Exploring candidate genes for human brain diseases from a brain-specific gene network

Bing Liu, Tianzi Jiang *, Songde Ma, Huizhi Zhao, Jun Li, Xingpeng Jiang, Jing Zhang

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, PR China

Received 27 August 2006
Available online 7 September 2006

Abstract

It is believed that large numbers of genes are involved in common human brain diseases. Here, we propose a novel computational strategy for simultaneously identifying multiple candidate genes for genetic human brain diseases from a brain-specific gene network-level perspective. By integrating diverse genomic and proteomic datasets based on Bayesian statistical model, we built a large-scale human brain-specific gene network. Based on this network and minor prior knowledge of a specific brain disease, we can effectively identify multiple candidate genes for this disease. When four known Alzheimer’s disease genes were used as the prior knowledge, among the top 46 high-scoring genes that we have found, 37 were previously reported to be associated with Alzheimer’s disease. And the higher score a gene has, the more likely this gene is a disease-related one. The results suggest that the proposed method is effective, convenient, and applicable in the future genetic studies.

© 2006 Elsevier Inc. All rights reserved.

Keywords: Bayesian statistical model; Brain-specific gene network; Subnetwork; Candidate genes

Many common human brain diseases, such as schizophrenia, Alzheimer’s disease, Parkinson’s disease, depression, etc., have prominent genetic components [1,2]. Most researchers think that large numbers of genes each act in a small but significant manner to predispose an individual to disease [3,4]. One of the main aims of current genetic studies is to find these genes and to clarify their role in disease [5]. However, for the abundant numbers of genetic factors involved in polygenic diseases, the limited number of candidate genes selected only from biological studies have been far fewer than what is needed. Many approaches have been proposed for solving the problem of selecting candidate genes. Recently, genome-wide disease-association mapping, which genotypes the whole genome bypassing the selection of candidate genes, has been heralded as one of the main, next-generation study designs [6]. However, it is even more time and labor consuming than genetic candidate study.

New computational methods for predicting susceptibility genes for common diseases have emerged with the public availability of various large-scale human genome datasets. Most of computational methods predicted such gene groups either by literature analysis [7,8] or by high-throughput data mining techniques based on diverse public databases [9,10]. These methods have afforded many valuable clues for genetic studies of certain common diseases and undoubtedly could save much labor. However, all of these studies have attempted to find candidate genes from those associated with disease phenotypes and have heavily depended on either previously reported results or on definite phenotype characterizations. One recent study [11] has presented a molecular-triangulation method, which used a species-specific molecular network built by text mining to identify candidate genes in Alzheimer’s disease. Here, we attempted to explore susceptibility genes for genetic human brain diseases from the human organ-specific, that is, human brain-specific, gene network, which...
was built by integrating diverse genomic and proteomic datasets. From this, we developed a strategy to predict a list of susceptibility genes computationally based on a brain-specific gene network. Our hypothesis was that the multiple susceptibility genes of a specific brain disease might be extracted as a subnetwork of the entire network. This idea has been implemented by two key processes. First, by integrating diverse genomic and proteomic datasets based on Bayesian statistical model, we constructed a gene interactome network of the human brain, which consisted of those genes expressed in the human brain. The framework for constructing the brain-specific gene network can be seen from Fig. 1. Second, we developed an effective scheme for scoring each gene by estimating the importance of each gene relative to seed genes which are the prior knowledge of a specific disease throughout the entire network. And those top-scoring interacting genes were thought to construct the disease-related subnetwork. We applied this approach to identify susceptibility genes for Alzheimer’s disease and obtained a great success. In addition, we further tested the stability of our method by adding random noise and validated the sensitivity and specificity of our computational prediction.

Materials and methods

Integrated diverse datasets. We collected 10 individual, publicly available, microarray experiments of 367 human brain samples (Supplementary Table 1), to identify co-expression relationships between brain genes. According to the expression variations of 8283 brain genes through different experiments, we used an algorithm similar to ARACNe [12] to construct a brain-specific gene network. This is implemented as follows: (1) preprocess the raw datasets, including mapping brain genes and normalization; (2) estimate the mutual information of each gene pair; (3) do 1000,000 random permutation tests to get the significant threshold of mutual information for the network; (4) applied data processing inequality (DPI) to delete redundant edges from the primary mutual information network.

