TOWARDS INTEGRATIVE GENE PRIORITIZATION IN ALZHEIMER'S DISEASE

JANG H. LEE[1] AND GRACIELA H. GONZALEZ[2]

[1]*School of Computing, Informatics, and Decision Systems Engineering, Arizona State University,Tempe, AZ 85287-8809*
[2]*Department of Biomedical Informatics, Arizona State University,425 N. 5th Street, Phoenix, AZ 85004*

Many methods have been proposed for facilitating the uncovering of genes that underlie the pathology of different diseases. Some are purely statistical, resulting in a (mostly) undifferentiated set of genes that are differentially expressed (or co-expressed), while others seek to prioritize the resulting set of genes through comparison against specific known targets. Most of the recent approaches use either single data or knowledge sources, or combine the independent predictions from each source. However, given that multiple kinds of heterogeneous sources are potentially relevant for gene prioritization, each subject to different levels of noise and of varying reliability, each source bearing information not carried by another, we claim that an ideal prioritization method should provide ways to discern amongst them in a true integrative fashion that captures the subtleties of each, rather than using a simple combination of sources. Integration of multiple data for gene prioritization is thus more challenging than its single data type counterpart. What we propose is a novel, general, and flexible formulation that enables multi-source data integration for gene prioritization that maximizes the complementary nature of different data and knowledge sources in order to make the most use of the information content of aggregate data. Protein-protein interactions and Gene Ontology annotations were used as knowledge sources, together with assay-specific gene expression and genome-wide association data. Leave-one-out testing was performed using a known set of Alzheimer's Disease genes to validate our proposed method. We show that our proposed method performs better than the best multi-source gene prioritization systems currently published.

1. Introduction

Of particular relevance to researchers trying to track the molecular basis of disease is to be able to increase the selectivity and sensitivity when predicting the potential association of a phenotype or function with specific genes, an area referred to as "gene prioritization". Genome sizes of species of interest are typically large, and gene prioritization is an effective means for data reduction. By ranking genes in terms of their relevance to a disease, and with an appropriate thresholidng, a select set of genes can be generated by gene prioritization. Time and cost considerations in disease research usually favor a reduced gene set which enables more focused research and facilitates more effective use of the limited resources.

Over the years, many methods have been proposed for this purpose, with molecular biologists usually favoring those that focus on the statistical analysis and consequent ranking of lists of genes from the output data of high-throughput experiments. Thus, significance analysis of microarrays (SAM), analysis of variance (ANOVA), empirical Bayes t-statistic, between group analysis (BGA), and other methods are used with the help of biostatisticians, and are sometimes provided with commonly used commercial and open-source bioinformatics tools such as Illumina's Genome Studio or caBIG's geWorkbench. Knowledge about the significant genes is sometimes provided by the tools or by sought out separately by researchers only as a way to annotate the genes, but is not used to prioritize them. Researchers have to pick and choose using their own intuition and experience.

Integrating multiple kinds of heterogeneous data and knowledge sources is a challenging problem for which formulation of a flexible and general approach is sought. A number of approaches employing protein interaction as a single knowledge source[8,19,22] have been published. Other systems, the best of which are Kohler *et al*'s[12] GeneWanderer and Aerts *et al*'s[2] Endeavour, use heterogeneous knowledge and data sources. GeneWanderer was shown to outperform many existing network-based gene prioritization algorithms.[31] It assumes a set of seed genes known to be disease genes as input and proposes a method where nodes in a protein interaction network are randomly visited (restarting the walk randomly during the process), ranking candidates with respect to their relevance to the given seed gene set. Aerts *et al* proposed Endeavour, a similarity-based approach that uses heterogeneous data to calculate the similarity between a set of candidate genes and a set of 'training' or seed genes. It was successfully employed in various biological studies. Candidate genes are ranked independently by using a selection of knowledge sources. An N-dimensional order statistics is used for combining the multiple rankings. de Bie *et al*[6] used similarity measures and kernels corresponding to each data source and integrated rankings from multiple sources by weighting kernels. Li *et al*[13] employed GO-derived gene similarity networks and a PPI network, applied random walk with restart to each and combined the multiple rankings by using a discounted rating system.

