The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

Modularized Random Walk with Restart for Candidate Disease Genes Prioritization*

Xing Chen[1,2,†] Guiying Yan[1] Wei Ren[1]
Ji-Bin Qu[1,2]

[1]Academy of Mathematics and Systems Science, CAS, Beijing 100190,China
[2]Graduate University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Identifying disease genes is very important not only for better understanding of gene function and biological process but also for human medical improvement. Many previous methods are based on modular nature of human genetic disease and the similarity between known disease genes and candidate genes. In this paper, we propose the method of Modularized Random Walk with Restart (MRWR) based on the functional module partition and module importance. Genes are prioritized in each functional module and then the gene ranking in each module is fused into a global ranking in the entire network. MRWR is applied to prostate cancer network. It is surprising that twenty-eight out of top fifty ranking genes are confirmed by PDGB or KEGG or literatures. MRWR significantly improves the performance of previous classical algorithms.

Keywords Candidate disease genes prioritization; Modularized random walk with restart; Functional module partition

1 Introduction

Connecting genotype with phenotype is a fundamental aim of the human genetic studies [1]. Linkage analysis can identify a genomic region which often contains up to hundreds of genes but it is very difficult to exactly find the location of disease genes [2].

Since the Human Genomic Project started in 1990 [3], the molecular bases of 2514 phenotype descriptions have been clarified but still 1743 Mendelian phenotypes or loci unknown [4]. Phenotypically similar diseases are often caused by functionally related genes, being referred to as the modular nature of human genetic disease [5, 6]. So far, most existing methods for candidate disease genes prioritization have been based on this idea. Franke et al constructed a functional human gene network that integrated information on genes and the functional relationships between them [7]. They used this network to rank genes based on their functional interactions. ENDEAVOUR used similarity between disease genes and candidate genes to prioritize candidate genes [8]. Up to now the method has integrated more than 10 types of genomic data. Recently a new method

[*] This work was supported by the National Natural Science of Foundation of China under Grant Nos. 10531070, 10721101, and KJCX-YW-S7.

[†] Corresponding author. Email: xingchen@amss.ac.cn

called CIPHER has integrated human protein -protein interactions (PPIs), disease phenotype similarities, and known gene-phenotype associations to uncover the relation between phenotype and genotype [9]. Kohler et al used the global network distance measure, random walk with restart (RWR) and diffusion kernel (DK), instead of simple direct neighborhood and shortest path measure to prioritize candidate disease genes [10]. Human disease genes were also predicted based on Human-Mouse conserved co-expression analysis and module-based interpretation of phenotypic effects [11].

Actually those known genes causing phenotypically similar diseases are not always in the same functional module in the molecular network. When similarities between candidate disease genes and known disease genes are used to identify potential new disease genes, some candidate genes which are relevant to certain disease may not get a high similarity score when training genes are not in the same functional module with them. In fact all the methods mentioned above will get a good performance unless candidate genes are in the same module with all the training genes. Otherwise some casual genes may not be identified because they are not located in the same functional module with all the training genes. Hence it is very important to rank candidate genes using training genes located in the same module with them. Based on this idea we propose the method of Modularized Random Walk with Restart (MRWR) using the functional module partition and module importance. Genes are prioritized in each functional module and then the gene ranking in each module is fused into a global ranking in the entire network. MRWR is applied to prostate cancer network. It is surprising that twenty-eight out of top fifty ranking genes are confirmed by Prostate Gene DataBase (PDGB) [12] or Kyoto Encyclopedia of Genes and Genomes (KEGG) [13] or literatures. MRWR significantly improves the performance of previous classical algorithms.

2 Materials and Methods

2.1 Data

The prostate cancer network contains 233 genes and 1207 interactions [14]. This gene-interaction network is extracted using automatic literature mining based on Support vector machine (SVM) and dependency parsing. Because another two disease genes (ZFHX3 and HNF1B) have been added to Online Mendelian Inheritance in Man (OMIM) [4] after the publication, these two new disease genes and those genes which interact with new disease genes in Biological General Repository for Interaction Datasets (BioGRID) [15] are included into our network. We also consider the interaction between added genes and previous genes.

