Published in final edited form as:

Stat Med. 2010 August 30; 29(19): 1998–2011. doi:10.1002/sim.3962.

Discovery of Complex Pathways from Observational Data

James Baurley, David V. Conti, James Gauderman, and Duncan Thomas

Department of Preventive Medicine, University of Southern California, Los Angeles, California

Abstract

Unraveling complex interactions has been a challenge in epidemiologic research. We introduce a pathway modeling framework that discovers plausible pathways from observational data, and allows estimation of both the net effect of the pathway and the types of interactions occurring among genetic or environmental risk factors. Each discovered pathway structure links combinations of observed variables through intermediate latent nodes to a final node, the outcome. Biologic knowledge can be readily applied in this framework as a prior on pathway structure to give preference to more biologically plausible models, thereby providing more precise estimation of Bayes factors for pathways of greatest interest by Markov Chain Monte Carlo (MCMC) methods.

Data was simulated for binary inputs of which only a subset was involved in different pathway topologies. Our algorithm was then used to recover the pathway from the simulated data. The posterior distributions of inputs, pair-wise and higher order interactions, and topologies were obtained by MCMC methods. The evidence in favor of a particular pathway or interaction was summarized using Bayes factors. Our method can correctly identify the risk factors and interactions involved in the simulated pathway. We apply our framework to an asthma case-control dataset with polymorphisms in 12 genes.

Introduction

The etiology of complex diseases may involve a network of biological interactions, genetic and environmental. Although these diseases may aggregate in families, there is no clear Mendelian mode of inheritance, and environmental exposures often play an important role in susceptibility [1]. Disease risk may be influenced by multiple factors including environment exposures, genetic factors, and different types of interactions. The polymorphisms involved in complex diseases may be common in a population, but only have small effects on protein function, thereby having only modest effects on disease risk [2].

Although many complex diseases are common, such as asthma, diabetes, and cancer, they are also the most challenging to investigate. Risk factors found to be associated with a complex disease in one study often fail to replicate in another. With the availability of high-throughput genotyping platforms for single nucleotide polymorphisms (SNPs), epidemiologists can thoroughly evaluate the genetic component of complex diseases either by pathway-driven approaches or systematic genome-wide association scans. These approaches are complimentary as new susceptibility genes discovered in a genome-wide scan can ultimately be included in a candidate gene study. As gene-gene and gene-environment interactions are known to play a crucial role in complex diseases, analysis

Software Implementation

The software implementation of this analysis framework is called ALPS (Algorithm to Learn Pathway Structure) and is developed in the C++ programming language. This software is available upon request. The current version is ALPS v0.8 available for 32 and 64 bit architectures.

methods that focus on pathways and incorporate prior biological knowledge can ultimately improve detection of true interactions [3].

Traditional analysis approaches such as logistic regression have long been used to test for main effects and simple pairwise interactions. Modern studies of complex diseases genotype a large panel of markers and may also collect an extensive exposure history. Even with large sample sizes, these studies lack the statistical power to detect high order interactions using traditional methods. Though computationally feasible, testing all possible pairwise or higher interactions would require adjustment for multiple comparisons. In the process, real effects may be eliminated along with spurious ones. Plausible interactions are sometimes ignored because traditional techniques break down as dimensionality increases.

Model uncertainty is central in the study of disease pathways. There are many circumstances, especially with high dimensional data, in which alternative models fit equally well. Because there is uncertainty about the true model, it is desirable for a method to return a set of plausible models rather than a single best model.

Stochastic approaches can incorporate uncertainty about the pathway mechanisms, structure, and intra-individual variation [4]. Additionally, prior knowledge can be used either to constrain the search space to be computationally feasible or to focus on biologically plausible regions. As analysis of complex pathways is becoming more common, pathway-based methods such as stochastic search variable selection (SSVS) [5], Monte Carlo logic regression [6], and Bayesian networks [7] are emerging.

We introduce a pathway modeling framework that stochastically discovers a class of plausible pathways from observational data. More specifically, our approach summarizes features of the posterior distribution of pathways rather than a single model. Prior knowledge in the form of a prior topology can be incorporated into the method to help guide the discovery algorithm. As it is a Markov Chain Monte Carlo (MCMC) approach, the posterior distribution of pairwise and higher-order interactions, as well as discovered topologies can be summarized. Additionally under each structure, the type of interaction can be estimated along with the net effect of the pathway on probability of disease.

Methods

Our framework for modeling pathways is shown in Figure 1. Let Y_i denote a phenotype (binary or continuous) for subject $i=1,\dots,I$ and $\mathbf{Z} = (Z_j)_{j=1,\dots,J}$ a vector of risk factors (genotypes and environmental exposures). We assume the two are related through a set of latent nodes (unobserved variables) $\mathbf{X} = (X_n)_{n=1,\dots,N}$ that have some joint conditional distribution given the observed vector $\mathbf{Z}$. This conditional distribution comprises a pathway structure, which we call the topology Λ , and a set of parameters specific to the topology θ_Λ . Additionally there may be prior biological knowledge about a pathway, and the assumed structure of the pathway is represented as Λ_0 , the “prior topology”. For computational simplicity we assume that Λ is a directed acyclic graph (DAG). Multiple pathways can be readily incorporated by adding nodes that represent the composite effect on the phenotype of each of the subpathways.

By integrating over the latent variables $\mathbf{X}$, we obtain a marginal probability distribution,

$$p(Y|\mathbf{Z};\Lambda,\theta_\Lambda) = \int p(Y|\mathbf{Z},\mathbf{X};\Lambda,\theta_\Lambda) d p(\mathbf{X}|\mathbf{Z};\Lambda,\theta_\Lambda) \quad (1)$$

However if we assume that the outcome Y is conditionally independent of the rest of the pathway given the last latent node of the topology X_N , the integrand can be rewritten as,

$$p(Y|Z, \mathbf{X}; \Lambda, \theta_\Lambda) = p(Y|X_N; \Lambda, \beta_\Lambda) p(X_N | \mathbf{X}_{(-N)}, Z; \Lambda, \theta_\Lambda) \quad (2)$$

where $\mathbf{X}_{(-N)}$ represents all the intermediate latent variables and β_Λ are regression coefficients specific to the topology. If the outcome Y is binary, the first conditional probability on the right side can be given by a logistic regression model of the form $\text{logit } \Pr(Y|X_N; \Lambda, \beta_\Lambda) = \beta_0 + \beta_1 X_N$ where β_1 represents the net effect of the pathway. Since we are assuming the topology is a DAG, the second term of equation 2 can be simplified by exploiting the conditional independencies in the topology, that is:

$$p(X_N | \mathbf{X}_{(-N)}, Z; \Lambda, \theta_\Lambda) = \prod_{n=1}^N p(X_n | \text{par}(X_n); \Lambda, \theta_\Lambda) \quad (3)$$

