proposed method to make the search algorithm more biologically oriented.

References

[1] Genetic analysis workshop 15, http://www.gaworkshop.org/gaw15data.htm. 2006.

[2] E. C. Anderson and E. A. Thompson. A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, 160:1217–1229, 2002.

[3] K. Bhasi, L. Zhang, D. Brazeau, A. Zhang, and M. Ramanathan. Vizstruct for visualization of genome-wide snp analyses. *Bioinformatics*, 22(1569-1576), 2006.

[4] W. S. Bush, S. M. Dudek, and M. Ritchie. Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions. *Bioinformatics*, 22:2173–2174, 2006.

[5] P. Chanda, A. Zhang, D. Brazeau, L. Sucheston, J. L. Freudenheim, and M. Ramanathan. Information-theoretic metrics for visualizing gene-environment interactions. *Am J Hum Genet*, 81:939–963, 2007.

[6] H. J. Cordell. Epistasis: what it means, what it doesnt mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, 2002.

[7] R. Culverhouse, B. K. Suarez, J. Lin, and T. Reich. A perspective on epistasis: limits of models displaying no main effects. *Am J Hum Genet*, 70(461-471), 2002.

[8] T. S. Han. Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, 46:26–45, 1980.

[9] J. Hoh and J. Ott. Mathematical multi-locus approaches to localising complex human trait genes. *Nat Rev Genet*, 4:701–709, 2003.

[10] A. Jakulin. Machine learning based on attribute interactions. *Phd Thesis, Computer Science,University of Ljubljana, Ljubljana, Slovenia*, 2005.

[11] A. Jakulin and I. Bratko. Testing the significance of attribute interactions. *Proceedings of the Twenty-first International Conference on Machine Learning (ICML 2004)*, pages 409–416, Banff, Canada, 2004.

[12] W. Li and J. Reich. A complete enumeration and classification of two-locus disease models. *Hum Hered*, 50:334–339, 2000.

[13] Z. Liu and L. T. Multilocus ld measure and tagging snp selection with generalized mutual information. *Genetic Epidemiology*, 29:353–364, 2005.

[14] X. Y. Lou, G. B. Chen, L. Yan, J. Z. Ma, and J. Zhu. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am J Hum Genet*, 80:1125–1137, 2007.

[15] J. Marchini, P. Donnelly, and L. R. Cardon. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, 37:413–417, 2005.

[16] W. J. McGill. Multivariate information transmission. *Psychometrika*, 19:97–116, 1954.

[17] M. B. Miller, G. R. Lind, N. Li, and S. Y. Jang. Genetic analysis workshop 15: Simulation of a complex genetic model for rheumatoid arthritis in nuclear families including a dense snp map with linkage disequilibrium between marker loci and trait loci. *BMC Genetics*, 2007.

[18] J. H. Moore, J. C. Gilbert, C. T. Tsai, F. T. Chiang, and T. Holden. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol*, 241:252–261, 2006.

[19] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, and W. D. Dupont. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 69:138–147, 2001.

[20] N. A. Rosenberg, L. M. Li, R. Ward, and P. J. K. Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet*, 73:1402–1422, 2003.

[21] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

[22] J. Zhao, E. Boerwinkle, and M. Xiong. An entropy-based statistic for genomewide association studies. *Am J Hum Genet*, 77:27–40, 2005.