2.2 Approach

We choose those genes which are confirmed to be relevant to prostate cancer by OMIM and also located in our network under investigation as training genes. The aim of our algorithm is to obtain more genes associated with prostate cancer. So we seek to get the ranking of non-training genes in each functional module and further give each gene an overall ranking in the entire network to select the most probable disease gene. ENDEAVOUR, Diffusion Kernel (DK) and Random Walk with Restart (RWR) have shown their effectiveness in previous research [8,10]. Here we put forward Modularized Random Walk with Restart (MRWR) method for candidate disease genes prioritization.

Firstly we use the module partition method based on maps of random walks [16]. It is hard for most existing clustering algorithms to automatically determine the number of modules in biological network. The particular advantage of this algorithm is to automatically detect the number of modules in the process of simulated annealing. Moreover the method has been successfully illustrated in scientific communication network and can uncover the community structure in the complex network. There are two important reasons for using this algorithm to implement functional modules partition. One is that the coverage of the gene functional annotation is limit. It has been pointed out that only two thirds of all the genes are annotated by at least one functional annotation [17]. Hence using functional annotation to implement functional modules partition is unreasonable. The other is that PPINs and social networks share scale-free and small-world properties and methods used for both social and Web networks have been successfully used for biological networks, even directly for disease genes prioritization [17].

Then well-known Random Walk with Restart [10] is adopted to prioritize the genes in each functional module. When implementing RWR in certain module k, we don't select all the known disease genes as training genes. Only disease genes in module k will be used as training genes.

Thirdly modules are scored by the module importance measure. The concept of importance here is used for the global prioritization. The high rank in the entire network should be given to genes in the important module. We define the score of the module as the product of the number of training genes and all genes in that module. Most people think disease genes deploy their functions as part of sophisticated functional modules [5,18]. Recent studies have also reported this type of module-based interpretation of phenotypic effects [19]. So we can conclude that disease genes will always appear near the known disease genes and the module which has many known disease genes is likely to have more unknown disease genes. On the other hand, the number of genes in the module is obviously relevant to the importance of module.

Finally the global ranking of candidate genes will be obtained. A new algorithm is proposed to fuse the ranking in every module into a global ranking in the entire network. Algorithm is showed below.

Step 1: It is assumed that we want to get the top N ranking genes in the entire network. M is denoted as the number of modules. We choose the top 1 ranking gene in the most important module as the top 1 ranking gene in the entire network (i=1).

Step 2: We denote m(i,j) as the number of top i ranking genes (global rank) located in the module j and s(j) as the score of the module j (j=1,2,···,M)

We calculate

$$e(i,j) = (i+1)\frac{s(j)}{\sum\limits_{k=1}^{M} s(k)} - m(i,j)$$

and assume the module with the biggest e(i,j) as module p. We choose the gene which is ranked m(i,p) +1 in the module p as the i+1 ranking gene in the entire network.

i=i+1

Step 3: If i is less than N, we return to step 2, otherwise we will end our algorithm.