where $\text{par}(X_n)$ denotes the parents of the latent variable (either a measured risk factor Z or another latent variable). If the relationships between the latent nodes are deterministic in nature, we can compute $p(Y|Z; \Lambda, \theta_\Lambda)$ by their predicted values denoted $\tilde{\mathbf{X}}$. This type of simplification has been used in physiologically based pharmacokinetic (PBPK) models where intermediate metabolites are predicted using a system of deterministic toxicokinetic differential equations [8]. The predicted values of each latent variable are now determined by a system of equations. We parameterize the model in the form,

$$\tilde{X}_n = \theta_{n,1} \text{par}_1(X_n) + \theta_{n,2} \text{par}_2(X_n) + (1 - \theta_{n,1} - \theta_{n,2}) \text{par}_1(X_n) \text{par}_2(X_n) \quad (4)$$

where $\text{par}_1(X_n)$ and $\text{par}_2(X_n)$ return the values of the parents of X_n and $\theta_{n,1}$ and $\theta_{n,2}$ are parameters representing a range of interaction types. The predicted value $\tilde{X}_n$ can be continuous or binary. With binary parents, θ_n can emulate logical operators (Table 1), whereas for continuous inputs represent many different types of interactions. For an 'AND' type interaction ($\theta_1 = \theta_2 = 0$), the risk of disease increases with the environmental factor only when the genetic factor is present. An example of this type of interaction is given in Ottman 1990 where those lacking glucose-6-phosphate dehydrogenase (G6PD) and consumed fava beans develop hemolytic anemia [1,9]. For an additive model (ADD $\theta_1 = \theta_2 = 1/2$), the effect of the environmental risk factor is not altered by the genetic factor at all (i.e there is no interaction). Synergistic and antagonistic relationships can also be represented. For an 'OR' ($\theta_1 = \theta_2 = 1$), the environmental effect is suppressed by a genetic factor whereas in the 'XOR +OR' ($\theta_1 = \theta_2 = 2$), the genetic factor reduces the effect of the environmental exposure on Y .

Estimation

Pathway Topology Discovery Algorithm

We use a Markov Chain Monte Carlo (MCMC) method for fitting the model [10]. We call our method the Algorithm for Learning Pathway Structure (ALPS). ALPS involves the following steps, the details of which are provided in sections below and the supplemental material.

1. At each cycle, propose a change to the topology Λ to Λ' with probability $Q(\Lambda \rightarrow \Lambda')$. The change is either an addition or removal of a latent node.

2. Under the new topology, approximate the marginal likelihood $p(Y|\mathbf{Z};\Lambda')$ by integrating over θ_Λ numerically using MCMC or using the BIC (Bayesian Information Criterion) approximation [11].
3. Compute the new topology prior, $\pi(\Lambda')$.
4. Accept the new topology with Metropolis-Hastings probability

$$\min\left(1, \frac{\Pr(Y|\mathbf{Z};\Lambda')\pi(\Lambda')Q(\Lambda' \rightarrow \Lambda)}{\Pr(Y|\mathbf{Z};\Lambda)\pi(\Lambda)Q(\Lambda \rightarrow \Lambda')}\right)$$

5. If rejected, the previous topology is restored, along with the previous value of the parameters.
6. Repeat for U iterations.

Marginal Likelihood of a Topology

The marginal likelihood of a topology is obtained by integrating over θ_Λ ,

$$p(Y|\mathbf{Z};\Lambda) = \int p(Y|\mathbf{Z};\Lambda, \theta_\Lambda) p(\theta_\Lambda|\Lambda) d\theta_\Lambda$$

where $p(Y|\mathbf{Z};\Lambda, \theta_\Lambda)$ is the likelihood of θ_Λ for the topology Λ (equation 1) and $p(\theta_\Lambda|\Lambda)$ is the prior on θ_Λ . Though computationally intensive, we can compute the marginal likelihood numerically using a nested MCMC by averaging over the samples from the posterior distribution of θ_Λ . For larger problems, however, we recommend using the BIC approximation of the marginal likelihood and reserve the nested MCMC to summarize the posterior distribution of θ_Λ for the topologies of interest. The BIC approximation is given by

$$\log p(Y|\mathbf{Z}, \Lambda') \approx \log p(Y|\mathbf{Z}, \Lambda', \hat{\theta}_\Lambda) - 0.5 k \log(I)$$

where $p(Y|\mathbf{Z}, \Lambda', \hat{\theta}_\Lambda)$ is the maximized likelihood (ML), k is the number of parameters, and I is the sample size [11]. If there is a prior on θ_Λ the maximum a posteriori (MAP) can be used instead of the ML [12].

Prior on Pathway Structure

Recall that Λ_0 denotes a “prior topology”; some assumed form of the topology given by prior biological knowledge (or beliefs). We assume a prior on Λ of the form

$$\pi(\Lambda) = \exp(-\psi d(\Lambda, \Lambda_0) - Y \log(T(S_j))) / C(\psi, \Lambda_0) \text{ where } 0 \leq Y \leq 1$$

where $d(\Lambda, \Lambda_0)$ denotes a distance metric between the topologies Λ and Λ_0 , $\tau(S_j)$ is the size of S_j the set of all possible topologies with j risk factors, γ is a tuning parameter that adjusts the prior on topology size, $C(\psi, \Lambda_0)$ is an additive normalizing constant, and ψ is a decay parameter that adjusts the rate the Potts prior decreases as distance between topologies increases [13–16]. For the distance metric, we use an approximation to the Hamming distance — the minimum number of changes required to convert one topology to another [17]. If the proposed model is “further” away from the prior, the model is more likely to be rejected in the Metropolis-Hastings step. $C(\psi, \Lambda_0)$ is the sum of the numerator over all

possible topologies, which we estimated over a finite grid of ψ values using a sample of 1000 randomly permuted topologies. For greater generality, one could specify the prior in terms of a set of plausible topologies with corresponding prior probabilities. The parameter ψ is also updated by another Metropolis-Hastings step using a random walk proposal on ψ given the topology Λ .

Ignoring the Potts prior for the moment, the prior probability for a topology with j risk factors is proportional to $\tau^{-\gamma}(S_j)$ [16]. If $\gamma=0$, all topologies have equal prior probability regardless of the number of risk factors in Λ . When the number of variants J is large compared to the number of individuals I , we discourage complex topologies that may overfit the data in favor of parsimony by setting γ to either 0.5 or 1 [16].