In order to mine co-citation relationships between brain genes, we downloaded and analyzed a set of N = 189,406 Medline abstracts, and the gene pairs co-cited in the same abstract and the times of co-citation were saved for further integration. We searched the Database of Interacting Proteins (DIP: http://dip.doe-mbi.ucla.edu/) to download the interaction data of three model organisms (Caenorhabditis elegans; Saccharomyces cerevisiae; Drosophila melanogaster), and then mapped model organism proteins to human orthologs using the Inparanoid database (http://inparanoid.cgb.ki.se/). All interaction data were downloaded from the Human Protein Reference Database (HPRD http://www.hprd.org/) on October of 2005. Donaldson et al. [13] computationally predicted large-scale potential gene interactions and these data can be publicly downloaded from BIND (http://prebind.bind.ca/). We also used the recently published datasets by Stelzl et al. [14] which used automated yeast two-hybrid (Y2H) interaction mating. For these datasets, it needs to map genes and only selected the interactions in which both genes are included in our human brain gene set.

Benchmark sets. We downloaded biological process annotation from the Gene Ontology Consortium (http://www.geneontology.org/), and then classified 8283 genes into 393 categories based on the 8th level annotation. We assign pairs of genes that shared at least one annotation to be positive.

Fig. 1. Flowchart of the construction of a human brain-specific gene network.
sets, and those pairs of genes that shared no annotations are chosen as negative sets. These benchmark sets are used for the unified rescoring schemes.

Unified scoring schemes. The odds ratio representing the likelihood that if a pair of genes is linked in each experiment is defined as [15]:

\[ \text{OR}(L, E) = \frac{P(L|E)}{P(L)} \]

\[ \text{Lscore} = \ln(\text{OR}(L, E)). \]

Here \( P(L|E) \) represents the conditional probability of linkage between a pair of genes given evidence \( E \), \( \sim P(L|E) \) represents the complementary probability of \( P(L|E) \), and \( P(L) \) is the probability of linkage between a pair of genes. If \( \text{Lscore} \) is greater than zero, it indicates that the experiment tends to link genes within the same pathway, with higher scores indicating more confident linkages. According to this criterion, we can rescore diverse datasets and integrate them into one brain-specific gene network.

Finding the subnetwork of a specific brain disease. We developed an effective method to find the subnetwork related to a specific brain disease based on the entire brain-specific gene network. It includes the following steps: (1) define one or more seed genes of a certain brain disease; (2) compute the shortest graph-theory distances between each node and seed nodes; (3) score the relative importance of each gene based on the following scoring functions:

\[ \text{Score} = \text{Evidence} \times \text{Zscore}, \]

\[ \text{Evidence} = \sum_{j} W \times f_{ij}, \]

\[ f_{ij} = \frac{1}{1 + d_{ij}}, \]

\[ \text{Zscore} = \frac{\text{Evidence} - \text{mean}(\text{RndEvidence})}{\text{std}(\text{RndEvidence})}. \]

where \( W \) is the weight of each seed gene (in this study, \( W = 1 \)), \( d_{ij} \) is the shortest graph-theory distance between gene \( i \) and seed gene \( j \), \( \text{RndEvidence} \) is the Evidence values when seeding the equivalent number of random genes 100 times. Here, we introduced a z-score into scoring function, so that it can avoid the likely possibility that higher degree genes in the network get higher scores (4) sort genes based on the scores of each gene and choose the top \( n \) genes as candidate genes of this disease, and the interactions between them formed the disease-related subnetwork. It is assumed that the higher score that one gene has, the more likely this gene is a susceptibility one of the disease.

Validating of the AD related subnetwork. We searched the OMIM database with the keyword “Alzheimer” and selected 114 genes (Supplementary Table 2) included in our brain gene set from the searching result. Then 46 genes previously reported to be associated with AD in various PubMed listings were selected as an AD positive gene set (Supplementary Table 3). The percentage of AD positive genes in the top 46 high-scoring genes was computed based on our defined 46 AD positive genes. We examined the stability of our prediction system by comparing the results of sequentially adding random seed genes with the results from using the same number of random seed genes, and also we analyzed the prediction results by randomly selecting 4 genes in the 46 AD positive gene set as seed nodes.

