The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009
Copyright © 2009 ORSC & APORC, pp. 319–328

Gene Prioritization for Type 2 Diabetes in Tissue-specific Protein Interaction Networks*

Biao-Bin Jiang[1] Ji-Guang Wang[2] Jing-Fa Xiao[3]
Yong Wang[2,†]

[1]School of Chemical Engineering & Environment, Beijing Institute of Technology,
Beijing 100081, China
[2]Academy of Mathematics and Systems Science, Chinese Academy of Science,
Beijing 100190, China
[3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China

Abstract

Computationally prioritizing disease genes by large-scale bio-experimental data can provide important insights into the underlying mechanism of complex diseases. Here, we explore the topology of the protein-protein interaction network and apply the PageRank algorithm to identify candidate genes relating to type 2 diabetes. Importantly, a novel idea is introduced to rank the disease genes in tissue-specific protein-protein interaction networks instead of the global protein interaction network. To this end, we extend the original PageRank algorithm by adopting a block-based strategy. The leave-one-out cross validation is conducted to evaluate the performances of all ranking algorithms. The resulting ROC curves show that the proposed method with tissue-specific information performs better than original PageRank algorithm in the global protein-protein interaction network and each subnetwork of single tissue. Finally, four candidate genes are highlighted for further experimental validation due to their higher scores.

Keywords PageRank, protein-protein interaction network, tissue-specificity, type 2 diabetes

1 Introduction

With the completion of the Human Genome Project and the great advances in high-through biotechnologies, biology has become a data-intensive discipline. Thus a great challenge for biologists is to uncover the underlying mechanisms of complex biological activities behind vast experimental data. It gives rise to the close cooperation between biologists and data analysts with backgrounds in computer science, mathematics or physics, and the birth of a new interdiscipline, bioinformatics. A typical subject in bioinformatics is to identify key disease genes among a large amount of candidate genes by computational analysis of bio-experimental data such as gene expression profiles [12] and protein-protein interaction networks (PPINs) [3]. The aim of this study is to help biologists to

*This work is partly supported by the National Natural Science Foundation of China (No. 10801131) and the Beijing Natural Science Foundation (No. 1092011).

†Corresponding author. E-mail: ywang@amss.ac.cn

identify and highlight potential disease gene candidates for further bio-experimental validation.

During recent decades, several gene prioritization methods have been developed [17, 16, 8, 1, 2]. Roughly speaking, these methods can be divided into two categories according to their input data. The first class is text mining and genomics data mining based methods. For instance, GeneSeeker [17] is a web tool that selects candidate genes of interest based on expression and phenotypic data from both human and mouse through several online databases. eVOC system [16] performs candidate gene selection based on the association between each eVOC anatomy term and the disease name according to their co-occurrence in PubMed abstracts through a combination of text and data-mining. DPG (Disease Gene Prediction) [8] and PROSPECTS [1] require only basic sequence information to classify genes as likely or unlikely to be involved in disease. The extended version of PROSPECTS, SUSPECTS [2], is built by integrating annotation data from Gene Ontology (GO), InterPro and expression libraries. More details about the above methods have been reviewed by Nicki Tiffin *et al.* [15].

The second class is biological network based methods. For example, Ma *et al.* [9] develop a system for gene prioritization associated with a phenotype by Combining Gene expression and protein-protein Interaction network (CGI) using Markov random field theory. Another paradigm is CANDID [5] which is designed to rank candidate genes by eight evaluative criteria (publication, protein domains, conservation, expression, interaction, linkage, association, and custom). The other attractive research tendency is to borrow analytical methodology from other network study (like the internet and the social networks) to explore biological networks since all these networks share several common characteristics such as scale-free and small-world properties [19]. For instance, Morrison *et al.* [10] design a gene prioritization system to analyze microarray data by using the famous PageRank algorithm [11], which is developed by Google's founders Larry Page and Segrey Brin to rank the importance of web pages on the internet. Furthermore, Chen *et al.* [4] employ three Web networks algorithms, PageRank, HITS and k-step Markov, to identify disease genes in protein-protein interaction networks.