Posterior Distribution of Pathway Structures

For all iterations after convergence, features of the pathway are recorded and later summarized. Let λ be a feature of Λ . λ could be the risk factors or some combination of risk factors contained in Λ , for example the risk factor 1 or an interaction (1,2) between risk factor 1 and 2. We defined pairwise interactions as internal nodes with both parents directly connected to inputs. We summarize the following: the topology Λ , risk factors contained in Λ , pairwise interactions, higher order interactions (>2 factors), the number of inputs and internal nodes, the distance $d(\Lambda|\Lambda_0)$, and the ψ parameter. Bayes factors are computed for features of the topology λ , giving the evidence in favor of models including that feature in Λ [18]. The Bayes factor is the ratio of posterior and prior odds,

$$BF(\lambda) = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{\Pr(\lambda|\mathbf{D})/(1 - \Pr(\lambda|\mathbf{D}))}{\Pr(\lambda)/(1 - \Pr(\lambda))}$$

where $\Pr(\lambda|\mathbf{D})$ is the posterior probability of including feature λ and $\Pr(\lambda)$ is the prior probability of including feature λ . The evidence in favor of λ is defined in terms of the Bayes factor: 1–3 Weak, 3–20 Positive, 20–150 Strong, and >150 Very Strong [11]. The prior odds are computed by a run in which the likelihood contribution is omitted from the Metropolis-Hastings step.

Simulation

To simulate pathway data, we specify a topology Λ containing N latent nodes and j risk factors. For each input, we generate binary indicators for each individual. For each internal node, we specify topology-specific parameters θ_Λ in order to simulate a type of interaction. We generate the predictive values $\tilde{X}$ for each intermediate node in order of the topology. The predicted value of the last node, $\tilde{X}_N$ and the predefined regression coefficients β_Λ are used to calculate the probability of an individual being a case; that is,

$$p_i = \Pr(Y_i=1|\beta_\Lambda, \tilde{X}_{N_i}) = \frac{\exp(\beta_0 + \beta_1 \tilde{X}_{N_i})}{1 + \exp(\beta_0 + \beta_1 \tilde{X}_{N_i})}$$

The outcome Y_i is then sampled as $Y_i \sim \text{Bernoulli}(p_i)$.

Data was simulated under different topologies and pathway-specific parameters. Logistic regression was utilized as a comparison method to find scenarios that were realistic for our approach. In the first stage of our investigation, data was simulated under the five interaction

types in Table 1 (ADD, AND, OR, XOR+OR, and MINUS), with pathway effects $\beta_1=0.5, 0.6, 0.3, 0.3,$ and 0.5 respectively to yield a statistically significant association between the simulated value of X_N and Y ($p \leq 0.01$). In addition to the true risk factors, an additional set of four inputs uninvolved in the pathway were included.

We then simulated more complicated structures with mixed interaction types to determine if ALPS could learn the topology and estimate the pathway-specific parameters. Data were simulated for two topologies with ten binary inputs, four which were risk factors for the disease Y and six which had no effect on the outcome (null). The topologies had different structures with a mixture of AND and OR node types (Figure 3). In scenario 1 OR(AND,AND), the pathway specific parameters were zero for the AND nodes, i.e. $\theta_{n,0} = \theta_{n,1} = 0$ and one for the OR nodes, i.e. $\theta_{n,0} = \theta_{n,1} = 1$ (Figure 3 top). For example, if either risk factor pair 1 and 2, or pair 3 and 4 were present, then the predicted value X_N would be one. In scenario 2, we simulated a different topology and arrangement of AND and ORs. (Figure 3 bottom). To investigate the scalability of our approach, the scenario 1 dataset was expanded to 100 and 1000 variants with the same four true risk factors nested within each. In the analysis of the 1000 variant dataset, a prior was placed on topology complexity by setting $\gamma=0.5$. We compared the time (in iterations) until risk factors and their interactions were included in the topology and the Bayes factors for features of Λ across the 10, 100, and 1000 variant runs.

We simulated 1,000 individuals under each scenario, each variant having a population prevalence of 0.25. The intercept β_0 was set at -0.9 , giving a baseline disease prevalence of 0.29 and the net effect of the pathway β_1 was set to 1.5 (odds ratio of 4.5). For scenario 1, there were 344 cases and 656 controls, while for scenario 2 there were 295 cases and 705 controls.

The odds ratios and 95% confidence intervals for the data simulated under these scenarios are displayed in Table 3. In a univariate analysis of the scenario 1 dataset, risk factors 1, 2, and 4 were statistically significantly associated with Y ($p < 0.05$). For the scenario 2 dataset only risk factor 3 and 4 were statistically significantly associated with Y .

For comparison, stepwise logistic regression (using a BIC penalty) was run on the simulated datasets. For scenario 1 considering main effects and up to four-way interactions, stepwise regression selected a model including only risk factors 1 and 4, each with corresponding p -values less than 0.001. Interestingly in the scenario 2 dataset stepwise regression selected the intercept only model.

Model Settings and Convergence

The ALPS software was executed for 103,000 topology updates, throwing out the first 3000 for the chain to converge (burn-in period). Visual investigation of time-series plots of the marginal likelihood from multiple chains was used to determine burn-in and total number of iterations. In the 100 and 1000 variant runs, the number of iterations was increased to more than 250,000 iterations. The algorithm was initialized to a random topology. An annealing process described in the supplemental material was applied, an initial temperature of $T=4$ was set at the beginning of the simulation and the temperature annealed according to a schedule to reach $T=1$ at the end of the burn-in period. The BIC approximation to the marginal likelihood was utilized. For the top topologies, a MCMC of 15,000 iterations (5,000 burn-in) was used to summarize the posterior distribution of θ_Λ . Traces from multiple chains indicated convergence and adequate mixing.

Results

The results for the simple two-way structures under different interaction scenarios are shown in Table 2. The Bayes factors showed evidence in favor of including the true risk factors and interaction in all scenarios. In general including a prior topology yielded larger Bayes factors for the risk factors and interactions, except for (1,2) in the OR scenario where the Bayes factor was slightly lower. Using Raftery's characterization, there was "very strong" evidence for including (1,2) in the XOR+OR scenario, "strong" for ADD, and "positive" evidence for AND, OR, and MINUS [11].

Recall that scenario 1 and 2 were mixtures of AND and OR interaction types (Figure 3). The 103,000 iterations completed in 5.7 hours using a 2.3 GHz processor. Variants not in Λ_0 had a prior probability of approximately 18% while inputs contained in Λ_0 had slightly higher prior probabilities (23% on average in scenario 1 for all four inputs and 29% for inputs 1 and 2 in scenario 2). For scenario 1, there was strong evidence that input 1 and 2 were involved in the pathway with Bayes factors 100 and 56 respectively. For scenario 2 there was very strong evidence in favor of models including risk factors 3 and 4 with corresponding Bayes factors approximately 450 and 160 respectively (Table 3).

