Candidate Gene Prioritization Using Network Based Probabilistic Models

Shuguang Wang, MSc, Milos Hauskrecht, PhD, Shyam Visweswaran, MD, PhD
University of Pittsburgh, Pittsburgh, PA

Abstract. *We present a new gene prioritization method that learns a probabilistic knowledge model from a knowledge base and free text research documents and exploits it to prioritize candidate genes. The knowledge model is represented by a network of associations among the domain entities (e.g., genes) and is extracted from a domain knowledge base (e.g., protein-protein interaction database) or a corpus of text documents (e.g., PubMed documents). This knowledge model is then used to perform probabilistic inferences and applied to the task of candidate gene prioritization. We evaluate our new method on five diseases and show that it outperforms a recently described network based method for candidate gene prioritization.*

Introduction

A primary goal of high-throughput genomic studies is to identify candidate genes that are related to a specific disease or a biological condition from many thousands of genes in the genome; these are then validated by more labor-intensive low-throughput methods. Typically, the high-throughput studies produce large sets of candidate genes, and methods that further prioritize the genes in these candidate sets are needed. Thus, numerous computational methods for candidate gene prioritization have been developed.

Many gene prioritization methods rank candidate genes based on the similarity between the candidate genes and the genes known to be associated with the disease or the biological condition of interest. Similarity between genes is derived from known information about genes such as gene function derived from functional annotations [1]. Such approaches are limited by the low annotation coverage due to the lack of information and bias in the annotation. Though more than a thousand human disease genes have been identified and documented, a large number of them are still not characterized according to function. In addition, there is bias in the existing annotations as genes with little information are sparsely annotated.

To overcome the limitations of functional annotation, researchers have explored the use of alternative or additional sources of information for prioritizing genes. Such information include sequence data [2], combination of sequence data and biological annotation information [2], functional information, and gene expression data [3]. A recent study showed that incorporation of mouse phenotype information was helpful in prioritization of human genes [7]. Chen et al. [6,8] have explored the use of link analysis methods on gene networks derived from protein-protein interactions to identify and prioritize candidate genes for a disease of interest. In particular, Chen et al. used Hyperlink-Induced Topic Search (HITS) [11] and PageRank [12] as the link analysis methods; these methods have been extensively used for analyzing social, Web and citation networks.

In this paper, we present an improved link analysis method and apply it to the task of gene prioritization. Our method uses more accurate distributional assumptions than those implemented in the HITS [11] and PageRank methods [12]. We evaluate our method on prioritizing genes related to five diseases, and compare the performance of our link analysis method to that of PagerRank. In addition, we explore the combination of a curated knowledge base and knowledge extracted from free text research abstracts to develop the gene networks on which the link analysis methods can be applied. Previous studies have shown that knowledge bases (e.g., protein-protein interaction knowledge bases) are very useful for candidate gene prioritization. However, such knowledge bases are likely to be incomplete and can potentially be enhanced with knowledge mined from the research literature. We evaluate our improved link analysis method on networks obtained from (1) a knowledge base only, and, (2) from a combination of a knowledge base and the research literature.

The rest of this paper is organized as follows. In the next section, we describe the structure of the network, our link analysis method, and the construction of the network from different sources of knowledge. Finally, we present and analyze the results of the evaluation of our method on five diseases.

Methods

Knowledge in a scientific domain can be represented as a rich network of relations among the domain entities. Based on this view, we build a knowledge model that is represented by a graph (network)

structure, where nodes represent domain entities and arcs between nodes represent pair-wise relations among the domain entities. For the knowledge model we consider in this paper, the nodes represent genes (or the corresponding proteins) and the arcs represent association relations that abstract a variety of relations that may exist among the genes. We refer to this knowledge model as an association network.

A. Probabilistic Knowledge Model

The arcs in an association network represent association relations (or associations) that represent a variety of relations among domain entities. Such a simple representation has several advantages. One advantage is that the different types of relations that may exist among the domain entities are treated uniformly, which simplifies the analysis. Another advantage is that the association relations are relatively easy to mine both from structured databases and from free text.

We use the association network to infer the relevance among the domain entities. Studies [13,14,15] have shown that relevance can be inferred by analyzing the interconnectedness of nodes in the association network using link analysis methods. The intuition is that domain entities that are semantically interconnected in terms of their roles or functions should be considered more relevant to each other, and this interconnectedness can be inferred from the topology of an association network that is derived from pair-wise relations that are contained in the databases (e.g., protein-protein interactions) and in free text (e.g., co-occurrence).

