Inference of Disease-specific Gene Interaction Network Using a Bayesian Network learned by Genetic Algorithm

Daye Jeong
Yonsei University
50 Yonsei-ro, Seodaemun-gu,
Seoul, Korea
+82 2 2123 7757
rabilish@cs.yonsei.ac.kr

Yunku Yeu
Yonsei University
50 Yonsei-ro, Seodaemun-gu,
Seoul, Korea
+82 2 2123 7757
yyk@cs.yonsei.ac.kr

Jaegyoon Ahn
UCLA
Los Angeles, California 90095,
USA
+1 310 825 3891
jgahn@ucla.edu

Youngmi Yoon
Gachon University
191 Hambakmoero, Yeonsu-gu
Incheon, Korea
+82 32 820 4393
ymyoon0719@gmail.com

Sanghyun Park
Yonsei University
50 Yonsei-ro, Seodaemun-gu,
Seoul, Korea
+82 2 2123 5714
sanghyun@cs.yonsei.ac.kr

ABSTRACT
An important goal of systems biology is to understand and identify mechanisms of the human body system. Genes play functional roles in the context of complex pathways. Analysis of genes as networks is therefore important to understand whole system mechanisms. Biological activities are governed by various signaling networks. The advent of high-throughput technologies has made it possible to obtain biological information on a genome-wide scale. Genetic interactions have been identified from high-throughput data such as microarray data using Bayesian networks.

In this paper, we infer the disease-specific gene interaction network using a Bayesian network, which is robust to noise in the data. We apply a genetic algorithm to learn the Bayesian network. We use heterogeneous data, including microarray, protein-protein interaction (PPI), and HumanNET data to learn and score the network. We also exploit single nucleotide polymorphism (SNP) data to infer disease-specific genetic interaction network. We included SNPs as this data may help detect weak signals related to genetic variation.

In this paper, we reconstruct interactions between pathway genes using our method. We confirm that our method has statistically significant reconstruction power by applying it to Type II diabetes data. Importantly, using Alzheimer disease data, we infer an unreported interaction between a SNP and a disease-related gene.

SAC'15, April 13-17, 2015, Salamanca, Spain.

ACM 978-1-4503-3196-8/15/04…$15.00
http://dx.doi.org/10.1145/2695664.2695944

Categories and Subject Descriptors
H.2.8 **[Database Management]**: Data mining; J.3 [**Life and medical sciences**]: Biology and Genetics

General Terms
Algorithms, Experimentation, Human Factors

Keywords
Disease-specific, Gene interaction, Bayesian network, genetic algorithm

1. INTRODUCTION
The human body is a complex system. Interactions among DNA, RNA, and proteins underlie this complex system. Identifying disease mechanisms in humans is a major goal of systems biology, as an understanding of these mechanisms would make it possible to understand and potentially prevent and/or treat diseases.

High-throughput technologies such as microarrays and RNA-seq allow parallel analysis of gene expression profiles. It is challenging to infer biologically meaningful results from these high-throughput data. Genes function in the context of complex pathways rather than in isolation. Therefore, it is critical to understand interactions between genes as a network. Such genetic interactions have been inferred from gene expression data by constructing Bayesian networks [6]. However, small sample sizes and noise in the data have hampered the extraction of meaningful results. Hence, it was proposed that not only gene expression data, but also biological knowledge data be used to construct Bayesian

networks [8]. However, previous Bayesian network analyses using "realistic" data, such as tumor vs. normal gene expression patterns, failed to provide useful insights [3].

In this paper, we infer disease-specific genetic interaction network using a Bayesian network learned by genetic algorithm, which simulates the survival of the fittest among individuals. A Bayesian network is a directed acyclic graph (DAG) with a conditional probability distribution for each node. It is one of the models that can be used to represent causal interactions between genes. Furthermore, it is an appropriate model to apply to noisy data like high-throughput data. Nodes in a Bayesian network represent genes, and edges represent genetic interactions between two genes. The structure of a Bayesian network implies conditional probability, and the distribution of the child node is determined by its parent nodes.