Albeit intended on a genomic scale, most of the currently available knowledge sources and experimental platforms have rather low sensitivity. For example, current PPI databases are estimated to capture only 10% of true interactions.[9] Often times data and knowledge sources are orthogonal, with pieces of information absent in one being provided in another. Thus, distinct sources tend to have a complementary nature such that a holistic perspective on genes can be gained by appropriately complementing and integrating distinct sources. Existing approaches for multiple sources take data and knowledge sources separately, whereby their complementarity can be easily lost. Also, many involve rather high computational cost or assume specific types of data and limit the applicability to other data types.

Given a known group of genes associated with a specific disease as a "seed", we hypothesized that the degree of association of a candidate gene with the seed genes signifies its relevance to the disease. All knowledge about the genes was represented in a single network, which can be appropriately configured based on types of data, availability and reliability. Here, we used protein-protein interactions (PPIs), Gene Ontology annotations, gene expression data and SNP data from a Genome-Wide Association Study for validating our approach. Application to a large number of diseases of distinct kinds showed uniform performance level and hence no bias for particular kinds of diseases. We report the results of this general experiment, as well as a more extensive evaluation using genes related to Alzheimer's Disease (AD).

2. Material and methods

PPI and Gene Ontology associations were used as knowledge sources in building an integrated gene-gene association network used for gene prioritization. This is what we called the base scheme (BS) for purposes of evaluation. Additionally, gene expression and GWAS data were used as empirical data sources and incorporated in the prioritization by adding a value (level of significance) to each node in the integrated network above. This is what we called the incorporated scheme (IS). In the following subsections, we outline how the associations for each component of the network are defined and integrated, and present two experimental setups (the base scheme and the incorporated scheme) to validate the approach.

2.1. *Establishing Gene Ontology associations*

The Gene Ontology (GO) consists of a directed graph of terms organized under three main categories: biological process, cellular component and molecular function. Genes are annotated with those terms that apply to them. Resnik[21] defined similarity between two GO terms t_0, t_1 under the same category as

$$sim(t_0, t_1) = \mathrm{IC}_{ms}(t_0, t_1) = \max \mathrm{IC}(t_p) \tag{1}$$

where $t_p \in parents(t_0, t_1)$, and IC(t) is the information content of term t which is defined as $\mathrm{IC}(t) = -logP(t)$ with $P(t)$ being the probability of occurrence of the term across a genome.

Couto *et al*[5] defined similarity between two genes g_0, g_1 with respective terms $t_a \in \{terms(g_0)\}$ and $t_b \in \{terms(g_1)\}$ as

$$sim(g_0, g_1) = \max_{a,b} sim(t_a, t_b)\mathrm{IC}(t_a)\mathrm{IC}(t_b) \tag{2}$$

Term similarity is a normalized quantity ranging between 0 and 1. We used GO annotations[11] of the human genome, which included a total of 14,685 genes annotated with biological process terms, with a total term occurrence count of 60,792 for an average of 4.140 terms per gene. In establishing a gene-gene association based on GO annotations, we varied the similarity threshold from 0.30 to 0.70 in increments of 0.10 to retain gene pair similarity only above or equal to the given threshold, obtaining five nested sets of associations.

2.2. *Protein-protein interactions*

Three protein interaction databases were employed, to match those used by Kohler *et al* in [12] and allow a fair comparison: HPRD,[20] STRING[16] and NCBI yeast protein interactions. HPRD is a manually annotated protein interaction data set: the one we used had 2,125 homomeric interactions and 36,631 heteromeric interactions. The STRING database contains information from four sources (genomic context, high-throughput experiments, coexpression, and derived from text), including direct (physical) and indirect (functional) associations. We used version 8.3, which covers 2.6 million proteins from 630 organisms. Each interaction in STRING is assigned a significance score (non-linear) in the range between 150 and 1000. In addition, known protein interactions in yeast were downloaded from NCBI.[17] Each yeast protein was mapped to a human ortholog using InParanoid.[18] Only interactions involving protein pairs that have a 100% match score to human orthologs were retained (a total of 39,665).