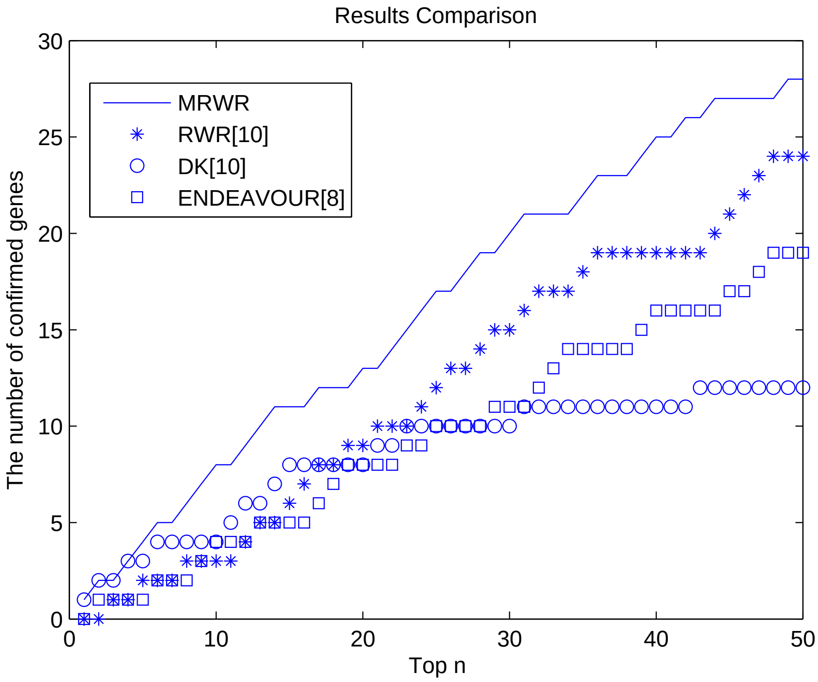

Figure 1: The number of genes confirmed by PGDB or KEGG or literatures in the top n ranking

3 Results

After module partition, known disease genes from OMIM are indeed located in different modules. MRWR is evaluated by examining how many genes are confirmed by PGDB or KEGG or literatures in the top n ranking. In large scale cross-validations, area under the curve (AUC) is an important evaluation measure of gene prioritization methods. Here we propose another measure, which is still the area under the curve. But the curve is the top n curve (the curve describing the relation between n and the number of confirmed genes in the top n ranking) not Receiver Operating Characteristic curve (ROC). The comparison between the result of our algorithms and previous algorithms is showed in Figure 1 [12,13,20-42]. All the genes inferred by our methods and previous classical method are showed in Table 1. These Results (Figure 1 and Table 1) confirm the superiority of our algorithms to previous classical algorithms.

Four out of top five ranking genes (AKT1, BRCA1, MYC and TP53) in the prioritization of MRWR are confirmed by various evidences [12,13,20,22,29,33]. Among the top ten ranking genes in the prioritization of MRWR, there are surprisingly eight true disease genes. On the contrary, classical algorithms only can find at most four confirmed genes. Twenty-eight out of top fifty ranking genes are confirmed by PGDB, KEGG or literatures in the term of MRWR.

Table 1: Genes inferred by various methods

Gene	ENDEAVOUR	DK	MRWR	RWR	Evidence
AKT1	-	-	+	+	[12,20]
APC	-	-	+	+	[12]
ATM	+	+	-	-	[12,21]
BCL2	-	-	+	-	[12,13]
BRCA1	+	+	+	+	[12,22]
CCND1	-	+	+	+	[12,13]
CDK4	+	-	-	-	[23]
CDKN1B	-	-	+	-	[12,13,24]
CDKN2A	-	+	+	-	[12,25]
INS	-	-	+	-	[13,26]
MAPK8	-	-	-	+	[12]
MAPK1	-	-	-	+	[27]
MAX	+	-	-	-	[28]
MYC	-	+	+	+	[12,29]
NCOA3	+	-	-	-	[12]
NR3C1	+	-	+	-	[30]
PDPK1	-	+	+	-	[13,31]
RB1	+	-	-	-	[12,13,32]
TNF	-	+	+	+	[12]
TP53	+	+	-	+	[12,13,33]
VEGFA	-	-	+	-	[12]

'+' indicates that the given gene is ranked within the top 20 and '-' indicates that the gene is not among the top 20 ranking genes inferred by the method.

MRWR is tested under different back probability values to show the robustness of algorithm to the parameter values. The results of MRWR in 11 different test conditions (MRWR with back probability 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9) are listed in Table 2. We observe that the performance of MRWR is robust to the back probability. The best performance may be obtained when back probability is 0.2.

4 Discussion and Conclusion

4.1 Discussion

The shortcoming of MRWR is that we can't find disease genes from those modules which don't have any known disease gene. If we can obtain more training genes, the value of our modularization algorithm will be further showed. The effectiveness of MRWR strongly depends on the construction of disease-specific network. Sometimes modules in complex network are overlapping [43]. A good overlapping clustering algorithm which can automatically determine the number of clusters in the network will give a further modification of our algorithm. If we can get the network integrating more data sources, the result may be even better. Biological information can be converted to the weight of vertices and edges in the network. A good module importance measurement may also improve the results of our algorithm.