Our proposed approach has two advantages compared to current methods. First, we exploit single nucleotide polymorphism (SNP) data to incorporate genotypic variation. SNPs can influence promoter activity, mRNA stability, and the amino acid sequence, all of which may contribute to the activation and function of genes and proteins [17]. Even though not all SNPs affect the expression of genes, and not all disease-related gene expression is related to SNPs, including SNP information may help reduce the search space for disease-specific module inference and allow detection of weak signals related to SNPs.

Second, we construct a Bayesian network considering biological characteristics. We use a genetic algorithm, which is based on natural selection and evolution, to learn the Bayesian network. Because most gene expression data are noisy, we discretize gene expression data and merge multiple gene expression data though the Bayesian network to deal with uncertainty and avoid overfitting. Because gene expression data are not sufficient to infer a genetic interaction network, we use heterogeneous biological data including microarray, protein-protein interaction (PPI), and HumanNET [13] data as biological knowledge data to construct the Bayesian network. Furthermore, we score the network in a different way from that usually used to avoid overfitting.

To estimate the robustness of our approach, we reconstruct a network with nodes for a "known" disease pathway in KEGG [11]. Robustness is measured by the Pearson correlation coefficient (PCC) between the vector of the fitness scores of the each learned network and the vector of the ratios of correctly recovered edges in each network. To analyze real data, we construct a disease-specific genetic interaction network using a Bayesian network with nodes that include SNPs in genes and differentially expressed genes (DEGs).

We use the graphics package in R to implement the Bayesian network [16].

The remainder of this paper is organized as follows. In Section 2, we describe how we implemented a genetic algorithm to learn the Bayesian network. In Section 3, we apply our approach to reconstruct a KEGG pathway. In Section 4, we infer a disease-specific genetic interaction network using Alzheimer's disease data. Finally, in Section 5, we conclude by summarizing our work and discussing future directions.

2. METHODS

2.1 System overview

An overview of our process is provided in Figure 1 .

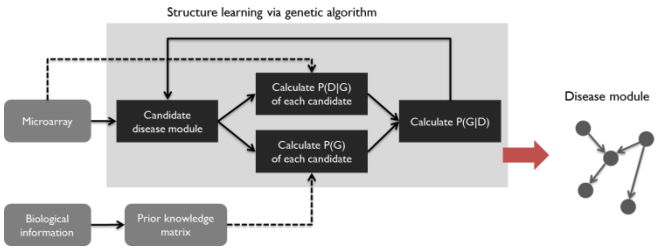

Figure 1. System overview

In this paper, we propose a novel method to infer the structure of disease-specific genetic interaction network using heterogeneous biological data. We selected a Bayesian network as a data model to represent causal interactions between genes. We learned the topology of the Bayesian network through a genetic algorithm, then evaluated the learned Bayesian network using various biological data sources, and generated a disease-specific genetic interaction network.

To evaluate the learned network, we calculated the fitness of the network. The fitness function is based on the Bayes theorem:

$$P(G|D) \propto P(D|G) \times P(G)$$

where G is the network and D is biological data. $P(G|D)$ is the probability of network G given data D; we therefore maximized $P(G|D)$ to get the best network from data D . $P(G|D)$ is proportional to multiplication of $P(D|G)$ and $P(G)$ according to the Bayes theorem. Therefore, we found the graph G that maximizes $P(D|G)$, which indicates how accurately the network G represents the gene expression data D , and $P(G)$, which indicates how well the network G represents biological knowledge. Calculation of $P(D|G)$ and $P(G)$ by our fitness function is described in more detail in Section 2-3.

2.2 Biological knowledge and preprocessing

We used two types of data - gene expression data and biological knowledge data – to learn the Bayesian network. Gene expression data were obtained from microarray experiments. To apply gene expression data to a multinomial Bayesian network, we discretized gene expression values into -1, 0, and 1 using the modified Friedman's method [6]. In this paper, the average expression level of a gene across all experiments was used as the control expression level of the gene. We also chose a threshold value of 0.5 in logarithmic scale base 2. Because only a few samples can be run on microarrays and the data is extremely noisy, we merged multiple microarray studies after discretization.