Table 4 summarizes the pairwise interactions discovered in scenario 1 and 2. For scenario 1 ALPS identified (1,2) and (3,4) with Bayes factors 140 and 1.3 respectively. In scenario 2, there was very strong evidence for including pair (3, 4) with a posterior probability of 96% and a Bayes factor of 1,300. All possible higher-order interactions among the true risk factors are tabulated in Table 5. The prior probabilities for topologies representing higher order interactions were very small. For instance (1,2,3) and (1,2,3,4) in scenario 1 had prior probabilities of $6.5E-5$ and $1.5E-6$ respectively. In the scenario 1 dataset, ALPS discovered the true four way interaction (1,2,3,4) with a posterior of 2.4% and very strong evidence in favor of the corresponding topologies (Bayes factor $1.7E+4$). In scenario 2 higher order interactions were included rarely, though there was positive evidence for including (1,3,4) and (2,3,4), Bayes factors 31 and 21 respectively. Although the true 4-way interaction was not found when we simulated data at $\beta_1=1.5$, when we increased β_1 to 2.0, the interaction (1,2,3,4) had a posterior probability of 41%.

The posterior distribution of topologies was summarized. For scenario 1, 95% of the samples had two inputs, and 5% had three or four inputs. The Hamming distance to the prior topology ranged from 0–6, with 95% of the topologies having a distance of 2 or 3 and 2% having a distance of zero (equivalent to Λ_0). For scenario 2, 99% had two inputs and 1% had three or four inputs. Most topologies had a Hamming distance of 4 or 5 (99%). The top five topologies for each scenario are shown in Figure 2. For scenario 1 the topology with the largest posterior probability involved risk factors 1 and 2 (90%, Bayes factor: 95). The true topology had a posterior probability of 2% with a very large Bayes factor of 29,000. In scenario 2, a topology involving risk factors 3 and 4 had a posterior probability of 96% and another involving risk factors 1 and 3 had a posterior probability of 2%. Topologies with more than two inputs had posterior probabilities less than 1%. In scenario 2 the true topology was not recovered in the $\beta_1=1.5$ dataset but had a posterior probability of 41% in the $\beta_1=2$ dataset.

In Figure 3 the posterior means and standard deviation for the pathway-specific parameters are given for the true topologies. The estimates are close to the simulated values in general. For scenario 1, the parameters θ_1 and θ_2 had posterior means close to zero (logical AND relationships) and θ_3 had a posterior mean close to one (logical OR). The estimate of the net effect of the pathway β_1 was 1.11 (odds ratio 3.03). For scenario 2, the posterior means of θ_1 and θ_3 were close to zero representing AND's, and for θ_2 there was an OR or XOR+OR

type of interaction. The pathway effect β_1 appeared to be underestimated with an odds ratio of 2.3.

Averaging over all the visited topologies the estimates of β for scenario 1, were close to the simulated values: β_0 had a mean of -0.89 and standard deviation of 0.59 , and β_1 had a mean of 1.4 and standard deviation of 0.59 . The decay parameter ψ had a mean of 0.93 indicating the prior decreased rather rapidly with increasing distance. In the scenario 2 dataset, the estimates of β_0 had a mean of -1.01 and standard deviation of 0.04 (very close to the simulated value), and the β_1 estimates had a mean of 1.03 and standard deviation of 0.64 . The posterior mean of ψ was 0.62 , indicating the topology prior had less impact than in scenario 1.

Further simulations (not shown) demonstrated that with increased sample size, the posterior probability and Bayes factors for the true topologies increased and the posterior variance of the pathway-specific parameter estimates decreased. In small sample sizes with small pathway effects, ALPS yielded topologies with simple structures representing subsets of interactions found in the simulated pathway. ALPS also performed as expected under the null, estimating β_1 close to zero over all the sampled topologies. It is worth noting that under the null, ALPS samples many simple topologies but the net pathway effect is estimated to be zero (odds ratio of one).

To investigate the impact of the prior on topologies, scenario 1 was run without the prior topology. The prior probabilities of including features of Λ_0 were decreased. For example, the interaction (1,2) had a prior probability of 0.02 instead of 0.09 , a posterior probability of 0.84 rather than 0.93 , and a Bayes factor of 250 instead of 140 . The sensitivity to the prior was related both to the sample size and scale of the problem. In small sample situations the prior on topology prevented false positives by reducing the posterior probability of such interactions. In situations with many inputs, the prior has an impact on the ranking of posterior probability, moving biologically plausible topologies towards the top. Our method appears robust to misspecified Λ_0 . Again utilizing the scenario 1 dataset, the prior was incorrectly placed on four inputs that had no involvement in the pathway. For the interaction (1,2), the prior and posterior probability decreased (0.02 and 0.88 respectively). We also swapped the scenario 1 prior with that of scenario 2. For (1,2), the posterior probability increased to 0.99 , but for (3,4) the posterior probability decreased from $4.9\text{E}-2$ to $4.3\text{E}-4$.

The scalability of our method was assessed by expanding the scenario 1 dataset to reach 100 and 1,000 variants, the same four risk factors being involved in each case. For these datasets, ALPS performed three to four topology updates per second on a 2.3 GHz processor. The Bayes factors and iterations required to find these risk factors and pairwise interactions are given in Table 6. With 10 variants, risk factor 1 and 2 and their interaction had strong Bayes factors (100 , 56 , and 64 respectively) and were discovered rapidly, i.e they were already in Λ by the end of burn-in. With 100 variants these features and risk factor 3 were again contained in Λ at the end of burn-in. The Bayes factors for 1, 2, and (1,2) were much larger than the 10 variant case: 1600 , 800 , and 15000 respectively. The posterior probability for (1,4) remained about the same but the Bayes factor increased to 160 . The number of iterations to discovery increased with 1,000 variants (Table 6). Risk factor 1 was included in the topology after risk factors 2, 3, and 4. The Bayes factors for 1, 2, and (1,2) were 380 , 210 , and $3.0\text{E}+5$ respectively. The remaining interactions had very small prior and posterior probabilities resulting in large and unstable Bayes factors. As the number of variants analyzed increases there was a corresponding increase in the computational burden in terms of the number of interactions required to discovery topologies with true variants. However, across the scenarios investigated, ALPS was able to identify the same risk factors and interactions.

Application to Asthma Case-Control Study

Asthma is a respiratory disease influenced by both genetic and environmental factors. A case/control dataset was selected from the Children's Health Study (CHS) [19] to investigate twelve candidate genes believed to be involved in the oxidative stress and inflammatory pathways [20,21]. The dataset contained 642 children, 321 with physician-diagnosed asthma at study entry. Using the focused interaction testing framework (FITF), Millstein et al. identified a three-way interaction between *NQO1*, *MPO*, and *CAT* [21]. Odds ratios and 95% confidence intervals were computed for the association of each gene with asthma (Table 7). *NQO1* and *MPO* were statistically significantly associated with asthma ($p < 0.05$).