Interacting proteins were each mapped to coding genes and then a set of interacting genes were obtained. Some common interactions in the databases derive from single experimental evidence and hence there exists a degree of duplicity among the three databases. The three PPI networks were combined into a single network by counting edges only once irrespective of their duplicity:

$$\{e'(g_1, g_2)\} = \cup\{e_{N_i}(g_1, g_2)\}, 1 \leq i \leq N \tag{3}$$

with $e'(g_1, g_2)$ being the edge between nodes g_1 and g_2 in the combined network and N being the total number of PPI networks. Five distinct sets of associations were obtained by using nested sets of interactions with different STRING significance score thresholds (300, 400, 500, 600 and 700).

2.3. *Gene expression*

For this paper, we used microarray expression data sets by Webster *et al*,[23] comprised of control and AD case samples. Genes showing significantly distinct levels between normal and disease cells were identified by using differential expression analysis. Wilcoxon rank sum test was applied to expression levels from the two groups of samples and a P-value of each gene's differential expression was obtained. The P-value threshold was set to 0.05. The significance of a gene G, $S(G)$, from differential expression was calculated as:

$$S(G) = -log(\text{P-value}) \tag{4}$$

2.4. *Genome-wide association study*

SNP genotyping is performed on genomes from normal and disease samples. Certain SNP may show distinct presence in one group vs the other e.g., allele A constitutes 80% of disease samples at a certain locus while it constitutes 30% in normal samples. A P-value can be calculated for each SNP and hence for a corresponding gene if the locus of the SNP is within or close to the gene, which would imply the gene is strongly relevant to a specific disease. If a SNP is too distant from genes (more than 20kb away upstream or 5kb downstream), then it was not included in our experiments. Similar to expression data, disease significance P-values were calculated and assigned to genes by using Eq 4.

2.5. *Network representation*

To construct the networks used for the base (BS) and incorporated (IS) schemes, the PPI and GO associations described above were used as edges, with genes mapped to nodes. If more than one knowledge source associated two genes g_1 and g_2, then the edge is weighted according to the multiplicity of the number of associating sources. Thus, if N sources were associating the two genes then weight$(e(g_1, g_2)) = N$.

Gene g may be completely missing or may not have a P-value above a threshold in the outcome of some experimental data, and have P-values above thresholds only in N_e number of effective sources. Given a significance $S_i(g)$ from empirical data source i $(1 \leq i \leq N_e)$ for a given disease, gene g's overall empirical significance is calculated as

$$S(g) = \sum_{i=1}^{N_e} S_i(g) \tag{5}$$

That is, the sum of all significance values is assigned as a combined significance score for the gene (its aggregate experimental significance).

2.6. *Base scheme*

Given a set of training seed genes $\{S_i\}$, candidate gene C was scored as follows:

$$score(C,S) = \sum^{\forall S} e(C,S_i) \tag{6}$$

where $e(C,S_i)$ is a non-zero value if an edge exists between C and S and 0 otherwise. Either only the edge presence between C and S can be recognized for scoring, or its weight from the aggregate network can be considered, i.e.,

$$e(C,S)_{BS1} = \mathbf{1}\{e(C,S)\} \tag{7a}$$

$$e(C,S)_{BS2} = \text{weight}(e(C,S)) \tag{7b}$$

with **1** being an indicator function corresponding to edge presence. If only the presence of an edge is considered, then Eq 7a is used together with Eq 6. This will be referred to as base scheme 1 (BS1). If edge weight is considered, then Eqs 7b and 6 are used which will be referred to as base scheme 2 (BS2). Candidate genes are ranked according to their scores.

2.7. *Empirical data incorporation scheme*

The network topology used in the empirical data incorporation scheme (IS) is the same as the one in the base scheme. Candidate gene C can have an edge to jth seed gene T_j of an overall empirical significance $S(T_j)$. Then T_j's contribution to the score of C is calculated as

$$e(C,T_j) + kS(T_j) \tag{8}$$

where k is a scaling factor, the value of which is to be set according to data reliability. If an edge does not exist between them, then T_j's contribution is 0. The contribution from each training gene T_j, $1 \leq j \leq |T|$, in the training set to candidate gene C is added up for its combined score:

$$score(C) = k_1 S(C) + \sum_{j=1}^{|T|} [e(C,T_j) + k_2 S(T_j)] \tag{9}$$

where k_1 and k_2 are scaling factors and $|T|$ the total number of training genes. The ranking of the candidate genes corresponds to the combined scores of the candidate genes.