We use functional similarity and HumanNET [13] as biological knowledge data to construct the Bayesian network. We measured functional similarity as closeness in a PPI network [15]:

$$S(p, p') = Ae^{-bD(p,p')}$$

where $S(p, p')$ computes the similarity between two proteins, corresponding to two genes in our network. $D(p, p')$ is the shortest path between these proteins in the PPI network, A is 0.9, and b is 1. Self-similarity is set to 1.

HumanNET is a probabilistic functional gene network constructed using various genetic data. In HumanNET, each interaction between two genes has a log-likelihood score (LLS) that measures the probability that linkage is true or not. In this paper, we calculated the minimum and maximum LLS according to the generated network, and then normalized the LLS values to be in the range [0, 1].

2.3 Bayesian network learned by genetic algorithm

We used a genetic algorithm to learn the Bayesian network. We defined the structure of the network as a chromosome in the genetic algorithm. A Bayesian network has the following characteristics:

1. It is a directed network, and edges have a direction.

2. It is an acyclic network, so there is no cycle in the network.

In this paper, we defined a chromosome of an individual as shown in Figure 2.

$E_{1,2}$	$E_{1,3}$	$E_{1,4}$	...	$E_{1,n}$	$E_{2,3}$	...	$E_{n-2,n-1}$	$E_{n-2,n}$	$E_{n-1,n}$

Figure 2. Chromosome of an individual

To minimize computational costs and fix specific nodes upstream, we arranged chromosomes as shown in Figure 2 instead of in a $n * m$ matrix. We fixed specific nodes upstream to identify the relationship between genotype variations and phenotype entries. Furthermore, there was no bi-directional edge between two nodes in the network when we used the chromosome shown in Figure 2. Genes in the chromosome, which represents edges between two nodes in the network, had one of three values (-1, 0, 1). For example, if $E_{1,2}$ was 1, then there was an edge going from (gene node 1) to (gene node 2). If $E_{1,2}$ was -1, then the edge was going from (gene node 2) to (gene node 1). If $E_{1,2}$ was 0, there was no edge between (gene node 1) and (gene node 2).

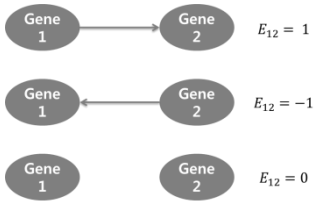

Figure 3 Representation of a network according to the chromosome structure shown in Figure 2.

We constructed Bayesian networks using the chromosome described above. First, we generated the initial population randomly. We evaluated the fitness of every individual using gene expression data and biological knowledge data. Then, we selected a group of individuals based on their fitness using a roulette wheel algorithm. Individuals with higher fitness scores were more likely to be selected. The selected group became parents for the next generation, and each chromosome of the selected individuals could be mutated or crossed over. This process was repeated until termination conditions were reached. In this paper, we defined the mutation of chromosome as an edge generation, deletion, or alteration. Each type of the mutation had uniform probability. To construct a minimum and reliable network, we limited the maximum number of edges in the network. When the number of edges in the network exceeded the limitation of edges during the learning process, edges in the network were deleted or altered direction during the mutation procedure.

Because a Bayesian network is a directed acyclic graph, the generated network might have cycles due to the characteristics of the genetic algorithm, where structural alterations in the network happen randomly. We performed a depth-first search (DFS) to remove the cycles.

2.4 Fitness of the network

The fitness function calculates the probability of network G for the given data D. Based on the Bayes theorem, the posterior probability of network G for given data D is proportional to multiplication of the likelihood $P(D|G)$ and the structure prior probability $P(G)$:

$$P(G|D) \propto P(D|G) \times P(G).$$

In this paper, we computed $P(D|G)$ as the ability to predict the pattern of gene expression in a given network. We divided gene expression data into a training set and test set. We determined a parameter (CPT) of the Bayesian network using the structure of network G and the training gene expression dataset. Then we predicted the most probable state of gene expression (one of -1, 0, 1) for each node. This was predicted by the local Markov assumption implied by the Bayesian network G. Next, we evaluated the network by comparing predicted states of predefined nodes from the training set with states of the same nodes from the test set. The more closely matched the two group of states were, the higher the $P(G|D)$ score. We defined predefined nodes as leaf nodes and randomly selected nodes, excluding upstream nodes. We prevent overfitting by selecting nodes randomly for every individual network.