However, we note that there is still plenty of room for further improvement of gene prioritization methods and it is in pressing need to achieve more accurate and convincing results. A plausible way is to consider disease-related information on tissue level in gene prioritization since some diseases possess tissue-specific traits. And it is irrational to rank all disease candidate genes in a single list on the whole PPIN since some of genes may localize and function in different tissues though they involve in the same pathogenesis. That is the basic start point of our work in this paper.

In this paper, we propose an approach motivated from PageRank and BlockRank [6] algorithms to prioritize Type 2 Diabetes (T2D) candidate genes using tissue-specific protein interaction networks which are assembled through the analysis of gene expression in human cells or tissues [14]. Firstly, we apply the PageRank in the prioritization of all T2D candidate genes on the global PPIN obtained from Bossi's literature [3]. Then we propose ensemble weighting and BlockRank algorithm to rank T2D candidate genes co-expressed in five T2D-related tissues respectively, pancreas, pancreatic islets (insulin secretion), liver, skeletal muscle, and adipose tissue (insulin action). They are all main diabetes-related metabolic environments and well demonstrated in authoritative clinical literature [13, 18]. Finally, we optimize the parameters of the algorithms and compare the performance of our method with other alternative methods in literature [4].

2 Materials and methods

2.1 Data sources

The tissue-specific protein-protein interaction data used in this study was downloaded from literature [3] at the Molecular Systems Biology website. It comprises 80,922 physical interactions that occur among 10,229 human proteins [3]. To determine the tissue specificity of human protein interactions, Bossi *et al.* used gene expression data from Microarray experiments in [14]. The basic philosophy is that if two genes are co-expressed in a cell, then under some conditions their protein products can physically interact [3], vice versa. As a result, all the co-expression relationships in 79 tissues are marked by binary variables. And a matrix of 80,922 interactions by 79 tissues mathematically represents the tissue-specific protein-protein interaction network.

2.2 Seed genes for type 2 diabetes

We propose a general method which can be widely used to study various complex diseases by highlighting their related candidate genes. In this paper, we use type 2 diabetes as a proof-of-concept study to show the efficiency of our new method. It is well known that Type 2 Diabetes (T2D) is a complex disease with polygenic traits differing from Mendelian diseases with monogenic traits [18]. Thus, it is unrealistic to make any explanation on the pathogenesis of T2D by means of a case-control study on any single genetic variant. Rational study should be performed from the system level such as network-based analysis which asks for the construction of gene networks relating to T2D. Therefore, we query “Type 2 Diabetes” and “non-insulin-dependent diabetes” in Morbid Map of OMIM database at the NCBI website. There are 91 genes matching these records. We manually remove non-coding genes, overlapped genes involved in multiple phenotypes, and absent genes in the tissue-specific PPINs, then left 34 independent known genes as the gold-standard positive dataset (seed genes). Mining the tissue-specific PPINs, we extract 285 immediate interactors (1-order neighbors) of these 34 seed genes as the approximate gold-standard negative dataset.

2.3 PageRank algorithm

The whole procedure of gene prioritization is illustrated in Figure 1A. And Figure 1B displays the network modeling using PageRank algorithm that is described mathematically in this section.

Generally, the protein-protein interaction network is represented as an un-weighted and undirected graph, $G(V,E)$ where proteins (genes) are nodes (vertex) and interactions are edges. For example, a small PPIN consisting of six proteins is shown in Figure 1 B1.

The topology of this PPIN can be formulated as a square symmetric matrix $L = (L_{ij})$ (adjacent matrix of graph G), where $L_{ij} = 1$ if protein p_i can interact with protein p_j, and $L_{ij} = 0$ otherwise (see Figure 1 B2).

