# Gene Prioritization Using a Probabilistic Knowledge Model

Shuguang Wang
*Intelligent Systems Program*
*University of Pittsburgh*
*Pittsburgh, PA 15260*
*Email: swang@cs.pitt.edu*

Milos Hauskrecht
*Department of Computer Science*
*University of Pittsburgh*
*Pittsburgh, PA 15260*
*Email: milos@cs.pitt.edu*

Shyam Visweswaran
*Department of Biomedical Informatics*
*University of Pittsburgh*
*Pittsburgh, PA 15260*
*Email: shv3@pitt.edu*

***Abstract*—We are interested in exploiting domain knowledge for the task of candidate gene prioritization. In this paper, we present a new gene prioritization method that learns a probabilistic knowledge model and exploits it to prioritize candidate genes. The knowledge model is represented by a network of associations among domain concepts (e.g., genes) and is extracted from a domain database (e.g., protein-protein interaction database). This knowledge model is then used to perform probabilistic inferences and applied to the task of 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.**

***Keywords*-Gene Prioritization, Probabilistic Knowledge Model**

## I. Introduction

Candidate gene prioritization (or gene prioritization) is the process of ranking genes as potential candidates of being associated with a disease or biological condition of interest. High-throughput genomic studies produce data on many thousands of genes in the genome, and the candidate genes identified from such studies are 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 according to 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 one or more types of known information about genes such as functional annotations [2]. However these approaches are limited in at least two ways: considerably low annotation coverage due to the lack of information on many of the genes and bias in the annotation. Over the past few decades, more than a thousand human disease genes have been identified and documented. However, a large number of them are yet to be characterized according to function. In addition, there is bias in the existing annotations as genes with little information are sparsely annotated.

Due to the limitations of functional annotations, researchers have started to explore alternative or additional sources of information for prioritizing genes. Such information include sequence data [1], combination of sequence data and biological annotation information [2], functional information, and gene expression data [3]. In [8], Chen et al. defined a heuristic relevance scoring function on a protein-protein interaction network and applied it to prioritize candidate genes for Alzheimer's disease. More recently Chen et al. [6] explored link analysis methods to study similar interaction networks to identify and prioritize candidate genes given a list of seed genes known to be associated with the disease of interest. The link analysis methods used in this work were Hyperlink-Induced Topic Search (HITS) [11] and PageRank [12] which have been extensively applied for analyzing social, Web and citation networks. However these models assume the data to be normally distributed which is not necessarily valid [9]. Other studies [15][16] have shown the benefit of knowledge of relations among proteins and genes for the purpose of document retrieval.

In this paper, we develop an improved link analysis method that uses more appropriate distributional assumptions than those implemented in the HITS and PageRank methods. We apply it to prioritize candidate genes for a disease of interest from knowledge obtained

978-1-4244-5121-0/09/$25.00 ©2009 IEEE

from a curated protein-protein interaction database and a set of seed genes known to be associated with the disease.

Our method works by learning a protein-protein interaction network from the database, and then applying the improved link analysis method to infer new genes that are most likely to be associated with the disease given the seed genes. We demonstrate the utility of our method by prioritizing genes related to five diseases. We show that our method outperforms the existing link analysis methods that were used by Chen et al. to prioritize candidate genes [6].

This rest of this paper is organized as follows. In Section II, we describe in detail the probabilistic knowledge model that was originally developed for document retrieval [14][15]. In section III, we describe the application of this model to the candidate gene prioritization task. In section IV, we experimentally evaluate this method for prioritizing genes for five diseases and compare its performance to that of a recently described method called PageRank with Priors. We review related work in section V and in the final section we provide conclusions and some directions for future work.

## II. The Probabilistic Knowledge Model

In this section, we provide details of our knowledge model and the probabilistic inferences that it supports.

### A. *Association Network*

The knowledge in a scientific domain can be represented as a rich network of relations among the domain concepts. Based on this view, we build a knowledge model that is represented by a graph (network) structure, where nodes represent domain concepts and arcs between nodes represent pair-wise relations among associated proteins) as domain concepts and abstract a variety of relations that may exist among the genes as association relations. We refer to this knowledge model as an association network.

Association relations (or associations) that represent a variety of relations among domain concepts have several advantages. One advantage is that the different types relations that may exist are treated uniformly, which simplifies the analysis. Another advantage is that the associations are relatively easy to mine from sources such as structured databases and free text in research documents.