The fitness function of the Bayesian network, because it uses the joint probability table, would benefit if patterns of states of the parent nodes and child node had consistency in the network. Edges that connect two nodes with state 0 (baseline) in the training set would result in a network with a higher score. However, those edges would be biologically meaningless interactions. Therefore, those networks with a large number of edges with a node of which the state was non-zero received more weight.

$P(G)$, the prior probability of the network structure, measures how closely the network represents biological knowledge. When the number of nodes is n, the prior knowledge n * n matrix is constructed by adding a weight matrix of functional similarity and HumanNET. We used the prior knowledge matrix with relative scaling values of 0 to 1. We calculated $P(G)$ as the average of the prior knowledge matrix in the given network:

$$prior(G) = \sum_{i,j \in V, i \neq j} \frac{prior(i,j)}{|E|}$$

where E is a set of edges in the given network, and $prior(i,j)$ denotes the weight of the edge between node i and node j, respectively, in the set of nodes V.

Because we used only selected nodes to calculate the fitness of individual networks and compared predicted states of nodes from the training set and test set, overfitting of the test set might have occurred. Furthermore, gene expression data were available for only a few samples; we therefore performed 10-fold cross validation.

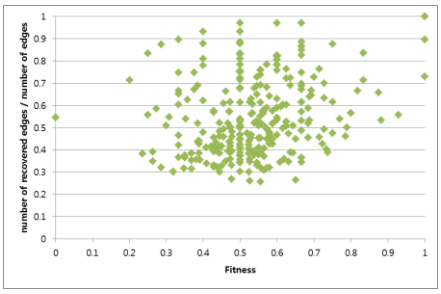

Figure 5.
Fitness ratio of recovered edges.

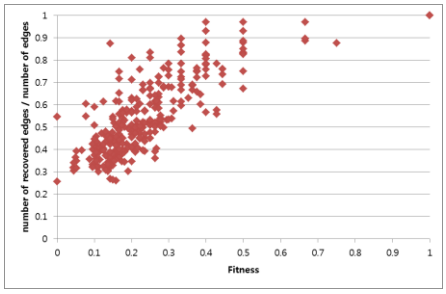

Figure 4.
Fitness and ratio of recovered direct edges

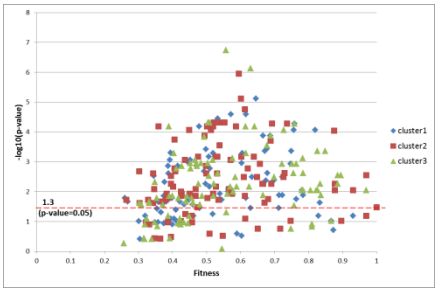

Figure 6.
Fitness and p-values for each cluster.

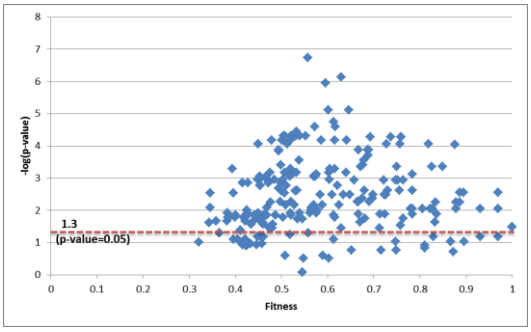

Figure 7. Fitness and p-values

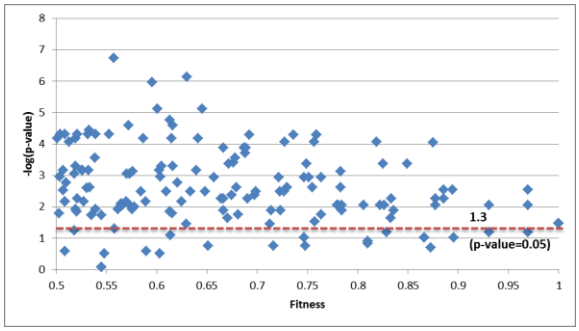

Figure 8. Fitness and p-values over 0.05.