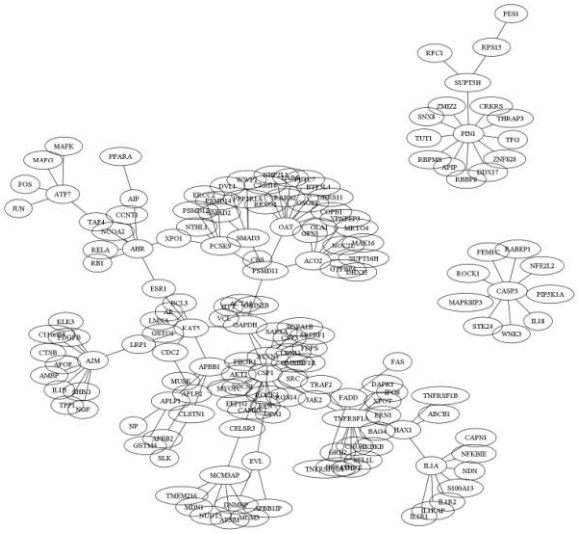

Figure 1 Section of an association network constructed from OPHID.

Figure 1 shows a section of the association network constructed from the Online Predicted Human Interaction Database (OPHID) [5], which is an online database of curated protein-protein interactions. Each entry in the OPHID database consists of two proteins that are known to interact. In the network shown in the figure, a node represents a gene (for the corresponding protein in OPHID), and an arc represents an interaction between the two genes it connects. From the figure, it can be seen that there are well-defined clusters of genes in the network, and we make the assumption that genes that belong to a cluster are more likely to be related to each other than genes that belong to different clusters. We next describe our link analysis method that takes into account the clustering of nodes in an association network. In contrast, existing method such as PageRank and HITS do use information about clustering

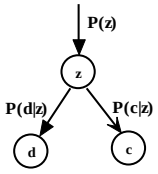

Figure 2 Graphical representation of the PHITS model.

Our link analysis method is a modification of a link analysis method called PHITS (probabilistic version of HITS) [9] to analyze the mutual connectivity of domain entities in an association network and derive a probabilistic model that reflects the mutual relevance among domain entities. Figure 2 shows a graphical representation of the original PHITS model that was developed for inference in a co-citation network where a node represents a document and an arc between two documents represent the association that one document cites the other. In the figure,, variable *d* represents documents, *z* is a latent (and not observed) variable and variable *c* represents citations. The model defines the joint probability of a document *d* and a citation *c*, $P(d, c)$ as $\sum_z P(z)P(c|z)P(z|d)$. In a co-citation network, the latent variable *z* can be interpreted as topics into which the documents cluster. In our modification of PHITS that is applied to a gene association network, both variables *d* and *c* represent genes (and hence we represent both a common variable *e*) and variable *z* is interpreted as gene clusters. The PHITS model assumes that the data follow a multinomial distribution. This is a more suitable distribution for modeling association networks than the normal distribution assumptions made by PageRank and HITS [11].

B. Probabilistic Inference

Given an association network and the PHITS model, one useful inference task that can be done is the calculation of the probability of seeing an unobserved entity *e* (e.g., a candidate disease gene) given a set of observed entities (e.g., genes already know to be associated with a disease) $o_1, o_2, \ldots, o_k$. In the model, entities are treated as alternatives and the conditional probability is defined by the following distribution:

$P(e=b_1|o_1, o_2, \ldots, o_k)$

$P(e=b_2|o_1, o_2, \ldots, o_k)$

…

$P(e=b_n|o_1, o_2, \ldots, o_k)$

where *e* is a random variable and $b_1, b_2, \ldots, b_n$ are its values that denote individual domain entities. Intuitively, this conditional distribution defines the probability of seeing an unobserved entity after we have observed some evidence entities. To calculate the conditional probability of *e*, we use the following approximation:

$$P(e| o_1, o_2, \ldots, o_k) \sim \sum_z P(e|z) \prod_{j \in [1,k]} P(z|o_j)$$

where $o_1, o_2, \ldots, o_k$ are observed (known) entities and *z* is the latent factor. More details on our probabilistic inference method can be found in [15].