2.8. *Validation*

The disease gene sets from Kohler *et al*[12] were used. Leave one out testing was performed by holding out one disease gene as a true test gene to be (ideally) recalled from the disease gene set by taking the remainder genes as a training gene set, and this was repeated for each gene over all disease gene sets. Sensitivity and specificity values were calculated as defined in (2). Specifically, ranking results were aggregated and the number of true test genes above a given ranking threshold was counted as true positives. The number of test genes below the threshold, non-test genes below the threshold and non-test genes above the threshold were respectively counted as false negatives, true negatives and false positives. As frequently done in literature, a narrowed-down set of genes (e.g., 100) in closest proximity to the true test gene along its chromosome is given as a candidate set. We also show the ranking obtained over all genes in the genome.

Current knowledge sources may involve degrees of incompleteness and incorrectness. This would correspond to false positive and negative edges in networks. Facing this, we randomly perturbed 10% of network edges by randomly reassigning them in an experiment. Eight such instances of randomly perturbed networks were generated and the base scheme was applied to each of them.

Table 1. AD gene prioritization

	Base		Endeavour	GeneWanderer		Incorp.	
Gene	Rank	Rk100	Rk100	Rank	Rk100	Rank	Rk100
APOC1	93	2	5	275	7	1	1
APOE	1	1	4	17	4	1	1
APP	382	1	4	264	1	156	1
CLU	7	1	9	102	2	17	1
CR1	437	2	44	1158	3	352	2
GAB2	202	1	31	496	3	452	2
MSRA	-	100	24	6511	11	-	100
PICALM	444	1	8	978	3	95	1
PSEN1	1	1	2	14	1	1	1
PSEN2	7	1	4	84	1	25	1
PVRL2	7	1	47	67	4	15	2
RELN	439	1	43	957	5	413	1
TOMM40	1261	10	86	3319	18	34	2

3. Results

Genes implicated in AD were collected from the literature (1, 10, 15, 14, 24, 25) (Table 1). For comparison of performance, gene prioritization based on random walk with restart (RWR) as described by Kohler *et al* (12) was implemented. In RWR, nodes are navigated in a random fashion starting from a gene randomly selected from a given set of seed genes. Gene ranking in RWR is according to the visit frequency at the conclusion of iteration following a convergence criteria. In addition, Endeavour^2 was downloaded from the authors' website. It randomly selects 99 genes other than true test gene to produce a 100 gene candidate set together with the test gene. Even though the candidate gene sets used for Endeavour are different from the ones used for base scheme and RWR, we reasoned the set size is sufficiently large from a statistical sense to facilitate sound comparisons and show the rankings under the column name of Rk100.

The base gene prioritization scheme was applied to the AD gene set. The same set was also used for Endeavour and GeneWanderer. When gene APOC1 was left out as a true test gene to be recalled and the other genes were used as a training seed gene set (row 1 in Table 1), there were 92 other genes from the human genome which ranked more significantly (column Base-Rank in Table 1). When the candidate gene set was reduced to the 100 genes of closest proximity (Loc100 set), APOC1 ranked 2nd highest (column Base-Rk100). Endeavour's ranking of the gene was 5th out of 100 genes and RWR's ranking was 275th among