3. EXPERIMENTAL RESULTS

To evaluate the robustness of our approach, we reconstructed a known pathway from the KEGG database. In this experiment, we reconstructed the type-2-diabetes (T2D) pathway in KEGG (hsa04930) using T2D microarray data.

Microarray data used in our experiment are listed in Table 1. We use five GEO studies with a total of 46 T2D samples.

Diseases are often classified into several subtypes or states. Although T2D is a diabetes subtype, there could also be different subtypes within T2D. Our gene expression data, however, contained no information regarding subtypes or states, and may not have been internally consistent due to integration of multiple microarray studies. Therefore, we performed hierarchical clustering after discretization for each microarray. We assumed that samples belonging to similar subtypes or states would tend to cluster together. When we trained the network, we calculated the fitness of the network using sample data in the same cluster. There were three clusters, and the number of samples in each cluster was 26, 14, and 6, respectively.

Table 1. Microarray data used in the validation experiment

GEO accession	Platform	Total samples	T2D samples	Control samples
GSE9006	GPL96	117	12	24
GSE16415	GPL2986	10	5	5
GSE23343	GPL570	17	10	7
GSE25462	GPL570	50	10	40
GSE26168	GPL6883	23	9	8
	Total		**46**	**84**

The PPI dataset was obtained from STRING [10] and I2D [1]. The T2D KEGG pathway has 22 nodes and 30 edges (19 direct edges, 11 indirect edges). In this experiment, we used all 22 known nodes in the KEGG pathway and 14 SNPs associated with T2D. All known edges were hidden. SNP data were obtained from the GWAS catalog (11-28-08) [7]. Because we integrated multiple microarray datasets across different platforms, the number of common genes was limited. Fifteen were common of 22 genes from KEGG, and six SNPs were common of 14 SNPs in the integrated microarray datasets. Seventeen edges were recovered (7 direct edges, 10 indirect edges).

Parameters for our genetic algorithm were as follows in Table 2

Table 2. Parameters of the genetic algorithm

Parameter	Value
Mutation rate	0.1
Crossover rate	0.5
Maximum number of edges	Number of nodes
Population size	20, 30, 40, 50, 70, 100
Generation size	100, 200, 300, 500, 700, 1000

PPI and HumanNET data are not directional. Gene expression data is also not directional data. Even though a Bayesian network is a directed network, we used directionless data to learn the Bayesian network. Edges may not have been in the right directions; we therefore removed edge direction after network generation.

To validate that our proposed fitness function was capable of extracting genetic interactions, we calculated the PCC of the ratio of recovered edges to the fitness score. We assumed that generated interactions present in the KEGG pathway were "recovered" edges, and those not present in the KEGG pathway

were "unconfirmed" edges. A PCC value of 1 indicates a perfect positive correlation, so if PCC value close to 1 would indicate that our fitness function was successful at identifying biologically relevant genetic interactions. PCC values of clusters 1, 2, and 3 were 0.32, 0.27, and 0.27 respectively, indicating weak positive correlations, as shown in Figure 6. X-axis and Y-axis represent fitness and the ratio of recovered edges, respectively.

Two types of edges exist in a KEGG pathway: direct edges and indirect edges. Indirect edges represent unknown interactions between two nodes, indicating that there may be many mediators between two nodes. The number of KEGG nodes for the T2D pathway was 22, and the maximum number of edges of the network was set to the number of nodes. An average of 15 edges was reconstructed for each cluster. It is challenging to recover indirect edges when the number of edges is small; we therefore focused only on direct edges (Figure 4).

PCC values of the clusters without indirect edges were 0.77, 0.78, and 0.81, with an average value of 0.78. There was a strong positive relationship between the fitness score and the number of recovered edges. These results demonstrated that our proposed fitness function can reliably identify direct genetic interactions.