From Markov chain perspective, the PPIN can be explained by a probability transition matrix that one protein may interact with other proteins in this network with a certain degree of probability. Each row of the transition matrix represents the connection probability

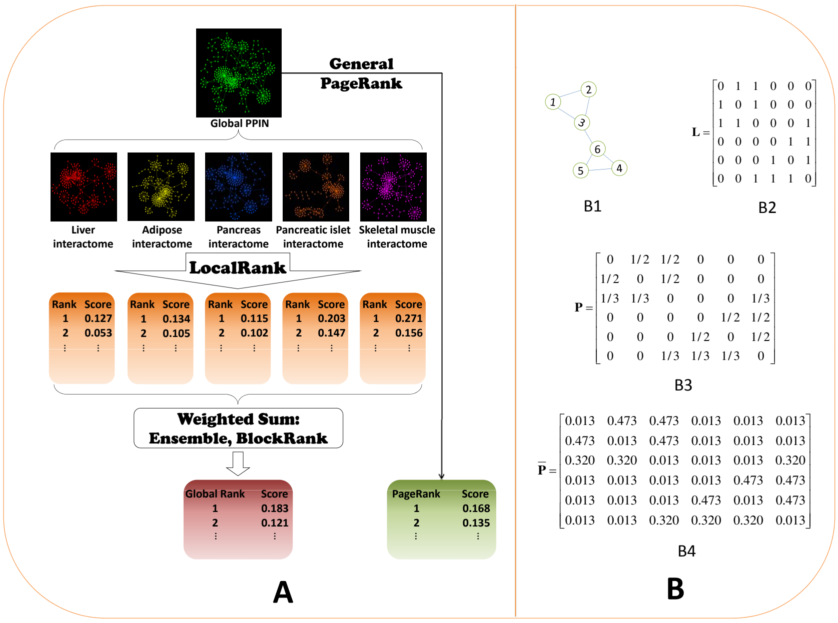
General PageRank
Global PPIN
Liver interactome
Adipose interactome
Pancreas interactome
Pancreatic islet interactome
Skeletal muscle interactome
LocalRank
Rank Score 1 0.127 2 0.053
Rank Score 1 0.134 2 0.105
Rank Score 1 0.115 2 0.102
Rank Score 1 0.203 2 0.147
Rank Score 1 0.271 2 0.156
Weighted Sum: Ensemble, BlockRank
Global Rank Score 1 0.183 2 0.121
PageRank Score 1 0.168 2 0.135
A
1 2 3 4 5 6
B1
L = [0 1 1 0 0 0; 1 0 1 0 0 0; 1 1 0 0 0 1; 0 0 0 0 1 1; 0 0 0 1 0 1; 0 0 1 1 1 0]
B2
P = [0 1/2 1/2 0 0 0; 1/2 0 1/2 0 0 0; 1/3 1/3 0 0 0 1/3; 0 0 0 0 1/2 1/2; 0 0 0 1/2 0 1/2; 0 0 1/3 1/3 1/3 0]
B3
P̄ = [0.013 0.473 0.473 0.013 0.013 0.013; 0.473 0.013 0.473 0.013 0.013 0.013; 0.320 0.320 0.013 0.013 0.013 0.320; 0.013 0.013 0.013 0.013 0.473 0.473; 0.013 0.013 0.013 0.473 0.013 0.473; 0.013 0.013 0.320 0.320 0.320 0.013]
B4
B

Figure 1: (**A**) Schematic overview of our gene prioritization framework. (**B**) Illustration for the ranking algorithm. B1: The graph representation of the example network. B2: Adjacency matrix L. B3: Transition matrix P. B4: Irreducible transition matrix $\bar{P}$

of one protein to other proteins in this network. Any suitable probability distribution may be used across the rows [7]. We use uniform distribution in this work since the PPIN is an unweighted network. Thus, we obtain the transition matrix of Markov model $P = (P_{ij})$ from connection matrix L as follows (see an example in Figure 1 B3):

$$P_{ij} = \frac{L_{ij}}{\sum_j L_{ij}}$$

Since the PPIN has been constructed from seed genes and their neighbors from the whole human protein interactome, it guarantees that there is no isolated protein (indicated as dangling node [7]) in this network. Thus, no row in the matrix P contains all zeros.