Table 2: The number of confirmed genes in the top n ranking under different test conditions

Back probability	10	20	30	40	50
0.01	8	14	17	21	22
0.01	8	14	20	25	27
0.05	8	14	20	24	27
0.1	8	13	20	25	28
0.2	8	13	20	24	29
0.3	8	13	19	25	28
0.4	8	13	19	24	22
0.5	8	12	19	24	28
0.6	8	11	20	23	27
0.7	8	12	19	23	27
0.8	9	12	19	23	27
0.9	9	13	20	24	28

4.2 Conclusion

The similarity measurement is used widely in the prioritization of candidate disease genes. Satisfactory results can't be obtained using traditional algorithms if known disease genes are located in different modules. We propose the Modularization Random Walk with Restart based on module partition, genes prioritization per module and rank fusion.

MRWR is applied to prostate cancer network. Twenty-eight out of top fifty ranking genes are confirmed to be associated with prostate cancer by the PGDB or KEGG or literatures. On the contrary, traditional gene prioritization methods respectively find only 24, 19 and 12 correct genes. We conclude that similarity-based algorithm after being processed by modularization, gene prioritization per module and rank fusion can get much better result than previous algorithm. Our algorithm is a framework which allows integrating many previous similarity-based methods and can improve the results of them. Our validation experiments also show that disease genes are indeed likely to be located in different modules.

Acknowledgement

We would like to thank Prof. Rosvall Martin for giving the code of the module partition algorithm and Prof. Fengzhu Sun for helpful discussion.

References

[1] Botstein,D.et al. Discovering genotypes underlying human phenotypes: past successes for mendelian disease,future approaches for complex diseases. Nat.Genet. 2003,33: 228-237

[2] Botstein,D.et al. Constructing of a genetic linkage map in man using restriction fragment length polymorphisms. Am.J.Hum.Genet. 1980,32: 314-331.

[3] 2003 Release: International Consortium Completes HGP [http://www.genome.gov/11006929]

[4] Hamosh,A.et al. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res,2005,33:514-517.

[5] Oti,M.et al. The modular nature of genetic diseases. Clin Genet,2007,71:1-11.

[6] van Driel,MA.et al. A text-mining analysis of the human phenome. Eur J Hum Genet,2006,14:535-542.

[7] Franke,L.et al. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am.J.Hum.Genet,2006,78:1011-1025.

[8] Aerts,S.et al. Gene prioritization through genomic data fusion. Nat.Biotechnol,2006,24:537-544.

[9] Wu,XB.et al. Network-based global inference of human disease genes. Mol Syst Biol,2008,4:189-199.

[10] Kohler,S.et al. Walking the Interactome for Prioritization of Candidate Disease Genes. Am. J. Hum. Genet, 2008,82:949-958.

[11] Ala,U.et al. Prediction of Human Disease Genes by Human-Mouse Conserved Coexpression Analysis. PLoS Comput Biol,2008,3:e1000043.

[12] Li,LC.et al. Pgdb: a curated and integrated database of genes related to the prostate. Nucleic Acids Res,2003,31:291-293.

[13] Kanehisa,M.et al. The KEGG resource for deciphering the genome. Nucleic Acids Res,2004,23,277-280

[14] Ozgur,A.et al. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics,2008:24,277-285.

[15] Stark,C.et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Res,2006,34:535-539.

[16] Rosvall,M.et al. Maps of random walks on complex networks reveal community structure. Proc Natl. Acad. Sci,2008,105:1118-1123.

[17] Chen,J.et al. Disease candidate gene identification and prioritization using protein interaction networks. BMC bioinformatics,2009,10:73.

[18] Barabasi,A.et al. Network biology: understanding the cell's functional organization. Nat.Rev.Genet,2004,5: 101-113.

[19] Wang,L.et al. Prioritizing functional modules mediating genetic perturbations and their phenotypic effects: a global strategy. Genome Biol,2008:12,537-544.