### B. *Probabilistic Knowledge Model*

We use the knowledge model represented as an association network to support inferences on relevance among concepts. We do so by analyzing the interconnectedness of concepts in the association network. More specifically, our hypothesis is that domain concepts are more likely to be relevant to each other if they belong to the same, well defined, and highly interconnected group of concepts. The intuition behind our method is that concepts that are semantically interconnected in terms of their roles or functions should be considered more relevant to each other. And, we expect that these semantically distinct roles and functions are embedded in the knowledge databases and are represented in our association network.

The analysis of networks is typically done using link analysis methods. For example, in Web hyperlink networks, two commonly used methods are PageRank and HITS. PageRank uses a probability distribution to represent the likelihood that a person randomly clicking on links will arrive at any particular page [12]. HITS [11] computes two scores for each document: a hub score and an authority score. Documents that have high authority scores are expected to have relevant content, whereas documents with high hub scores are expected to contain links to relevant content. However both models assume that the data are normally distributed which may be invalid [9].

We adopt a modification of PHITS (probabilistic HITS) [9] to analyze the mutual connectivity of domain concepts in an association network and derive a probabilistic model that reflects the mutual relevance among the domain concepts. We note that our application of the link analysis method at the domain concept level is very different from the typical application of such methods for analyzing the links among documents in the co-citation, social or web hyperlink networks [11][12].

Figure 1 shows a graphical representation of the original PHITS model that represents documents related by citations. Variable $d$ represents documents, $z$ is a latent variable, and variable $c$ represents citations. This model was originally used to study co-citation networks, where $z$ defines topics into which documents cluster. Using this symmetric parametrization, the model defines the joint probability of a document $d$ and a citation $c$, $P(d, c)$ as $\sum_z P(z)P(c|z)P(z|d)$.

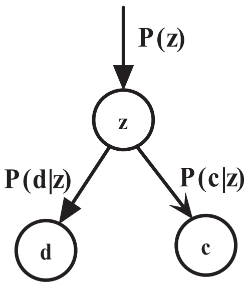

Figure 1. Graphical representation of PHITS.

In our our method, we use PHITS to analyze relations among domain concepts such as genes. Hence, both $d$ (documents) and $c$ (citations) represent the same domain concept, namely, genes. To make this distinction clear we denote domain concepts by $e$.

### C. Probabilistic Inference

Our probabilistic knowledge model defines various distributions among domain concepts and latent factors, e.g., $P(e_1, e_2)$ where $e_1$ and $e_2$ are two domain concepts [15][16]. We now describe the inference we do with this model in general terms. The basic inference task we perform is the calculation of the probability of seeing an absent (unobserved) concept $e$ given a list of observed concepts $o_1, o_2, ..., o_k$. In the model, concepts are treated as alternatives and the conditional probability is defined by the following distribution:

$$\begin{aligned} &P(e = b_1 | o_1, o_2, \ldots, o_k) \\ &P(e = b_2 | o_1, o_2, \ldots, o_k) \\ &\ldots \\ &P(e = b_n | o_1, o_2, \ldots, o_k), \end{aligned}$$

where $e$ is a random variable and $b_1, b_2, \ldots, b_n$ are its values that denote individual domain concepts. Intuitively, this conditional distribution defines the probability of seeing an absent concept after we have observed some evidence concepts.

To calculate the conditional probability of $e$, we use the following approximation:

$$\begin{aligned} &P(e|o_1, o_2, \ldots, o_k, M) \\ &\sim \sum_z P(e|z, M) \prod_{j=1}^{k} P(z|o_j, M) \qquad (1) \end{aligned}$$

where $o_1, o_2, \ldots, o_k$ are observed (known) concepts and $M$ is the knowledge model mined with PHITS. We take an approximation in this derivation due to the feature that PHITS does not represent individual citations with multiple random variables.

## III. Application to Gene Prioritization

In this section we demonstrate the application of the probabilistic knowledge model and the inference described in the previous section to the task of gene prioritization. We apply the method to prioritize candidate genes for a disease of interest given a set of genes that are known to be associated with the disease.

### A. Extraction of the Association Network

The association network, which is the knowledge model, can be constructed from many different sources [16]. In this study, we use a curated protein-protein interaction database as the source of knowledge.

Our association network was constructed from the Online Predicted Human Interaction Database (OPHID) which is an online database of protein-protein interactions [5]. Proteins in this network were mapped to their corresponding genes, and the pair-wise associations in the network then represent interactions between genes. Figure 2 presents a section of the association network constructed from OPHID. In the network, each node represents a gene, and each 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. This provides support to our hypothesis that domain concepts cluster into highly interconnected groups. We believe that domain concepts such as genes are more likely to be related to each other if they belong to the same highly interconnected group of concepts. We quantify the relevance among genes probabilistically that takes into account the group membership; we do so using PHITS.