However, this transition matrix P cannot ensure the existence of the stationary vector of the Markov chain, i.e., the PageRank vector. If the transition matrix were irreducible, the PageRank vector is guaranteed to exist. Thus, one more adjustment, to make P irreducible, is implemented [7]. The matrix P is revised into an irreducible matrix $\bar{P}$ (see Figure 1 B4) as below

$$\bar{P} = (1-\beta)P + \beta e v^T$$

where $0 \leq \beta \leq 1$. β is the back probability indicating that one web surfer may open a webpage via internet browser other than hyperlinks [7]. It seems that the surfer stop browsing via hyperlinks and reconstruct a root website. v^T is the personalization vector used as the hint of prior knowledge. e is the vectors of all ones. Here, we give the explanation of β from gene prioritization perspective that all proteins are inclined to interact with seed proteins based on the proportion of β in each time step. In our implementation, the element of v^T, v_i equals to $1/34$ if protein i comes from a seed gene (totally 34 seed genes); $v_i = 0$ otherwise, with the elements of v^T summed up as 1.

The aim of solving Markov chains is to compute the stationary vector π^T , which can be viewed as the eigenvector problem [7]:

$$\pi^T \bar{P} = \pi^T$$

where, the ith element of π^T, π_i , is the PageRank score of webpage i which indicates the importance of protein i in our work

We use an iterative method to solve the above equation. For any starting vector $x^{(0)T}$ (generally, $x^{(0)T} = e^T/n$, indicating uniform distribution), we use the following power method to $\bar{P}$ as follows,

$$x^{(k)T} = x^{(k-1)T} \bar{P}$$

We iterate the above formula until the residual $\tau = x^{(k)T} - x^{(k-1)T}$ is less than some predetermined tolerance. In this study, we set the maximal iteration time as 20. Eventually, the stationary vector π^T containing the PageRank score of each protein will be obtained once the residual τ is small enough or total 20 iterations are accomplished.

2.4 BlockRank algorithm

Previous studies rank genes in the whole-genome protein-protein interaction data. However we note that the general PPIN is a composite of tissue-specific protein-protein interaction networks. In this paper, we aim to further investigate the tissue specific structure of the PPIN to improve disease gene identification accuracy. For example, T2D is closely related to five tissues. We need to perform gene prioritization in each T2D-related tissue: pancreas, pancreas islet, liver, adipose, and skeletal muscle. To meet this demand, we use the BlockRank algorithm motivated from basic PageRank.

The basic idea of the BlockRank algorithm is to exploit this structure to speed up the computation of PageRank by a 3-stage algorithm. Suppose there are several blocks in the network. Firstly the local PageRank scores for each blocks are computed independently using the link structure of that block. As a second step these local PageRanks are then weighted by the “importance” of the corresponding block. Finally the standard PageRank algorithm is then run using the weighted aggregate of the local PageRanks as its starting vector. Empirically, this algorithm speeds up the computation of PageRank twice in realistic scenarios.

Specifically our procedure of gene prioritization is summarized as follows [6]:

1. Splitting the PPIN into blocks by tissues.
2. Computing the local PageRank score for each block.
3. Estimating the relative importance of each block (BlockRank scores).

4. Weighting the local PageRank scores in each block by the BlockRank scores of that block, and concatenate the weighted local PageRank of each protein to form an approximate global PageRank vector z.
5. Using z as a starting vector for standard PageRank.

Assuming there are k blocks in the PPIN ($k = 5$ in this study), we need to construct a $k \times k$ transition matrix B where the element B_{ij} represents the transition probability of block i to block j. Firstly we count the number of protein-protein interactions between block i and block j as the initial value of B_{ij} and normalize each row of matrix B as uniform-distributed probability vector with the sum of elements as 1. Then we use standard PageRank algorithm to compute the BlockRank scores of each block with block-based transition matrix B and two uniform-distributed k-vectors as starting vector and personalization vector respectively.