In this paper, the set of evidence entities is represented by a set of genes known to be associated with the disease of interest (seed genes), and we infer the probability of observing a candidate (non-seed) gene being associated with the disease given the seed genes.

C. Enhancing the Model with Associations Extracted from Text

One shortcoming of existing knowledge bases is that they are incomplete. We believe that we can obtain additional knowledge by mining associations from research articles. Such mining of domain knowledge from free text documents have been shown to be beneficial for the tasks of document search and document retrieval [13,14,15]. Here, we extract associations among pairs of genes from abstracts of documents indexed by PubMed. We enhance the association network derived from OPHID with these associations extracted from text and apply our modified link analysis method to the new association network.

To extract associations from the text in abstracts, we first identify the genes using a dictionary lookup approach presented in [13] and then identify associations based on the co-occurrence of genes. If two genes occur in the same abstract, we add an arc between them in the association network to indicate that there is an association between them. However, co-occurrence of a pair of genes in an abstract may not represent an interaction among them. Due to this uncertainty, we weight the associations extracted from abstracts with 0.5, and weight the associations obtained from OPHID with 1. The selection of 0.5 was arbitrary; we have not explored how to further optimize the weights for the associations extracted from text.

Evaluation

We constructed probabilistic knowledge models for five diseases, and utilized them to prioritize candidate genes for those diseases given a known set of genes associated with each disease. The diseases included Alzheimer's disease, autism, Grave's disease, migraine, and systemic scleroderma. We call our approach the Knowledge Model Ranking (KM Ranking) method, and compare it to a network based method that has been recently shown to be useful for the gene prioritization task, namely, PageRank with Priors [6].

For each disease we constructed two knowledge models. The first knowledge model was extracted from OPHID and the same knowledge model was used for all diseases. OPHID contains more than 40,000 curated protein-protein interactions involving approximately 9,000 human proteins. The second knowledge model consisted of the OPHID model enhanced with associations extracted from a set of abstracts obtained from PubMed that was relevant to the disease under consideration. Thus, the second knowledge model was different for each disease.

For each disease, we also obtained a set of seed genes known to be associated with the disease. For Alzheimer's disease, we obtained the seed genes from the online AlzGene database [4], and for the remaining four diseases, we obtained them from OMIM (Online Mendelian Inheritance in Man) as described in a previous study [6].

For evaluation, we used the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) as described in previous studies [1,6,7]. A ROC curve for a particular prioritization method and a specific disease was generated as follows:

1. Randomly select 20 seed genes from the seed set of the disease.

2. Remove one gene (termed the target gene) from the 20 seed genes and mix it with 99 genes chosen at random from OPHID that are not seed genes.

3. Apply a prioritization method to rank the 100 genes and record the rank of the target gene.

4. Repeat Steps 2 and 3 for each of the 20 seed genes.

5. Repeat Steps 1 though 4 for a total of 10 runs.

The above protocol generated a total of 200 (20 x 10) sets of rankings for a prioritization method and a disease. Sensitivity was calculated as the frequency at which the target genes are ranked above a particular threshold position, and specificity as the percentage of genes ranked below that threshold. For example, if target genes are ranked in the top 10% in 80% of the set of rankings, the sensitivity is 80% and specificity is (100% - 10% = 90%). The ROC curve was plotted using the sensitivity and specificity values computed from all 200 sets of rankings.

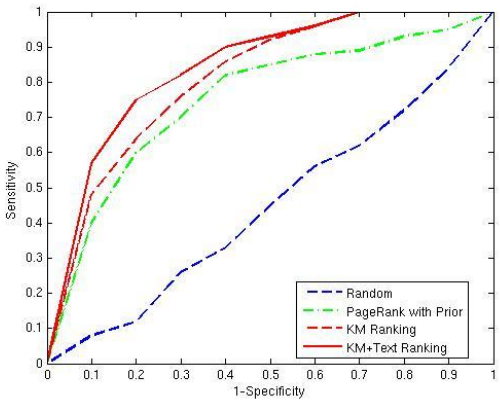

Figure 3 ROC curves for various gene prioritization methods for Alzheimer's disease.