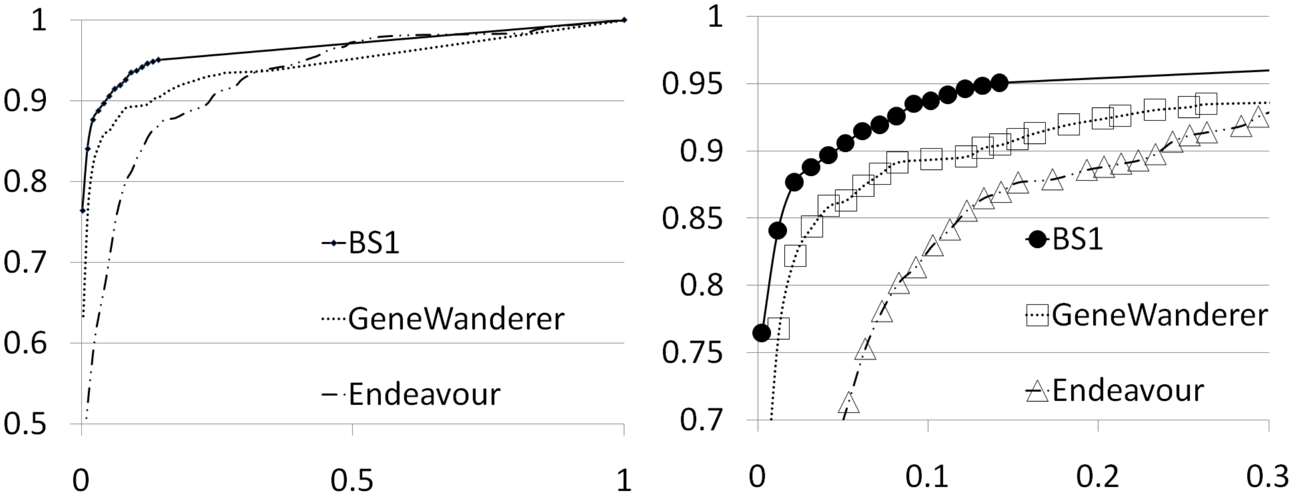

Fig. 1. ROC curves of specificity vs. 1-sensitivity (a) Base scheme has a larger AUC than Endeavour and RWR. (b) Close-up of higher sensitivity range

Table 3. AUC difference between base scheme 1 and base scheme 2; BS1 - BS2

GO $\backslash PPI$	300	400	500	600	700
30	-5.603	-12.101	-12.751	-14.561	-14.451
40	-0.265	+5.767	+9.010	+8.708	+4.035
50	+1.815	+5.567	-0.048	+6.935	+7.971
60	+0.986	+1.065	-0.157	+0.390	+0.000
70	+0.532	+0.464	+0.000	+0.165	+0.398

Units in 10^{-4}

entire genome and 7th among Loc100 genes. Each subsequent row can be read in a similar fashion. Thus, the base gene prioritization scheme ranked the AD set genes more significantly than RWR (signed rank test P-value=6.836×10^{-3}.) and Endeavour (P-value= 2.148×10^{-2}).

In order to assess the applicability of the base scheme (BS1) to other diseases besides AD, we applied it to disease gene set of Li *et al*[13] (Li10) which was derived from the complete Kohler *et al* set. It includes 36 diseases and genes implicated therein. The receiver operating characteristic (ROC) curve of the base scheme BS1 is shown in Fig 1 together with the curves of Endeavour and RWR for the same set. AUC value of the base scheme was 0.9655 while, for Endeavour and RWR, the AUC values respectively were 0.9287 and 0.9442. The reasonable AUC value means the base scheme is applicable to other diseases in general as well. Base schemes 1 and 2 were compared over the Li10 set and their AUC values showed a marginal difference possibly suggesting edge multiplicity does not greatly contribute in distinguishing true test gene from the other candidate genes (Table 3). Subsequently, we used only base scheme 1 and will refer to it as the base scheme.

Knowledge sources such as PPI or GO may entail some levels of false and missing annotations. In order to evaluate the influence of such noise on the performance of the base scheme, 10% of the edges in the combined network were randomly rewired. Eight such instances of the perturbed networks were generated, and then the base scheme was applied. In all cases, AUC values decreased by small degrees, but consistently from that of the un-perturbed network; average AUC value was 0.96070 and standard deviation 0.00223 (Fig. 2 and Table 4). Only a slight degradation in the AUC of the perturbed network means our base scheme is robust with respect to a noticeable amount of possible mis-curations in the knowledge sources and corresponding noise in the network.