To assess the reliability of our approach, we calculated the p-value of the constructed network, and compared it to the p-value of a random network. We calculated the probability of obtaining at least k recovered edges from a randomly generated network with n edges. When the number of nodes in a network is fixed and the maximum number of edges is N, a set of M edges is recovered, while $N - M$ number of edges is unconfirmed. According to the hypergeometric distribution, the probability of obtaining m recovered edges and $n - m$ uncertain edges can be expressed as

$$P_x = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

where N is $\binom{n}{2}$ when n nodes are given. Therefore, the p-value of a network is given by

$$p - value = \sum_{m=0}^{k} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

Note that the probability for indirect edges cannot be calculated. P-values and fitness scores across experiments are displayed in Figure 6. Two hundred forty-seven of 313 experiments had a p-value less than 0.05.

The p-value for the individual clusters was 0.035, 0.036, and 0.059 respectively, with an average value of 0.043, which is statistically significant.

If the population is too small, non-optimal and poor solutions may result [12]. Therefore, we excluded experiments with a population size under 20 and recalculated p-values; results are displayed in Figure 7. Average p-value of the experiments was 0.032, which is statistically significant at the 0.05 level. Two hundred eleven of 250 experiments had a p-value under 0.05.

To confirm that our fitness function generated a reliable and statistically significant network, we excluded experiments with a fitness value of less than 0.5. As shown in Figure 8, most experiments had a p-value greater than 0.05. The average p-value of the experiments was 0.033, and 136 of 157 experiments had a p-value less than 0.05.

4. CASE STUDY AND DISCUSSION

We confirmed that the proposed method was able to recover reliable and statistically significant genetic interactions via T2D analysis. In this section, we inferred a disease-specific genetic interaction network for Alzheimer's disease using our reconstruction method. No information about pathway nodes and edges was used in this experiment.

First, we selected genes for network reconstruction. We choose three DEGs with the lowest p-values from microarrays GSE1297 and GSE28146, respectively. Because microarray GSE5281 has information about brain regions, we selected three DEGs from each brain region. With these seed DEGs, we created eight gene sets from three DEGs in the PPI network (distance from DEG <3). Then, we set network nodes as the intersection of these 8 gene sets.

We fixed SNPs to be at upstream of the network because genomic variations might be the root causes in the organism. In other words, SNPs only had outgoing edges. We obtained SNPs related

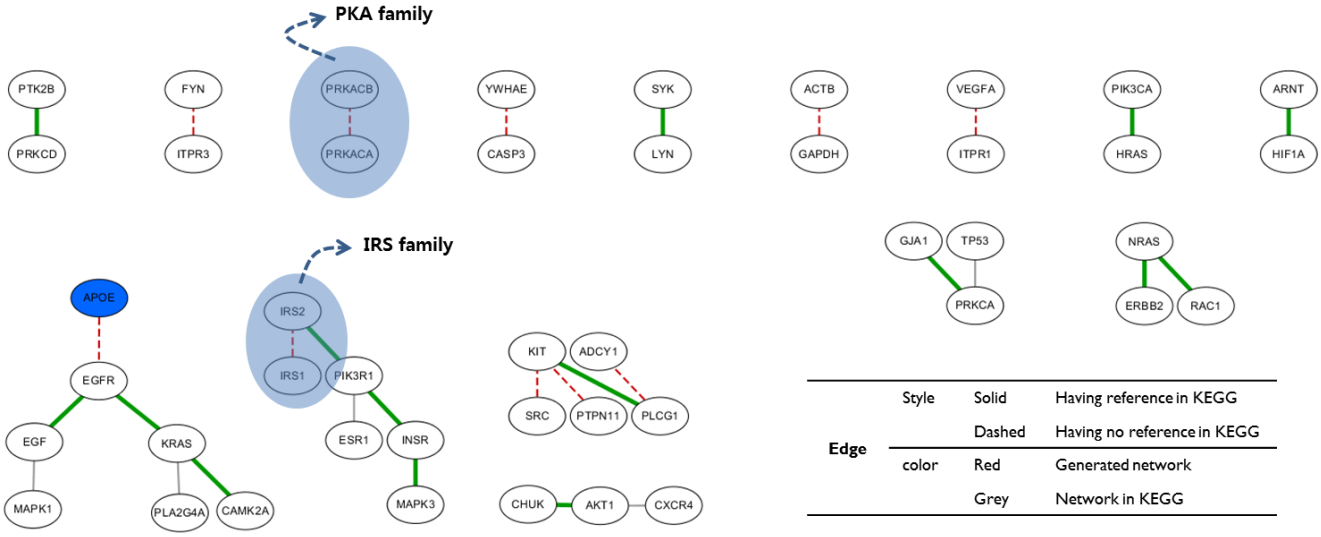

Figure 9. Network with the maximum fitness value

to Alzheimer's disease from GWAS. The final number of nodes was 98.