### B. Prioritizing Candidate Genes

We now describe the application of the inference method detailed in Section II for prioritization of genes that are candidates for being associated with a disease.

Although we have incomplete information about human disease genes, such as their functional annotations, we often have a list of genes that are known to be associated with a disease of interest. For many diseases, several genes have been identified that are associated with the development of the disease. We

978-1-4244-5121-0/09/$25.00 ©2009 IEEE

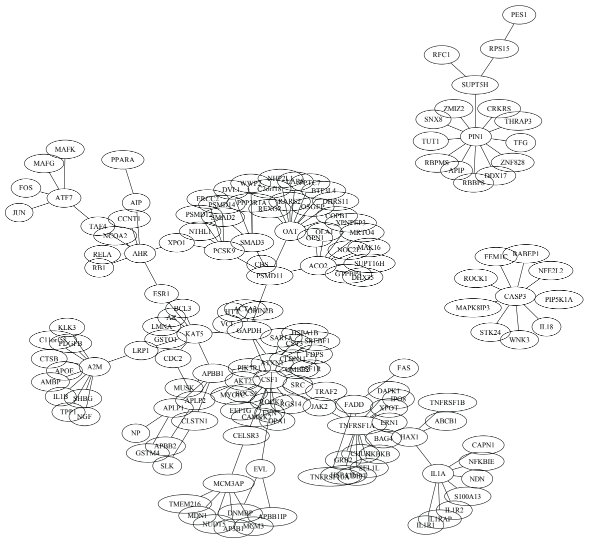

Figure 2. Section of the association network constructed from OPHID.

use this information to compute the relevance between the candidate genes and the diseases.

Equation 1 gives the probability of seeing a new concept given a list of evidence concepts. If we substitute the evidence concepts with genes known to be associated with disease (the seed genes), we can infer the probability of seeing a candidate gene given these seed genes. Then the prioritization of candidate genes is done by ranking them according to these probabilities.

### C. Other Network-Based Methods

Our network-based knowledge model method is very different from previous approaches that use networks such as those presented in [6]. There are at least three major differences:

- Our framework is more general in that it works with an arbitrary association network and this network can be extracted from many different knowledge sources.
- Our model does not make unnecessary independence assumptions about the relations among genes.
- The probabilistic nature of the model makes it very flexible and supports various types of inferences without having to train the model multiple times.

We now elaborate on the last two differences. Equation 2 defines the importance or relevance of a gene $t$ with respect to a set of $n$ seed genes $R = (r_1 ... r_n)$ as used in [6].

$$I(t|R) = \frac{1}{n}\sum_{i}^{n} I(t|r_i)) \tag{2}$$

The scores $I(t|r_i)$ are assumed to be independent and mutually exclusive. This assumption is often invalid, and hence the scores do not reflect the actual relevance of genes to the seed set. Our model is less restrictive and assumes that the genes can be dependent.

Another issue with the method used in [6] is that the training of the model must be repeated for a each new set of seed genes. Equation 3 defines the PageRank with Priors method used in [6].

$$\pi(\nu)^{(i+1)} = (1-\beta)(\sum_{\mu=1}^{d_{in}(\nu)} p(\nu|\mu)\pi^{(i)}(\mu)) + \beta p_\nu \tag{3}$$

Clearly, probability $p(\nu)$ is defined only for the seed genes, which means that the score $\pi(\nu)$ for every node $\nu$ in the network has to be recalculated every time the seed genes change. In contrast, our method constructs the model once and different sets of seed genes are handled directly by the inference procedure.

## IV. Experiments

To demonstrate the power of our method, we learned a probabilistic knowledge model, and applied it to prioritize candidate genes for five diseases: Alzheimer's disease, autism, Grave's disease, migraine, and systemic scleroderma. We call our approach the Knowledge Model Ranking (KM Ranking) method. We compared the performance of our method to that of a recently used network-based method that uses PageRank with Priors [6].

### A. Data

We learned the knowledge model from OPHID that contains more than 40,000 protein-protein interactions involving approximately 9,000 human proteins. Although our model is capable of incorporating other sources of knowledge, we used only this interaction database to conduct a fair comparison with the PageRank method.

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 used the seed genes obtained from OMIM (Online Mendelian Inheritance in Man) as described in a previous study [6].

## B. Evaluation Metric

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].

An ROC curve for a particular prioritization method and a specific disease was generated as follows:

1) Randomly select 25 seed genes for the disease.
2) Remove one gene (termed the target gene) from the 25 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 list of 100 genes and record the rank of the target gene.
4) Repeat Steps 2 and 3 for each of the 25 seed genes.
5) Repeat Steps 1 though 4 for a total of 10 times.