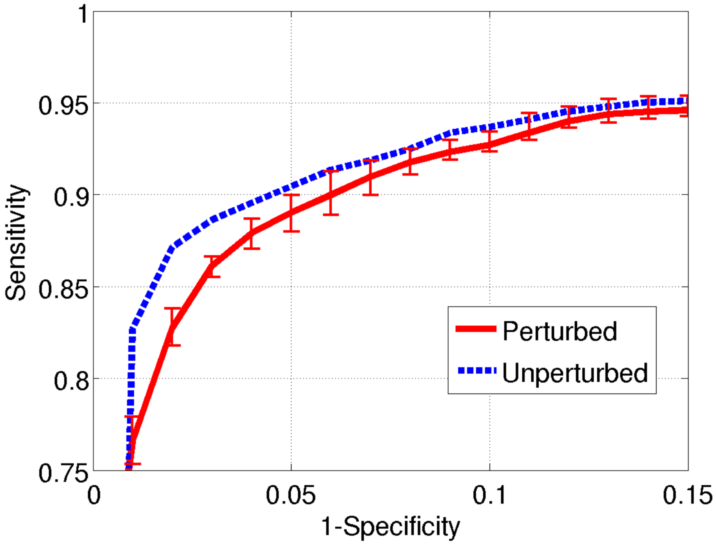

Fig. 2. ROC's from perturbed and unperturbed networks

Table 4. AUC values from perturbed networks

Instance	1	2	3	4	5	6	7	8	Average	St.dev.
AUC	0.96119	0.95875	0.96216	0.95869	0.96533	0.95993	0.96101	0.95908	0.96070	0.00223

Table 5. AUC values from application to different disease categories

Type	Cancer	Monogenic	Polygenic	Average
BS1	0.95727	0.96677	0.98025	0.96810
RWR	0.95414	0.90535	0.94978	0.95890
Endeavour	0.87947	0.94471	0.88191	0.90203

Diseases were categorized as belonging to one of three types by Kohler *et al*: cancer, monogenic and polygenic. Cancer and polygenic categories each included 12 diseases, and monogenic 86 diseases. We chose the 6 largest disease gene sets from each category to form categories balanced in count and applied the base scheme, Endeavour and RWR to each. AUC values were similar across disease categories (Table 5), thus suggesting that the base scheme is not biased to a particular category of diseases. Higher AUC values were produced by BS throughout the different categories.

The contribution of individual knowledge sources was assessed by using either PPI or GO associations alone and by comparing the resulting AUC values with the ones obtained with aggregate sources. Specifically, 5 sets of GO associations were produced with distinct thresholds of 0.30 to 0.70 in increments of 0.10, and also 5 sets of PPIs with thresholds 300 to 700 in increments of 100. A total of 35 networks resulted; 5 with only GO associations as edges, 5 PPI only, and aggregate networks in 25 different combinations of GO and PPI thresholds. The Base Scheme was applied to the Li10 set for each of the networks. The AUC value monotonically increased as GO or PPI thresholds were lowered (resulting in more network edges) (Figs 3(a), 3(b)). The highest AUC value was produced with the aggregate network of least stringent threshold combination (PPI 300 and GO 0.30).

The PPI network alone shows reasonable AUC values under varying thresholds (bottom-most curve of Fig. 3(a)). Aggregation with GO network consistently improves the AUC values. However, GO networks alone show rather low AUC values especially at high thresholds, but aggregation with PPIs, even at the highest threshold, drastically improves AUC values. Clearly, aggregation of networks from distinct knowledge sources is an effective way of comprehensively utilizing their respective information content, and our base scheme