We performed hierarchical clustering and observed two clusters. A set of parameters for the genetic algorithm is almost same as prior experiment (in Table 2) except that maximum number of edges is 20, 50, 70, and 100. This experiment was repeated approximately 250 times.

The network that had the maximum fitness value regardless of genetic algorithm parameters is shown in Figure 10.

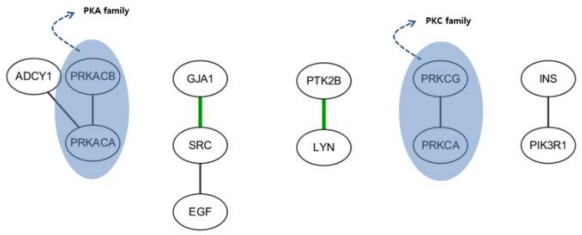

Figure 10. Network with the maximum fitness value

Thick solid lines between nodes represent edges that have one or more references in KEGG, while thin solid lines are edges represented directly in our network but indirectly in KEGG. In other words, if edges are thin solid lines, there are some genes between the two nodes in KEGG. For cases like this, we considered only sequential interactions considering the direction of the edges, regardless of the number of nodes between them. PRKCA is associated with memory capacity [5]. Patients who suffer from Alzheimer's disease have been reported to have abnormalities in protein kinase C (PKC) expression and activity in the brain. Furthermore, PKCs, which include PRKCG and PRKCA, have been implicated in the pathophysiology of Alzheimer's disease. More specifically, PKA activity is down-regulated in the brains of Alzheimer's disease patients [14].

Interactions with a KEGG reference are described in Table 3. An interaction between PIK3R1 and INS is represented in our network, but not directly in KEGG. This indicates the presence of one or more genes such as IRS1, GFR, etc. between PIK3R1 and INS in the KEGG pathway. Furthermore, an interaction between EGF and SRC is represented by the RTK gene in-between them in the gap junction pathway. An interaction between PRKACA and ADCY1 is evidenced by the presence of cAMP in a number of pathways. Therefore, all interactions in our network are reflected in KEGG in some way.

Table 3. Edges in KEGG

Node1	Node2	Reference (KEGG pathway)
SRC	**GJA1**	Gap junction (hsa04540)
LYN	**PTK2B**	Chemokine signaling pathway (hsa04062)

We examined the best network among the networks, which had more than 30 edges. This network is displayed in Figure 9. The network had 30 edges, of which 15 had one or more references in KEGG. Dashed lines indicate edges with no references in KEGG. When we excluded interactions between protein families such as PKA and IRS, only eight edges did not have a reference. In our network, there was an interaction between APOE and EGFR. APOE refers to a SNP in the gene encoding apolipoprotein E that is genetically associated with Alzheimer disease [2]. In the transcriptional of EGFR has been reported to be altered in Alzheimer disease patients [4], and some researchers have reported that EGFR is one of the most significant Alzheimer disease risk genes [18]. The interaction shown in our network might be a novel interaction that connects genotype to phenotype. There are some interactions that are represented directly in our network but indirectly in KEGG. This may have been due to the nodes we selected. Together, these results indicate that our proposed approach was able to identify a set of genetic interactions related to Alzheimer's disease.

5. CONCLUSION & FUTURE WORK

How genotype maps to phenotype remains largely unknown and is usually investigated in biological experiments. A goal of systems biology is to construct the phenomenon of life as a system from fragmentary data.