For each node (protein) in the PPIN, the initial value in the starting vector z for global PageRank is its local PageRank scores weighted by its block's BlockRank scores. Computing the Global PageRank, we rank all proteins using the global PageRank scores, and compare this ranking with that generated by general PageRank in cross validation experiment.

In addition, a simple ensemble method is also applied to weight each block based on the percentage of disease genes in each tissue. Simply, a tissue in which more disease genes are co-expressed is expected to be more relevant to the disease progression. Thus, compared with the BlockRank algorithm to computationally weight tissues by the network topology, the ensemble method uses a more straightforward strategy and yields a set of weights directly from seed gene distribution perspective.

3 Results

As a pilot study, we perform a proof-of-concept analysis by comparing our gene prioritization method with other alternative methods. We designed the leave-one-out cross-validation procedure to benchmark these methods. Specifically, we remove one gene from the personalization vector in each test and guarantee the personalization vector is still a probability vector with the sum of its elements as 1. Then we use our proposed method to rank the 34 seed genes and their 285 1-order neighbors together. The receiver operating characteristic (ROC) curve is created for evaluating each ranking algorithm by setting various cutoffs. We choose leave-one-out method since it has relatively smaller variance when gold-standard positive samples are scarce.

After plenty of trials, we choose the back probability as 0.08 in the PageRank algorithm. Firstly we perform block-based PageRank algorithm on tissue-specific PPINs. Then we compare the performance of our methods by designing two control experiments. One control is that we directly apply the standard PageRank algorithm in the global protein-protein interaction network. The second experiment is to apply PageRank algorithm in the five tissue specific protein-protein interaction networks individually. All the ROC curves of BlockRank, standard PageRank in global PPIN, and local PageRank in five tissue-specific PPINs are shown together in Figure 2.

From Figure 2, we can see that the ROC for single tissue outperforms the random control, which clearly demonstrates that every tissue-specific protein interaction network

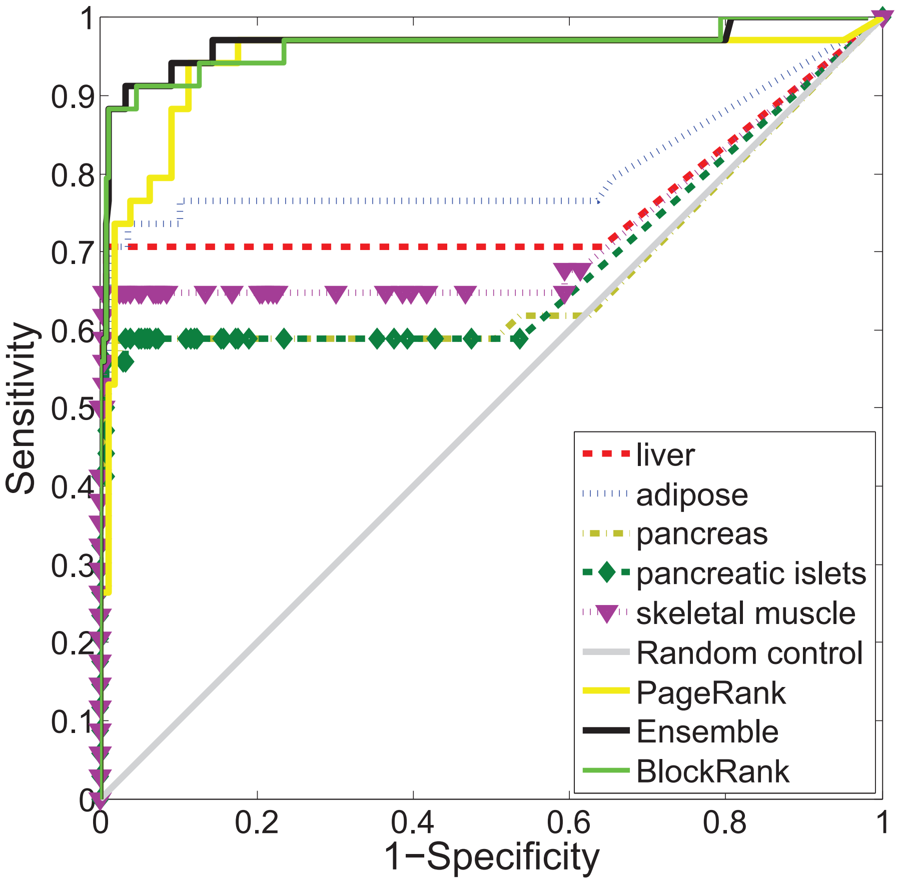

Figure 2: Comparison of different methods for gene prioritization by the ROC curves from leave-one-out cross validation.