ALPS was applied to this dataset. The model was run for 103,000 topology updates with a burn-in of 3,000. The annealing process started at a temperature of 4 and the sampling temperature was set at $T=1$. In a first run, no prior topology was specified. In a second run, we constructed a prior topology from the Gene Ontology (GO) database [22]. A Pearson correlation distance matrix was constructed from the 448 GO terms associated with the 12 genes. Afterwards a hierarchical clustering algorithm was applied to produce the prior topology (Figure 4).

Traces indicated adequate mixing and convergence. The prior and posterior probabilities were used to compute Bayes factors (Table 8). For main effects, there was strong evidence for including the *NQO1* and *MPO* genes (Bayes factors 50 and 26 respectively). When a GO prior was applied, the prior and posterior probabilities for including *NQO1* decreased yielding a lower Bayes factor (36). For the genes *MPO* and *CAT*, however, the GO prior increased the Bayes factors (26 to 30 for *MPO* and 0.97 to 1.3 for *CAT*).

The top pairwise and three-way interactions are shown in Table 8. There was evidence that pairwise interaction (*NQO1*, *MPO*) was involved in the pathway (Bayes factor 200 with no prior and 170 with the GO prior). There was some evidence in favor of models with other risk factors interacting with either *NQO1* or *MPO*, such as *CAT* and *GSTM1*. With the GO prior, the prior probability for the pair (*CAT*, *MPO*) increased from 1% to 4% but the Bayes factor was relatively unchanged. As seen in Figure 4, this pair is closely related by shared GO terms.

Three-way interactions occurred much less frequently but had substantial Bayes factors. The (*CAT*, *NQO1*, *MPO*) interaction found by Millstein, et al. had a Bayes factor of 340 with no prior topology. With the GO prior, the data supported the prior and increased the Bayes factor to 530. There were other notable Bayes factors sharing *NQO1* and *MPO* in common, reflecting the uncertainty in the discovered topology. The majority of the topology sampled had two inputs (96%). The Hamming distance to the prior topology ranged between 9 and 15 indicating the GO structure was never proposed and the proposed topologies had few inputs. The posterior mean of ψ was 0.56 indicating the prior drops off rapidly with distance.

The topology of the top three-way interaction had a posterior probability of 0.03 and is shown in Figure 5. This interaction has essentially the same structure as that identified by the FITF method. The posterior means and standard deviation were computed. The θ parameters for node 1 connecting to *MPO* and *CAT* appear to have an AND mode of interaction, that is a change in risk when both are present. This sub-pathway and *NQO1* have a behavior similar to a logical OR. These two sub-pathways lead to a protective effect (odds ratio 0.46).

Discussion

We demonstrated that our ALPS method is capable of discovering potentially complex pathway structures from observational data and that this approach scales well to large datasets. Topology uncertainty was accounted for by sampling from the posterior distribution of topologies. Instead of obtaining a single “best” model as in stepwise regression, ALPS discovers a set of plausible models that can be ranked by their posterior probabilities or Bayes factors. With mixed simulated OR and AND parameters, our approach identified key features of the simulated nodes.

Our approach includes a natural way to incorporate prior knowledge about the structure of a pathway as seen in the asthma application. A prior topology represents biological relationships between proteins, genes, RNA, and metabolites. These biological relationships could be conceptualized and modeled using systems biology, perhaps using a unified language like Systems Biology Markup Language (SBML) to schematically represent the current state of knowledge [23]. Databases exist for cataloging and annotating -omic data such as KEGG and Gene Ontology [22,24]. Although systems biology models represent a simplification of a biological pathway, their abstract properties can be readily utilized as priors on topologies and pathway-specific parameters.

In our framework, the pathway structure constrains the number of free parameters. Consider a logistic regression model with J variants and all possible interactions. The full model would have 2^J coefficients. If we add the requirement that the pathway adheres to a graph structure where each child has two inputs, the number of nodes becomes $N=J-1$ and the number of coefficients is reduced to $4(J-1)$. Utilizing the parameterization proposed in our framework with N denoting the number of internal nodes, there are $2+2N$ parameters. This allows our method to explore a large space of pathways with relatively few parameters. For example a pathway with four risk factors and three latent nodes has only eight parameters.

There are a number of other approaches that aim to reduce the number of parameters [25]. One such method is stochastic search variable selection (SSVS) [5]. The space of possible regression models is searched by including an indicator random variable for whether a term is included in the model [26]. Pathway structure knowledge can be integrated into priors on the indicator variables. This constrains the search space to adhere to properties of a pathway and informs about the relative importance of the priors. Like our approach, this method is capable of finding higher-order interactions and returns a set of plausible models. This method, however, does not estimate the type of interaction.

Logic regression is a method that searches for Boolean combinations of variants [27]. A limitation of logic regression is that it identifies a single model rather than a set of plausible models. Monte Carlo logic regression overcomes this by using MCMC to identify a set of models that may be associated with the outcome and then summarizing features of those models [6]. Similar to our approach, they summarized the posterior distribution of risk factors and interactions. Logic regression is limited to binary input and intermediate nodes whereas our approach works with continuous variables as well. Additionally our pathway specific θ 's are continuous and thus can represent a much wider range of interaction types. There is currently no prior on the tree structures in MCMC logic regression. For comparison, MCMC logic regression was applied to the 10 variant scenario 1 and 2 datasets. MCMC logic regression was run for 103,000 iterations (3000 as burn-in). For scenario 1, the prior probability of including (1,2) was 1.7% and the posterior probability was 90% (Bayes factor: 550). For scenario 2, predictor 3 and 4 occurred together 84% of the iterations after burn-in (Bayes factor: 330). Both ALPS and MCMC logic regression discovered similar features in the scenario 1 and scenario 2 datasets.

Bayesian networks have been widely used for exploring relationships between variables in *in silico* experiments and analysis of gene expression properties and patterns [7,28,29]. A Bayesian network consists of a directed acyclic graph (DAG) representing dependencies between the random variables in a pathway and a joint probability distribution for these random variables. Bayesian networks are similar to our approach in methods for parameter estimation, structural learning, and natural incorporation of prior knowledge. However, Bayesian network learning is usually restricted to finding relationships between risk factors, *observed* intermediates, and the outcome. The absence of unobserved intermediates hints at an incomplete pathway representation. Although one can add a hidden node to a Bayesian network, computing the joint probability distribution would require integrating over all possible values for each such hidden node. For directed acyclic graphs (DAG) this integration can be implemented using a peeling algorithm [30].