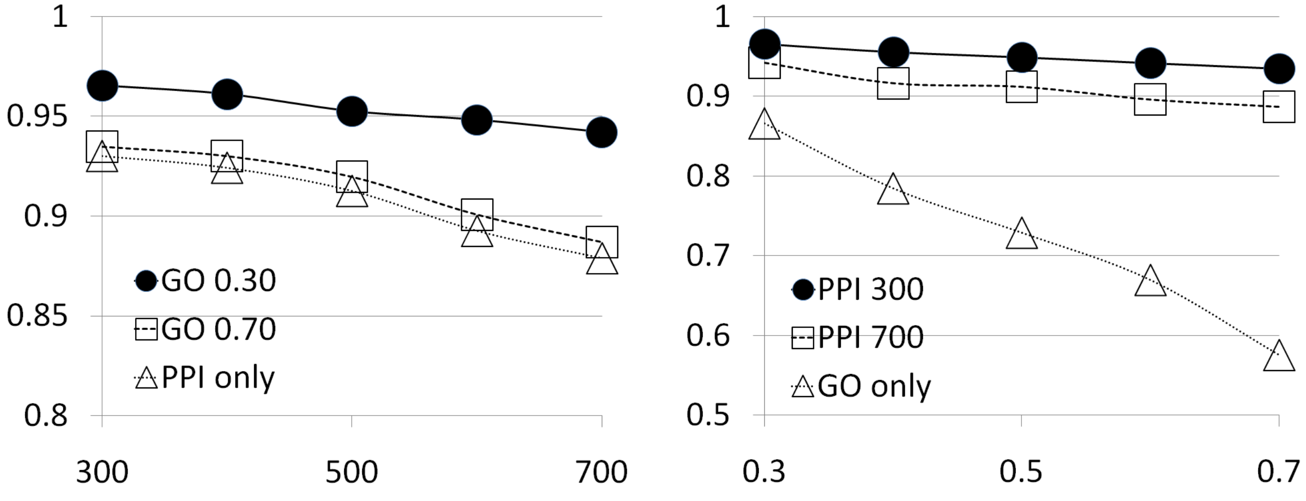

Fig. 3. AUC values from different knowledge source combinations (a) AUC vs. PPI threshold (b) AUC vs. GO threshold

indeed utilizes the higher information content.

3.1. *Incorporation of empirical data*

Alzheimer's Disease GWAS and differential expression data were incorporated in the gene prioritization process (Table 1 column Incorp.) as explained in the incorporated scheme. Improvement over the base scheme was rather marginal (P-value=0.1934). This may be attributable to a rather low reproducibility of significant genes between experiments, especially expression data.[7,27,28] A number of approaches have been suggested for an appropriate interpretation and extraction of useful information from experimental data including shifting of focus towards groups of genes rather than on individual genes.[29] A new formulation of the incorporated scheme is left as a future work, which considers the difference in nature of experimental data.

4. Discussion and conclusion

Two different knowledge sources were each represented in a network and unified in a model that allows for additional sources to be added in a similar fashion. Each independent knowledge source is likely incomplete and missing many associations between genes.[9] The proposed knowledge integration method (base scheme) complements incomplete knowledge sources to produce a more comprehensive view of genes. For example, among well known AD genes, APOE has edges to genes APP, CLU, PSEN1 and PSEN2 in PPI network and lacks an edge to PICALM (Fig 4). The GO network does not have the APOE-APP edge but contains the APOE-PICALM edge. We compared our proposed method to two of the best multi-source gene prioritization algorithms. Endeavour utilizes knowledge sources separately and tended to produce the lowest AUC values among the compared algorithms. The method proposed here effectively integrates individual knowledge sources to overcome the incompleteness of each.

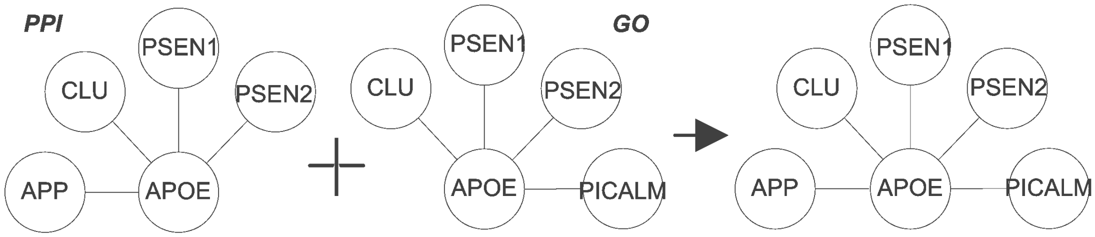

Fig. 4. Network aggregation

The base scheme alone showed better performance than Endeavour and RWR. Rankings based on combined networks were consistently better than rankings based on individual networks. There is a degree of overlap between the two knowledge sources (PPI and GO), since the same information from literature is frequently used to annotate genes. Still there is information content in one source which is not captured in the other. The edge formation by similarity criterion in the GO network can associate genes that are highly related in pathways or from biological perspectives which do not directly interact through their protein products and hence is missed in a PPI network. The described schemes rely on the association between genes to infer disease genes from known genes. The effectiveness of this approach was shown through a series of experiments. The information from knowledge sources and experimental data vary in reliability, degree of curation and level of acceptance. For example, many protein interactions have been verified over time and are well accepted, while high throughput interaction data tends to involve a high rate of false positives.