Results

Construction of the human brain-specific gene network

Various genome-wide complex biological networks have been successfully constructed in relative simple genomes by large-scale experimental techniques and computational methods [16–20]. However, few have attempted to construct such a network in higher-order eukaryotic organisms, especially in mammalian genomes, because of their complexity and resulting difficulties. Here, we constructed a human brain-specific gene network (Fig. 1) by integrating diverse genomic and proteomic datasets based on a Bayesian statistics model, which has previously been successfully applied to constructing high-quality yeast and C. elegans gene networks [15,21]. First, we collected and filtered various public datasets that we thought would potentially be useful in predicting a human brain-specific gene network. Then we rescored these disparate data types into union scores (Supplementary Fig. 1) and integrated them based on the naive Bayesian classifier. Eventually, a global view of the functional interactions among human brain genes was presented.

This network consisted of ~8000 genes and ~45,000 interactions. Its global connectivity properties (Supplementary Fig. 2) suggested a scale-free behavior, since the results showed a power-law phenomenon between the number of genes in the network and the number of their interactions when the interaction number was over twelve. Using BiNGO, a plugin of Cytoscape software, we assigned the biological process from gene ontology (GO) to highly connected genes, those whose connectivity degree \( \geq 100 \), in the network, called hubs, and found enriched functions that were the keys to understanding the effects of hubs (Supplementary Table 4).

Identification of the AD related subnetwork

Based on the above brain-specific gene network and prior or knowledge about the candidate genes for a specific complex brain disease (that is, one or more susceptibility genes of the disease had previously been identified), we attempted to predict a list of potential candidate genes computationally and thereby to find the subnetwork of the brain disease. In this way, the problem could be reduced to defining and measuring the importance of nodes in a graph relative to one or more root nodes. In this paper, we propose an effective scoring scheme to evaluate the relative importance of genes in a specific brain disease based on the linkage characteristics of the entire network. We select the top-scoring interacting genes, which appear to construct the disease-related subnetwork, as candidate genes for this brain disease. And we assume that the higher score one gene has, the more likely this gene is the disease-related candidate one.

We applied the brain-specific network-level computational strategy to predict a list of candidate genes for Alzheimer’s disease, that is, to identify the AD related subnetwork. First, 46 AD positive genes were defined. Next, we chose four known AD genes (APOE, APP, PSEN1, and PSEN2) [11,22] as seed nodes to score each gene based on the brain-specific gene network. Afterwards, each gene was ranked by its corresponding score, and the top 46 high-scoring genes (Table 1) were predicted as AD candidate genes. These 46 genes and their interactions with each other enabled us to construct an AD related gene subnetwork (Supplementary Fig. 3). We were very pleased to
OMIM as AD susceptibility genes and not reported previously to be associated with AD. BACE, ESR1, LPL, and TAU) excluding four seed genes; however, among top 50 genes predicted by us, there are 11 expert AD genes (Table 1) excluding four seed genes. Thus, from the above two comparisons, we both can clearly see the significant improvements in our prediction results. So, we believed that our method can be more effectively used to identify candidate genes for human brain diseases.

Table 1 indicates that the twelve highest score genes are all in our AD positive gene set and only three genes out of the top twenty are not AD positive genes. This finding supported our hypothesis that the higher score a gene has, the more likely that the gene is an AD susceptibility one. We also assigned biological processes based on gene ontology using BiNGO for the predicted top 46 genes (Supplementary Table 6), and found that most of the significant biological processes, including copper ion transport [23], amyloid precursor protein catabolism [24], notch receptor processing [25], etc., were important and potentially pathological AD pathways.

Validation of the AD related subnetwork

To evaluate the performance of our prediction system, we roughly investigated the false positive rate and the false negative rate using known evidences. From Table 1, assuming that only those genes with known evidences are actually AD related, only 9 false positive genes exist out of 46 predicted AD genes. We also examined the 6000 lowest score gene, and found only 5 AD positive genes (GAL,
HLA-A, HTR2A, IGF1, and PIN1) among them. It means that, the other 41 AD positive genes were in the set of the 2000 top score genes. We divided the top 1000 genes according to their scores into ten equal bins to obtain the distribution of the AD positive genes. The distribution (Fig. 2A) showed that with a decrease in scores, the percentage of AD positive genes included in each bin is reduced dramatically and therefore this further supports our hypothesis that the higher score a gene has, the more potential the gene is to be identified as a susceptibility one.