There are a number of exploratory techniques that aim to identify interactions though they are not pathway-based. Multifactor dimensionality reduction (MDR) is an approach to identify factors that are associated with the outcome. MDR constructs a new variable that is a combination of several risk factors and the new variable is evaluated for its ability to classify the outcome [31]. Classification and regression tree (CART) is a binary partitioning method that builds a tree by repeatedly stratifying data into risk groups based on how well each variable predicts the outcome [32]. Although CART can represent interactions graphically, it does not represent biological pathways in an intuitive way. In CART, the split closest to the root of the tree represents the best predictor of the outcome and the following splits represent interactions conditional on previous splits. In our approach the root represents the net effect of the pathway and types of interactions are represented by latent nodes. Though useful in exploratory analysis, these methods are neither pathway-driven nor provide a posterior distribution of plausible models. Further, these methods do not estimate the prior probability of finding a feature, so it is difficult to evaluate if the models and effects are false discoveries.

The asthma application demonstrates the ability of ALPS to identify a set of pathways for follow-up. These topologies were undetectable by stepwise regression. In fact, considering main effects and up to four way interactions, only *NQOI* was selected in the model using stepwise regression. Topology priors based on biological knowledge, such as the Gene Ontology database in the asthma example, can alter the prior probability of a pathway and thus alter the ordering of posterior probabilities and Bayes factors. In the case of the (*CAT*, *NQOI*, *MPO*) topology, the prior and posterior probability increased when a GO topology prior was utilized. This follows by examining the clustering in Figure 4. Genes sharing branches have more shared GO terms and a closer relationship in biology. We acknowledge that replication of findings across populations and analytic technique is very important in the study of complex diseases. Millstein et. al. showed evidence of replication of the (*CAT*, *NQOI*, *MPO*) interaction in an independent population [21].

In genome-wide association studies (GWAS), the number of potential genetic factors increases substantially. For a complex disease, there may be many candidate genes identified, but also genes that are not known. Future research is needed to determine the feasibility of scaling our approach to these much larger problems. This includes taking advantage of high performance computing and parallel programming, partitioning the inputs into a hierarchical structure, and possibly applying a screening step to reduce the number of variants. Also manually extracting prior information on such large pathways could be impractical. Technologies to automatically extract pathway priors from databases are another area of future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grants U01ES15090, P30ES07048, R01HL087680, P50CA084735, R01HL061768, P01ES011627, and U01DA020830.

References

1. Ottman R. Gene-environment interaction: definitions and study designs. *Prev Med.* 1996; 25(6): 764–70. [PubMed: 8936580]
2. Thomas PD, Kejariwal A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc Natl Acad Sci U S A.* 2004; 101(43):15398–403. [PubMed: 15492219]
3. Thomas DC. The Need for a systematic approach to complex pathways in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev.* 2005; 14(3)
4. Conti DV, et al. Bayesian modeling of complex metabolic pathways. *Hum Hered.* 2003; 56(1–3): 83–93. [PubMed: 14614242]
5. Conti D, et al. Using ontologies in hierarchical modeling of genes and exposures in biological pathways. Phenotypes and Endophenotypes: Foundations for Genetic Studies of Nicotine Use and Dependence. 2009. NCI Tobacco Control Monograph Series. :539–584.
6. Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol.* 2005; 28(2):157–70. [PubMed: 15532037]
7. Friedman N, et al. Using Bayesian networks to analyze expression data. *J Comput Biol.* 2000; 7(3–4):601–20. [PubMed: 11108481]
8. Cortessis V, Thomas DC. Toxicokinetic genetics: an approach to gene-environment and gene-gene interactions in complex metabolic pathways. *IARC Sci Publ.* 2004; (157):127–50. [PubMed: 15055294]
9. Ottman R. An epidemiologic approach to gene-environment interaction. *Genet Epidemiol.* 1990; 7(3):177–85. [PubMed: 2369997]
10. Robert, CP.; Casella, G. Monte Carlo Statistical Methods. 2. New York: Springer; 2004.
11. Raftery AE. Bayesian Model Selection in Social Research. *Sociological Methodology.* 1995; 25:111–163.
12. Sparacino G, Tomblato C, Cobelli C. Maximum-likelihood versus maximum a posteriori parameter estimation of physiological system models: the C-peptide impulse response case study. *IEEE Trans Biomed Eng.* 2000; 47(6):801–11. [PubMed: 10833855]
13. Thomas DC, et al. Bayesian spatial modeling of haplotype associations. *Hum Hered.* 2003; 56(1–3):32–40. [PubMed: 14614236]
14. Descombes X, et al. Estimation of Markov Random Field Prior Parameters Using Markov Chain Monte Carlo Maximum Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 1999; 8(7):954–964.
15. Green PJ, Richardson S. Hidden Markov Models and Disease Mapping. *Journal of the American Statistical Association.* 2002; 97(460):1055–1070.
16. Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika.* 2008; (95):759–771.
17. Hamming R. Error Detecting and Error Correcting Codes. *Bell System Technical Journal.* 1950; 26(2):147–160.
18. Gelman, A., et al. Texts in Statistical Science. 2. Chapman & Hall/CRC; 2004. Bayesian Data Analysis.
19. Peters JM, et al. A study of twelve Southern California communities with differing levels and types of air pollution. I. Prevalence of respiratory morbidity. *Am J Respir Crit Care Med.* 1999; 159(3): 760–7. [PubMed: 10051248]

20. Gilliland FD, et al. A theoretical basis for investigating ambient air pollution and children's respiratory health. *Environ Health Perspect.* 1999; 107(Suppl 3):403–7. [PubMed: 10346989]
21. Millstein J, et al. A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet.* 2006; 78(1):15–27. [PubMed: 16385446]
22. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000; 25(1):25–9. [PubMed: 10802651]
23. Thomas PD, Mi H, Lewis S. Ontology annotation: mapping genomic regions to biological function. *Curr Opin Chem Biol.* 2007; 11(1):4–11. [PubMed: 17208035]
24. Kanehisa M, et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 2006; 34(Database issue):D354–7. [PubMed: 16381885]
25. O'Hara RB, Sillanpaa MJ. A Review of Bayesian Variable Selection Methods: What, How, and Which. *Bayesian Analysis.* 2009; 4(1):85–118.
26. Conti DV, Gauderman WJ. SNPs, haplotypes, and model selection in a candidate gene region: the SIMPL analysis for multilocus data. *Genet Epidemiol.* 2004; 27(4):429–41. [PubMed: 15543635]
27. Ruczinski I, Kooperberg C, LeBlanc M. Exploring interactions in high-dimensional genomic data: an overview of Logic Regression with Applications. *Journal of Multivariate Analysis.* 2004; 90:178–195.
28. Yu J, et al. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics.* 2004; 20(18):3594–603. [PubMed: 15284094]
29. Khalil IG, Hill C. Systems biology for cancer. *Curr Opin Oncol.* 2005; 17(1):44–8. [PubMed: 15608512]
30. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 1981; 17(6):368–76. [PubMed: 7288891]
31. Moore JH, et al. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol.* 2006; 241(2):252–61. [PubMed: 16457852]
32. Cook NR, Zee RY, Ridker PM. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat Med.* 2004; 23(9):1439–53. [PubMed: 15116352]

Figure 1. Graphical representation of the model. The risk factors Z and the outcome Y are observed. The X 's are latent nodes determined by topology specific parameters θ_Λ and the topology Λ .