is informative regarding to identification of T2D disease genes. Also their accuracies are different tissue by tissue. For example, the gene ranking result in the PPIN of adipose is better than that of pancreatic islets. As is known that adipose tissue is derived from lipoblasts and more closely relates to T2D. Its main function is to store energy in form of fat. Obesity or being overweight in humans and most animals does not depend on body weight but on the amount of body fat-specifically, adipose tissue.

We expect that higher accuracy can be achieved by integrating all these tissue specific PPINs together. Figure 2 shows that directly applying PageRank in the global PPIN achieves a pretty good accuracy by outperforming all the results obtained in any single tissue-specific PPIN, which well demonstrates the effectiveness of the data integration strategy without the tissue-specific information. The underlying reason is that different tissue relates to T2D in different ways. If we comprehensively consider all the tissues in human PPIN, we could have more confidence in disease gene identification.

Importantly, we observe that our BlockRank method in tissue-specific PPINs outperforms the direct application of PageRank in global PPIN. We calculate their area under curve (AUC) values to compare their performances. The AUC values of BlockRank and PageRank are 0.9254 and 0.8930 respectively. Especially when the false positive rate is less than 0.2, we can see significant improvement of our method in the increase of sensitivity. This further demonstrates the effectiveness of our newly introduced idea to consider tissue-specificity in the global PPIN. The possible reason is that we extract the useful information from the T2D related tissues and discard the information in other tissues which may become noise in our gene identification procedure. In the comparison of literature [4], we find that the AUC value of our block-based PageRank algorithm is higher than that of the PageRank with back probability as 0.3 which just achieves the AUC of 0.8 or less. It indicates that the performance of PageRank algorithm is sensitive to the variation of the parameters (back probability and iterative times). Surprisingly, we also see in Figure 2 that the simple ensemble method performs a little bit better than BlockRank weighting, which demonstrates that the utilization of percentage of of disease genes to weight tissues is more informative in gene prioritization than purely using the topology of global protein interaction network.

In Table 1, we list 10 disease genes and 4 candidate genes related to diabetes as an example and their ranks. We want to emphasize that four candidate genes which are deserved further experimental validation. Because they are currently not in the gold-standard positive dataset (seed genes) and they have even higher ranks than some of the disease seed genes.

Table 1: The 14 genes related to diabetes and their ranks by different ranking methods

Gene Name	Seed Label	PageRank	BlockRank	Ensemble	Liver	Adipose	Pancreas	Pancreas islet	Skeletal muscle
INSR insulin receptor	1	1	1	1	1	1	1	1	1
SLC2A4 solute carrier family 2 (facilitated glucose transporter),	1	5	7	2	3	3	3	2	2
IRS2 insulin receptor substrate 2	1	4	3	3	6	4	4	4	3
IRS1 insulin receptor substrate 1	1	2	2	4	2	2	2	180	203
HNF4A hepatocyte nuclear factor 4,	1	3	5	5	5	5	5	5	4
PPARG peroxisome proliferator-activated receptor gamma	1	6	19	7	212	219	6	3	206
TCF7L2 transcription factor 7-like 2 (T-cell specific, HMG-box)	1	18	54	11	216	10	9	8	8
INS insulin	1	7	4	16	209	217	203	7	6
KLF11 Kruppel-like factor 11	1	63	202	30	217	55	212	17	17
ABCC8 ATP-binding cassette, sub-family C (CFTR/MRP), member 8	1	27	59	31	214	221	211	14	16
EIF6 eukaryotic translation initiation	0	12	8	20	22	20	16	16	18
ALDOA aldolase A, fructose-	0	10	9	22	21	18	15	15	219
B2M beta-2-microglobulin	0	36	38	29	50	31	23	23	25
YWHAB tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, beta polypeptide	0	62	16	34	29	32	26	47	76