Our Gene ontology annotation of genes reflects a relatively high level of verification and curation. On the other hand, experimental data is subject to a high level of noise and variance and has not been extensively and thoroughly verified. Hence a network was not directly formed from experimental evidence at this stage, and only node significance was adjusted in accordance with the experimental significance. Our schemes are robust against false positives and missing knowledge as shown in the perturbation experiment. Future work will be directed at incorporating empirical data from experiments in a way that is more consistent with the way knowledge sources are used. While particular knowledge sources and experimental data were used for illustration, the described schemes are sufficiently general to be used with other data types as well. After the preparation of our manuscript, a gene prioritization method[30] was noted for its use of diverse data with a Bayesian approach. While a readily accessible version of their algorithm was unavailable, it will be interesting to perform a comparative study involving it.

5. Acknowledgement

Authors are grateful to Dr. Matthew Huentelman for AD GWAS data. This work was partially funded by the NIH/NIA under grant P30 AG019610-09.

References

1. R. Abraham *et al.* A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. *BMC Med Genomics*, 1:44 (2008).
2. S. Aerts *et al.* Gene prioritization through genomic data fusion. *Nature Biotechnology*, **24(5)**, 537-44 (2006).
3. M. Ashburner *et al.* Gene ontology: tool for the unification of biology. *Nat. Genet*, **25**, 25-29 (2000).
4. J. Chen, B. J. Aronow and A. G. Jegga. Disease candidate gene identification and prioritization using protein interaction networks. *BMC bioinformatics*, **10** (2008).
5. Couto *et al.* Implementation of a functional similarity measure between gene-products. *Technical report DI/FCUL TR 03-29* (2003).
6. T. de Bie *et al.* Kernel-based data fusion for gene prioritization. *Bioinformatics*, **23(13)**, i25-32 (2007).
7. L. Ein-Dor *et al.* Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171-178 (2005).
8. G. Gonzalez *et al.* Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. *Pacific symposium on biocomputing* (2007).
9. G.T. Hart *et al.* How complete are current yeast and human protein-interaction networks? *Genome Biol.*, **7**, 120 (2006).
10. D. Harold *et al.* Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet*, **41(10)**, 1088-93 (2009).
11. ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/.
12. S. Kohler *et al.* Walking the interactome for prioritization of candidate disease genes. *the American journal of human genetics*, **82(4)**, 949-58 (2008).
13. Y. Li *et al.* Integration of multiple data sources to prioritize candidate genes using discounted rating system. *BMC Bioinformatics*, **11** Suppl 1:S20 (2010).
14. H. Li *et al.* Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol*, **65(1)**, 45-53 (2008).
15. J. Lambert *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet*, **41(10)**, 1094-9 (2009).
16. C. von Mering *et al.* String 7 - recent developments in he integration and prediction of protein interactions. *Nucleic acids research*, **35**, D358-62 (2003).
17. NCBI ftp://ftp.ncbi.nih.gov/gene/GeneRIF/interactions.gz (2010).
18. G. Ostlund *et al.* InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196-203 (2010).
19. M. Oti *et al.* Predicting disease genes using protein-protein interactions. *Journal of medical genetics*, **43(8)**, 691-8 (2006).
20. T.S.K. Prasad *et al.* Human protein reference database - 2009 update. *Nucleic acids research*, **37**, D767-72 (2009).
21. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *IJCAI* (1995).
22. A. Sharma *et al.* Gene prioritization in type 2 diabetes using domain interactions and network analysis. *BMC genomics*, **11** :84 (2010).
23. J.A. Webster *et al.* Genetic control of human brain transcript expression in Alzheimer disease. *American journal of human genetics*, **84(4)**, 445-58 (2009).
24. E. Reiman *et al.* GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. *Neuron*, **54(5)**, 713-20 (2007).
25. P. Kramer *et al.* Alzheimer disease pathology in cognitively healthy elderly: A genome-wide study. *Neurobiol Aging*, May 6 (2010).

26. S. Frantz. An array of problems. *Nat. Rev. Drug Discov.*, **4**, 362-363 (2005).
27. G. L. Miklos and R. Maleszka. Microarray reality checks in the context of a complex disease. *Nat. Biotechnol.*, **22**, 615-621 (2004).
28. E. Marshall. Getting the noise out of gene arrays. *Science*, **306**, 630-631 (2004).
29. D. Yang. Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. *Bioinformatics*, **24**, 265-271 (2008).
30. B. Linghu *et al.* Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol. **10** (9):R91 (2009).
31. S. Navlakha and C. Kingsford. The power of protein interaction networks for associating genes with diseases, *Bioinformatics*, **26(8)**, 1057-63 (2010).