Figure 2.
Top 5 discovered topologies by posterior probability.

Figure 3. These two different topologies are constructed of AND and OR node types. Scenario 1 OR(AND,AND) in the upper panel and AND(OR(AND)) in the lower panel. Each edge is labeled with the corresponding simulated value in brackets and the posterior mean and standard deviation for θ_{Λ} and β_{Λ} .

Figure 4.
Cluster dendrogram of asthma genes from Gene Ontology.

Figure 5.
Discovered asthma pathway involving three genes.

Figure 6. Topology moves. From left to right, a node is removed deleting the edge to input 3. A new node is then added connecting input 1 and 2.

Table 1

Truth table for predicted value of X_N given binary parent inputs under different θ_V parameters.

$X_{pa1(n)}$	$X_{pa2(n)}$	$\theta_{n,0} = \theta_{n,1} = 0.5$ ADDITIVE	$\theta_{n,0} = \theta_{n,1} = 0$ AND	$\theta_{n,0} = \theta_{n,1} = 1$ OR	$\theta_{n,0} = \theta_{n,1} = 2$ XOR + OR	$-\beta_1, \theta_{n,0} = \theta_{n,1} = 1$ MINUS
1	1	1	1	1	1	-1
1	0	0.5	0	1	2	-1
0	1	0.5	0	1	2	-1
0	0	0	0	0	0	0

Table 2

Prior and posterior inclusion probabilities and Bayes factors for the true risk factors and their interaction under different simulated, θ_{λ} with a prior topology (left) and without a prior topology (right).

			Prior Topology			No Prior Topology		
Interaction Type	β_1	Term	$\Pr(\lambda)$	$\Pr(\lambda D)$	Bayes Factor	$\Pr(\lambda)$	$\Pr(\lambda D)$	Bayes Factor
ADD	0.5	1	0.47	0.93	1.6E+01	0.34	0.66	3.7E+00
		2	0.50	0.93	1.4E+01	0.32	0.60	3.2E+00
		(1,2)	0.28	0.89	2.2E+01	0.06	0.34	8.6E+00
AND	0.6	1	0.47	0.92	1.2E+01	0.34	0.83	9.7E+00
		2	0.50	0.87	6.7E+00	0.32	0.50	2.1E+00
		(1,2)	0.28	0.80	1.1E+01	0.06	0.35	9.1E+00
OR	0.3	1	0.47	0.76	3.6E+00	0.34	0.53	2.2E+00
		2	0.50	0.83	4.9E+00	0.32	0.64	3.7E+00
		(1,2)	0.28	0.73	7.0E+00	0.06	0.34	8.6E+00
XOR+OR	0.3	1	0.47	0.99	1.0E+02	0.34	0.94	2.9E+01
		2	0.50	1.00	1.1E+03	0.32	1.00	5.0E+02
		(1,2)	0.28	0.98	1.6E+02	0.06	0.90	1.5E+02
MINUS	0.5	1	0.47	1.00	8.9E+02	0.34	0.98	8.2E+01
		2	0.50	0.93	1.3E+01	0.32	0.68	4.4E+00
		(1,2)	0.28	0.93	3.3E+01	0.06	0.66	3.2E+01

Table 3

Odds ratios and 95% confidence intervals for scenario 1 OR(AND, AND) and 2 AND(OR(AND)) simulated data $I=1000$. For both scenarios, the net effect of the pathway was $\beta_1=1.5$. Inputs 1–4 were risk factors and 5–10 had no effect on the pathway. The Bayes factor is the ratio of posterior to prior odds.

Inputs	Node	OR(AND,AND)			$\beta_1=1.5$			AND(OR(AND))			$\beta_1=1.5$		
		Logistic regression		Node	ALPS		Node	Logistic regression		Node	ALPS		Node
		OR	(95% CI)		Pr(λ)	Pr(λ D)		OR	(95% CI)		Pr(λ)	Pr(λ D)	
Risk Factor	1	1.64	(1.23, 2.20)		0.24	0.97		1.25	(0.92, 1.72)		0.28	0.02	
	2	1.46	(1.09, 1.95)		0.25	0.95		1.04	(0.75, 1.44)		0.30	0.00	
	3	1.09	(0.81, 1.47)		0.22	0.05		1.53	(1.12, 2.10)		0.19	0.99	
	4	1.64	(1.22, 2.20)		0.23	0.09		1.45	(1.07, 1.96)		0.17	0.97	
Null	5	0.98	(0.72, 1.32)		0.18	0.00		0.89	(0.65, 1.22)		0.18	0.01	
	6	1.06	(0.79, 1.43)		0.18	0.00		0.97	(0.71, 1.32)		0.17	0.00	
	7	1.05	(0.78, 1.42)		0.18	0.00		0.90	(0.66, 1.21)		0.17	0.00	
	8	0.81	(0.59, 1.11)		0.19	0.00		1.15	(0.85, 1.56)		0.19	0.01	
	9	0.92	(0.68, 1.25)		0.17	0.00		0.95	(0.70, 1.30)		0.18	0.00	
	10	1.19	(0.88, 1.61)		0.18	0.00		0.91	(0.68, 1.25)		0.18	0.00	

Table 4

Prior and posterior probability and Bayes factors for pairwise interactions.

	OR(AND,*AND)			AND(OR(AND))		
Pairwise Interaction	Pr(λ)	Pr(λ D)	Bayes Factor	Pr(λ)	Pr(λ D)	Bayes Factor
(1,2)	0.09	0.93	1.4E+02	0.12	0.00	2.6E-03
(1,3)	0.02	0.00	5.2E-02	0.02	0.02	8.2E-01
(1,4)	0.02	0.02	1.3E+00	0.02	0.00	2.0E-01
(2,3)	0.02	0.00	5.2E-02	0.02	0.00	9.1E-02
(2,4)	0.02	0.01	3.1E-01	0.02	0.00	8.8E-03
(3,4)	0.06	0.05	8.3E-01	0.02	0.96	1.3E+03

Table 5

Scenario 1 and 2 higher order interactions (>2 way).