In this paper, we inferred disease disease-specific genetic interaction networks using Bayesian networks. Parameters and the structure of the network were learned via genetic algorithm using heterogeneous data. We confirmed that the proposed fitness function was capable of inferring genetic interactions and that the structure of the learned network was statistically significant. Furthermore, we inferred the disease disease-specific genetic interaction network for Alzheimer's disease. We demonstrated that our approach is robust by performing two different types of analyses. Our method detected not only genetic interactions, but specific modules for disease.

Disease-specific genetic interaction networks could be markers of specific diseases, and help our understanding of disease. Interest in personalized care is increasing at present. Networks can potentially be generated using gene expression data from single individuals, and the disease-specific genetic interaction network can be used as a classification model.

In the future, we intend to use time-series microarray or RNA-seq data to construct Bayesian networks to provide more reliable and accurate results than those obtained using microarray data alone. In particular, time series gene expression data may help determine the exact direction of edges. We also intend to incorporate more prior biological information in our model to improve its reliability and accuracy [9]. We use a genetic algorithm to learn the Bayesian network. The efficiency of the genetic algorithm relies on parameters of genetic algorithm [12]. In this paper, we do not optimize parameters for the genetic algorithm. If we select proper parameter for the genetic algorithm, the result might be more improve.

6. ACKNOWLEDGMENTS

This Research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIP) (2012R1A2A1A01010775). Sanghyun Park is corresponding author of this paper.

7. REFERENCES

[1] Brown, Kevin R., and Igor Jurisica. Online predicted human interaction database. *Bioinformatics, 21.9* (2005): 2076-2082.

[2] Corder, E. H., et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer disease in late onset families. *Science, 261.5123* (1993): 921-923.

[3] Djebbari, Amira, and John Quackenbush. Seeded Bayesian Networks: constructing genetic networks from microarray data. *BMC systems biology, 2.1* (2008): 57.

[4] Doecke, James D., et al. Blood-based protein biomarkers for diagnosis of Alzheimer disease. *Archives of neurology, 69.10* (2012): 1318-1325.

[5] Dominique, J-F., et al. PKCα is genetically linked to memory capacity in healthy subjects and to risk for posttraumatic stress disorder in genocide survivors. *Proceedings of the National Academy of Sciences, 109.22* (2012): 8746-8751.

[6] Friedman, Nir, et al. Using Bayesian networks to analyze expression data. *Journal of computational biology, 7.3-4* (2000): 601-620.

[7] Hindorff LA, MacArthur J (European Bioinformatics Institute), Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed [date of access]

[8] Imoto, Seiya, et al. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Journal of Bioinformatics and Computational Biology, 2.01* (2004): 77-98.

[9] Isci, Senol, et al. Bayesian network prior: network analysis of biological data using external knowledge. *Bioinformatics, 30.6* (2014): 860-867.

[10] Jensen, Lars J., et al. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research,* 37.suppl 1 (2009): D412-D416.

[11] Kanehisa, Minoru, and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research, 28.1* (2000): 27-30

[12] Koumousis, Vlasis K., and Christos P. Katsaras. A saw-tooth genetic algorithm combining the effects of variable population size and reinitialization to enhance performance. *Evolutionary Computation, IEEE Transactions on 10.1* (2006): 19-28.

[13] Lee, Insuk, et al. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research, 21.7* (2011): 1109-1121.

[14] Liang, Zhihou, et al. Down-regulation of cAMP-dependent protein kinase by over-activated calpain in Alzheimer disease brain. *Journal of neurochemistry, 103.6* (2007): 2462-2470.

[15] Perlman, Liat, et al. Combining drug and gene similarity measures for drug-target elucidation. *Journal of computational biology, 18.2* (2011): 133-145.

[16] R. Gentleman, Elizabeth Whalen, W. Huber and S. Falcon (2006). graph: graph: A package to handle graph data structures. R package version 1.38.3.

[17] Shastry, Barkur S. SNPs: impact on gene function and phenotype. Single Nucleotide Polymorphisms. *Humana Press, 2009*. 3-22.

[18] Talwar, Puneet, et al. Genomic convergence and network analysis approach to identify candidate genes in Alzheimer disease. *BMC genomics, 15.1* (2014): 199.