In order to examine the importance of seed genes in predicting a list of susceptibility genes, we rescored each gene, by adding random genes as seed nodes, to predict AD susceptibility genes and thus evaluate the stability of our method (Fig. 2B). Using the four known AD genes as seed genes, the percentage of AD positive genes in the top 46 high-scoring genes was 50%. But when the four known AD genes were substituted by four random genes, and the process repeated 100 times, the mean percentage was only 2.60% (95% CI: 2.05–3.15%). As the number of non-relevant random genes increased, the mean percentage decreased. Even after adding twenty additional non-relevant random genes to the four seed genes, for a total of 24 genes as seeds, the percentage of AD positive genes in top the 46 high-scoring genes was still significantly higher than the result obtained by using the same number of random genes as seed genes. Furthermore, we randomly selected four genes from our defined 46 AD positive gene set to use as seeds and repeated the test 100 times; in this case the mean percentage was still 31.48% (95% CI: 29.95–33.01%), which also overwhelmingly exceeded the results obtained by using four random seeds. We found that four known AD genes (APP, APOE, PSEN1, and PSEN2) and several other genes (GSK3B, LRP1, APBB1 BCHE, and LDLR) could be successfully located with over 80% probability when we randomly selected four genes from the 46 positive genes as seeds. We hypothesize that these genes may play pivotal roles in AD pathology. For example, four known AD genes, and GSK3B, LRP1, and APBB1 are included in the AD pathway of KEGG.

Discussion

This is the first time to our knowledge that a large-scale human brain-specific gene network has ever been constructed. We have integrated diverse genomic and proteomic data to construct a system-level human brain gene network. Then based on this network, we have developed an effective method to computationally predict a list of candidate genes for specific brain diseases and find the disease-related gene subnetwork. This method has been used to effectively identify multiple candidate genes of AD and their AD related subnetwork, and also has good stability when random noises are introduced. So, we conclude that this brain-specific network-level strategy can be used to effectively predict multiple candidate genes for common brain diseases, and this computational method can save time and effort compared with traditional genetic methods.

It should be noted that we cannot be completely certain that the human brain genes defined by us are complete and accurate nor that the human brain-specific gene network we constructed based on various large-scale public datasets is sufficient and accurate. This potential for incompleteness and inaccuracy of this network undoubtedly greatly lowers the certainty of the performance of this method of predicting susceptibility genes of common brain diseases. When more such public datasets are available and more biological experiments have been done, a more nearly accurate human brain-specific gene network can be constructed. At that time we believe that this model and method can be used to obtain even more exciting results.
In this paper, we have only concentrated on Alzheimer’s disease. In fact, based on the human brain-specific gene network we constructed, we believe that it is possible to effectively predict large-scale susceptibility genes of any other common brain disorders using the same approach. If more prior genetic knowledge, such as more accurate seed genes and the weights of each seed gene, which can, for example, be determined by previous reports of linkage analysis or association studies, is available, the results can be improved to be more accurate. We have defined 46 AD genes based on the results of searching the OMIM database and related articles referenced in PubMed. But these genes may not be held to be true AD susceptibility genes with absolute certainty. Thus when these genes are used as true AD genes to test the performance of our method, it could potentially cause some bias. Thus, further biological experiments using our predicted results need to be done in order to obtain a more conclusive validation.

Acknowledgments

The authors are grateful to Dai Zhang, Institute of Mental Health, Peking University, and Michael Zhang of the Cold Springs Laboratory in Cold Spring Harbor, NY, USA, for some valuable discussions. We also appreciate the extensive editing and English language assistance provided by Dr. Rhoda E. Perozzi of the Beijing University of Technology. This work was partially supported by the Natural Science Foundation of China, Grant Nos. 30425004, 60121302, and the National Key Basic Research and Development Program (973) Grant 2004CB318107.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbr.2006.08.168. And also large-scale supplementary data can be found by the below URL: http://www.nlpr.ia.ac.cn/english/mic/2006.htm.

References