Sensitivity was calculated as the frequency of target genes that are ranked above a particular threshold position, and specificity as the percentage of genes ranked below the threshold. For example, if target genes are ranked in the top 10% in the list 80% of the time, the sensitivity is 80% and specificity is (1-10%=90%). The ROC curve was plotted using the sensitivity and specificity values computed from all 10 runs.

## C. Evaluation Results

Figure 3 plots the ROC curves for the KM Ranking method and the PageRank with Priors method on systemic scleroderma. For completeness, we also plot the ROC obtained by applying the KM Ranking method to randomly selected seed genes. For the PageRank with Priors method, we used the settings that gave the best results in [6]. We do not show the plots for the remaining diseases as the trends for them are similar. Table I gives the AUCs for both the methods for the five diseases.

Our method had superior performance when compared to the PageRank with Priors method and the improvement was significant even at high specificity ($\geq$0.8). In the KM Ranking method, over 90% of the time, the target gene was located in the top half of the ranked list. Based on the AUC, the KM Ranking method shows improved performance by up to 6% when compared to the PageRank with Priors method.

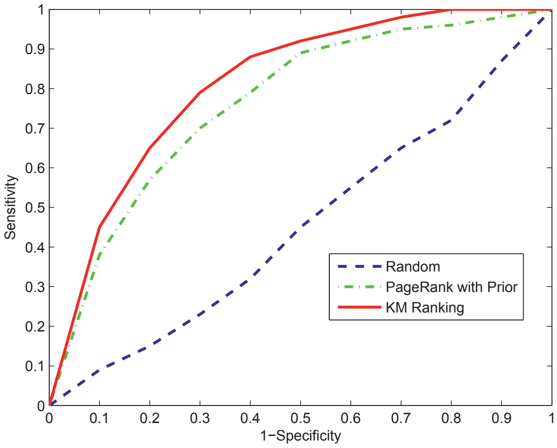

Figure 3. ROC curves for the KM Ranking method and the PageRank with Priors method on systemic scleroderma.

Our method has additional advantages that are not directly shown in this evaluation. These include its ability to seamlessly incorporate multiple knowledge sources, and its ability to perform inferences without having to re-learn the model when the seed genes change.

Table I
AUCs OF VARIOUS PRIORITIZATION METHODS.

| **Diseases** | **PageRank** | **KM Ranking** |
|---|---|---|
| Alzheimer's disease | 0.795 | 0.848 |
| Autism | 0.810 | 0.852 |
| Grave's disease | 0.815 | 0.858 |
| Migraine | 0.805 | 0.840 |
| Systemic scleroderma | 0.811 | 0.853 |

# V. Related Work

Many of the existing gene prioritization methods use functional annotations and a commonly used source of functional annotations is the Gene Ontology (GO) [10]. For example, PROSPECTR [1] uses genes sequence information, and SUSPECTS [2] additionally uses annotation data from GO. ENDEAVOUR [3] uses several data sources such as GO annotations, protein-protein interactions, regulatory information, expression data, and sequence based data. Disadvantages of using functional annotations include incompletely annotated data sources since many disease associated genes are yet to be functionally characterized.

Our method is flexible in that it can incorporate multiple sources of knowledge including free text from research document collections, which can complement the functional annotation information. Some recent studies [13] have explored using multiple sources of knowledge and have attempted to find optimal weights to combine the evidence from different sources. In our method, heterogeneous data sources can be combined at the network level and we do not need to learn various weights for the model; thus the model is less complex in terms of combining data sources.

## VI. Conclusion and Future Work

We have presented a probabilistic knowledge model learned from a protein-protein interaction database and applied it to the task of gene prioritization. We showed that our method can reduce the prioritization error by 6% when compared to a recently published PageRank with Priors method when applied to five diseases. To the best of our knowledge this is the first study that attempts to learn probabilistic relations among genes from protein-protein interaction data and performs 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 multiple times when the seed genes change.

In this study, we only used only a single source of knowledge to learn the model for a fair comparison. But the model is robust and flexible enough to integrate knowledge from various sources and combine them with existing document retrieval methods. In future work, we plan to explore the utility of integrating different sources of knowledge such as the research literature and curated databases.

## 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] J. Sun, P. Jia, A. H. Fanous, B. T. Webb, E. J. C. G. van den Oord, X. Chen, J. Bukszar, K. S. Kendler, and Z. Zhao. A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases - schizophrenia as a case. *Bioinformatics*, 2009.

[14] 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.

[15] S. Wang and M. Hauskrecht. Improving biomedical document retrieval by mining domain knowledge. In *FLAIRS '09: Proceeding of the 22nd international FLAIRS conference*, 2009.

[16] 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.