Figure 3 plots the ROC curves of four gene prioritization methods for Alzheimer's disease. We do not show similar plots for the remaining diseases due to the space constrains, but the trends for them are similar to those seen in Alzheimer's disease. For each disease, we applied the following four methods:

1. PageRank with Priors. This method is described in [6] and we use the settings that gave the best results in [6] on the association network extracted from OPHID.

2. KM Ranking. This uses our link analysis method on the same association network extracted from OPHID as in 1.

3. KM+Text Ranking. This uses our link analysis method on the association network extracted from OPHID that has been enhanced with associations extracted from from abstracts of research documents obtained from PubMed that are relevant to the disease (e.g., Alzheimer's disease).

4. Random. For completeness, we also plot the ROC obtained by applying the KM Ranking method to seed genes that are randomly selected from OPHID.

Table 1 gives the AUCs for the first three methods (excluding Random) for all five diseases. The results demonstrate that our method performed better than PageRank with Priors on all five diseases when evaluated on the association network extracted only from OPHID. In addition, enhancing the association network with associations extracted from the abstracts of research documents improves further the performance of our method. And this improvement is significant even at high specificity (≥80%).

Table 1 AUCs for various gene prioritization methods on five diseases.

	PageRank with Priors	**KM Ranking**	**KM+Text Ranking**
Alzheimer's disease	0.795	0.848 (+6%)	0.871 (+9%)
Autism	0.810	0.852 (+5%)	0.869 (+7%)
Grave's disease	0.815	0.858 (+5%)	0.886 (+7%)
Migraine	0.805	0.840 (+5%)	0.860 (+7%)
Systemic scleroderma	0.811	0.853 (+5%)	0.870 (+8%)

Conclusions and Future Work

We have presented a new network based candidate gene prioritization method that uses a probabilistic knowledge model that can be learnt from multiple knowledge sources such as structured knowledge bases and free text research documents. Our method outperformed a recently described network based gene prioritization method by up to 9% on the AUC when evaluated on five different diseases. To the best of our knowledge this is the first study that attempts to learn probabilistic relations among genes from a structured knowledge base and free text documents and uses the knowledge so obtained to perform candidate gene prioritization.

Our knowledge model can be mined fully automatically and is flexible enough to support

various probabilistic inferences without the need to re-learn the model when the seed genes change.

In this study, we have demonstrated one simple way to combine two different sources of knowledge. In future work, we will explore several different ways of integrating multiple sources of knowledge.

References

1. E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 55(6), 2005.
2. E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard. SUSPECTS: enabling fast and effective prioritization of positional candidates. *BMC Bioinformatics*, 22(6):773–774, 2006.
3. S. Aerts, D. Lambrechts, S. Maity, P. V. Loo, B. Coessens, F. D. Smet, L.-C. Tranchevent, B. D. Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau. Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24(5):537–544, 2006.
4. L. Bertram, M. B. McQueen, K. Mullin, D. Blacker, and R. E. Tanzi. Systematic meta-analyses of Alzheimer disease genetic association studies: the alzgene database. *Nature Genetics*, 39:17–23, 2007.
5. K. R. Brown and I. Jurisica. Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082, 2005.
6. J. Chen, B. J. Aronow, and A. G. Jegga. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, 10(73), 2009.
7. J. Chen, H. Xu, B. J. Aronow, and A. G. Jegga. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, 392(8), 2007.
8. J. Y. Chen, C. Shen, and A. Y. Sivachenko. Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pacific Symposium on Biocomputing*, pages 378–378, 2006.
9. D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *Proc. 17th International Conf. on Machine Learning*, pages 167–174. Morgan Kaufmann, San Francisco, CA, July 2000.
10. T. G. O. Consortium. Gene ontology: tool for the unification of biology. *Nature Genet*, 25:25–29, 2000.
11. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *The Journal of ACM*, 46(5):604–632, 1999.
12. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. *Technical Report*, 1998.
13. S. Wang and M. Hauskrecht. Improving biomedical document retrieval using domain knowledge. In *SIGIR '08: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 785–786. ACM, 2008.
14. S. Wang and M. Hauskrecht. Improving biomedical document retrieval by mining domain knowledge. In *FLAIRS '09: Proceeding of the 22nd international FLAIRS conference*, 2009.
15. S. Wang, S. Visweswaran, and M. Hauskrecht. Document retrieval using a probabilistic knowledge model. In *KDIR '09: Proceeding of the international conference on knowledge discovery and information retrieval*, 2009.