[20] Chen,X.et al. Constitutively active Akt is an important regulator of TRAIL sensitivity in prostate cancer. Oncogene,2001,42:6073-6083.

[21] Angéle,S.et al. ATM polymorphisms as risk factors for prostate cancer development. Br J Cancer,2004,91:783-787.

[22] Gao,X.et al. Elevated 12-lipoxygenase mRNA expression correlates with advanced stage and poor differentiation of human prostate cancer. Urology,1995,46:227-237.

[23] Agarwal,R.et al. Cell Signaling and Regulators of Cell Cycle as Molecular Targets for Prostate Cancer Prevention by Dietary Agents. Biochem Pharmacol,2000,60:1051-1059.

[24] Chang,BL.et al. A Polymorphism in the CDKN1B Gene Is Associated with Increased Risk of Hereditary Prostate Cancer. Cancer Res, 2004,64:1997-1999.

[25] Isaacs,W.et al. Oncogenes and Tumor Suppressor Genes in Prostate Cancer. Epidemiol Rev,2001,32:779-790.

[26] Ho,G.et al. Polymorphism of the insulin gene is associated with increased prostate cancer risk. Br. J. Cancer,2003,88:263-269.

[27] Hao,F.et al. Lysophosphatidic acid induces prostate cancer pc3 cell migration via activation of lpa(1), p42 and p38alpha. Biochim. Biophys.Acta,2007,1771:883-892.

[28] Ponzielli,R.et al. Cancer therapeutics: Targeting the dark side of Myc. Eur J Cancer,2005,41:2485-2501.

[29] Ellwood-Yen,K.et al. Myc-driven murine prostate cancer shares molecular features with human prostate tumors. Cancer Cell,2003,3: 223-238.

[30] Wei,Q.et al. Global analysis of differentially expressed genes in androgen-independent prostate cancer. Prostate Cancer Prostatic Dis,2007,10:167-174.

[31] David,M.et al. mTOR and cancer: insights into a complex relationship. Net Rev Cancer,2006,6:729-734.

[32] Latil,A.et al. Loss of heterozygosity at chromosome arm 13q and RB1 status in human prostate cancer. Hum Pathol,1999,7:809-815.

[33] DeMarzo,AM.et al. Pathological and molecular aspects of prostate cancer. Lancet,2003,9361:955-996.

[34] Majumder,PK.et al. Akt-regulated pathways in prostate cancer. Oncogene,2005,24:7465-7474.

[35] Palmer,DH.et al. CD40 expression in prostate cancer: a potential diagnostic and therapeutic molecule. Oncol Rep,2004,4:679-682.

[36] Olapade-Olaopa,E O.et al. Evidence for the differential expression of a variant EGF receptor protein in human prostate cancer. Br J Cancer,2000,82:186-194.

[37] Matsuyama,M.et al. Expression of lipoxygenase in human prostate cancer and growth reduction by its inhibitor. Int J Oncol,2004,24: 821-827.

[38] Wang,H.et al. Experimental therapy of human prostate cancer by inhibiting mdm2 expression with novel mixed-backbone antisense oligonucleotides: in vitro and in vivo activities and mechanisms. Prostate,2003,54:194-205.

[39] Linda,B.et al. Constitutive Activation of Stat3 in Human Prostate Tumors and Cell Lines. Cancer Res,2002,62:6659-6666.

[40] Xu,WS.et al. Intrinsic apoptotic and thioredoxin pathways in human prostate cancer cell response to histone deacetylase inhibitor. Proc. Natl Acad. Sci,2006,42:15540-15545.

[41] Chien,J.et al. A role for candidate tumor-suppressor gene TCEAL7 in the regulation of c-Myc activity, cyclinD1 levels and cellular transformation. Oncogene,2008,27:7223-7234.

[42] Stubbs,AP.et al. Differentially Expressed Genes in Hormone Refractory Prostate Cancer: associated with chromosomal regions involved with genetic aberrations. Am. J. pathol,1999,5:1335-1343.

[43] Palla,G.et al. Uncovering the overlapping community structure of complex networks in nature and society. Nature,2005,9:814-818.