Interaction	OR(AND,AND)			AND(OR(AND))		
	Pr(λ)	Pr(λ D)	Bayes Factor	Pr(λ)	Pr(λ D)	Bayes Factor
(1,2,3)	6.5E-05	1.1E-03	1.8E+01	4.5E-04	3.7E-04	8.3E-01
(1,2,4)	6.3E-05	1.6E-02	2.6E+02	7.0E-05	0.0E+00	0.0E+00
(1,3,4)	5.9E-05	1.3E-03	2.2E+01	2.7E-05	8.3E-04	3.1E+01
(2,3,4)	5.7E-05	7.0E-03	1.2E+02	3.2E-05	6.6E-04	2.1E+01
(1,2,3,4)	1.5E-06	2.4E-02	1.7E+04	2.2E-06	0.0E+00	0.0E+00

Comparison of iterations to discovery, prior and posterior probabilities, and Bayes factors for risk factors and pairwise interactions across the 10, 100, and 1000 variant datasets.

Table 6

	10 Variants						100 Variants						1000 Variants					
	Input	Iterations to Discovery	Pr(λ)	Pr(λ D)	Bayes Factor	Iterations to Discovery	Pr(λ)	Pr(λ D)	Bayes Factor	Iterations to Discovery	Pr(λ)	Pr(λ D)	Bayes Factor	Iterations to Discovery	Pr(λ)	Pr(λ D)	Bayes Factor	
Risk factor	1	1	2.4E-01	9.7E-01	1.0E+02	1	2.3E-02	9.7E-01	1.6E+03	11680	1.8E-03	4.0E-01	3.8E+02					
	2	1	2.5E-01	9.5E-01	5.6E+01	1	2.3E-02	9.5E-01	8.0E+02	834	2.1E-03	3.2E-01	2.1E+02					
	3	530	2.2E-01	5.2E-02	2.0E-01	1	2.4E-02	4.8E-03	2.0E-01	596	2.5E-03	1.1E-02	4.5E+00					
	4	489	2.3E-01	9.4E-02	3.6E-01	1761	2.2E-02	3.9E-02	1.8E+00	457	1.9E-03	9.6E-02	5.7E+01					
Pairwise Interactions	(1,2)	1	1.8E-01	9.3E-01	6.4E+01	1	7.2E-04	9.2E-01	1.5E+04	45104	1.4E-06	3.0E-01	3.0E+05					
	(1,3)	608	1.8E-01	1.1E-03	5.0E-03	1902	3.6E-04	5.4E-04	1.5E+00	Not found	1.8E-05	0.0E+00	0.0E+00					
	(1,4)	494	1.8E-01	2.4E-02	1.1E-01	986	8.5E-05	1.3E-02	1.6E+02	11680	<2.0E-07	8.3E-03	>4.2E+04					
	(2,3)	7388	1.9E-01	1.1E-03	4.9E-03	10522	1.3E-04	6.8E-04	5.3E+00	128943	<2.0E-07	3.6E-05	>1.8E+02					
	(2,4)	489	1.7E-01	6.9E-03	3.4E-02	1258	1.7E-04	4.5E-03	2.6E+01	834	<2.0E-07	2.3E-03	>1.1E+04					
	(3,4)	530	1.8E-01	4.9E-02	2.4E-01	1763	1.2E-03	2.9E-03	2.4E+00	595	8.0E-07	8.0E-03	1.0E+04					

Table 7

Odds ratios and 95% confidence intervals from logistic regression, prior and posterior probabilities, and Bayes factors from ALPS for the asthma case-control dataset.

Gene	Logistic regression			No Prior Topology			Prior Topology (GO)		
	OR	(95% CI)	Pr(λ)	Pr(λ D)	Bayes Factor	Pr(λ)	Pr(λ D)	Bayes Factor	
<i>GSTM1</i>	1.28	(0.94, 1.75)	0.16	0.01	3.5E-02	0.19	0.01	6.0E-02	
<i>GSTT1</i>	0.84	(0.58, 1.24)	0.16	0.04	2.3E-01	0.19	0.04	1.7E-01	
<i>CAT</i>	0.83	(0.63, 1.10)	0.17	0.16	9.7E-01	0.17	0.21	1.3E+00	
<i>SOD2</i>	0.90	(0.73, 1.12)	0.17	0.01	4.1E-02	0.15	0.01	6.1E-02	
<i>ICAM1 (241)</i>	0.83	(0.59, 1.16)	0.17	0.04	2.2E-01	0.17	0.02	7.7E-02	
<i>ICAM1 (469)</i>	0.90	(0.72, 1.12)	0.16	0.01	4.7E-02	0.18	0.02	7.7E-02	
<i>TGFB1</i>	0.97	(0.77, 1.21)	0.17	0.01	4.0E-02	0.15	0.00	1.1E-02	
<i>TNF</i>	1.14	(0.84, 1.54)	0.17	0.00	1.4E-02	0.14	0.00	2.9E-02	
<i>NQO1</i>	0.67	(0.52, 0.88)	0.17	0.91	5.0E+01	0.14	0.85	3.6E+01	
<i>MPO</i>	0.74	(0.57, 0.97)	0.17	0.84	2.6E+01	0.17	0.86	3.0E+01	
<i>ADRB2 (16)</i>	1.09	(0.88, 1.36)	0.17	0.01	3.1E-02	0.18	0.02	9.1E-02	
<i>ADRB2 (27)</i>	1.01	(0.81, 1.25)	0.17	0.00	1.2E-02	0.17	0.00	5.2E-03	

Table 8

Pairwise and 3-way interactions discovered in the asthma case-control dataset

Top pairwise interactions by posterior probability						
No Prior Topology			Gene Ontology Prior			
Interaction	Pr(λ)	Pr(λ D)	Bayes Factor	Pr(λ)	Pr(λ D)	Bayes Factor
(<i>NQOI,MPO</i>)	0.01	0.74	2.0E+02	0.01	0.69	1.7E+02
(<i>CAT,MPO</i>)	0.01	0.07	4.9E+00	0.04	0.15	4.6E+00
(<i>CAT,NQOI</i>)	0.01	0.07	5.5E+00	0.01	0.06	5.5E+00
(<i>GSTM1,NQOI</i>)	0.02	0.02	1.4E+00	0.01	0.02	1.5E+00

Top 3-way interactions by posterior probability						
No Prior Topology			Gene Ontology Prior			
Interaction	Prior	Posterior	Bayes Factor	Prior	Posterior	Bayes Factor
(<i>CAT,NQOI,MPO</i>)	6.5E−05	2.2E−02	3.4E+02	5.8E−05	3.0E−02	5.3E+02
(<i>GSTT1,NQOI,MPO</i>)	8.0E−05	8.9E−03	1.1E+02	4.4E−05	3.7E−03	8.5E+01
(<i>SOD2,NQOI,MPO</i>)	7.1E−05	1.6E−03	2.2E+01	3.9E−05	1.4E−03	3.5E+01
(<i>ICAM1-469,NQOI,MPO</i>)	6.8E−05	1.8E−03	2.6E+01	4.2E−05	1.3E−03	3.0E+01
(<i>CAT,SOD2,MPO</i>)	7.1E−05	2.7E−04	3.8E+00	1.8E−04	7.2E−04	4.0E+00