4 Conclusions

In this paper, we use the tissue-specific PPINs data to implement disease gene prioritization and achieve better accuracy in the preliminary experiments with T2D as an application example. Here, we can conclude that tissue-specific PPINs data can offer deeper insight to disease gene priorities. In the research of complex diseases like T2D, a gene may participate in different pathways and function in different tissues. Simply ranking the importance of gene in a single global network cannot capture this kind of complex situation. In this study, we also verify that the methodological achievements in the internet and the social networks are potentially powerful tools in the research of biological networks. It is expected that more analytical methodologies in complex networks would bring bio-molecular network study a brilliant vision. In the future, one ongoing direction is to improve our method for deeper exploration to T2D or other complex diseases by highlighting more potential disease genes. Another ongoing effort is to develop a general methodology by extensively integrating other data sources.

References

[1] Euan A. Adie, Richard R. Adams, Kathryn L. Evans, David J. Porteous, and Ben S. Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6:55, 2005.

[2] Euan A. Adie, Richard R. Adams, Kathryn L. Evans, David J. Porteous, and Ben S. Pickard. Suspects: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 22(6):773–774, 2006.

[3] Alice Bossi and Ben Lehner. Tissue specificity and the human protein interaction network. *Mol Syst Biol*, 5, April 2009.

[4] Jing Chen, Bruce Aronow, and Anil Jegga. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, 10(1), 2009.

[5] Janna E. E. Hutz, Aldi T. T. Kraja, Howard L. L. Mcleod, and Michael A. A. Province. Candid: a flexible method for prioritizing candidate genes for complex human traits. *Genetic epidemiology*, July 2008.

[6] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Exploiting the block structure of the web for computing pagerank. Technical report, Stanford University, 2003.

[7] Amy N. Langville and Carl D. Meyer. Deeper inside pagerank. *Internet Math.*, 1(3):335–380, 2004.

[8] N. López-Bigas and C. A. Ouzounis. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*, 32(10):3108–3114, 2004.

[9] X. Ma, H. Lee, and F. Sun. Cgi: a new approach for prioritizing genes by combining gene expression and proteinprotein interaction data. *Bioinformatics*, 23(2):215–221, January 2007.

[10] Julie Morrison, Rainer Breitling, Desmond Higham, and David Gilbert. Generank: Using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, 6(1), 2005.

[11] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[12] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–97, Dec 1998.

[13] M. Stumvoll, B. J. Goldstein, and T. W. van Haeften. Type 2 diabetes: principles of pathogenesis and therapy. *Lancet*, 365(9467):1333–1346, 2005.

[14] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101(16):6062–6067, April 2004.

[15] N. Tiffin, E. Adie, F. Turner, H. G. Brunner, M. A. van Driel, M. Oti, N. Lopez-Bigas, C. Ouzounis, C. Perez-Iratxeta, M. A. Andrade-Navarro, A. Adeyemo, M. E. Patti, C. A. Semple, and W. Hide. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res*, 34(10):3067–3081, 2006.

[16] Nicki Tiffin, Janet F. Kelso, Alan R. Powell, Hong Pan, Vladimir B. Bajic, and Winston A. Hide. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Research*, 33(5):1544–1552, 2005.

[17] M. A. van Driel, K. Cuelenaere, P. P. Kemmeren, J. A. Leunissen, and H. G. Brunner. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur J Hum Genet*, 11(1):57–63, 2003.

[18] Jin W. and Patti M.E. Genetic determinants and molecular pathways in the pathogenesis of type 2 diabetes. *Clinical Science*, 116:99–111, 2009.

[19] Scott White and Padhraic Smyth. Algorithms for estimating relative importance in networks. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–275. ACM Press, 2003.